TechRxiv
Powered by IEEE

A Discriminative Neighborhood-Based Collaborative Learning for Remote Sensing Scene Classification

This paper was downloaded from TechRxiv (https://www.techrxiv.org).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

25-08-2021 / 31-08-2021

CITATION

Muhammad, Usman; Hoque, Md. Ziaul; Wang, Weiqiang; Oussalah, Mourad (2021): A Discriminative Neighborhood-Based Collaborative Learning for Remote Sensing Scene Classification. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.16441593.v1

DOI

10.36227/techrxiv.16441593.v1

A Discriminative Neighborhood-Based Collaborative Learning for Remote Sensing Scene Classification

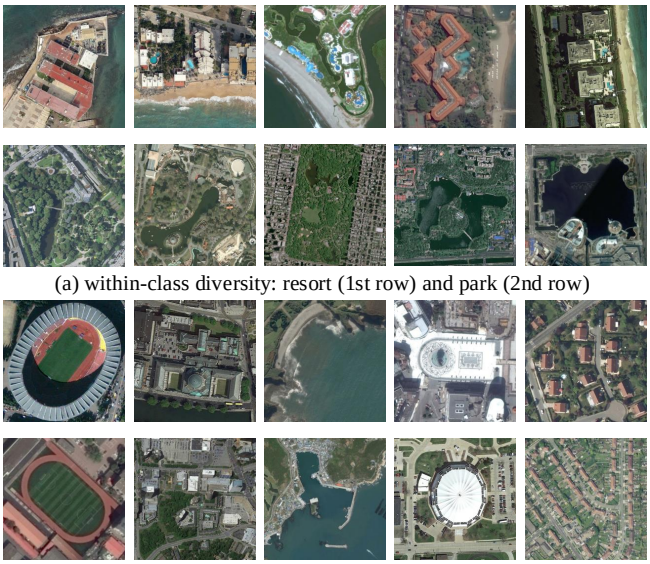
Usman Muhammad, Md. Ziaul Hoque, Weiqiang Wang, and Mourad Oussalah, *Senior Member, IEEE*

***Abstract*—The bag-of-words (BoW) model is one of the most popular representation methods for image classification. However, the lack of spatial information, change of illumination, and 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 introduce a neighborhood-based collaborative learning (NBCL) for remote-sensing scene classification. Particularly, our proposed method employs multilevel features 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 where each image is represented by four independent image region sequences. Apart from multilevel learning, we explicitly impose Gaussian 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 multilevel and multiscale 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 these features, and the fused features are incorporated into a Bidirectional Long Short-Term Memory (BiLSTM) network for constructing the final representation 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.**

***Index Terms*—Scene classification, Bag-of-words (BoW) model, Gaussian pyramids, Neighborhood-based learning, Bidirectional Long Short-Term Memory (LSTM).**

I. 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 earth not only provide high-resolution scene images but also give us an opportunity to study the spatial information with a fine-grained detail. However, within-class diversity among scene categories is one of the main challenges that brings new obstacles in image analysis as shown in Fig.1 (a). The first and second row represents resort and park scenes, respectively. A large diversity can be observed even within the same class. Here, the "scenes" belong to different types of subareas extracted from large satellite images. These subareas could be different types of land covers or objects and possess specific semantic meaning, such as commercial area, dense residential, sparse residential and parking lot in a typical urban area satellite image [68]. 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.



(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 [68]. (a) the intraclass diversity and (b) interclass similarity are the main obstacles that limit the scene classification performance. This encourages us to learn multilevel spatial features that have small within-class scatter but large between-class separation.

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) [26], scale invariant feature transform [42], color Histogram [56], and histogram oriented

Usman Muhmmad is with Center for Machine Vision and Signal Analysis, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland, and School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.

Md. Ziaul Hoque and Mourad Oussalah are with Center for Machine Vision and Signal Analysis, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland, and Weiqiang Wang is with School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.

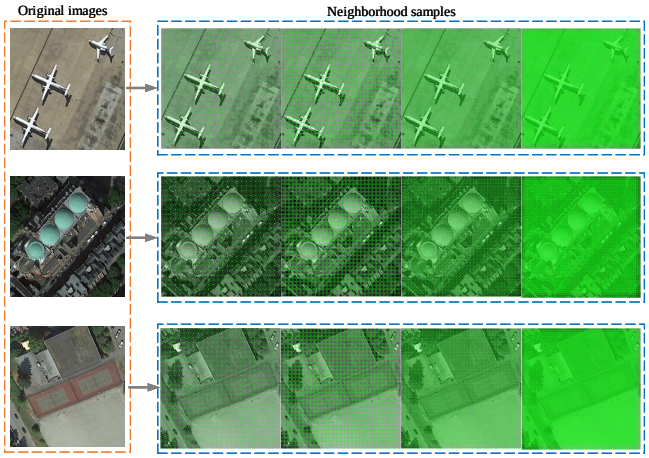


Figure 2. The main idea of the proposed work. Given a remote sensing image, BoW model is designed with different neighborhood regions to incorporate spatial information. Left column: Example images from NWPU dataset. Right column: SURF features of light, medium and dark green colors represent different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

gradients (HOG) [18]. A major shortcoming of the low-level descriptors is that not fulfilling the entire scene understanding due to high diversity and non-homogeneous spatial distributions of the scene classes. Moreover, they require complex engineering skills that rely on expert experience. 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 [33]. 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 BoW model has limited performance in 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, Wang et al. [66] proposed a domain adaptation method based on the deep neural network, where the manifold alignment is adopted on the target domain to avoid distortion. Chen et al. [12] proposed an associative learning-based domain adaptation that does not require target labeled information and can achieve unsupervised classification of the target image. A combined CNN-based recurrent neural network is proposed in [57] to exploit both local and long-range spatial relation information to enhance the classification performance of the model. However, the backpropagation process, hyperparameter setting, or training a CNN from the scratch remains challenging, and a small number of training samples often cause overfitting in deep learning models. To address these limitations, most works tend to use a pre-trained model as a feature extractor, such as VGGNet, which can easily get deep features without high computational cost [44]. However, the 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. To illustrate this observation, some images from AID dataset are displayed as an example in Fig.1 (b). One can see that both stadium and playground scenes display high appearance similarity and the diversity within the scene varies largely. Moreover, natural images can be mainly captured by cameras with manual or auto-focus options and it makes them to be center-biased [28]. 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 [44, 6] attempted to focus on the critical local image patches and discard the useless information, they still only utilize the visual information [37].

