Adaptive Deep Co-Occurrence Feature Learning Based on Classifier-Fusion for Remote Sensing Scene Classification

Ronald Tombe and Serestina Viriri, *Senior Member, IEEE*

***Abstract*—Remote sensing scene classification has numerous applications on land cover land use. However, classifying the scene images into their correct categories is a challenging task. This challenge is attributable to the diverse semantics of remote sensing images. This nature of remote sensing images makes the task of effective feature extraction and learning complex. Effective image feature representation is essential in image analysis and interpretation for accurate scene image classification with machine learning algorithms. The recent literature shows that convolutional neural networks are mighty in feature extraction for remote sensing scene classification. Additionally, recent literature shows that classifier-fusion attains superior results than individual classifiers. This article proposes the adaptive deep co-accordance feature learning (ADCFL). The ADCFL method utilizes a convolutional neural network to extract spatial feature information from an image in a co-occurrence manner with filters, and then this information is fed to the multigrain forest for feature learning and classification through majority votes with ensemble classifiers. An evaluation of the effectiveness of ADCFL is conducted on the public datasets Resisc45 and Ucmerced. The classification accuracy results attained by the ADCFL demonstrate that the proposed method achieves improved results.**

***Index Terms*—Adaptive deep co-occurrence learning, deep feature extraction, ensemble learning, machine learning, multigrained forests, scene classification.**

I. Introduction

THE key problem in computer vision is to develop algorithms for effective image feature processing to detect and group objects into categories independent of scale, illumination, clutter, and pose positions. The fundamental question is, how can a vision system learn image feature representations effectively, given the huge volumes of image data with diverse contents? Several image analysis and interpretation techniques extract features from images [1]–[3]; these features are given to learning classifiers which apply similarity and dissimilarity rules to solve pattern recognition problems with positive and negative examples.

Manuscript received September 22, 2020; revised October 30, 2020 and November 21, 2020; accepted December 4, 2020. Date of publication December 11, 2020; date of current version January 6, 2021. This work was supported in part by the University of KwaZulu-Natal. *(Corresponding author: Ronald Tombe.)*

The authors are with the School of Computer Science, University of KwaZuLu-Natal, Durban 4000, South Africa (e-mail: ronaldtombe@gmail.com; viriris@ukzn.ac.za).

Digital Object Identifier 10.1109/JSTARS.2020.3044264

Effective feature learning is of high significance for the construction of reliable applications. These applications can be in various contexts such as management and conservation of natural resources [4], urban planning [5], precision agriculture [6], and disaster management [7]. Convolutional neural networks (CNNs) extract high capacity image feature parameters through convolution and pooling processes to yield image feature representations. In this regard, there are various deep learning architectures in literature [3], [8] which have been adopted for remote sensing image classifications [9], [10]. The deep learning feature representation strategies [11], [12] demonstrate impressive classification accuracies compared to handcrafted [9] and mid-level methods [10] in remote sensing image scene classification. Whereas the performance of deep neural networks is of significant improvements with high accuracy classification results on small datasets, this performance degrades with huge datasets that contain diverse image contents [9], [10]. Owing to the aforementioned observations, computer vision challenge is attributed to differences in image statistics such as viewpoints, scale, and semantics, among other factors [13]. Deep learning [14] provides a means for models that comprise multiple processing layers to learn feature representations of data with several levels of abstractions. Deep learning unearths complex structure in large datasets by using unique algorithms to depict how a machine should adjust its internal parameters that apply to compute the feature representations in every layer based on those of the previous layer. Indeed, the effectiveness of deep feature extraction is evident in recent literature [9]–[11] on remote sensing scene classification.

In machine learning, it is common for standard feature learning algorithms to exhibit performance variations on different datasets. This implies that the application of a particular algorithm can result in powerful classifiers with some databases; however, with the same classifiers trained, different datasets utilizing the identical algorithm may be unsteady. In remote sensing scene classification, a standard learning technique might not be capable to effectively learn specific features of the different scene classes. This is because the different class scene images contain very diverse semantics. The softmax [15] and support vector machines [16] are popular machine learning techniques which apply in remote sensing scene classification. Recent literature [17]–[19] shows that when multiple classifiers apply in feature learning, they attain improved classification accuracy. In their work [17], they utilize CNNLeNet-5 to extract deep features from digital handwritten images. Then, they apply multiple classifiers to learn the features of digit for multiclass classification problem.

This research proposes the adaptive deep co-occurrence feature learning technique based on classifier-fusion that learns scene image semantic features at different levels (layers) while considering the spatial-relative feature arrangements. The rest of this article is structured as follows. Section II provides a concise review of works related to this article. Section III presents the methodology of this article and the operation mechanism of the proposed method, while Section IV discusses the experiment setup, dataset, results, analysis, and discussions. Finally, Section VI concludes the article.

II. Related Work

This section reviews works in literature that are closely related to this article. First, this work reviews the literature in remote sensing image scene classification to highlight the developments, challenges, and opportunities in this area. Second, a critical analysis of computer vision methods in the remote sensing literature is given. To this end, the review is classed into five aspects, that is, the developments and challenges in remote sensing, conventional feature representation methods, deep learning and CNNs, feature learning through CNNs weights, and deep forests.

A. Remote Sensing Image Scene Classification

Remote sensing images are a valuable source of data that can be utilized to determine and visualize detailed information on the Earth's cover. The exponential increase of remote sensing images is due to improvements in satellite and sensor technologies [20], [21], and this has prompted the need for intelligent earth observations [22], [23]. The corresponding effects are improvements in remote sensing images quality spatial resolutions because of the sensor technology advances. With these gradual improvements, the recent literature [20] groups remote sensing image classification into three levels: 1) pixel-level, 2) object-level, and 3) scene-level. Here, the concept "remote sensing image classification" is general, encompassing all the three mentioned levels. Specifically, the initial literature [24], [25] majorly focused on human-engineered methods (pixel-level, also called semantic) to classify remote sensing images. Research is active in this area of semantic analysis for hyper-spectral and multi-spectral image analysis [26], [27]. The emergence deep learning is shifting the research efforts to scene-level classification, where CNNs apply for scene image feature extraction [9], [11], [28].

