Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network

Zhenfeng Shao, Weixun Zhou, Xueqing Deng, Maoding Zhang, and Qimin Cheng

***Abstract*—Conventional remote sensing image retrieval (RSIR) system usually performs single-label retrieval where each image is annotated by a single label representing the most significant semantic content of the image. In this scenario, however, the scene complexity of remote sensing images is ignored, where an image might have multiple classes (i.e., multiple labels), resulting in poor retrieval performance. We therefore propose a novel multilabel RSIR approach based on fully convolutional network (FCN). Specifically, FCN is first trained to predict segmentation map of each image in the considered image archive. We then obtain multilabel vector and extract region convolutional features of each image based on its segmentation map. The extracted region features are finally used to perform region-based multilabel retrieval. The experimental results show that our approach achieves state-of-the-art performance in contrast to handcrafted and convolutional neural network features.**

***Index Terms*—Fully convolutional networks (FCN), multilabel retrieval, multilabel vector, region convolutional features (RCFs), remote sensing image retrieval (RSIR), single-label retrieval.**

I. Introduction

THE RECENT development in remote sensing (RS) technology has resulted in a considerable volume of RS archives, which makes it a significant challenge of searching images of interest from a large-scale RS archive in the literature. Remote sensing image retrieval (RSIR) is a simple yet effective method to solve this problem. An RSIR system generally has two main parts: feature extraction and similarity measure. The goal of feature extraction is to develop the representation of the images which can be used to measure the similarity later. In similarity measure, the query image is matched with other images in the archive to retrieve the most similar images. However, RS community has been focused mainly on developing discriminative image features due to the fact that retrieval performance greatly depends on the effectiveness of extracted features.

Conventional RSIR approaches are based on low-level visual features extracted either globally or locally. Color (spectral) and texture features are commonly used as global features for RSIR problem. In [1], the morphology-based spectral features are proposed and explored for RS image retrieval. In [2], morphological texture descriptors are computed and combined color channels in [3]. In contrast to global features, local features are generally extracted from image patches of interest, with bag-of-visual-words (BoVW) [4] framework to build the feature representations. An improved color texture descriptor is proposed by incorporating discriminative information and usually achieves better performance than global features. This is due to the fact that local representations are able to narrow down the semantic gap since the RS image content is characterized in a small neighborhood region [5]. As an example, the scale invariant feature transform (SIFT) descriptors are extracted and aggregated by BoVW to generate compact features for RSIR in [6]. Although these RSIR methods mentioned above are able to achieve reliable performance, they are essentially single-label approaches. For single-label RSIR, each image in the archive is labeled by a single, coarse class label. However, the scene complexity of RS images is ignored in this scenario, where an image is likely to have multiple classes (i.e., multiple labels). Single labels are sufficient to address RS problems where the image contains simple content, such as to distinguish between building and river image categories, while multiple labels are required for distinguishing more complex image categories, such as dense residential and medium residential, where the differences only lie in the density of buildings. Thus, in the case of RSIR problem with such complex image categories, multilabel RSIR approaches are needed.

Multilabel analysis is an effective method for image retrieval and classification in computer vision field [7]–[12]. Inspired by these works, RS community has recently focused on developing multilabel approaches to overcome the limitations of single-label RSIR methods. In [13], an image scene semantic matching scheme is proposed for multilabel RSIR, in which an object-based support vector machine classifier is used to obtain classification maps of images in the archive, and in the

Manuscript received August 24, 2019; revised November 25, 2019; accepted December 16, 2019. Date of publication January 8, 2020; date of current version February 12, 2020. This work was supported in part by the National Key Research and Development Plan on Strategic International Scientific and Technological Innovation Cooperation Special Project under Grant 2016YFE0202300, in part by the National Natural Science Foundation of China under Grants 61671332, 41771452, and 41771454, and in part by the Natural Science Fund of Hubei Province in China under Grant 2018CFA007. *(Corresponding author: Weixun Zhou.)*

Z. Shao and M. Zhang are 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; zhangmaoding@whu.edu.cn).

W. Zhou is 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).

X. Deng is with the Electrical Engineering and Computer Science, School of Engineering, University of California at Merced, Merced, CA 95343 USA (e-mail: xdeng7@ucmerced.edu).

Q. Cheng is with the School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: chengqm@hust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2019.2961634

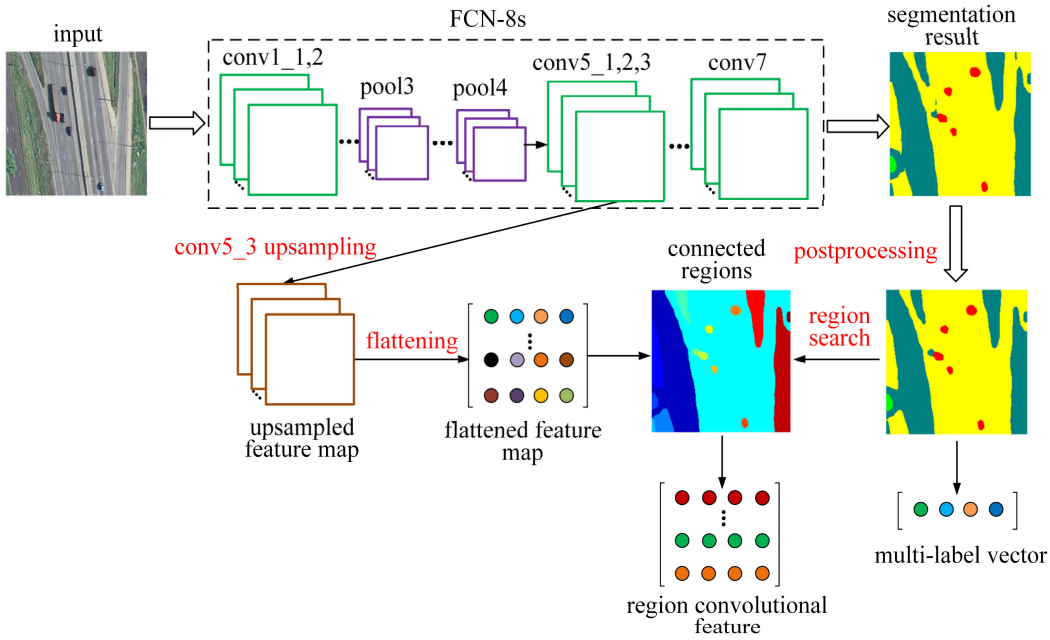


Fig. 1. Process of multilabel analysis and SSRCF extraction. Conv1_1,2 and conv5_1,2,3 represent two (conv1_1, conv1_2) and three (conv5_1, conv5_2, conv5_3) convolution layers, respectively. Each color in the connected regions (also called components or objects) represents one connected region.

other work [14], image visual, object, and semantic features are combined to perform a coarse-to-fine retrieval of RS images from multiple sensors. In [15] and [16], a novel multilabel RSIR system combining spectral and spatial features is presented for hyperspectral image retrieval. In the recent work [17], a semisupervised graph-theoretic method is introduced for multilabel RSIR, in which only a small number of images are manually labeled for training. These approaches, however, greatly depend on the preliminary segmentation results, and generally rely on a set of combined handcrafted features to perform retrieval. There are few study in multilabel RSIR using deep learning and particularly convolutional neural networks (CNN) features which have been proved to be more effective on tasks including image super-resolution [18], hyperspectral imagery classification [19], [20], and RSIR [21]–[26].

