Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation

Xin-Yi Tong, Gui-Song Xia, *Senior Member, IEEE*, Fan Hu, Yanfei Zhong, *Senior Member, IEEE*, Mihai Datcu, *Fellow, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

Abstract—Remote sensing (RS) image retrieval is of great significant for geological information mining. Over the past two decades, a large amount of research on this task has been carried out, which mainly focuses on the following three core issues: feature extraction, similarity metric, and relevance feedback. Due to the complexity and multiformity of ground objects in high-resolution remote sensing (HRRS) images, there is still room for improvement in the current retrieval approaches. In this article, we analyze the three core issues of RS image retrieval and provide a comprehensive review on existing methods. Furthermore, for the goal to advance the state-of-the-art in HRRS image retrieval, we focus on the feature extraction issue and delve how to use powerful deep representations to address this task. We conduct systematic investigation on evaluating correlative factors that may affect the performance of deep features. By optimizing each factor, we acquire remarkable retrieval results on publicly available HRRS datasets. Finally, we explain the experimental phenomenon in detail and draw conclusions according to our analysis. Our work can serve as a guiding role for the research of content-based RS image retrieval.

Index Terms—Content-based image retrieval, remote sensing, feature extraction, similarity metric, relevance feedback

1 Introduction

With the explosive development of earth observation technologies, both the quantity and quality of remote sensing (RS) data are growing at a rapid pace [1]. Millions of RS images captured by various satellite sensors have been stored in massive archives [2], [3]. To make full use of big RS data, efficient information management, mining and interpretation methods are urgently needed. During the past decades, significant efforts have been made in developing accurate and efficient retrieval methods to search data of interest from large RS archives [4], [5], [6], [7].

Primal RS image retrieval systems generally used geographical area, time of acquisition or sensor type as queries [8], [9]. These approaches might be very imprecise and inefficient because text-based image retrieval rely largely on manually annotated keywords [10], which are less relevant to the visual content of RS images. As content-based image retrieval [11], [12] was proposed in the early 1990s, the performance of RS image retrieval approaches has been remarkably improved. New architectures for RS image archives were constructed, where RS images were stored [13] and retrieved [14], [15] based on visual content. So far, several mature RS retrieval systems have come into service [1], [2], [3], [10], [16], [17], [18], [19], [20], [21].

Content-based image retrieval takes images as queries, rather than keywords, whose performance therefore is extremely dependent on the visual features [22], [23]. For promoting the accuracy of RS image retrieval, early studies mainly focused on seeking various feature representation methods, hoping to find more discriminating image features [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] or feature combinations [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. Nevertheless, due to the drastically increasing volume and complexity of RS data, visual features may become subjective and ambiguous in some situations [4]. Consequently, the performance of the basic RS retrieval systems was no longer satisfactory. To solve this problem, on the one hand, researchers proposed to select or design the most suitable similarity metric for some specific tasks [54], [55], [56], [57], which can adaptively amend the degree of similarity between image feature vectors. On the other hand, researchers applied relevance feedback to RS retrieval system [1], [2], [3], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], aiming to capture the exact intentions of the users and return retrieval results that meet user demand. As described above, feature extraction, similarity metric and relevance feedback constitute the three core issues of modern RS image retrieval framework.

In this paper, the main issue we focus on is the visual features. According to the approaches of feature extraction, the existing retrieval methods can be divided into three

- *X.-Y. Tong, Y. Zhong, and L. Zhang are with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China E-mail: {xinyi.tong, zhongyanfei, zlp62}@whu.edu.cn.*
- *G.-S. Xia is with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China E-mail: guisong.xia@whu.edu.cn.*
- *F. Hu is with Electronic Information School, Wuhan University, Wuhan 430072, China. E-mail: hfmelizabeth@gmail.com.*
- *M. Datcu is with Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen D-82234, Germany E-mail: mihai.datcu@dlr.de.*

Manuscript received 22 Dec. 2017; revised 25 May 2019; accepted 15 Oct. 2019. Date of publication 23 Oct. 2019; date of current version 29 Aug. 2020. (Corresponding author: Gui-Song Xia).
Recommended for acceptance by xxxx.
Digital Object Identifier no. 10.1109/TBDATA.2019.2948924

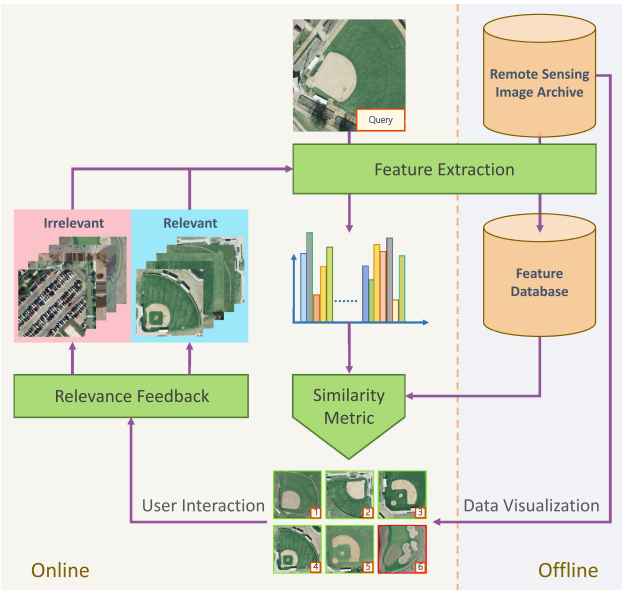


Fig. 1. The framework of content-based RS image retrieval system.

categories: methods based on low-level features, methods based on mid-level features and methods based on high-level features. Low-level features are always designed by human on the basis of engineering skills and domain expertise. Diverse low-level features have been exploited for RS retrieval, mainly including spectral features [24], [25], [26], [31], [32], texture features [30], [33], [34], [39], [40], [41], [42], [44], [47], [70] and shape features [35], [37], [38], [71]. In contrast, mid-level features can represent more discriminating information by encoding raw features using bag-of-words (BoW) [72], Fisher vector (FV) [73], vector locally aggregated descriptors (VLAD) [74] or their variants. However, owing to the changes of photography scale, orientation and illumination, which makes the relevant ground objects have quite different appearance, the above hand-crafted features lose their effectiveness of discriminating high resolution remote sensing (HRRS) images.

To overcome this difficulty, researchers made use of high-level features derived from Convolutional Neural Networks (CNNs) for HRRS image retrieval. In recent literatures [75], [76], [77], CNN features have proved to be of strong discrimination ability and able to dramatically improve the retrieval performance. CNNs are deep hierarchical architectures with parameters (or weights) of each layer learned from large-scale datasets [78]. The pre-trained CNNs can be well transferred to relatively small datasets for feature learning [79], [80]. Nevertheless, when a CNN model trained for classification is used for domain-specific retrieval, its transferability and adaptability to the target data is likely to be unreliable. Various factors involved with transferability may limit the performance of deep feature-based retrieval methods. With this in mind, we intend to further investigate how to better use deep features for content-based HRRS image retrieval task.

We first analyze the retrieval framework and present a comprehensive review on the following three core issues: feature extraction, similarity metric and relevance feedback, so as to complement existing surveys in literatures [4], [5], [6], [7]. Then, we focus on the feature extraction issue and delve into deep features to fully advance the state-of-the-art in HRRS image retrieval. We investigate almost all influencing factors concerned to the property of deep features, including CNN architectures, depth of layers, aggregation method for feature maps, dimension of features and fine-tuning. In addition, we propose multi-scale concatenation and multi-patch pooling methods to further promote the retrieval performance.

In summary, this paper mainly contributes in the following aspects:

- We provide a comprehensive review of content-based RS image retrieval, covering the three key issues: feature extraction, similarity metric and relevance feedback.
- We investigate systemically how to utilize deep features for HRRS image retrieval. Comprehensive influencing factors are assessed, and noteworthy experimental results are achieved on three public HRRS datasets.
- We analyze the experimental phenomena in detail, some of which are generalized, and some are data-dependent. We draw many instructive conclusions from thorough analysis, which can play a guiding role for domain-specific retrieval problems.

2 Review on RS Image Retrieval

In this section, we first make a detailed introduction of content-based RS image retrieval, and then comprehensively review the existing research in this field. We take three key aspects, feature extraction, similarity metric and relevance feedback, into consideration and analyze their role in the retrieval task.

2.1 An Overview of RS Image Retrieval

The goal of content-based RS image retrieval is to find a set of images that contain the content desired by the users from RS archives. We indicate the images stored in database to be retrieved as reference images. If the retrieval system returns reference images containing relevant visual content, we regard them as correct retrieval results.

A content-based image retrieval framework at least consists of two stages [22], [81]. The first stage extracts image features to describe the physical object and scene of both the input query image and reference images. The second stage calculates the visual similarity between the query image and each reference image based on feature vectors, and then returns a ranked list of relevant images ordered by the degree of similarity. Moreover, if visual features and similarity metrics have limited ability to accurately measure the relationship between image contents, relevance feedback can be used to interactively revise the initial ranking [82], [83]. The overall framework of a RS image retrieval system is illustrated in Fig. 1.

To gain an insight into content-based image retrieval of RS imagery, we present a comprehensive review focusing on the above three core issues. Though there has been a few surveys on broad content-based RS image retrieval research [4], [5], [6], [7], they particularly give attention to some single aspects of RS data mining. Our work can serve as a thorough complement for the previous surveys.

2.2 How to Represent RS Images

RS image retrieval methods can be divided into three categories based on the way of feature extraction: methods based on low-level features, methods based on mid-level features and methods based on high-level features. We introduce the relative literatures at length in the following.

2.2.1 Methods Based on Low-Level Features

Visual content can be typically defined by a set of hand-crafted features that describe the spectral, texture or shape information of RS images.

Spectral features are one of the simplest features, yet they describe the most prominent information of RS images [84]. Spectral features have been utilized for various RS retrieval works [24], [25], [26], [31], [32]. They encode the reflectance of the corresponding areas of the Earths surface, resulting in serious sensitivity to noise and illumination change.

