Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5843816, 18 pages
https://doi.org/10.1155/2021/5843816

Research Article

Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50

Amsa Shabbir,[1] Nouman Ali,[1] Jameel Ahmed,[2] Bushra Zafar,[3] Aqsa Rasheed,[1] Muhammad Sajid,[4] Afzal Ahmed,[1] and Saadat Hanif Dar[1]

[1]*Department of Software Engineering, Mirpur University of Science & Technology (MUST), Mirpur 10250, AJK, Pakistan*
[2]*Department of Electrical Engineering, RIPHAH International University, Islamabad 75300, Pakistan*
[3]*Department of Computer Science, Government College University, Faisalabad 38000, Pakistan*
[4]*Department of Electrical Engineering, Mirpur University of Science & Technology (MUST), Mirpur 10250, AJK, Pakistan*

Correspondence should be addressed to Nouman Ali; nouman.ali@live.com

Received 10 May 2021; Revised 17 June 2021; Accepted 3 July 2021; Published 13 July 2021

Academic Editor: Muazzam Maqsood

Image classification has gained lot of attention due to its application in different computer vision tasks such as remote sensing, scene analysis, surveillance, object detection, and image retrieval. The primary goal of image classification is to assign the class labels to images according to the image contents. The applications of image classification and image analysis in remote sensing are important as they are used in various applied domains such as military and civil fields. Earlier approaches for remote sensing images and scene analysis are based on low-level feature representations such as color- and texture-based features. Vector of Locally Aggregated Descriptors (VLAD) and orderless Bag-of-Features (BoF) representations are the examples of mid-level approaches for remote sensing image classification. Recent trends for remote sensing and scene classification are focused on the use of Convolutional Neural Network (CNN). Keeping in view the success of CNN models, in this research, we aim to fine-tune ResNet50 by using network surgery and creation of network head along with the fine-tuning of hyperparameters. The learning of hyperparameters is tuned by using a linear decay learning rate scheduler known as piecewise scheduler. To tune the optimizer hyperparameter, Stochastic Gradient Descent with Momentum (SGDM) is used with the usage of weight learn and bias learn rate factor. Experiments and analysis are conducted on five different datasets, that is, UC Merced Land Use Dataset (UCM), RSSCN (the remote sensing scene classification image dataset), SIRI-WHU, Corel-1K, and Corel-1.5K. The analysis and competitive results exemplify that our proposed image classification-based model can classify the images in a more effective and efficient manner as compared to the state-of-the-art research.

1. Introduction

Image classification and analysis is an active research area and there are many applications of automatic image classification in computer vision domains such as pattern recognition, image retrieval, object recognition, remote sensing, face recognition, textile image analysis, automatic disease detection, geographic mapping, and video processing [1–3]. In any image classification-based model, the primary objective of research is to assign the class labels to images. A group of images are used as training samples and learning of classification-based model is done by using a training dataset. After training, the test dataset is assigned to the trained model to predict the class labels of images. On the basis of prediction of test dataset, images can be arranged in a semantic and meaningful order. Selection of discriminating and unique features is always beneficial as it can enhance the performance of any classification-based system [4–6]. In remote sensing, the problem of image classification is more challenging as objects are rotated within a view and background is usually more complex [7]. Satellites, unmanned aerial vehicles, and aerial systems are used to capture the image datasets that are used to evaluate the research of remote sensing [7]. According to the recent

reviews [8, 9], there are three main approaches that can be used to classify digital images and they are based on (i) low-level features representation [10], (ii) mid-level features representation [11–14], and (iii) approaches based on Convolutional Neural Network (CNN) [7].

Figure 1 represents a block diagram of a CCN which consists of multiple hierarchical layers including feature map layers, classification layers, and fully connected layers. CNN takes an input image, processes it, and classifies it under certain categories/class labels, for example, elephant, flower, cat, and dog. In a deep CNN, input image is passed through a series of layers called convolution layers with certain filters (kernels), pooling layers, fully connected layers, and finally classification layers. Typically, the first layer in CNN is convolution layer, which generates the feature maps with the help of filters [15, 16]. The filters that are used in convolution layers can perform operations such as edge detection, blurring, and sharpening. The feature maps generated by the convolution layers are passed to the sampling layers to reduce the size of the impending layers. They help to reduce the size of parameters when the size of the input image is large. The size is reduced in such a way that important information is preserved while omitting the information that is not necessary. Then, the feature maps are converted into vectors and passed to the fully connected layers. Finally, activation function and classification function classify the images into respective categories. Backpropagation is followed by CNN to carry out the process of classification in a more efficient way [8].

Figure 2 represents different levels for remote sensing image classification which are (i) pixel level, (ii) object level, and (iii) scene level [8]. According to the literature [8, 17], the early research models for remote sensing image classification are based on pixel level or subpixel level. The reason for this classification is the low resolution of satellite image as capturing devices are not that capable to create a high-resolution image as available information is in the form of small pixels [18, 19]. Due to recent advancement in imaging technology, the spatial resolution of remote sensing images is increasing, and it is possible to capture the visuals in more semantic way [8]. Due to this reason, in satellite image classification, it is not much beneficial to focus more on pixel level [8]. Blaschke and Strobl [20] concluded that, for remote sensing image classification, it is more beneficial to focus on object-level classification instead of pixel-level analysis. The authors suggested that object-level analysis for remote sensing images is more efficient and semantic as compared to the previous approaches based on pixel-level analysis. Since the last two decades, significant research has been published by considering the object-level classification for remote sensing images [18, 19]. Later on, due to advancement in technology of image-capturing devices, remote sensing images may contain many object classes [8]. So, in this case, the former two pixel-level and object-level approaches may not be significant. Due to this reason, it is considered to classify the images in a global context, and the focus of research is shifted to the use of scene-level remote sensing image classification. The scene-level classification of images is considered as a significant approach to represent visual information as discriminating features [8]. In last two decades, extensive efforts are exerted by computer vision research community to develop the discriminating features such as Scale-Invariant Feature Transformation (SIFT) [21], Speeded-up Robust Features (SURF) [22], Histogram of Oriented Gradients (HOG) [23], and Maximally Stable Extremal Regions (MSER) [24]. Bag-of-Features (BoF), Spatial Pyramid Matching (SPM), and Vector of Locally Aggregated Descriptors (VLAD) are the examples of simple and efficient encoding models and they have been used in various fields of remote sensing and scene classification [25, 26]. Due to recent increase in the size and number of training images, the use of CNN models and Graphics Processing Unit (GPU) are considered as current research trends. The concept presented by Hinton and Salakhutdinov by using multilayered neural networks has provided a foundation for deep learning research [27].

The comprehensive literature reviews about remote sensing image classification and use of recent trends of deep learning models can be found in [8, 17, 28, 29]. According to the literature, the most popular CNN architectures are AlexNet [30], VGG network [31], Residual Network (ResNet) [32], and GoogLeNet [33]. There are 08 layers in AlexNet [30], 19 layers in VGG network, and 22 layers in GoogLeNet [34]. ResNet50 is based on ResNet with 50 layers and is inspired from the idea to make deeper layers with a higher value of classification accuracy for complex tasks [35]. Usually in neural networks, when we increase the number of layers, the classification accuracy begins to degrade, while this problem is handled by residual training [35]. Here are the main contributions of this research:

(i) We fine-tuned ResNet50 by using network surgery and creation of network head along with the fine-tuning of hyperparameters.

(ii) The learning of hyperparameters is tuned by using a linear decay learning rate scheduler known as piecewise scheduler. To tune the optimizer hyperparameter, Stochastic Gradient Descent with Momentum (SGDM) is used with the usage of weight learn and bias learn rate factor.

(iii) Experiments and analysis are conducted on five different datasets, that is, UC Merced Land Use Dataset (UCM), RSSCN (the remote sensing scene classification image dataset), SIRI-WHU, Corel-1K (1000 images), and Corel-1.5K (1500 images). The analysis and competitive results exemplify that our proposed image classification-based model can classify the images in a more effective and efficient manner as compared to the state-of-the-art research.

The remainder of the paper is organized as follows: Section 2 is about literature review and discussion about relevant research based on remote sensing image classification, Section 3 presents the proposed fine-tuned ResNet50 and provides details of ResNet50 parameters, Section 4 is about the description of image benchmarks that are used for evaluation of this research, Section 5 is about results,

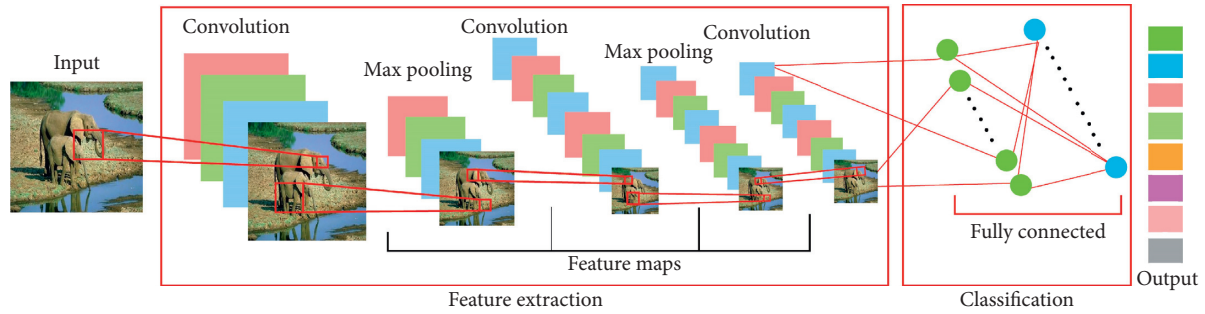

Figure 1: Image classification-based framework of a CNN.

