Patch-Based Discriminative Learning for Remote Sensing Scene Classification

Usman Muhammad [1], Md Ziaul Hoque [2,2], Weiqiang Wang [1], and Mourad Oussalah [1]

[1]Affiliation not available
[2]University of Oulu

November 8, 2023

Abstract

The bag-of-words (BoW) model is one of the most popular representation methods for image classification. However, the lack of spatial information, the intra-class diversity, and the inter-class similarity among scene categories impair its performance in the remote-sensing domain. To alleviate these issues, this paper proposes to explore the spatial dependencies between different image regions and introduces patch-based discriminative learning (PBDL) for remote-sensing scene classification. Particularly, the proposed method employs multi-level feature learning based on small, medium, and large neighborhood regions to enhance the discriminative power of image representation. To achieve this, image patches are selected through a fixed-size sliding window and sampling redundancy, a novel concept, is developed to minimize the redundant features while sustaining the relevant features for the model. Apart from multi-level learning, we explicitly impose image pyramids to magnify the visual information of the scene images and optimize their position and scale parameters locally. Motivated by this, a local descriptor is exploited to extract multi-level and multi-scale features that we represent in terms of codewords histogram by performing k-means clustering. Finally, a simple fusion strategy is proposed to balance the contribution of individual features, and the fused features are incorporated into a Bidirectional Long Short-Term Memory (BiLSTM) network for classification. Experimental results on NWPU-RESISC45, AID, UC-Merced, and WHU-RS datasets demonstrate that the proposed approach not only surpasses the conventional bag-of-words approaches but also yields significantly higher classification performance than the existing state-of-the-art deep learning methods used nowadays.

Patch-Based Discriminative Learning for Remote Sensing Scene Classification

Usman Muhammad[a], Md Ziaul Hoque[a], Weiqiang Wang[b] and Mourad Oussalah[a,c]

[a]Center for Machine Vision and Signal Analysis, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland
[b]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
[c]Medical Imaging, Physics, and Technology (MIPT), Faculty of Medicine, University of Oulu, Finland

ARTICLE INFO

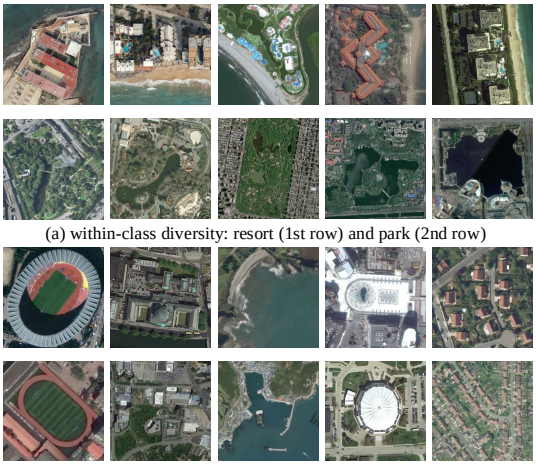
Keywords:
Scene classification
Bag-of-words model
Gaussian pyramids
Patch-based learning
BiLSTM

ABSTRACT

The bag-of-words (BoW) model is one of the most popular representation methods for image classification. However, the lack of spatial information, the intra-class diversity, and the inter-class similarity among scene categories impair its performance in the remote-sensing domain. To alleviate these issues, this paper proposes to explore the spatial dependencies between different image regions and introduces patch-based discriminative learning (PBDL) for remote-sensing scene classification. Particularly, the proposed method employs multi-level feature learning based on small, medium, and large neighborhood regions to enhance the discriminative power of image representation. To achieve this, image patches are selected through a fixed-size sliding window and *sampling redundancy*, a novel concept, is developed to minimize the redundant features while sustaining the relevant features for the model. Apart from multi-level learning, we explicitly impose image pyramids to magnify the visual information of the scene images and optimize their position and scale parameters locally. Motivated by this, a local descriptor is exploited to extract multi-level and multi-scale features that we represent in terms of *codewords* histogram by performing k-means clustering. Finally, a simple fusion strategy is proposed to balance the contribution of individual features, and the fused features are incorporated into a Bidirectional Long Short-Term Memory (BiLSTM) network for classification. Experimental results on NWPU-RESISC45, AID, UC-Merced, and WHU-RS datasets demonstrate that the proposed approach not only surpasses the conventional bag-of-words approaches but also yields significantly higher classification performance than the existing state-of-the-art deep learning methods used nowadays.

1. Introduction

Remote sensing has received unprecedented attention due to its role in mapping land cover, geographic image retrieval, natural hazards detection, and monitoring changes in land cover. The currently available remote sensing satellites and instruments (e.g., IKONOS, unmanned aerial vehicles (UAVs), synthetic aperture radar, etc.,) for observing the Earth not only provide high-resolution scene images but also give us an opportunity to study the spatial information with a fine-grained detail [1]. However, within-class diversity and between-class similarity among scene categories are the main challenges that make it extremely difficult to distinguish the scene classes. For instance, as shown in Fig.1 (a), a large intra-class or within-class diversity can be observed such as the resort scenes appearing in different building styles but all of them belong to the same class. Similarly, the park scenes show large differences within the same semantic class. In addition, satellite imagery data can be influenced by differences in color or radiation intensity due to different factors such as weather, cloud coverage, mist, etc., which may also cause within-class diversity [2, 3]. In terms of inter-class or between-class similarity, the challenge exists in the appearance of the same ground objects within different scene classes as illustrated in Fig.1 (b). One can see that stadium and playground are different classes but represent the high semantic overlapping between scene categories. Here, the "scenes" belong to a different type of subareas extracted from large satellite images. These subareas could be different types of land covers or objects and possess specific semantic



(a) within-class diversity: resort (1st row) and park (2nd row)

(b) between-class similarity: stadium vs. playground, church vs. commercial, beach vs. port, square vs. center, medium residential vs. dense (from top to bottom)

Figure 1: The challenging scene images of AID dataset [2]. (a) the intra-class diversity and (b) inter-class similarity are the main obstacles that limit the scene classification performance. This encourages us to learn multi-level spatial features that have small within-class scatter but large between-class separation.

muhammad.usman@oulu.fi (U. Muhammad)
ORCID(s):

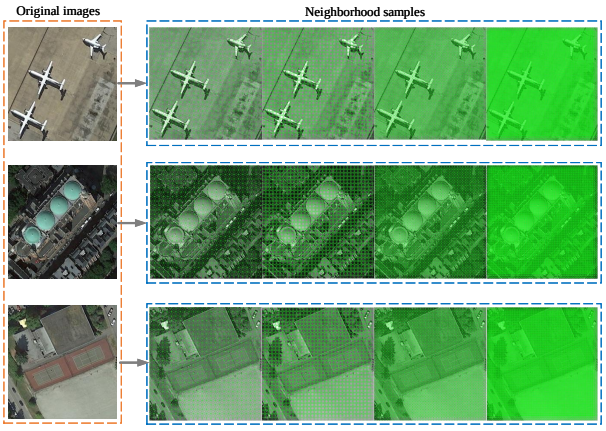


Figure 2: The main idea of the proposed work. Given a remote sensing image, BoW model is designed with different image patches to incorporate spatial information. Left column: Example images from NWPU dataset. Right column: SURF features of light, medium and dark green colors represent different spatial locations. These samples can significantly improve the scene classification performance.

meaning, such as commercial area, dense residential, sparse residential, and parking lot in a typical urban area of satellite image [2]. With the development of modern technologies, scene classification has been an active research field, and correctly labeling it to a predefined class is still a challenging task.

In the early days, most of the approaches focused on hand-crafted features, which can be computed based on shape, color, or textual characteristics where commonly used descriptors are local binary patterns (LBPs) [4], scale invariant feature transform [5], color Histogram [6], and histogram oriented gradients (HOG) [7]. A major shortcoming of these low-level descriptors is their inability to fulfill scene understanding due to the high diversity and non-homogeneous spatial distributions of the scene classes. In comparison to handcrafted features, the bag-of-words (BoW) model is one of the famous mid-level (global) representations and is extremely popular in image analysis and classification, while providing an efficient solution for aerial or satellite image scene classification. It was first proposed for text analysis and then extended to images by a spatial pyramid method (SPM) because the vanilla BoW model does not consider spatial and structural information. The SPM method divides the images into several parts and computes BoW histograms from each part based on the structure of local features. The histograms are then concatenated from all image parts to make the final representation [8]. Although these mid-level features are highly efficient, they may not be able to characterize detailed structures and distinct patterns. For instance, some scene classes are represented mainly by individual objects, e.g., runway and airport in remote-sensing datasets. As a result, the performance of BoW model remains limited when dealing with complex and challenging scene images.

Recently, deep learning-based methods have been successfully utilized in scene classification and proven to be promising in extracting high-level features. For instance, Shi *et al.* [9] proposed a multi-level feature fusion method based on a lightweight convolution neural network to improve the classification performance of scene images. Yuan *et al.* [10] proposed a multi-subset feature fusion method to integrate the global and local information of the deep features. A dual-channel spectral feature extraction network is introduced in [11] that employs a 3D convolution kernel directly to extract multi-scale spatial features and then an adaptive fusion of spectral and spatial features is performed to improve the performance. These methods have proved the importance of deep learning-based feature fusion, but patch-based global feature learning has been never deeply investigated in the BoW framework. Moreover, deep learning-based methods generally analyze an individual patch and treat different scene categories equally. Thus, they fail to capture contextual dependencies for better representation. One of the reasons is that natural images can be mainly captured by cameras with manual or auto-focus options and it makes them center-biased [12]. However, in the case of remote sensing scene classification, images are usually captured overhead. Therefore, using a CNN as a "black box" to classify remote sensing images may be not good enough for complex scenes. Even though several works [13, 14] attempted to focus on critical local image patches, the role of the spatial dependency among objects in remote sensing scene classification task remains an unsolved problem [15].