In this article, we therefore propose a novel multilabel RSIR method based on fully convolutional network (FCN), which has several advantages over the existing standard multilabel RSIR methods. In general, standard methods such as [17] contain a couple of steps including image segmentation, region feature extraction, label annotation, etc., while our approach integrates these steps in a single framework based on FCN (see Fig. 1) which simplifies the process of multilabel analysis and feature extraction. More importantly, standard methods need to fuse different handcrafted features in order to achieve good performance, while our approach can learn CNN features that are simple yet effective for multilabel RSIR problems. In our approach, we first train a FCN network adapted from the pretrained CNN to generate semantic segmentation map of each image in the considered archive. We then compute multilabel vector and extract region convolutional features (RCFs) of each image based on corresponding segmentation map. For the computation of multilabel vector, it can be directly obtained from the segmentation map, while for the extraction of RCFs, feature map upsampling and flattening, and region search are needed to find the local convolutional features that lie in each connected region. We finally apply the extracted RCFs to perform multilabel retrieval using a region-based similarity measure.

Our contributions in this article are summarized as follows.

1) We propose a novel multilabel RSIR method based on FCN, which combines image segmentation, label annotation, and region feature extraction in a single framework.
2) We propose to extract the learned RCFs instead of handcrafted features for multilabel RSIR.
3) We release a dense labeling RS dataset that can be used for pixel-based problems such as multilabel retrieval, multilabel classification, semantic segmentation, etc., and provide baseline results for multilabel retrieval.

The remaining article is organized as follows. Section II presents our multilabel RSIR approach based on FCN. Section III first introduces the dataset used for training FCN and evaluating multilabel retrieval performance, and then shows the experimental results. Section IV draws the conclusion.

II. Multilabel Retrieval Based on FCN

Multilabel RSIR system can lead to finer retrieval benefiting from image multilabel information and generally contains three parts, i.e., multilabel analysis, feature extraction, and similarity measure. Multilabel analysis and feature extraction are more important compared to similarity measure, since the performance significantly relies on the effectiveness of the extracted multilabel vector and RCFs.

We explore using FCN to perform multilabel analysis and feature extraction. Given an image, FCN is able to predict the label for each pixel which has been widely used in semantic segmentation since [27]. In order to perform pixelwise prediction, FCN consists of downsampling and upsampling paths where semantic or contextual information are captured, and spatial information is recovered, respectively. Features can be extracted from convolutional layers and multilabel prediction can be obtained based on the prediction map from FCN. Therefore, FCN provides a potential solution to achieve our main goals (i.e., multilabel analysis and feature extraction) within one unified model.

A. Architecture of FCN

In practice, FCN is usually adapted from the CNNs pretrained on ImageNet [28], for example, the very deep network (VGG-16) [29] by using convolutional layers instead of fully connected layers. There are three variants of FCN [27]; three networks, namely, FCN-32s, FCN-16s, and FCN-8s have been used in semantic segmentation. In this article, we choose FCN-8s (termed FCN hereafter) for multilabel analysis and RCF extraction since FCN-32s and FCN-16s achieve worse segmentation performance than FCN-8s in our preliminary experiments described in Section III. The better segmentation performance can provide more accurate extracted region and thus better features can be extracted.

We follow the steps in [27] to build our FCN and the details are as followed. The first two fully connected layers of VGG-16 are converted into convolution layers (i.e., conv6 and conv7 layers in Fig. 1). The last fully connected layer (i.e., classifier layer) is modified to output *N* (the number of classes in our dataset) classes, followed by a transposed convolution layer (also inappropriately called deconvolution layer sometimes) to upsample the coarse predictions to dense predictions. The upsampled predictions are fused with the outputs of pool3 and pool4 layers to provide further precision via skip connection. We refer the readers to [11] for more details on how to build FCN with the pretrained VGG-16 network.

To train FCN, our dense labeling dataset is divided into training set D_T and retrieval set D_R. Training set is used for training FCN, and retrieval set is used for multilabel retrieval performance evaluation.

B. Multilabel Analysis

The goal of multilabel analysis is to obtain image multilabel vector achieved by the following two steps: 1) semantic segmentation, and 2) postprocessing, as shown in Fig. 1.

Semantic segmentation refers to generating the segmentation prediction map by feeding through the image in D_R to FCN, and can be denoted as follows:

$$Y_i = f(D_R^i) \tag{1}$$

where D_R^i is the *i*th image in D_R; $f(\bullet)$ is the function that maps image to segmentation map, which is FCN in this article; Y_i is the segmentation map that has the same size as D_R^i, each pixel in Y_i has a label in $\{1, 2, \ldots, N\}$; and *N* is the total number of classes. Once the initial segmentation map is obtained, we use morphological operation and region merging as two postprocessing steps to improve segmentation result. Specifically, morphological opening and closing operations are first used to eliminate small objects, and then each connected region with the area smaller than 10 pixels is merged into its largest neighbor region. Afterward, given a segmentation map, we can build the *N*-D multilabel vector L_i for the corresponding image

$$L_i = [l_1, l_2, \ldots, l_N] \tag{2}$$

where $l_j (j = 1, 2, \ldots, N)$ equals to 1 or 0, indicating whether D_R^i contains the *j*th class. It is noted that by performing semantic segmentation, each image is able to obtain pixelwise labels; however, these labels in total only form one multilabel vector for the image by checking the occurrence of the labels. For example, an image divided into regions with three classes (labels) 1, 3, 4 and five classes in total, and then the multilabel vector will be $[1, 0, 1, 1, 0]$.

C. RCF Extraction

The convolutional layers of CNNs have been demonstrated to generate local descriptors for RSIR [22]. Like SIFT, these local features can also be postprocessed using feature aggregation approaches to generate a compact feature vector. We take inspiration from [22], and propose to extract single-scale and multiscale RCFs (called SSRCF and MSRCF, respectively) for multilabel RSIR.

"Region convolutional features (RCFs) are region-based features extracted from convolutional layers." As aforementioned, segmentation map is not only used to output the multilabel for image but also used to extract the RCFs. But the problem is that the segmentation map is pixel based, which contradicts region. To address the above challenge, we propose to use connected component analysis to group pixels into regions which is termed connected regions. And thus, in general, to extract RCF for both single scale and multiscale, three steps are needed namely region search, feature map upsampling, and feature map flattening. We describe the details for extracting SSRCF and MSRCF in similar approach. The pipeline for the extraction of SSRCF_c5 is shown in Fig. 1.

1) Region Search: This step aims to find n connected regions $R_i = [R_i^1, R_i^2, \ldots, R_i^n]$ in Y_i using a connectivity of 8. This results in that D_R^i is divided into a set of connected regions based on its segmentation map Y_i, which transfers the pixels in a segmentation map into groups. Fig. 2 shows some example images of region search.

2) Feature Map Upsampling: It is impossible to directly extract SSRCF since the size of feature map does not match that of the segmentation map. The feature map from conv layer is downsampled by pooling, resulting in its size reduced. Therefore, to extract SSRCF, the feature map is needed to be upsampled so that it has the same size as segmentation map does. In this article, feature maps from conv5_3, conv6, and conv7 layers are extracted, and the corresponding SSRCF features are termed as SSRCF_c5, SSRCF_c6, and SSRCF_c7, respectively.