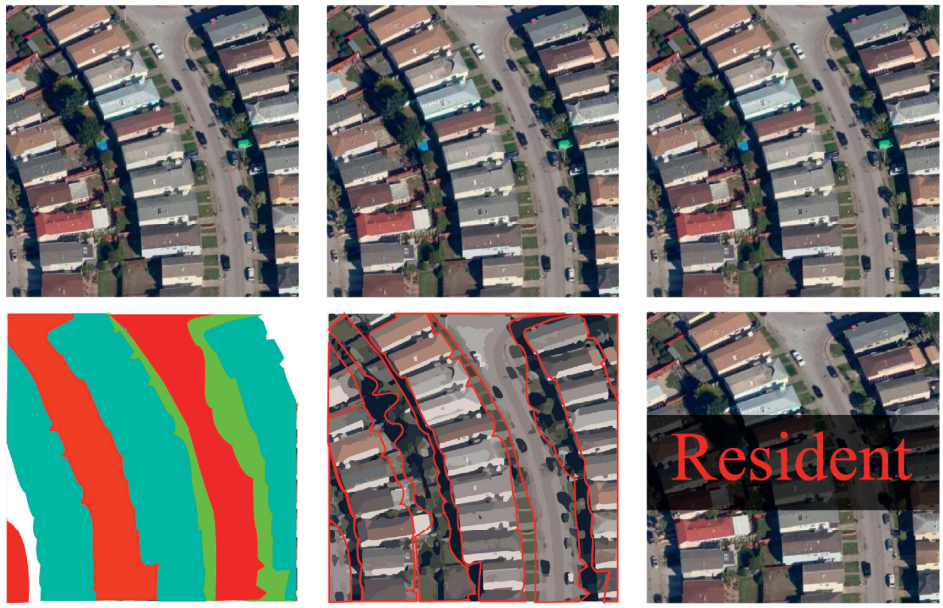

Figure 2: Different levels of remote sensing classification [8]. (a) Pixel-level image classification. (b) Object-level image classification. (c) Scene-level image classification.

experimental values, discussion, and comparisons, and Section 6 concludes the proposed research based on fine-tuned ResNet50.

2. Related Work

Content-based image analysis is widely used in various applied and real-time domains of computer vision [36, 37]. Classification of images according to the image contents, visual appearance, and human visual perception is considered as an open research problem [38]. Remote sensing image classification approaches are broadly categorized into three groups based on the type and the usage of visual clues, that is, approaches based on low-level visual features, approaches based on mid-level features, and high-level feature extraction approaches [11, 39]. We have hand-picked recent state-of-the-art approaches from the above-mentioned categories, which have reported results on similar image benchmarks. The earlier research for remote sensing and scene classification is formulated on the use of low-level visual features [40, 41]. Khalid et al. [40] reduced the semantic gap and proposed an efficient feature vector-based image representation. Histogram-based approach is used to compute the feature vector of images. The authors extracted the autocorrelogram by using RGB format that is followed by a moment's extraction. The efficiency is enhanced by applying Discrete Wavelet Transform (DWT) on multiple resolutions and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used to compute the codebook. Different variants of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT) are used to classify images, and the authors have presented a comprehensive comparison while using different classifiers. The proposed research based on DBSCAN is evaluated on three publicly available datasets, that is, Corel-1K, Corel-1.5K, and Corel-5K [40]. Raja et al. [41] proposed an approach for content-based image analysis which is based on feature extraction from color images. The region of

interest in an image is computed with the help of first-order derivatives. Due to closeness with respect to human visual perception, the HSV (Hue, Saturation, Value) histograms are used to represent the color space. Neural networks (NN) are used for the purpose of image classification/class label assignments, and the results are reported while using Corel-1K and Corel-5K image benchmarks [41]. Desai et al. [42] proposed an image representation based on fusion of different features. The authors selected a combination of low-level visual features, which are DWT, Edge Histogram Descriptor (EHD), Sobel operator, Moment Invariant (MI), Histogram of Oriented Gradients (HoGs), and Local Binary Pattern (LBP). Different combinations of low-level visual features are evaluated to sort the most reliable image representation. According to the published results values [42], a combination of low-level features with SVM outperforms all other features combination. Shikha et al. [43] proposed a hybrid image representation and low-level attributes of images are computed by using a combination of color, shape, and texture. The authors computed a hybrid feature vector (HFV) by using a feature integration of three different visual attributes. A feed-forward neural network known as Extreme Learning Machine (ELM) is trained while using input as HFV. To enhance the performance of system, Relevance Feedback (RF) is applied to ELM. The performance of the proposed system is evaluated while using Corel-1K, Corel-5K, Corel-10K, and GHIM-10 image benchmarks.

Aslam et al. [14] proposed a late fusion of mid-level features based on BoF model. According to the authors, mid-level image representation late fusion can enhance the performance of image classification-based model. In this research [14], the late fusion of Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) is proposed by using BoF representation model. Support Vector Machine (SVM) is applied for the classification of histograms that are created on the basis of late fusion of two mid-level features. The proposed late fusion is evaluated while using Corel-1K and Corel-1.5K image benchmarks. Yu et al. [44] proposed High-order Distance-based Multiview Stochastic Learning (HD-MSL) approach for classification. According to the authors, the proposed learning approach (HD-MSL) is based on the features combination and labeling information is computed by applying a probabilistic framework. Spatial Pyramid Matching (SPM) and BoF model are used to represent various mid-level image categorization-based approaches. Zafar et al. [12] stated that SPM can only capture the absolute spatial distribution of visual words and is not robust to image transformations such as translation, flipping, and rotations. The discriminating power of SPM degrades if images are not well aligned and, due to this reason, Zafar et al. [12] proposed an image representation that can compute the relative spatial information based on histogram of Bag of Visual Words (BoVW) model. Global relationship of identical visual words with image centroid was explored by the authors to achieve the objective. Five image benchmarks are used for the evaluation of this research [12]. Ali et al. [11] stated that the classification accuracy of orderless BoF-based histograms suffers due to unavailability of image spatial clues. The approaches that are centered on splitting of images into subblocks to capture spatial clues cannot handle rotations. In case of remote sensing image classification, these spatial clues can increase the learning ability and classification accuracy of the trained model [11]. The authors proposed in [11] a rotation invariant feature vector-based image representation that can compute spatial clues with the help of orthogonal vectors histograms. The results are computed while using three publicly available satellite image benchmarks (SIRI-WHU, RSSCN, and AID) [11]. Figure 3 shows an example of image classification based on a CNN model. Fine-tuning is used with transfer learning to adjust the parameters of a pretrained CNN model by using a new dataset with different number of classes. This process is beneficial as the training is done with small learning rate by reducing number of training epochs [7, 45]. According to Petrovska et al. [7], the recent focus of research for image classification is on the use of a pretrained CNN. The authors of [7] used a CNN for features extraction and then training was performed by using these extracted features. Transfer learning was implemented by the authors for the purpose of fine-tuning using pretrained CNNs. Support Vector Machine (SVM), Radial Basis Function (RBF) kernels are used for the purpose of image classification. Linear decay learning rate scheduler and cyclical learning rates are used to tune the hyperparameter of the network and label smoothing regularization is used to avoid the overfitting. Shafaey et al. [46] explored a deep learning model performance for remote sensing image classification. A comprehensive review is presented by considering the deep learning models such as AlexNet, VGGNet, GoogLeNet, Inception-V3, and ResNet101. Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and SVM are used for predicting the class labels, and the results are compared with the above-mentioned deep learning models. A detailed quantitative comparison in terms of results is presented by considering seven publicly available datasets [46]. In another research, Zhao et al. [47] stated that Residual Dense Network (RDN) is with more learning ability as it can utilize the information available in convolutional layers. The authors designed an RDN that is based on channel-spatial attention for the classification of remote sensing images. In the first step, multilayer convolution features are fused by using residual dense blocks and, in the next step, channel-spatial attention module is applied to enhance the effectiveness of features. By considering the training requirements, data augmentation is applied, and classification is done with the help of softmax classifier. The proposed research of Zhao et al. [47] is evaluated while using UCM and AID image benchmark.

3. Proposed Method of Research

The proposed methodology aims to enhance the image classification accuracy while using CNN model. Keeping in view the robust performance of the model, we selected Residual Network (ResNet50) for evaluation. ResNet50 is the short form of Residual Network with 50 layers. When researchers started to follow the phrase "the deeper the

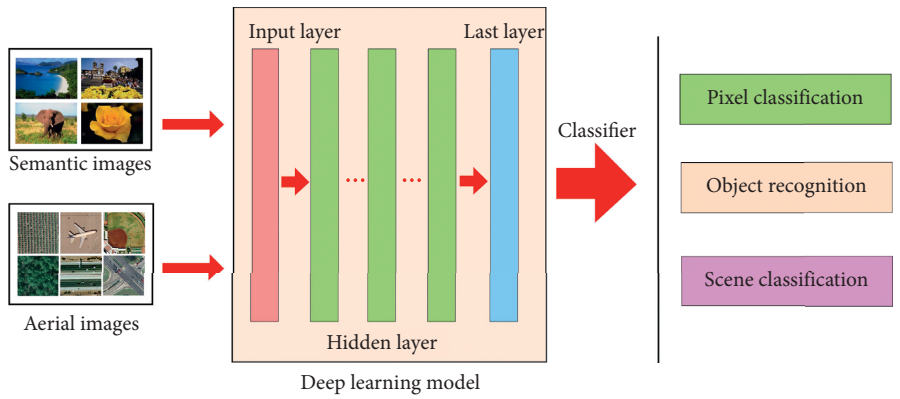

Figure 3: An example of image classification based on CNN.

better" with deep learning models, they encountered some problems. "The deeper the network is the performance of the network should be better"; this theory was proved wrong when a deep network with 52 layers generated bad results as compared to the networks with 20–30 layers [32]. Multiple predictions are reported about this decrease in the performance of the model and the most appropriate reason for this is the vanishing gradients. When the network is too deep, the gradient value shrinks to 0, which causes the weights not to update, and as a result no learning is performed. Figure 4 shows the phenomena of vanishing gradients.