Despite the success of deep learning and BoW algorithms, the *intraclass* diversity and *interclass* similarity are two big challenges that still needed to be addressed. To alleviate these issues, we propose a multilevel learning approach to extract image features region by region based on small, medium, and large neighborhood patches to fully exploit spatial structure information in BoW model. This is motivated by the fact that even patch sizes are different in size, they exhibit good learning ability of spatial dependencies between image region features that may help to interpret the scene [60]. Our proposed method also magnifies the visual information by utilizing Gaussian pyramids and combine these two approaches to solve the problem of remote-sensing scene classification. In order to balance the contribution of these two types of features, we propose a simple fusion strategy based on three motivations. First, even though a considerable amount of literature on fusion of multiscale or multilayer features is available in scene classification, there is currently no clear consensus on the best features for large-scale scene recognition. Our second motivation is to introduce a simple fusion strategy that can surpass the previous performance without utilizing state-of-the-arts fusion methods such as DCA [10], PCA [35], CCA [46], etc., as previously utilized in remote sensing domain. The third 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 diagonalize the covariance matrix, and the lack of guarantee that distances in the original and projected spaces are well retained. In summary, we want to develop a simple bag-of-words method to make full use of spatial and visual information, which can effectively improve the classification performance.

Therefore, in this paper, we propose a simple, yet very effective approach called neighborhood-based collaborative learning (NBCL) that builds on fusion of the corresponding

multilevel and multiscale features. In particular, the NBCL encodes spatial and visual information based on small, medium, and large neighborhood regions. The extracted spatial locations which are used in our work are visualized in Fig 2. The final representation for an image is achieved by fusing small, medium, and large scale spatial and visual histograms. For classification purpose, the BiLSTM approach is adopted due to its proven efficient classification performance. In contrast with previous works which choose a certain sampling (sparse, dense, random, etc.) approach [29, 28], the proposed work defines an artificial fixed-size sliding window not only for the original image but also extend in scale space pyramid to accommodate all the multiscale patches of the image and extract local features from each patch. In summary, our main contributions in this paper are summarized as follows:

1) We present a neighborhood-based collaborative learning (NBCL) to combine all the surrounding features into a new single vector and address the problem of intraclass diversity and interclass similarity.
2) To improve the visual information, a smoothing and downsampling is preformed by convolving the image with Gaussian kernels. Simultaneously, we propose to integrate the fixed neighborhood regions into multiple downscaled versions of the input image in a scale space pyramid. In this way, we can explore more content and important information.
3) 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 II discusses the related literature work of this study. Section III introduces the proposed NBCL for remote sensing scene classification. Section IV shows the experimental results of the proposed NBCL on several public benchmark datasets. Section V summarizes the entire work and gives suggestions for future research.

II. Related work

Early attempts heavily depend on the hand-crafted features and focus on different types of color features for remote sensing scene image analysis. Since only spectral information can be utilized, 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 the local descriptors such as local binary patterns (LBPs) [26], scale invariant feature transform (SIFT) [42], color histogram [56], and histogram oriented gradients (HOG) [18]. Yu *et al.* [75] proposed a new descriptor called color-texture-structure (CTS) to encode color, texture and structure features. In their work, 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. Chen *et al.* [11] evaluated the performance of 13 features consists of color, structure, and texture features. To perform classification, 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. Tokarczyk *et al.* [59] proposed to use integral images and extract discriminative textures at different scale levels of scene images. The features are named as Randomized Quasi-Exhaustive (RQE) which are capable of covering a large range of texture frequencies.

In bag-of-words (BoW) framework, conventional coding methods focus on a single grid size or block based response to extract spatial features while not taking into account the other useful filter responses. An image often consists of a single object, but sometimes several objects or particles also appear in the surrounding. Although researchers showed that by incorporating relative spatial information, such as distance or angular directions, the technique becomes less effective as the codebook size increases. To address these challenges, Savarese *et al.* [51] introduced integral correlograms to capture spatial co-occurrences of features, which are easy to compute with respect to basic geometric transformations. The combination of correlograms and visual words is explored by forming a co-occurrence matrix of visual words as a function of distance. The co-occurrence matrix models the typical spatial correlations between visual words in object classes, yielding invariance to rotation and translation. In order to improve the classification performance, the co-occurrence matrix with the orderless BoW is also utlized in [74]. However, the correlogram matrix takes expensive computation power and memory cost [51].

Khan *et al.* [31] investigated multiple hand-crafted color features in bag-of-word model. In their work, color and shape cues are used to enhance the performance of the model. Yang *et al.* [73] improved the BoW model based on the spatial co-occurrence kernel, where two spatial extensions are proposed to emphasize the importance of spatial structure in geographic data. Vigo *et al.* [63] proved that incorporating color and shape in both feature detection and extraction significantly improves the bag-of-words based image representation. Sande *et al.* [61] 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.* [33] 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 is merged in the traditional SPM [72]. To further enhance the spatial information, a new coding algorithm called Locality-constrained Linear Coding (LLC) is introduced that utilizes the locality constraints to project each descriptor into its local-coordinate system. Then, a max-pooling is used to integrate projected coordinates [64]. Computational efficiency is achieved by introducing a soft-assignment coding that computes the distance from a local feature to each word, and allows each feature vector to belong to multiple histogram bins [38]. However, the classification performance is limited to the newly generated sparse or local coding schemes. Boureau *et al.* [8] analyzed the various combinations of coding and pooling schemes through a comprehensive cross evaluation of several types of pooling and coding stages: hard assignment vs. soft assignment, the linear kernel vs. the histogram intersection

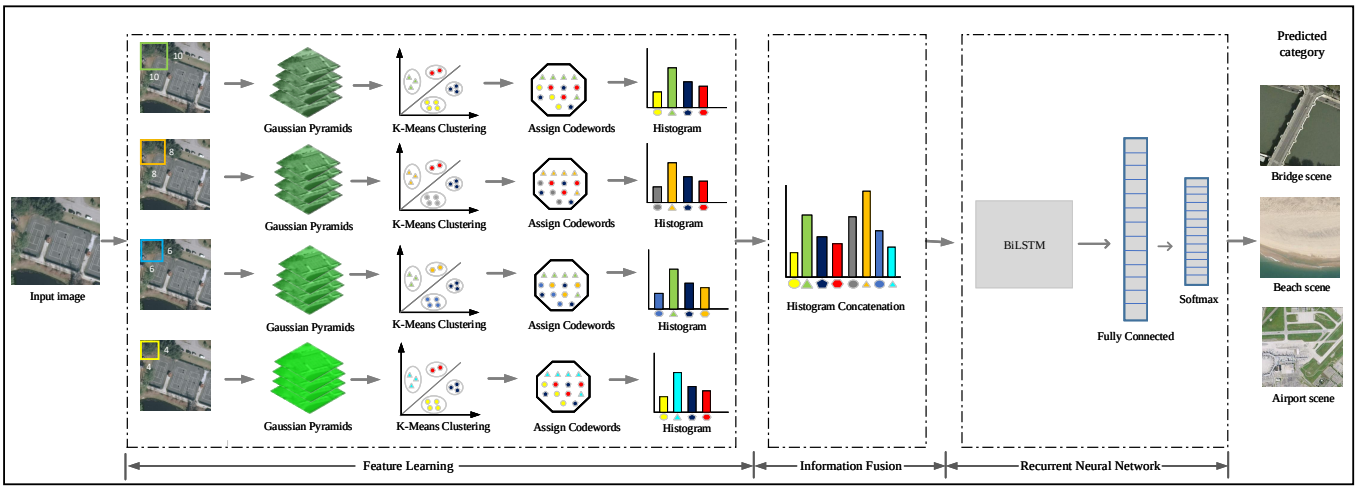


Figure 3. Flowchart of the proposed method. The local patches are selected by a fixed-size sliding window, where green, orange, blue, and yellow rectangles represent the sizes of 10×10, 8×8, 6×6, and 4×4, respectively. The dense interest points are encoded to a specific codeword through the k-means clustering process. Finally, a concatenated histogram is used as an input for training the BiLSTM network.

kernel and average pooling vs. max-pooling.