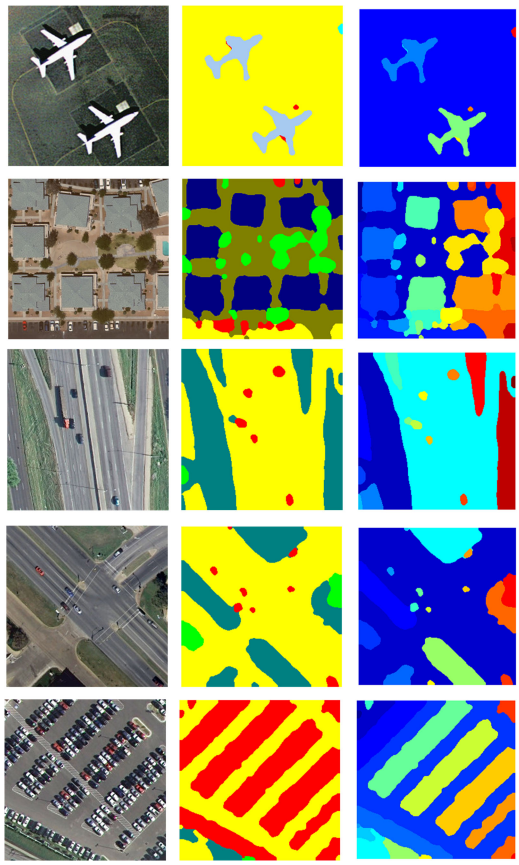


Fig. 2. Some example images of connected region search. (a) Source image. (b) Segmentation map. (c) Eight-connected regions (each connected region is illustrated in different color).

Inspired by the transposed convolution in FCN, we upsample the feature map using the same technique to make it have the same size as D_R^i, and it can be defined as follows:

$$F_S = f_b(F) \tag{3}$$

where F is the feature map of D_R^i extracted from conv5_3 layer, $f_b(\bullet)$ is bilinear interpolation function, and F_S is the upsampled $W \times H \times D$ feature map

$$F_S = \begin{bmatrix} F_S(1,1) & F_S(1,2) & \cdots & F_S(1,W) \\ F_S(2,1) & F_S(2,2) & \cdots & F_S(2,W) \\ \vdots & \vdots & \vdots & \vdots \\ F_S(H,1) & F_S(H,2) & \cdots & F_S(H,W) \end{bmatrix} \tag{4}$$

where W and H are the width and height of feature map, respectively. $F_S(p,q)(p = 1, 2, \ldots, H; q = 1, 2, \ldots, W)$ is a row feature vector with the dimension of D.

3) Feature Map Flattening: This step aims to flatten the upsampled feature map to a two-dimensional feature map which will be overlapped with the connected regions to extract RCFs. The process of flattening the $W \times H \times D$ feature map F_S to $WH \times D$ feature map F'_S is defined as follows:

$$F'_S = [F_S(1,1)\ F_S(1,2) \cdots\ F_S(H,W)]^T. \tag{5}$$

To obtain SSRCF_c5 of D_R^i, we first obtain the convolutional feature matrix $F'_{S,t}$ ($F'_{S,t}$ consists of convolutional feature vectors in F'_S that are located in R_i^t) of region $R_i^t (t = 1, 2, \ldots, n)$ by comparing the pixel coordinates of R_i^t and $F_S(p,q)$, and then SSRCF_c5 can be achieved

$$X_i = [x_1\ x_2 \cdots\ x_n]^T \tag{6}$$

where $x_t (t = 1, 2, \ldots, n)$ is the max convolutional feature vector of R_i^t and is defined as follows:

$$x_t = f_{\max}(F'_{S,t}) \tag{7}$$

where $f_{\max}(\bullet)$ is a function computing the maximum value of each column in $F'_{S,t}$.

Regarding the extraction of MSRCF, it is similar to that of SSRCF, and the difference only lies in that MSRCF is extracted based on multiscale feature map. In this article, we propose two strategies to obtain multiscale feature maps, as shown in Fig. 3. The first strategy is fusing the feature maps of two different layers. In our approach, conv6 and conv7 layers are selected to achieve the multiscale feature map due to the fact that the higher layers of CNN tend to learn better (more powerful) features. Specifically, the feature maps of conv6 and conv7 are first upsampled to have the same size as the input image and then combined to obtain the fused feature map

$$F_{\text{fuse}} = F_{\text{conv6}} + F_{\text{conv7}} \tag{8}$$

where "+" sign represents summation, and F_{fuse} is the fused feature map, and F_{conv6} and F_{conv7} are the upsampled feature maps of conv6 and conv7, respectively. MSRCF extracted based on F_{fuse} is termed as MSRCF_c67. For the second strategy, the multiscale feature maps are directly extracted from two multiscale layers (termed as sum1 and sum2 hereafter) of FCN, as shown in Fig. 3. It can be observed that sum1 fuses the feature maps of pool4 and output layers, and sum2 fuses the feature maps of sum1 and pool3 layers. Therefore, both sum1 and sum2 can extract multiscale feature maps for the extraction of MSRCF. We refer the readers to [27] for more details on these two layers. MSRCF extracted based on feature maps of sum1 and sum2 are termed as MSRCF_s1 and MSRCF_s2, respectively.

D. Similarity Measure and Performance Evaluation

After the extraction of SSRCF and MSRCF, each image in D_R is represented by a RCF matrix. We choose the region-based distance [5] to measure image similarity. Let I_q and I_r be the query image and another image in the archive, respectively; the region-based distance is defined as follows:

$$D(X_q, X_r) = \frac{1}{m} \sum_{i=1}^{m} \min(D_{L_2}^i(x_i, x_j)) \tag{9}$$

where $X_q = [x_1, x_2, \ldots, x_m]^T$ and $X_r = [x_1, x_2, \ldots, x_n]^T$ are the RCFs of I_q and I_r, respectively, $x_i (i = 1, 2, \ldots, m)$ is the feature vector of region R_q^i in I_q, and $x_j (j = 1, 2, \ldots, n)$ is the

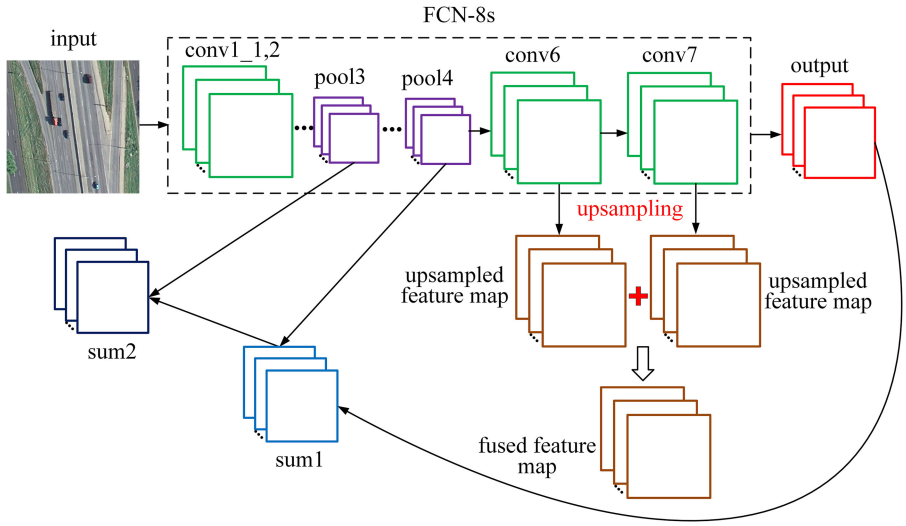


Fig. 3. Two strategies for multiscale feature map extraction.

feature vector of region R_r^j in I_r, and $D_{L_2}^i(\bullet)$ is $1 \times n$ Euclidean distance matrix between feature vectors of R_q^i and R_r^j.