Deep networks faced many complications including the optimization of networks, degradation, and most importantly vanishing gradients. According to literature, fine-tuning of a pretrained CNN network can increase the classification accuracy in the respective domain [48, 49]. ResNet50 is trained on ImageNet, which consists of almost 1.2 million images whose features and weights are transferred to the next task using the same pretrained network. Fine-tuning works and processes a new task with different numbers of classes and categories. The number of epochs referred to as iterations used to train a fine-tuned network is less compared to training the model from scratch. The motivation behind the usage of pretrained networks is to intensify the accuracy by using the concept of "transfer learning." Transfer learning refers to machine learning technique, which allows the transfer of information learnt from one domain to similar problems in related domain. It is recommended to use the model developed and trained for a task as a starting point of the task that is similar to the trained one [50]. Researchers have used diverse notations to describe different concepts of transfer learning to define it. Domain and task are the two basic concepts of transfer learning, which are explained mathematically. Transfer learning is defined arithmetically to make the picture clearer [51]. Domain D consists of two parts, that is, a feature space F and a marginal distribution $P(F)$ [51].

$$D = \{\mathrm{F}, P(F)\}, \tag{1}$$

Here, F represents an occurrence set (called instance set), which is explained as $F = \{x | x_i \in \mathrm{F}, i = 1, \ldots, n\}$. A task T comprises a decision function t and a label space L; that is,

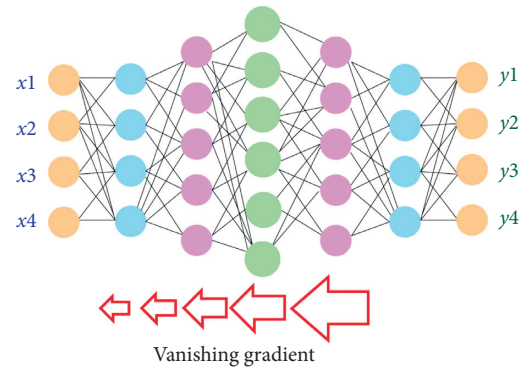

Figure 4: Vanishing gradients in CNN.

$$\mathrm{T} = \{\mathrm{L}, t\}. \tag{2}$$

A starting domain referred to as source D_S related to a main task (source) T_S is analyzed by the number of occurrence-label pairs; that is, $D_S = \left\{(x, y) | x_j \in \mathrm{F}^S, y_j \in \mathrm{T}^S, j = 1, \ldots, q^S\right\}$; target domain observation usually comprises unassigned occurrences and/or limited labeled occurrences.

Here, we report some observation(s) related to $m^S \in \mathbb{N}^+$ source domain(s) and task(s), that is, $\left\{(D_{S_k}, \mathrm{T}_{S_k}) | k = 1, \ldots, m^S\right\}$, and observation(s) corresponding to $m^{\mathrm{T}} \in \mathbb{N}^+$ target domain(s) and task(s), that is, $\left\{(D_{\mathrm{T}_j}, \mathrm{T}_{\mathrm{T}_j}) | j = 1, \ldots, m^{\mathrm{T}}\right\}$; based on the knowledge implied in the source domain(s) the learned decision functions performance amplifies with the help of transfer learning $f^{T_j}(j = 1, \ldots, m^T)$ on the target domain(s).

Deep Neural Network (DNN) ResNet50 is fine-tuned by doing "network surgery." In the process of network surgery, final layers of the pretrained network are removed. The layers removed from the network are "fc1000," "fc1000 softmax," and "ClassificationLayer fc1000" layers. These layers are than replaced with the new layers. The new layers introduced into the architecture establish a "network head." The composition of network head is the combination of three layers: A fully connected layer with WeightLearnRateFactor given a value of 20 and BiasLearnRateFactor given a value of 20. The second layer added is a new softmax layer

and finally a new classification layer is added to the network head. Learning rate is said to be the step size (which is the number of weights updated during training) at each iteration in the model. It is perhaps the most important hyperparameter to tune the neural network. It is a configurable hyperparameter that can be altered according to the needs to enhance the performance of the model. The learning curve which is also known as a function is expressed as [52]

$$\gamma = a\chi^{b}, \tag{3}$$

where γ represents the progressing average time called cumulative (or cost) per unit, χ is the progressing/growing number of units manufactured, a shows the time necessary to obtain the first unit, and b = log of the learning rate/log2. Learning rate in our model is modified and an initial learning rate is assigned to the model, which is 0.001, while a learning rate schedule is applied which will be used to modulate how the learning rate of the optimizer changes over time [53]. While training neural network models, it is suggested to reduce the learning rate with respect to training progress. The learning rate is reduced using predefined schedule; in our case, we used piecewise learning rate schedule. With the increase in epochs or iterations, the learning rate decreases using the predefined schedule. The mathematical form that is used to calculate the learning rate (decreasing) is given as [54]

$$\eta_{n+1} = \frac{\eta_n}{1 + \mathrm{d}n'}, \tag{4}$$

where n is iteration step, η_n is learning rate at the nth step, and d is decay rate. As the learning progresses, the rule updates the learning rate by reducing the denominator. Since n is initialized at zero, 1 is added to the denominator in order to prevent it from being zero.

We used Stochastic Gradient Descent with Momentum (SGDM) as optimizer. This helps gradient vectors to accelerate into the direction in which they are supposed to. Usage of SGDM enhances the converging process. The mathematical representation of SGDM is given as follows [55]:

$$\begin{aligned} m_{t,i} &= \beta m_{t-1,i} + g_{t,i}, \\ \theta_{t+1,i} &= \theta_{t,i} - \alpha m_{t,i}. \end{aligned} \tag{5}$$

The momentum gained at the tth recurrence for the ith parameter is $m_{t,i}$. The hyperparameter that controls the momentum is β. SGDM is an improved version of SGD with better convergence rate than the former one. Figure 5 shows the proposed research methodology, while Figure 6 demonstrates the process of fine-tuning.

The Residual Network (ResNet) has solved the problems associated with deep networks with the addition of new neural network layer called the Residual Block. The idea of solving identity function through neural network seemed easy and hence the output of the function becomes the input itself. The following equation represents the identity function which is considered to be of prime importance in solving the problem of deep architectures [32].

$$\mathsf{F}(x) = x. \tag{6}$$

By providing the input of the initial layer of the model as the output of the last layer, it is assumed that the model will learn and predict whatever it was learning before the addition of input.

$$\mathsf{F}(x) + x = H(x). \tag{7}$$

The above equations are important, and they formulate the concept of "skip connection" and identity mapping. Identity mapping is a simple concept and has no parameters. Its main function is to add the output from the descending layers to the preceding layers. The diagram below shows the architecture of ResNet50 with all the layers. When x and $\mathsf{F}(x)$ have the same dimensions, the process follows the same equations; however, sometimes the dimensions of both $\mathsf{F}(x)$ and x are not the same. In that case, a multiplication factor $W$is introduced to match the shortcuts or skip connection. By doing so, x and $\mathsf{F}(x)$ become the input of next layer as explained by the following equation:

$$y = \mathsf{F}\left(x, \left\{W_j\right\}\right) + W_s x. \tag{8}$$

This equation is used when $\mathsf{F}(x)$ and x are of different dimensions. W_s adds extra parameters to the model which helps to avoid the problems of dual dimensionality. With the help of ResNet, gradients can flow using skip connections back to initial layers without touching all the layers. In ResNet50 architecture, there are different groups of identical layers, and each group is distinguished by a different color used in Figure 7. The curve lines represent the skip connection or identity mapping through which the input of previous layer is passed into the next layers. These skip connections are the key features that help ResNet to overcome the problems of degradation and vanishing gradients. The figure illustrates that the first layer is a convolution layer with 7×7 size and 64 kernels followed by 3×3 max pooling layer. Next there is a block of identical layers separated by different colors. The curves in Figure 7 represent the skip connections. The overall parameters of ResNet50 are 23.521 M. Multidimensional input problem is handled by introducing two shortcuts. These shortcuts are identity shortcut and projection shortcut. The identity shortcut does a simple operation of bypassing the input to the addition operator. Projection shortcut makes sure that the inputs at addition operation are of the same size and performs the convolution operation to make this possible.

To escalate the efficiency and competence of the model, the process of fine-tuning is performed. This is a very critical process and small modifications with careful observations are done to get the better accuracy and optimization. The changes that are made for the purpose of fine-tuning are so crucial that they affect the training process a lot. We repeated the process of fine-tuning over and over again to increase the accuracy of our model. Table 1 illustrates the parameters that affected the accuracy and performance of our model.

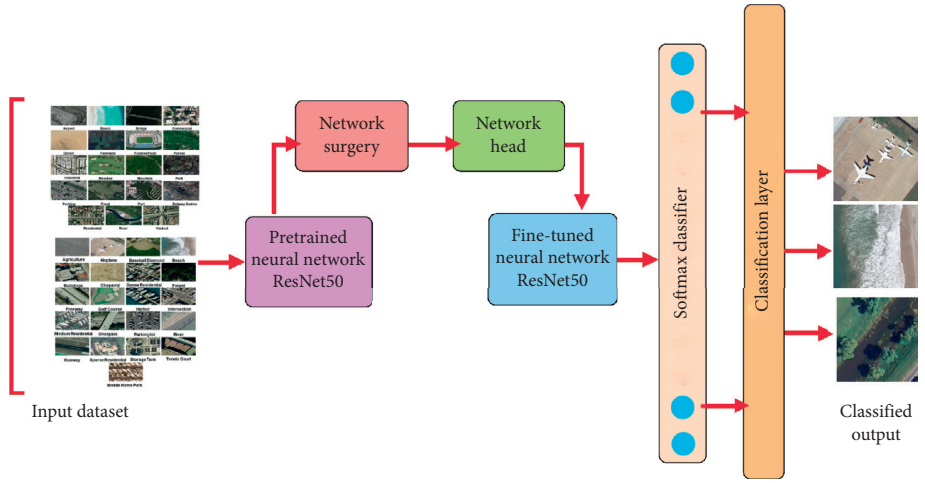

Figure 5: The proposed method of research.