Zhou *et al.* [79] introduced a hierarchical Gaussian mixture model for feature vectors at difference levels, and several Gaussian maps for its spatial layout. To alleviate noise directions and to further enhance the discrimination power, a supervised dimension reduction technique is introduced called Discriminant Attribute Projection (DAP). In contrast to k-means clustering, a modified coding approach is introduced in [32] called Fisher vector coding. They extend BoW by using a Gaussian mixture model to encode spatial layout. To encode the appearance of local features, representation of spatial layout is combined with the use of Fisher kernels. While their representation is computationally efficient and compact, their evaluations indicates marginal improvement over SPM. Gabriel *et al.* [47] presented a new method, called Sparse Spatial Coding (SSC). In their work, they build the dictionary with a set of random patches and code a descriptor using a spatial constraint. Their method is closely related to LLC [64].

In recent years, a large number of studies have been conducted to merge attention mechanism and multiscale learning based on deep learning models. Ghanbari *et al.* [21] proposed a 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 global spatial pyramid pooling module to obtain more abundant multiscale features at different levels. Zhao *et al.* [78] proposed a deep learning model which simultaneously captures spectral-spatial features of the target pixel and its neighboring pixels for classification. The authors used a center attention module that pays more attention to the features correlated to target pixels and reduce the number of parameters in the network via weighted sum of the spectral-spatial features. Zuo *et al.* [81] 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 processed to generate mid level features and RNN layer is learned to encode spatial dependencies.Xue *et al.* [70] proposed a hierarchical residual network to extract multiscale and spectral features at a granular level. The authors used an attention mechanism to set adaptive weights for spatial and spectral features of different scales for the further improvement of the discriminative ability of extracted features. Ran *et al.* [50] proposed a multiscale context and enhanced channel attention model that employs PeleeNet as the backbone network. The authors improved the characterization ability of the convolutional neural network by proposing channel attention approach. Huang *et al.* [30] proposed an end-to-end deep learning model and employ multiscale feature fusion, a channel-spatial attention, and a label correlation extraction module. Specifically, a channel-spatial attention mechanism is used to fuse and refine multiscale features from different layers of the CNN model. Moreover, a label co-occurrence matrix is utilized to extract the label correlation information and embedded into the multiscale attentive features which increases the discriminative ability of their proposed model.

Mei *et al.* [43] proposed a sparse representation-based model with deep feature fusion. Multilevel features are extracted from different layers of convolutional neural networks to exploit the feature learning ability. Zhang *et al.* [77] proposed a multi-scale deep feature representation and the region-based features selection. The model first filters the multi-scale deep features extracted from pre-trained convolutional networks and then fuses those features via their proposed fusion strategy. The authors utilized the complementarity between local and global features by exploiting the features of different scales and discarding the redundant information in features. Liang *et al.* [37] introduced a novel two-stream architecture combining global-based visual features and object-based features. The model first extracts the appearance visual features from the scene image using convolutional neural network

and later detects the ground objects and finally constructs a graph to learn spatial features using a graph convolutional network. Tian *et al.*[58] proposed a multiscale dense network based on squeeze and excitation attention, dense connections and squeeze-and-excitation attention mechanism. The authors imposed two settings with computational constraints including budgeted batch classification (a fixed computational budget setting for sample classification) and prediction module that forces the network to predict the output.

Fu *et al.* [19] proposed a feature fusion architecture to generate a multiscale features hierarchy that augments the features of shallow layers with semantic representations and combine the feature maps of top layers with low-level information. The authors built a unified framework upon the region-based convolutional neural network for arbitrary-oriented and multi-scale object detection. Cheng *et al.* [17] proposed a multi model fusion neural network with the combination of a convolutional neural network and a multilayer perceptron to estimate a fine-resolution population mapping. This model takes the local spatial information and global information from multisource data to estimate the fine-resolution population where a first-order space matrix of a geographic unit is used to characterize these information. Qu *et al.* [49] proposed a novel multiscale deeply supervised convolutional feature fusion module. The authors used multiscale feature fusion by using the high-level features and the low-level features with deep supervision that provides direct supervision to improve the performance of the model. Li *et al.* [36] proposed an adaptive multilayer 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.

In overall, patch sampling or feature learning is a key step for building up an intelligent system either for CNN model or BoW-based approaches. In deep learning CNN models, convolutional layers convolve the local image regions independently, and pass its result to the next layer, whereas pooling layers summarize the dimensions of data. Due to 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 has been shifted to multiscale and fusion methods in scene image classification domain, and existing deep learning methods are making full use of multiscale information and fusion for better representation.

A sampling step is generally required to select uniform, sparse, random, and dense representative subset of the images for subsequent analysis. The most popular BoW methods utilize a fixed grid size and extract features at different scales [39, 24, 3]. This is because it is still not clear which sampling strategy is suitable, and even no optimal patch size can be set for the scene classification of high-resolution remote sensing images. Due to the unique nature of remote sensing images, discriminative patch sizes are needed that can focus on the significant regions within the image itself in the procedure of feature learning. Thus, it stimulates us to research whether the correlation of scene categories can increase the discriminative ability of BoW features. NBCL provides a natural approach to model the spatial dependencies among scene categories, where we propose to merge the fixed patch size for all downscaled versions of the original image in a scalespace pyramid.

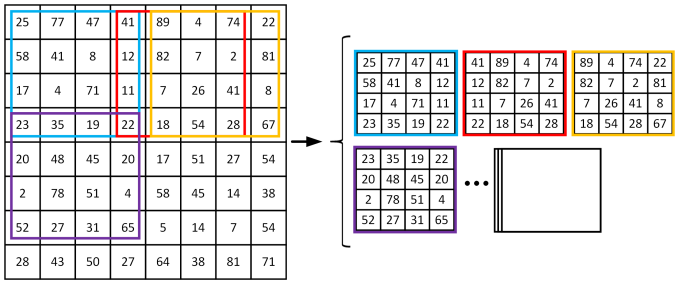


Figure 4. Predefined patch size (4 × 4) with the sliding step is set to be 1 pixel before representing features over entire image.

