ISPRS Journal of Photogrammetry and Remote Sensing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval

Weixun Zhou [a], Shawn Newsam [b], Congmin Li [a], Zhenfeng Shao [a,*]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, No. 129 Luoyu Road, Wuhan 430079, China
[b] Electrical Engineering and Computer Science, University of California, Merced, CA 95343, USA

ARTICLE INFO

Article history:
Received 11 June 2017
Received in revised form 8 December 2017
Accepted 5 January 2018
Available online xxxx

Keywords:
Remote sensing
Content based image retrieval (CBIR)
Benchmark dataset
Handcrafted features
Deep learning
Convolutional neural networks

ABSTRACT

Benchmark datasets are critical for developing, evaluating, and comparing remote sensing image retrieval (RSIR) approaches. However, current benchmark datasets are deficient in that (1) they were originally collected for land use/land cover classification instead of RSIR; (2) they are relatively small in terms of the number of classes as well as the number of images per class which makes them unsuitable for developing deep learning based approaches; and (3) they are not appropriate for RSIR due to the large amount of background present in the images. These limitations restrict the development of novel approaches for RSIR, particularly those based on deep learning which require large amounts of training data. We therefore present a new large-scale remote sensing dataset termed "PatternNet" that was collected specifically for RSIR. PatternNet was collected from high-resolution imagery and contains 38 classes with 800 images per class. Significantly, PatternNet's large scale makes it suitable for developing novel, deep learning based approaches for RSIR. We use PatternNet to evaluate the performance of over 35 RSIR methods ranging from traditional handcrafted feature based methods to recent, deep learning based ones. These results serve as a baseline for future research on RSIR.

1. Introduction

Over the past several decades, remote sensing has experienced dramatic changes in the increased spatial resolution of the imagery as well as the increased rate of acquisition. These changes have had profound effects on the way that we use and manage remote sensing images. The increased spatial resolution provides new opportunities for advancing remote sensing image analysis and understanding, making it possible to develop novel approaches that were not possible before. The increased acquisition rate enables us to acquire a considerable volume of remote sensing data on a daily basis. But this has resulted in the significant challenge of how to efficiently manage the large data collections, particularly so that the data of interest can be accessed quickly.

Content based image retrieval (CBIR) is a useful technique for the fast retrieval of images of interest from a large-scale dataset (Agouris et al., 1999). The remote sensing community has invested significant effort to adapt CBIR to remote sensing images in recent years, making remote sensing image retrieval (RSIR) an active and challenging research topic. The remote sensing community has been particularly focused on developing powerful feature extraction methods since retrieval performance is heavily dependent on the effectiveness of the features.

Traditional RSIR methods use low-level visual features to represent the content of the images. These features can be either global or local. Global features are extracted from the whole image, e.g. color (spectral) features, texture features and shape features. In contrast to global features, local features like Scale Invariant Feature Transform (SIFT) Lowe (2004) are extracted from image patches that are centered at interesting points. Local features enjoy several advantages over global ones such as robustness to occlusion as well as invariance to viewing angle and lighting conditions. The remote sensing community has sought to exploit these properties of local features for remote sensing image retrieval (Özkan et al., 2014). These local and global features are hand-crafted though. Their development is time consuming and often involves ad-hoc or heuristic design decisions, making them suboptimal for the task at hand. Deep learning has dramatically advanced the state-of-the-art in various computer vision problems (LeCun et al., 2015) as well as remote sensing problems such as simultaneous roads and buildings extraction (Alshehhi et al., 2017), very-high-resolution optical image (Zhao et al., 2017) and hyperspectral image (Ma et al., 2016) classification. Unlike hand-crafted features,

* Corresponding author.
E-mail address: shaozhenfeng@whu.edu.cn (Z. Shao).

https://doi.org/10.1016/j.isprsjprs.2018.01.004
0924-2716/

deep learning is capable of discovering intricate structure in large data sets and can automatically learn optimal and powerful feature representations. Inspired by the great success of deep learning, researchers have begun exploiting the potential of deep learning techniques for image retrieval (Wan et al., 2014).

Though the remote sensing community has achieved notable progress in RSIR in recent years, particularly through deep learning-based methods, a comprehensive survey of the existing methods on a benchmark dataset is lacking. The existing evaluations are deficient in that they are performed using different performance metrics, on different datasets, and/or under different experimental configurations. There are two fundamental challenges to performing a consistent evaluation. First is the effort and technical challenges of re-implementing the methods to produce results that can be meaningfully compared. Second is establishing consistent experimental conditions, central to which is having a rich evaluation dataset. Actually, there are a number of publicly available remote sensing benchmark datasets; however, they were collected for land use/land cover classification and not RSIR.

Image classification and image retrieval differ in terms of their goals, how they accomplish those goals, and how they are employed by users. They also require different types of datasets to develop and evaluate novel methods.

The goal of image classification is to assign one or more semantic labels to a given image (Han et al., 2017). The goal of image retrieval is to identify images in a target set that are similar to a query image. Classification is typically performed using a classifier that is trained using a set of labeled images. Retrieval is performed by comparing features extracted from the query image to features extracted from the target images. These comparisons are used to rank the target images in order of decreasing similarity.

When performing retrieval, a user usually selects a query image that contains *only* the pattern/object/scene of interest. For this case, the query image does not contain other patterns/objects/scenes by construction since they are irrelevant to the task at hand. Therefore, image retrieval methods should be developed and evaluated using datasets in which sample images contain only a single pattern/object/scene of interest without any background. In contrast, a user performing classification might want to assign multiple labels to an image including a background label. Classification methods can be developed and evaluated using sample images with multiple patterns/objects/scenes of interest including background in which all instances are labeled. (This generally makes classification the more difficult task.)

From a dataset perspective, classification methods can be developed and evaluated using image retrieval datasets but not vice versa. Image retrieval images should not have distracting patterns/objects/scenes or background.

In this paper, we first introduce a novel, large-scale remote sensing dataset, named PatternNet. PatternNet provides the remote sensing community with a publicly available benchmark dataset to develop novel algorithms for RSIR. We then provide a comprehensive review of the existing RSIR approaches ranging from traditional handcrafted feature-based methods to recently developed deep learning feature-based ones. The main contributions of this paper are as follows:

- We construct a large-scale remote sensing benchmark dataset, PatternNet, for RSIR. PatternNet is a publicly available, high-resolution dataset which contains more classes and more images than the current RSIR datasets.
- We provide a comprehensive review of the state-of-the-art methods for RSIR, ranging from traditional handcrafted feature based methods to recently developed deep learning feature based ones.
- We evaluate more than 35 methods on PatternNet under consistent experimental conditions. This provides the literature with extensive baseline results for future research on RSIR.

The rest of this paper is organized as follows. We provide a comprehensive review of several publicly available remote sensing datasets and introduce our large-scale dataset PatternNet in Section 2. Section 3 reviews the existing methods including handcrafted features and deep learning features for RSIR. The results and comparisons of these methods are shown in Section 4. Section 5 draws some conclusions.

2. PatternNet: A large-scale dataset for remote sensing image retrieval

In addition to the limitations (i.e. datasets are collected for classification problem) of the existing datasets mentioned above, most of those datasets are also small, making them unsuitable for developing deep learning based RSIR methods. Further, there tends to be a large amount of background present in the images which makes them inappropriate for RSIR. A large scale benchmark dataset collected specifically for RSIR is needed to advance the field. Such a dataset is particularly important to exploit deep learning based approaches which have shown tremendous success on other computer vision problems.

This section first reviews several publicly available remote sensing datasets and then introduces the proposed large-scale PatternNet dataset for RSIR.

2.1. The existing remote sensing datasets

UC Merced dataset (http://vision.ucmerced.edu/datasets/landuse.html). The UC Merced dataset (UCMD) Yang and Newsam (2013) is a land use/land cover classification dataset which contains 100 images of the following 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks and tennis courts. Each image measures 256 × 256 pixels. The images are cropped from large aerial images downloaded from the United States Geological Survey (USGS) and the spatial resolution is around 0.3 m. The UCMD dataset has several highly overlapping classes (i.e. sparse residential, medium residential and dense residential), which makes it a challenging dataset. It was also the first publicly available remote sensing evaluation dataset and has been used extensively to develop and evaluate RSIR methods.

WHU-RS19 dataset (http://dsp.whu.edu.cn/cn/staff/yw/HRSscene.html). The WHU-RS19 remote sensing dataset (RSD) Sheng et al. (2012) is manually collected from Google Earth Imagery and labeled into 19 classes: airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking, pond, port, railway station, residential area, river, and viaduct. The dataset consists of a total of 1005 images and each image has the size of 600 × 600 pixels. The images in the RSD dataset have a wide range of spatial resolutions with a maximum of 0.5 m.

RSSCN7 dataset (https://www.dropbox.com/s/j80iv1a0mvhonsa/RSSCN7.zip?dl=0). The RSSCN7 dataset (Zou et al., 2015) is sampled on four different scale levels from Google Earth imagery and consists of 7 classes: grassland, forest, farmland, parking lot, residential region, industrial region, river, and lake. There are 400 images in each class and each image has size of 400 × 400 pixels.