To evaluate the multilabel retrieval performance of SSRCF and MSRCF, we follow [17] and choose accuracy, precision, and recall as performance evaluation metrics. In addition, hamming loss (HL) and F1-measure (F1) are also used. These metrics are defined as follows:

$$P_{\mathrm{HL}} = \frac{1}{M}\sum_{i=1}^{M}\frac{|L_Q \Delta L_{R_i}|}{N} \tag{10}$$

$$P_{\mathrm{Accuracy}} = \frac{1}{M}\sum_{i=1}^{M}\frac{|L_Q \wedge L_{R_i}|}{|L_Q \vee L_{R_i}|} \tag{11}$$

$$P_{\mathrm{Precision}} = \frac{1}{M}\sum_{i=1}^{M}\frac{|L_Q \wedge L_{R_i}|}{|L_{R_i}|} \tag{12}$$

$$P_{\mathrm{Recall}} = \frac{1}{M}\sum_{i=1}^{M}\frac{|L_Q \wedge L_{R_i}|}{|L_Q|} \tag{13}$$

$$P_{F1} = \frac{1}{M}\sum_{i=1}^{M}\frac{2\,|L_Q \wedge L_{R_i}|}{|L_Q| + |L_{R_i}|} \tag{14}$$

where Δ, $\wedge$, and $\vee$ are logical XOR, logical AND and logical OR operations, respectively. $|\bullet|$ is the number of nonzeros, L_Q is the multilabel vector of query image, L_{R_i} is the multilabel vector of the ith returned image R_i, N is the number of labels (i.e., image classes), and M is the number of returned images in one query.

III. Experiments and Analysis

This section first introduces two dense labeling datasets and experimental settings for multilabel RSIR, and then presents the retrieval performance of our SSRCF and MSRCF features.

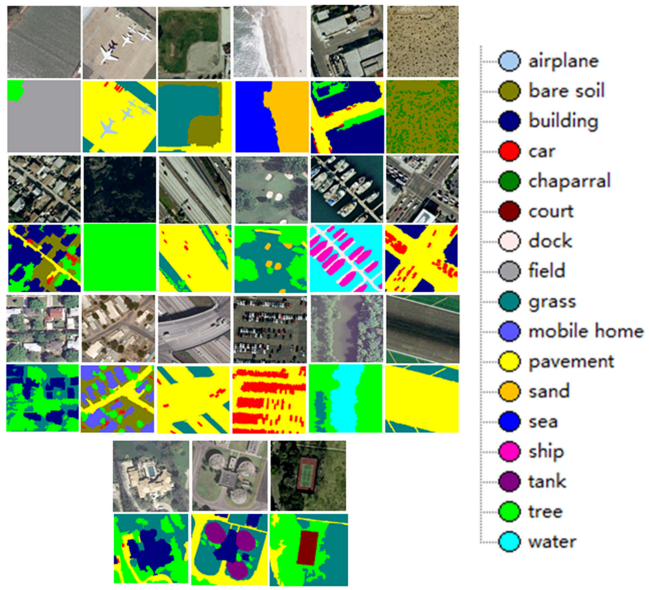


Fig. 4. Example images and corresponding labeling results. (The first, third, five, and seventh rows are source images, and the second, fourth, sixth, and eighth rows are corresponding labeling results, respectively.)

A. Datasets

Dense labeling remote sensing dataset (DLRSD) is the first dense labeling dataset used in our experiment. It is manually labeled based on the UC Merced archive [6] and first presented in our previous work [30]. Specifically, each of the image in UC Merced archive is labeled per pixel with the following 17 classes, i.e., airplane, bare soil, building, car, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tank, tree, and water, which are proposed in [17]. DLRSD contains 2100 RGB images with the spatial size of 256 $\times$ 256 and the resolution is 0.3 m. Fig. 4 shows some example images with the corresponding pixelwise labeling results, and Table I shows

TABLE I
OVERVIEW OF DLRSD DATASET

Class	Number of Images	Label
airplane	100	1
bare soil	754	2
building	713	3
car	897	4
chaparral	116	5
court	105	6
dock	100	7
field	103	8
grass	977	9
mobile home	102	10
pavement	1331	11
sand	291	12
sea	101	13
ship	103	14
tank	100	15
tree	1021	16
water	208	17

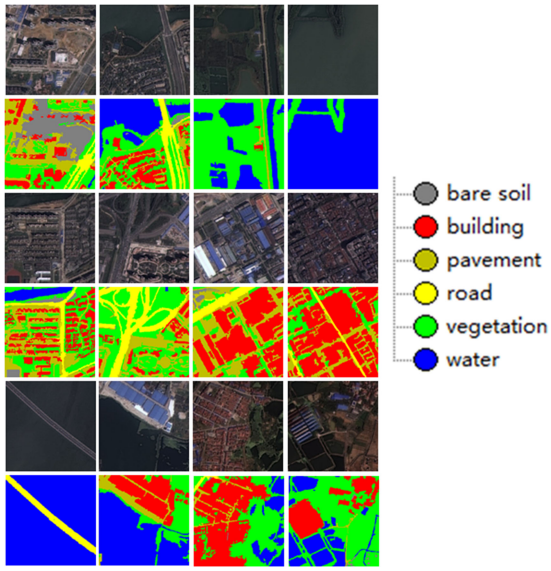


Fig. 5. Example images and corresponding labeling results. (The first, third, and five rows are source images, and the second, fourth, and sixth rows are corresponding labeling results, respectively.)

the overview of DLRSD dataset including the 17 classes as well as the number of images and label corresponding to each class. The readers can be referred to [30] for more details.

Wuhan dense labeling dataset (WHDLD) is the second dense labeling dataset used in our experiment. It is cropped from a large RS image of Wuhan urban area and the pixels of each image in WHDLD is manually labeled with the following six classes, i.e., building, road, pavement, vegetation, bare soil, and water. WHDLD contains 4940 RGB images with the spatial size of 256×256 and the resolution is 2 m. Fig. 5 shows some example images with the corresponding pixelwise labeling results, and Table II shows the overview of WHDLD dataset including the contained six classes as well as the number of images and label corresponding to each class.