In general, patch sampling or feature learning is a critical component for building up an intelligent system either for the CNN model or BoW-based approaches. Ideally, special attention should be paid on the image patches that are the most informative for classification. This is due to the fact that objects can appear at any location in the image. Recent studies address this issue by sampling feature points based on a regular dense grid [16] or a random strategy [17] because there is no clear consensus about which sampling strategy is suitable for natural scene images. Although multiscale keypoint detectors (Harris-affine, Laplacian of Gaussian, etc.) as samplers [18] are well studied in the computer vision community, they were not designed to find the most informative patches for scene image classification [17]. In this paper, instead of working towards a new CNN model or a local descriptor, we introduce patch-based discriminative learning (PBDL) to extract image features region by region based on small, medium, and large neighborhood patches to fully exploit the spatial structure information in the BoW model. This is motivated by the fact that different patch sizes still exhibit good learning ability of spatial dependencies between image region features that may help to interpret the scene [19]. Figure 2 illustrates the extracted regions used in our work. Moreover, the proposed method also magnifies the visual information by utilizing Gaussian pyramids in a scale-space setting to improve the classification performance. Although the proposed multi-level learning is based on different image patch sizes, spatial receptive fields

may overlap due to unique nature of remote sensing scene images (e.g. buildings, fields, etc.). Thus, we also consider the *sampling redundancy* problem to minimize the presence of nearby or neighboring pixels. We show that overlapping pixels can be minimized by setting pixel stride equal to the pixel width of the feature window.

Next, we balance the contribution of individual patch features by proposing a simple fusion strategy based on two motivations. Firstly, the proposed method introduces a simple fusion strategy that can surpass the previous performance without utilizing state-of-the-art fusion methods such as DCA [20], PCA [21], CCA [22], etc., as previously utilized in remote sensing domain (we further discuss this aspect in section 4.3). The second motivation is to evade the disadvantages of traditional dimensionality reduction techniques such as principle component analysis (PCA): its data-dependent characteristic, the computational burden of diagonalizing the covariance matrix, and the lack of guarantee that distances in the original and projected spaces are well retained. Finally, the BiLSTM network is adopted after combining small, medium, and large scale spatial and visual histograms to classify the scene images. We demonstrate that the collaborative fusion of different regions (patch sizes) addresses the problem of *intra-class* difference, and the aggregated multi-scale features in scale-space pyramids can be used to solve the problem of *inter-class* similarity. To this end, our main contributions in this paper are summarized as follows:

1. We present patch-based discriminative learning to combine all the surrounding features into a new single vector and address the problem of *intra-class* diversity and *inter-class* similarity.
2. We demonstrate the effectiveness of patch-based learning in the BoW model for the first time. Our method suggests that exploring visual descriptor on image regions independently can be more effective than random sampling for the remote sensing scene classification.
3. To enlarge the visual information, smoothing and stacking is performed by convolving the image with Gaussian second derivatives. In this way, we integrate the fixed regions (patches) into multiple downscaled versions of the input image in a scale-space pyramid. By doing so, we explore more content and important information.
4. The proposed method not only surpasses the previous BoW methods but also several state-of-the-art deep learning-based methods on four publicly available datasets and achieves state-of-the-art results.

The rest of this work is organized as follows. Section 2 discusses the related literature work of this study. Section 3 introduces the proposed PBDL for remote sensing scene classification. Section 4 shows the experimental results of the proposed PBDL on several public benchmark datasets. Section 5 summarizes the entire work and gives suggestions for future research.

2. Literature Review

In the early 1970s, most of the early methods in remote sensing image analysis focused on per-pixel analysis, through labeling each pixel in the satellite images (such as the Landsat series) with a semantic class, because the spatial resolution of Landsat images acquired by satellite sensor is very low- the size of a pixel is close to the sizes of the objects of interest [3]. With the advances in remote sensing technology, the spatial resolution of remote sensing images is increasingly finer than the typical object of interest, and the objects are usually composed of many pixels, such that single pixels lost their semantic meanings. In such cases, it is difficult or sometimes impoverished to recognize scene images at the pixel level solely. In 2001, Blaschke and Strobl [23] raised the critical question "What's wrong with pixels?" to conclude that analyzing remote sensing images at the object level is more efficient rather than the statistical analysis of single pixels. Afterward, a new paradigm of approaches to analyze remote sensing images at the object level has dominated for the last two decades [3].

However, pixel and object-level classification methods may not be sufficient to always classify them correctly because pixel-based identification tasks carry little semantic meanings. Under this circumstances, semantic-level remote sensing image scene classification seeks to classify each given remote sensing image patch into a semantic class that contains explicit semantic classes (e.g., commercial area, industrial area, and residential area). Thus, a majority of remote sensing image scene classification is developed and categorized into three main classes according to the features they used: human engineering-based methods, unsupervised feature learning or global-based methods, and deep feature learning-based methods. The early works for scene classification require a considerable amount of engineering skills and are mainly based on handcrafted descriptors [24, 25, 4, 6]. These methods mainly focus on texture, color histograms, shape, spatial and spectral information, and are invariant to translation and rotation about the viewing axis.

In brief, handcrafted features have their benefits and disadvantages. For instance, the color features are more convenient to extract in comparison with texture and shape features. The color histograms and color moments provide discriminative features and can be computed based on local descriptors such as local binary patterns (LBPs) [4], scale invariant feature transform (SIFT) [5], color histogram [6], and histogram oriented gradients (HOG) [7]. Although color-based histograms are easy to compute, these methods do not convey spatial information and the high resolution of scene images makes it very difficult to distinguish the images with the same colors. Yu *et al.* [26] proposed a new descriptor called color-texture-structure (CTS) to encode color, texture, and structure features. In their work, a dense approach is used to build the hierarchical representation of the images. Finally, the co-occurrence patterns of regions are extracted and the local descriptors are encoded to test the discriminative capability. Tokarczyk *et al.* [25] proposed to use integral images and extract discriminative textures at different scale

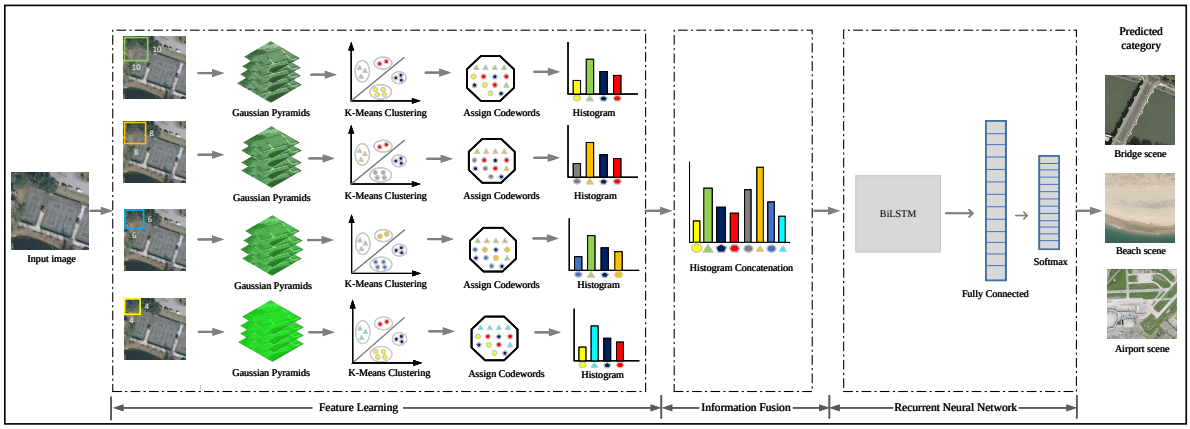


Figure 3: Flowchart of the proposed method. The local patches are selected by a fixed-size sliding window, where green, orange, blue, and yellow rectangles represent the patch sizes of 10×10, 8×8, 6×6, and 4×4, respectively. Then, the dense interest points are extracted with Gaussian second derivatives without changing the size of the original image and encoded to a specific codeword through the k-means clustering process. Finally, a concatenated histogram is used as an input for training the BiLSTM network.

levels of scene images. The features are named Randomized Quasi-Exhaustive (RQE) which are capable of covering a large range of texture frequencies. The main advantage of extracting these spatial cues such as color, texture, or spatial information is that they can be directly utilized by classifiers for scene classification. On the other hand, every individual cue focuses only on one single type of feature, so it remains challenging or inadequate to illustrate the content of the entire scene image. To overcome this limitation, Chen *et al.* [24] proposed a combination of different features such as color, structure, and texture features. To perform classification, the k-nearest-neighbor (KNN) classifier and the support vector machine classifiers (SVM) are employed and the decision level fusion is performed to improve the performance of scene images. Zhang *et al.* [27] focused on the variable selection process based on random forests to improve land cover classification.

To further improve the robustness of handcrafted descriptors, the bag-of-words (BoW) framework has made significant progress for remote sensing image scene classification [28]. By learning global features, Khan *et al.* [29] investigated multiple hand-crafted color features in the bag-of-word model. In their work, color and shape cues are used to enhance the performance of the model. Yang *et al.* [30] utilized the BoW model using the spatial cooccurrence kernel, where two spatial extensions are proposed to emphasize the importance of spatial structure in geographic data. Vigo *et al.* [31] proved that incorporating color and shape in both feature detection and extraction significantly improves bag-of-words based image representation. Sande *et al.* [32] proposed a detailed study about the invariance properties of color descriptors. They concluded that the addition of color descriptors over SIFT increases the classification accuracy by 8 percent. Lazebnik *et al.* [8] proposed a spatially hierarchical pooling stage to form the spatial pyramid method (SPM). To improve the SPM pooling stage, sparse codes (SC) of SIFT features are merged into the traditional SPM [33]. Although, researchers have proposed several methods to achieve good performance for land use classification, especially compared to handcrafted feature-based methods, one of the major disadvantages of BoW is that it neglects the spatial relationships among the patches, and the performance remains unclear or localization is not well understood.

Recently, most of the current state-of-the-art approaches generally rely on end-to-end learning to obtain good feature representations. Specifically, the use of convolutional neural networks (CNN) is the state-of-the-art framework for scene image classification. In deep learning models, convolutional layers convolve the local image regions independently, and pass their results to the next layer, whereas pooling layers summarize the dimensions of data. Due to the wide range of image resolution and various scales of detail textures, fixed-sized kernels are inadequate to extract scene features of different scales. Therefore, the focus of current literature has been shifted to multi-scale and fusion methods in the scene image classification domain, and existing deep learning methods are making full use of multi-scale information and fusion for better representation. For instance, Ghanbari *et al.* [34] proposed a multi-scale method called dense-global-residual network to reduce the loss of spatial information and enhance the context information. The authors used a residual network to extract the features and a global spatial pyramid pooling module to obtain dense multi-scale features at different levels. Zuo *et al.* [35] proposed a convolutional recurrent neural network to learn the spatial dependencies between image regions and enhance the discriminative power of image representation. The authors trained their model in an end-to-end manner where CNN layers are formed to generate mid-level features and RNN is used for learning contextual dependencies. Huang *et al.* [36]

proposed an end-to-end deep learning model with multi-scale feature fusion, channel-spatial attention, and a label correlation extraction module. Specifically, a channel-spatial attention mechanism is used to fuse and refine multi-scale features from different layers of the CNN model.

