A Remote Sensing Scene Classification Model Based on EfficientNet-V2L Deep Neural Networks

Atif A. Aljabri[1], Abdullah Alshanqiti[1], Ahmad B. Alkhodre[1], Ayyub Alzahem[2], and Ahmed Hagag[3,*]
421001356@stu.iu.edu.sa , amma@iu.edu.sa , aalkhodre@iu.edu.sa , aalzahem@psu.edu.sa , ahagag@fci.bu.edu.eg
[1]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia.
[2]Robotics and Internet-of-Things Lab, Prince Sultan University.
[3]Faculty of Computers and Artificial Intelligence, Benha University, Benha, 13518, Egypt.

Summary
Scene classification of very high-resolution (VHR) imagery can attribute semantics to land cover in a variety of domains. Real-world application requirements have not been addressed by conventional techniques for remote sensing image classification. Recent research has demonstrated that deep convolutional neural networks (CNNs) are effective at extracting features due to their strong feature extraction capabilities. In order to improve classification performance, these approaches rely primarily on semantic information. Since the abstract and global semantic information makes it difficult for the network to correctly classify scene images with similar structures and high interclass similarity, it achieves a low classification accuracy. We propose a VHR remote sensing image classification model that uses extracts the global feature from the original VHR image using an EfficientNet-V2L CNN pre-trained to detect similar classes. The image is then classified using a multilayer perceptron (MLP). This method was evaluated using two benchmark remote sensing datasets: the 21-class UC Merced, and the 38-class PatternNet. As compared to other state-of-the-art models, the proposed model significantly improves performance.

Keywords:
VHR, Remote sensing, scene classification, Deep learning, EfficientNet.

1. Introduction

Our ability to obtain images with very high resolution (VHR) has advanced with the advancement of remote sensing technologies [1]. We can now measure the Earth's surface in detail with VHR images, which are a valuable data source for Earth observation. Additionally, VHR images can provide accurate insights into man-made infrastructures, like streets, because they allow for their depiction.VHR images have been included in several datasets, including the UC Merced dataset [2], which consists of 21 different scene classes each with 100 images; the PatternNet dataset [3] containing 38 classes with 800 images per class; and the WHU-RS19 dataset [4, 5], which contains 950 images arranged into 19 categories.While low-level attributes such as spectral, textural, and geometrical characteristics can be used in scene classification, they rarely yield satisfactory results. Several algorithms, including machine learning and data-driven approaches, have been proposed in the field of remote sensing in recent years. The accuracy of classification has been improved by researchers using advanced classification approaches. It has been found that there are three types of VHR imagery [6]: low level, middle level, and high level. A low-level method [7-11] extracts information from local or global locations to design artificial features, such as colors. As a result, this method performs poorly in terms of classification [6]. In order to obtain statistical representations, scientists have designed middle-level methods [12-15] , such as the improved Fisher vector. The accuracy of classification can also be improved by using deep learning-based methods [16-21].

The field of computer vision has recently proven the benefits of deep learning. Image categorization and object recognition are greatly improved using convolutional neural networks (CNNs), such as AlexNet [22], VGGNet [23], Inception Net [24], and ResNet [25]. By training CNN-based frameworks, high-level discriminative features can be automatically extracted, which are commonly used in the past. Meanwhile, CNN-based algorithms have been employed in remote sensing and shown to be effective. The multilevel improved circle pooling (MICP) method was proposed by Kunlun et al. [26] to improve the discriminative power of CNN activations. A CNN-based classification algorithm based on multilayer perceptrons (MLPs) was proposed by Osama et al. [27]. In this case, the features are created using a pre-trained CNN without fully connected layers. Due to the limited number of training images available in each class, this method employs data augmentation techniques to increase the number of training images available. MLPs were used to classify the resulting feature maps. BRBM stands for best representation branch model, created by Zhang et al. [28]. In order to obtain the final classification accuracy, the BRBM uses a classifier using CapsNet that extracts feature maps using ResNet50. On both benchmarks of remote sensing datasets, the BRBM achieves good results when compared with state-of-the-art methods.