Texture features are generally understood as ordered structures composed of a group of pixels [85]. A number of texture features have been applied to RS image retrieval in the form of single feature [30], [33], [34], [39], [40], [41], [42] or combination of multiple features [44], [47], [70]. Commonly used texture features include gray level co-occurrence matrices (GLCM) [85], wavelets [86], [87], Gabor filters [88], [89] and local binary patterns (LBP) [90]. However, they do not fully reflect the essential features of objects because texture is only a characteristic of surface.

Shape features are important cue for content recognition of RS images [35], [37], [38], [71]. They have been used for infrared image retrieval [35] and object retrieval in optimal image [71]. Shape features describe the outline or area information of ground objects, but have little ability to capture their spatial relationship information.

Other types of features are also proposed for RS image retrieval. Scale-invariant feature transform (SIFT) [91] has been proved to be more effective than texture features in scene retrieval [36]. Structural features derived from shape ensembles and relationships [92], [93] also provide satisfactory performance [94], [95], [96]. In addition, researchers have explored combinations of diverse low-level features to improve the retrieval results [17], [18], [19], [20], [43], [45], [46], [48], [49], [50], [51], [52], [53], [97]. Different visual features make up for each others defects, hence their combinations have stronger discriminating ability.

2.2.2 Methods Based on Mid-Level Features

In contrast with low-level features, mid-level features embed raw descriptors into visual vocabulary space and encode feature spatial distribution to capture semantic concepts. Mid-level features are more invariant to changes of scale, rotation or illumination, and they can better represent the complex textures and structures with more compact feature vectors. The general pipeline to extract mid-level features is first obtaining local image descriptors, such as spectral, texture or local invariant features, and then aggregating them into holistic representations using encode methods, e.g., BoW [72], FV [73], and VLAD [74].

BoW [72] is a widely used basic encoding method, it employs k-means clustering to construct visual codebook and counts local features into the histogram of codebook. It has been utilized in some RS image retrieval research and has achieved desired results. Concretely, [98], [99], [100] have shown the effectiveness of encoded features compared with local low-level features.

VLAD [74] is an advanced version of BoW, apart from feature distribution, it additionally counts the distance between local features and cluster centers. VLAD is applied to encode local pattern spectra [101] and obtains high-precision retrieval results on HRRS images [102]. In [103], the experimental results demonstrate that BoW behaves better in calculation speed while VLAD behaves better in indexing accuracy.

Multi-scale spatial information has also been exploited for feature encoding. For instance, spatial pyramid matching based on sparse codes (ScSPM) [104] fuses holistic and local features to enhance the discrimination of mid-level features [105]. Except for the above methods, other unsupervised feature learning methods have also been employed to construct features with higher level of semantic information. Such as auto-encoder [106] and hierarchical neural networks [107].

2.2.3 Methods Based on High-Level Features

The hierarchical architecture of CNN models can simulate very complex nonlinear functions and automatically learn parameters during the training process [78]. Therefore, CNN models are able to capture the essential characteristics of training data so as to represent discriminating visual features [108].

Some RS image retrieval works based on high-level features have been represented up till now. The approaches include obtaining features by existing CNNs from convolutional (conv.) layers or fully-connected (FC) layers [76], fine-tuning off-the-shelf CNN models with domain-relevant datasets [77], or developing tailored CNN architectures [75] and training it with large scale RS dataset [109], etc.

Nevertheless, there is no a comprehensive investigation on deep feature-based HRRS image retrieval. Besides, whether the research conclusions of specific retrieval contexts are transferable to other situations is unknown, since the diversity of data domain has some degree of impact on the feature description. Therefore, how to optimize the performance of deep feature-based HRRS image retrieval is still a problem need to be solved.

2.3 How to Measure Feature Similarity

Similarity metric (or distance function) is a function that defines the distance between visual feature vectors [55], which is one of the foundations of pattern recognition. In a RS image retrieval task, different similarity metric may lead to different ranking results. In [54], eight similarity metrics are investigated. Similarity metrics examined in this work can be divided into two major categories: general feature vector-based metrics and histogram vector-based measures. This work intuitively demonstrates the importance of similarity metrics in retrieval process.

Apart from selecting the appropriate similarity metrics, distance functions can also be manually constructed for specific retrieval situation. For instance, in [55], an informational similarity metric is introduced for compressed RS data mining. In [56], a hyper-spectral image distance is developed. In [57], dictionary-based similarity metrics are

employed for retrieval in different hyper-spectral image datasets, demonstrating the applicability of dictionary-based similarity metrics for hyperspectral image retrieval.

However, manually constructing a similarity metric may be inefficiency and not robust to different data source, metric learning can be an ideal alternative. In contrast to hand-crafted similarity metrics, metric learning is capable of automatically learning distance function for a specific retrieval situation according to task requirement [110], [111], [112], [113]. Unsupervised metric learning has been successfully applied to RS retrieval, for example, [114] models RS images with graphs and uses an unsupervised graph-theoretic method to measure the similarity between the query graph and the graphs of images in the archive. Besides, deep learning-based metric learning approaches have been investigated. In [115], geographic coordinates are treated as weakly supervised information and used to train a triplet network for street view image retrieval.

2.4 How to Optimize Ranking Result

In the case where the visual features are discriminating, and the similarity metric is adaptive, the ranking results of content-based RS image retrieval may still be unsatisfactory [60]. The intelligent feedback techniques therefore become essential for RS retrieval systems.

Relevance feedback can iteratively optimize the retrieval results according to the previous ranking. Once the ranking of the initial retrieval is returned, there are two ways to select a subset of relevant images: automatically selection and manually selection, which are applied to pseudo relevance feedback and explicit relevance feedback respectively [116].

In pseudo relevance feedback [116], the top several returned results are regarded as relevant images, and their features are used for query expansion [105], [117]. Then the fused feature vector is considered as a new query and able to generate more exact ranking list.

In contrast, in explicit relevance feedback [116], the retrieved images are marked as relevant or irrelevant manually by the users at every feedback round. There are three different methods to re-estimate the target query, namely query-point movement and re-weighting method [83], [118], probability distribution-based method [82], [119] and machine learning-based method [120], [121].

The idea of query-point movement and re-weighting method is to adjust the query point in the feature space according to the users feedback, and then use the adjusted query point to re-calculate the ranking list [61].

The probability distribution-based method aims to minimize the probability of retrieving irrelevant images. Specifically, assume there is a mapping from the visual features to the image categories, and the purpose of probability distribution-based relevance feedback is to find the optimal mapping that can minimize the error probability [1], [15], [58], [62], [69].

Machine learning-based relevance feedback can be considered as a binary-classification problem: the relevant retrieved images are positives while the irrelevant retrieved images are negatives [60]. In each iteration, the classifier can be trained with the feedback samples of the current round, or with the combination of the current and the former feedback samples via incremental learning. It returns image ranking according to the category scores derived from the classifier. Commonly used classifiers include decision tree [63], [68], Bayesian networks [66], support vector machine (SVM) [59], [60], [64], [65] and so on.

Apart from the above literatures, there are also many works aiming at improving retrieval efficiency for RS images, including taking advantages of distributed computation [122], applying tree structures [17], [18], [19], [20], [50], [71] and utilizing hash codes [35], [123], [124], [125], [126].

3 Deep Features for RS Image Retrieval

Although some literatures have made advantages of deep features for RS retrieval task, there still no comprehensive research on how to optimize the transferability of CNN models to RS retrieval. With this in mind, we investigate almost all variables concerned to the property of deep features on several public HRRS datasets and analyze the effects of each variable.

An elementary content-based image retrieval framework is at least composed of two stages. For an image dataset which contains N images, the first stage extracts the visual features $\boldsymbol{f}^q$ and $\boldsymbol{f}^r$ respectively from the query image $\boldsymbol{I}^q$ and all the reference images $\boldsymbol{I}^r$, for $r = 1, \ldots, N$. The second stage calculates the distances $D_{qr} = d(\boldsymbol{f}^q, \boldsymbol{f}^r)$ between extracted feature vectors and then ranks retrieved images according to the values of D_{qr}, i.e., the more similar $\boldsymbol{f}^q$ and $\boldsymbol{f}^r$ are, the lower $\boldsymbol{I}^r$ ranks, where $d(\cdot, \cdot)$ stands for a distance function.

Because conv. and FC features derived from different depths in CNN architecture, they are at different representation levels. Conv. features correspond to local responses of every image region, while FC features contain global information of the holistic image, this diversity may lead to different retrieval performance. Moreover, the off-the-shelf CNN models may have limited capability of transferring to RS domain, which is likely to degrade the performance of HRRS image retrieval.

On account of the aforementioned considerations, we make use of five representative CNN models: CaffeNet [127], VGG-M [128], VGG-VD16 [129], VGG-VD19 [129], GoogLeNet [130], and exploit various approaches for feature extraction.

We mainly conduct three schemes: extracting deep features from conv. and FC layers of the pre-trained CNNs respectively, the process of which is illustrated in Fig. 2; and fine-tuning the off-the-shelf CNN models for targeted feature extraction, demonstrated in Fig. 3.

3.1 Scheme (I): Employing Conv. Features

3.1.1 Convolutional Features

When passing an image I through a CNN, the outputs from conv. layers are feature maps, in which each element corresponds to a receptive field of the input image. Suppose the responses of a certain conv. layer form L feature maps and the size of each feature map is $W \times H$. The activations of this layer can be interpreted as $W \times H$ L-dimensional vectors, where the channel number L depends on the inherent structure of CNN, and the spatial resolution $W \times H$ depends on the architecture of CNN, the adopted layer, and the size of image I.

