



Compact Deep Color Features for Remote Sensing Scene Classification

Rao Muhammad Anwer^{1,2} · Fahad Shahbaz Khan¹ · Jorma Laaksonen²

Accepted: 8 February 2021 / Published online: 27 February 2021
© The Author(s) 2021

Abstract

Aerial scene classification is a challenging problem in understanding high-resolution remote sensing images. Most recent aerial scene classification approaches are based on Convolutional Neural Networks (CNNs). These CNN models are trained on a large amount of labeled data and the *de facto* practice is to use RGB patches as input to the networks. However, the importance of color within the deep learning framework is yet to be investigated for aerial scene classification. In this work, we investigate the fusion of several deep color models, trained using color representations, for aerial scene classification. We show that combining several deep color models significantly improves the recognition performance compared to using the RGB network alone. This improvement in classification performance is, however, achieved at the cost of a high-dimensional final image representation. We propose to use an information theoretic compression approach to counter this issue, leading to a compact deep color feature set without any significant loss in accuracy. Comprehensive experiments are performed on five remote sensing scene classification benchmarks: UC-Merced with 21 scene classes, WHU-RS19 with 19 scene types, RSSCN7 with 7 categories, AID with 30 aerial scene classes, and NWPU-RESISC45 with 45 categories. Our results clearly demonstrate that the fusion of deep color features *always* improves the overall classification performance compared to the standard RGB deep features. On the large-scale NWPU-RESISC45 dataset, our deep color features provide a significant absolute gain of 4.3% over the standard RGB deep features.

Keywords Remote sensing · Deep learning · Scene classification · Color features · Feature compression

1 Introduction

Remote sensing scene classification is a challenging research problem, where the task is to associate each aerial image, comprising a variety of land cover types and ground objects,

✉ Rao Muhammad Anwer
rao.anwer@inceptioniai.org

¹ Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

² Department of Information and Computer Science, Aalto University School of Science, Espoo, Finland

with its respective semantic scene category. The problem is important for understanding remote sensing image data and has many potential applications, including disaster monitoring, vegetation mapping, land resource management, urban planning, traffic supervision, and environmental monitoring. Most previous methods either rely on employing hand-crafted visual features [13,24,47,77], such as color and shape, or using mid-level holistic image representations [11,33,44,75,76] constructed by encoding hand-crafted visual features. Recently, deep Convolutional Neural Networks (CNNs) have revolutionized computer vision by significantly advancing the state-of-the-art in many areas such as, image classification [8,20,31,38,43,64], object detection/segmentation [19,28,30,40,53,57,68,69,81] and action recognition [26,48,51,63]. Similarly, deep learning techniques have also made an impact on satellite image analysis, including aerial scene classification [2,32,54,73] and hyperspectral image analysis [14,15,66].

Generally, deep convolutional neural networks or CNNs take a fixed-sized image as input to a series of convolution, local normalization and pooling operations (termed as layers). The final layers of the convolutional neural network are fully connected (FC), and are typically used to extract deep features that are generic and used for a variety of vision applications [5], including remote sensing scene classification [54,73]. The standard input to a deep convolutional neural network is *RGB* pixel values, with training performed on the large-scale ImageNet dataset. Most existing remote sensing scene categorization approaches employ these CNNs, pre-trained on the ImageNet dataset, as feature extractors. The exploration of different color spaces and their combination for remote sensing scene classification is still an open research problem. The work of [56] investigated different color spaces for vehicle color identification. The work of [23] explored YCbCr and RGB color channels for image super resolution. A collaborative facial color feature learning approach, combining multiple color spaces, was proposed by [41] for face recognition. Here, we investigate a variety of color features within a deep learning framework for remote sensing scene classification.

Prior to deep learning, the impact of multiple color features was well studied for object recognition and detection [3,4,39,67]. The work of [67] studied the invariance properties of different color descriptors and showed that different color features are complementary and their combination provides a consistent improvement in overall classification performance. Khan et al. [39] proposed an attention based framework to combine multiple color features with shape features. Within the deep learning framework, the complementary characteristics of these color features are yet to be investigated for remote sensing scene classification. Figure 1 shows visualizations of filter weights from the first convolutional layer of CNNs, trained from scratch on ImageNet, using different color features. Visualizations are shown for RGB, Opponent, YCbCr, and Lab based CNNs. The visualizations show that the corresponding activation maps for different color spaces have different responses due to variations in filter weights. Further, the trained image representations in different color features based CNNs are represented in different feature subspaces that likely provide complementary information. In this work, we investigate the effectiveness of combining multiple color features, within a deep learning framework, for remote sensing scene classification.