Deep learning models are designed to increase classification accuracy at the expense of speed. Although scene classification could be made more accurate, and the time complexity should be reduced, both need to be

Manuscript received October 5, 2022
Manuscript revised October 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.10.53

improved. Hence, we developed a deep transfer learning method for classifying VHR images. The main contributions of this study are as follows:

- Generate robust features for a VHR image by utilizing EfficientNetV2L.
- Our framework uses the extracted features to determine the classification of scenes from remote sensing. In addition, we used the Adagrad optimizer to classify the VHR image.
- Two public benchmark VHR datasets were used to demonstrate the effectiveness of the proposed model. The proposed model achieves better classification accuracy than other related models after extensive experiments.

In the remainder of this paper, we organize our findings as follows. In section 2, we discuss related work on classifying VHR images. In section 3, each component of the proposed deep-learning-based classification model is explained in detail. In Section 4, the dataset used, the experimental setup, and the results are summarized. Section 5 concludes with concluding remarks.

2. Related Works

Satellite images have been the subject of primary research in the literature. In terms of developed approaches, the following three groups can be distinguished: those based on low-level features [7-11] (scale invariant feature transform (SIFT), local binary pattern (LBP)), those based on mid-level visual representations [12-15] (e.g., bag-of-visual-words (BoVW) and extreme value theory (EVT)-based normalization), and those relying on high-level vision information [16-21] (e.g., deep learning methods).

Deep-learning-based techniques have been developed for a number of applications in the last decade, including satellite image classification. The development of SRSCNNs was motivated by the drawbacks of BoVWs and CNNs [29]. The DLGFF framework provides a unique end-to-end CNN by combining global context features (GCFs) with local object features (LOFs) for VHR imaging scene classification [30]. For object-based categorization, the proposed network has two branches: the local object branch (LOB) and the global semantic branch (GSB) [31]. Based on the preceding thematic maps and satellite images of the study regions, feature weight maps were derived for each terrain type [32].

Recently, a method for analyzing VHR scenes based on saliency features was proposed in [33]. Global CNNs and saliency features are confused by this method. In the subsequent step, an enhanced MLP classifier was used. [34] Alhichri et al. proposed a deep learning model to classify remote-sensing scenes. The effectiveness of their experience was determined by a combination of an efficient Net CNN and an attention mechanism. Two versions of their model have been tested: EfficientNet-B3-Attn-1 and EfficientNet-B3-Attn-2.

In EfficientNet-B3-Attn-1, the attention mechanism was added to the last feature map. However, in EfficientNet-B3-Attn-2, the attention model was added at the end of layer 262. According to Peng et al. [35], an efficient architecture search framework was proposed. Due to limitations in the extracted features from pre-trained CNN models, as well as data deficiencies from the extracted scene images, they also investigated a new paradigm that automatically designed a suitable CNN architecture. An image classification algorithm based on deep learning was developed by Xiaowei et al. [36]. Optimising cross-validation was improved using recurrent neural networks (RNNs) and random forests. The researchers demonstrated that training neural networks can be improved by using multiscale views. Testing and training costs can be reduced while achieving optimal results with a random forest. A new CNN based on the Siamese network was proposed by Tang et al. [37].

3. Methodology

In this study, a remote sensing scene classification model based on deep CNNs is proposed, which includes four main tasks. The framework of our classification model is illustrated in Fig. 1. First, data pre-processing is performed. Second, the pre-trained model EfficientNetV2L was used to extract global features from the original VHR image. Finally, to classify the VHR image, the proposed model used an MLP classifier based on the Adagrad optimizer. Each step of the proposed model is described in the following subsections.

3.1 Pre-processing

Some models use images with values ranging from 0 to 1. Others from -1 to +1. Therefore, in the proposed model, data pre-processing is performed as a part of the model (i.e., Rescaling layer).