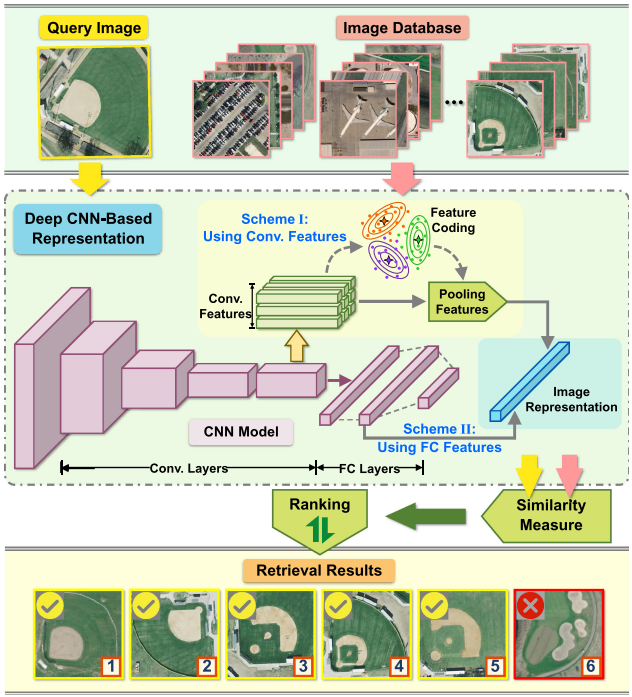


Fig. 2. The pipelines of scheme (I) and scheme (II).

The similarity metric between images is calculated based on their feature vectors, hence it is necessary to aggregate the 3-dimensional feature maps into 2-dimentional feature vectors. We apply several promising aggregation strategies to process feature maps, which can be divided into pooling and encoding methods.

3.1.2 Pooling Methods

We utilize five pooling methods for conv. feature aggregation: max pooling, mean pooling, hybrid pooling [131], sum-pooled convolutional features (SPoC) [132], and cross-dimensional weighting and pooling features (CroW) [133]. When dealing with feature maps using pooling methods, we treat conv. activations as L two-dimensional matrices and each of them composed of $W \times H$ elements: $a^z_{1,1}, a^z_{1,2}, \ldots, a^z_{W,H}$.

– *Max Pooling [131]*: Max pooling generates L-dimensional feature vector $\boldsymbol{f} = [f_1, \ldots, f_z, \ldots, f_L] \in \mathbb{R}^L$, where every element in resulting representation is simply the max activation of a corresponding feature map:

$$f_z = \max_{1 \le x \le W, 1 \le y \le H} a^z_{x,y}. \tag{1}$$

– *Mean Pooling [131]*: Analogous, the L-dimensional output $\boldsymbol{f} = [f_1, \ldots, f_z, \ldots, f_L] \in \mathbb{R}^L$ of mean pooling is a set of average values yielded from corresponding feature maps:

$$f_z = \frac{\sum_{y=1}^{H} \sum_{x=1}^{W} a^z_{x,y}}{WH}. \tag{2}$$

– *Hybrid Pooling [131]*: The feature vector produced by hybrid pooling is the intuitional concatenation of max pooling and mean pooling representations, therefore the hybrid pooling representation is with a dimension of $2L$.

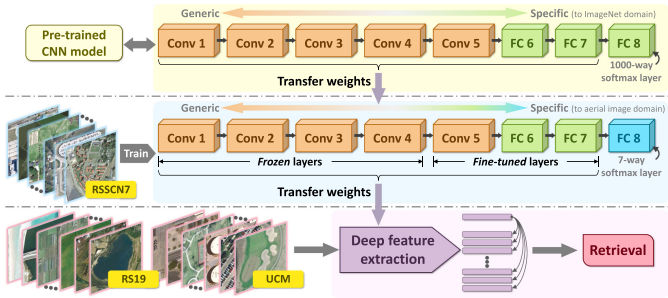


Fig. 3. Overview of fine-tuning for scheme (III).

– *SPoC [132]*: SPoC representation is acquired with center-prior Gaussian weighting on spatial of feature maps as $\boldsymbol{f} = [f_1, \ldots, f_z, \ldots, f_L] \in \mathbb{R}^L$, followed by sum pooling:

$$f_z = \sum_{y=1}^{H} \sum_{x=1}^{W} g_{(x,y)} a^z_{x,y}. \tag{3}$$

The function $g_{(x,y)}$ is constructed based on Gaussian weighting scheme:

$$g_{(x,y)} = \exp\left\{ -\frac{(y - \frac{H}{2})^2 + (x - \frac{W}{2})^2}{2\sigma^2} \right\}, \tag{4}$$

where σ is set to be $1/3$ of the distance between the center and the closest boundary of the input image.

– *CroW [133]*: CroW is a promotion version of SPoC with specific non-parametric schemes for both spatial and channel wise weighting. First, activations with positive values of each feature map channel are counted as:

$$\phi^z = \frac{\sum_{y=1}^{H} \sum_{x=1}^{W} \theta^z_{x,y}}{WH}, \tag{5}$$

where $\theta^z_{x,y} = 1$ if $a^z_{x,y} > 0$, and $\theta^z_{x,y} = 0$ if $a^z_{x,y} \le 0$. The spatial weighting $\alpha_{x,y}$ and channel weighting β^z are defined as follows:

$$\begin{aligned} \alpha_{x,y} &= \frac{\sum_{z=1}^{L} a^z_{x,y}}{\left(\sum_{y=1}^{H} \sum_{x=1}^{W} \left(\sum_{z=1}^{L} a^z_{x,y}\right)^2\right)^{\frac{1}{2}}}, \\ \beta^z &= \begin{cases} \log\left(\sum_{z=1}^{L} \phi^z / \phi^z\right) & \phi^z > 0 \\ 0 & \phi^z = 0. \end{cases} \end{aligned} \tag{6}$$

thereby the final feature $\boldsymbol{f} = [f_1, \ldots, f_z, \ldots, f_L] \in \mathbb{R}^L$ is obtained by weight-summing:

$$f_z = \sum_{y=1}^{H} \sum_{x=1}^{W} \alpha_{x,y} \beta^z a^z_{x,y}. \tag{7}$$

3.1.3 Encoding Methods

We employ three traditional encoding methods to aggregate feature maps into compact feature vectors: bag-of-words (BoW) [72], improved Fisher kernel (IFK) [73] and vector locally aggregated descriptors [74]. When processing feature

maps with encoding methods, conv. activations are treated as $W \times H$ feature vectors: $\{v_{1,1}, v_{1,2}, \ldots, v_{x,y}, \ldots, v_{W,H}\}$, here $v_{x,y} = [a^1_{x,y}, \ldots, a^L_{x,y}] \in \mathbb{R}^L$.

- *BoW [72]*: BoW describes image information with statistics on the spatial distribution of local feature vectors. A codebook of k centroids $\{c_1, \ldots, c_i, \ldots, c_k\} \in \mathbb{R}^{L \times k}$, for $i = 1, \ldots, k$, is learned from local feature set $\{v_{1,1}, v_{1,2}, \ldots, v_{W,H}\}$ via k-means clustering, then every local feature is assigned to its closest centroid. The output of BoW is a k-dimensional vector $\boldsymbol{f} = [f_1, \ldots, f_i, \ldots, f_k] \in \mathbb{R}^K$, where f_i denotes the amount of local features that is assigned to c_i.
- *IFK [73]*: IFK is a combination of generative and discriminative approaches, it utilizes Gaussian mixture model (GMM) with k Gaussian components to construct a probability density distribution of local features. Parameters of GMM is denoted as $\lambda = \{\omega_i, \mu_i, \Sigma_i\}, i = 1, \ldots, k$, where ω_i, μ_i and Σ_i are respectively the mixture weight, mean vector and covariance matrix of Gaussian distributions. Then the L-dimensional gradient vectors $\mathcal{G}_{\mu,i}$ and $\mathcal{G}_{\Sigma,i}$, which are separately with respect to the mean vector and covariance matrix of ith Gaussian component, are derived based on the feature set $\{v_{1,1}, v_{1,2}, \ldots, v_{W,H}\}$. The final IFK representation is a $2L \times k$-dimensional vector indicated as $\boldsymbol{f} = [\mathcal{G}_{\mu,1}, \mathcal{G}_{\Sigma,1}, \ldots, \mathcal{G}_{\mu,k}, \mathcal{G}_{\Sigma,k}] \in \mathbb{R}^{2L \times k}$.
- *VLAD [74]*: VLAD is similar to BoW, yet it considers the statistical distribution of local features as well as the vector difference between local features and centroids simultaneously. The set of feature vectors $\{v_{1,1}, v_{1,2}, \ldots, v_{x,y}, \ldots, v_{W,H}\}$ is clustered into a codebook $\{c_1, \ldots, c_i, \ldots, c_k\} \in \mathbb{R}^{L \times k}$ of k visual words with k-means, where $i = 1, \ldots, k$. A local feature $v_{x,y}$ is assigned to its nearest visual word $c_i = NN(v_{x,y})$ and the vector difference $v_{x,y} - c_i$ between them is recorded and accumulated, ultimately, a VLAD descriptor with a dimension of $L \times k$ is represented as $\boldsymbol{f} = [\sum_{NN(v_{x,y})=c_1}(v_{x,y} - c_1), \ldots, \sum_{NN(v_{x,y})=c_k}(v_{x,y} - c_k)] \in \mathbb{R}^{L \times k}$.

For both the query and reference images, before aggregating their conv. activations into global descriptors, we preprocess the feature maps with l_2-normalization. Since the dimensionality of the pooling and encoding features are different, for fair comparison, we compress aggregated feature vectors to unified dimensions with PCA dimensionality reduction. The final image representations are l_2-normalized again with the purpose of stronger robustness against noise.

3.1.4 Multi-Scale Concatenation

We propose to utilize the fusion of conv. features derived from different scales to enhance the discriminating ability of image descriptors.

The query image and database images are first resized to a sequence of different scales and then separately input to CNN model to obtain multi-scale deep features. For every image, each set of feature maps are encoded into a feature vector. Finally, we simply concatenate the multi-scale feature vectors of a same original image into a long feature vector. These fused features are with multiplied dimensions due to vector concatenation, thus we compress them into a unified low dimension using PCA before retrieval.