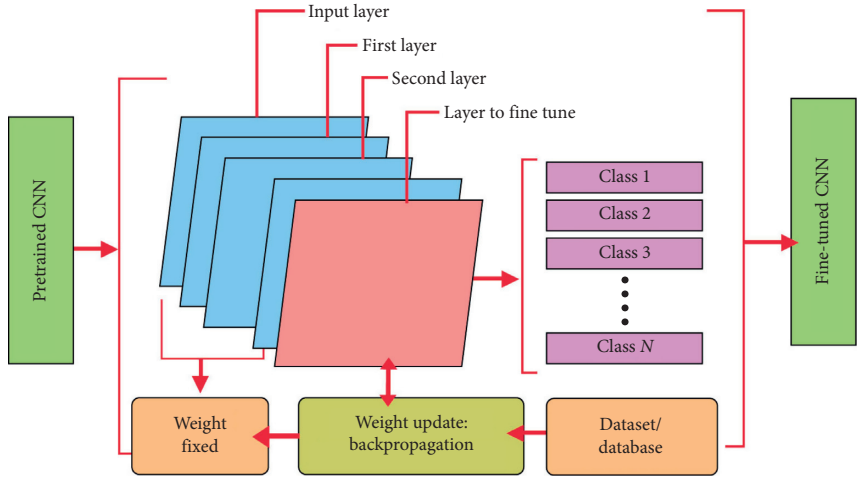

Figure 6: Details about fine-tuning of CNN.

4. Dataset Description

To analyze the effectiveness of the implemented technique, diverse image classification benchmarks which are widely used in literature have been utilized. Table 2 summarizes details regarding the total number of classes, images per class, number of images per class and total number of images in the benchmark, image spatial resolution, and dimensions:

(i) RSSCN: the remote sensing scene classification dataset [59] comprises images gathered from Google Earth Engine and covers widespread areas. RSSCN consists of 7 classes of quintessential scene images having a size of 400×400 pixels. Figure 8 shows indiscriminately selected samples of those classes and areas. Further description about this image benchmark can be found in [59].

(ii) SIRI-WHU: the description such as image size, total number of images, images per class, and date of creation can be found in [56]. The images have a spatial resolution of 2 m with image size of 200×200 pixels. Figure 9 shows randomly selected images taken from each class of SIRI-WHU dataset.

(iii) UC Merced Land Use Dataset: the description such as image size, total number of images, images per class, and date of creation can be found in [57]. There are a total of 21 distinctive scene categories with 100 images per class and dimensions of 256×256 pixels. Figure 10 shows indiscriminately selected examples of each category included in the dataset.

(iv) Corel-1K: the third dataset used for experimentation is Corel–1K [58], comprised of 1000 varying images. Wang's image dataset is organized into 10 semantic categories. Each category consists of 100 instances with image size of either 256×384 for portrait or 384×256 for landscape orientation. Figure 11 demonstrates indiscriminately selected images from Corel–1K image benchmark.

(v) Corel-1.5K: the last dataset used in our experiments is the Corel-1.5K image benchmark, which is a

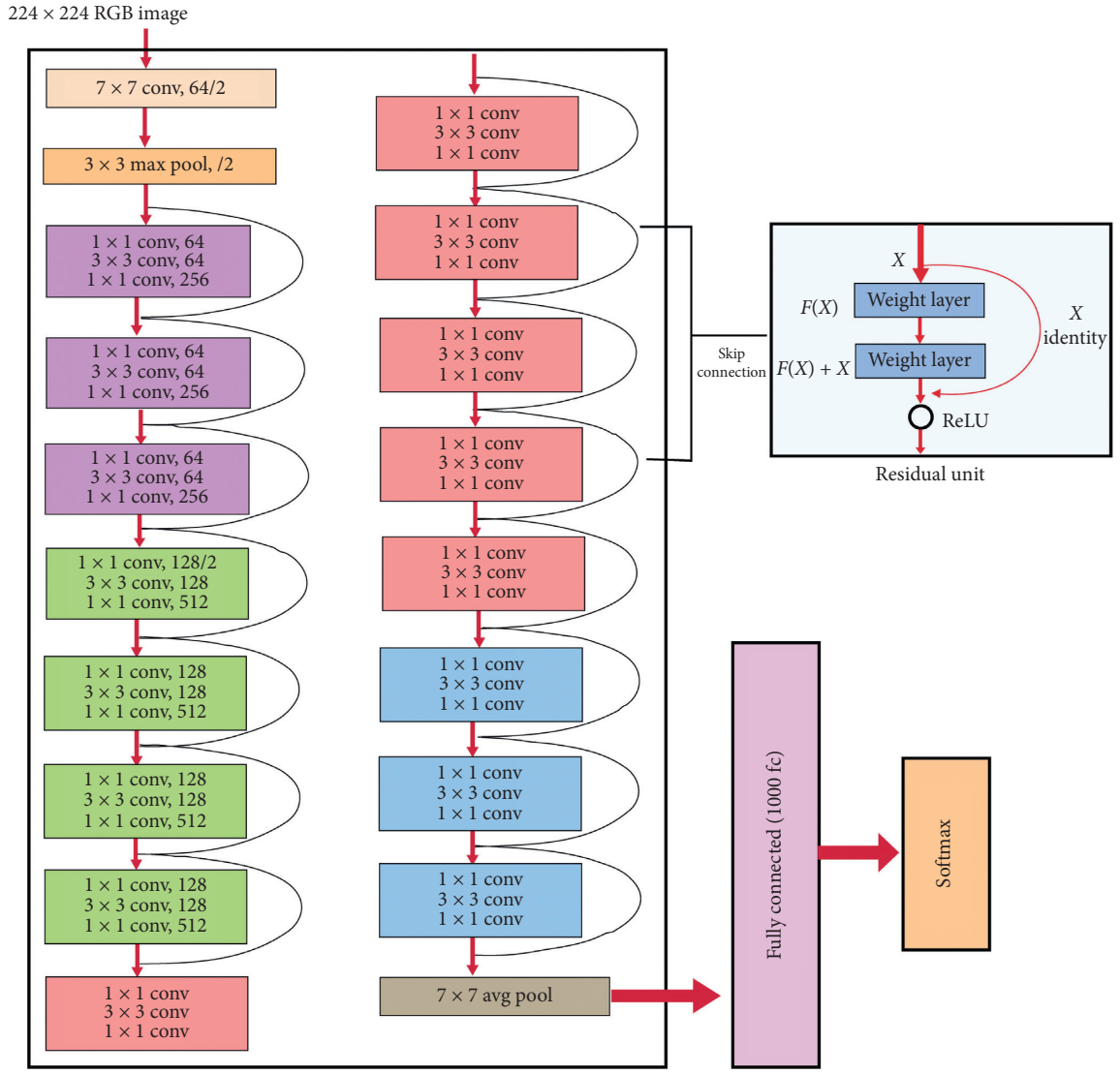

Figure 7: ResNet50 architecture with internal layerwise detail.

Table 1: Details about ResNet50 parameters.

Parameter	Value
Epochs	100
Validation step	1
Optimizer	SGDM (Stochastic Gradient Descent with Momentum)
Learning rate	Piecewise scheduler
Decay	Default
Momentum	Default

subset of Corel image dataset [58]. The dataset is comprised of 1500 images organized into 15 semantic categories. Figure 12 shows indiscriminately selected samples from each class of the dataset.

5. Performance Evaluation

All the experiments have been performed while using HP-ENVY-x360, with Intel Core-i7-7500U CPU, 2.7 GHz 2.9 GHz, 16 GB RAM, 64-bit Windows 10 OS, and 256 GB SSD as primary storage for OS; and a training : testing ratio of 70 : 30 is used for all experiments. This section provides details of the evaluation metrics used and presents a comprehensive discussion on results. The most widely used metric for evaluation of classification performance is the classification accuracy (A), defined as total instances (images) correctly classified and fractionated by total number of instances (images) within the dataset under consideration. It is mathematically expressed as

$$A = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}, \tag{9}$$

Table 2: Summary of standard image benchmarks used for evaluation of the proposed research.

Dataset	Classes	Images per class	Total images	Spatial resolution	Image size
RSSCN [12]	7	400	2800	—	400×400
SIRI-WHU [56]	12	200	2400	2 m	200×200
UCM [57]	21	100	2100	0.3 m	256×256
Corel-1K [58]	10	100	1000	—	256×384 or 384×256
Corel-1.5K [58]	15	100	1500	—	256×384 or 384×256

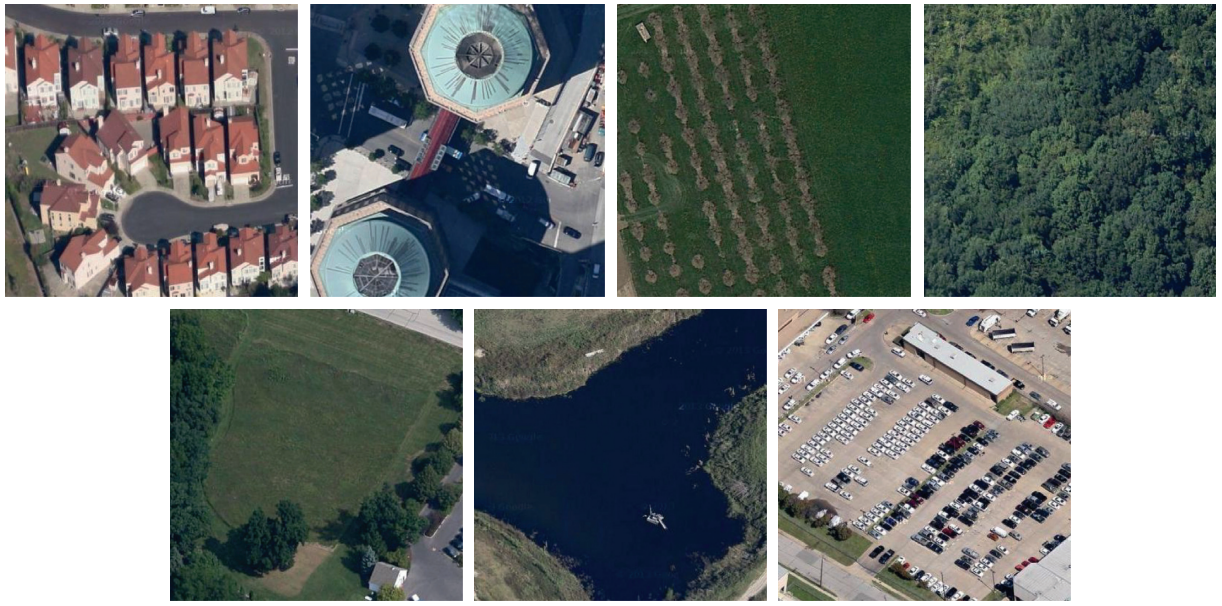

Figure 8: The photo gallery based on a random selection of images taken from each class of RSSCN image benchmark [59].