3.2 Feature Extraction

In the proposed model, an EfficientNetV2L pre-trained is used to extract features from the original VHR image. According to [38], CNN models are characterized by their width and depth. Using this approach, CNN models could be designed with fewer parameters and achieve better classification accuracy. Their original paper proposed seven such models, called EfficientNetB0 to EfficientNetB7, called EfficientNetV1 CNN models. Scaling up CNN models is the basis of the EfficientNetV1 family.

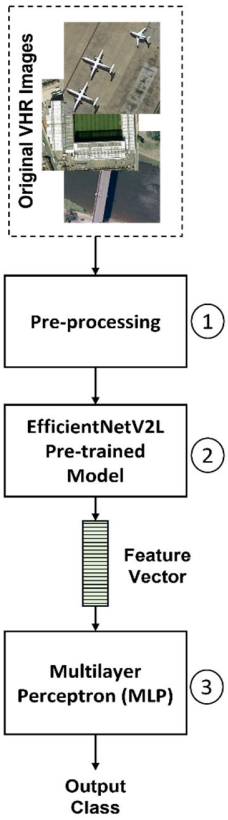

Fig. 1. A framework of the proposed model for VHR imagery scene classification.

The compound coefficient used in this method is simple and highly effective. In contrast to traditional methods, EfficientNetV1 uniformly scales each dimension based on a fixed set of scaling coefficients. The performance of a model improves when each dimension is scaled. When all network dimensions are balanced according to the available resources, overall performance is improved. Compared to EfficientNetV1, EfficientNetV2 [39] provides faster training speeds and more efficient parameters. By combining training-aware neural architecture search with scaling, the authors optimized training speed at the same time. Fused-MBConv was added to the search space to enhance the search process. EfficientNetV2 extensively utilized the main architectural differences between MBConv and fused-MBConv in its early layers. EfficientNetV2 prefers smaller expansion ratios for MBConv since smaller expansion ratios result in fewer memory access overheads. The smaller kernel size in EfficientNetV2 results in a smaller receptive field, so more layers are added to compensate. EfficientNetV2 lacks the last stride-1 stage from EfficientNetV1 due to its large parameter size and memory access overhead. Based on [38], EfficientNetV2L employs compound scaling similar to [38], with some additional optimizations, as shown in [39]: (1) the inference image size is 480, as large images often result in high memory and training speed overhead; (2) to increase network capacity without increasing runtime overhead, additional layers were added to stages 5 and 6 of the EfficientNetV1 architecture [38].

3.3 Multilayer Perceptron (MLP)

Multilayer perceptrons (MLPs), commonly known as feed-forward ANNs, are the most widely used ANN architectures. There are at least three layers: an input layer, a hidden layer, and an output layer. Perceptrons are modeled after brain neuron cells and are an essential component of MLP design. Perceptrons receive inputs from previous layers and transfer outputs to next layers after performing certain mathematical operations.

There are four layers in the proposed MLP design. First, all image features were normalized through a normalization layer. A normalization process ensures that the data distribution for each pixel is uniform. During network training, it converges rapidly. Second, 50 units of density were used. Third, we used a dropout layer, as described in [27]. To reduce overfitting, neural networks employ dropout regularization to avoid complicated co-adaptations to training data. Standard layouts can be performed using neural networks using this method. Fourth, we used a logistic regression algorithm (SoftMax) to normalize input values into a vector of value vectors representing classes in the VHR dataset that follow a probability distribution. By incorporating prior observations, the learning rate was modified using the Adagrad algorithm [40]. Adagrad optimizers are suitable for sparse data because they adapt a larger learning rate update for infrequent parameters and a smaller update for frequent parameters. In addition, the Adagrad optimizer takes considerably less time to predict than other optimizers (e.g., Adadelta, RMSprop, Adam, Nadam, and SGD).