Aerial image dataset (http://www.lmars.whu.edu.cn/xia/AID-project.html). The aerial image dataset (AID) Xia et al. (2017) is a

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

large-scale dataset which was collected with the goal of advancing the state-of-the-art in scene classification of remote sensing images. It is notably larger than the three datasets mentioned above and contains 30 classes: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. There are a total of 10,000 images in the AID dataset and each class has 220–420 images of size 600 × 600 pixels. The spatial resolution of this dataset varies greatly between approximately 0.5 to 8 m.

NWPU-RESISC45 dataset (https://1drv.ms/u/s!AmgKYzARBl5-ca3HNaHIlzp_IXjs). The NWPU-RESISC45 dataset (NWPU45) Cheng et al. (2017) is currently the largest publicly available benchmark dataset for remote sensing scene classification. It is constructed by first investigating all scene classes of the existing datasets and then selecting a list of 45 representative ones: airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snow berg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. Each class has 700 images of size 256 × 256 pixels and the spatial resolution of the images varies from approximately 0.2 to 30 m.

The UCMD dataset has been used the most widely as a benchmark for RSIR as it is the oldest. It is too small, however, for developing deep learning based approaches. The NWPU45 and AID datasets are larger compared to UCMD, RSD, and RSSCN7. However, their images contain significant amounts of background and thus are not suitable for retrieval. See the sample images in Fig. 1. The top 12 images are from NWPU45. The class of interest only represents a small portion of each image. The remainder of the image contains widely varying background which will dominate the image representation and distract retrieval. The bottom three images in Fig. 1 are from AID and are also seen to contain a large amount of background. The third image of this set is also potentially mislabeled as it contains a wastewater treatment plant and not storage tanks. Finally, the spatial resolution of the images in the NWPU45 and AID datasets vary greatly, approximately 0.2 to 30 m for NWPU45 and 0.5 to 8 m for AID. Such a large range is not appropriate for retrieval as the images will cover significantly differently sized regions and thus are unlikely to be visually or semantically similar. It also increases the chance they are mislabeled.

2.2. A New, Large-scale dataset for remote sensing image retrieval

A new dataset is needed for RSIR, one which overcomes the limitations of the datasets above. We thus present the new RSIR dataset PatternNet.

The PatternNet dataset. PatternNet[1] is a large-scale[2] high-resolution remote sensing dataset collected for RSIR. It contains 38 classes: airplane, baseball field, basketball court, beach, bridge, cemetery, chaparral, Christmas tree farm, closed road, coastal mansion, crosswalk, dense residential, ferry terminal, football field, forest, freeway, golf course, harbor, intersection, mobile home park, nursing home, oil gas field, oil well, overpass, parking lot, parking space, railway, river, runway, runway marking, shipping yard, solar panel, sparse residential, storage tank, swimming pool, tennis court, transformer station and wastewater treatment plant. Each class contains 800 images measuring 256 × 256 pixels. The images in PatternNet are collected from Google Earth imagery or via the Google Map API for US cities. Table 1 compares the PatternNet dataset with the existing datasets in terms of the number of images per class, the number of classes, the total number of images, the image spatial resolution and size, and the target application. PatternNet is seen to be a better dataset for RSIR, especially for deep learning which requires large amounts of labeled data.

Table 2 provides additional details about PatternNet including the source and spatial resolution of the images for each class. Similar to the AID and the NWPU45 datasets, the PatternNet images vary in resolution. However, they tend to be much higher resolution. Even at the lowest resolution, 4.693 m, they cover a much smaller region, approximately 1.4 km^2, versus 59.3 km^2 for the NWPU45 and 23.0 km^2 for the AID datasets. This again makes the NWPU45 and AID images more likely to contain background regions and be visually or semantically different.

Fig. 2 shows sample images from the PatternNet dataset. Note that the class of interest covers most of the image—there is very little background. In summary, our proposed PatternNet dataset has the following notable characteristics.

- RSIR dataset. The PatternNet dataset is the largest publicly available remote sensing dataset collected specifically for RSIR.
- Large scale. PatternNet has a large number of images per class and a large number of images overall making it more suitable for deep learning based RSIR approaches than the existing datasets.
- High resolution. The AID and NWPU45 datasets have spatial resolutions ranging from 0.5 m to 8 m and from 0.2 m to 30 m respectively. Many images thus cover a large area and contain a large amount of background which is not appropriate for RSIR. In contrast, PatternNet has a higher spatial resolution so that the classes of interest constitute a larger portion of the images.

3. Remote sensing image retrieval methods

The retrieval performance of RSIR methods depends greatly on the representation strength of the image features. Significant effort has therefore been undertaken to develop effective features over the past few decades. Existing feature representations for RSIR can be generally categorized into two groups, handcrafted features and deep learning features. Note that the two categories are not strictly distinct—hybrid or combinations have also been proposed.

3.1. Handcrafted feature based methods

3.1.1. Methods based on low-level features

Early RSIR methods relied on handcrafted low-level visual features to represent the content of remote sensing images. This includes globally extracted features (global features) and locally extracted features (local features).

Generally, there are three kinds of global features: color (spectral) features Bosilj et al. (2016), texture features (Aptoula, 2014; Zhu and Shao, 2011; Shao et al., 2014), and shape features (Zhang et al., 2013; Scott et al., 2011). Color and texture features have been used more widely than shape features for RSIR. Remote sensing images typically have several spectral bands (e.g. multispectral imagery) and sometimes even have hundreds of bands (e.g. hyper-spectral imagery) and therefore spectral information is crucial for remote sensing image analysis. Bosilj et al. explored both global and local pattern spectral features for geographical image retrieval, and implemented pattern spectra features for the first time with a dense strategy (Bosilj et al., 2016). The

[1] PatternNet is available at https://sites.google.com/view/zhouwx/dataset.
[2] In this paper, "large-scale" means "large amount".

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

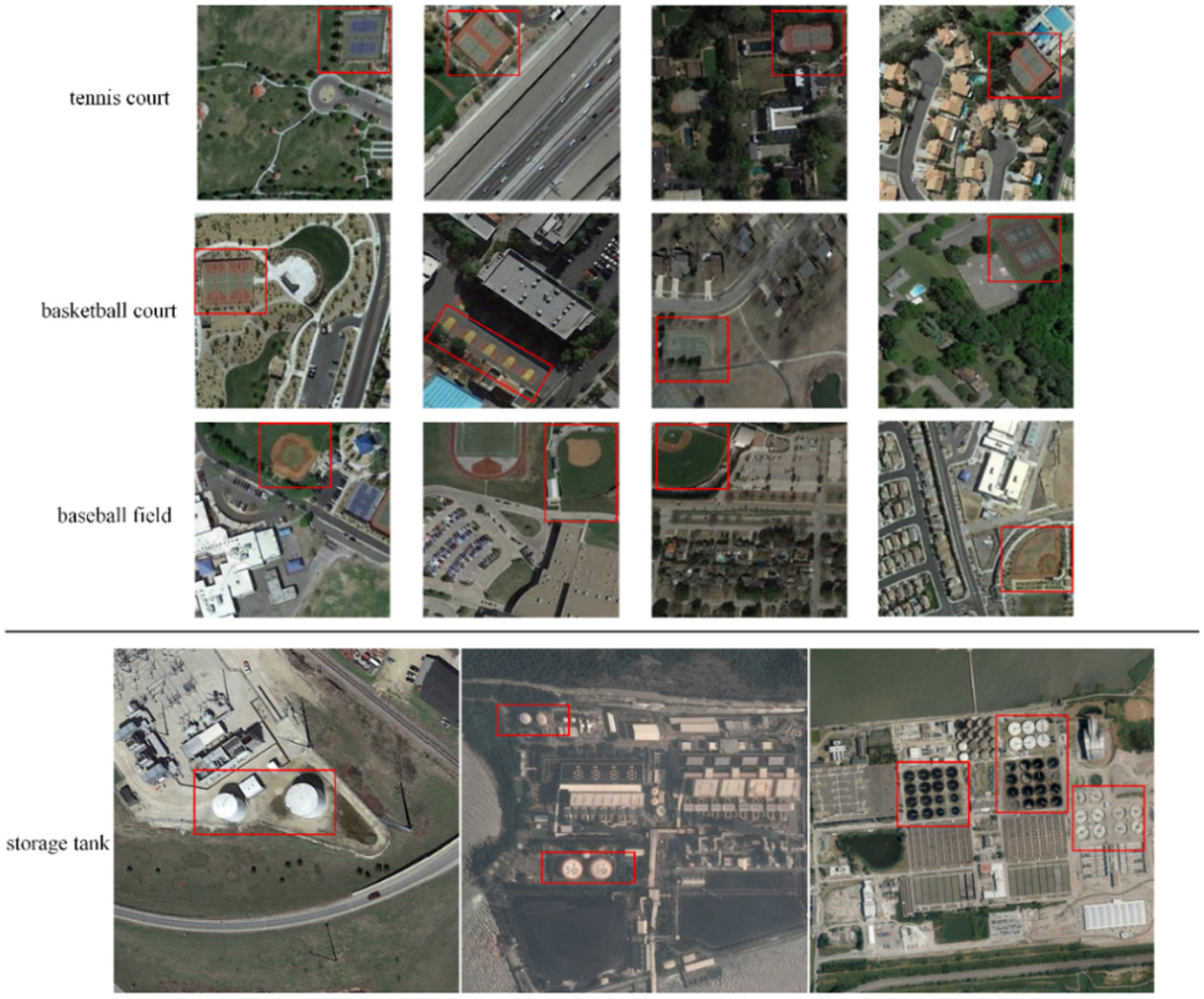

Fig. 1. Some example images from the NWPU45 (top row) and the AID (bottom row) datasets.