Remote sensing image scene classification aims to correctly annotate the remote sensing images based on their semantic contents, for instance, classifying a remote sensing image to agriculture, or airport or dense residential. Ideally, the remote sensing images comprise various objects from the ground; these may include buildings, trees, and roads on a residential scene. Scene classification of remote sensing images is a challenging problem due to their complex nature; that is, they are characterized by 1) high interclass similarity, 2) high intraclass diversity, 3) multiple-scale variances, and 4) coexistence of several ground objects, as depicted in Figs. 1 and 2. The driving force for remote sensing image scene classification is its broad application on real-world applications, such as vegetation mapping [29], [30], natural hazard detection [31], urban planning [32], [33], and environmental monitoring [34]–[36].

The challenging research problem in remote sensing is to develop computer vision techniques that can effectively apply to interpret and classify remote sensing images accurately. Elaborate researches have been conducted in remote sensing images scene classification; however, there is still no algorithm that attains satisfactory accuracy results.

B. Conventional Feature Representation Methods

The majority of recent scene classification techniques use the pipeline of bag-of-visual features [37]–[39] in encoding features. The bag of visual words (BOVWs) feature representation records the feature occurrences in the image, i.e., BOVWs $= [k_1, k_2, \ldots, k_T]$, where k_t is the number of feature occurrences. This is normally a histogram representation. The SIFT method [24] is quantized using the bag-of-visual feature through the k-means clustering algorithm. Spatial feature pooling [40], histogram feature encoding [39], and fisher vector feature encoding [41] are popular methods for feature assembly. Whereas these feature-representation techniques have been proven to work, it is not clear whether they are optimal for the tasks. This is a question of great interest in feature learning [42].

C. Deep Learning and Convolution Neural Networks

Deep learning is a multilayer feature learning and representation technique that transforms image data, i.e., pixels, to a feature vector that the system can detect and classify patterns. Deep learning models use nonlinear functions such as rectified linear units [43] for feature extraction in multiple levels [44], [46]. Deep learning initializes the network through parameter-tuning in a supervised version [47] where high-level abstract and invariant features in deep layers are learned from low-level features of the network lower layers. Examples of deep learning models in literature include deep belief networks [48] and CNNs [8]. CNNs are a type of deep learning strategy for image feature learning which applies in task classification. Generally, the CNNs apply in the following three ways on feature extraction in the context of remote sensing images.

1) Spectral feature extraction: In these CNN models, pixels are annotated to individual land-use-land-cover type [49]. CNNs use the raw image data to represent spectral feature directly as input feature vectors [50] to obtain a 1-D CNN architecture that receives (N) feature vectors as inputs, N being the number of spectral bands [47].
2) Spatial feature extraction: In this category, the CNN models use neighboring pixels of a given pixel in the original scene image to extract spatial features [50]. 2-D CNN architectures are applied for neighboring input data patch of dimensions $P \times P$ pixels [7]. Several methods are implemented to extract high-level spatial features [51], [52].

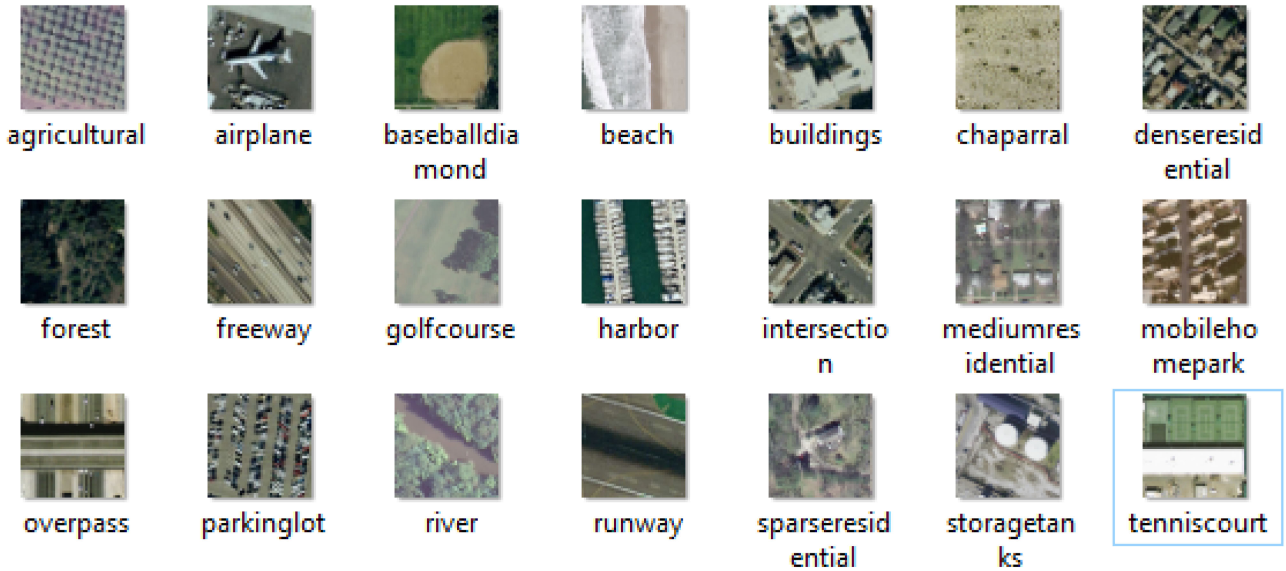

Fig. 1. Sample images of Ucmerced dataset [38].