Li *et al.* [37] proposed an adaptive multi-layer feature fusion model to fuse different convolutional features with feature selection operation, rather than simple concatenation. The authors claimed that their proposed method is flexible and can be embedded into other neural architectures. Few-shot scene classification is introduced by proposing an end-to-end network, called discriminative learning of adaptive match network (DLA-MatchNet) in [38]. The authors addressed the issues of the large intraclass variances and interclass similarity by introducing the attention mechanism into the feature learning process. In this way, discriminative regions were extracted, which helps the classification model to emphasize valuable feature information. Xiwen *et al.* [39] proposed a unified annotation framework based on a stacked discriminative sparse autoencoder (SDSAE) and weakly supervised feature transferring. The results demonstrate the effectiveness of weakly supervised semantic annotation in remote sensing scene classification. Rosier *et al.* [40] find that fusing Earth observation and socioeconomic data lead to increases the accuracy of urban land use classification.

Due to the wide range of image resolution and various scales of detail textures, fixed-sized CNN kernels are inadequate to extract scene features of different scales. Therefore, the focus has been shifted to multi-scale and fusion methods in the scene image classification domain, and existing deep learning methods are making full use of multi-scale information and fusion for better representation. However, we pay particular attention to the previous work [21] where the authors claimed that a simple combination strategy achieves less than 1% accuracy when the fusion of deep features (AlexNet, VGG-M, VGG-S and CaffeNet) is applied. Thus, a natural question arises: can we combine different region features effectively and efficiently to address scene image classification? With the exception [20], to our knowledge, this question still remains mostly unanswered. Experimental results on four public remote sensing image datasets demonstrate that combining the proposed discriminative regions can improve performance up to 20%, 15%, 10% and 6% for NWPU, AID, WHU-RS and UC Merced datasets, respectively.

3. The Proposed Method

The proposed approach is divided into four indispensable components: (a) estimation of patch-based regions (b) scale-space representation (c) information fusion and (d) a BiLSTM based sub-network for classification purpose. We first describe the procedure of patch-based learning. Next, we describe the proposed fusion along the classification process of BiLSTM network. The overall procedure of the proposed approach is illustrated in Fig 3.

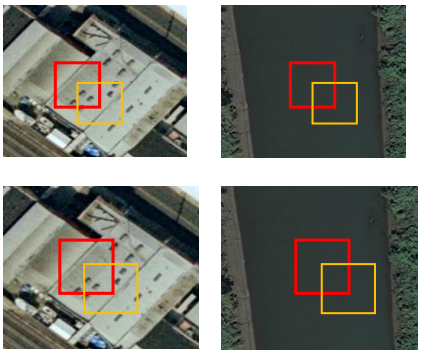


Figure 4: Illustration of overlapping sample windows at two sizes. In both images, the pixel offset kept the same between the yellow window and the red window. A large overlapping can be observed in the bottom images.

3.1. Features extraction using patch-based regions

In order to explore the spatial relationship between scenes or sub-scenes, we propose to extract multi-level features with the objective that different regions contain discriminative characteristics that can be used to extract more meaningful information. Based on our observation, the size of the neighborhood has a great impact on the scene representations and classification performance. To demonstrate this, we first define a region over the entire image, where the patch sizes used are (4×4), (6×6), (8×8), (10×10), with the sliding steps corresponding to patch sizes. Here, the definition of different neighborhood sizes is considered to be small, medium, or large regions. More specifically, given an image $I : \Omega \longmapsto R^Q$ where $\Omega = \{0, 1, ..., G-1\} \times \{0, 1, ..., H-1\}$, G and H represent the number of rows and columns of an image, respectively. The sampling patch g is the number of sampled grids divided by the number of pixels in an image; the objective is to determine a subset D of Ω for a given sampling patch g, such that:

$$D = \left\{ c | \, c \, \epsilon \, \Omega, \quad j(x) \; is \; informative, \quad \frac{\#C}{G \times H} = g \right\} \qquad (1)$$

where c denotes the local patches (i.e., grids) defined at the image pixel x, $j(x)$ is the response map at x and $\#C$ represents the number of grids. In our work, we set the size of the sampling patch g to be the number of sampled patches partitioned by the number of pixels in an image. Therefore, an image is represented by the same number of patches that defines the representative area of the same size. Thus, four kinds of grid sampling size (g) as mentioned above are used for each image to ensure that the output is full of content information.

Moreover, we adopt multi-scale representation by utilizing different scale σ sizes. However, the natural question is whether the large scale images can provide salient features from every scale σ, or small scaled images are enough for the classifier. For instance, taking an equal 4×4 pixel stride at the lowest scale $\sigma = 1.6$, should the proposed sampling at a 4 pixel stride be able to recognize objects at a wide variety

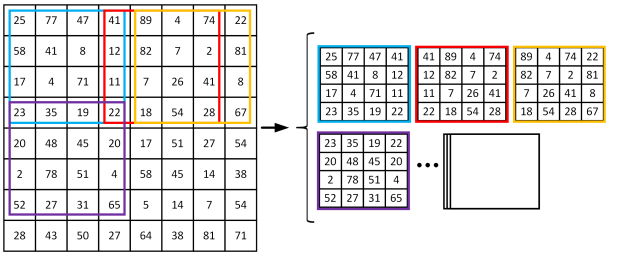


Figure 5: Predefined patch (4 × 4) size before representing features over entire image.

of scales? Figure 4 interprets the ambiguity. It can be easily observed how large scales exhibit larger redundancy. Both images share significant overlapping even at the large scale. Can it consider redundancy between the smooth pair of samples? This is an open question that must be addressed during feature extraction stage. Recent studies generally address this issue by sampling the feature points either uniformly or randomly [41, 42]. For uniform sampling, local patches are sampled densely within regular sampling grids across an image with certain pixel spacing. For instance, an example image with the neighborhood patch (4 × 4) size is provided in Fig.5 to show how the local descriptor can be exploited using fixed-size window with a constant stride 1. Such an approach would be sub-optimal if:

- There is not much spatial information available at the larger scales. This suggests that larger scales should not be weighted equally.
- A large number of scale images provide more redundancy at the same pixel stride. Since the fixed pixel stride can share overlapping, spatial closeness must be taken into account before employing the local descriptor.

Perhaps surprisingly, the proposed strategy has the potential to be more efficient, exploring the salient features at a wide variety of scales. Specifically, if the proposed sampling uses a 4 pixel strides for $\sigma = 1.6$, then it would also utilize other pixel stride of 6, 8 and 10 for higher scales $\sigma = 6.0$ to avoid ambiguity. Thus, the proposed sampling method overcome the overlapping or redundancy problem by, first, setting the different patch regions, e,g., (4×4), (6×6), (8×8), (10×10), and then keeping the pixel stride equal to the pixel width of the feature window (i.e., 4, 6, 8 and 10). By doing this, a bias, if exist at all, would then only be applicable at the borders of such a region, but not for the central pixels (we further discuss this argument in section 4.3).

3.2. Scale-Space Representation

To achieve multi-scale information of each region, we propose to use multi-scale filtering motivated by the fact that it can adaptively integrate the edges of small and large structures referring as image pyramids. Inspired by the Gaussian scale-space theory [43], Hessian matrix-based extractor is used by enlarging the size of the box filter without compromising on the size of the original image. In this way,

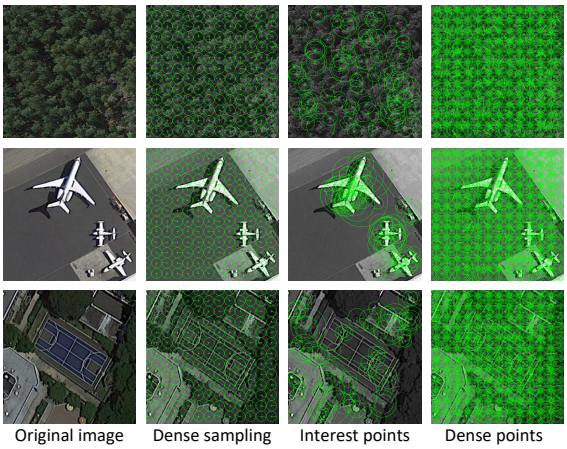


Figure 6: Scene recognition with dense sampling, sparse interest points and the proposed dense interest points.

multi-scale information could be achieved based on a second derivative Gaussian filter and a convolution operation as follows in (Eq. (2) and (3)):

$$H(X,\mu) = \left[\frac{L_{xx}(X,\mu), L_{xy}(X,\mu)}{L_{xy}(X,\mu), L_{yy}(X,\mu)}\right] \tag{2}$$

$$L_{xx}(X,\mu) = I(x,y) \times \frac{\partial^2}{\partial x^2} g(\mu) \tag{3}$$

where $L_{xx}(X,\mu)$ represents a second-order differentiated Gaussian filter along the xx direction while $L_{xy}(X,\mu)$ and $L_{yy}(X,\mu)$ denotes second order differentiated Gaussian filters and convolution operations in xy direction (diagonal) and yy direction (vertical), respectively [44]. Since the the Gaussian filter has a drawback due to a large amount of computation, this issue is addressed by using box filters [45] which have been particularly employed for fast implementation such as:

$$det(H_{app}) = D_{xx}D_{yy} - (\lambda D_{xy})^2 \tag{4}$$

where λ required to balance the Hessian determinant and is acquired using the Frobenius Norm. In this way, computation amount can be significantly decreased as:

$$\lambda = \frac{\left|L_{xy}(1,2)\right|_F \left|D_{yy}(9)\right|_F}{\left|L_{yy}(1,2)\right|_F \left|D_{xy}(9)\right|_F} = 0.912 \cong 0.9 \tag{5}$$

Hence, the proposed idea takes the advantage of a hybrid feature extraction scheme, i.e. multi-scale interest points and dense sampling, where we start from dense sampling on regular grids with the repeatability of interest points at multiple scales. Figure 6 displays the dense sampling, sparse interest points, and hybrid (dense interest points) scheme. Once the scale space has been built, we utilize SURF descriptor [45] to extract the features within a bounded search area. For an image I, image scales $m_i = (i = 1, 2, ..., n)$ are denoted as

$x_{mi} = (i = 1, 2, ..., n)$. Formally, for each smoothed image the feature extracted from the SURF is illustrated as follows:

$$f_{m_i} = SURF(x_{mi}), \qquad i = 1, 2, \dots, n \tag{6}$$

where n is the number of scales, i is the index of scale, x_{mi} is the i^{th} scale, x_{mi} is the region at i^{th} scale, and f_{mi} is the SURF feature for x_{mi}.