4. Experiments Results and Discussion

Several experiments were conducted on two public remote sensing image datasets to assess the performance of the classification models. The first part of this section describes the datasets. The experimental setup and performance measures are presented in the second and third parts. In the fourth part of this section, we describe our experiments and discuss the results for each VHR dataset. Finally, we discuss the ablation study of the proposed model.

4.1 Description of Datasets

Two benchmark datasets for land use scene classification were used: UC Merced [2], and PatternNet [3]. The details of these datasets are listed in Table 1. Fig. 2 shows the samples of the selected VHR datasets. All two datasets were used for benchmark comparison.

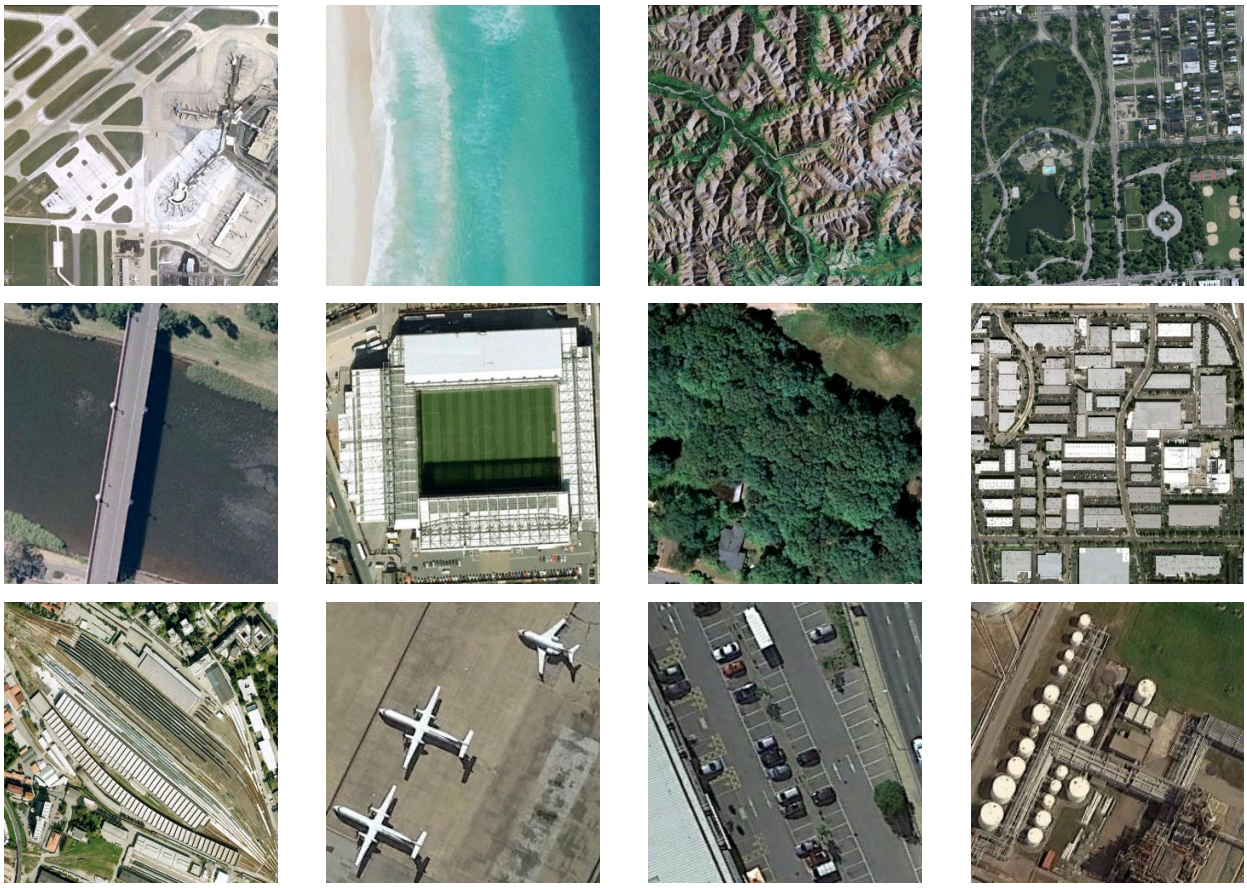

Fig. 2. Example of VHR images.