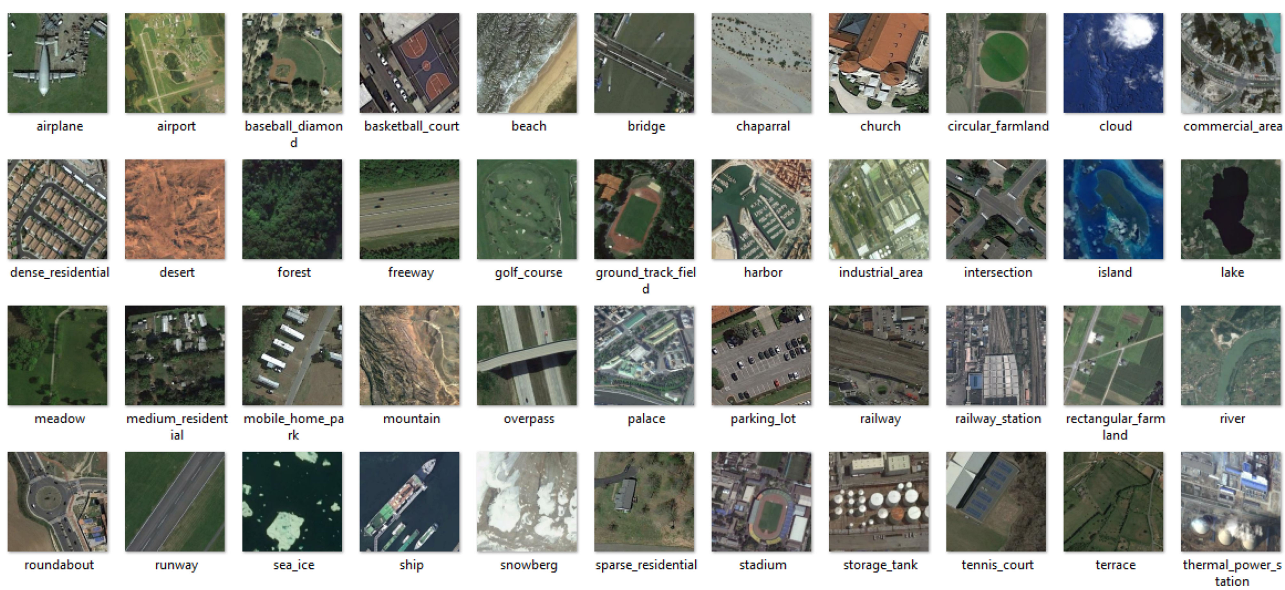

Fig. 2. Image samples from Resisc45 dataset [9].

3) Spatial–spectral feature extraction: This strategy entails a fusion of spectral and spatial features for improved classification accuracy [53].

The popular CNNs that are utilized in remote sensing include the following.

1) AlexNet: AlexNet [43] architecture won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. This network comprises five convolutional layers and three fully connected layers. Additionally, it has normalization layers after the first and second convolution layers. The pooling layers are put after the normalization layers and at the first convolutional layer. This network has been applied in remote sensing and it has demonstrated to achieve impressive results [9] in scene classification.

2) GoogLeNet: The GoogLeNet [8] architecture attained state of the art for object detection and classification tasks in the ILSVRC 2014. The main attribute of this architecture is the improved efficiency in the usage of computing resources within the network. The width and depth of the network are increased while maintaining the computational-budget constant. The main advantages of this network are as follows: 1) employ different filter sizes in the same layer; this keeps most of the spatial information, and 2) network parameter reduction, thus making it less prone to overfitting and permitting it to be deeper. Compared to AlexNet, GoogLeNet has 12 times fewer parameters.

3) VGGNet: The VGGNet [54] won in tracks of localization and classification with the ILSCVRC in 2014. VGGNet has two popular architectures, VGG-16 and VGG-19. The VGG-16 is common in the remote sensing literature. It comprises 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. The architecture commonly applies transfer learning in feature extraction of remote sensing imagery.

D. Feature Learning Through CNN Weights

A working CNN step comprises a convolution, pooling, and fully connected layers. A deep CNN is developed by stacking multiple convolution and pooling layers together to create deep architecture. The convolution layer is the first layer of the network. Neuron k_{lm}^{x} at *x* position of the *m*th feature map in layer *n*th is depicted by the following equation:

$$k_{lm}^{x} = g\left(b_{lm} + \sum_{q}\sum_{d=0}^{D_i-1} w_{lmq}^{d} k_{(l-1)q}^{x+d}\right) \quad (1)$$

where *q* is the feature-map index of the previous layer ($(l-1)$th) connecting the current feature map, w_{lmq}^{d} is the position *d* weight connecting the *q*th feature-map, D_i is the filter-kernel dimensions, and b_{lm} is the bias of *m*th feature map in the *n*th layer. The pooling layer reduces the feature-map resolution, thereby offering invariance [56]. Every pooling layer communicates with the previous convolution layer. The max-pooling operation is depicted in the following equation:

$$a_m = \max_{N\times 1}(a_l^{n\times 1} v(n,1)) \quad (2)$$

where a_m is the maximum value in the neuron neighborhood and $v(n,1)$ is a window function for the convolutional layer. The fully connected layers aggregate all the feature-map features generated by successful pooling layers to form robust feature representations usable for classification tasks by various machine learning algorithms.

E. Deep Forests

The deep forest [19] combines different classifiers to form a cascading training structure whereby each level obtains feature information in a cascaded manner through processing its previous level, and then the results are outputs to the following level. Every level represents an ensemble of classifiers. The classifier hyper-parameter is the number of trees in every forest. Each forest generates a prediction on the distribution of classes via probabilities of the different training classes at the leaf nodes where the involved sample falls, then an average for all the trees in the same forest is performed. The objective is to learn and determine the feature relationship from feature maps of different scene classes that can apply to categorize new features of unknown remote sensing scene images.

TABLE I
PARAMETERS OF VGG16 ARCHITECTURE

Blocks	Parameters
Block1	224×224 conv, 64
	224×224 conv, 64
	112×112 Max-pooling
Block2	112×112 conv, 128
	112×112 conv, 128
	56×56 Max-pooling,128
Block3	56×56 conv, 256
	56×56 conv, 256
	56×56 conv, 256
	28×28 Max-pooling, 512
Block4	28×28 conv, 512
	28×28 conv, 512
	28×28 conv, 512
	14×14 Max-pooling, 512
Block5	14×14 conv, 512
	14×14 conv, 512
	14×14 conv, 512
	7×7 Max-pooling, 512

III. METHODOLOGY

Consider a training set $X = \{\text{Image}_i, y_i\}_{i=1}^{n}$, where $\text{Image}_i \in R^x$ is a training instance and y_i is the image label $y \in Y$ representing scene class *C*; $Y = \{1, 2, \ldots, C\}$. To perform a remote sensing scene classification with test images $X' = \{(\text{image}'_t, y'_t)\}_{t=1}^{\infty}$, this research proposes a deep adaptive co-occurrence feature learning method for RS scene classification. The proposed strategy consists of two major steps which are depicted in Fig. 3: 1) spatial feature extraction with a pretrained convNet, and 2) ensemble learning that entails multilevel classifier-fusion on multigrain features [19] to learn co-occurrence deep features from the feature maps with sliding windows (SLWs). To train multiple classifiers, z learning algorithms apply in training primary classifiers on every feature set FeatureMap_i, thus creating a primary ensemble E_i. Eventually, the n primary ensembles that learn the n feature sets fuse via majority voting to make a classification prediction. The more discriminating feature information generated with CNN combined with effective feature learning with ensemble classifiers can lead to improved remote sensing scene classification accuracy.