In order to construct the visual vocabulary, SURF features are clustered through the k-means clustering process and mapped to a specific codeword, thus, can be represented by a histogram of visual words. The histogram becomes a final representation of the image.

3.3. Information fusion

Information fusion is the process of combining multiple pieces of information to provide more consistent, accurate, and useful information than a single piece of information. In general, it is divided into four categories: decision level, scale level, feature level, and pixel-level [46]. Among them, feature-level fusion has comparatively a shorter history but is an emerging topic in a remote-sensing domain. The spatial relation between the proposed regions can improve scene classification in two aspects. First, aggregating the information of a neighborhood and its adjacent neighborhoods assists in recognizing the features that accurately represent the scene type of the image. For instance, determining whether farmland belongs to a forest field or a meadow requires information about its neighboring area. Second, the natural relationship of the spatial distribution pattern of a scene helps to infer the scene category. An industrial area, for instance, is likely planar, and the runway is always linear. Therefore, we select to combine four different regions based on multiscale features, with the aim to obtain more informative and relevant features to represent the input image. Each input image I produced four sets of a histogram of visual words, which are generated by different pixel strides through the k-means clustering process as previously mentioned and denoted as Q_1, Q_2, Q_3, and Q_4. Specifically, the first set of histogram of visual words extracted from the image is $Q_1 = (q_{e_1}, q_{e_2}, \dots, q_{e_n}) \in R^z$; R^z represents z-dimensional vector. The second set is represented as $Q_2 = (q_{w_1}, q_{w_2}, \dots, q_{w_n}) \in R^w$; R^w represents w-dimensional vector. Q_1 and Q_2 are the outputs of two different patch sizes. Similarly, the third and fourth sets are represented as $Q_3 = (q_{y_1}, q_{y_2}, \dots, q_{y_n}) \in R^y$; R^y represents y-dimensional vector, and $Q_4 = (q_{u_1}, q_{u_2}, \dots, q_{u_n}) \in R^u$; R^u represents an u-dimensional vector, respectively. Information fusion is performed by the concatenation of Q_1,Q_2,Q_3 and Q_4, and result is denoted by Q_f that is an $(z+w+y+u)$-dimensional vector. Thus, fusion is achieved by the following formula:

$$Q_f = Q_1 \oplus Q_2 \oplus Q_3 \oplus Q_4, Q_f \in R^{z+w+y+u}, \tag{7}$$

where the elements $(q_{e_1}, q_{e_2}, \dots, q_{e_n})$ of Q_1, the elements $(q_{w_1}, q_{w_2}, \dots, q_{w_n})$ of Q_2, the elements $(q_{y_1}, q_{y_2}, \dots, q_{y_n})$ of Q_3, and the elements of $(q_{u_1}, q_{u_2}, \dots, q_{u_n})$ of Q_4 construct a new vector Q_f to express the fused feature vector.

Table 1
Neighborhood-based analysis on each dataset.

	Different Neighborhood combinations	Accuracy(%)
UC Merced dataset		
1	4×4	88.10
2	6×6	86.79
3	8×8	85.43
4	10×10	84.52
WHU-RS dataset		
1	4×4	86.10
2	6×6	88.70
3	8×8	91.52
4	10×10	89.52
NWPU-RESISC45 dataset		
1	4×4	67.10
2	6×6	65.61
3	8×8	64.52
4	10×10	62.52
AID dataset		
1	4×4	75.60
2	6×6	77.13
3	8×8	78.52
4	10×10	76.90

3.4. Recurrent Neural Network (RNN)

The Earth observation satellites normally capture consecutive images of the same ground by visiting the same area every few days. Thus, the time elapsed between consecutive images complement the temporal resolution (i.e., the time when it was acquired) [47]. Our motivation for using bidirectional long short-term memory (BiLSTM) [48] is to take advantage of the temporal pattern of the scenes across image time series. BiLSTM determines the input sequence $i = i_1, i_2, ..., i_n$ from the opposite order to a forward hidden sequence $\overrightarrow{\mathrm{f_t}} = (\overrightarrow{\mathrm{f_1}}, \overrightarrow{\mathrm{f_2}}, ..., \overrightarrow{\mathrm{f_n}})$ and a backward hidden sequence $\overleftarrow{\mathrm{f_t}} = (\overleftarrow{\mathrm{f_1}}, \overleftarrow{\mathrm{f_2}}, ..., \overleftarrow{\mathrm{f_n}})$. The encoded vector v_t is computed by the accumulation of the final forward and backward outputs $v_t = [\overrightarrow{\mathrm{f_t}}, \overleftarrow{\mathrm{f_t}}]$.

$$\overrightarrow{\mathrm{f_t}} = \delta(W_{\overrightarrow{\mathrm{f}}i} i_t + W_{\overrightarrow{\mathrm{f}}\overrightarrow{\mathrm{f}}} \overrightarrow{\mathrm{f}}_{t-1} + q_{\overrightarrow{\mathrm{f}}}), \tag{8}$$

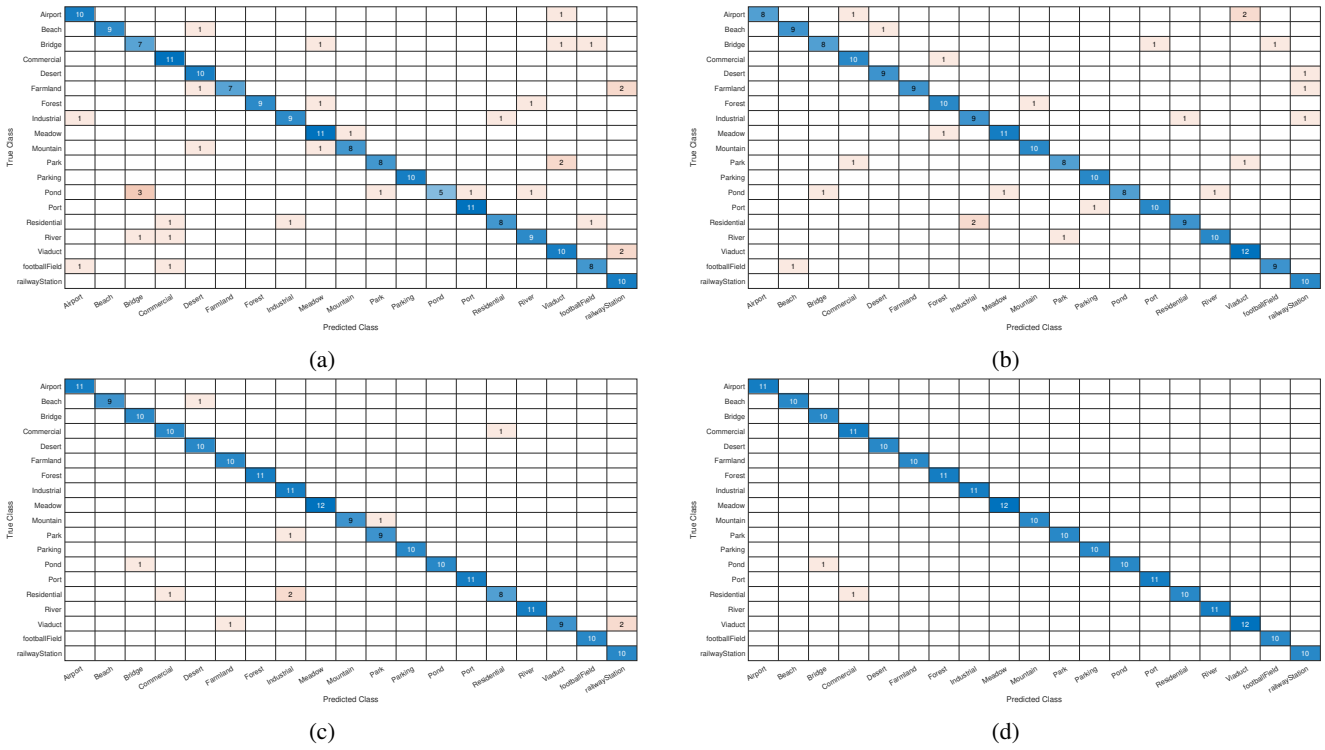
$$\overleftarrow{\mathrm{f_t}} = \delta(W_{\overleftarrow{\mathrm{f}}i} i_t + W_{\overleftarrow{\mathrm{f}}\overleftarrow{\mathrm{f}}} \overleftarrow{\mathrm{f}}_{t+1} + q_{\overleftarrow{\mathrm{f}}}), \tag{9}$$

$$v_t = W_{v\overrightarrow{\mathrm{f}}} \overrightarrow{\mathrm{f_t}} + W_{v\overleftarrow{\mathrm{f}}} \overleftarrow{\mathrm{f_t}} + q_v \tag{10}$$

where δ is the logistic sigmoid function and $v = (v_1, v_2, ...v_t..., v_n)$ is the output sequence of the first hidden layer.

4. Datasets And Experimental Setup

In this section, we first provide a brief description of four databases that are used to evaluate our method. Then, the implementation details and ablation analysis are discussed and the results are compared with state-of-the-art methods.



(a) (b) (c) (d)

Figure 7: Confusion matrix of our proposed method on WHU-RS dataset by fixing the training ratio as 80% (a) with one-stage learning, (b) with two-stage learning, (c) with three-stage learning, and (d) with multi-stage learning. Zoom in for a better view.

4.1. Datasets

UC Merced Land Use Dataset (UC-Merced): This dataset was obtained from the USGS National Map Urban Area with a pixel resolution of one-foot [30]. It contains 21 distinctive scene categories and each class consists of 100 images of size $256 \times 256 \times 3$. Inter-class similarity, for example, highway and architecture scenes can be easily mixed with other scenes, such as freeways and buildings, which makes this dataset a challenging one.

WHU-RS Dataset: It was collected from satellite images of Google Earth [49]. This dataset consists of 950 scene images and 19 classes with a size of 600×600. Each image varies greatly in high resolution, scale, and orientation, which makes it more complicated than the UCM dataset.