3.2 Scheme (II): Employing FC Features

3.2.1 Full-Connected Features

After removing the softmax layer, the rest portion of a CNN can be regarded as a feature extractor. In contrast with conv. layers, which can process images with any size and aspect ratio, FC layers can only process images with a fixed size. Inputting an image $\boldsymbol{I}$, FC layer straightforwardly generates single vector with a settled dimension (shown in Fig. 2), which depends on the inherent structure of CNN model.

We make use of the first two FC layers of all examined CNN models for layer comparison. To achieve higher effectiveness, we also preprocess FC activations with l_2-normalization, PCA reduction and another l_2-normalization procedure prior to retrieval.

3.2.2 Multi-Patch Pooling

We implement multi-patch pooling method to promote the discriminating ability of deep features with multi-position information.

We crop patches with the required size of CNN models at the center and four corners from an input image. Then the horizontal, the vertical, and the horizontal-vertical reflections of these five patches are gathered, thereby, total 20 subpatches are generated from each image. We pass those subpatches one by one through FC layer to extract multi-patch feature vectors, which are with similar form of feature maps but without spatial distribution relationship. We aggregate such 20 feature vectors into holistic feature using pooling method, but note that SPoC and CroW cant be adopted because they are both spatial weighting based. Likewise, multi-patch pooling features are PCA reduced at end.

3.3 Scheme (III): Fine-Tuning Off-the-Shelf CNN Models

Fine-tuning is a supervised retraining process that improves the performance of CNNs for domain specific recognition tasks. The process of fine-tuning is first initializing the CNN model except for the softmax layer with parameters learned on the source training set, next updating the parameters of part or full CNN model via stochastic gradient descent (SGD) using the target retraining set. The dimension of the softmax layer of fine-tuned CNNs is same as the number of retraining dataset's categories, and the remaining architecture is identical with that of the original CNNs. For saving computing resources, we only fine-tune the last conv. layer of each CNN model. Specifically, for GoogLeNet, we fine-tune all 6 conv. layers between the penultimate Inception and the last Inception, as well as the last FC layer.

Concretely, we first change the original 1000-dimensional softmax layer into a Gaussian distributively initialized softmax layer that contains t nodes, where t is the number of retraining dataset's classes. Then, we perform SGD to update the weights of CNN models. The fine-tuning hyper-parameters are set as follows: epoch number 20; mini-batch size 50; momentum 0.9; initial learning rate 0.1, which is decreased to 0.05, 0.005 and 0.001 when every 5 epochs are iterated. The variation of learning rate mitigates the validation error when

it tends to convergence. Moreover, dropout is applied in FC layers, and the activations are randomly set to be zero with probability 0.5 to address the problem of overfitting.

4 Experimental Setup

We perform experiments on three publicly available HRRS image datasets: RS19 [94], RSSCN7 [134] and UCM [135]. Two standard retrieval measures are used to evaluate the results: ANMRR [136] and MAP [23].

4.1 HRRS Image Datasets

- RS19 [94]:The High-resolution Satellite Scene dataset is constituted by 19 categories of satellite scene images with a size of 600×600 pixels collected on diverse orientations and scales from Google Earth. Each category contains slightly different numbers of images for a total of 1005.
- RSSCN7 [134]: The Remote Sensing Scene Classification dataset is composed of 7 categories of typical scene images with a size of 400×400 pixels gathered from Google Earth. Each category contains 400 images, which are sampled on 4 different scales with 100 images per scale.
- UCM [135]: The UC Merced Land Use/Land Cover dataset comprises 21 categories of land-use aerial images with a size of 256×256 pixels selected from aerial orthoimagery. Each category includes 100 images, each of which has a pixel resolution of 30 cm.

In view of the relatively large scale of RSSCN7 [134], we choose it as retraining dataset. Note that fine-tuned CNN features will be correlated with the class information of the retraining data. If such features are adopted for retrieval, the retrieval system would incline to return images with the same class label as that of the query image. However, in unsupervised retrieval task, category labels are with no practical meaning and only used for accuracy assessment. To avoid evaluation bias, we don't conduct retrieval experiment on RSSCN7 with the fine-tuned CNNs.

4.2 Standard Retrieval Measures

- ANMRR [136]: The average normalized modified retrieval rank (ANMRR) takes into account the number of ground truth items and the ranks obtained from the retrieval. Note that ANMRR takes values between 0 and 1, and lower value of ANMRR indicates better retrieval performance.
- MAP [23]: The mean average precision (MAP) is the most common tool to evaluate the rank positions of all ground truth. The average precision (AvePr) for a single query image I is the mean over the precision scores of each relevant item. Different from ANMRR, the value of MAP and the performance of retrieval system are positive correlated.

4.3 Preprocessing and Parameter Settings

In multi-scale concatenation scheme, we define each dataset to three scales as follows: for RS19, scale1, scale2 and scale3 are 300×300 pixels, 600×600 pixels and 1200×1200 pixels respectively; for UCM, scale1, scale2 and scale3 are 256×256 pixels, 512×512 pixels and 1024×1024 pixels separately. For conv. feature aggregation, the number of K-means clustering centroids is empirically set to be 1000 and 100 respectively for BoW and VLAD, and the number of Gaussian components in the GMM for IFK is empirically set to be 100. Apart from Section 5.3, the similarity measure we use in experiments is euclidean distance.

5 Results and Analyses

In this section, we present the results of experiments and analyze how the variables affect the retrieval performance. The variables include architecture of CNN model, depth of CNN layer, aggregation method for feature map, dimension of feature vector and fine-tuning.

In all of the following experiments, CNN layers are denoted using their numerical orders, such as "*conv5*", "*conv5_3*" referring to conv. layers and "*fc6*", "*fc7*" referring to FC layers.

5.1 Convolutional Layers

5.1.1 Aggregation Method

We examine the performance of different aggregation methods for conv. layers. All feature vectors are compressed to be 32-dimensional using PCA before similarity calculation.

The performance comparisons of different aggregation methods and different CNN models are shown in Table 1. It can be clearly observed that IFK outstands among all aggregation methods, and GoogLeNet normally outperforms other CNNs based on its deep architecture.

5.1.2 Dimensionality Reduction

We use hybrid pooling and IFK on feature maps generated from the last conv. layers, and then reduce each feature vector with PCA to some continuously changed dimensions: $\{8, 16, 32, 64, 128, 256, 512, \cdots\}$. The dimension of hybrid pooling features is the maximum among all pooling features (it is the concatenation of max pooling and mean pooling) so that hybrid pooling enables us to test on a wider range of varying dimensions.

We plot the change curves of MAP for different PCA compression rates in Fig. 4, where "OD" denotes that PCA compression is not performed. Apparently, the best accuracies of all datasets and methods are achieved in the range of 16-64 dimensions, since the redundant information is discarded along with the secondary components.

5.1.3 Depth of Conv. Layer

We use IFK to encode feature maps extracted from conv. layers and compress the feature vectors to 32 dimensions uniformly. Fig. 5 shows MAP value of the corresponding conv. layers.

It demonstrates that deeper layers usually perform better, since activations obtained from deeper layers correspond to bigger receptive fields, which contain more information of the original image. But different from image classification, performance of retrieval is not always optimized by deeper

TABLE 1
Comparison of Different Aggregation Methods

(a) RS19

Aggregation Method		CaffeNet		VGG-M		VGG-VD16		VGG-VD19		GoogLeNet	
		ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)
Pooling	*Max Pooling*	0.353	56.90	0.316	61.49	0.286	64.38	0.288	64.29	0.274	65.51
	Mean Pooling	0.389	53.03	0.386	53.43	0.280	65.57	0.293	63.83	0.250	68.42
	Hybrid Pooling	0.350	57.31	0.316	61.59	0.284	64.65	0.286	64.57	0.269	66.05
	SPoC	0.411	50.53	0.412	50.21	0.294	64.17	0.312	61.84	0.263	67.18
	CroW	0.346	57.87	0.334	59.30	0.238	70.37	0.246	69.39	0.246	69.05
Encoding	*BoW*	0.319	61.22	0.298	63.52	0.209	73.85	0.210	73.73	**0.168**	**78.49**
	IFK	**0.244**	**69.64**	**0.233**	**71.52**	**0.190**	**76.51**	**0.188**	**76.59**	0.174	77.93
	VLAD	0.270	66.35	0.260	68.02	0.232	71.59	0.232	71.45	0.277	64.84

(b) RSSCN7

Aggregation Method		CaffeNet		VGG-M		VGG-VD16		VGG-VD19		GoogLeNet	
		ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)
Pooling	*Max Pooling*	0.422	46.27	0.396	49.41	0.408	47.51	0.403	47.94	0.388	49.94
	Mean Pooling	0.388	50.10	0.377	51.03	0.394	49.22	0.382	50.44	0.367	52.42
	Hybrid Pooling	0.420	46.56	0.396	49.44	0.407	47.61	0.402	48.05	0.386	50.17
	SPoC	0.387	49.94	0.382	50.17	0.392	48.82	0.383	49.71	0.392	49.30
	CroW	0.379	51.14	0.398	49.02	0.380	50.67	0.371	51.66	0.370	52.12
Encoding	*BoW*	0.378	51.54	0.375	51.87	0.368	52.29	0.360	53.22	0.354	53.86
	IFK	**0.345**	**55.51**	**0.338**	**55.79**	**0.352**	**54.01**	**0.336**	**55.61**	**0.346**	**54.97**
	VLAD	0.381	51.53	0.395	49.90	0.379	51.09	0.376	51.52	0.423	45.97

(c) UCM

Aggregation Method		CaffeNet		VGG-M		VGG-VD16		VGG-VD19		GoogLeNet	
		ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)	ANMRR	MAP(%)
Pooling	*Max Pooling*	0.469	44.92	0.444	47.60	0.385	53.71	0.390	53.19	0.387	53.13
	Mean Pooling	0.535	38.75	0.495	42.22	0.413	50.81	0.416	50.22	0.381	53.94
	Hybrid Pooling	0.468	45.05	0.443	47.67	0.384	53.83	0.389	53.29	0.384	53.49
	SPoC	0.532	38.61	0.477	43.39	0.384	53.25	0.385	52.79	**0.339**	**58.50**
	CroW	0.493	42.85	0.473	44.63	0.376	54.94	0.379	54.36	0.349	57.26
Encoding	*BoW*	0.485	43.52	0.450	46.71	0.372	55.34	0.371	55.19	0.349	57.44
	IFK	**0.422**	**50.27**	**0.417**	**50.40**	**0.343**	**58.30**	**0.351**	**57.58**	0.367	55.03
	VLAD	0.484	43.49	0.471	44.41	0.425	49.35	0.414	50.39	0.498	39.85

All features are compressed into 32 dimensions. The lower is the value of ANMRR the better is the accuracy and that for MAP is opposite.

layers, especially for the CNNs which are very deep in structure. This is because the receptive fields of extremely deep layers are with considerable large scales and unable to grasp the image details.