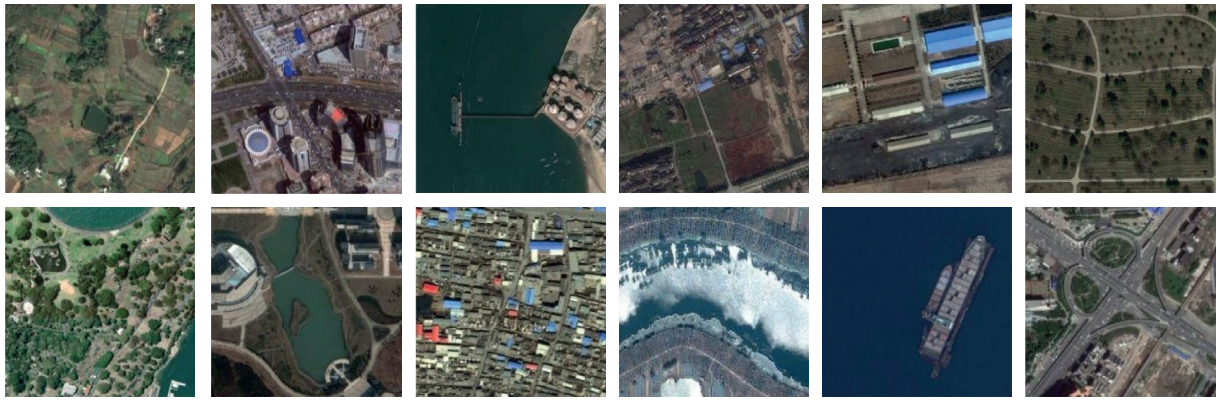

Figure 9: The photo gallery based on a random selection of images taken from each class of SIRI-WHU image benchmark [56].

where tp denotes true positives, tn denotes true negatives, fp denotes false positives, and fn denotes false negatives.

Precision (P) and recall (R) are used very commonly for the performance assessment of image classification systems. Precision is the equivalence of the ratio of correctly classified images to the total number of classified images.

$$P = \frac{\text{tp}}{\text{tp} + \text{fp}}. \tag{10}$$

Here, tp represents the correctly classified image and fp represents misclassified images, also known as false positives.

The recall is the fraction of correctly classified images to the total number of related images present in the database. The mathematical form of recall is

$$R = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{11}$$

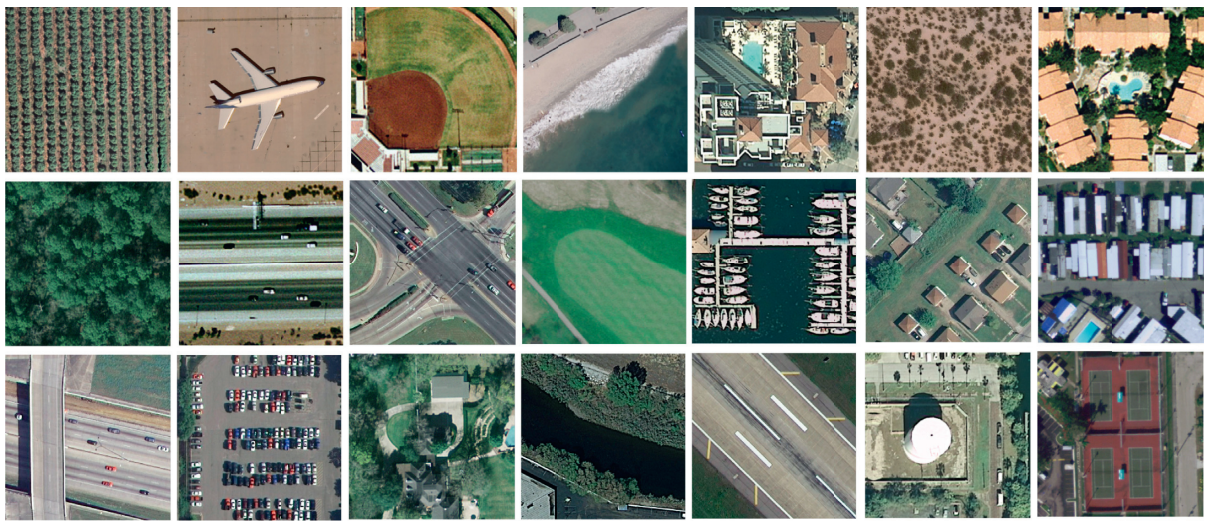

Figure 10: The photo gallery based on a random selection of images taken from each class of UCM image benchmark [57].

Figure 11: The photo gallery based on a random selection of images taken from each class of Corel-1K image dataset [58].

Figure 12: The photo gallery based on a random selection of images taken from each class of Corel-1.5K image dataset [58].

Here, fn denotes false negatives, the images which belonged to the correct class but were misclassified by the classifier.

F-score is the result of the harmonic mean of precision and recall; a higher value of it is a symbol of the better predictive power of the system. Alone precision or recall is not adequate to evaluate the performance of systems. *F*-score could be expressed mathematically as

$$F - \text{score} = 2 \times \left(\frac{(P.R)}{(P + R)} \right). \tag{12}$$

Here, P and R represent precision and recall, respectively. *F*-score is used for comparing the performance in those scenarios, where one approach has higher precision but a lower recall rate than the comparative approach.

5.1. Results for RSSCN Image Benchmark. The classification accuracy and performance of the proposed approach in comparison with the state-of-the-art research are shown in Table 3. Here, the proposed research based on fine-tuned ResNet50 outperforms the approaches based on mid-level features, that is, RGSIR [12] and POVH [11], by 10.56% and 7.93%, respectively, which are based on low-level handcrafted features. Table 3 shows a quantitative analysis and comparison of the proposed fine-tuned ResNet50 with the methods based on deep learning architectures. It can be evidently seen that the proposed research achieves highest classification accuracy as compared to the methods based on deep learning models, that is, AlexNet, GoogLeNet, Inception-V3, VGG-VD-16, and CaffeNet, outperforming these methods by 6.4%, 6.16%, 5%, 4.82%, and 3.75%, respectively.

Figure 13 demonstrates the precision, recall, and *F*-score for RSSCN image dataset using the proposed research. *F*-score is important since if precision or recall values are very low, *F*-score helps balance the two metrics. The higher the *F*-score, the better the results, with 0 being the worst possible and 1 being the best. A good *F*-score is indicative of a good precision and recall value. The average precision, recall, and *F*-score for RSSCN image benchmark are 92.74%, 92.84%, and 92.76%, respectively.

Figure 14 shows confusion matrix from RSSCN image benchmark. The confusion matrix summarizes the performance of a classification algorithm and provides an insight into how correct the predictions were and how they hold up against the actual values. On the confusion matrix plot, the rows correlate to the true class and columns conform to the predicted class. The diagonal values correspond to correctly classified observations. The off-diagonal values indicate the observations incorrectly classified.

5.2. Results for SIRI-WHU Image Benchmark. The experimental results for the SIRI-WHU image dataset are presented in Table 4. It can be evidently seen that the overall classification accuracy of the proposed research is higher than that of the research selected for comparison. POVH [11] uses mid-level attributes or features and captures the spatial attributes, which are considered very important for classification of satellite imagery. The proposed research based on high-level features outperforms POVH by 13.89%. Further the comparison of the proposed research is presented against deep learning models. The proposed research based on ResNet50 surpasses the state-of-the-art deep learning models VGGNet, Inception-V3, GoogLeNet, and AlexNet by 7.43%, 5.03%, 4.73%, and 3.83%, respectively.

Table 5 shows the precision, recall, and *F*-score for each class of SIRI-WHU image benchmark. The average precision, recall, and *F*-score for the SIRI-WHU image dataset are 94.03%, 94.19%, and 94.02%, respectively.

Figure 15 demonstrates the confusion matrix for the SIRI-WHU image dataset.

5.3. Results for UCM Image Benchmark. In this subsection, we will discuss the result of UCM image benchmark. Table 6 presents a comparison of proposed fine-tuned ResNet50 with recently published research and deep learning models.

Table 3: A quantitative comparison with recently published research in terms of classification accuracy for RSSCN image benchmark.

Name of algorithm/model	Classification accuracy (%)
RGSIR [12]	81.44
POVH [11]	84.07
AlexNet [46]	85.6
GoogLeNet [39]	85.84
Inception-V3 [46]	87
VGG-VD-16 [39]	87.18
CaffeNet [39]	88.25
ResNet50	92

It can be clearly seen that the proposed approach based on ResNet50 achieves the highest classification accuracy as compared to the related research. In [46], the authors used Inception-V3 deep learning model, and their reported accuracy was 6.68% times low as compared to the proposed research. The authors in [60] proposed an approach based on fusion of low-level features with high-level ResNet features and used SVM as classifier. The proposed approach achieves 3.97% higher classification accuracy as compared to the feature fusion-based approach [60]. The proposed research outperforms AlexNet, GoogLeNet, CaffeNet, and VGG-VD-16 by 3.58%, 3.47%, 2.76%, and 2.57%, respectively.

Table 7 shows the precision, recall, and *F*-score for each class of UCM image benchmark.

The average precision, recall, and *F*-score for the UCM image dataset are 97.78%, 97.83%, and 97.77%, respectively. Figure 16 demonstrates the confusion matrix for the UCM image dataset. Here, we can see that most of the classes are correctly classified, and the major confusion is observed between classes storage tanks and buildings, medium residential, and dense residential. This is because the classes medium residential and dense residential are overlapped and vary in the density of structures.