Aerial Image Dataset (AID): There are 10000 images in AID dataset, which are categorized into 30 scene classes [2]. Each class contains images ranging from 220 up to 420 with the fixed size of 600×600 pixels in the RGB space. The pixel resolution changes from about 8 m to about half a meter.

NWPU-RESISC45 Dataset: It consists of 31,500 remote sensing images divided into 45 scene classes, covering more than 100 countries and regions all over the world [50]. Each class contains 700 images with the size of 256×256 pixels. This dataset is acquired from Google Earth (Google Inc.), where the spatial resolution varies from 30 to 0.2 m per pixel. This is one of the largest datasets of remote sensing images and is 15 times larger than the most widely-used UC Merced dataset. Hence, the rich image variations, high inter-class similarity, and the large scale make the dataset even more challenging.

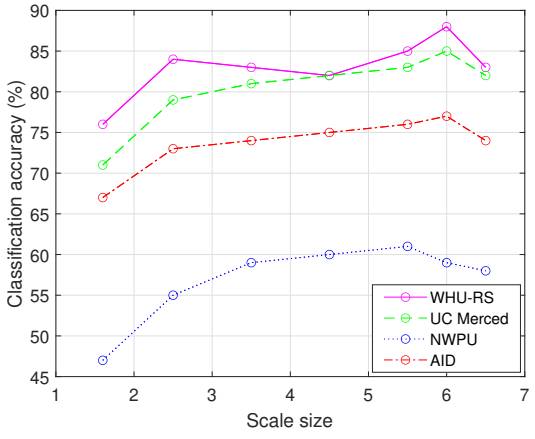


Figure 8: Classification accuracy of the proposed method under different Gaussian scales for four datasets.

4.2. Implementation details

To evaluate the performance on the above-mentioned datasets, the BoW is used as the base architecture with four distinct image regions and seven adjacent Gaussian scaled images, i.e., $[1.6, 2.5, 3.5, 4.5, 5.5, 6.0, 6.4]$. The vocabulary size of k in the remote-sensing domain varies from a few hundred to thousands. We set the size of visual vocabulary to 15000 for UC Merced, AID, NWPU, and 10000 for the

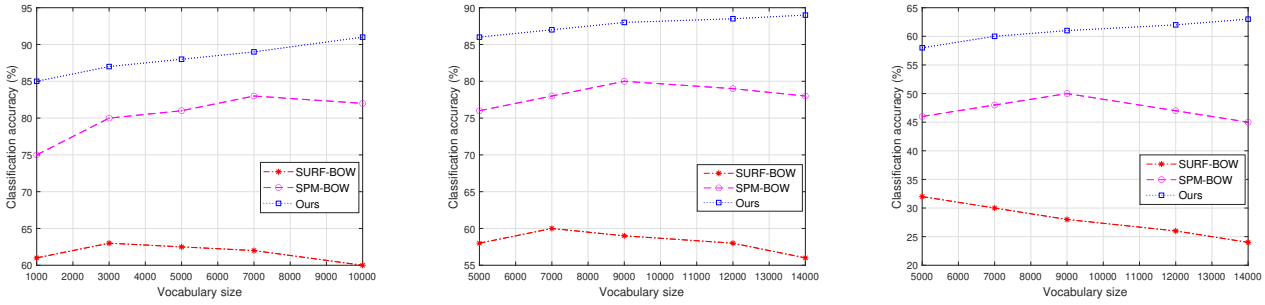


Figure 9: Comparison of classification on three datasets. (a) Comparing the performance on WHU-RS dataset with SURF-BOW [45], SPM-BOW [8], and ours. (b) Comparing the performance on UC Merced dataset using SURF-BOW [45], SPM-BOW [8], and ours. (c) Comparing the performance on NWPU dataset using SURF-BOW [45], SPM-BOW [8], and ours.

Figure 10: Right: the proposed sample windows show negligible redundancy. Center: sample windows with a pixel stride 2 display overlapping. Left: sample windows shows high redundancy.

Table 2
The general comparison of the proposed method after information fusion in terms of accuracy (%), training and testing time per second.

Dataset	Neighborhood-Based Fusion	Accuracy	Training (s)	Testing (s)
UC-Merced	BoW(1+2)+BiLSTM	94.10	8671.74	253.16
	BoW(1+2+3)+BiLSTM	97.76	15007.61	480.74
	BoW(1+2+3+4)+BiLSTM	99.57	19343.48	601.32
WHU-RS	BoW(1+2)+BiLSTM	95.49	10452.38	676.3
	BoW(1+2+3)+BiLSTM	98.20	17678.57	1104.5
	BoW(1+2+3+4)+BiLSTM	99.63	22904.76	1452.6
NWPU	BoW(1+2)+BiLSTM	89.32	21271.08	7798.06
	BoW(1+2+3)+BiLSTM	94.72	32906.62	12695.59
	BoW(1+2+3+4)+BiLSTM	97.13	44542.16	16594.12
Aerial Image	BoW(1+2)+BiLSTM	92.77	40035.02	9131.96
	BoW(1+2+3)+BiLSTM	96.51	62152.53	14697.94
	BoW(1+2+3+4)+BiLSTM	98.43	82170.04	19263.92

WHU-RS dataset. The BiLSTM is trained using the Adam optimizer with a gradient threshold 1, while the minibatch size of 32 with a hidden layer dimension of 80. Initializing the BiLSTM with the right weights is a challenging task because standard gradient descent from random initialization can hamper the training of BiLSTM. Therefore, we set the recurrent weights with Glorot initializer (Xavier uniform) [51] which performs the best in all scenarios of our experiments. To decrease the computation complexity on AID and NWPU datasets, we only use four Gaussian scaled images where the highest filter image takes a weight of 4.5, and the lowest 1.6.

Table 3
Comparison of classification accuracy (%) with feature-level fusion methods under different training sizes on the UC Merced dataset.

Train data	DCA[20]	PCA [21]	CCA[22]	Ours
30%	93	93	91	92
40%	94	93	92	93
50%	95	97	94	95
80%	96	98	98	99

Table 4
Comparison of classification accuracy (%) based on different pixel strides with NWPU, AID, and UC Merced datasets.

Dataset	PS1	PS2	PS
NWPU	94.21	95.33	97.13
WHU-RS	95.54	97.94	99.63
UC Merced	98.44	98.91	99.57

4.3. Ablation study

We thoroughly validate the performance of each neighborhood size by performing an ablation study. In Table 1, we have reported the results of estimating PBDL on UC Merced, WHU-RS, NWPU, and AID datasets. Our one-stage detection method on the WHU-RS dataset with the neighborhood size of (4×4) achieves 86.10% accuracy and the numerical results of each category are shown in Fig.7 (a). The diagonal elements represent the number of images for which the classifier predicted correctly. It can be seen that several classes such as bridge (three images), pond (six images), farmland (three images), residential (three images), and viaduct (two images) are misclassified. In Fig.7 (b), we show that when the neighborhood size (10×10) increases, the overall classification is improved from 86% to 89%, which is 3% higher than the (4×4) size. After combining both kinds of features, we notice that images of the bridge, pond, farmland, and residential are predicted correctly up to 99% and achieve an overall classification accuracy of 95% as shown in Fig.7 (c). The final results are obtained by

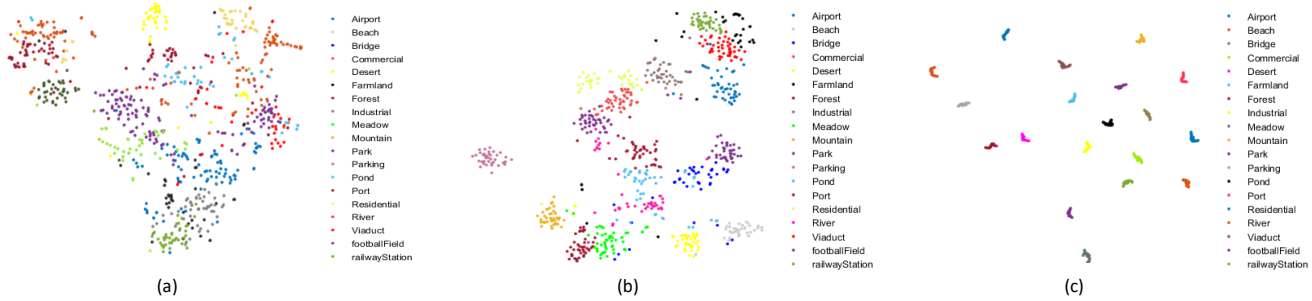


Figure 11: Two-dimensional scatterplots of SURF-based BoW features genertaed with t-SNE over the WHU-RS dataset. (a) Scatterplot of one-stage multi-scale features. (b) Scatterplot of features extracted and combined from four-stage learning. (c) Features extracted from the last fully-connected layer of BiLSTM. All points in the scatterplots are class coded.

combining all the neighborhood features and are displayed in Fig.7 (d). A significant improvement can be observed in overall classification performance and only two images are misclassified in the WHU dataset. Based on these results, we conclude that a single BoW model cannot provide state-of-the-art results without aggregating the features of discriminative regions. From the Table 2 findings, it is evident that the BoW(1+2)+BiLSTM obtains good performance on the UC Merced dataset right from the start. When we integrate the features of different neighborhood(1+2+3) sizes, the model further improves the performance up to 9% than the single grid-sized BoW model. By combining all the neighborhood features, we achieved the best performance i.e., 99%. Similarly, for NWPU and Aerial Image datasets, a significant difference can be seen even with combining two neighborhood(1+2) sizes, and the performance is boosted with increasing number of neighborhood(1+2+3) sizes, surpassing 90% with just 10% of all samples as a training sample. This is a remarkable improvement compared with the previous methods. In addition, UC-Merced, WHU-RS, NWPU, and Aerial Image take 19343.48 s, 22904.76 s, 44542.16 s, and 82170.04 s for training, and 601.32 s, 1452.6 s, 1452.6 s, and 19263.92 s for testing, respectively. Thus, results demonstrate that different neighborhood sizes play different roles in classifying remote sensing scene images, and the proposed patch-based discriminative learning plays an essential role in significantly improving the feature representation for remote sensing scene classification.