5.2 Full-Connected Layers

5.2.1 Dimensionality Reduction

We as well investigate the effect of dimensionality reduction on FC features. It can be seen in Fig. 6 that the optimized dimensions of all datasets are in the range of 8-32. This demonstrates that PCA compression is also effective for FC features in performance improvement.

5.2.2 Depth of FC Layer

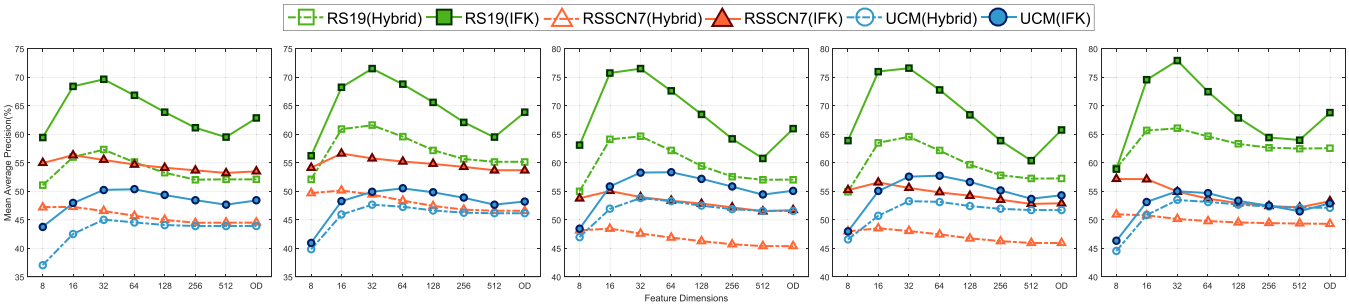
As with conv. layers, we perform retrieval test on FC layers. Fig. 6 clearly demonstrates that the peak of MAP of every dataset is achieved by the lower layer whatever the CNN model is. It is verified again that the deeper layers are not always the better since representations from deeper layers may be too semantically specific to the pre-training natural dataset.

5.2.3 Convolutional versus Full-Connected Layers

We pick out layers offering the highest MAP from both conv. and FC layers, as well as the dimension which wins out for each selected layer. A comprehensive assessment is presented in Table 2. The aggregation method is IFK. The results show that, in most cases, features from conv. layers are more outstanding on the RS19 and RSSCN7 datasets though features from FC layers works better on UCM.

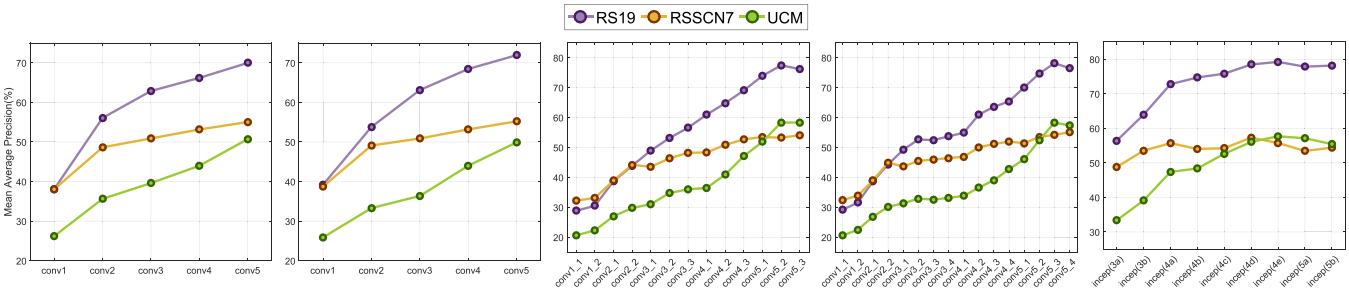
For further verification, we separately show top 5 images retrieved by the best conv. and FC layers of GoogLeNet at the opposite sides of a dotted line in Fig. 7. Where correct results are surrounded by green rectangle while red denotes wrong.

The results can be explained based on the characteristics of the data. Query images from RS19 and RSSCN7 are mainly covered by texture and massive structure, for example, blocky structure in farmland and canopy texture in forest.



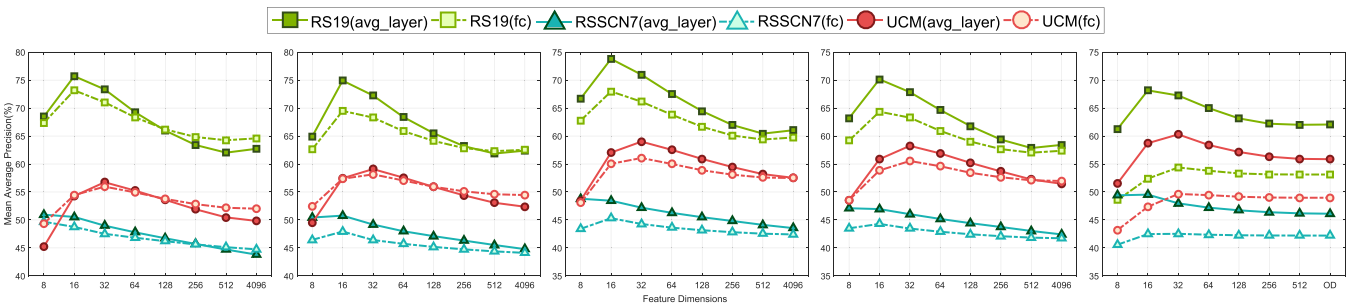
(a) CaffeNet (b) VGG-M (c) VGG-VD16 (d) VGG-VD19 (e) GoogLeNet

Fig. 4. Performance of varying feature dimensions of both hybrid pooling and IFK. The best results are obtained with feature dimensions in the range of 16-64, showing that PCA compression can promote retrieval accuracy and reduce computational cost.



(a) CaffeNet (b) VGG-M (c) VGG-VD16 (d) VGG-VD19 (e) GoogLeNet

Fig. 5. Performance of different conv. layers. Feature maps are aggregated with IFK, and the final feature vectors are compressed to 32 dimensions. It can be observed that intermediate or higher conv. layers produce better results.



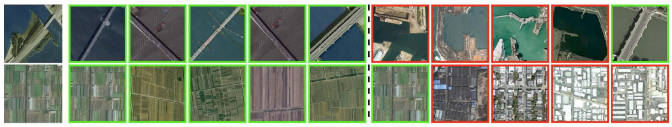
(a) CaffeNet (b) VGG-M (c) VGG-VD16 (d) VGG-VD19 (e) GoogLeNet

Fig. 6. Performance of different FC layers with different dimensions. "OD" means the original dimensions, which are 1024 and 1000 for *avg_layer* and *fc* separately. PCA reduction also helps for accuracy promotion for FC layer, and the highest MAP is achieved by the relatively lower FC layers.

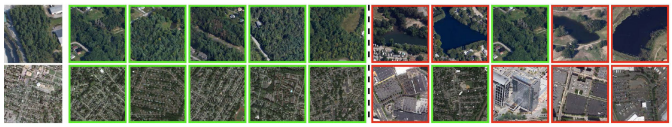
TABLE 2
Comparisons Between Conv. and FC Layers

Nets	RS19				RSSCN7				UCM			
	Layer	Dim	ANMRR	MAP(%)	Layer	Dim	ANMRR	MAP(%)	Layer	Dim	ANMRR	MAP(%)
CaffeNet	conv5	32	0.241	69.90	conv5	16	0.341	55.78	conv5	32	0.416	50.73
	fc6	16	0.190	75.70	fc6	8	0.376	50.94	fc6	32	0.364	56.74
VGG-M	conv5	32	0.230	71.94	conv5	16	0.337	55.99	conv5	64	0.419	50.51
	fc6	16	0.197	74.91	fc6	16	0.383	50.77	fc6	32	0.340	59.09
VGG-VD16	conv5_2	16	0.174	78.52	conv5_3	16	0.340	55.30	conv5_2	32	0.350	58.34
	fc6	16	0.203	73.79	fc6	8	0.393	48.79	fc6	32	0.339	59.00
VGG-VD19	conv5_3	16	**0.163**	**79.48**	conv5_4	16	0.331	56.12	conv5_3	64	0.349	58.44
	fc6	16	0.232	70.14	fc6	8	0.409	47.09	fc6	32	0.345	58.25
GoogLeNet	incep(4e)	32	0.165	79.24	incep(4d)	16	**0.314**	**59.04**	incep(4e)	64	0.349	57.97
	avg_layer	16	0.243	68.16	avg_layer	16	0.387	49.53	avg_layer	32	**0.320**	**60.29**

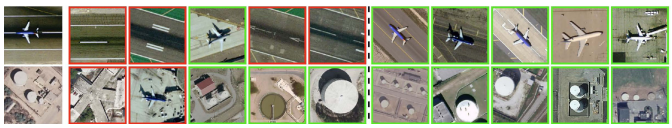
Feature maps of conv. layers are aggregated by IFK. The lower is the value of ANMRR the better is the accuracy, that for MAP is opposite.



(a) RS19



(b) RSSCN7



(c) UCM

Fig. 7. Retrieval results produced by conv. and FC layers of GoogLeNet.