III. Proposed Method

The proposed approach is divided into three indispensable components: (1) estimation of multi-neighborhood regions (2) information fusion and (3) a BiLSTM based sub-network for classification purpose. We first describe the procedure of neighborhood estimation with multi-scale filtering. 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.

A. Features extraction using multi-neighborhood regions

In order to explore the spatial relationship between scenes or sub-scenes, we propose to extract multilevel features with the objective that different regions contain discriminative characteristics that can be used to extract more meaningful information and to correctly classify target samples. Based on our observation, the size of the neighborhood has a great impact on the scene representations and classification performance. To achieve this, we first define a region over the entire image, where the patch sizes used are (4×4), (6×6), (8×8), (10×10), and the sliding step is set to be 1 pixel. An example image with the neighborhood patch (4×4) size is provided in Fig.4 to show how the local descriptor is exploited in each part of the region. Here, the definition of different neighborhood size is considered to be small, medium, or large regions. Thus, four kinds of sizes are used for each image to ensure that the output is full of content information. To achieve multiscale 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 by removing image noise. Inspired by the Gaussian scale-space [67], images are repeatedly smoothed with appropriately sized Gaussian kernels by convolving over image:

$$I(a, b, \sigma_i) = I(a, b) * G(a, b, \sigma_i), \tag{1}$$

where * is the convolution operator and $G(a, b, \sigma_i)$ is the Gaussian kernel with the standard deviation * defined by $G(a, b, \sigma_i) = \frac{1}{2\pi\sigma_i^2}\exp-\frac{a^2+b^2}{2\sigma_i^2}$. In this way, noise and illumination factors are suppressed by using these smoothed images

Table I
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

and coarser structures are emphasized since no new structures are created after smoothing. Therefore, the proposed idea takes the advantages of both schemes. We experimentally study the outcome of this choice in ablation study section. Once the scale space has been built, we utilize SURF descriptor [4] 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 below:

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

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.

B. 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 [55]. Among them, the 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. 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, in the aim to obtain more informative and relevant features to represent the input image. Each input image I produced four sets of multiscale features, that is Q_1, Q_2, Q_3, and Q_4 representing four sets of features. The final fused vector illustration is obtained as:

$$Q_q(I) = Q_1(I) + iQ_2(I) + iQ_3(I) + iQ_4(I), \tag{3}$$

where i is the imaginary unit.

C. Recurrent Neural Network (RNN)

The multiscale histograms from different regions provide crucial information for understanding the spatial structure. We concatenate them into a final representation, which is then used as an input to Bidirectional neural network [52]. The Bidirectional Long Short-Term Memory Networks learns the correlations of features and encodes the feature histograms based on the memory cell (M_t), and are competent for retaining track of the dependencies between the elements in the input sequence. It consists of an input gate (i_t), an output gate (o_t) and a forget gate (f_t). The input gate controls the information flow into the cell by multiplying the cell's non-linear transformation of inputs n_t. The output gate governs how much information from the cell is employed to calculate the output activation of the LSTM unit. The forget gate decides the amount to which a value remains in the cell. The LSTM unit updates for time step t are:

$$\begin{bmatrix} f_t \\ i_t \\ n_t \\ o_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \\ \sigma \end{bmatrix} H \cdot [p_{t-1}, x_t] \tag{4}$$

$$M_t = f_t \odot M_{t-1} + n_t \odot i_t \tag{5}$$

$$p_t = \tanh(M_t) \odot o_t \tag{6}$$

where x_t denotes the input at the current time-step, i_t denotes the current cell state. f, i and n represent the input gate activation, forget gate activation and output gate activation respectively, σ illustrates the logistic sigmoid function and $\odot$ represents element-wise multiplication.

IV. 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. 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 [73]. 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 [54]. 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

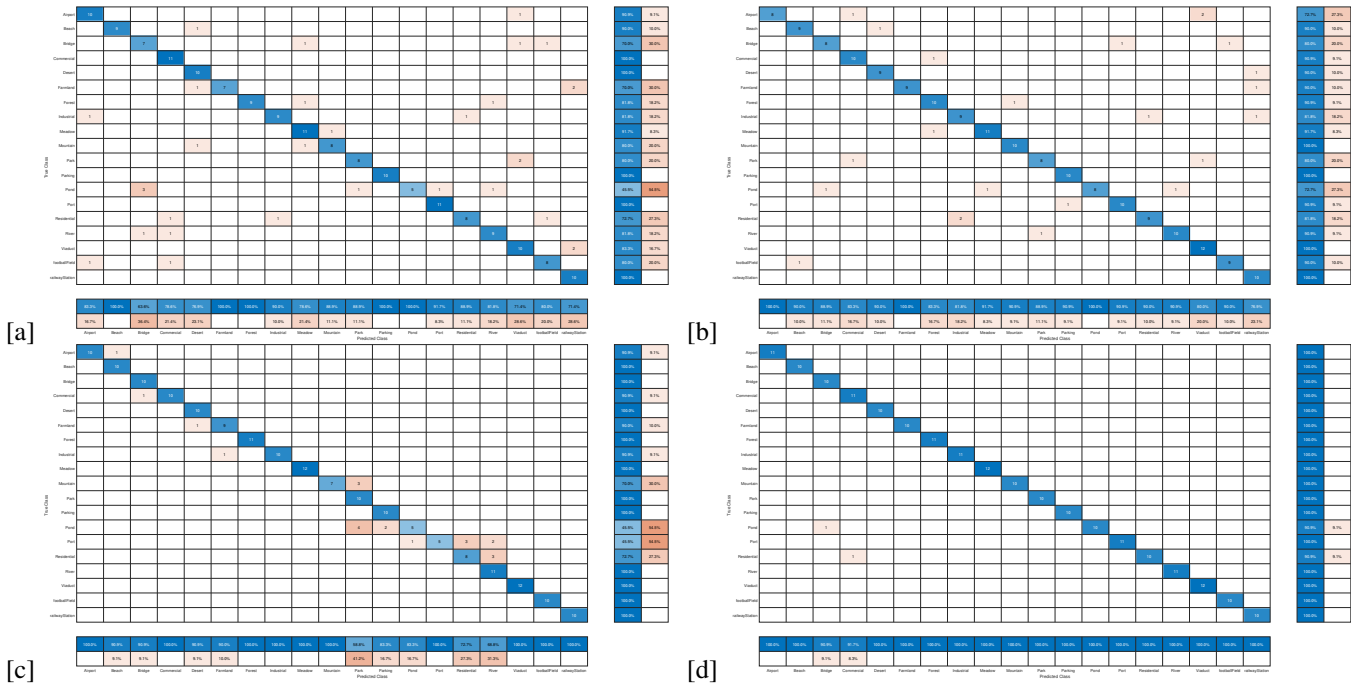


Figure 5. Confusion matrix of our proposed method on WHU-RS dataset by fixing the training ratio as 80% (a) with one-stage learning, (b) with two-stage learning, (c) with three-stage learning, and (d) with multi-stage learning. Zoom in for a better view.