TABLE II
OVERVIEW OF WHDLD DATASET

Class	Number of Images	Label
building	3722	1
road	3162	2
pavement	3881	3
vegetation	4631	4
bare soil	3539	5
water	3886	6

B. Experimental Settings

For DLRSD and WHDLD datasets, we randomly select 80% images as training set D_T and the rest 20% images as retrieval set D_R. With respect to FCN architecture, we modify the last fully connected layer of VGG-16 to have 17 and 6 output classes for DLRSD and WHDLD datasets, respectively. During training, the weights of VGG-16 are transferred to FCN and the weights of the transposed convolutional layer are fixed to bilinear interpolation. The learning rate, batch size, and the number of epochs are set to 0.0001, 10, and 120 for DLRSD dataset, and set to 0.001, 32, and 120 for WHDLD dataset. Our FCN is implemented by using MatConvNet [31] as the deep learning framework, and trained with stochastic gradient decent as the optimizer. It is worth noting that we perform 420 queries for DLRSD dataset and 988 queries for WHDLD dataset, and the query image is also regarded a similar image during one query.

In terms of successful matching in multilabel RSIR, we define that two similar images must contain at least one overlapped label. Therefore, we propose to perform a two-step retrieval process to avoid invalid search where query image cannot even find one overlapped label with any images in D_R. In detail, we first remove images without any overlapped labels with query images from D_R, and second query image is compared with each of the remaining images using the region-based distance (9) to find the top 20 most similar images.

The proposed RCFs (i.e., SSRCF and MSRCF) are compared with handcrafted and CNN features to evaluate their multilabel retrieval performance. For handcrafted features, we select color histogram (CH), local binary pattern (LBP) [32], Gabor texture (GT) [33], GIST [34], BoVW [4], and the recent multilabel RSIR approach MLIR [30], while for CNN features, we select pretrained Fc (P-Fc), single-label Fc (S-Fc), multilabel Fc (M-Fc), and the pretrained region convolutional (P-Conv) features.

Tables III and IV show the detailed implementations of handcrafted and CNN features including feature dimension, parameter setting, feature normalization, and distance used for similarity measure. It is worth noting that SSRCF and MSRCF are compared with both image-level and region-level handcrafted and CNN features to demonstrate their advantages for multilabel retrieval. Specifically, CH, LBP, GT, GIST, BoVW, P-Fc, S-Fc, and M-Fc are image-level features represented with feature vectors, while MLIR and P-Conv are region-level features represented with feature matrixes. In Table IV, VGGM [35] and VGG-16 are used as pretrained CNNs to extract image-level and region-level CNN features, respectively. In addition, Caffe [36] is selected as the deep learning framework to fine-tune

TABLE III
DETAILED IMPLEMENTATIONS OF HANDCRAFTED FEATURES

Features	Dimension	Implementation	
		Extraction	Similarity Measure
Color Histogram (CH)	96-D	Color histogram is extracted by quantizing each channel of the RGB color space into 32 bins and concatenating the three histograms.	L_2 normalization is used before measuring the similarity. L_1 distance is used for color histogram and BoVW, and L_2 distance is used for LBP, Gabor texture and GIST.
LBP	10-D	The uniform rotation invariant histogram is extracted with 8 pixel circular neighborhood of radius 1.	
Gabor Texture (GT)	80-D	Five scales and eight orientations are considered and the Gabor filter window size is 32×32.	
GIST	512-D	The default parameters of original implementation are considered [34].	
BoVW	*K*-D	*K* is empirically set as 150 and 200 for DLRSD and WHDLD datasets, respectively.	
MLIR	The default parameters of original implementation are considered [30].		

TABLE IV
DETAILED IMPLEMENTATIONS OF CNN FEATURES

Features	Dimension	Implementation	
		Extraction	Similarity Measure
P-Fc	4096-D	P-Fc is extracted from the Fc layer of pre-trained VGGM.	L_2 normalization is used before measuring the similarity with L_2 distance.
S-Fc		S-Fc is extracted from the Fc layer of fine-tuned VGGM using single-label dataset.	
M-Fc		M-Fc is extracted from the Fc layer of fine-tuned VGGM using multi-label dataset.	
P-Conv	512-D	P-Conv is similar to the proposed SSRCF, but the feature maps are extracted from the conv5_3 layer of pre-trained VGG-16.	L_2 normalization is used before measuring the similarity with region-based distance (9).

TABLE V
SEGMENTATION PERFORMANCE OF OUR FCN, FCN-32S, AND FCN-16S ON TWO DENSE-LABELING DATASETS

Metrics	DLRSD			WHDLD		
	FCN-32s	FCN-16s	FCN-8s	FCN-32s	FCN-16s	FCN-8s
pacc	0.7425	0.7905	**0.8054**	0.7656	0.7958	**0.8106**
macc	0.7267	0.7886	**0.8234**	0.6454	**0.6849**	0.6765
miu	0.5503	0.6369	**0.6770**	0.5076	0.5528	**0.5603**

TABLE VI
MULTILABEL RETRIEVAL PERFORMANCE OF SSRCF FEATURE

Features	HL	Accuracy	Precision	Recall	F1
DLRSD					
SSRCF_c5	0.1164	0.6928	0.7916	**0.8483**	0.7936
SSRCF_c6	0.1108	0.6909	**0.8192**	0.8191	0.7911
SSRCF_c7	**0.1103**	**0.6939**	0.8191	0.8246	**0.7943**
WHDLD					
SSRCF_c5	**0.1561**	**0.8119**	0.8818	**0.9195**	**0.8842**
SSRCF_c6	0.1703	0.7862	**0.8876**	0.8843	0.8647
SSRCF_c7	0.1668	0.7925	0.8785	0.9005	0.8691

single-label and multilabel CNN to extract S-Fc and M-Fc features.

C. Experimental Results

1) Multilabel Retrieval Performance of SSRCF and MSRCF: Multilabel retrieval performance relies on multilabel vectors and RCFs; at the meantime, the accurate segmentation map results in extracting good features and providing accurate multilabel vectors mentioned above, as described in Section II. We therefore first evaluate the segmentation performance of different FCN networks before retrieval. Following the performance metrics used in [27], we use pixel accuracy (pacc), mean accuracy (macc), and mean IU (miu) to report the segmentation performance of our FCN (i.e., FCN-8s), FCN-32s, and FCN-16s, as shown in Table V. As can be seen, our FCN (FCN-8s) is able to achieve the best performance on both DLRSD and WHDLD datasets in terms of pacc, macc, and miu values, and thus it is able to provide optimal solution for multilabel retrieval.

The multilabel retrieval performance of our SSRCF and MSRCF features on DLRSD and WHDLD datasets is shown in Tables VI and VII, respectively. For SSRCF feature, we evaluate features extracted from conv5_3 (SSRCF_c5), conv6 (SSRCF_c6), and conv7 (SSRCF_c7) layers. In Table VI, it can be observed that SSRCF_c7 achieves better performance than SSRCF_c5 and SSRCF_c6 on DLRSD dataset in terms of HL (0.1103), Accuracy (0.6939) and F1 (0.7943) values, and SSRCF_c5 outperforms SSRCF_c6 and SSRCF_c7 on WHDLD dataset in terms of HL (0.1561), Accuracy (0.8119), Recall (0.9195) and F1 (0.8842) values. The results indicate that there are slight differences among the performances of SSRCF extracted from different layers, and SSRCF extracted

TABLE VII
MULTILABEL RETRIEVAL PERFORMANCE OF MSRCF FEATURE ON TWO DENSE-LABELING DATASETS

Features	HL	Accuracy	Precision	Recall	F1
DLRSD					
MSRCF_c67	0.1111	0.6913	**0.8178**	0.8213	0.7916
MSRCF_s1	**0.1051**	**0.7193**	0.7996	0.8830	**0.8150**
MSRCF_s2	0.1072	0.7172	0.7916	**0.8870**	0.8133
WHDLD					
MSRCF_c67	0.1675	0.7909	0.8815	0.8956	0.8679
MSRCF_s1	0.1280	0.8493	0.8818	0.9625	0.9084
MSRCF_s2	**0.1151**	**0.8628**	**0.8912**	**0.9684**	**0.9172**

from higher layer does not show distinct advantages over SSRCF extracted from lower layers.