5.4. Results for Corel-1K Image Dataset. Corel-1K image benchmark is the third dataset used for the experimentation in this research. Table 8 presents a comparison of the research proposed with the state-of-the-art research. It can be manifestly seen that the proposed research provides the highest accuracy and outperforms the state-of-the-art approaches based on mid-level and high-level features. In [43], a hybrid feature vector is created by integrating three visual attributes, that is, color, texture, and shape. The experimental evaluation and analysis illustrate that the implemented technique outstrips many state-of-the-art related approaches based on varied hybrid systems. The proposed research achieves the highest accuracy as compared to the state-of-the-art research, thereby outperforming the researches of Li et al. [61], Aslam et al. [14], SCNN-ELM [61], MKSVM-MIL et al. [62], Raja et al. [41], Desai et al. [42], Yu et al. [44], and Shikha et al. [43] by 26.16%, 15.74%, 12.68%, 11.8%, 10.34%, 8.8%, 1.02%, and 0.5%, respectively.

Table 9 demonstrates the classwise performance for Corel-1K image benchmark in terms of precision, recall, and *F*-score. The average precision, recall, and *F*-score values for

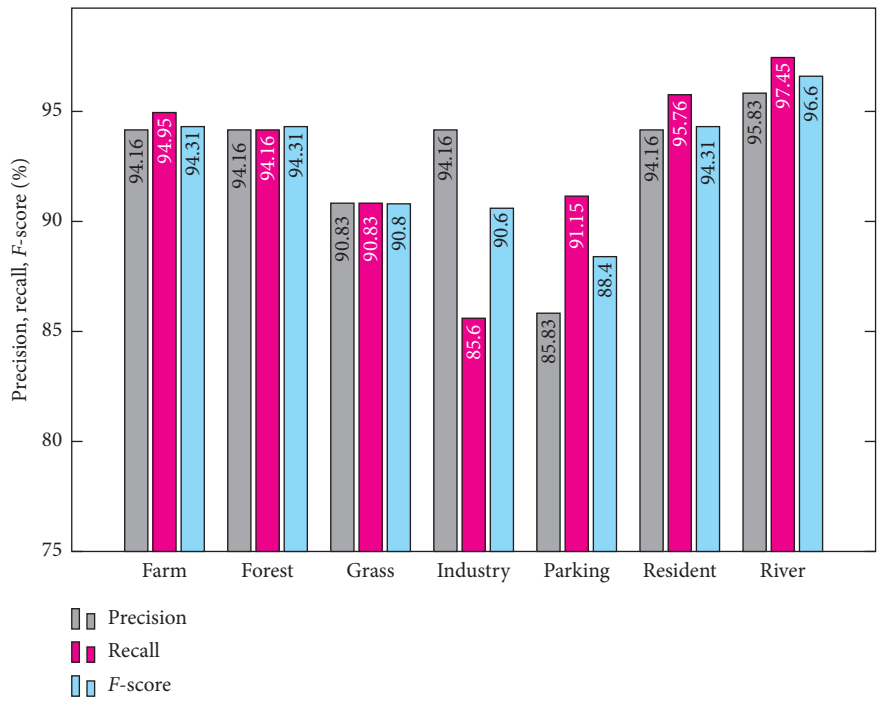

Figure 13: Precision, recall, and *F*-score for each class of RSSCN image dataset.

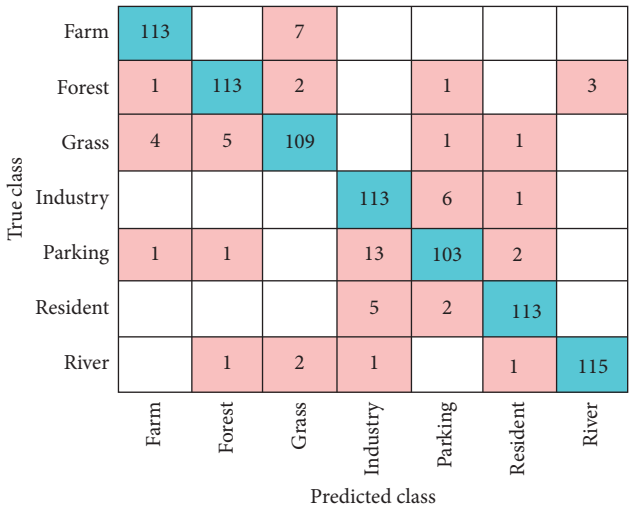

Figure 14: Confusion matrix for RSSCN image dataset.

Corel-1K image benchmark are 97%, 97%, and 96.99%, respectively, which demonstrate the good prediction performance of the proposed research.

Figure 17 demonstrates the confusion matrix computed while using Corel-1K image benchmark. It can be seen that all classes are correctly classified except for African, Beach, and Mountain. The major confusion exists between categories African and Beach, since similar objects can be observed between both classes.

5.5. Results for Corel-1.5K Image Dataset. Table 10 shows the experimental results for Corel-1.5K image benchmark. The numerical values presented in this table show that the classification accuracy obtained from the proposed fine-tuned ResNet50 is higher than the research based on hybrid feature techniques. The proposed research based on ResNet50 achieves 33.2% higher accuracy as compared to SIFT [14], 27.6% higher accuracy as compared to HOG [14],

Table 4: A quantitative comparison with recently published research in terms of classification accuracy for SIRI-WHU image benchmark.

Name of algorithm/model	Classification accuracy (%)
POVH [11]	80.14
VGGNet [46]	86.6
Inception-V3 [46]	89
GoogLeNet [46]	89.3
AlexNet [46]	90.2
ResNet50	94.03

Table 5: Classwise performance for SIRI-WHU image benchmark.

Class name	Precision (%)	Recall (%)	*F*-score (%)
Agriculture	100	100	100
Commercial	95	95	95
Harbor	95	95	95
Idle land	88.3	91.38	89.83
Industrial	95	96.61	95.8
Meadow	86.67	88.14	87.39
Overpass	98.33	90.77	94.4
Park	88.33	96.36	92.17
Pond	98.33	85.5	91.47
Residential	96.67	95.08	95.87
River	88.33	98.15	92.98
Water	98.33	98.33	98.33
Average	94.03	94.19	94.02

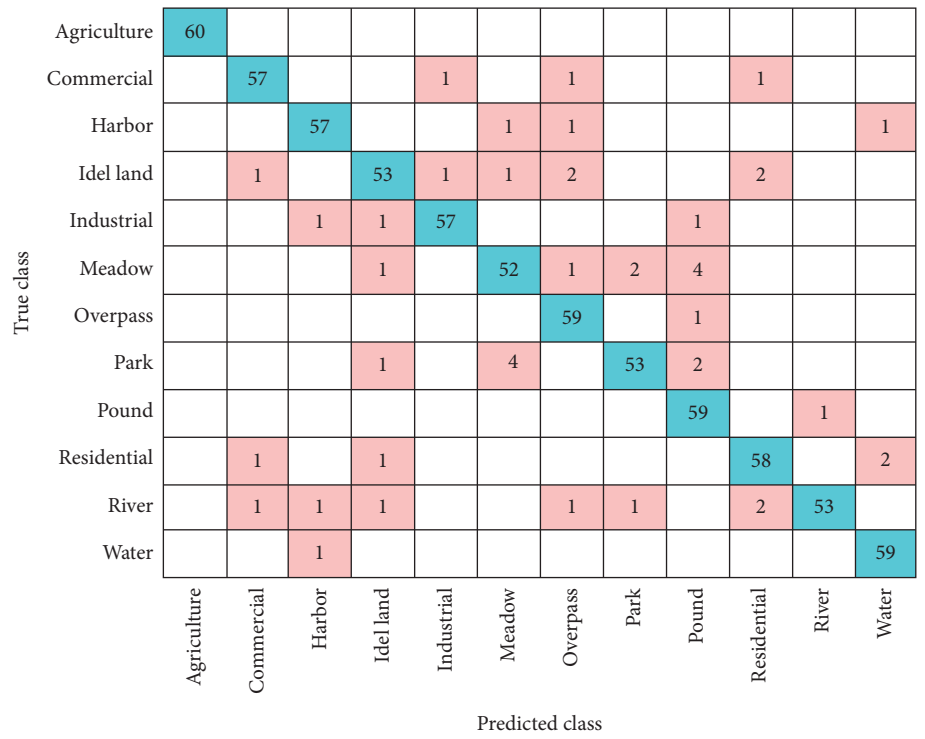

Figure 15: Confusion matrix for SIRI-WHU image dataset.

Table 6: A quantitative comparison with recently published research in terms of classification accuracy for UCM dataset.

Name of algorithm/model	Classification accuracy (%)
Inception-V3 [46]	91.1
Feature$_{RCG}$SVM [60]	93.81
AlexNet [46]	94.2
GoogLeNet [39]	94.31
CaffeNet [39]	95.02
VGG-VD-16 [39]	95.21
ResNet50	97.78

and 18.41% higher accuracy as compared to the approach presented in [14] and outperforms [40] by 0.66%. Hence, it can be safely concluded that the proposed research based on ResNet50 provides better performance for scene classification as compared to the related state-of-the-art research.

Table 11 provides classwise comparison of precision, recall, and *F*-score for Corel-1.5K image benchmark. The average precision, recall, and *F*-score for the Corel-1.5K image dataset are 99.56%, 99.78%, and 99.66%, respectively. High precision depicts a low false positive rate, and high recall depicts a low false negative rate. A good *F*-score is indicative of low false positives and low false negatives, as well as the capability of the model to correctly identify instances. An *F*-score of 1 is considered perfect, while an *F*-score of 0 indicates that the model is a total failure.

Figure 18 demonstrates the confusion matrix for Corel-1.5K image benchmark. Here, we can see that almost all classes are correctly classified with only one misclassified instance in each of categories Africa and Model.

Table 7: Classwise performance for UCM image benchmark.