As discussed above, the common strategy employed by most remote sensing scene classification methods is to extract deep features from the activations of the FC layers of a pre-trained deep convolutional neural network. However, such a strategy will encounter the inherent problem of a high-dimensional final image representation when combining activations from multiple deep color convolutional neural networks. Therefore, it is desired to obtain a compact final image representation without sacrificing the improvements obtained from the complementary characteristics of multiple deep color features. Recently, Khan et al. [37] proposed to use a divisive information theoretic clustering (DITC) technique [22] to

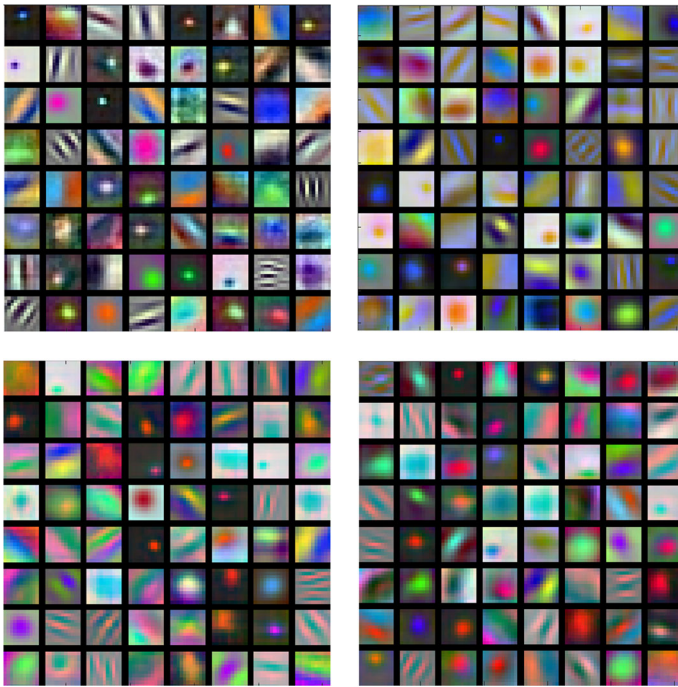


Fig. 1 Visualization of filter weights from different color features. Top row: RGB space-based deep convolutional neural network (left) and Opponent color space-based deep convolutional neural network (right). Bottom row: YCbCr color space-based deep convolutional neural network (left) and Lab color space-based deep convolutional neural network (right). All networks employ the same architecture. (Color figure online)

combine heterogeneous texture descriptors for texture classification. Their work showed that a notable reduction in the dimensionality of the final image representation can be obtained using the DITC technique, without significant loss in classification performance. Motivated by this, we propose to use the DITC technique to compress the dimensionality of a multi-color deep representation for remote sensing scene classification. The DITC approach has previously been employed to compress the high-dimensionality of bag-of-words based spatial pyramids and hand-crafted heterogeneous texture representations [25,37]. To the best of our knowledge, we are the first to investigate the DITC technique to compress deep multi-color image representations for scene classification in remote sensing images.

Contributions In this work, we study the problem of remote sensing scene classification with the following contributions.

- We investigate the contribution of color in a deep learning framework for scene classification in remote sensing images. We further demonstrate the effectiveness of combining multiple color features within the deep learning framework. Furthermore, we propose the usage of an information theoretic compression approach to compress high-dimensional multi-color deep features into a compact image representation.
- Comprehensive experiments are performed on several remote sensing datasets: UC-Merced with 21 scene classes, WHU-RS19 with 19 scene types, RSSCN7 with 7 categories, AID with 30 aerial scene classes, and NWPU-RESISC45 with 45 categories. The results of our experiments clearly demonstrate that combining multi-color deep features significantly improves the classification performance compared to standard RGB

deep features alone. Furthermore, our results show that multi-color deep features can be efficiently compressed without any significant loss in classification performance. Finally, our compact multi-color deep features provide competitive classification results compared to existing remote sensing image classification approaches in the literature.