For MSRCF feature, we evaluate features extracted with the two proposed strategies. In Table VII, it can be observed that MSRCF extracted using the second strategy (extracting multiscale feature maps directly from sum1 or sum2 layers, i.e., MSRCF_s1 and MSRCF_s2) outperforms MSRCF extracted using the first strategy (extracting multiscale feature maps by fusing conv6 and conv7 layer, i.e., MSRCF_c67) on DLRSD and WHDLD datasets. More specifically, MSRCF_s1 achieves better performance than MSRCF_c67 and MSRCF_s2 on DLRSD dataset in terms of HL (0.1051), Accuracy (0.7193) and F1 (0.8150) values, and MSRCF_s2 outperforms MSRCF_c67 and MSRCF_s1 on WHDLD dataset in terms of all five performance metrics. From the results in Tables VI and VII, we can conclude that MSRCF generally outperforms SSRCF, specifically with an improvement of 0.0254 in accuracy and of 0.0207 in F1 on DLRSD dataset, and with an improvement of 0.0509 in accuracy and of 0.033 in F1 on WHDLD dataset.

2) Performance Comparison of RCF and Handcrafted Features: SSRCF and MSRCF that achieve the best performances in Tables VI and VII are compared against handcrafted features, as shown in Table VIII. RCF, and MSRCF in particular, improves multilabel retrieval performance of handcrafted features including single-label RSIR approaches such as CH, LBP, GT, GIST, BoVW, and recent multilabel RSIR approach MLIR. For example, SSRCF shows an improvement of 0.1485 in accuracy and 0.1484 in F1 over BoVW (the best performing handcrafted feature) on DLRSD dataset, and for MSRCF, the improvement is even more significant, specifically 0.1739 higher in accuracy and 0.1691 higher in F1. For WHDLD dataset, SSRCF and MSRCF also result in significant improvement over the best performing handcrafted feature (i.e., GT). The results in Table VIII indicate that our approach achieves a remarkable improvement over handcrafted features for multilabel retrieval.

3) Performance Comparison of RCF and CNN Features: SSRCF and MSRCF are further compared against CNN features, and the results are shown in Table IX. For both DLRSD and WHDLD datasets, RCF and particularly MSRCF outperform other CNN features including pretrained CNN features P-Fc and P-Conv, and fine-tuned CNN features S-Fc (S-Fc is not available for WHDLD dataset since the single label of each image is unknown.) and M-Fc. As an example, SSRCF shows an improvement of 0.0264 in accuracy and 0.0262 in F1 over M-Fc (the best performing CNN feature) on DLRSD dataset, and for MSRCF, the improvement is even more remarkable and is of 0.0518 in accuracy and of 0.0469 in F1. For WHDLD dataset,

TABLE VIII
COMPARISONS OF REGION CONVOLUTIONAL AND HANDCRAFTED FEATURES ON TWO DENSE-LABELING DATASETS

Features	HL	Accuracy	Precision	Recall	F1
DLRSD					
CH	0.1980	0.4895	0.6359	0.6291	0.5932
LBP	0.2087	0.4904	0.6213	0.6492	0.5983
GT	0.1931	0.5129	0.6484	0.6527	0.6167
GIST	0.2217	0.3926	0.5961	0.4866	0.4800
BoVW	0.1718	0.5454	0.6952	0.6603	0.6459
MLIR	0.2017	0.5440	0.6095	0.7717	0.6539
SSRCF	0.1103	0.6939	**0.8191**	0.8246	0.7943
MSRCF	**0.1051**	**0.7193**	0.7996	**0.8830**	**0.8150**
WHDLD					
CH	0.2159	0.7328	0.8543	0.8499	0.8252
LBP	0.2261	0.7245	0.8477	0.8473	0.8202
GT	0.2271	0.7338	0.8507	0.8478	0.8268
GIST	0.2550	0.6924	0.8475	0.7881	0.7907
BoVW	0.2513	0.7013	0.8216	0.8328	0.7991
MLIR	0.2651	0.6974	0.7881	0.8614	0.7963
SSRCF	0.1561	0.8119	0.8818	0.9195	0.8842
MSRCF	**0.1151**	**0.8628**	**0.8912**	**0.9684**	**0.9172**

TABLE IX
COMPARISONS OF REGION CONVOLUTIONAL AND CNN FEATURES ON TWO DENSE-LABELING DATASETS

Features	HL	Accuracy	Precision	Recall	F1
DLRSD					
P-Fc	0.1259	0.6582	0.7823	0.7848	0.7584
S-Fc	0.1255	0.6580	0.7845	0.7831	0.7582
M-Fc	0.1164	0.6675	0.8163	0.7782	0.7681
P-Conv	0.1310	0.6630	0.7629	0.8189	0.7650
SSRCF	0.1103	0.6939	**0.8191**	0.8246	0.7943
MSRCF	**0.1051**	**0.7193**	0.7996	**0.8830**	**0.8150**
WHDLD					
P-Fc	0.1942	0.7599	0.8889	0.8442	0.8457
S-Fc	N/A	N/A	N/A	N/A	N/A
M-Fc	0.1887	0.7675	0.8881	0.8557	0.8518
P-Conv	0.2003	0.7678	0.8557	0.8865	0.8526
SSRCF	0.1561	0.8119	0.8818	0.9195	0.8842
MSRCF	**0.1151**	**0.8628**	**0.8912**	**0.9684**	**0.9172**