Table 1
Comparison of the proposed PatternNet dataset and the existing datasets.

Dataset	Images/class	Classes	Images	Resolution (m)	Size	Application
UCMD	100	21	2100	0.3	256 × 256	Classification
RSD	~50	19	1005	up to 0.5	600 × 600	Classification
RSSCN7	400	7	2800	N/A	400 × 400	Classification
AID	220–420	30	10,000	0.5– 8	600 × 600	Classification
NWPU45	700	45	31,500	0.2–30	256 × 256	Classification
PatternNet	800	38	30,400	0.062–4.693	256 × 256	Retrieval

performance of the global spectral features as well as its new counterpart were evaluated and compared to state-of-the-art approaches on a benchmark dataset, resulting in the best morphology-based results thus far. Color features, however, perform poorly when instances of an object/class vary in spectra or spectra are shared between different objects/classes. Texture features have therefore been applied to capture the spatial variation of pixel intensity, and, indeed, they have demonstrated remarkable performance on a range of remote sensing tasks including RSIR. Aptoula explored the potential of recently developed multiscale texture descriptors, the circular covariance histogram and the rotation-invariant point triplets, for the problem of geographic image retrieval, and introduced several new descriptors based on the Fourier power spectrum (Aptoula, 2014). These descriptors were shown to outperform the best retrieval scores in spite of their low dimensions. However, most existing texture features are extracted from greyscale images, discarding the useful color information of remote sensing images. Shao et al. therefore proposed improved color texture descriptors for RSIR which incorporate discriminative information among color channels (Shao et al., 2014) and thus outperform other texture features like Gabor texture and local binary pattern (LBP) Ojala et al. (2002). Zhu et al. proposed a multi-scale, multi-orientation texture transform spectrum to perform two-level coarse-to-fine rotation- and scale-invariant texture image retrieval (Zhu and Shao, 2011). Experiments on a benchmark texture dataset show that the proposed approach captures the primary orientation of an image and generates an informative descriptor. There are other works that focus on combining color and texture features to improve the performance of hyperspectral imagery retrieval (Shao et al., 2015). Other global features like simple statistics (Yang and Newsam, 2013), and GIST features (Oliva and Torralba, 2001) have also been used for RSIR.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

Table 2
Details of PatternNet dataset. "GMA" means the images are collected using the Google Maps API and "GE" means the images are collected from Google Earth imagery.

Class	Resolution (meter/pixel)		Source	
	GMA	GE	GMA	GE
Airplane	N/A	0.217	No	Yes
Baseball field	0.233–0.293	0.124	Yes	Yes
Basketball court	0.116–0.146	0.161	Yes	Yes
Beach	N/A	0.158	No	Yes
Bridge	0.465–0.586	0.466	Yes	Yes
Cemetery	0.233–0.293	N/A	Yes	No
Chaparral	0.233–0.293	N/A	Yes	No
Christmas tree farm	N/A	0.124	No	Yes
Closed road	0.233–0.293	0.217	Yes	Yes
Coastal mansion	0.233–0.293	N/A	Yes	No
Crosswalk	0.233–0.293	N/A	Yes	No
Dense residential	0.233–0.293	N/A	Yes	No
Ferry terminal	0.465–0.586	0.311	Yes	Yes
Football field	0.931–1.173	0.817	Yes	Yes
Forest	0.233–0.293	N/A	Yes	No
Freeway	N/A	0.311	No	Yes
Golf course	0.233–0.293	0.233	Yes	Yes
Harbor	0.233–0.293	N/A	Yes	No
Intersection	0.465–0.586	N/A	Yes	No
Mobile home park	N/A	0.248	No	Yes
Nursing home	0.465–0.586	N/A	Yes	No
Oil gas field	3.726–4.693	N/A	Yes	No
Oil well	N/A	0.062	No	Yes
Overpass	N/A	0.466	No	Yes
Parking lot	0.233–0.293	N/A	Yes	No
Parking space	0.116–0.146	0.102	Yes	Yes
Railway	0.233–0.293	N/A	Yes	No
River	0.931–1.173	N/A	Yes	No
Runway	0.465–0.586	N/A	Yes	No
Runway marking	0.233–0.293	N/A	Yes	No
Shipping yard	0.233–0.293	N/A	Yes	No
Solar panel	0.233–0.293	N/A	Yes	No
Sparse residential	0.233–0.293	N/A	Yes	No
Storage tank	0.465–0.586	N/A	Yes	No
Swimming pool	0.116–0.146	N/A	Yes	No
Tennis court	0.116–0.146	0.158	Yes	Yes
Transformer station	0.233–0.293	N/A	Yes	No
Wastewater treatment plant	0.233–0.293	0.124/0.189/0.248	Yes	Yes

Unlike global features, local features are extracted from image patches centered at interesting points in an image. SIFT is one of the most popular local feature descriptors and has been used widely for various remote sensing tasks including scene classification, RSIR, etc. Yang et al. investigated the use of local invariant features to perform an extensive evaluation of geographic image retrieval on the UCMD data which was, at the time, the only publicly available remote sensing benchmark dataset (Yang and Newsam, 2013). The local invariant features are compared to several global features, such as simple statistics, color histograms, and texture features. The extensive experiments indicate the superiority of local invariant features over global features. In Özkan et al. (2014), the performance of various image representations for image search problems for geographic image retrieval are investigated. The results demonstrate the suitability of local features for RSIR. Shechtman et al. proposed a local self-similarity (SSIM) descriptor Shechtman and Irani (2007) to measure the similarity between images or videos based on internal similarities. This descriptor is shown to be efficient and effective for deformable shape retrieval. Other popular local features include histogram of oriented gradient (HOG) Dalal and Triggs (2005) and its variant, descriptor pyramid histogram of oriented gradient (PHOG) (Bosch et al., 2007).

3.1.2. Methods based on mid-level features

In general, local features like SIFT are of high dimension and numerous, making them impractical for large-scale RSIR. Methods have therefore been developed to transform the local, low-level features into mid-level features of intermediate complexity through feature encoding techniques such as bag of visual words (BOVW) Sivic and Zisserman (2003), vector of locally aggregated descriptors (VLAD) Jégou et al. (2010), and improved fisher kernel (IFK) Perronnin et al. (2010). BOVW is one of the most popular mid-level features and has been widely used to encode local features into a compact global image representation. BOVW and its variants have shown remarkable performance in image retrieval (Özkan et al., 2014; Yang and Newsam, 2013; Aptoula, 2014; Yang et al., 2015). In Yang and Newsam (2013), BOVW features obtained by encoding saliency and grid based SIFT descriptors are compared to several global features. The extensive experiments demonstrate the superiority of BOVW over these global features. In Özkan et al. (2014), BOVW is compared with VLAD and its more compact variation, product quantized VLAD (VLAD-PQ), for the purpose of geographic image retrieval from satellite imagery. The results show that VLAD-based representations are more discriminative than BOVW in almost all the land cover classes.

BOVW is not only an image representation but also a framework that can be combined with other features to extract even more powerful representations. For instance, in Aptoula (2014), morphological texture descriptors are combined with the BOVW paradigm in order to extract bags of morphological words for content-based geographic image retrieval. The existing global morphological texture descriptors are adapted to local sub-windows. These local descriptors are then used to form a vocabulary of "visual morphological words" through clustering.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