AID dataset, which are categorized into 30 scene classes [68]. 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 [13]. 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.

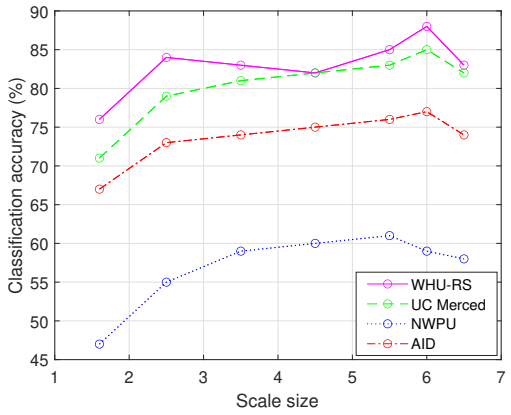


Figure 6. Classification accuracy of the proposed method under different Gaussian scales for four datasets.

B. Implementation details

To evaluate the performance on the above-mentioned datasets, the BoW is used as the base architecture with four distinct neighborhood sizes 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 WHU-RS dataset. The BiLSTM is trained using the Adam optimizer with gradient threshold 1, while the minibatch size of 32 with 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) [22] 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.

C. Ablation study

We thoroughly validate the performance of each neighborhood size by performing an ablation study. In Table I, we have reported results of estimating the proposed NBCL on UC Merced and WHU-RS datasets. Note that, these experiments are performed by using four different neighborhood sizes with six Gaussian scaled $[1.6, 2.5, 3.5, 4.5, 5.5, 6.0]$ images. Our one-stage detection method on WHU-RS dataset with the size of (4×4) achieves 86.10% accuracy and the numerical results of each category are shown in Fig.5 (a). It can be

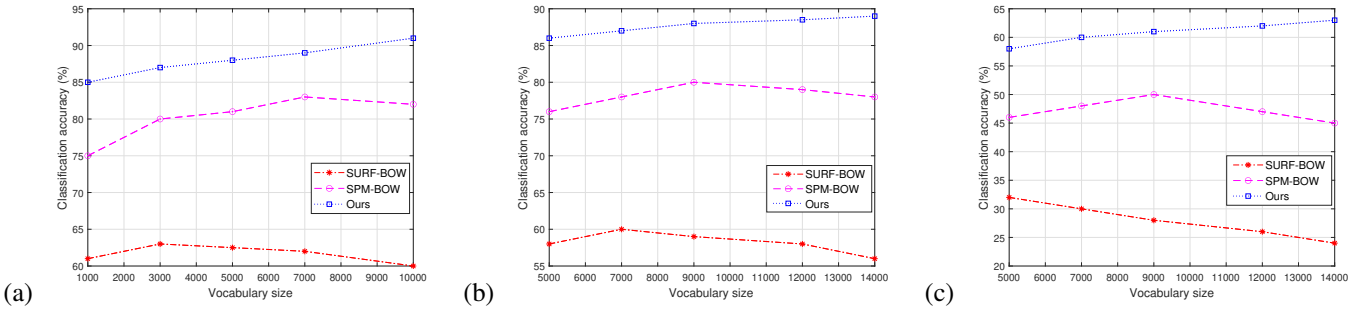


Figure 7. Comparison of classification on three datasets. (a) Comparing the performance on WHU-RS dataset with SURF-BOW [4], SPM-BOW [33], and ours. (b) Comparing the performance on UC Merced dataset using SURF-BOW [4], SPM-BOW [33], and ours. (c) Comparing the performance on NWPU dataset using SURF-BOW [4], SPM-BOW [33], and ours.

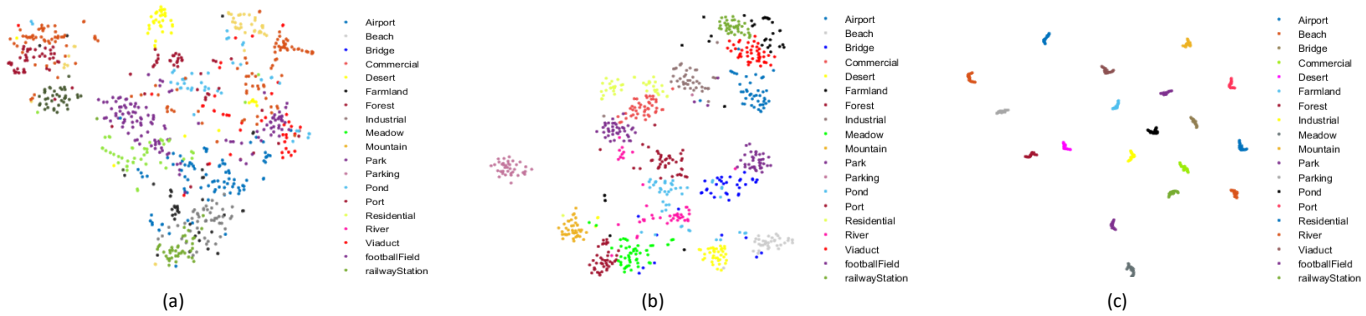


Figure 8. Two-dimensional scatterplots of SURF-based BoW features genertaed with t-SNE over the WHU-RS dataset. (a) Scatterplot of one-stage multiscale features. (b) Scatterplot of features extracted and combined from four-stage learning. (c) Features extracted from the last fully-connected layer of BiLSTM. All points in the scatterplots are class coded.

seen that several classes such as bridge 63%, viaduct 71%, railway station 71%, meadow 78%, and desert 76% are highly misclassified. In Fig.5 (b), we show that when the neighborhood size (6×6) increases, the classification performance is improved up to 88%, 80%, 77%, and 90% for bridge, viaduct, railway station, meadow, and desert, respectively. The overall classification is improved up to 89%, which is 3% higher than the (4×4) size. We further increase the size up to (8×8) and notice that viaduct, railway station, meadow are 100% correctly classified and achieve 92% classification accuracy. Results demonstrate that different neighborhood sizes play different roles in classifying remote sensing scene images, and it is hard to determine an optimal size to obtain the best result. On the other hand, 4×4 block size outperforms other block sizes in UC Merced dataset while 8×8 block size is more effective for the WHU-RS dataset as shown in Table I.

1) Scale Factor of Gaussian Kernel: Fig.6 shows the classification performance of each scaled image based on 10×10 neighborhood size. The NBCL extracts multiscale dense features according to the scale factor σ to control the Gaussian kernel. It can be observed that with the increase of scale factor, the performance first improves and than gradually decreases after the 6.0 scaled image. We conclude that including a certain range of Gaussian smoothed images can improve the performance, but too much of them degrade the performance.

2) Codebook learning: We quantitatively analyze the performance with the SURF descriptor and 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 multiscale 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.7 (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 with ours on all databases.