Table 1: Key information on the benchmark VHR scenes.

Dataset	*Images*	*Classes*	*Images per class*	*Image size*	*Spatial resolution (m)*
UC Merced	2,100	21	100	256 × 256	0.3
PatternNet	30,400	38	800	256 × 256	0.062~4.693

4.2 Experimental Setup

In this study, the features of the images with pre-trained CNNs (i.e., EfficientNetV2L) were extracted from the original VHR images. All the images were resized to 480×480 pixels. The output global feature vector size was 1280. After that, the MLP classifier was used for the classification step. The learning rate was 0.1. A dropout of 0.4 is utilized in the proposed model. The proposed model used 100 epochs for the UC Merced, WHU-RS19, and PatternNet datasets. The batch size is 32~64. The experiment was repeated five times, and the average values were tabulated. The software environment used to conduct the proposed model was Python 3. All experiments were conducted on the Linux-based Google Colab (an online browser-based platform) with a Pro subscription. The platform provides 25 GB of RAM and a Tesla K80 12GB GPU as an accelerator.

4.3 Experimental Results

In the following subsections, two public VHR datasets are used in in-depth experiments to test the efficacy of the proposed model against the most advanced scene classification models under identical settings and training ratios. The same experimental setup was used five times, and the mean of the five outcomes was computed to evaluate the effectiveness of the proposed model. Various evaluation metrics, including confusion matrix, accuracy, precision, recall, and F1-Score, can be used to assess the performance of the proposed classification method. The

experimental findings show that the proposed model outperforms other state-of-the-art models.

4.3.1 Classification results of the UC Merced dataset

In Table 2, the performance of the proposed model on the UC Merced dataset was quantitatively measured using commonly used parameters, such as overall OA, precision, recall, Cohen's kappa coefficient (κ), MCC, ROC, and F1-score. An overall accuracy (OA) comparison between the proposed model and state-of-the-art methods on the UC Merced dataset is reported in Table 3 to evaluate the effectiveness of our method. With training ratios (TR) of 50% and 80%, the OAs of our model reach 97.38% and 96.87%, respectively, exceeding those of all the methods listed in the table. Compared with the related methods reported in Table 3, the OAs of our model are approximately 1.12% and 0.71% higher than those of STHP [41], with a TR of 50% and 80%, respectively. For method using GoogLeNet [6] with a TR of 50% and 80%, the OA was 92.70% and 94.31%, respectively. The results show that the proposed model obtains better classification accuracy than the other related models.

Table 2: Evaluation of the proposed VHR scene classification model on the UC Merced with TR = 80%.

Metric	***Results***
OA	0.9738
Precision	0.9751
Recall	0.9750
Kappa (κ)	0.9724
MCC	0.9725
ROC	0.9869
F1-Score	0.9745

Table 3: Comparison of overall accuracy (%) on the UC Merced dataset.

Method	***Year***	***Training Ratio (TR)***	
		50%	***80%***
CaffeNet [6]	2017	93.98	95.02
VGG-VD-16 [6]	2017	94.14	95.21
GoogLeNet [6]	2017	92.70	94.31
SalM3LBP-CLM [42]	2017	94.21	95.75
PMWMFF [43]	2021	-	97.14
SAFF [44]	2021	-	97.02
LCPB [45]	2021	-	96.66
AlexNet+MICP [26]	2021	-	96.13
GAN [46]	2022	93.22	96.10
STHP [41]	2022	95.75	96.67
Proposed Model	2022	**96.87**	**97.38**

4.3.2 Classification results of the PatternNet dataset

Compared to the UC Merced, the PatternNet dataset provides a larger data size and more scene categories. The performance of the proposed model on the PatternNet dataset is reported in Table 4. The OAs of the proposed classification model and related works are listed in Table 5. Our model outperformed all the models and reached the highest OA of 98.76%, 98.46%, and 98.16% when the TRs were set to 80%, 60%, and 40%, respectively. The experimental results at TR=60% demonstrate that the proposed model increases the OA by 2.34%, 1.75%, and 0.15% over GoogleNet, ResNet-50, and VGG-16 [35], respectively.