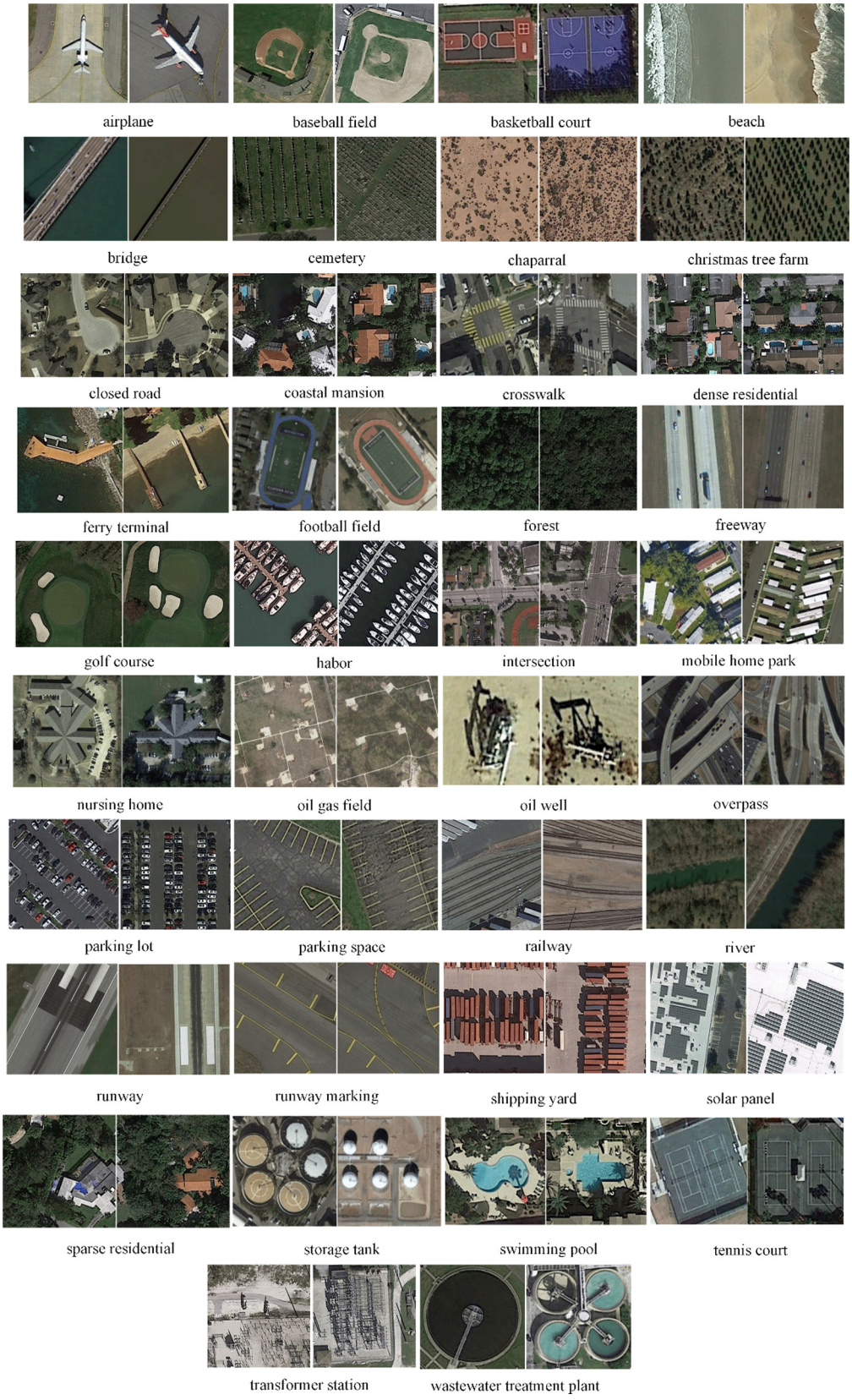

Fig. 2. Two sample images from each class of the PatternNet dataset.

Other works focus on improving the BOVW framework in order to achieve better performance. For instance, in Yang et al. (2015), an improved BOVW framework is proposed for remote sensing image retrieval in large-scale image databases. It is shown to have better performance than the standard BOVW framework yet requires less storage.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

Though BOVW and its variants have achieved remarkable performance on various tasks, the major limitation of such approaches is that the spatial distribution of local features is ignored, which has been shown to be very helpful in improving retrieval performance. Therefore, methods have been proposed to incorporate the spatial arrangement of local features. Cao et al. proposed spatial bags of features to encode the geometric information of objects within an image (Cao et al., 2010) for large scale image retrieval. Compared with BOVW, the spatial bags of features work well for image retrieval since the spatial information is encoded.

3.2. Deep learning feature based methods

As mentioned above, image retrieval performance depends greatly on the effectiveness of the features. Deep learning has demonstrated that it is capable of deriving powerful feature representations.

3.2.1. Unsupervised feature learning based methods

Unsupervised feature learning (UFL) aims to directly learn powerful feature representations from large volumes of unlabeled data. It is therefore attractive for remote sensing since the field has relatively little labeled data compared with many other image analysis areas. In (Cheriyadat, 2014), an unsupervised feature learning approach combining SIFT and sparse coding is proposed to learn sparse feature representations for aerial scene classification. Since then, a number of unsupervised feature learning approaches have been proposed for various remote sensing applications like RSIR (Zhou et al., 2015; Wang et al., 2016; Li et al., 2016). In Zhou et al. (2015), an unsupervised feature learning framework based on an auto-encoder is proposed. It consists of the four steps shown in Fig. 3: (1) local feature extraction, (2) unsupervised feature learning, (3) feature encoding and (4) sparse feature extraction and pooling. The local features extracted from the training images are first fed into an auto-encoder network for unsupervised feature learning. Once trained, the auto-encoder network is used to encode the local feature descriptors to obtain the learned feature set. The final feature representation is then generated by pooling the learned feature descriptors into a global feature vector. The learned sparse feature representation shows better performance than handcrafted BOVW features for high-resolution remote sensing imagery retrieval.

In a recent work Wang et al. (2016) developed a novel graph-based learning method for effectively retrieving remote sensing images based on a three-layer framework. This framework integrates the strengths of query expansion and the fusion of holistic and local features, achieving remarkable performance on a benchmark dataset. In Li et al. (2016), a novel content-based remote sensing image retrieval approach is proposed via multiple feature representation and collaborative affinity metric fusion. This approach can generate four types of unsupervised features that outperform several handcrafted features on two publicly available datasets.

In contrast to traditional handcrafted features, unsupervised feature learning based methods directly learn powerful feature representations from the data for RSIR. The performance improvement, however, has been limited. This is because the unsupervised feature learning methods mentioned above are often based on shallow networks (e.g. the three-layer auto-encoder in Zhou et al. (2015)) which cannot learn higher-level information. It is therefore worth investigating deeper networks in order to extract more discriminative features for RSIR.

3.2.2. Convolutional neural networks based methods

Convolutional neural networks (CNNs) have proven to be the most successful deep learning approach to image analysis based on their remarkable performance on the ImageNet (Deng et al., 2009) and other problems. CNNs learn high-level feature representations that are more discriminative than unsupervised features via a hierarchical architecture consisting of convolutional, pooling, and fully-connected layers. However, large numbers of labeled images are needed to train effective CNNs from scratch.

Transfer learning is often used to remedy the lack of enough labeled images by treating the CNNs pre-trained on ImageNet as feature extractors, possibly fine-tuning the pre-trained CNNs on the target dataset to learn domain-specific features. This is very helpful for some domains (e.g. remote sensing) where large-scale publicly available datasets are lacking. In Penatti et al. (2015), the generalization power of deep features extracted by CNNs is

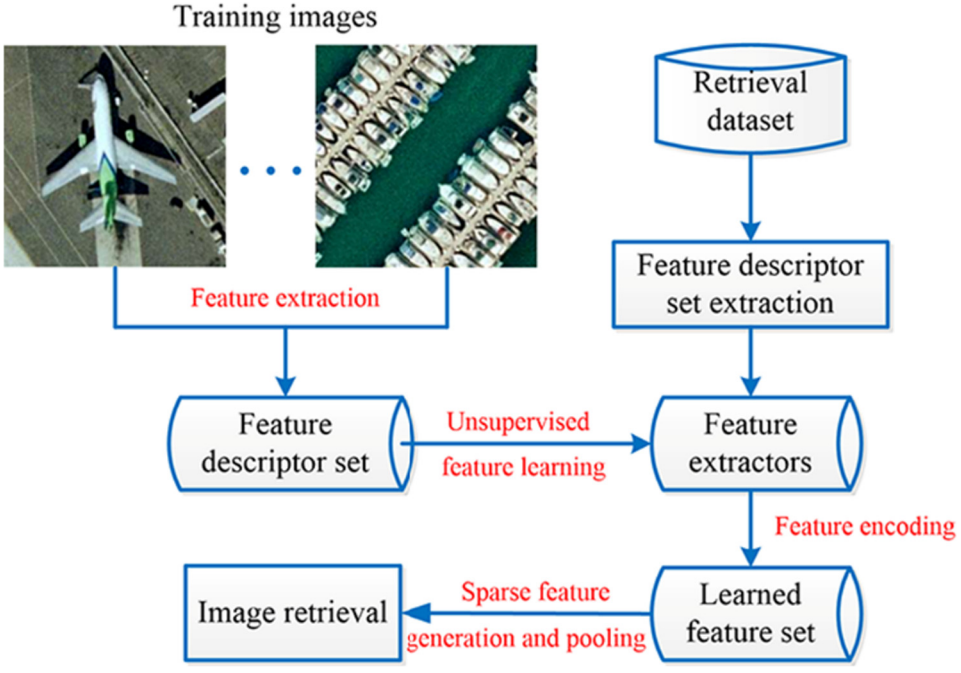

Fig. 3. The flowchart of the unsupervised feature learning method (UFL). The figure is adapted from previous work.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

investigated by transferring deep features from everyday objects to remote sensing. Experiments demonstrate that transfer learning is an effective approach for cross-domain tasks. There are a number of pre-trained CNNs that can be used for transfer learning, such as the baseline model AlexNet (Krizhevsky et al., 2012), the Caffe (Convolutional Architecture for Fast Feature Embedding) reference model (CaffeRef) Jia et al. (2014), the VGG networks (Chatfield et al., 2014) including VGGF, VGGM and VGGS, the VGG-VD networks (Simonyan and Zisserman, 2014) including VGG-VD16 and VGG-VD19, and the recently developed deeper models, GoogLeNet (Szegedy et al., 2015) and Residual networks (ResNet) He et al. (2016) including ResNet-50, ResNet-101 and ResNet-150.