4.3.1. Scale Factor of Gaussian Kernel

Figure 8 shows the classification performance of each scaled image based on 10×10 neighborhood size. The PBDL extracts multi-scale dense features according to the scale factor to control the Gaussian kernel. It can be observed that with the increase of scale σ factor, the performance first improves and then gradually decreases after the $\sigma = 6.0$ scaled image. We conclude that including a certain range of Gaussian smoothed images can improve the performance, but too many of them degrade the performance.

4.3.2. Codebook learning

We quantitatively analyze the performance with the SURF descriptor and standard SPM method in the bag-of-words framework. An engaging question is how much the performance can be improved by defining the proposed spatial locations with multi-scale information. With this in mind, we set different vocabulary sizes for WHU, UC Merced, and NWPU datasets. The respective outcomes can be found in Fig.9 (a) (b) and (c). One can see that even the proposed one-stage detection method with the neighborhood size of (4×4) significantly outperforms the SPM method. Similarly, using the SURF descriptor in the BoW framework cannot achieve the best performance and provides more than 20% lower accuracy than ours on all databases.

4.3.3. Quantitative comparison of different fusion methods

Table 3 provides the quantitative analysis based on the different sizes of the training data. All the compared methods such as [20, 21, 22] perform feature-level fusion based on DCA, PCA, or CCA to improve the scene classification performance. For instance, the authors in [20] fuse the deep neural network features based on discriminant correlation analysis (DCA). To make the deep learning features more discriminant, features of different models are combined based on PCA [21]. The global features under the BoW framework are fused based on canonical correlation analysis (CCA) [22]. In comparison with these state-of-the-art fusion methods, our proposed fusion performs best with an accuracy of 99%.

4.3.4. Performance comparison of different pixel strides

During patch-based learning, we consider the problem of sampling redundancy. Although different number of image patch sizes have been used in a traditional dense feature sampling approach, the optimal pixel strides are not deeply investigated in the literature. We show that same pixel stride (4) corresponding to the pixel width of the feature window (4×4), is better suited to the domain of remote sensing scene image classification. In this way, it allows the classifiers to

Table 5
Classification accuracy (%) for the NWPU dataset with two training ratios.The results are obtained directly from the corresponding papers.

Method	10%	20%
BoW with dense SIFT [53]	41.72±0.21	44.97±0.28
BOCF [53]	82.65±0.31	84.32±0.17
BoVW+SPM [50]	27.83±0.61	32.96±0.47
D-CNN [54]	89.22±0.50	91.89±0.22
Triple networks [55]	-	92.33±0.20
MDFR [56]	83.37±0.26	86.89±0.17
APDC-Net [57]	85.94±0.22	87.84±0.26
BoWK [22]	-	66.87±0.90
SFCNN [58]	89.89±0.16	92.55±0.14
Attention GANs [59]	86.11±0.22	89.44±0.18
MDFR [56]	83.37±0.26	86.89±0.17
CNN + GCN [15]	90.75±0.21	92.87±0.13
Color fusion [60]	-	87.50±0.00
Graph CNN [61]	91.39±0.19	93.62±0.28
AlexNet+SAFF [62]	80.05±0.29	84.00±0.17
VGG-VD16+SAFF [62]	84.38±0.19	87.86±0.14
IDCCP [63]	91.55±0.16	93.76±0.12
SEMSDNet [64]	91.68±0.39	93.89±0.63
PBDL+SVM (ours)	91.11±0.77	93.33±1.13
PBDL (The proposed)	**94.20±0.81**	**97.13±0.92**

consider more scales with minimal increase in overlapping or redundancy. Table 4 shows the impact of this effective parameter tuning. The PS represent same pixel stride corresponding to the proposed patch sizes while pixel stride 1 and 2 are used for comparison purpose and expressed as PS1 and PS2, respectively. One can see that this basic modification provides improved results on all datasets and minimize overlapping in (x,y) space. In addition, Figure 10 visualizes the point of redundancy.

4.3.5. Visualization of Feature Structures

One of the advantages of the proposed approach is that we can interpret the classification process of the model. Especially for each stage, we can see how the features are structured into data space and their impact along the different classification stages. Taking this into consideration, we employed the "t-distributed stochastic neighboring embedding" (t-SNE) algorithm [52] and illustrated the derived embeddings into three separated processing stages: 1) one-stage learning, 2) combined learning (PBDL), and 3) BiLSTM classified features for the WHU dataset. The features with the patch size of 4×4 in Fig.11 (a) show that most classes are strongly correlated, which makes the classifier (BiLSTM) hard to separate them. We also visualize the clusters by fusing all the neighborhood features in Fig.11 (b). The derived clusters indicate that the proposed fusion reduces the correlation between similar classes and can capture more variability in the feature space. Moreover, it could be noticed from Fig.11 (c) that all the classes are well separable which could potentially lead to better performance when training BiLSTM on remote sensing dataset.

Table 6
Classification accuracy (%) for the AID dataset with two training ratios.The results are obtained directly from the corresponding papers.

Method	20%	50%
Fusion by addition [20]	-	91.87±0.36
D-CNN [54]	90.82±0.16	96.89±0.10
MDFR [56]	90.62±0.27	93.37±0.29
APDC-Net [57]	88.56±0.29	92.15±0.29
SFCNN [58]	94.93±0.31	96.89±0.10
Attention GANs [59]	93.97±0.23	96.03±0.16
CNN + GCN [15]	94.93±0.31	96.89±0.10
Color fusion [60]	-	94.00±0.00
AlexNet+SAFF [62]	87.51±0.36	91.83±0.27
VGG-VD16+SAFF [62]	90.25±0.29	93.83±0.28
Graph CNN [61]	93.06±0.26	95.78±0.37
IDCCP [63]	94.80±0.18	96.95±0.13
SEMSDNet [64]	94.23±0.63	97.64±0.51
PBDL+SVM (ours)	91.83±0.23	94.31±0.59
PBDL (The proposed)	**96.11±0.81**	**98.43±0.33**

Table 7
Comparison of classification accuracy (%) for the UC-Merced dataset with 80% ratios. The results are obtained directly from the corresponding papers.

Method	Accuracy (Mean±std)
AlexNet+sum pooling [65]	94.10±0.93
VGG-VD16+sum pooling [65]	91.67±1.40
SPP-Net [66]	96.67±0.94
GoogleNet [2]	94.31±0.89
VGG-VD16 [2]	95.21±1.20
DCA fusion [20]	96.90±0.77
MCNN [67]	96.66±0.90
D-CNN [54]	98.93±0.10
Triple networks [55]	97.99±0.53
VGG-VD16 +AlexNet [21]	98.81±0.38
Fusion by concatenation [68]	98.10±0.20
MDFR [56]	98.02±0.51
APDC-Net [57]	97.05±0.43
BoWK [22]	97.52±0.80
Attention GANs [59]	97.69±0.69
AlexNet+SAFF [62]	96.13±0.97
VGG-VD16+SAFF [62]	97.02±0.78
Color fusion [60]	98.10±0.00
Graph CNN [61]	99.00±0.43
IDCCP [63]	99.05±0.20
SEMSDNet [64]	99.41±0.14
PBDL+SVM (ours)	98.11±0.54
PBDL (The proposed)	**99.57±0.36**

4.4. Performance comparison with state-of-the-art methods

4.4.1. NWPU-RESISC45 Dataset

To demonstrate the superiority of the proposed method, we evaluate the performance against several state-of-the-art classification methods on the NWPU dataset as shown in Table 5. Especially, we choose mainstream BoW and deep learning-based methods and compare the performance

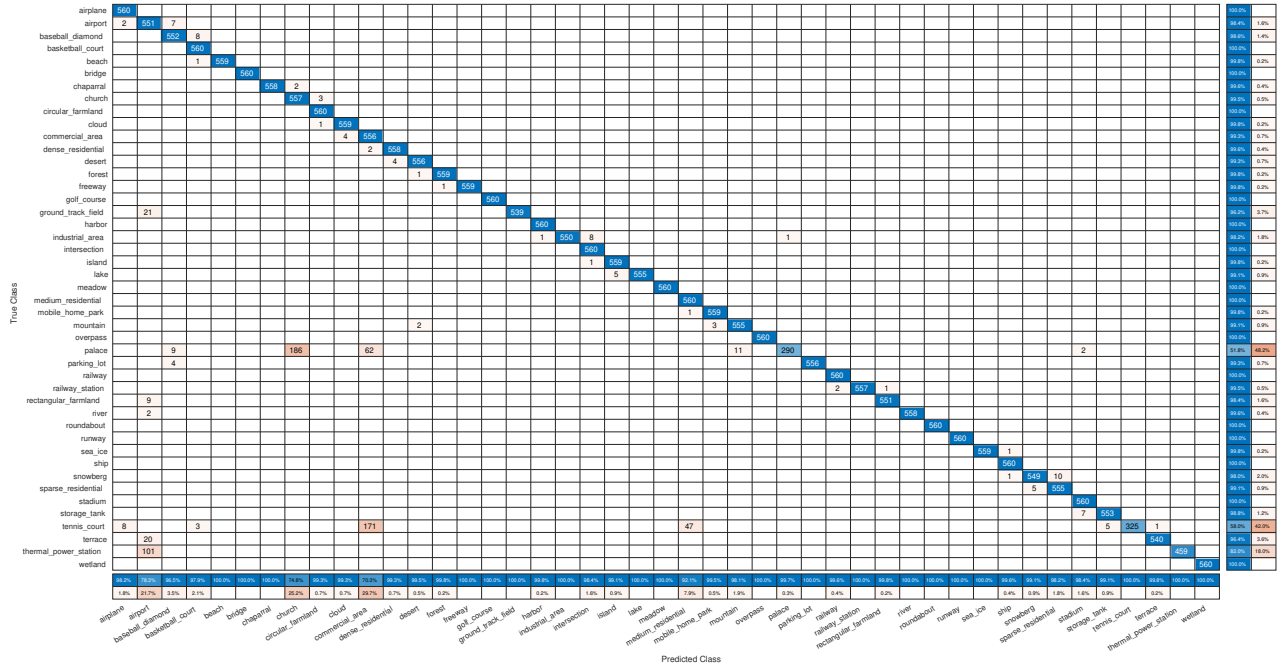


Figure 12: Confusion matrix of our proposed method on NWPU-RESISC45 Dataset by fixing the training ratio as 20%. Zoom in for a better view.