A. Spatial Features Extraction

ConvNets are effective on spatial feature extraction from images [11], [43]. This work utilizes VGG16 [54], a pretrained CNN for feature extraction. Table I shows the VGG16 architecture and its parameters that include 13 convolutional layers and 5 pooling layers divided into 5 sections, and 3 fully connected layers. This work utilizes feature maps rather than fully connected layer features.

Assume $\text{conv}_l(\text{Image}_i) = \text{FeatureMaps}_i$, where $\text{FeatureMaps}_i \in^{h\times w\times d}$ are output feature maps of size $(h \times w \times d)$ obtained from the layer *l*th by conv_l of a pretrained convNet. For input FeatureMaps_i that characterizes the input image Image_i, the SLW with dimensions $(h_s \times w_s)$ slides on the FeatureMaps_i with *s* strides to generate feature samples α

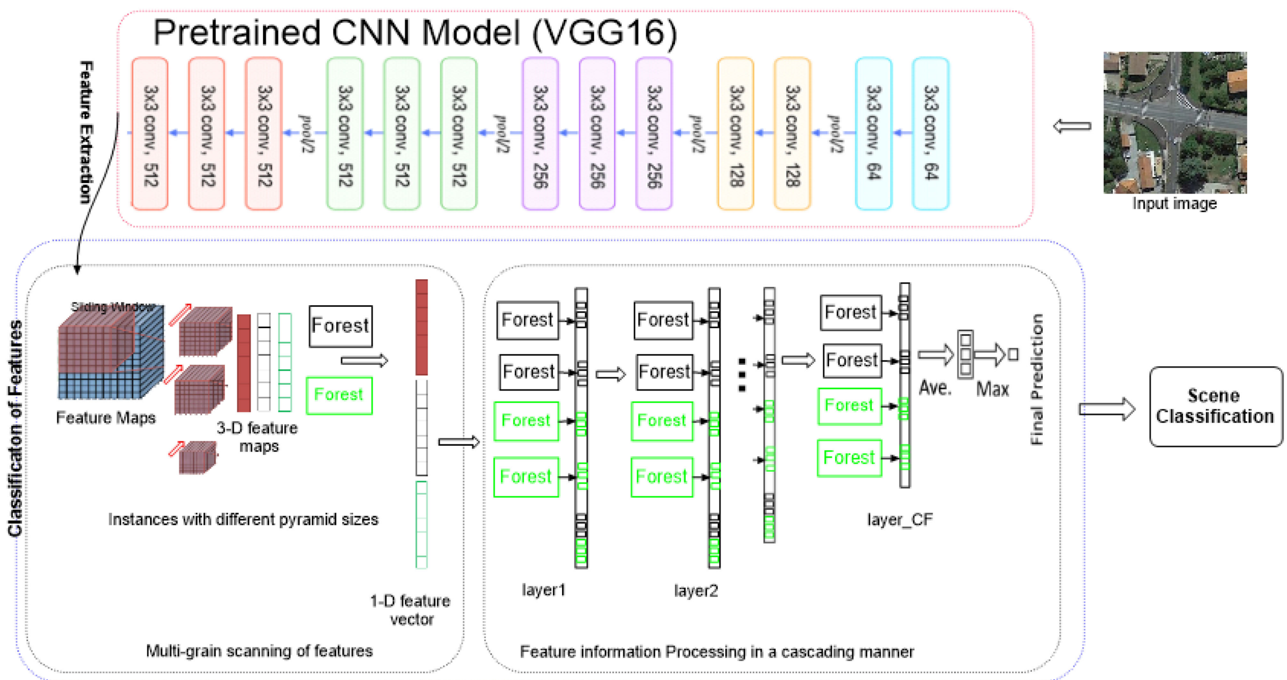

Fig. 3. Feature extraction with transfer learning and co-occurrence feature learning by multigrained forest classifiers for remote sensing scene classification.

that are of size SLW

$$\alpha = \left(\frac{h - w_s}{s} + 1\right) \times \left(\frac{w - w_s}{s} + 1\right). \quad (3)$$

B. Deep Co-occurrence Feature Learning With Multigrained Cascade Forests

Let $\mathbf{M}_{j,n} \in \mathbb{R}^{n_j \times n_j \times a_j}$ be a 3-D matrix, where jth is a feature map of an image Image_i, then, for every location $(u, v), j \leq u \leq n_j$ and $j \leq v \leq n_j, m_{1,j}^{u,v}$ forms an $\mathbf{a}_j$-dimensional feature representation for a local patch of Image_i. Following this process obtains $h_s \times w_s$ local feature maps of an image Image_i in conv_l layer l of size SLW.

Let m_1^r and m_2^r be two image feature patches satisfying a predicate condition in the visual words dictionary [40]. The deep features spatial-pyramid co-occurrence can be computed as per the following equation:

$$X(m_1^r, m_2^r) = \sum_{l=0}^{L} w_l \sum_{x}^{D} \sum_{u,v \in M} \min(m_1^p(x, u, v), m_2^u(y, u, v)) \quad (4)$$

where x is the relative arrangements of spatial feature patches. Combining predicates to characterize different spatial relationships, for instance, combining orientation and proximity predicates may represent the spatial distribution of features and the shape of local response regions. Remotely sensed images generally do not contain an absolute referencing frame; therefore, "relative spatial arrangements" of image contents which are key discriminating features are captured by the multigrained scanning windows (SLW). At this stage, the multigrain feature scanning transforms the 3-D feature maps into a 1-D feature vector (**FV**) representation.