Class name	Precision (%)	Recall (%)	*F*-score (%)
Agriculture	100	100	100
Airplane	100	100	100
Baseball diamond	100	100	100
Beach	100	100	100
Building	100	85.71	92.31
Chaparral	100	100	100
Dense residential	90	100	94.74
Forest	100	96.77	98.36
River	100	96.77	100
Freeway	100	100	100
Golf course	100	100	100
Harbor	96.67	93.55	95.08
Intersection	96.67	87.88	92.06
Mobile home parks	100	100	100
Medium residential	96.67	100	98.31
Sparse residential	100	100	100
Overpass	96.67	100	98.31
Parking lot	100	100	100
River	100	93.75	96.77
Storage tanks	80	100	88.89
Tennis court	96.67	100	98.31
Average	97.78	97.83	97.77

5.6. Time Performance Analysis. Besides classification accuracy, time performance analysis of the proposed system is an important parameter to be considered to determine its efficiency. Here, the time analysis is done during testing the model which is based on the testing time of the complete proposed model. Figure 19 shows the time comparison for

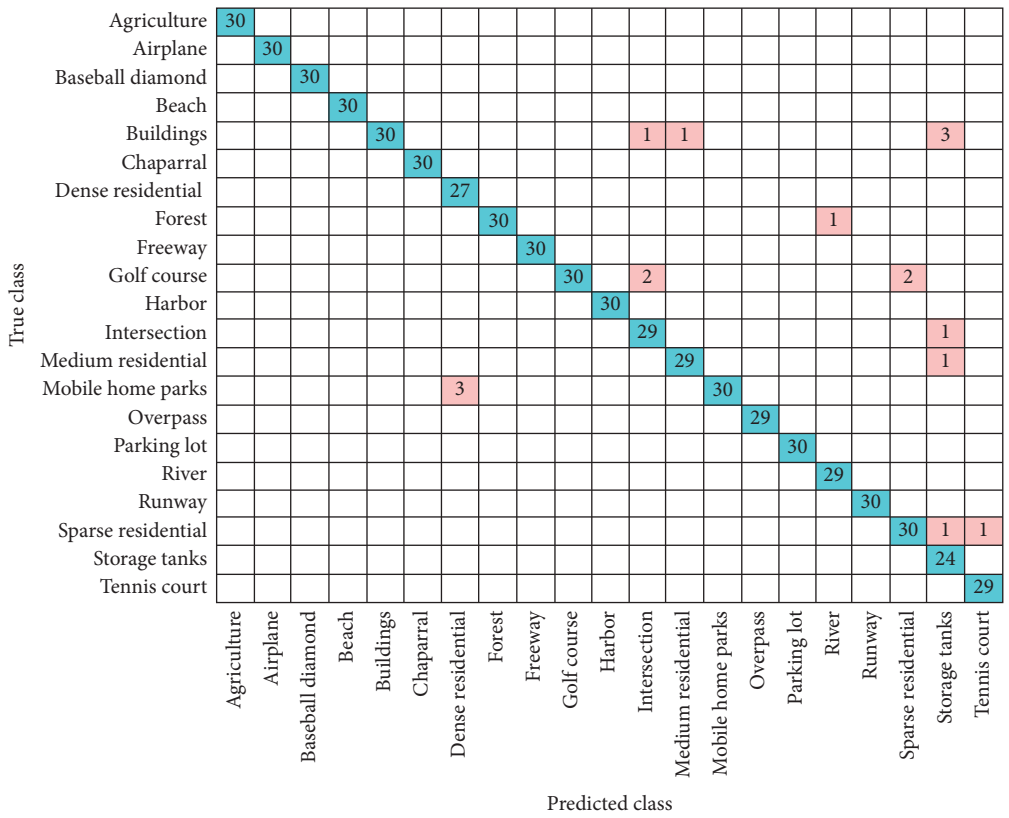

Figure 16: Confusion matrix for UCM image dataset.

Table 8: A quantitative comparison with recently published research in terms of classification accuracy for Corel-1K image benchmark.

Name of algorithm/model	Classification accuracy (%)
Li et al. [61]	70.84
Aslam et al. [14]	81.26
SCNN-ELM [61]	84.32
MKSVM-MIL et al. [62]	85.2
Raja et al. [41]	86.66
Desai et al. [42]	88.2
Yu et al. [44]	95.98
Shikha et al. [43]	96.5
ResNet50	97

Table 9: Classwise performance for Corel-1K image benchmark.

Class name	Precision (%)	Recall (%)	*F*-score (%)
African	90	90	90
Beach	86.67	86.67	86.67
Building	100	100	100
Bus	100	100	100
Dinosaur	100	100	100
Elephant	100	100	100
Flower	100	100	100
Food	100	96.77	98.36
Horse	100	100	100
Mountain	93.3	96.55	94.92
Average	97	97	96.99

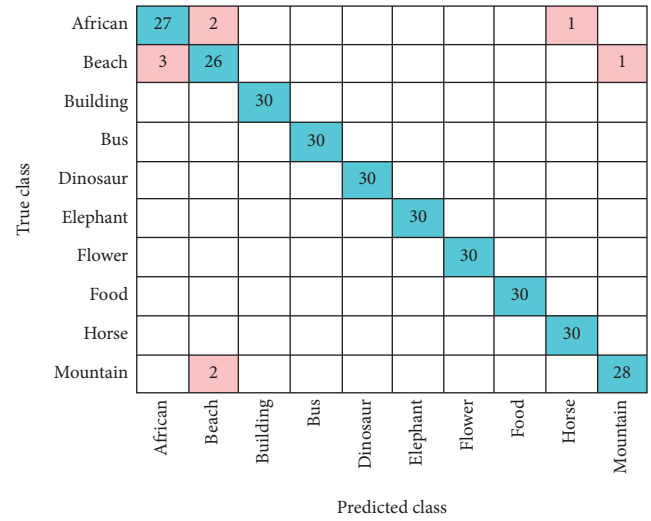

Figure 17: Confusion matrix for Corel-1K image benchmark.

Table 10: A quantitative comparison with recently published research in terms of classification accuracy for Corel-1.5K image benchmark.

Name of algorithm/model	Classification accuracy (%)
Aslam et al. [14]	66.36
Aslam et al. [14]	71.69
Aslam et al. [14]	81.15
Khalid et al. [40]	98.9
ResNet50	99.56

Table 11: Classwise performance for Corel-1.5K image dataset.

Class name	Precision (%)	Recall (%)	*F*-score (%)
African	96.67	96.67	96.67
Beach	100	100	100
Building	100	100	100
Bus	100	100	100
Cave	100	100	100
Dinosaur	100	100	100
Elephant	100	100	100
Flower	100	100	100
Food	100	100	100
Horse	100	100	100
Model	96.67	100	98.31
Mountain	100	100	100
Painting	100	100	100
Sunset	100	100	100
Tiger	100	100	100
Average	99.56	99.78	99.66

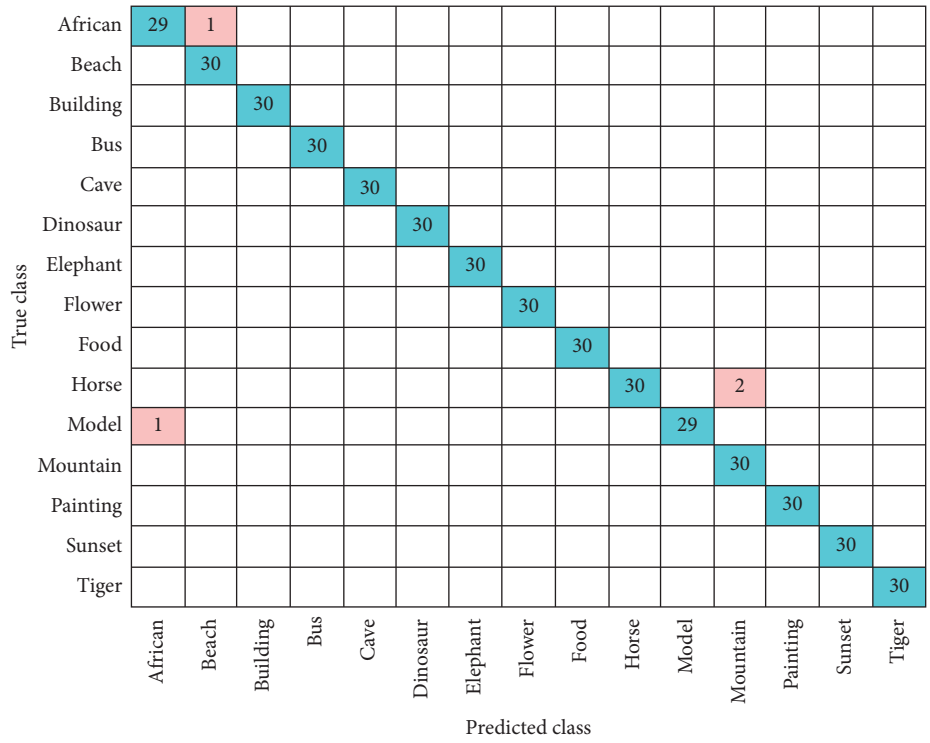

Figure 18: Confusion matrix for Corel-1.5K image dataset.

all the image datasets used for experimentation representing time per image, time per class, and time for the entire image dataset. From Figure 19, it can be deduced that, with the increase in number of images or with the data being more complicated, the time utilized for testing the model increases. Hence, it can be concluded that the training time is directly proportional to the size of the image datasets. Table 12 shows the time comparison of the proposed approach with the state-of-the-art research in terms of time per image for classification. It can be evidently seen that the proposed approach is computationally efficient as compared to the state-of-the-art research.

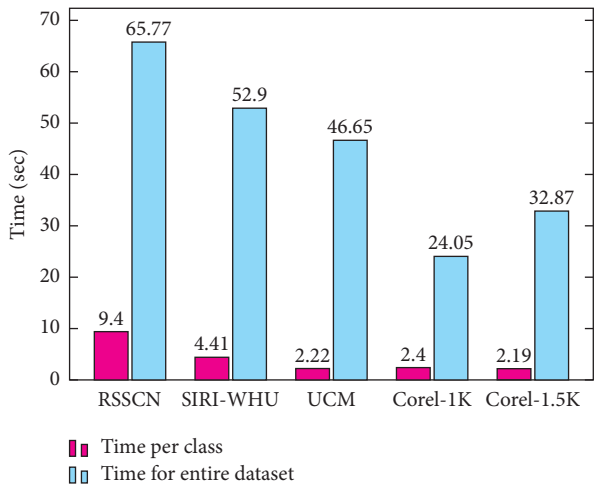

Figure 19: Time comparison for each image benchmark.