3) 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 [62] and illustrated the derived embeddings into three separated processing stages: 1) one-stage learning, 2) combined learning (NBCL), and 3) BiLSTM classified features for the WHU dataset. The features with the patch size of 4×4 in Fig.8 (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.8 (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.8 (c) that all the classes are well separable

Table II
CLASSIFICATION ACCURACY (%) FOR THE NWPU DATASET WITH TWO TRAINING RATIOS.

Method	10%	20%
BoW with dense SIFT [15]	41.72±0.21	44.97±0.28
BOCF [15]	82.65±0.31	84.32±0.17
BoVW+SPM [14]	27.83±0.61	32.96±0.47
D-CNN [16]	89.22±0.50	91.89±0.22
Triple networks [40]	-	92.33±0.20
MDFR [77]	83.37±0.26	86.89±0.17
APDC-Net [5]	85.94±0.22	87.84±0.26
BoWK [46]	-	66.87±0.90
SFCNN [69]	89.89±0.16	92.55±0.14
Attention GANs [76]	86.11±0.22	89.44±0.18
MDFR [77]	83.37±0.26	86.89±0.17
CNN + GCN [37]	90.75±0.21	92.87±0.13
Color fusion [1]	-	87.50±0.00
Graph CNN [20]	91.39±0.19	93.62±0.28
AlexNet+SAFF [9]	80.05±0.29	84.00±0.17
VGG-VD16+SAFF [9]	84.38±0.19	87.86±0.14
IDCCP [65]	91.55±0.16	93.76±0.12
SEMSDNet [58]	91.68±0.39	93.89±0.63
NBCL (The proposed)	**94.20±0.81**	**97.13±0.92**

Table III
CLASSIFICATION ACCURACY (%) FOR THE AID DATASET WITH TWO TRAINING RATIOS.

Method	20%	50%
Fusion by addition [10]	-	91.87±0.36
D-CNN [16]	90.82±0.16	96.89±0.10
MDFR [77]	90.62±0.27	93.37±0.29
APDC-Net [5]	88.56±0.29	92.15±0.29
SFCNN [69]	94.93±0.31	96.89±0.10
Attention GANs [76]	93.97±0.23	96.03±0.16
CNN + GCN [37]	94.93±0.31	96.89±0.10
Color fusion [1]	-	94.00±0.00
AlexNet+SAFF [9]	87.51±0.36	91.83±0.27
VGG-VD16+SAFF [9]	90.25±0.29	93.83±0.28
Graph CNN [20]	93.06±0.26	95.78±0.37
IDCCP [65]	94.80±0.18	96.95±0.13
SEMSDNet [58]	94.23±0.63	97.64±0.51
NBCL (The proposed)	**96.11±0.81**	**98.43±0.33**

which could potentially lead to a better performance when training BiLSTM on remote sensing dataset.

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

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 is shown in Table II. Especially, we choose mainstream deep learning and BoW based methods and compare the performance of scene classification. It could be observed from Table II, 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%. The classification performance of the proposed NBCL shows the effectiveness of combining global-based visual features on the NWPU dataset.

Fig.10 illustrates the confusion matrix produced by our proposed method (NBCL) 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 [37]. However, one common challenge is found that the church and palace are two confusing categories which limits many existing works to surpass the performance [37]. In our case, 25% of images from church are mistakenly classified as a palace which is 1% high misclassification than the CNN + GCN [37]. 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% [69] and 70% [37] performance for the palace class. By analyzing the confusion matrix on NBCL, 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.

Table IV
COMPARISON OF CLASSIFICATION ACCURACY (%) FOR THE UC-MERCED DATASET WITH 80% RATIOS.

Method	Accuracy (Mean±std)
AlexNet+sum pooling [2]	94.10±0.93
VGG-VD16+sum pooling [2]	91.67±1.40
SPP-Net [25]	96.67±0.94
GoogleNet [68]	94.31±0.89
VGG-VD16 [68]	95.21±1.20
DCA fusion [10]	96.90±0.77
MCNN [41]	96.66±0.90
D-CNN [16]	98.93±0.10
Triple networks [40]	97.99±0.53
VGG-VD16 +AlexNet [35]	98.81±0.38
Fusion by concatenation [45]	98.10±0.20
MDFR [77]	98.02±0.51
APDC-Net [5]	97.05±0.43
BoWK [46]	97.52±0.80
Attention GANs [76]	97.69±0.69
AlexNet+SAFF [9]	96.13±0.97
VGG-VD16+SAFF [9]	97.02±0.78
Color fusion [1]	98.10±0.00
Graph CNN [20]	99.00±0.43
IDCCP [65]	99.05±0.20
SEMSDNet [58]	99.41±0.14
NBCL (The proposed)	**99.57±0.36**

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 III. It could be observed that NBCL achieved the overall accuracy of 96.11% and 98.43% using 20% and 50% training ratios, respectively. As can be seen from Table III, our method outperformed the SEMSDNet [58] 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 multilevel and multiscale feature fusion.

Fig.11 represents the confusion matrix generated by NBCL with the 50% training ratio. As can be seen from Fig. 11, the

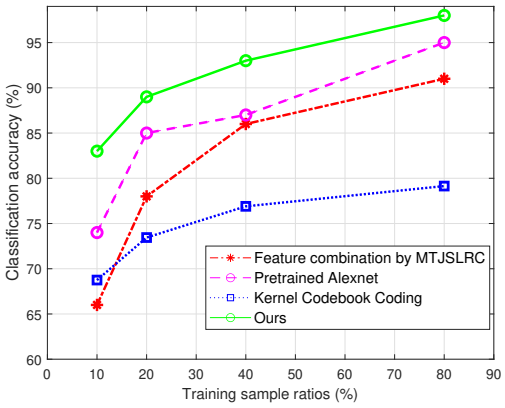


Figure 9. The influence of the training sample ratios with different methods such as feature combination by MTJSLRC [80], Pretrained Alexnet [25], and kernel codebook coding [53].

Table V
COMPARISON OF CLASSIFICATION ACCURACY (%) FOR THE WHU-RS19 WITH 80% RATIOS.

Method	Accuracy (Mean±std)
Transferring CNNs (Case I) [27]	96.70±0.00
Transferring CNNs (Case II) [27]	98.60±0.00
Two-Step Categorisation [71]	93.70±0.57
CaffeNet [68]	94.80±0.00
GoogleNet [68]	92.90±0.00
VGG-VD16 [68]	95.10±0.00
MDDC [48]	98.27±0.53
salM^3LBP-CLM [7]	96.38±0.76
AlexNet-SPP-SS [25]	95.00±1.12
VGG-VD19 [35]	98.16±0.77
DCA by addition [10]	98.70±0.22
MLF [34]	88.16±2.76
Fusion by concatenation [45]	99.17±0.20
D-DSML-CaffeNet [23]	96.64±0.68
BoWK [46]	99.47±0.60
Color fusion [1]	96.60±0.00
NBCL (The proposed)	**99.63±0.42**

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, 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 [58]. For instance, SFCNN [69] and the CNN + GCN [37] 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.