Table 4: Evaluation of the proposed VHR scene classification model on the PatternNet.

Dataset	***PatternNet TR = 80%***
OA	0.9876
Precision	0.9877
Recall	0.9876
Kappa (κ)	0.9873
MCC	0.9873
ROC	0.9936
F1-Score	0.9876

Table 5: Comparison of overall accuracy (%) on the PatternNet dataset.

Method	***Year***	***Training Ratio (TR)***		
		40%	***60%***	***80%***
Full-trained VGG-16 [35]	2021	-	97.31	-
Full-trained GoogleNet [35]	2021	-	96.12	-
Full-trained ResNet-50 [35]	2021	-	96.71	-
Fine-tuning VGG-16 [35]	2021	-	98.31	-
Fine-tuning GoogleNet [35]	2021	-	97.56	-
Fine-tuning ResNet-50 [35]	2021	-	98.23	-
STHP [41]	2022	-	-	98.67
Proposed Model	2022	**98.16**	**98.46**	**98.76**

4.3.3 Time Complexity

In this section, we examine the time cost of the proposed model. We compared the average training and test times for the two selected VHR datasets. All experiments were performed in the same environment described in Section 4.2. The comparison results are presented in Table 6.

Table 6: Comparison of computation times on the two selected datasets.

Metric	*UC Merced*	*PatternNet*
TR	80%	80%
Training images	1,680 images	24,320 images
Feature size	1,280	1,280
OA (%)	97.38	98.76
Training time (min)	4.66	64.00
Test time (sec)	0.073	0.489

5. Conclusion

This paper proposes a scene classification method based on transfer deep CNNs to improve the performance of very high-resolution (VHR) imagery scene classification. In the classification process, an adaptive gradient algorithm (i.e., Adagrad) was used in conjunction with a multilayer perceptron (MLP) to improve the classification accuracy. We tested the proposed model against two selected VHR scene datasets, and it achieved high classification accuracy of 97.38% for UC Merced, and 98.76% for PatternNet. By comparing the proposed model to traditional CNN-based models, the accuracy is improved. In order to make VHR scene classification more practical for practical applications, further research is needed to address the shortcomings of the proposed model. To reduce the computational cost of our trained model, we aim to combine features into short vector lengths more than the extracted from EfficientNetV2L.

References

[1] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing,* vol. 7, pp. 14680-14707, 2015.

[2] 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*, 2010, pp. 270-279.

[3] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS journal of photogrammetry and remote sensing,* vol. 145, pp. 197-209, 2018.

[4] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium-100 Years ISPRS*, 2010, pp. 298-303.

[5] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and remote sensing letters,* vol. 8, pp. 173-176, 2010.

[6] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong*, et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, pp. 3965-3981, 2017.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, pp. 91-110, 2004.

[8] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 6, pp. 1899-1912, 2013.

[9] B. Luo, J.-F. Aujol, Y. Gousseau, and S. Ladjal, "Indexing of satellite images with different resolutions by wavelet features," *IEEE Transactions on Image Processing,* vol. 17, pp. 1465-1472, 2008.

[10] B. Luo, J.-F. Aujol, and Y. Gousseau, "Local scale measure from the topographic map and application to remote sensing images," *Multiscale modeling & simulation,* vol. 8, pp. 1-29, 2009.

[11] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, image and video processing,* vol. 10, pp. 745-752, 2016.

[12] 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*, 2011, pp. 385-388.

[13] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2933-2940.

[14] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 8, 2015.

[15] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, pp. 2175-2184, 2014.

[16] F. P. Luus, B. P. Salmon, F. Van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 12, pp. 2448-2452, 2015.