Currently, these pre-trained CNNs and corresponding variants have been widely used for various retrieval tasks ranging from computer vision (Chandrasekhar et al., 2016; Yandex and Lempitsky, 2016; Gordo et al., 2016) to remote sensing (Napoletano, 2016; Zhou et al., 2017). In Napoletano (2016), an extensive evaluation of visual descriptors including handcrafted global and local features as well as CNN features is conducted for content-based retrieval of remote sensing images. The results demonstrate that CNN-based features usually outperform handcrafted features except for remote sensing images that have more heterogeneous content. In Zhou et al., (2017) investigated how to extract powerful feature representations based on the pre-trained CNNs for high-resolution remote sensing imagery retrieval. In one scheme, the fully-connected layers of pre-trained CNNs are regarded as feature extractors. Though these features can achieve remarkable performance, they have 4096 dimensions which presents computational and storage challenges for large-scale RSIR. Therefore, in a second scheme, a novel, low dimensional CNN (LDCNN) is proposed to learn low dimensional features. LDCNN consists of five convolutional layers and an mlpconv layer (three-layer perceptron) as shown in Fig. 4. A global average pooling layer is used to compute the average of each feature map in the previous layer, leading to an n-dimensional feature vector (n is the number of image classes).

It should be noted that although deep learning feature based methods can directly learn powerful feature representations and often outperform handcrafted feature based methods for RSIR, they still have limitations. A large amount of data is needed to train the models. Supervised models such as CNNs require this data to be labeled. There is relatively little labeled data in remote sensing. The other limitation is that "tricks" are often necessary to speed up the training and to achieve satisfactory performance. This makes it difficult and time consuming to determine the optimal model for a particular task.

4. Experiments and results

In this section, we evaluate a large number of state-of-the-art handcrafted and deep learning feature based RSIR methods on the proposed PatternNet dataset. The methods and details used to extract these features are shown in Table 3.

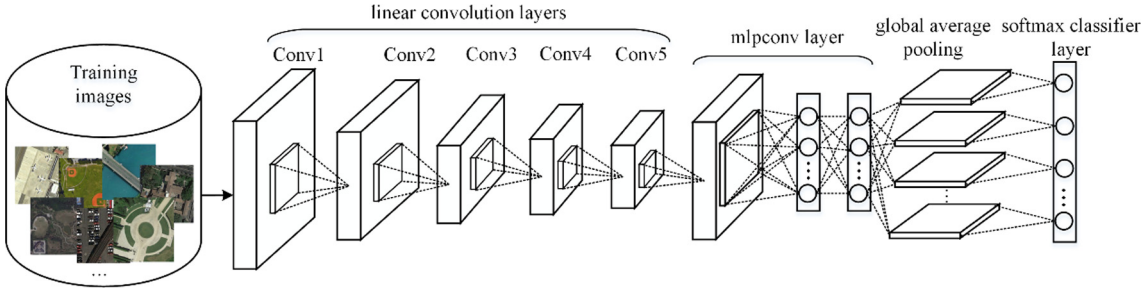

Fig. 4. The flowchart of the low dimensional CNN (LDCNN). The figure is adapted from previous work.

Table 3
The handcrafted and deep learning features and the details used to extract them. K is the size of the dictionaries, M is the dimension of the local descriptors, N is the number of hidden units, and P is the number of image classes.

		Feature dimension	Implementation details
Low-level Handcrafted Features	Simple statistics	2-D	Mean and standard deviation of corresponding grayscale image
	Color histogram	96-D	Quantize each channel of the RGB color space into 32 bins and concatenate the three histograms
	Gabor texture	80-D	Five scales and eight orientations; Gabor filter window size 32×32
	GIST	512-D	Default parameters of original implementation
	LBP	10-D	8 pixel circular neighborhood of radius 1; uniform rotation invariant histogram
	PHOG	680-D	Default parameters of original implementation
Mid-level Handcrafted Features	BOVW	K-D	k-means clustering with $L2$ distance
	VLAD	KM-D	VLAD is extracted with default parameters
	IFK	$2KM$-D	IFK is extracted with default parameters
Unsupervised Feature Learning	UFL	2N-D	UFL is extracted using original implementation
Convolutional Neural Networks	AlexNet	4096-D	Features are extracted from the first and second fully-connected layers.
	CaffeRef	4096-D	
	VGG	4096-D	
	VGG-VD	4096-D	
	GoogLeNet	1024-D	Features are extracted from the last pooling layer.
	ResNet	2048-D	Features are extracted from the fifth pooling layer.
	LDCNN	P-D	VGGF model is used as the basic block.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

4.1. Experimental setup

CNNs require images to have fixed dimensions. The PatternNet images are therefore resized to 227 × 227 pixels for AlexNet and CaffeRef and to 224 × 224 pixels for the other CNNs. In addition, average images provided by the pre-trained CNNs are subtracted from the resized images. Recent work (Zhou et al., 2017) demonstrates that including element-wise rectified linear units (ReLU) as activation functions affects the performance of features extracted from the fully-connected layers. In particular, the features extracted from the first fully-connected layer (Fc1 feature) achieve better performance without the use of ReLU while features extracted from the second fully-connected layer (Fc2 feature) benefit from the use of ReLU. Therefore, in our experiments, Fc1 features are extracted without ReLU and Fc2 features are extracted with ReLU.

With respect to LDCNN, the weights of the five convolutional layers are transferred from VGGF and are also kept fixed during training in order to speed up training. The weights of the mlpconv layer are initialized from a Gaussian distribution (with a mean of 0 and a standard deviation of 0.01). We randomly select 80% of the images from each class of PatternNet as the training set and the remaining 20% of the images are used for retrieval performance evaluation.

For the three mid-level features (i.e. BOVW, VLAD and IFK), the dictionary is constructed by aggregating the 128-D SIFT descriptors extracted at the salient points within the image. The dictionary sizes of VLAD and IFK are set to 64. For BOVW, a set of dictionary sizes (i.e. 64, 128, 256, 512, 1024, 2048, and 4096) are used. For the unsupervised feature learning method (UFL), the number of neural units in the hidden layer is set to 400, 600 and 800, and the sparsity value is set to 0.4 to generate sparse features.

We empirically select $L1$ as the distance function to compute image similarity for the histogram features including color histogram, BOVW and UFL, and select $L2$ as the distance function for the remaining features including simple statistics, Gabor texture, GIST, LBP, PHOG and the CNNs. All the features are $L2$ normalized before the similarity measure is applied. Four commonly used performance metrics, average normalized modified retrieval rank (ANMRR), mean average precision (mAP), precision at k (*P@k*

Table 4
The results of the handcrafted low-level features on PatternNet. For ANMRR, lower values indicate better performance, while for mAP and *P@k*, larger is better. Bold values mean the best performance of each permormance metric.

Features	ANMRR	mAP	*P@5*	*P@10*	*P@50*	*P@100*	*P@1000*
Simple Statistics	0.8968	0.0662	0.0739	0.0741	0.0739	0.0738	0.0701
Color Histogram	0.6697	0.2510	0.7475	0.7032	0.5733	0.5062	0.2349
Gabor Texture	**0.6422**	**0.2769**	**0.8021**	**0.7631**	**0.6393**	**0.5674**	**0.2556**
GIST	0.7511	0.2001	0.6429	0.5957	0.4645	0.4013	0.1773
LBP	0.6470	0.2583	0.6358	0.6027	0.5115	0.4646	0.2505
PHOG	0.8162	0.1312	0.4852	0.4430	0.3376	0.2903	0.1295

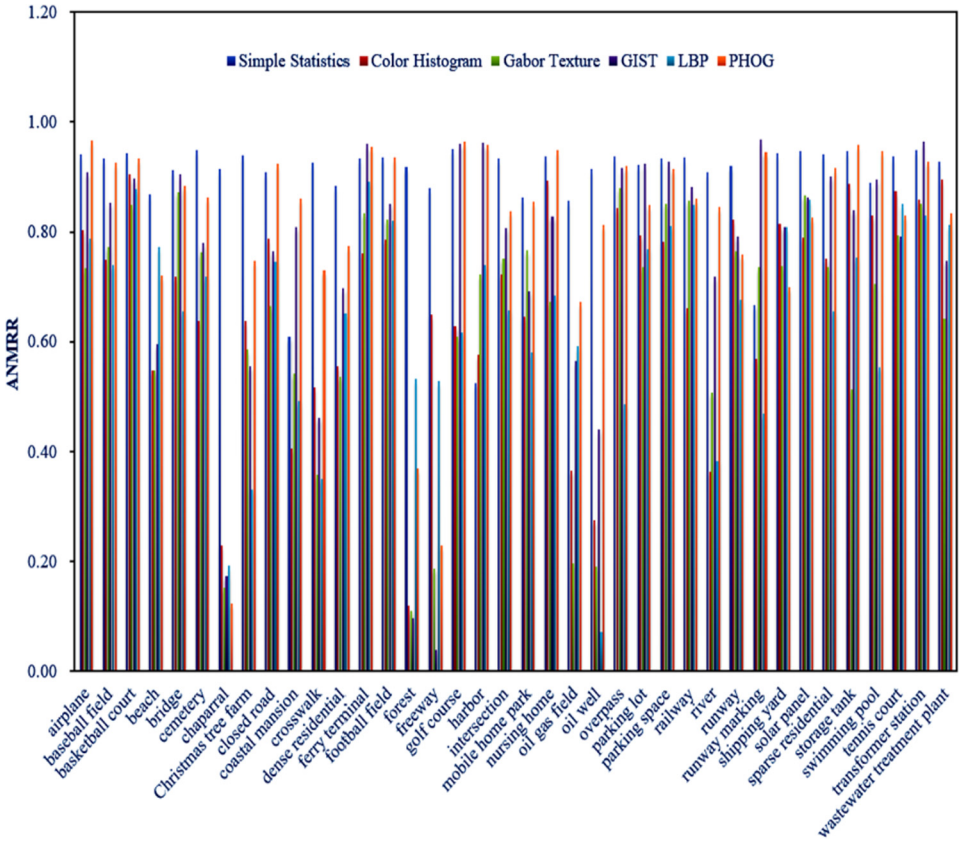

Fig. 5. The results of low-level features for each class in the PatternNet dataset.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