The rest of this paper is organized as follows. We present related work in Sect. 2. Our method is described in detail in Sect. 3. We present the remote sensing scene classification experiments and results in Sect. 4. Finally, our conclusions are drawn along with potential future research directions in Sect. 5.

2 Related Work

The impact of color features for remote sensing scene analysis has been extensively studied [11,24,59,65,76,78]. The work of [76] investigated integrating color information within Gabor features for remote sensing scene classification. dos Santos et al. [24] evaluated a variety of hand-crafted color and texture feature description approaches for remote sensing image classification and retrieval. Chen et al. [11] performed an evaluation of local features, such as structure, texture and color for remote sensing scene classification. The work of [65] studied a variety of feature description baselines in different color spaces for remote sensing images.

Combining multiple hand-crafted color features has also been investigated in the literature [39,60,67]. The work of [67] investigated integrating color and shape features, within the bag-of-words framework, for object recognition. Their evaluation recommends employing opponent color features with SIFT descriptor for object recognition and also showed the importance of fusing multiple color representations to achieve further improvement in classification performance. Khan et al. [39] proposed an approach where multiple hand-crafted color features are employed to modulate shape features for object recognition. The work of [1] proposed an approach to combine color models that are learned to achieve color invariance for object detection. The work of [21] investigate the impact of color information for joint set registration. Different to these previous works using hand-crafted features, recent works have also investigated combining multiple color features within a deep learning framework for face recognition [41], image super resolution [23] and vehicle classification [56]. However, to our knowledge, the effectiveness of combining multiple color features within a deep learning framework is yet to be investigated for remote sensing scene analysis.

In recent years, several deep learning-based approaches have been introduced for remote sensing scene classification. The work of [54] evaluated off-the-shelf CNNs features and compared their performance with low-level descriptors for remote sensing scene classification. Marmanis et al. [49] proposed a two-stream approach where pre-trained CNNs features are first used to represent the images. Then, the extracted representations were input to shallow CNN classifiers. The work of [72] introduced a hybrid architecture where multi-column stacked denoising sparse auto-encoder is combined with Fisher vector to learn features in a hierarchical manner for land use scene classification. Yan et al. [74] proposed an approach based on improved category-specific codebook using kernel collaborative representation based classification which is integrated with Spatial pyramid matching. Their approach then employed SVM classifier to classify remote sensing images. The work of [73] introduced a large-scale remote sensing scene classification dataset and also evaluated several pre-trained CNNs on their large dataset. Cheng et al. [17] introduced a method based on bag of convolutional features where CNNs features are employed in place of hand-crafted local features

to construct bag-of-words based image representation. The work of [2] investigated a fusion approach where standard RGB deep features are combined with LBP-based deep texture features to classify remote sensing images. Chen et al. [10] proposed a CNN based approach under the guidance of a human visual attention mechanism. In their approach, a computational visual attention model is utilised to extract salient regions. Afterwards, sparse filters are employed for learning features from extracted regions. A superpixel-guided layer-wise embedding CNN based approach is introduced by [46] to exploit information from both labeled and unlabeled examples. The work of [29] introduced a center-based structured metric learning approach where both deep metrics and center points are taken into account to penalize pairwise correlation and class-wise information between categories. Most of these approaches employ CNN models trained using RGB patches as input. Different to these approaches based on the de facto practice of using RGB patches for CNN training, we investigate the contribution color in deep learning framework and demonstrate the effectiveness of integrating multiple deep color features. Our extensive experiments on five benchmarks demonstrate the effectiveness of combining multiple deep color features for remote sensing scene classification.

3 Our Approach

Here, we first discuss the motivation behind the proposed approach. Then, we describe how the deep color models are constructed. Afterwards, we investigate the fusion of deep color features and counter the problem of the high-dimensionality of fused deep color features for classification.

Motivation As discussed earlier, most existing state-of-the-art remote sensing image classification methods are based on CNNs. Here, CNNs are generally pre-trained on a large-scale generic object recognition dataset (ImageNet) using raw RGB pixel values as an input. Previously, combining multiple hand-crafted local color features have been investigated, within the bag-of-words framework, for object recognition. Motivated by these previous works, we investigate the contribution of color within a deep learning framework (CNNs) and demonstrate the impact of integrating multiple deep color features for remote sensing scene analysis. To the best of our knowledge, we are the first to investigate the effectiveness of integrating multiple color features, within a deep learning framework (CNNs), for remote sensing scene classification.