[17] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 54, pp. 1793-1802, 2015.

[18] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 12, pp. 2321-2325, 2015.

[19] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, pp. 4775-4784, 2017.

[20] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geoscience and Remote Sensing Letters,* vol. 14, pp. 1735-1739, 2017.

[21] M. Shahriari and R. Bergevin, "Land-use scene classification: a comparative study on bag of visual word framework," *Multimedia Tools and Applications,* vol. 76, pp. 23059-23075, 2017.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.
[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[26] K. Qi, C. Yang, C. Hu, H. Zhai, Q. Guan, and S. Shen, "A multi-level improved circle pooling for scene classification of high-resolution remote sensing imagery," *Neurocomputing,* vol. 462, pp. 506-522, 2021.
[27] O. A. Shawky, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "Remote sensing image scene classification using CNN-MLP with data augmentation," *Optik,* vol. 221, p. 165356, 2020.
[28] X. Zhang, W. An, J. Sun, H. Wu, W. Zhang, and Y. Du, "Best Representation Branch Model for Remote Sensing Image Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 9768-9780, 2021.
[29] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sensing,* vol. 10, p. 444, 2018.
[30] Q. Zhu, Y. Zhong, Y. Liu, L. Zhang, and D. Li, "A deep-local-global feature fusion framework for high spatial resolution imagery scene classification," *Remote Sensing,* vol. 10, p. 568, 2018.
[31] D. Zeng, S. Chen, B. Chen, and S. Li, "Improving remote sensing scene classification by integrating global-context and local-object features," *Remote Sensing,* vol. 10, p. 734, 2018.
[32] Y. Chen, L. Feng, X. Zhang, Z. Shen, and X. Zhou, "Supervised and adaptive feature weighting for object-based classification on satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 11, pp. 3224-3234, 2018.
[33] O. A. Shawky, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "A very high-resolution scene classification model using transfer deep CNNs based on saliency features," *Signal, Image and Video Processing,* vol. 15, pp. 817-825, 2021.
[34] H. Alhichri, A. S. Alswayed, Y. Bazi, N. Ammour, and N. A. Alajlan, "Classification of Remote Sensing Images Using EfficientNet-B3 CNN Model With Attention," *IEEE Access,* vol. 9, pp. 14078-14094, 2021.
[35] C. Peng, Y. Li, L. Jiao, and R. Shang, "Efficient Convolutional Neural Architecture Search for Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 59, pp. 6092-6105, 2021.
[36] X. Xu, Y. Chen, J. Zhang, Y. Chen, P. Anandhan, and A. Manickam, "A novel approach for scene classification from remote sensing images using deep learning methods," *European Journal of Remote Sensing,* vol. 54, pp. 383-395, 2021.
[37] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 2030-2045, 2021.
[38] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105-6114.
[39] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, 2021, pp. 10096-10106.
[40] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of machine learning research,* vol. 12, 2011.
[41] X. Gu, C. Zhang, Q. Shen, J. Han, P. P. Angelov, and P. M. Atkinson, "A Self-Training Hierarchical Prototype-based Ensemble Framework for Remote Sensing Scene Classification," *Information Fusion,* vol. 80, pp. 179-204, 2022.
[42] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 10, pp. 2889-2901, 2017.
[43] B. Yuan, L. Han, X. Gu, and H. Yan, "Multi-deep features fusion for high-resolution remote sensing image scene classification," *Neural Computing and Applications,* vol. 33, pp. 2047-2063, 2021.
[44] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 18, pp. 43-47, 2020.
[45] X. Sun, Q. Zhu, and Q. Qin, "A multi-level convolution pyramid semantic fusion framework for high-resolution remote sensing image scene classification and annotation," *IEEE Access,* vol. 9, pp. 18195-18208, 2021.
[46] S. Ansith and A. Bini, "Land use classification of high resolution remote sensing images using an encoder based modified GAN architecture," *Displays,* vol. 74, p. 102229, 2022.