3) UC Merced Dataset: The evaluation results on the UC Merced dataset are presented in Table IV by using 80% training ratio. The proposed method achieves 99.57% accuracy and competes the previous BoW [46] approach by a margin of 2.05%. 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. 9. 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. For further evaluation, a confusion matrix of the UC Merced dataset is shown in Fig.12. 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. Thus, the proposed method is effective to classify most of the scene categories.

4) WHU-RS Dataset: Table V reports the comparison results of the WHU-RS dataset. As shown in Table V, the NBCL 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.5 (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.

V. CONCLUSION

In this paper, we explore the spatial dependencies between different image regions and propose a discriminative neighborhood-based collaborative learning to address the problem of interclass similarity in remote sensing scene images. Multi-neighborhood local patches are firstly proposed to preserve the spatial information between scenes or subscenes. Afterward, the preserved spatial locations are used in downscaled versions of the input image to get dense coverage of the entire scene in a scale-space pyramid. We show that multi-neighborhood learning in BoW model significantly improves the recognition performance compared to using the single level BoW alone. By combining BoW and RNN, we can learn discriminative feature representations. Experiments are conducted on four publicly available datasets, and the results demonstrate 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. We concluded the spatial information plays an important role and encourages multi-neighborhood strategy in scene classification. Our future work is to design an end-to-end method that can automatically obtain multilevel and multi-scale features without human intervention.

VI. DECLARATION OF COMPETING INTEREST

The authors have no conflict of interest that could have appeared to influence the work reported in this paper.

VII. ACKNOWLEDGMENTS

This work is supported by the University of Chinese Academy of Sciences, Beijing, China and the Center for

Machine Vision and Signal Analysis (CMVS) in the Faculty of Information Technology and Electrical Engineering (ITEE) at University of Oulu, Finland. The financial support of the world academy of sciences (TWAS) is gratefully acknowledged.

APPENDIX

Figure 10, Figure 11, and Figure 12 represent the confusion matrices of NWPU-RESISC45, AID, and UC Merced datasets, respectively. The source code for reproducing the results of this research would be available to the research community upon the publication of this work.

REFERENCES

[1] R. M. Anwer, F. S. Khan, and J. Laaksonen. Compact deep color features for remote sensing scene classification. *Neural Processing Letters*, 53(2):1523–1544, 2021.

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

[3] J. E. Ball, D. T. Anderson, and C. S. Chan Sr. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4):042609, 2017.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[5] 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*, 17(9):1603–1607, 2019.

[6] 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*, 29:4911–4926, 2020.

[7] 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*, 10(6):2889–2901, 2017.

[8] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2559–2566. IEEE, 2010.

[9] 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*, 18(1):43–47, 2020.

[10] 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*, 55(8): 4775–4784, 2017.

[11] 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*, pages 385–388. IEEE, 2011.

[12] M. Chen, L. Ma, W. Wang, and Q. Du. Augmented associative learning-based domain adaptation for classification of hyperspectral remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6236–6248, 2020.

[13] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[14] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[15] 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*, 14(10):1735–1739, 2017.

[16] 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*, 56(5):2811–2821, 2018.

[17] L. Cheng, L. Wang, R. Feng, and J. Yan. Remote sensing and social sensing data fusion for fine-resolution population mapping with a multimodel neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5973–5987, 2021.

[18] 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)*, volume 1, pages 886–893. Ieee, 2005.

[19] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 161:294–308, 2020.

[20] 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*, 54(sup1): 141–155, 2021.

[21] 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*, 14:3602–3613, 2021.

[22] 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*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[23] 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*, 56(1):371–390, 2017.

[24] Y. Gu, Y. Wang, and Y. Li. A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Applied Sciences*, 9(10):2110, 2019.

[25] 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*, 9(8):848, 2017.

[26] D.-C. He and L. Wang. Texture unit, texture spectrum,

and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, 28(4):509–512, 1990.

[27] 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*, 7(11):14680–14707, 2015.

[28] 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)*, pages 2389–2392. IEEE, 2015.

[29] J. Hu, G.-S. Xia, F. Hu, and L. Zhang. Dense vs sparse: A comparative study of sampling analysis in scene classification of high-resolution remote sensing imagery. *arXiv preprint arXiv:1502.01097*, 2015.

[30] 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.

[31] F. S. Khan, J. Van de Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 98(1):49–64, 2012.

[32] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *2011 International Conference on Computer Vision*, pages 1487–1494. IEEE, 2011.

[33] 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)*, volume 2, pages 2169–2178. IEEE, 2006.

[34] 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*, 10(3):1068–1081, 2016.

[35] 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*, 55(10):5653–5665, 2017.

[36] 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*, 12 (6):563–572, 2021.

[37] 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*, 13:4325–4338, 2020.

[38] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *2011 International Conference on Computer Vision*, pages 2486–2493. IEEE, 2011.

[39] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019.

[40] Y. Liu and C. Huang. Scene classification via triplet networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1):220–237, 2017.

[41] Y. Liu, Y. Zhong, and Q. Qin. Scene classification based on multiscale convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12): 7109–7121, 2018.

[42] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[43] S. Mei, K. Yan, M. Ma, X. Chen, S. Zhang, and Q. Du. Remote sensing scene classification using sparse representation-based framework with deep feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[44] 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)*, pages 1622–1627. IEEE, 2018.

[45] 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)*, pages 249–253. IEEE, 2018.

[46] 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*, 13(4):395–403, 2019.

[47] G. L. Oliveira, E. R. Nascimento, A. W. Vieira, and M. F. Campos. Sparse spatial coding: A novel approach for efficient and accurate object recognition. In *2012 IEEE International Conference on Robotics and Automation*, pages 2592–2598. IEEE, 2012.

[48] 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*, 9(9):917, 2017.

[49] Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou. A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[50] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li. Lightweight oriented object detection using multi-scale context and enhanced channel attention in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[51] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2033–2040. IEEE, 2006.

[52] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[53] M. Shahriari and R. Bergevin. Land-use scene classification: a comparative study on bag of visual word

framework. *Multimedia Tools and Applications*, 76(21): 23059–23075, 2017.

[54] 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*, 33(8):2395–2412, 2012.

[55] 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*, 38(12):2437–2448, 2005.

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

[57] C. Tao, W. Lu, J. Qi, and H. Wang. Spatial information considered network for scene classification. *IEEE Geoscience and Remote Sensing Letters*, 18(6):984–988, 2020.

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

[59] 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*, 53(1): 280–295, 2014.

[60] T. Tuytelaars. Dense interest points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2281–2288. IEEE, 2010.

[61] 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*, 32(9):1582–1596, 2009.

[62] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[63] 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*, pages 1549–1553. IEEE, 2010.

[64] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3360–3367. IEEE, 2010.

[65] 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.

[66] Z. Wang, B. Du, Q. Shi, and W. Tu. Domain adaptation with discriminative distribution and manifold embedding for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 16(7):1155–1159, 2019.