where k is the number of retrieved images), and precision-recall (PR) curves, are used to evaluate the retrieval performance. In the following experiments, each image is taken as a query image, which means the ANMRR, mAP, and *P@k* are the averaged values over all the queries.

4.2. Experimental results

4.2.1. Results of handcrafted low-level features

Table 4 shows the performance of the handcrafted low-level features including simple statistics, color histogram, Gabor texture, GIST, LBP, and PHOG measured using ANMRR, mAP and *P@k* (k = 5, 10, 50, 100, 1000). We can see that Gabor texture features achieve the best performance and simple statistics features achieve the worst. Fig. 5 shows the results of these handcrafted features for each class. Simple statistics and PHOG perform worse than the other features for most of the classes in the PatternNet dataset.

4.2.2. Results of handcrafted mid-level features

The results of the mid-level features are shown in Table 5. For the BOVW features, a set of dictionary sizes (64, 128, 256, 512, 1024, 2048, 4096) are investigated. We can see BOVW with a dictionary size of 128 achieves better performance than BOVW with the other dictionary sizes. In contrast to BOVW, the higher dimensional features VLAD and IFK achieve about 7% and 4% improvement respectively in terms of ANMRR value. Though VLAD and IFK outperform BOVW, the main limitation is that they are of high dimension, resulting in high storage cost and low retrieval efficiency. The results of these mid-level features for each class are shown in Fig. 6. Generally, VLAD is the best mid-level feature for most of the classes.

4.2.3. Results of deep learning features

Table 6 shows the results of the deep learning feature based methods including the unsupervised feature learning method (UFL) and several pre-trained CNNs. For UFL features, we investigate the performance of UFL extracted with different numbers of neural units in the hidden layer. We can see UFL extracted with 400 hidden units performs better than the other UFL configurations. The pre-trained CNN features improve over the performance of UFL by more than 30% in terms of ANMRR values, indicating that supervised CNNs produce more discriminative features.

Table 5
The results of the handcrafted mid-level features on PatternNet. For ANMRR, lower values indicate better performance, while for mAP and *P@k*, larger is better. "BOVW-K" means the BOVW extracted with a dictionary size of K. Bold values mean the best performance of each permormance metric.

Features	ANMRR	mAP	*P@5*	*P@10*	*P@50*	*P@100*	*P@1000*
BOVW-64	0.6593	0.2536	0.5418	0.5158	0.4506	0.4172	0.2430
BOVW-128	0.6393	0.2729	0.5853	0.5564	0.4855	0.4489	0.2583
BOVW-256	0.6573	0.2613	0.5819	0.5498	0.4725	0.4323	0.2450
BOVW-512	0.7696	0.1781	0.3974	0.3596	0.2721	0.2323	0.1638
BOVW-1024	0.8604	0.1111	0.2068	0.1773	0.1213	0.1014	0.0973
BOVW-2048	0.9020	0.0820	0.1425	0.1205	0.0782	0.0639	0.0676
BOVW-4096	0.9231	0.0667	0.0938	0.0795	0.0565	0.0471	0.0525
VLAD	**0.5686**	**0.3367**	**0.6466**	**0.6204**	**0.5620**	**0.5318**	**0.3124**
IFK	0.6016	0.3093	0.6310	0.6049	0.5436	0.5114	0.2874

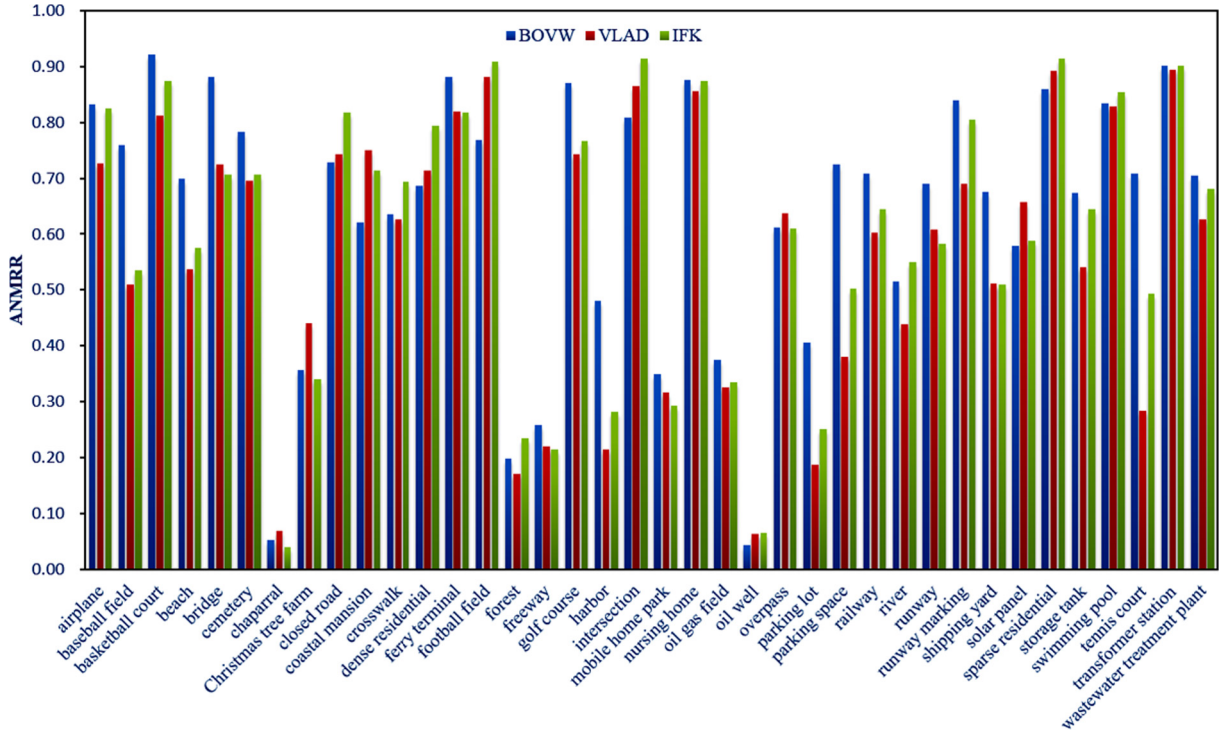

Fig. 6. The results of mid-level features for each class in the PatternNet dataset. For BOVW representation, BOVW-128 is selected.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

The best performance of the various CNNs is achieved by ResNet50, showing that the deeper networks tend to achieve better performance than the shallower networks (i.e. AlexNet, CaffeRef, VGG, VGG-VD and GoogLeNet). However, the increased depth does reduce the performance when the network is too deep (see the performance of ResNet101 and ResNet152). It can also be observed that the features extracted from the second fully-connected layer (Fc2 feature) outperform the features extracted from the first fully-connected layer (Fc1 feature) except for the VD19 network. A possible explanation is that the second fully-connected layer is connected to the classifier layer and hence learns higher-level information. The results of these deep learning features for each class are shown in Fig. 7.

Fig. 8 shows the precision-recall curves for the handcrafted features and deep learning features. For families of features, the configuration that achieves the best performance is selected, namely BOVW-128, UFL-400, AlexNet_Fc2, CaffeRef_Fc2, VGGF_Fc2, VGGM_Fc2, VGGS_Fc2, VD16_Fc2, VD19_Fc1, and ResNet50.

Though the pre-trained CNNs achieve remarkable performance, their features are usually thousands of dimensions which are not compact enough for large-scale RSIR. In contrast, LDCNN is able to generate low-dimensional features. LDCNN is compared with handcrafted low-level and mid-level features, as well as deep learning features including UFL and several pre-trained CNNs on 20% of the PatternNet images. As shown in Table 7, the results indicate that LDCNN outperform the pre-trained CNNs such as VGGF

Table 6
The results of deep learning features on PatternNet. For ANMRR, lower values indicate better performance, while for mAP and *P@k*, larger are better. "UFL-K" means UFL with K neural units in the hidden layer. Bold values mean the best performance of each permormance metric.