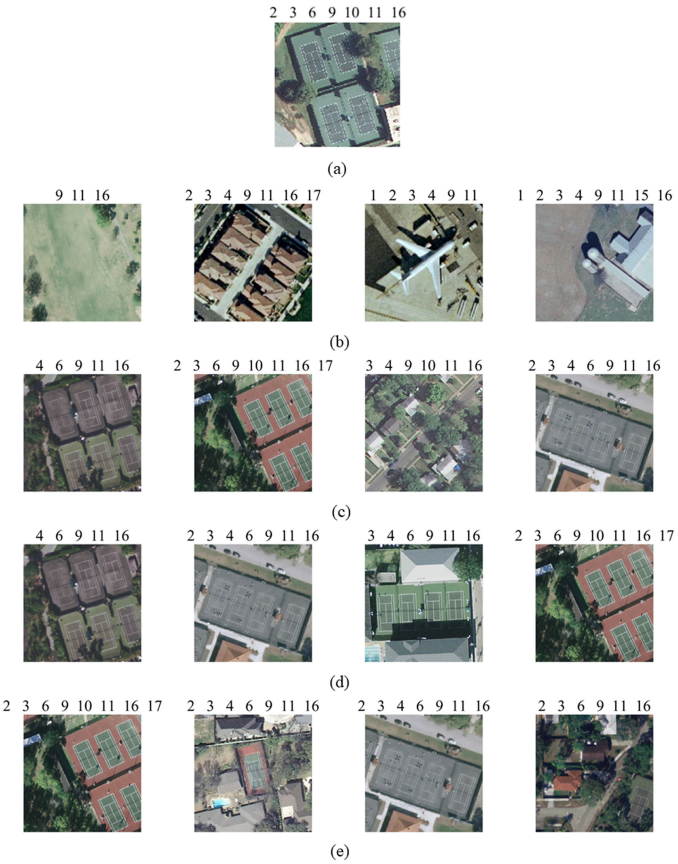


Fig. 6. Tennis court image retrieval in DLRSD dataset. (a) Query image. (b) Images retrieved by BoVW. (c) Images retrieved by M-Fc. (d) Images retrieved by SSRCF. (e) Images retrieved by MSRCF.

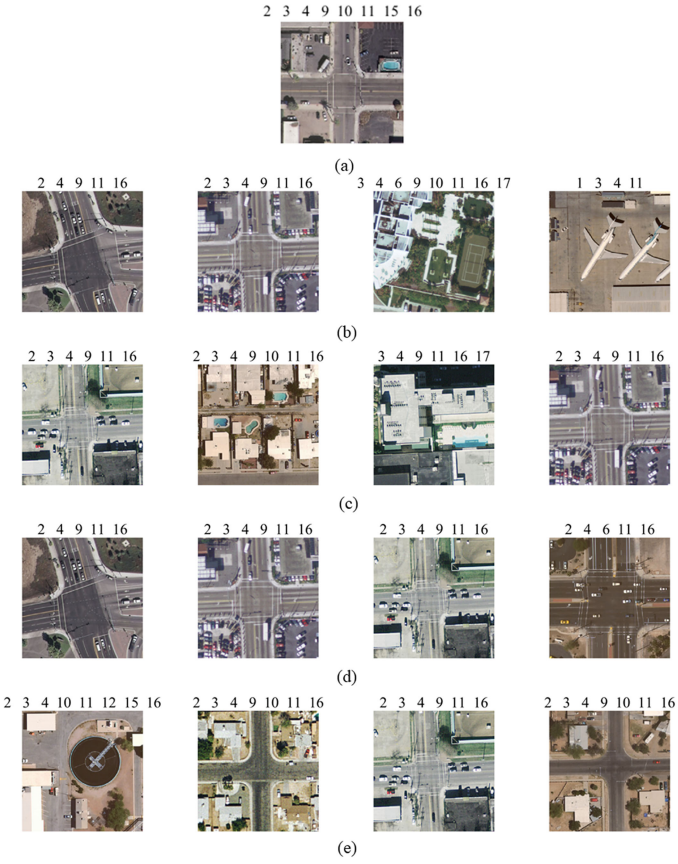


Fig. 7. Intersection image retrieval in DLRSD dataset. (a) Query image. (b) Images retrieved by BoVW. (c) Images retrieved by M-Fc. (d) Images retrieved by SSRCF. (e) Images retrieved by MSRCF.

SSRCF and MSRCF also result in significant improvement over the best performing CNN feature (i.e., M-Fc). The results in Table IX indicate that our approach achieves a remarkable improvement over CNN features for multilabel retrieval.

4) Multilabel Image Retrieval Instances: In order to further evaluate the performance of our proposed multilabel retrieval approach, we select BoVW, M-Fc, SSRCF, and MSRCF for DLRSD dataset, and select GT, M-Fc, SSRCF, and MSRCF for WHDLD dataset to perform multilabel retrieval and return the top four similar images, as shown in Figs. 6–9.

Fig. 6 shows an example of images in DLRSD dataset retrieved by BoVW, M-Fc, SSRCF, and MSRCF. The query image is selected from the tennis court category of UC Merced archive, and the multiple labels associated with each image are given above the related image. From the results, one can see that all the images retrieved by the proposed SSRCF [see Fig. 6(d)] and MSRCF [see Fig. 6(e)] contain tennis court. On the contrary, the images retrieved by BoVW [see Fig. 6(b)] and M-Fc [see Fig. 6(c)] have at least one image that does not contain tennis court. For example, the third image retrieved by M-Fc originally belongs to medium residential category of the UC Merced archive, and for BoVW, these four images originally belong to golf course, dense residential, airplane, and storage tank categories of the UC Merced archive, respectively.

Fig. 7 illustrates another example of images in DLRSD dataset retrieved by BoVW, M-Fc, SSRCF, and MSRCF with the query image taken from the intersection category. It is obvious that all the images (except for the first image retrieved by MSRCF) retrieved by SSRCF and MSRCF originally belong to intersection category of the UC Merced archive. From a visual analysis of the top four returned images, we can conclude that our proposed approach is able to accurately obtain the multiple primitive classes associated with each query image and thus retrieve those visually most similar images in DLRSD dataset.

Fig. 8 shows an example of images in WHDLD dataset retrieved by GT, M-Fc, SSRCF, and MSRCF. The query image is associated with two primitive classes, namely vegetation and water (see Table II). From these returned images, one can see that images retrieved by the proposed SSRCF and MSRCF contain all the two classes (i.e., vegetation and water). On the contrary, the images retrieved by GT and M-Fc mostly contain other primitive classes [see Fig. 8(b) and (c)]. Fig. 9 illustrates another example of images retrieved by GT, M-Fc, SSRCF, and MSRCF. The query image is associated with six primitive classes, namely building, road, pavement, vegetation, bare soil, and water. It is obvious that images retrieved by the proposed MSRCF contain all the six primitive classes [see Fig. 9(e)], and images retrieved by the proposed SSRCF contain most of the six primitive classes

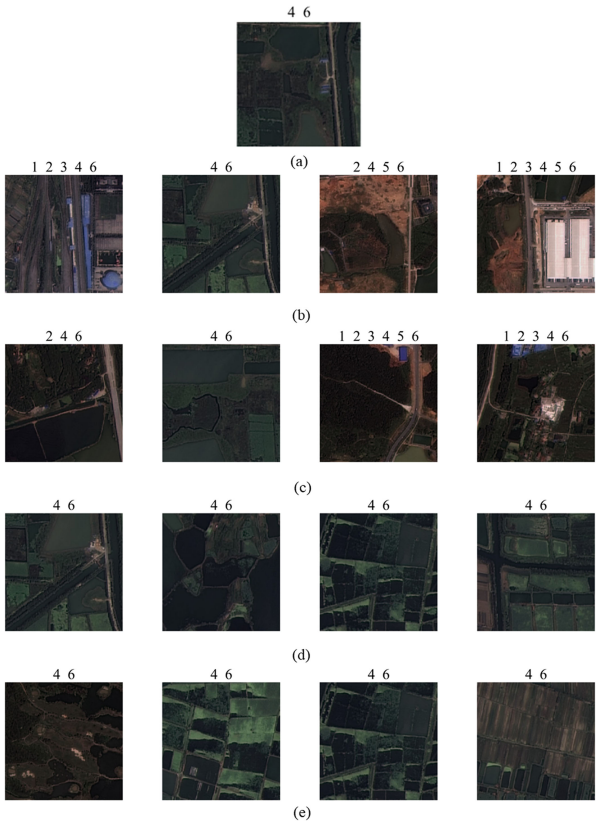


Fig. 8. Multilabel image retrieval in WHDLD dataset. (a) Query image. (b) Images retrieved by GT. (c) Images retrieved by M-Fc. (d) Images retrieved by SSRCF. (e) Images retrieved by MSRCF.