If the number of forests used is F, and patches_k, $k \in [1, N\text{patches}]$, every forest f_p, $p \in [1, F]$ generates the class outputs that correspond to probability vectors (**PV**): $f_p(\alpha_d) = PV_k^p$, where $|PV_k^p| = C$. All class probabilities, $C = |\mathbf{PV}_k^p|$ generated with f forests and $N\alpha$ samples, are concatenated to form a final feature vector (**FV**) output [see (5)] of the multigrain features scan patches. A flowchart of the presented method is given in Fig. 3.

$$|\mathbf{FV}| = N\alpha \times f \times C. \quad (5)$$

The multigrained forest provides a means to process the extracted feature vectors layer-by-layer, and the final layer performs the scene label prediction using a majority vote. Every level (layer) of the multigrain forest has decision trees T. Consider the cascade forest layer L_q, $q \in [1, Q]$, where qth is cascade layer while Q is the number of layers in the multigrain forest. Every layer comprises Z forest classifiers, $F_z^q, z \in [1, Z]$. The feature patches FV_q outputs by the L_q^{th} are inputs to the following layer $q + 1$. For each tree $(t_z^q)_{ft}$, $ft \in [1, T]$, the forest classier z in this layer f_z^q obtains the $\sqrt{d}$ features vector that are selected randomly [19] from the previous layer $(q-1)_{ft}$ with class probability $(C_z^q)_{ft}$ outputs. Each forest in F every level/layer generates a class distribution (**CD**) vector by computing the average class probabilities which are estimated by their total trees [see (6)]

$$\mathbf{CD}_z^q = \text{average}\left[\sum_{ft}^{T}(\{C_z^q\}ft)\right]. \quad (6)$$

Then, aggregation of the different $\mathbf{CD}_z^q$ generated with forests F is performed using the original feature vector input. This gives the final layer output; the final layer gets all the class probability

Algorithm 1: Co-Occurrence Feature Learning.

Require: features(α), forestTrees
  while $(f <= F)$ **do**
    $\mathbf{PV} \leftarrow f_p(\alpha_d) = \mathbf{PV}_k^p$
    $C \leftarrow |\mathbf{PV}_k^p|$
    $|\mathbf{FV}| = N\alpha \times f \times C$
    for $(t < T)$ **do**
      $\sqrt{d} \leftarrow (t_z^q)_{ft} + f_z^q$
      $\mathbf{CD}_z^q \leftarrow$ average$[\sum_{ft}^{T}(\{C_z^q\}ft)$
    end for
    $\mathbf{CD}_{\text{final}} \leftarrow$ average$[\sum_{z=1}^{N\text{Forests}}(\mathbf{CD}_z^q)z]$
  end while
  $\hat{y} \leftarrow \text{argmax}\mathbf{CD}^q(y), \{y \in [1, C]$

vectors and averages them (7), and by the majority-voting (8), a prediction of scene class $\hat{y}$ is performed.

$$\mathbf{CD}_{\text{final}} = \text{average}\left[\sum_{z=1}^{N\text{Forests}} (\mathbf{CD}_z^q)z\right] \quad (7)$$

$$\hat{y} = \text{argmax}\mathbf{CD}^q(y), \left\{y \in [1, C]\right. . \quad (8)$$

IV. Dataset Description and Experimental Setup

This section discusses the dataset, tools, and the experimental setups that apply in this research. Further, this section presents the result, analysis, and discussions.

A. Ucmerced Dataset

Ucmerced dataset [38] consist of 21 classes as shown in Fig. 1 and each class contains 100 images with three color channels. Each image dimension is 256 $\times$ 256 pixels and they have a spatial resolution of 1 ft. The classes are highly overlapped (e.g., agricultural and forest differ by vegetation cover; dense residence and medium residence differ by the number of units); this diverse image content pattern is a challenge for effective feature representations. Further, the images of Ucmerced dataset have many common low-level features with multipurpose visible images; hence, they are suitable candidates for fine-tuning with pretrained CNNs.

B. Resisc45 Dataset

The Resisc45 dataset [9] contains 31 500 scene images that are grouped to 45 classes; every class comprises 700 images with dimensions 256 $\times$ 256 pixels in three channels color space. Spatial resolutions of the images range approximately between 30 and 0.2 m per pixel. Image samples of RESISC45 dataset are shown in Fig. 2. The images in RESISC45 dataset are selected under varying conditions including different weather and seasons, various illuminations, and are varying resolutions and scales. Therefore, there are rich variations in object pose, translation, and appearance, viewpoint, illumination, occlusions, and background in this RESISC45 dataset. This dataset is more challenging, requiring innovative and sophisticated image feature analysis and representation mechanisms for effective feature characterization.

TABLE II
Overall Accuracy (OA%) Classification Performances on Resisc45 Dataset

Feature learning method	Resisc45 OA%
VGG- 16 with XGBoost[18]	83.37
VGG-16 with Bag of convolutional features [37]	84.32
VGG-16	82.12
VGG16-Classifier-fusion(proposed)	91.05

C. Experimental Setups

In the experiment study, the entire process entails feature extraction, selection of features, multilevel training, and classifier-fusion. For feature extraction, a pretrained VGG-16 is used with input images of size 224 $\times$ 224. For purposes evaluating the different classifiers, implementation strategies effectiveness in feature learning, two implementation strategies are adopted with VGG16 under different settings; i.e., 1) the multilevel fusion of classifiers for feature learning and RS scene classification; 2) We fine-tune the VGG16 with remote sensing datasets in Sections IV-A and IV-B and then apply the softmax classifier for RS scene classification. The experimental results for both strategies are reported. The experiments are implemented with python 3.7.5 and Keras on the googlecobab-GPU. For fair comparison on the classification results, the parameter settings for both experiments are the same, that is, the training, validation, and testing ratios are set to 70%:20%:10% on the Ucmerced dataset and 15%:80%:5% on Resisc45 dataset.

In the first implementation strategy, this research utilizes VGG-16 5-3 feature maps of size 14 $\times$ 14 (Table I). These features are then fed to multigrained forests for learning. The cascade forests of the deep forest are adaptively (automatically) established in the course of training, utilizing the early stopping strategy. As in [19], every cascade layer uses two forests (complete random forests and random forests); this increases the model balance between variance and bias. The classifier-fusion is accomplished by averaging their inner outputs (probabilities of every class), that is, mean algebraic fusion [55].

For the second implementation strategy, transfer learning of features with remote sensing datasets is conducted with 30 epochs in batches of 32, and then followed by a fine-tuning phase with the same settings. The learning rates and weight decays are the same as [9], that is, 0.001 and 0.0005, respectively.