Features	ANMRR	mAP	*P@5*	*P@10*	*P@50*	*P@100*	*P@1000*
UFL-400	0.6574	0.2525	0.5937	0.5646	0.4920	0.4516	0.2442
UFL-600	0.6588	0.2508	0.5903	0.5629	0.4898	0.4497	0.2430
UFL-800	0.6595	0.2501	0.5902	0.5619	0.4890	0.4489	0.2426
AlexNet_Fc1	0.3328	0.6003	0.9545	0.9438	0.8986	0.8617	0.4934
AlexNet_Fc2	0.3260	0.6042	0.9448	0.9331	0.8872	0.8529	0.4985
CaffeRef_Fc1	0.3134	0.6221	0.9602	0.9511	0.9121	0.8787	0.5083
CaffeRef_Fc2	0.3133	0.6171	0.9475	0.9370	0.8936	0.8604	0.5086
VD16_Fc1	0.3302	0.6020	0.9388	0.9268	0.8806	0.8459	0.4959
VD16_Fc2	0.3283	0.5986	0.9327	0.9204	0.8740	0.8404	0.4972
VD19_Fc1	0.3423	0.5869	0.9352	0.9210	0.8694	0.8320	0.4865
VD19_Fc2	0.3448	0.5789	0.9253	0.9113	0.8605	0.8247	0.4840
VGGF_Fc1	0.3184	0.6170	0.9592	0.9493	0.9080	0.8738	0.5033
VGGF_Fc2	0.3005	0.6309	0.9544	0.9442	0.9028	0.8714	0.5174
VGGM_Fc1	0.3124	0.6231	0.9576	0.9472	0.9055	0.8717	0.5086
VGGM_Fc2	0.3110	0.6188	0.9511	0.9405	0.8958	0.8627	0.5087
VGGS_Fc1	0.3070	0.6290	0.9595	0.9508	0.9112	0.8784	0.5129
VGGS_Fc2	0.2982	0.6333	0.9547	0.9449	0.9047	0.8734	0.5192
GoogLeNet	0.2983	0.6311	0.9445	0.9331	0.8918	0.8603	0.5202
ResNet50	**0.2606**	**0.6788**	**0.9665**	**0.9594**	**0.9274**	**0.9006**	**0.5533**
ResNet101	0.2624	0.6765	0.9638	0.9551	0.9208	0.8933	0.5525
ResNet152	0.2632	0.6757	0.9635	0.9550	0.9208	0.8939	0.5511

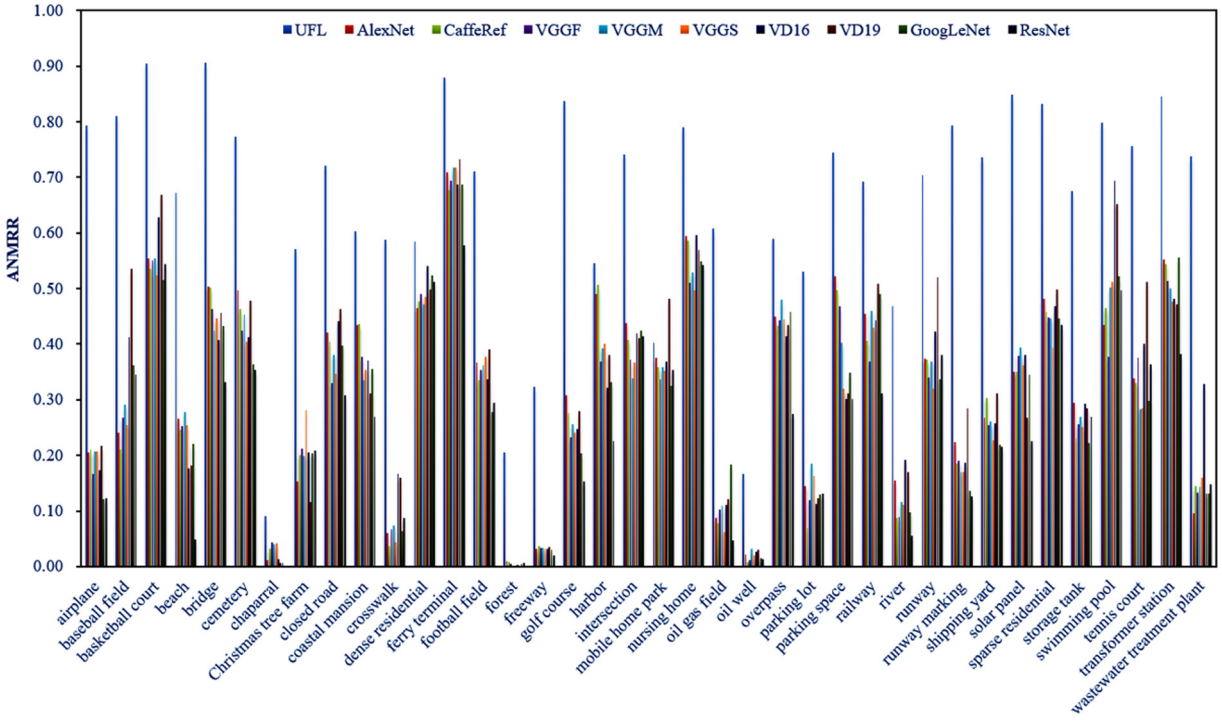

Fig. 7. The results of deep learning features for each class in the PatternNet dataset.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

(the basic block of LDCNN), VGGS and even ResNet50 which achieves the best performance on PatternNet. The features extracted by LDCNN are 38-D which is compact compared to the features extracted by the pre-trained CNNs.

The proposed PatternNet dataset is also compared to the existing datasets by evaluating the performance of the pre-trained CNNs on each dataset, as shown in Table 8. The NWPU45 dataset is not considered due to the time required to evaluate the 18

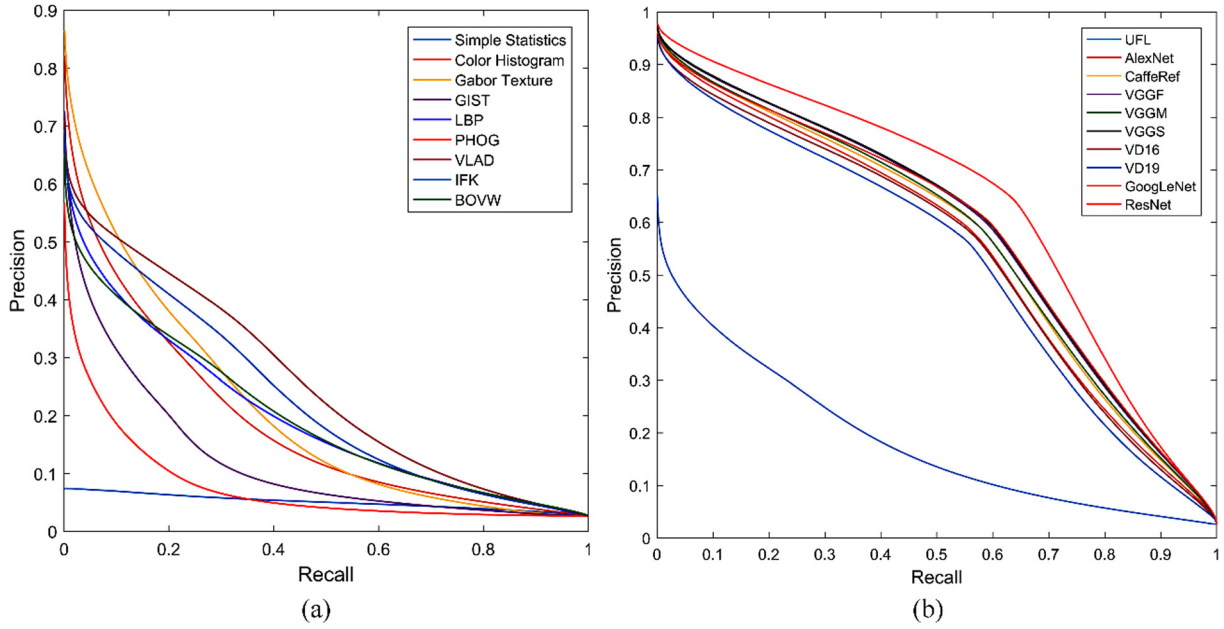

Fig. 8. The precision-recall curves of handcrafted feature based methods and deep learning feature based methods: (a) precision-recall curves of handcrafted features, and (b) precision-recall curves of deep learning features.

Table 7
Comparisons of LDCNN and other features. For ANMRR, lower values indicate better performance, while for mAP and *P@k*, larger is better. The handcrafted features and UFL are extracted under optimal configurations (i.e. the configurations that achieve the best performance on the entire PatternNet dataset).