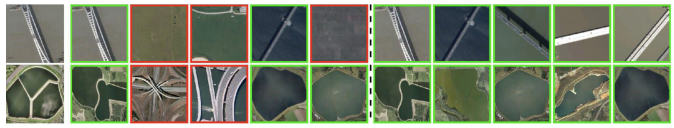
However, the most important elements in query images from UCM are ground targets: airplane and storage tanks. FC layers focus on global semantic information, whereas conv. layers extract the information from local regions. If we take away airplanes from UCM images in Fig. 7c, the results of conv. layers can be regarded to be better than FC layers on account of the very similar runway background. Since conv. features describe the structured information better than abstract semantic information, they are inadaptable to object-oriented dataset, such as UCM.

5.3 Fine-Tuning CNN Model

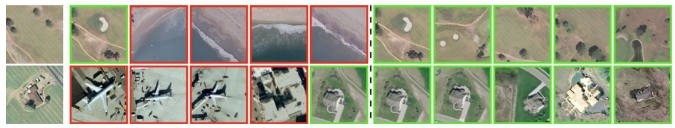
We retrain the last conv. layer and all FC layers of CNNs using RSSCN7 and then test modified models on RS19 and UCM.

The quantitative evaluation is shown in Table 3, we use the last conv. layer and the first FC layer of each CNN model and apply PCA to reduce image representations to 32 dimensions. IFK is the aggregation method for conv. feature maps on account of its prominent performance shown in Table 1. It can be observed that all fine-tuned models produce better MAP on both test datasets whether conv. layers or FC layers are used.

Furthermore, we display two sets of qualitative retrieval results in Fig. 8, from left to right displayed the top 5 images retrieved with original and fine-tuned GoogLeNet. Since the best performance on RS19 and UCM is derived by different layers in Table 3, here we specially use *inception (5b)* for RS19 and *avg_layer* for UCM. The results show that the retrained CNN model performs better.



(a) RS19



(b) UCM

Fig. 8. Retrieval results produced by original and fine-tuned GoogLeNet.

In Fig. 8a, it can be seen that pre-trained GoogLeNet erroneously returns a few viaduct images when querying by a pond image. This is because pond and viaduct images contain similar cross structures. And after fine-tuning GoogLeNet with HRRS images, the modified convolution filters are capable of capturing more specific semantic information from HRRS images. In Fig. 8b, the retrieved beach images have similar curve textures with the query meadow image. Although structure and texture information can always represent the intrinsic properties of natural images, HRRS images formed by similar structure and texture are likely to contain completely different semantic properties, so that fine-tuning can reinforce the transferability of CNNs for RS image retrieval.

5.4 Multi-Scale Concatenation and Multi-Patch Pooling

5.4.1 *Multi-Scale Concatenation*

We resize both query image and reference images into continuously varying scales: scale1, scale2, and scale3, and then pass them through fine-tuned GoogLeNet to obtain multi-scale feature maps from *inception(5b)*. And features with different scales are concatenated for more informative image representations, referred to as scale(1,2), scale(1,3), scale(2,3) and scale(1,2,3). Take note that the encoding method we apply here is BoW instead of IFK. This is because BOW achieves great performance similar to IFK for GoogLeNet in Table 1 and generates feature vectors with a much lower dimension (for *inception(5b)*, IFK features have a dimension of 204800, whereas BoW features are 1000-dimensional). Single and multiple scale feature vectors are uniformly compressed to be 32-dimensional by PCA.

Table 4 presents resulting MAP on RS19 and UCM. An explanation for the experimental phenomena is that the scale of receptive field varies with the scale of input images. CNN

TABLE 3
Comparisons Between Original and Fine-Tuned CNNs on RS19 and UCM With MAP

Dataset	Layer Type	CaffeNet		VGG-M		VGG-VD16		GoogLeNet	
		Original	Finetuned	Original	Finetuned	Original	Finetuned	Original	Finetuned
RS19	Conv.	69.90	70.13	71.94	72.51	76.18	77.20	78.16	**80.21**
	FC	73.35	75.47	72.25	75.55	70.97	75.79	67.25	72.96
UCM	Conv.	50.73	51.71	49.91	52.06	58.30	59.96	55.44	57.82
	FC	56.74	58.93	59.09	61.99	59.00	61.97	60.29	**62.23**

IFK is applied for feature aggregating. Dimensions of all features are compressed to 32.

TABLE 4
Performance of Multi-Scale Concatenation Method Applied on *Inception (5b)* of Fine-Tuned GoogleNet

Dataset	Single Scale			Multiple Scales			
	Scale1	Scale2	Scale3	Scale(1,2)	Scale(1,3)	Scale(2,3)	Scale(1,2,3)
RS19	78.95	81.08	68.47	**82.53**	69.67	72.84	73.70
UCM	**58.69**	50.04	35.30	52.97	36.14	37.83	38.49

TABLE 5
Performance of Multi-Patch Pooling Method Applied on *avg_Layer* of Fine-Tuned GoogleNet

Dataset	Full-size Image	Multiple Patches		
	Single feature	*Max pooling*	*Mean pooling*	*Hybrid pooling*
RS19	72.96	75.62	**76.80**	76.01
UCM	62.23	63.57	**64.56**	63.93

TABLE 6
Comparison With the Current Methods

Descriptors	Dim	Similarity Metrics			
		euclidean	Cosine	Manhattan	Chi-square
CCH+RIT+FPS_1+FPS_2[70]	62	0.640	-	0.589	0.575
CCH+RIT (BoW) [99]	128	0.640	-	0.613	0.585
Salient SIFT (BoW) [98]	128	0.607	0.607	0.591	0.599
Dense SIFT(VLAD) [103]	25600	-	0.460	-	-
Pyramid LPS-aug [102]	-	0.472	-	-	-
Manual RF VGG-M [76]	4096	0.316	0.316	0.333	0.315
Fine-tuned VGG-M [75]	4096	0.299	-	-	-
GoogLeNet(finetuned)+BoW	1000	0.423	0.423	0.685	0.639
GoogLeNet(finetuned)+MultiPatch	1024	0.314	0.314	0.323	**0.309**
GoogLeNet(finetuned)+BoW+PCA	32	0.335	0.335	0.337	-
GoogLeNet(finetuned)+MultiPatch+PCA	32	**0.285**	**0.285**	**0.303**	-

Several distance measures are evaluated with ANMRR, which indicates better performance with lower value.

activations focus more on global information of images with finer scale, but capturing local details from large scale. It makes sense again that RS19 and UCM are quite different in characteristics. RS19 is sensitive to low level visual features, such as edges, texture and graph structure, while UCM tends to be object-oriented. Hence the combination with enlarged scales detracts the discriminating ability of image representations on UCM.

5.4.2 Multi-Patch Pooling

We crop 20 sub-patches of 224×224 pixels from corner and center of each input image and extract deep features from *avg_layer* of retrained GoogLeNet. Sets of feature vectors extracted from different locations are aggregated into compact representations via max pooling, mean pooling and hybrid pooling. All features are reduced into a dimension of 32 using PCA.

As displayed in Table 5, multi-patch pooling significantly boosts MAP value compared to the second column, which shows retrieval accuracies of single feature vectors. And the best results on RS19 and UCM are both acquired by mean pooling.

5.5 Comparison With the Current Methods

We compare our proposed schemes with the recent HRRS image retrieval methods in Table 6. Because relative works are almost all assessed on UCM, we only compare the accuracies on UCM. We select methods yielding the highest MAP on UCM from Tables 5 and 4. Comparative methods are based on hand-crafted features or basic deep features.

We evaluate the performance of retrieval using ANMRR on several distance metrics: euclidean, Cosine, Manhattan and χ^2-square. On account of χ^2-square distance's computational condition that elements of the feature vectors must be non-negative, it cannot be used for features that are compressed by PCA.

It is can be seen that our methods outperform all of others. Especially, the overall best accuracy is acquired by compressed multi-patch mean pooling with euclidean distance, achieving ANMRR value of 0.285, which is about 1.4 percent better than the recent CNN-based method [75]. Apart from the precision, the feature dimension of our method is the lowest, which notably decreases the computation cost.

6 Conclusion

We comprehensively reviewed the existing research works on content-based RS image retrieval and explored how to use CNNs to address this issue with systematical experiments. We took exhaustive influencing variables into account and performed experiments on three public HRRS image datasets with five representative CNN models. By optimizing and

analyzing these variables, we achieved outstanding retrieval performance on the examined HRRS image datasets and drawn many instructive conclusions.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61922065, 61771350, 61871299, and 41820104006, in part by the Outstanding Youth Project of Hubei Province under Contract 2017CFA037.

REFERENCES

[1] M. Datcu, K. Seidel, S. D'Elia, and P. Marchetti, "Knowledge-driven information mining in remote-sensing image archives," *E. S. A. Bull.*, no. 110, pp. 26–33, 2002.

[2] M. Datcu et al., "Information mining in remote sensing image archives: System concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2923–2936, Dec. 2003.

[3] H. Daschiel and M. Datcu, "Information mining in remote sensing image archives: System evaluation," *IEEE Trans. Geosc. Remote Sens.*, vol. 43, no. 1, pp. 188–199, Jan. 2005.

[4] M. Quartulli and I. G. Olaizola, "A review of eo image information mining," *ISPRS J. Photogrammetry Remote Sens.*, vol. 75, pp. 11–28, 2013.

[5] M. Datcu, S. d'Elia, R. L. King, and L. Bruzzone, "Introduction to the special section on image information mining for earth observation data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 795–798, Apr. 2007.

[6] M. Datcu, R. L. King, and S. D'Elia, "Introduction to the special issue on image information mining: Pursuing automation of geospatial intelligence for environment and security," *IEEE Geosc. Remote Sens. Lett.*, vol. 7, no. 1, pp. 3–6, Jan. 2010.

[7] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, 2018.

[8] H. Lotz-Iwen and W. Steinborn, "The intelligent satellite-image information system isis," in *Proc. AIP Conf. Proc.*, 1993, vol. 283, no. 1, pp. 727–734.

[9] C. Chang, B. Moon, A. Acharya, C. Shock, A. Sussman, and J. H. Saltz, "Titan: A high-performance remote sensing database," in *Proc. 13th Int. Conf. Data Eng.*, 1997, pp. 375–384.