V. Results, Analysis, and Discussion

To evaluate the classification performance for the two datasets (Resisc45 and Ucmerced), overall accuracy (OA) [9] is computed as per (9) and the results are given in Tables II and III. In this research, for the initial experiments, the VGG-16 is fine-tuned to extract features from the Resisc45 and Ucmerced datasets; then application of the softmax function for scene classification of remote sensing images. Figs. 4 and 5 show the number of epochs versus train accuracy on Resisc45 and Ucmerced datasets with the fine-turned VGG-16. Fig. 8 provides

TABLE III
OVERALL ACCURACY (OA%) CLASSIFICATION PERFORMANCES ON UCMERCED DATASET

Feature learning method	Ucmerced OA%
VGG- 16 with XGBoost[18]	95.57
VGG-16 [28]	97.10
Adaptive deep pyramid matching Method [57]	94.92
VGG-16	96.92
VGG16-Classifier-fusion(proposed)	96.55

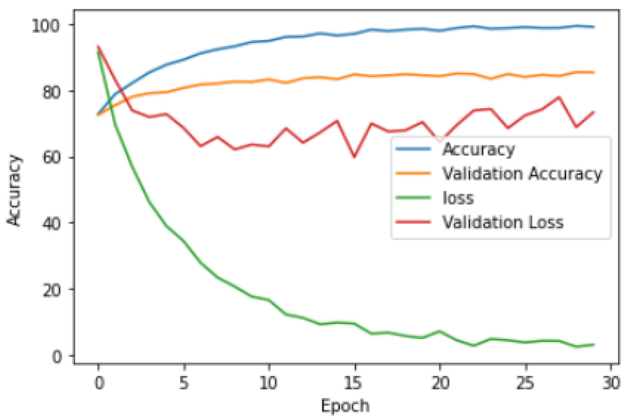

Fig. 4. Epochs versus accuracy of fine-tuned VGG-16 on Resisc45 dataset.

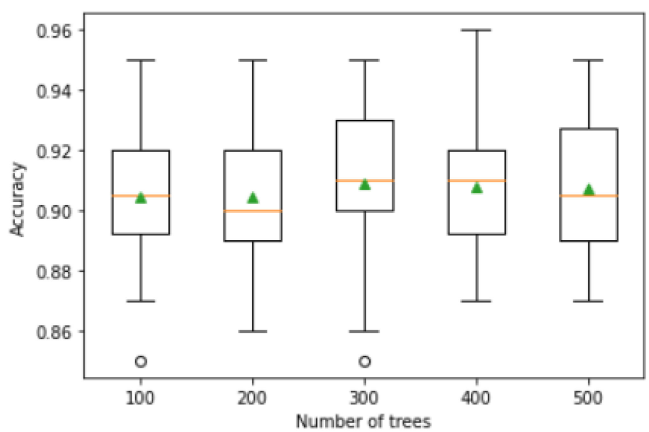

Fig. 5. Epochs versus accuracy of fine-tuned VGG-16 on Ucmerced dataset.

the confusion matrix that shows the predictions versus true label results with test images of the Ucmerced dataset with the fine-turned VGG-16. For ensemble learning, that is fusion of classifiers. Figs. 6 and 7 show the number of trees versus the training accuracy with Resisc45 and Ucmerced datasets

$$\text{OA} = \frac{\text{Correctly Classified Images}}{\text{Sampled Images}} \times 100. \tag{9}$$

It can be observed from Tables II and III that the fine-tuned VGG-16 in our experiments achieves more or less the same results with those attained by other works in the literature. This, therefore, sets a definitive benchmark in demonstrating that classifier-fusion achieves better classification results with remote sensing datasets. Comparing performance of the adaptive deep co-accordance feature learning (ADCFL) and the softmax

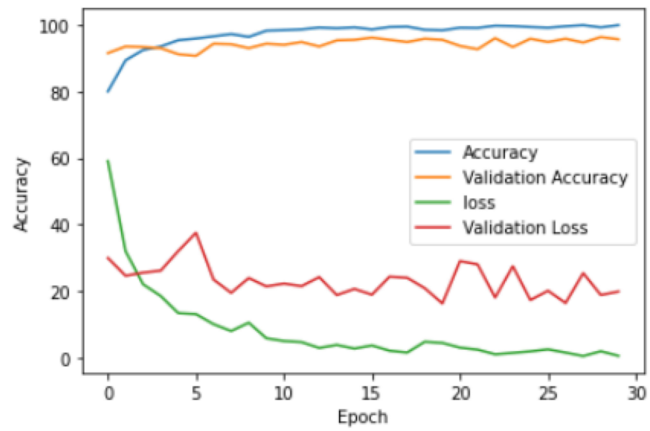

Fig. 6. Ensemble trees versus classification accuracy on Resisc45 dataset.

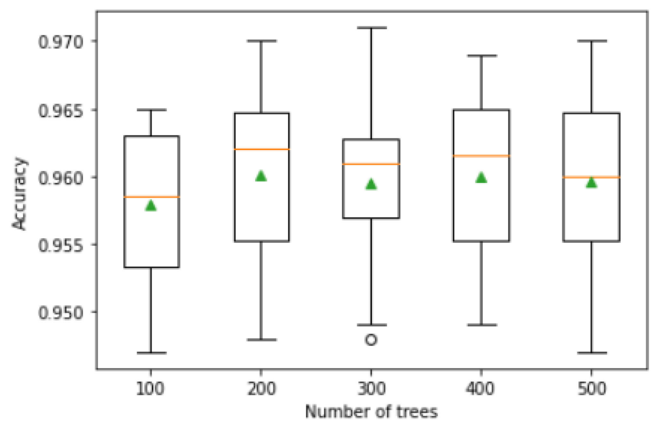

Fig. 7. Ensemble trees versus classification accuracy on Ucmerced dataset.

classifiers on two datasets (Resisc45 and Ucmerced) from Tables II and III , it can be observed that there is a significant improvement on the OA on Resisc45 dataset with the ADCFL. This implies that the application of a particular algorithm can result in powerful classifiers with some databases; however, the classifiers trained with more diverse datasets utilizing the identical algorithm may be unsteady. In remote sensing scene classification, a standard learning technique might not be capable to effectively learn all specific features of the different scene classes on more diverse datasets which contain high semantics variations. For pattern recognition problems from feature maps with machine learning, it is common for standard feature learning algorithms to exhibit performance variations on different datasets [17], [18]. This is evident from Fig. 8; for instance, there are confusions between the classes medium-residential and dense residence, resulting in low prediction results of 0.5. This research fuses complete random forest and random forest classifiers in multigrained feature learning [19] and as the experimental results demonstrate, the proposed method attains superior classification as compared to those of a single classifier. Furthermore, the adaptive deep co-occurrence feature learning method demonstrates superiority in terms of classification accuracy compared to the other methods in literature as summarized in Table II.