Features	ANMRR	mAP	*P@5*	*P@10*	*P@50*	*P@100*	*P@1000*
Gabor Texture	0.6439	0.2773	0.6855	0.6278	0.4461	0.3552	0.0899
VLAD	0.5677	0.3410	0.5825	0.5570	0.4757	0.4111	0.1104
UFL	0.6584	0.2535	0.5209	0.4882	0.3811	0.3192	0.0979
VGGF_Fc1	0.3177	0.6195	0.9246	0.9037	0.7926	0.6905	0.1425
VGGF_Fc2	0.2995	0.6337	0.9152	0.8964	0.7999	0.7047	0.1452
VGGS_Fc1	0.3050	0.6328	0.9274	0.9070	0.8003	0.7013	0.1436
VGGS_Fc2	0.2961	0.6374	0.9192	0.9009	0.8021	0.7073	0.1455
ResNet50	0.2584	0.6823	0.9413	0.9241	0.8371	0.7493	0.1464
LDCNN	0.2416	0.6917	0.6681	0.6611	0.6747	0.6880	0.1408

Table 8
The performance of the pre-trained CNNs on the PatternNet and four other datasets. The numbers are ANMRR values. The lower ANMRR values indicate better performance.

	UCMD	RSD	RSSCN7	AID	PatternNet
AlexNet_Fc1	0.411	0.313	0.437	0.537	0.333
AlexNet_Fc2	0.410	0.304	0.446	0.534	0.326
CaffeRef_Fc1	0.397	0.286	0.410	0.518	0.313
CaffeRef_Fc2	0.402	0.283	0.433	0.526	0.313
VD16_Fc1	0.380	0.311	0.436	0.545	0.330
VD16_Fc2	0.394	0.324	0.452	0.568	0.328
VD19_Fc1	0.386	0.326	0.441	0.546	0.342
VD19_Fc2	0.398	0.342	0.457	0.570	0.345
VGGF_Fc1	0.399	0.294	0.419	0.532	0.318
VGGF_Fc2	0.386	0.288	0.440	0.527	0.301
VGGM_Fc1	0.375	0.291	0.412	0.519	0.312
VGGM_Fc2	0.378	0.300	0.440	0.533	0.311
VGGS_Fc1	0.387	0.294	0.408	0.518	0.307
VGGS_Fc2	0.381	0.296	0.441	0.523	0.298
GoogLeNet	0.360	0.299	0.417	0.519	0.298
ResNet50	0.358	0.230	0.405	0.484	0.261
ResNet101	0.356	0.248	0.420	0.491	0.262
ResNet150	0.362	0.251	0.424	0.493	0.263

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004

features. The pre-trained CNNs achieve much better performance on PatternNet than on the existing datasets (with the exception of the RSD dataset which is small and thus not very challenging). The decreased performance on the UCMD, RSSCN7, and AID datasets is due to their images having a relatively large amount of background compared to PatternNet.

5. Conclusions

We present PatternNet, the largest publicly available remotely sensed evaluation dataset constructed for RSIR. We expect PatternNet help advance the state-of-the-art in RSIR, particularly deep learning based methods which require large amounts of labeled training data. We also surveyed a large number of RSIR approaches including traditional handcrafted features and recent deep learning features and evaluated them on PatternNet to establish baseline results to inform future research.

Acknowledgements

The authors would like to thank Paolo Napoletano for the code used in the performance evaluation. This work was supported by the National Key Technologies Research and Development Program (2016YFB0502603). Fundamental Research Funds for the Central Universities (2042016kf0179 and 2042016kf1019), Wuhan Chen Guang Project (2016070204010114), Guangzhou science and technology project (201604020070), Special task of technical innovation in Hubei Province (2016AAA018), and the National Natural Science Foundation of China (61671332, 41771452 and 41771454).

References

Agouris, P., Carswell, J., Stefanidis, A., 1999. An environment for content-based image retrieval from large spatial databases. ISPRS J. Photogramm. Remote Sens. 54, 263–272.

Alshehhi, R., Marpu, P.R., Woon, W.L., Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 130, 139–149.

Aptoula, E., 2014. Remote sensing image retrieval with global morphological texture descriptors. IEEE Trans. Geosci. Remote Sens. 52, 3023–3034.

Aptoula, E., 2014. Bag of morphological words for content-based geographical retrieval. In: Content-Based Multimed. Index. (CBMI), 2014 12th Int. Work.. IEEE, pp. 1–5.

Bosch, A., Zisserman, A., Munoz, X., 2007. Representing shape with a spatial pyramid kernel. In: Proc. 6th ACM Int. Conf. Image Video Retr.. ACM, pp. 401–408.

Bosilj, P., Aptoula, E., Lefèvre, S., Kijak, E., 2016. Retrieval of remote sensing images with pattern spectra descriptors. ISPRS Int. J. Geo-Inform. 5, 228.

Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L., 2010. Spatial-bag-of-features. In: Comput. Vis. Pattern Recognit. (CVPR), 2010 IEEE Conf.. IEEE, pp. 3352–3359.

Chandrasekhar, V., Lin, J., Morère, O., Goh, H., Veillard, A., 2016. A practical guide to CNNs and Fisher Vectors for image instance retrieval. Signal Process. 128, 426–439.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets, arXiv preprint arXiv:1405.3531.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE.

Cheriyadat, A.M., 2014. Unsupervised feature learning for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 52, 439–451.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Comput. Vis. Pattern Recognition, 2005. CVPR 2005. IEEE Conf.. IEEE, pp. 886–893.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conf. Comput. Vis. Pattern Recognit.. IEEE, pp. 2–9.

Gordo, A., Almazán, J., Revaud, J., Larlus, D., 2016. Deep image retrieval: learning global representations for image search. In: Eur. Conf. Comput. Vis.. Springer, pp. 241–257.

Han, W., Feng, R., Wang, L., Cheng, Y., 2017. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010. Aggregating local descriptors into a compact representation. In: IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3304–3311.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. In: ACM Int. Conf. Multimed.. ACM, pp. 675–678.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1–9.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, Y., Zhang, Y., Tao, C., Zhu, H., 2016. Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. Remote Sens. 8, 709.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Ma, X., Wang, H., Wang, J., 2016. Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning. ISPRS J. Photogramm. Remote Sens. 120, 99–107.

Napoletano, P., 2016. Visual descriptors for content-based retrieval of remote sensing images, arXiv preprint arXiv:1602.00970 (2016).

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24, 971–987.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. 42, 145–175.

Özkan, S., Ateş, T., Tola, E., Soysal, M., Esen, E., 2014. Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization. IEEE Geosci. Remote Sens. Lett. 11, 1996–2000.

Penatti, O.A.B., Nogueira, K., dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work, pp. 44–51.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the Fisher Kernel for large-scale image classification. In: Eur. Conf. Comput. Vis.. Springer, pp. 143–156.

Scott, G.J., Klaric, M.N., Davis, C.H., Shyu, C.R., 2011. Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases. IEEE Trans. Geosci. Remote Sens. 49, 1603–1616.

Shao, Z., Zhou, W., Zhang, L., Hou, J., 2014. Improved color texture descriptors for remote sensing image retrieval. J. Appl. Remote Sens. 8, 83584.

Shao, Z., Zhou, W., Cheng, Q., Diao, C., Zhang, L., 2015. An effective hyperspectral image retrieval method using integrated spectral and textural features. Sens. Rev. 35, 274–281.

Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. In: Comput. Vis. Pattern Recognition, 2007. CVPR 2007. IEEE Conf.. IEEE, pp. 1–8.

Sheng, G., Yang, W., Xu, T., Sun, H., 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int. J. Remote Sens. 33, 2395–2412.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556.

Sivic, J., Zisserman, A., 2003. Video Google: a text retrieval approach to object matching in videos. In: Proc. Ninth IEEE Int. Conf. Comput. Vis., pp. 1470–1477.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–9.

Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J., 2014. Deep learning for content-based image retrieval: a comprehensive study. In: Proc. 22nd ACM Int. Conf. Multimed.. ACM, pp. 157–166.

Wang, Y., Zhang, L., Tong, X., Zhang, L., Zhang, Z., Liu, H., Xing, X., Mathiopoulos, P.T., 2016. A three-layered graph-based learning approach for remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 54, 6020–6034.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens.

Yandex, A.B., Lempitsky, V., 2016. Aggregating local deep features for image retrieval. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 1269–1277.

Yang, Y., Newsam, S., 2013. Geographic image retrieval using local invariant features. IEEE Trans. Geosci. Remote Sens. 51, 818–832.

Yang, J., Liu, J., Dai, Q., 2015. An improved Bag-of-Words framework for remote sensing image retrieval in large-scale image databases. Int. J. Digit. Earth. 8, 273–292.

Zhang, M., Zhang, L., Mathiopoulos, P.T., Ding, Y., Wang, H., 2013. Perception-based shape retrieval for 3D building models. ISPRS J. Photogramm. Remote Sens. 75, 76–91.

Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017. Contextually guided very-high-resolution imagery classification with semantic segments. ISPRS J. Photogramm. Remote Sens. 132, 48–60.

Zhou, W., Shao, Z., Diao, C., Cheng, Q., 2015. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. Remote Sens. Lett. 6, 775–783.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2017. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. Remote Sens. 9, 489.

Zhu, X., Shao, Z., 2011. Using no-parameter statistic features for texture image retrieval. Sens. Rev. 31, 144–153.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12, 2321–2325.

Please cite this article in press as: Zhou, W., et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.01.004