[67] A. Witkin. Scale-space filtering: A new approach to multi-scale description. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 150–153. IEEE, 1984.

[68] 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*, 55(7):3965–3981, 2017.

[69] 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*, 57(9):6916–6928, 2019.

[70] Z. Xue, X. Yu, B. Liu, X. Tan, and X. Wei. Hresnetam: Hierarchical residual network with attention mechanism for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3566–3580, 2021.

[71] 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*, 9(3):223, 2017.

[72] 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*, pages 1794–1801. IEEE, 2009.

[73] 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*, pages 270–279, 2010.

[74] Y. Yang and S. Newsam. Geographic image retrieval using local invariant features. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2):818–832, 2012.

[75] H. Yu, W. Yang, G.-S. Xia, and G. Liu. A color-texture-structure descriptor for high-resolution satellite image classification. *Remote Sensing*, 8(3):259, 2016.

[76] Y. Yu, X. Li, and F. Liu. Attention gans: Unsupervised deep feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1): 519–531, 2019.

[77] 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*, 11(21):2504, 2019.

[78] Z. Zhao, D. Hu, H. Wang, and X. Yu. Center attention network for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3415–3425, 2021.

[79] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang. Hierarchical gaussianization for image classification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1971–1977. IEEE, 2009.

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

[81] 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*, pages 18–26, 2015.

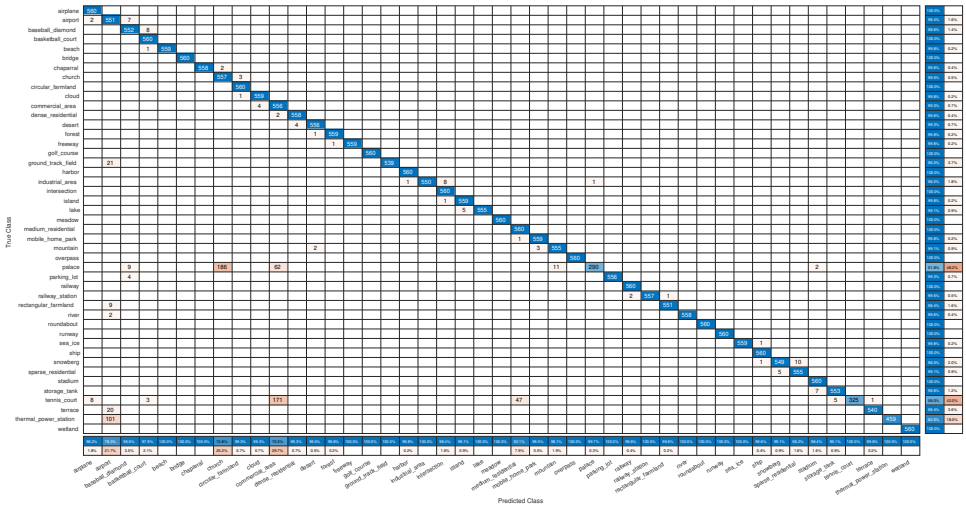


Figure 10. Confusion matrix of our proposed method on NWPU-RESISC45 Dataset by fixing the training ratio as 20%. Zoom in for a better view.

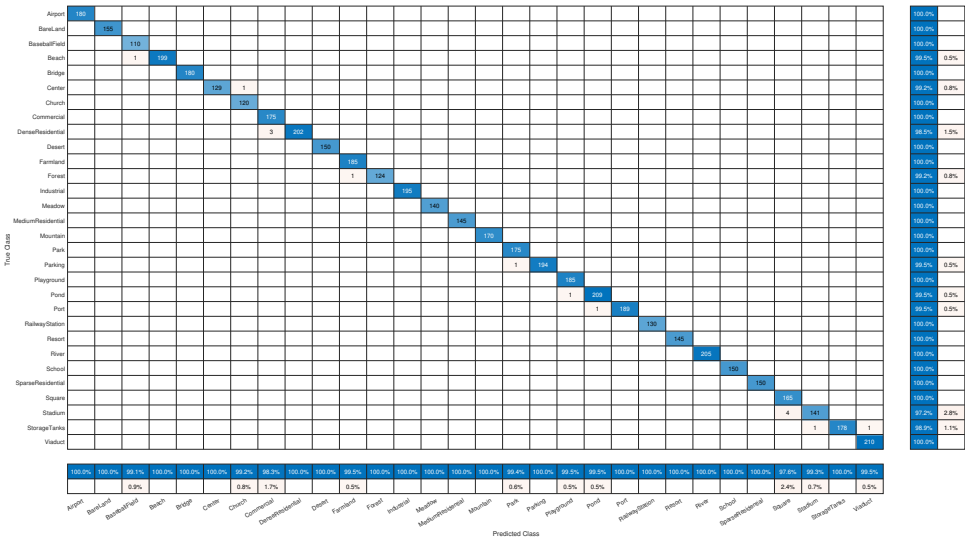


Figure 11. Confusion matrix of our proposed method on AID Dataset by fixing the training ratio as 50%. Zoom in for a better view.

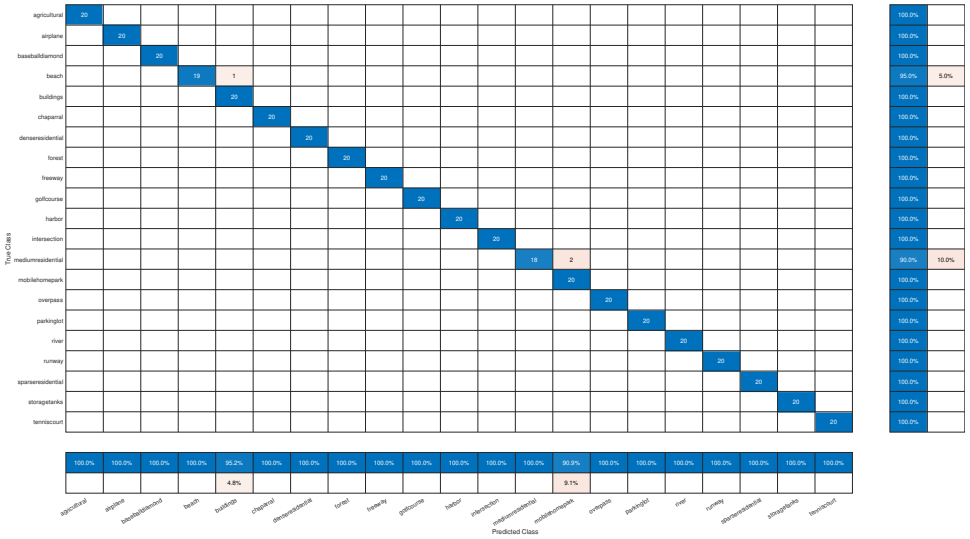


Figure 12. Confusion matrix of our proposed method on UC Merced Dataset by fixing the training ratio as 80%. Zoom in for a better view.