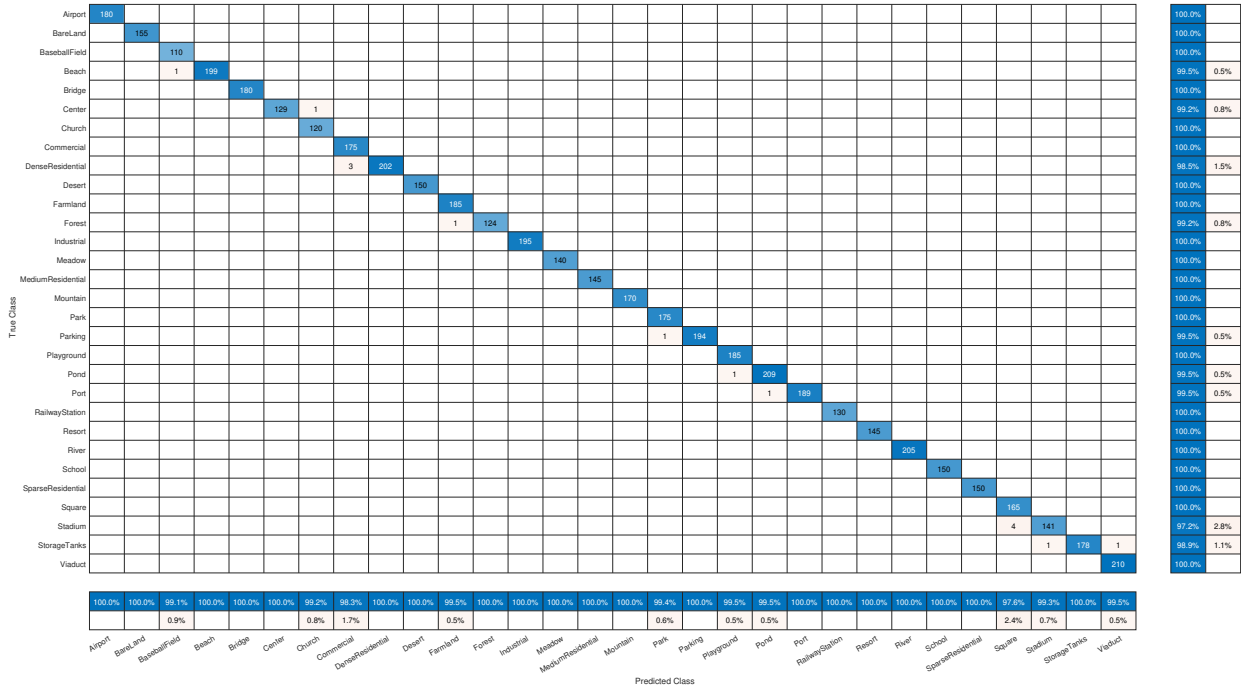


Figure 13: Confusion matrix of our proposed method on AID Dataset by fixing the training ratio as 50%. Zoom in for a better view.

of scene classification. It could be observed from Table 5, that the proposed approach, by combining all neighborhood-based features, achieved the highest overall performance of 94.20% and 97.13% using 10% and 20% training ratios, respectively. It is worth mentioning that NWPU is much more difficult than the other three datasets and our proposed method outperforms the previous state-of-the-art method by a margin of 4% under the training ratio of 20%. Even, the support vector machine (SVM) classifier provides competitive performance on both training ratios. Thus, the classification performance of the proposed PBDL shows the effectiveness of combining global-based visual features on

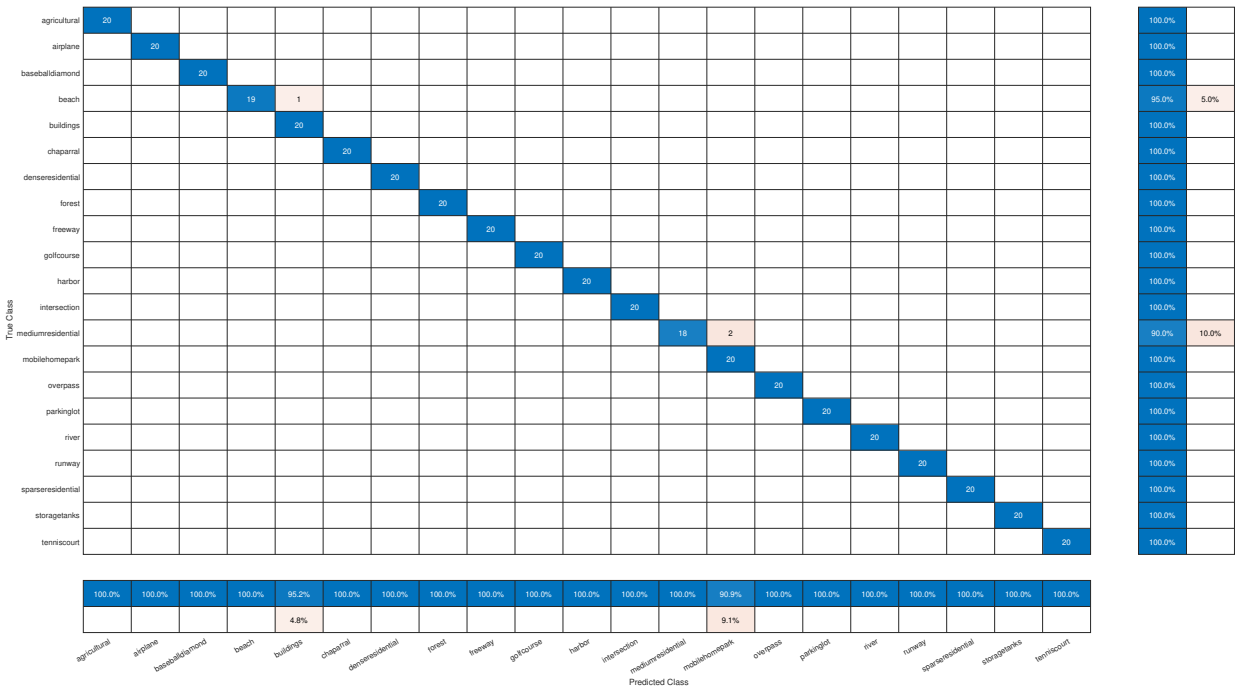


Figure 14: Confusion matrix of our proposed method on UC Merced Dataset by fixing the training ratio as 80%. Zoom in for a better view.

Table 8
Comparison of classification accuracy (%) for the WHU-RS19 with 80% ratios. The results are obtained directly from the corresponding papers.

Method	Accuracy (Mean±std)
Transferring CNNs (Case I) [69]	96.70±0.00
Transferring CNNs (Case II) [69]	98.60±0.00
Two-Step Categorisation [70]	93.70±0.57
CaffeNet [2]	94.80±0.00
GoogleNet [2]	92.90±0.00
VGG-VD16 [2]	95.10±0.00
MDDC [71]	98.27±0.53
salM^3LBP-CLM [72]	96.38±0.76
AlexNet-SPP-SS [66]	95.00±1.12
VGG-VD19 [21]	98.16±0.77
DCA by addition [20]	98.70±0.22
MLF [73]	88.16±2.76
Fusion by concatenation [68]	99.17±0.20
D-DSML-CaffeNet [74]	96.64±0.68
BoWK [22]	99.47±0.60
Color fusion [60]	96.60±0.00
PBDL+SVM (ours)	99.10±0.41
PBDL (The proposed)	**99.63±0.42**

the NWPU dataset.

Fig.12 illustrates the confusion matrix produced by our proposed method (PBDL) with the 20% training ratio. Each row represents the percentages of correctly and incorrectly classified observations for each true class. Similarly, each column displays the percentages of correctly and incorrectly classified observations for each predicted class. One can see that the classification performance of 41 categories is greater than 95% where only the 14 categories have achieved more than 95% in the previous methods [15]. However, one common challenge is found that the church and palace are two confusing categories, which limits many existing works to surpass the performance [15]. In our case, 25% of images from church are mistakenly classified as a palace which is 1% high misclassification than the CNN + GCN [15]. On the other side, only 0.3% of images from the palace are mistakenly classified as an industrial area where the previous methods achieve 67% [58] and 70% [15] performance for the palace class. By analyzing the confusion matrix on PBDL, the airport, church, and commercial area are the only challenging classes for our proposed method. Thus, the experimental results demonstrate the proposed method improves the discriminative ability of features and works well on the large-scale NWPU-RESISC45 dataset.

4.4.2. AID Dataset

We evaluate and report the comparison results against the existing state-of-the-art classification methods for the AID dataset in Table 6. It could be observed that PBDL achieved the overall accuracy of 96.11% and 98.43% using 20% and 50% training ratios, respectively. As can be seen from Table 6, our method outperformed the SEMSDNet [64] with increases in the overall performance of 1.88% and 0.79% under both training ratios. Thus, our proposed method, by combining all the neighborhood features, verifies the effectiveness of multi-level and multi-scale feature fusion.

Fig.13 represents the confusion matrix generated by PBDL with the 50% training ratio. As can be seen from

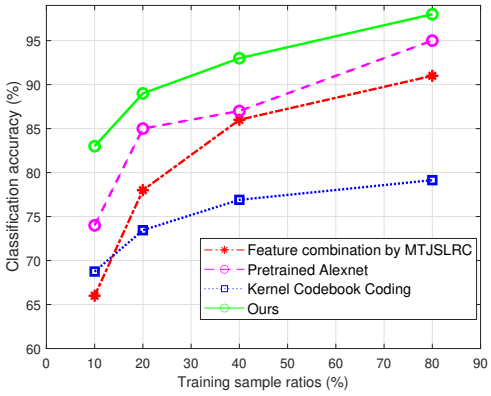


Figure 15: The influence of the training sample ratios with different methods such as feature combination by MTJSLRC [75], Pretrained Alexnet [66], and kernel codebook coding [76].

Fig. 13, the classification performance of all the categories is higher than 95% and only the square category provides the lowest accuracy up to 97%. Specifically, 4 of images from the square are mistakenly classified as stadium, and 3 of images from commercial are misclassified as dense residential. The five categories such as school, square, park, center, and resort are very confusing categories, which leads many existing works to be unable to get a competitive performance [64]. For instance, SFCNN [58] and the CNN + GCN [15] attain 70% to 91% accuracy for the class of resort while our method achieves 100% accuracy. It confirms that despite the high interclass similarity, the proposed method is capable of extracting robust spatial location information to distinguish these remote sensing scene categories.