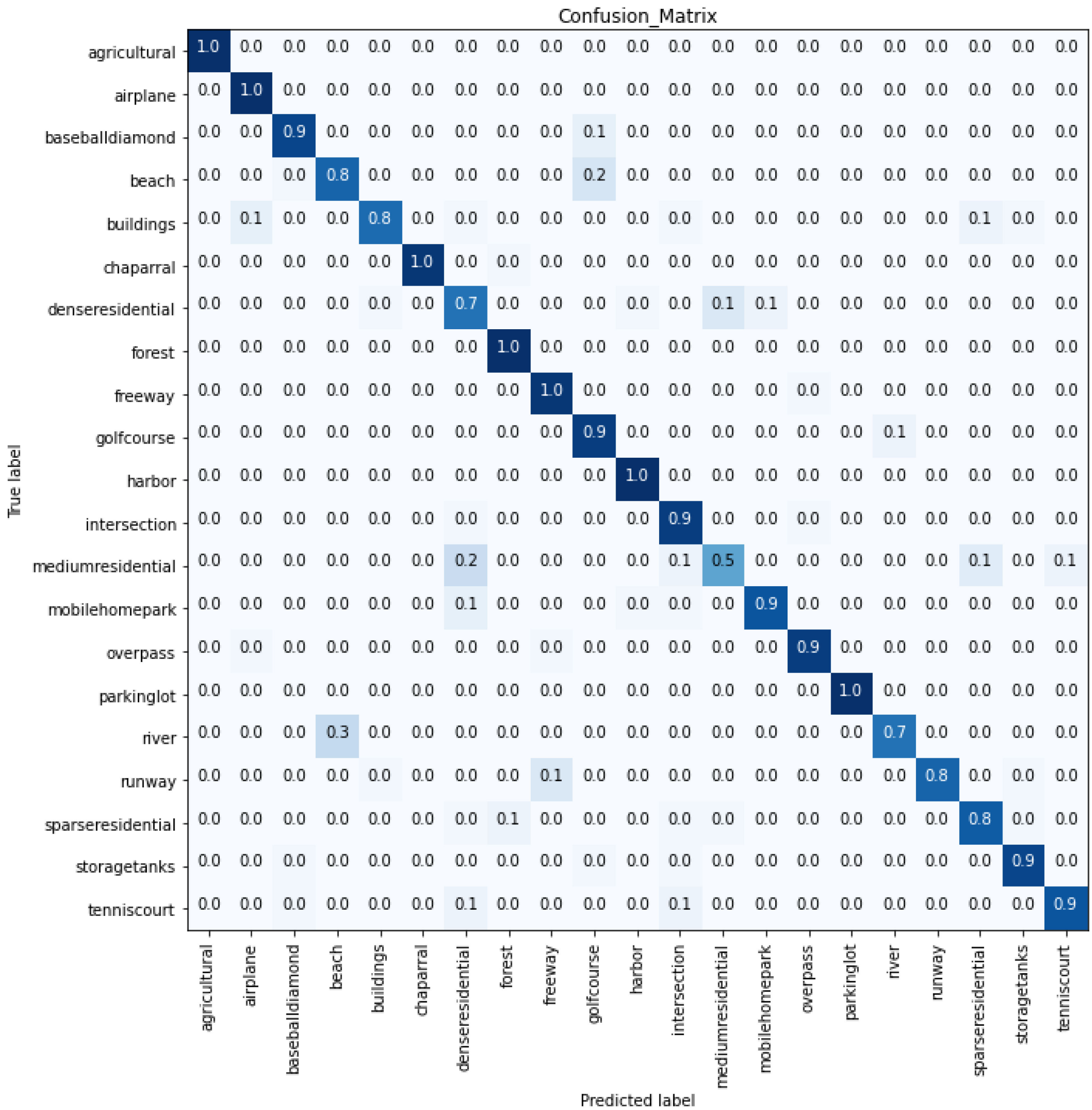

Fig. 8. Confusion matrix for Ucmerced dataset.

VI. Conclusion

This article proposes the ADCFL based on classifier-fusion for remote sensing scene classification. Specifically, this research utilizes the VGG-16 for spatial feature co-occurrence learning. These features are then fed to a deep multigrain (classifier-fusion) for feature learning and classification. To establish superiority of the proposed method, this research utilizes two different machine learning implementation approaches. The experimental results demonstrate that the classifier-fusion strategy attains superiority for remote sensing scene classification.

The future research investigation will investigate strategies for optimal classifier-fusion with different pretrained CNN features for remote sensing scene classification.

References

[1] R. Mehta, and K. Egiazarian, "Dominant rotated local binary patterns (DRLBP) for texture classification," *Pattern Recognit. Lett.*, vol. 71, pp. 16–22, 2016.

[2] O. A. Penatti, K. Nogueira, and J. A. Santos Dos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2016, pp. 2921–2929.

[4] K. S. Willis, "Remote sensing change detection for ecological monitoring in United States protected areas," *Biol. Conserv.*, vol. 182, pp. 233–242, 2015.

[5] W. Ji, J. Ma, R. W. Twibell, and K. Underhill, "Characterizing urban sprawl using multi-stage remote sensing images and landscape metrics," *Comput., Environ. Urban Syst.*, vol. 30, no. 6, pp. 861–879, 2006.

[6] M. Wojtowicz, A. Wójtowicz, and J. Piekarczyk, "Application of remote sensing methods in agriculture," *Commun. Biometry Crop Sci.*, vol. 11, no. 1, pp. 31–50, 2016.

[7] A. Vetrivel, M. Gerke, N. Kerle, F. Nex, and G. Vosselman, "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 45–59, 2018.