[10] G. B. Marchisio, W.-H. Li, M. Sannella, and J. R. Goldschneider, "Geobrowse: An integrated environment for satellite image retrieval and mining," in *Proc. IGARSS.*, pp. 669–673.

[11] C. Faloutsos et al., "Efficient and effective querying by image content," *J. Intell. Inf. Syst.*, vol. 3, no. 3/4, pp. 231–262, 1994.

[12] V. N. Gudivada and V. V. Raghavan, "Content-based image retrieval systems - guest editors' introduction," *IEEE Comput.*, vol. 28, no. 9, pp. 18–22, Sep. 1995.

[13] K. Seidel, R. Mastropietro, and M. Datcu, "New architectures for remote sensing image archives," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Proc. Remote Sens. - A Sci. Vis. Sustainable Develop.*, pp. 616–618.

[14] K. Seidel, M. Schroder, H. Rehrauer, G. Schwarz, and M. Datcu, "Query by image content from remote sensing archives," in *Proc. IGARSS.*, pp. 393–396.

[15] M. Datcu and K. Seidel, "Image information mining: exploration of image content in large archives," in *Proc. Aerosp. Conf. Proc.*, 2000, vol. 3, pp. 253–264.

[16] K. Koperski, G. Marchisio, S. Aksoy, and C. Tusk, "VisiMine: Interactive mining in image databases," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 1801–1812.

[17] K. W. Tobin et al., "Large-scale geospatial indexing for image-based retrieval and analysis," in *Proc. Int. Symp. Visual Comput.*, pp. 543–552.

[18] K. W. Tobin et al., "Automated feature generation in large-scale geospatial libraries for content-based indexing," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 531–540, 2006.

[19] M. Klaric, G. Scott, C.-R. Shyu, C. Davis, and K. Palaniappan, "A framework for geospatial satellite imagery retrieval systems," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2006, pp. 2457–2460.

[20] C.-R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "Geoiris: Geospatial information retrieval and indexing system content mining, semantics modeling, and complex queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 839–852, 2007.

[21] I. M. G. Muñoz and M. Datcu, "System design considerations for image information mining in large archives," *IEEE Geosc. Remote Sens. Lett.*, vol. 7, no. 1, pp. 13–17, Jan. 2010.

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

[23] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: an experimental comparison," *Inf. Retrieval*, vol. 11, no. 2, pp. 77–107, 2008.

[24] A. Vellaikal, C. J. Kuo, and S. K. Dao, "Content-based retrieval of remote-sensed images using vector quantization," in *Proc. Visual Inf. Process. IV*, 1995, pp. 178–189.

[25] J. E. Barros, J. C. French, W. N. Martin, and P. M. Kelly, "System for indexing multispectral satellite images for efficient content-based retrieval," in *Proc. IS&T/SPIE's Symp. Electron. Imaging: Sci. Technol.*, 1995, pp. 228–237.

[26] G. Healey and A. Jain, "Retrieving multispectral satellite images using physics-based invariant representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 842–848, Aug. 1996.

[27] K. Seidel, M. Schroder, H. Rehrauer, and M. Datcu, "Meta features for remote sensing image content indexing," in *Proc. IGARSS*, pp. 1022–1024.

[28] M. Datcu, K. Seidel, and M. Walessa, "Spatial information retrieval from remote-sensing images. I. information theoretical perspective," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 5, pp. 1431–1445, Sep. 1998.

[29] M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu, "Spatial information retrieval from remote-sensing images. II. gibbs-markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 5, pp. 1446–1455, Sep. 1998.

[30] M. Schröder, M. Walessa, H. Rehrauer, K. Seidel, and M. Datcu, "Gibbs random field models: A toolbox for spatial information extraction," *Comput. Geosci.*, vol. 26, no. 4, pp. 423–432, 2000.

[31] T. Bretschneider and O. Kao, "A retrieval system for remotely sensed imagery," in *Proc. Int. Conf. Imaging Sci., Syst., Technol.*, vol. 2, pp. 439–445, 2002.

[32] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2253–2255.

[33] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on gabor texture feature classification," in *Proc. 7th Int. Conf. Signal Process.*, 2004, pp. 733–736.

[34] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Appl. Opt.*, vol. 43, no. 2, pp. 210–217, 2004.

[35] A. Ma and I. K. Sethi, "Local shape association based retrieval of infrared satellite images," in *Proc. 7th IEEE Int. Symp. Multimedia*, 2005, pp. 551–557.

[36] S. D. Newsam and Y. Yang, "Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery," in *Proc. 15th Annu. ACM Int. Symp. Advances Geographic Inf. Syst.*, Art. no. 9.

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

[38] F. Dell 'Acqua and P. Gamba, "Query-by-shape in meteorological image archives using the point diffusion technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 9, pp. 1834–1843, Sep. 2001.

[39] V. P. Shah, N. H. Younan, S. Durba, and R. King, "Wavelet features for information mining in remote sensing archives," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2005, pp. 5630–5633.

[40] V. P. Shah, N. H. Younan, S. S. Durbha, and R. L. King, "A systematic approach to wavelet-decomposition-level selection for image information mining from geospatial data archives," *IEEE Trans. Geosc. Remote Sens.*, vol. 45, no. 4, pp. 875–878, Apr. 2007.

[41] 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, pp. 083584–083584, 2014.

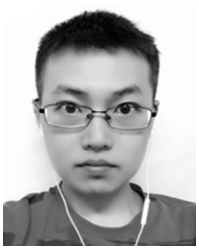
[42] S. Bouteldja and A. Kourgli, "Multiscale texture features for the retrieval of high resolution satellite images," in *Proc. Int. Conf. Syst., Signals Image Process.*, 2015, pp. 170–173.