4.4.3. UC Merced Dataset

The evaluation results on the UC Merced dataset are presented in Table 7 by using the 80% training ratio. The proposed method achieves 99.57% accuracy and competes with the previous BoW [22] approach by a margin of 2.05%. For further evaluation, a confusion matrix of the UC Merced dataset is shown in Fig.14. A total of 3 images are misclassified in this dataset where buildings and mobile home parks are found to be challenging categories for our proposed method. Moreover, the effect of the number of training samples on the UC Merced dataset is also examined by selecting 20%, 30%, 40%, and 80% as training samples and visualized in Fig. 15. It can be noticed that in comparison with other fusion methods, the proposed fusion method is superior from the start even with a 10% training sample ratio. Thus, the proposed method is effective to classify most of the scene categories.

4.4.4. WHU-RS Dataset

Table 8 reports the comparison results of the WHU-RS dataset. As shown in Table 8, the PBDL achieves the highest classification (99.63%) accuracy and outperforms all the previous methods for the 19 classes. In addition, a confusion matrix of the WHU-RS dataset is shown in Fig.7 (D). Tremendous improvements can be observed in some classes such as residential, industrial, port, pond, park, mountain, airport, and railway station. Only 2 images from commercial and bridge categories are misclassified in this dataset. Hence, based on experimental analysis, we argue that a combination of neighborhood sizes and multi-scale filtering is essential to produce robust feature representation for remote sensing scene classification.

5. Conclusion

This paper introduced a simple, yet very effective approach called patch-based discriminative learning (PBDL) for extracting discriminative patch features. The PBCL generates N patches for each image feature map and the individual patches of the same image are located at different spatial regions to achieve a more accurate representation. In particular, these regions focus on "where" is the discriminative information, whereas aggregation (fusion) of the neighborhood regions focuses on "what" is the scene semantic associated with, given an input image and taking into account the complementary aspect. We show that patch-based learning in the BoW model significantly improves the recognition performance compared to that obtained when using a single-level BoW alone. Experiments were conducted on four publicly available datasets, and the results demonstrate that the different distribution of spatial location and visual information is crucial for scene classification. The proposed approach is expected to have advantages over single-scale BoW or traditional CNNs methods, especially in the situation where a large number of training data is not available and classification accuracy is the prime goal. A drawback of PBDL is that it increases the computational complexity of the BoW model. Therefore, we plan to extend our work by developing computationally efficient methods to automatically obtain multi-level and multi-scale features without human intervention.

References

[1] R. Bishop-Taylor, R. Nanson, S. Sagar, and L. Lymburner, "Mapping australia's dynamic coastline at mean sea level using three decades of landsat imagery," *Remote Sensing of Environment*, vol. 267, p. 112734, 2021.

[2] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[3] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[4] D.-C. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509–512, 1990.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] M. Swain and D. Ballard, "Color indexing international journal of computer vision 7," 1991.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, Ieee, 2005.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178, IEEE, 2006.

[9] C. Shi, X. Zhang, J. Sun, and L. Wang, "Remote sensing scene image classification based on dense fusion of multi-level features," *Remote Sensing*, vol. 13, no. 21, p. 4379, 2021.

[10] B. Yuan, L. Han, X. Gu, and H. Yan, "Multi-deep features fusion for high-resolution remote sensing image scene classification," *Neural Computing and Applications*, vol. 33, no. 6, pp. 2047–2063, 2021.

[11] H. Gao, Z. Chen, and F. Xu, "Adaptive spectral-spatial feature fusion network for hyperspectral image classification using limited training samples," *International Journal of Applied Earth Observation and Geoinformation*, vol. 107, p. 102687, 2022.

[12] J. Hu, G.-S. Xia, F. Hu, H. Sun, and L. Zhang, "A comparative study of sampling analysis in scene classification of high-resolution remote sensing imagery," in *2015 IEEE International geoscience and remote sensing symposium (IGARSS)*, pp. 2389–2392, IEEE, 2015.

[13] U. Muhammad, W. Wang, S. P. Chattha, and S. Ali, "Pre-trained vggnet architecture for remote-sensing image scene classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1622–1627, IEEE, 2018.

[14] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4911–4926, 2020.

[15] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined cnn and gcn for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4325–4338, 2020.

[16] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 604–610, IEEE, 2005.

[17] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European conference on computer vision*, pp. 490–503, Springer, 2006.

[18] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[19] T. Tuytelaars, "Dense interest points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2281–2288, IEEE, 2010.

[20] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for vhr remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4775–4784, 2017.

[21] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017.

[22] U. Muhammad, W. Wang, A. Hadid, and S. Pervez, "Bag of words kaze (bowk) with two-step classification for high-resolution remote sensing images," *IET Computer Vision*, vol. 13, no. 4, pp. 395–403, 2019.

[23] T. Blaschke and J. Strobl, "What's wrong with pixels? some recent developments interfacing remote sensing and gis," *Zeitschrift für Geoinformationssysteme*, pp. 12–17, 2001.

[24] L. Chen, W. Yang, K. Xu, and T. Xu, "Evaluation of local features for scene classification using vhr satellite images," in *2011 Joint Urban Remote Sensing Event*, pp. 385–388, IEEE, 2011.

[25] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 280–295, 2014.

[26] H. Yu, W. Yang, G.-S. Xia, and G. Liu, "A color-texture-structure descriptor for high-resolution satellite image classification," *Remote Sensing*, vol. 8, no. 3, p. 259, 2016.

[27] F. Zhang and X. Yang, "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection," *Remote Sensing of Environment*, vol. 251, p. 112105, 2020.

[28] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818–832, 2012.

[29] F. S. Khan, J. Van de Weijer, and M. Vanrell, "Modulating shape features by color attention for object recognition," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 49–64, 2012.

[30] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

[31] D. A. R. Vigo, F. S. Khan, J. Van De Weijer, and T. Gevers, "The impact of color on bag-of-words based object recognition," in *2010 20th international conference on pattern recognition*, pp. 1549–1553, IEEE, 2010.

[32] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–1596, 2009.

[33] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *2009 IEEE Conference on computer vision and pattern recognition*, pp. 1794–1801, IEEE, 2009.

[34] H. Ghanbari, M. Mahdianpari, S. Homayouni, and F. Mohammadimanesh, "A meta-analysis of convolutional neural networks for remote sensing applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3602–3613, 2021.

[35] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 18–26, 2015.

[36] R. Huang, F. Zheng, and W. Huang, "Multi-label remote sensing image annotation with multi-scale attention and label correlation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[37] M. Li, L. Lei, X. Li, Y. Sun, and G. Kuang, "An adaptive multilayer feature fusion strategy for remote sensing scene classification," *Remote Sensing Letters*, vol. 12, no. 6, pp. 563–572, 2021.

[38] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "Dla-matchnet for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[39] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[40] J. F. Rosier, H. Taubenböck, P. H. Verburg, and J. van Vliet, "Fusing earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification," *Remote Sensing of Environment*, vol. 278, p. 113076, 2022.

[41] S. Chen, H. Liu, X. Zeng, S. Qian, W. Wei, G. Wu, and B. Duan, "Local patch vectors encoded by fisher vectors for image classification," *Information*, vol. 9, no. 2, p. 38, 2018.

[42] A. J. Chavez, *Image classification with dense SIFT sampling: an exploration of optimal parameters*. Kansas State University, 2012.

[43] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, pp. 150–153, IEEE, 1984.

[44] J.-H. Lee, "Panoramic image stitching using feature extracting and matching on embedded system," *Transactions on Electrical and Electronic Materials*, vol. 18, no. 5, pp. 273–278, 2017.

[45] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[46] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[47] V. Vijayaraj, N. Younan, and C. O'Hara, "Concepts of image fusion in remote sensing applications," in *2006 IEEE International Symposium on Geoscience and Remote Sensing*, pp. 3798–3801, IEEE, 2006.

[48] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[49] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International journal of remote sensing*, vol. 33, no. 8, pp. 2395–2412, 2012.

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

[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[52] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[53] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1735–1739, 2017.

[54] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[55] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 220–237, 2017.

[56] J. Zhang, M. Zhang, L. Shi, W. Yan, and B. Pan, "A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation," *Remote Sensing*, vol. 11, no. 21, p. 2504, 2019.

[57] Q. Bi, K. Qin, H. Zhang, J. Xie, Z. Li, and K. Xu, "Apdc-net: Attention pooling-based convolutional network for aerial scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 9, pp. 1603–1607, 2019.

[58] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6916–6928, 2019.

[59] Y. Yu, X. Li, and F. Liu, "Attention gans: Unsupervised deep feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 519–531, 2019.

[60] R. M. Anwer, F. S. Khan, and J. Laaksonen, "Compact deep color features for remote sensing scene classification," *Neural Processing Letters*, vol. 53, no. 2, pp. 1523–1544, 2021.

[61] Y. Gao, J. Shi, J. Li, and R. Wang, "Remote sensing scene classification based on high-order graph convolutional network," *European Journal of Remote Sensing*, vol. 54, no. sup1, pp. 141–155, 2021.

[62] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 1, pp. 43–47, 2020.

[63] S. Wang, Y. Ren, G. Parr, Y. Guan, and L. Shao, "Invariant deep compressible covariance pooling for aerial scene categorization," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[64] T. Tian, L. Li, W. Chen, and H. Zhou, "Semsdnet: A multi-scale dense network with attention for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[65] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, pp. 1269–1277, 2015.

[66] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sensing*, vol. 9, no. 8, p. 848, 2017.

[67] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7109–7121, 2018.

[68] U. Muhammad, W. Wang, and A. Hadid, "Feature fusion with deep supervision for remote-sensing image scene classification," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 249–253, IEEE, 2018.

[69] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, 2015.

[70] L. Yan, R. Zhu, N. Mo, and Y. Liu, "Improved class-specific codebook with two-step classification for scene-level classification of high resolution remote sensing images," *Remote Sensing*, vol. 9, no. 3, p. 223, 2017.

[71] K. Qi, C. Yang, Q. Guan, H. Wu, and J. Gong, "A multiscale deeply described correlatons-based model for land-use scene classification," *Remote Sensing*, vol. 9, no. 9, p. 917, 2017.

[72] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2889–2901, 2017.

[73] E. Li, P. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 1068–1081, 2016.

[74] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 371–390, 2017.

[75] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Information Sciences*, vol. 348, pp. 209–226, 2016.

[76] M. Shahriari and R. Bergevin, "Land-use scene classification: a comparative study on bag of visual word framework," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 23059–23075, 2017.