[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2015, pp. 1–9.

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

[10] G. S. H. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[11] Y. Bazi, Al Rahhal, M. M. H. Alhichri, and N. Alajlan, "Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2908.

[12] 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 Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[13] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. 2, Jun. 2011, Art. no. 7.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] C. M. Bishop, *Pattern Recognition Machine Learning.* Berlin, Germany: Springer, 2006.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[17] H. H. Zhao and H. Liu, "Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition," *Granular Comput.*, vol. 5, no. 3, pp. 411–418, 2020.

[18] S. Jiang, H. Zhao, W. Wu, and Q. Tan, "A novel framework for remote sensing image scene classification," *Int. Archives Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 3, pp. 657–663, 2018.

[19] Z. H. Zhou and J. Feng, "Deep forest," *Nat. Sci. Rev.*, vol. 6, no. 1, pp. 74–86, 2019.

[20] 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 J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020.

[21] Q. Hu *et al.*, "Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.

[22] D. Li, M. Wang, Z. Dong, X. Shen, and L. Shi, "Earth observation brain (EOB): An intelligent earth observation system," *Geo-Spatial Inf. Sci.*, vol. 20, no. 2, pp. 134–140, 2017.

[23] P. Gamba, "Human settlements: A global challenge for EO data processing and interpretation," *Proc. IEEE*, vol. 101, no. 3, pp. 570–581, Mar. 2013.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[25] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[26] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[27] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[28] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[29] N. B. Mishra and K. A. Crews, "Mapping vegetation morphology types in a dry savanna ecosystem: Integrating hierarchical object-based image analysis with random forest," *Int. J. Remote Sens.*, vol. 35, no. 3, pp. 1175–1198, 2014.

[30] X. Li and G. Shao, "Object-based urban vegetation mapping with high-resolution aerial photography as a single data source," *Int. J. Remote Sens.*, vol. 34., no. 3, pp. 771–789, 2013.

[31] Z. Y. Lv, W. Shi, X. Zhang, and J. A. Benediktsson, "Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1520–1532, May 2018.

[32] N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici, "Very high resolution multiangle urban classification analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1155–1170, Apr. 2012.

[33] A. Tayyebi, B. C Pijanowski, and A. H. Tayyebi, "An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran," *Iran. Landscape Urban Planning*, vol. 100, no. 1/2, pp. 35–44, 2011.

[34] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sens. Environ.*, vol. 196, pp. 56–75, 2017.

[35] T. Zhang and X. Huang, "Monitoring of urban impervious surfaces using time series of high-resolution remote sensing images in rapidly urbanized areas: A case study of Shenzhen," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2692–2708, Aug. 2018.

[36] F. Ghazouani, I. R. Farah, and B. Solaiman, "A multi-level semantic scene interpretation strategy for change interpretation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8775–8795, Nov. 2019.

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

[38] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th Sigspatial Int. Conf. Adv. Geographic Inf. Syst.*, ACM, Nov. 2010, pp. 270–279.

[39] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis.*, vol. 1, no. 1–22, May. 2004, pp. 1–2.

[40] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int.Conf. Comput. Vis.*, Nov. 2011, pp. 1465–1472.

[41] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[42] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 923–930.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1097–1105.

[44] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[45] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.

[46] Y. Bengio, "Learning deep architectures for AI," *Foundations Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[47] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[48] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[49] P. Fisher, "The pixel: A snare and a delusion," *Int. J. Remote Sens.*, vol. 18, no. 3, pp. 679–685, 1997.

[50] M. He, X. Li, Y. Zhang, J. Zhang, and W. Wang, "Hyperspectral image classification based on deep stacking network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 3286–3289.

[51] Q. Wu *et al.*, "Shape-based object extraction in high-resolution remote-sensing images using deep Boltzmann machine," *Int. J. Remote Sens.*, vol. 37, no. 24, pp. 6012–6022, 2016.

[52] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, 2016.

[53] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.

[56] Z. Zuo *et al.*, "Learning contextual dependence with convolutional hierarchical recurrent neural networks," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 2983–2996, Jul. 2016.

[57] Q. Liu, R. Hang, H. Song, F. Zhu, J. Plaza, and A. Plaza, "Adaptive deep pyramid matching for remote sensing scene classification," 2016, *arXiv:1611.03589*.

Ronald Tombe received the B.Sc. degree in information technology and the M.Sc. degree in software engineering and the B.Sc. degree in information technology from Jomo Kenyatta University of Agriculture and Technology (JKUAT), Juja, Kenya, in 2009 and 2015, respectively. He is currently working toward the Ph.D. degree in computer science at the Department of Computer Science, University of KwaZulu-Natal (UKZN), Durban, South Africa (2018–2020).

His research interests include computer vision, pattern recognition, image processing, remote sensing, and deep learning. He was/is a Lecturer with Kisii University, Kisii, Kenya, between 2014 and 2018, before taking a study leave to pursue the Ph.D. degree at University of KwaZulu-Natal. Between 2010 and 2013, he was a Teaching Assistant with Mount Kenya University, Thika, Kenya. Furthermore, he has presented parts of his research work on science computer and artificial intelligence at international conferences.

Mr. Tombe was a first place winner on a computer vision segmentation and classification problem during a "Deep Learning" workshop organized by South African Universities Scholars and Industry (April 14–17, 2019).

Serestina Viriri (Senior Member, IEEE) received the B.Sc. degree in mathematics and computer science, in 1998, and M.Sc. degree in computer science, in 2002, both from the University of Havana, and Ph.D. degree in computer science from the University of KwaZulu-Natal, in 2010.

He is a Full Professor of Computer Science with the University of KwaZulu-Natal, Durban, South Africa. He is currently the HoD of Computer Science with the School of Mathematics, Statistics, and Computer Science. He has been in academia for more than 20 years. His main research interests include computer vision, image processing, machine learning, pattern recognition, and other image processing related fields, such as biometrics, medical imaging, and nuclear medicine. He has authored or coauthored extensively in several computer vision related accredited journals and international and national conference proceedings. He has supervised the completion of several Ph.D. and M.Sc. students.

Prof. Viriri is a Reviewer for several computer vision related journals. He has also served on program committees for numerous international and national conferences. He is an NRF Rated Researcher.