Table 12: Classification with respect to run-time analysis.

UCM dataset	
Proposed	0.0740
Pretrained CNN with SVM [63]	7.76
Pretrained CNN with ELM [63]	0.89
Corel-1K dataset	
Proposed	0.08013
RSHD [64]	0.3750
EODH [65]	5.6

6. Conclusion

Remote sensing, distant perceiving, image classification, and categorization are considered as challenging research areas in the field of computer vision. The recent focus of research in this domain is to explore the novel deep learning model that can enhance the classification accuracy. In this research article, we fine-tuned the ResNet50 by using network surgery and creation of network head along with the fine-tuning of hyperparameters. The learning of hyperparameters was tuned by using a linear decay learning rate scheduler known as piecewise scheduler. To tune the optimizer hyperparameter, Stochastic Gradient Descent with Momentum (SGDM) was used with the usage of weight learn and bias learn rate factor. Experiments and analysis were conducted on five different datasets, that is, UC Merced Land Use Dataset (UCM), RSSCN (the remote sensing scene classification image dataset), SIRI-WHU, Corel-1K, and Corel-1.5K. The analysis and competitive results exemplified that our proposed image classification-based model can classify the images in a more effective and efficient manner as compared to the state-of-the-art research. The overall performance of any deep learning model is dependent on the availability of training samples. In the future, we aim to explore an efficient ResNet50 when there are a less number of training samples available. Most of the deep network models are trained while using natural images such as ImageNet, while remote sensing images are different from natural images as they are acquired from different remote sensors. To explore transfer learning while using a combination of natural images and remote sensing images is another possible future research direction.

Data Availability

The details about the data used are included within this manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Shakya, "Analysis of artificial intelligence based image classification techniques," *Journal of Innovative Image Processing (JIIP)*, vol. 2, no. 1, pp. 44–54, 2020.

[2] A. Shabbir, A. Rasheed, A. Rasheed et al., "Detection of glaucoma using retinal fundus images: a comprehensive review," *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2033–2076, 2021.

[3] A. Rasheed, B. Zafar, A. Rasheed et al., "Fabric defect detection using computer vision techniques: a comprehensive review," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8189403, 24 pages, 2020.

[4] H. Doreswamy, M. K. Hooshmand, and I. Gad, "Feature selection approach using ensemble learning for network anomaly detection," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 4, pp. 283–293, 2020.

[5] M. Sajid, N. Ali, N. I. Ratyal, S. H. Dar, and B. Zafar, "Facial asymmetry-based Feature extraction for different applications: a review complemented by new advances," *Artificial Intelligence Review*, pp. 1–41, 2021.

[6] M. Sajid, N. Ali, S. H. Dar, B. Zafar, and M. K. Iqbal, "Short search space and synthesized-reference re-ranking for face image retrieval," *Applied Soft Computing*, vol. 99, Article ID 106871, 2021.

[7] B. Petrovska, T. Atanasova-Pacemska, R. Corizzo, P. Mignone, P. Lameski, and E. Zdravevski, "Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing," *Applied Sciences*, vol. 10, no. 17, p. 5792, 2020.

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

[9] A. Latif, A. Rasheed, U. Sajid et al., "Content-based image retrieval and feature extraction: a comprehensive review," *Mathematical Problems in Engineering*, vol. 2019, Article ID 9658350, 21 pages, 2019.

[10] J. Zhang, W. Geng, X. Liang, J. Li, L. Zhuo, and Q. Zhou, "Hyperspectral remote sensing image retrieval system using spectral and texture features," *Applied Optics*, vol. 56, no. 16, pp. 4785–4796, 2017.

[11] N. Ali, B. Zafar, M. K. Iqbal et al., "Modeling global geometric spatial information for rotation invariant classification of satellite images," *PLoS One*, vol. 14, no. 7, Article ID e0219833, 2019.

[12] B. Zafar, R. Ashraf, N. Ali et al., "A novel discriminating and relative global spatial image representation with applications in CBIR," *Applied Sciences*, vol. 8, no. 11, p. 2242, 2018.

[13] N. Ali, B. Zafar, F. Riaz et al., "A hybrid geometric spatial image representation for scene classification," *PLoS One*, vol. 13, no. 9, Article ID e0203339, 2018.

[14] M. A. Aslam, M. N. Salik, F. Chughtai, N. Ali, S. H. Dar, and T. Khalil, "Image classification based on mid-level feature fusion," in *Proceedings of the 2019 15th International Conference on Emerging Technologies (ICET)*, pp. 1–6, IEEE, Peshawar, Pakistan, December 2019.

[15] M. Sajid, N. Ali, N. I. Ratyal et al., "Deep learning in age-invariant face recognition: a comparative study," *The Computer Journal*, 2020.

[16] M. Sajid, N. Ali, S. H. Dar et al., "Data augmentation-assisted makeup-invariant face recognition," *Mathematical Problems in Engineering*, vol. 2018, Article ID 2850632, 10 pages, 2018.

[17] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, p. e1264, 2018.

[18] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: a review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

[19] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: an overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, 2017.

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

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: speeded up robust features," in *Proceedings of the European Conference on Computer Vision*, pp. 404–417, Springer, Graz, Austria, May 2006.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection,"vol. 1, pp. 886–893, in *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, San Diego, CA, USA, June 2005.

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[25] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[26] W. Shao, W. Yang, G.-S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Proceedings of the International Conference on Computer Vision Systems*, pp. 324–333, Springer, Sydney, Australia, 2013.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[28] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: a state-of-the-art review," *Remote Sensing*, vol. 12, no. 9, p. 1444, 2020.

[29] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/14091556.

[32] 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*, pp. 770–778, Las Vegas, NV, USA, 2016.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, San Francisco, CA, USA, 2017.

[34] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, 2015.

[35] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia Computer Science*, vol. 132, pp. 377–384, 2018.

[36] S. Fatima, N. Aiman Aslam, I. Tariq, and N. Ali, "Home security and automation based on internet of things: a comprehensive review," *IOP Conference Series: Materials Science and Engineering*, vol. 899, Article ID 012011, 2020.

[37] B. Zafar, R. Ashraf, N. Ali et al., "Intelligent image classification-based on spatial weighted histograms of concentric circles," *Computer Science and Information Systems*, vol. 15, no. 3, pp. 615–633, 2018.

[38] C. Zhu, W. Yan, X. Cai, S. Liu, T. H. Li, and G. Li, "Neural saliency algorithm guide bi-directional visual perception style transfer," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 1–8, 2020.

[39] G.-S. Xia, J. Hu, F. Hu et al., "AID: a benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[40] M. J. Khalid, M. Irfan, T. Ali et al., "Integration of Discrete wavelet Transform, DBSCAN, and classifiers for efficient content based image retrieval," *Electronics*, vol. 9, no. 11, p. 1886, 2020.

[41] R. Raja, S. Kumar, and M. R. Mahmood, "Color object detection based image retrieval using ROI segmentation with multi-feature method," *Wireless Personal Communications*, vol. 112, no. 1, pp. 169–192, 2020.

[42] P. Desai, J. Pujari, C. Sujatha et al., "Impact of multi-feature extraction on image retrieval and classification using machine learning technique," *SN Computer Science*, vol. 2, no. 3, pp. 1–9, 2021.

[43] B. Shikha, P. Gitanjali, and D. P. Kumar, "An extreme learning machine-relevance feedback framework for enhancing the accuracy of a hybrid image retrieval system," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 2, 2020.

[44] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2431–2442, 2014.

[45] Y. Yang, L. Juntao, and P. Lingling, "Multi-robot path planning based on a deep reinforcement learning DQN

algorithm," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 3, pp. 177–183, 2020.

[46] M. A. Shafaey, M. A. M. Salem, H. Ebeid, M. Al-Berry, and M. F. Tolba, "Comparison of CNNs for remote sensing scene classification," in *Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pp. 27–32, IEEE, Cairo, Egypt, December 2018.

[47] X. Zhao, J. Zhang, J. Tian, L. Zhuo, and J. Zhang, "Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image," *Remote Sensing*, vol. 12, no. 11, p. 1887, 2020.

[48] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, Phoenix, AZ, USA, 2016.

[49] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral-spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.

[50] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[51] F. Zhuang, Z. Qi, K. Duan et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[52] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, IEEE, Santa Rosa, CA , USA, 2017.

[53] Z. Xu, A. M. Dai, J. Kemp, and L. Metz, "Learning an adaptive learning rate schedule," 2019, http://arxiv.org/abs/190909712.

[54] J. Park, D. Yi, and S. Ji, "A novel learning rate schedule in optimization for neural networks and it's convergence," *Symmetry*, vol. 12, no. 4, p. 660, 2020.

[55] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, "Diffgrad: an optimization method for convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4500–4511, 2019.

[56] B. Zhao, Y. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 2108–2123, 2015.

[57] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, ACM, 2010.

[58] J. Jia Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 985–1002, 2008.

[59] 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, no. 11, pp. 2321–2325, 2015.

[60] M. Wang, X. Zhang, X. Niu, F. Wang, and X. Zhang, "Scene classification of high-resolution remotely sensed image based on resnet," *Journal of Geovisualization and Spatial Analysis*, vol. 3, no. 2, pp. 1–9, 2019.

[61] D. Li, X. Qiu, Z. Zhu, and Y. Liu, "Criminal investigation image classification based on spatial CNN features and ELM,"vol. 2, pp. 294–298, in *Proceedings of the 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2, , IEEE, Hangzhou China, 2018.

[62] D. Li, J. Wang, X. Zhao, Y. Liu, and D. Wang, "Multiple kernel-based multi-instance learning algorithm for image classification," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1112–1117, 2014.

[63] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 704–708, 2017.

[64] S. R. Dubey, S. K. Singh, and R. K. Singh, "Rotation and scale invariant hybrid image descriptor and retrieval," *Computers & Electrical Engineering*, vol. 46, pp. 288–302, 2015.

[65] X. Tian, L. Jiao, X. Liu, and X. Zhang, "Feature integration of EODH and Color-SIFT: application to image retrieval based on codebook," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 530–545, 2014.