[43] G. B. Marchisio and J. Cornelison, "Content-based search and clustering of remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 1999, pp. 290–292.
[44] C. Li and V. Castelli, "Deriving texture feature set for content-based retrieval of satellite image database," in *Proc. Int. Conf. Image Process.*, 1997, pp. 576–579.
[45] K. Koperski and G. B. Marchisio, "Multi-level indexing and GIS enhanced learning for satellite imageries," in *Proc. Int. Workshop Multimedia Data Mining*, 2000, pp. 8–13.
[46] 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.
[47] S. D. Newsam and C. Kamath, "Retrieval using texture features in high-resolution multispectral satellite imagery," in *Proc. Int. Soc. Opt. Eng.*, 2004, pp. 21–32.
[48] Y. Li and T. R. Bretschneider, "Semantics-based satellite image retrieval using low-level features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, pp. 4406–4409.
[49] P. Maheswary and N. Srivastava, "Retrieval of remote sensing images using colour and texture attribute," *Int. J. Comput. Sci. Inf. Secur.*, vol. 4, no. 8, pp. 3–15, 2009.
[50] A. Samal, S. K. Bhatia, P. Vadlamani, and D. Marx, "Searching satellite imagery with integrated measures," *Pattern Recognit.*, vol. 42, no. 11, pp. 2502–2513, 2009.
[51] P. Maheshwary and N. Srivastava, "Prototype system for retrieval of remote sensing images based on color moment and gray level co-occurrence matrix," *Int. J. Comput. Sci. Issues*, pp. 20–23, vol. 3, Aug. 2009.
[52] Z. Shao, W. Zhou, and Q. Cheng, "Remote sensing image retrieval with combined features of salient region," *Int. Archives Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 40, no. 6, 2014, Art. no. 83.
[53] H. Sebai, A. Kourgli, and A. Serir, "Dual-tree complex wavelet transform applied on color descriptors for remote-sensed images retrieval," *J. Appl. Remote Sens.*, vol. 9, no. 1, pp. 095 994–095 994, 2015.
[54] Q. Bao and P. Guo, "Comparative studies on similarity measures for remote sensing image retrieval," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2004, pp. 1112–1116.
[55] L. Gueguen and M. Datcu, "A similarity metric for retrieval of compressed objects: Application for mining satellite image time series," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 4, pp. 562–575, Apr. 2008.
[56] M. Graña and M. A. Veganzones, "An endmember-based distance for content based hyperspectral image retrieval," *Pattern Recognit.*, vol. 45, no. 9, pp. 3472–3489, 2012.
[57] M. A. Veganzones, M. Datcu, and M. Grana, "Dictionary based hyperspectral image retrieval," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2012, pp. 426–432.
[58] M. Schröder, H. Rehrauer, K. Seidel, and M. Datcu, "Interactive learning and probabilistic retrieval in remote sensing image archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2288–2298, Sep. 2000.
[59] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818–826, Apr. 2007.
[60] 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.
[61] I. E. Alber, Z. Xiong, N. Yeager, M. Farber, and W. M. Pottenger, "Fast retrieval of multi-and hyperspectral images using relevance feedback," in *Proc. IGARSS*, 2001, pp. 1149–1151.
[62] S. Aksoy, G. Marchisio, K. Koperski, and C. Tusk, "Probabilistic retrieval with a visual grammar," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2002, pp. 1041–1043.
[63] S. Aksoy, K. Koperski, C. Tusk, and G. Marchisio, "Interactive training of advanced classifiers for mining remote sensing image archives," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 773–782.
[64] R. Datta, J. Li, A. Parulekar, and J. Z. Wang, "Scalable remotely sensed image mining using supervised learning and content-based retrieval," Pennsylvania State Univ., State College, PA, USA, Tech. Rep. CSE, pp. 06–019, 2006.
[65] M. Costache and M. Datcu, "Learning-unlearning for mining high resolution EO images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 4761–4764.
[66] Y. Li and T. R. Bretschneider, "Semantic-sensitive satellite image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 853–860, Apr. 2007.
[67] A. S. Barb and C.-R. Shyu, "Visual information mining and ranking using graded relevance assessments in satellite image databases," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 3398–3401.
[68] A. S. Barb and C. Shyu, "Visual-semantic modeling in content-based geospatial information retrieval using associative mining techniques," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 38–42, Jan. 2010.
[69] L. Gueguen, M. Pesaresi, and P. Soille, "An interactive image mining tool handling gigapixel images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 1581–1584.
[70] 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.
[71] G. J. Scott, M. N. Klaric, C. H. Davis, and C. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.
[72] 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, pp. 1470–1477.
[73] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[74] 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.
[75] 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.
[76] P. Napoletano, "Visual descriptors for content-based retrieval of remote sensing images," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1343–1376, 2018.
[77] 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.
[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[79] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 806–813.
[80] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 14–22.
[81] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, 2008, Art. no. 5.
[82] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval," in *Proc. 13th Int. Conf. Pattern Recognit.*, 1996, pp. 361–369.
[83] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "MindReader: Querying databases through multiple examples," in *Proc. 24th Int. Conf. Very Large Data Base*s, 1998, pp. 218–277.
[84] D. Peijun, C. Yunhao, T. Hong, and F. Tao, "Study on content-based remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 2, pp. 4–8, 2005.
[85] R. M. Haralick, K. S. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[86] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.
[87] A. Boggess, F. J. Narcowich, D. L. Donoho, and P. L. Donoho, "A first course in wavelets with fourier analysis," *Phys. Today*, vol. 55, no. 5, 2002, Art. no. 63.
[88] J. G. Daugman, "Complete discrete 2-d gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp. 1169–1179, Jul. 1988.
[89] B. S. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[90] M. Pietikäinen, T. Ojala, and Z. Xu, "Rotation-invariant texture classification using feature distributions," *Pattern Recognit.*, vol. 33, no. 1, pp. 43–52, 2000.
[91] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[92] V. Caselles, B. Coll, and J. Morel, "Topographic maps and local contrast changes in natural images," *Int. J. Comput. Vis.*, vol. 33, no. 1, pp. 5–27, 1999.
[93] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, 2010.
[94] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maitre, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC VII Symp.-100 Years*, 2010, vol. 38, pp. 298–303.
[95] G. Liu, G.-S. Xia, W. Yang, and L. Zhang, "Texture analysis by using shapes co-occurrence patterns," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 1–6.
[96] G. S. Xia, G. Liu, X. Bai, and L. Zhang, "Texture characterization using shape co-occurrence patterns," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5005–5018, Oct. 2017.
[97] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5–1, pp. 2874–2886, May 2013.
[98] Y. Yang and S. D. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.
[99] E. Aptoula, "Bag of morphological words for content-based geographical retrieval," in *Proc. 12th Int. Workshop Content-Based Multimedia Indexing*, 2014, pp. 1–5.
[100] 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.
[101] P. Maragos, "Pattern spectrum and multiscale shape representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 701–716, Jul. 1989.
[102] P. Bosilj, E. Aptoula, S. Lefèvre, and E. Kijak, "Retrieval of remote sensing images with pattern spectra descriptors," *Int. J. Geo-Inf.*, vol. 5, no. 12, 2016, Art. no. 228.
[103] S. Özkan, T. Ates, 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.
[104] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.
[105] 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.
[106] 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.
[107] 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, p. 709, 2016.
[108] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[109] 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.
[110] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu, "Collaborative image retrieval via regularized metric learning," *Multimedia Syst.*, vol. 12, no. 1, pp. 34–44, 2006.
[111] S. C. H. Hoi, W. Liu, and S. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 6, no. 3, pp. 18:1–18:26, 2010.
[112] C. Huang, S. Zhu, and K. Yu, "Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval," *arXiv preprint arXiv:1212.6094*, 2012.
[113] J. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[114] 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.
[115] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.
[116] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA, USA: ACM Press/Addison-Wesley, 1999.
[117] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2491–2521, 2008.
[118] C. Buckley and G. Salton, "Optimization of relevance feedback weights," in *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1995, pp. 351–357.
[119] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan. 2000.
[120] S. Tong and E. Y. Chang, "Support vector machine active learning for image retrieval," in *Proc. 9th ACM Int. Conf. Multimedia*, 2001, pp. 107–118.
[121] L. Zhang, F. Lin, and B. Zhang, "A neural network based self-learning algorithm of image retrieval," *Chin. J. Softw.*, vol. 12, no. 10, pp. 1479–1485, 2001.
[122] L. Mascolo, M. Quartulli, P. Guccione, G. Nico, and I. G. Olaizola, "Distributed mining of large scale remote sensing image archives on public computing infrastructures," *arXiv preprint arXiv:1501.05286*, 2015.
[123] N. Ramesh and I. Sethi, "A model based industrial part recognition system using hashing," in *Proc. 22nd Int. Symp. Ind. Robots, Int. Robots Vis. Automat. Conf*, 1991, pp. 37–51.
[124] I. K. Sethi and N. Ramesh, "Local association based recognition of two-dimensional objects," *Mach. Vis. Appl.*, vol. 5, no. 4, pp. 265–276, 1992.
[125] B. Demir and L. Bruzzo, "Kernel-based hashing for content-based image retrval in large remote sensing data archive," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 3542–3545.
[126] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Feb. 2016.
[127] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[128] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Mach. Vis. Conf.*, 2014, pp. 4–13.
[129] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Representations*, 2015, pp. 7–15.
[130] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[131] A. Mousavian and J. Kosecka, "Deep convolutional features for image based retrieval and scene categorization," 2015, *arXiv preprint arXiv:1509.06033*.
[132] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," 2015, *arXiv preprint arXiv:1510.07493*.
[133] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. Eur. Conf. Comput. Vis. 2016 Workshops*, 2016, pp. 685–701.
[134] 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.
[135] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2010, pp. 270–279.
[136] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

Xin-Yi Tong received the BS degree in remote sensing technology from Wuhan University, Wuhan, China, in 2014, where she is currently working toward the PhD degree in photogrammetry and remote sensing. Her research interests include high-resolution image retrieval, image classification, and deep learning.

Gui-Song Xia (M'10-SM'15) received the PhD degree in image processing and computer vision from CNRS LTCI, Telecom ParisTech, Paris, France, in 2011. From 2011 to 2012, he has been a post-doctoral researcher with the Centre de Recherche en Mathematiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently working as a full professor in computer vision and photogrammetry with Wuhan University. He has also been working as visiting scholar at DMA, École Normale Supérieure (ENS-Paris) for two months in 2018. His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing imaging. He is now serving as an associate editors for several journals, including the *Pattern Recognition*, *Signal Processing: Image Communications*, and the *EURASIP Journal on Image & Video Processing*. He is a senior member of the IEEE.



Fan Hu received the BS degree and the PhD degree in communication engineering from Wuhan University, Wuhan, China, in 2011 and 2017, respectively. He is currently a post-doctoral researcher with the Signal Processing Laboratory, Electronic Information School, Wuhan University. His research interests include high-resolution image classification, and machine learning, especially deep learning and their applications in remote sensing.

Yanfei Zhong (M'11-SM'15) received the BS degree in information engineering and the PhD degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively. He is currently the dean with the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, where he has been a full professor since 2010. He has authored more than 150 research papers, including more than 80 peer-reviewed articles in international journals, such as *ISPRS Journal of Photogrammetry and Remote Sensing*, *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Image Processing*, and *Pattern Recognition*. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data and applications. He was a recipient of the Excellent Young Scientist Foundation selected by the National Natural Science Foundation of China, the National Excellent Doctoral Dissertation Award of China, and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). He was also a Referee of more than 30 international journals. He is serving as an associate editor or an editor for the *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, the *International Journal of Remote Sensing*, and the *Remote Sensing*. He is a senior member of the IEEE.

Mihai Datcu (SM'04-F'13) received the MS and PhD degrees in electronics and telecommunications from the University Politechnica of Bucharest (UPB), Bucharest, Romania, in 1978 and 1986, respectively. Since 1981, he has been a professor in electronics and telecommunications with UPB. Since 1993, he has been a scientist with the German Aerospace Center (DLR), Munich, Germany. From 1991 to 1992, he was a visiting professor with the Department of Mathematics, University of Oviedo, Oviedo, Spain. From 1992 to 2002, he was a longer invited professor with the Swiss Federal Institute of Technology, Zrich, Switzerland. In 1994, he was a guest scientist with the Swiss Center for Scientific Computing, Manno, Switzerland. From 2000 to 2002, he was with Universitouis Pasteur, Strasbourg, France, and International Space University, Strasbourg. In 2003, he was a visiting professor with the University of Siegen, Siegen, Germany. He is currently a senior scientist and the Image Analysis Research Group leader with the Remote Sensing Technology Institute (IMF), DLR, a coordinator of the CNESDLR-ENST Competence Centre on Information Extraction and Image Understanding for Earth Observation, and a professor with the Paris Institute of Technology/GET Telecom Paris. His research interests include Bayesian inference, information and complexity theory, stochastic processes, model based scene understanding and image information mining for applications in information retrieval and understanding of high-resolution SAR, and optical observations. He is a member of the European Image Information Mining Coordination Group. In 1999, he received the title Habilitation á diriger des recherches from Universitouis Pasteur. He is a fellow of the IEEE.

Liangpei Zhang (SM'08-F'18) received the BS degree in physics from Hunan Normal University, Changsha, China, in 1982, the MS degree in optics from the Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xian, China, in 1988, and the PhD degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998. He is currently a Chang-Jiang scholar chair professor with Wuhan University, appointed by the Ministry of Education of China. His research interests include hyper spectral remote sensing, highresolution remote sensing, image processing, and artificial intelligence. He has authored or co-authored more than 500 research papers and five books. He holds 15 patents. He is a fellow of the Institution of Engineering and Technology, an executive member of the board of Governors of the China National Committee of International Geosphere-Biosphere Program, and an executive member of the China Society of Image and Graphics. He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He serves as a co-chair for the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition, the Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geo informatics symposiums. He also serves as an associate editor for the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geo-spatial Information Science*, and the *Journal of Remote Sensing*, and the guest editor for the *Journal of Applied Remote Sensing* and the *Journal of Sensors*. He serves as an associate editor for the IEEE Transactions on Geoscience and Remote Sensing. He is a fellow of the IEEE.