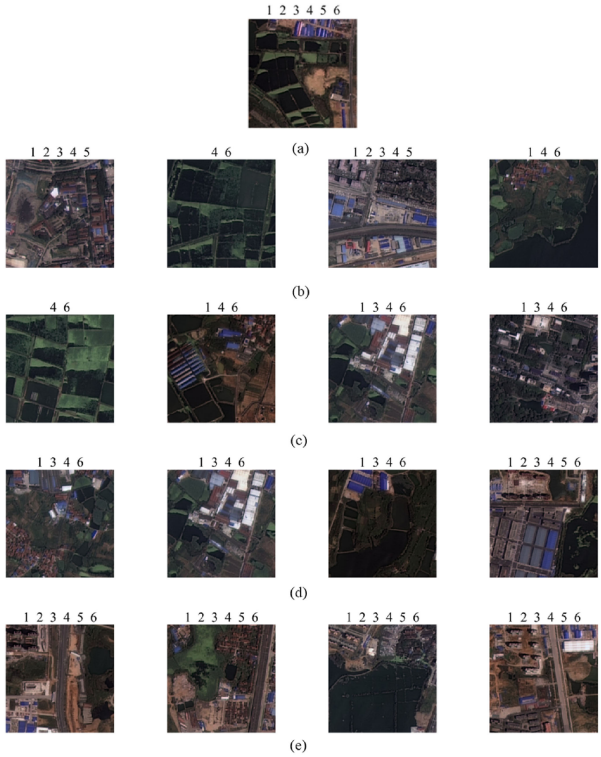


Fig. 9. Multilabel image retrieval in WHDLD dataset. (a) Query image. (b) Images retrieved by GT. (c) Images retrieved by M-Fc. (d) Images retrieved by SSRCF. (e) Images retrieved by MSRCF.

[see Fig. 9(d)]. From a visual analysis of the top four returned images, we can conclude that our proposed approach is able to accurately obtain the multiple primitive classes associated with each query image and thus retrieve those visually most similar images in WHDLD dataset.

IV. Conclusion

In this article, we present a novel multilabel RSIR approach based on FCN. In our approach, FCN is first trained to predict the segmentation map of each image in our retrieval archive. We then obtain multilabel vectors and extract single-scale and multi-scale RCFs of each image based on corresponding segmentation map. Finally, the region-based distance is used to measure the similarity between query image and other images in the archive.

Our multilabel RSIR approach is compared against handcrafted and CNN features using two dense-labeling datasets. Experimental results demonstrate that our approach outperforms state-of-the-art RSIR approaches including handcrafted features such as CH, LBP, GT, GIST, BoVW, and MLIR, and CNN features such as P-Fc, S-Fc, M-Fc, and P-Conv on both DLRSD and WHDLD datasets.

There are some limitations of our article. For example, the spatial relationships between different regions are not considered while measuring image similarity using the region-based distance. It is worth noting that the retrieval performance of our proposed approach can be further improved if the aforementioned problem can be addressed. In the future, we plan to design a FCN layer that can directly extract multiscale RCFs, and take the spatial relationships between different connected regions into consideration while computing image similarity.

Acknowledgment

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

References

[1] 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, pp. 228–243, 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

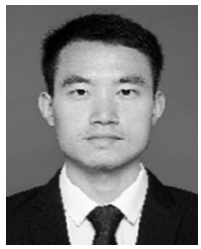
[7] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4094–4102.

[8] R. Li, Y. Zhang, Z. Lu, J. Lu, and Y. Tian, "Technique of image retrieval based on multi-label image annotation," in *Proc. 2nd Int. Conf. Multimedia Inf. Technol.*, 2010, pp. 10–13.

[9] G. Nasierding and A. Z. Kouzani, "Empirical study of multi-label classification methods for image annotation and retrieval," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl.*, 2010, pp. 617–622.
[10] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit*., vol. 37, no. 9, pp. 1757–1771, 2004.
[11] F. Li, Q. Dai, W. Xu, and G. Er, "Multilabel neighborhood propagation for region-based image retrieval," *IEEE Trans. Multimed.*, vol. 10, no. 8, pp. 1592–1604, Dec. 2008.
[12] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2015, pp. 1556–1564.
[13] 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.
[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] 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.
[16] 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 Obs. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.
[17] 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.
[18] L. Zhou, Z. Wang, Y. Luo, and Z. Xiong, "Separability and compactness network for image recognition and superresolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3275–3286, Nov. 2019.
[19] A. Sellami, M. Farah, I. R. Farah, and B. Solaiman, "Hyperspectral imagery classification based on semi-supervised 3-D deep neural network and adaptive band selection," *Expert Syst. Appl.*, vol. 129, pp. 246–259, 2019.
[20] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens*., vol. 9, no. 1, pp. 67–87, 2017.
[21] 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.
[22] 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, pp. 489–508, 2017.
[23] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval," *Remote Sens*., vol. 10, no. 5, pp. 709–753, 2018.
[24] 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.
[25] P. Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1343–1376, 2018.
[26] 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.
[27] 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.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2–9.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv: 1409.1556*.
[30] W. Shao, Z. Yang, and K. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens*., vol. 10, no. 6, pp. 964–976, 2018.
[31] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[32] 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.
[33] 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.
[34] 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.
[35] 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.*, 2014, pp. 1–11.
[36] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–6781.

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. His research interests include remote sensing and data mining.



Weixun Zhou received the B.S. degree in surveying and mapping engineering from the Anhui University of Science and Technology, Huainan, China, in 2012, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019.

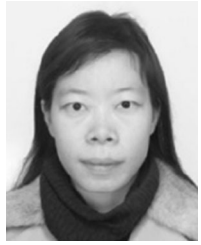
He is currently a Lecturer with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include urban remote sensing, image processing, deep learning, and computer vision.

Xueqing Deng received the B.Sc. degree in geographic information systems and remote sensing from Sun Yat-Sen University, Guangzhou, China, in 2016, and is currently working toward the Ph.D. degree in computer science with the University of California at Merced, Merced, CA, USA.

Her research interests include semantic segmentation, deep learning in the context of geospatial problems, and remote sensing. She has been working on problems such as synthesizing ground-level images from overhead images, land-cover classification using geotagged social media, generalizing deep learning by knowledge-guided models for semantic segmentation.

Maoding Zhang received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018, where he is currently working toward the M.S. degree in photogrammetry and remote sensing.

His research interests include remote sensing scene classification and retrieval.



Qimin Cheng received the Ph.D. degree in Cartography and Geographic Information System from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2004.

She is currently an Associate Professor with the Huazhong University of Science and Technology, Wuhan, China. Her research interests include image retrieval and annotation, remote sensing images understanding and analysis.