3.1 Deep Color Convolutional Neural Networks

Most existing state-of-the-art remote sensing image classification methods are based on deep models. These deep models are generally pre-trained on a large-scale generic object recognition dataset (ImageNet) using raw RGB pixel values as an input. Here, we analyze a variety of color features within deep learning framework to evaluate their impact on remote sensing scene classification. We investigate the contribution of color in a standard deep convolutional neural network (CNN) architecture [8,43,64].

To analyze the impact of color for remote sensing image classification, we employ a variety of color features popular in object recognition. The selected color features are based on different color space transformations: *HSV*, *YCbCr*, *Opponent*, *C*, *Lab*, and the *colornames*. Furthermore, the motivation of these color representations differ from photometric invariance to discriminative power.

RGB In this work, we use the standard three-channel RGB color space as the baseline.
HSV In the HSV color space, the H model is scale-invariant. Further, it is shift-invariant with respect to light intensity [67]. The HSV color space has been previously investigated with the hand-crafted SIFT descriptor for scene recognition [7].

YCbCr In the YCbCr color space, Y is the luminance component and Cb and Cr are the blue-difference and red-difference chroma components. This color space is approximately perceptually uniform and has been used previously for remote sensing images [34,65].

Lab The three dimensions of the Lab color space correspond to L for lightness and a and b for the color components green–red and blue–yellow. This color space is perceptually uniform implying that colors at an equal distance are also perceptually equally far apart.

Opponent In this color space, the O_1 and O_2 channels encode the color information in the image. The O_3 channel describes the intensity information. The image is transformed as in [45]:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{1}$$

The opponent color representation possesses invariance with respect to specularities. In the evaluation performed by van de Sande et al. [67], the opponent color space in conjunction with the hand-crafted SIFT feature descriptor, was shown to provide improved results for visual object recognition.

C The C representation, defined as $C = \left(\frac{O_1}{O_3} \frac{O_2}{O_3} O_3 \right)^T$, aims at adding photometric invariance with respect to shadow-shading to the opponent representation. The invariance is achieved by normalizing the first two dimensions with the luminance channel O_3 . The C representation was initially proposed by [27] and later employed with the SIFT descriptor by [67].

Color names Most of the aforementioned color representations aim at employing specific photometric invariance properties. Different to these representations, the color names are linguistic color labels assigned by humans to represent world colors. It involves the assignment of linguistic color labels to image pixels. A linguistic study by [6] identified that the 11 basic color terms of the English language are white, blue, grey, brown, orange, green, red, black, purple, yellow, and pink. The work of [70] proposed an approach to automatically learn from images retrieved with Google-image search. The descriptor is based on the 11 basic color terms. Color names representation [70] CN is defined as a feature vector comprising the probability of a color name for an image *Img*:

$$CN = \{p(cns_1|Img), p(cns_2|Img), \dots, p(cns_{11}|Img)\} \tag{2}$$

with

$$p(cns_j|Img) = \frac{1}{P} \sum_{x,y \in Img} p(cns_j|\mathbf{f}(x,y)) \tag{3}$$

here cns_j is the *j*-th color name, $\mathbf{f} = \{L^*a^*, b^*\}$ and x, y are the spatial coordinates of the P pixels in the image *Img*, Further, $p(cns_j|\mathbf{f})$ is the probability of a color name given a pixel value in the image, computed from an image dataset collected from Google. Since the images

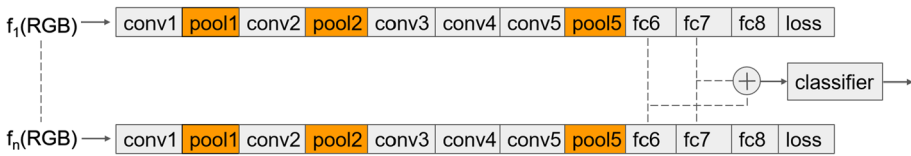


Fig. 2 The proposed color fusion in a deep convolutional neural network architecture. Here, the function $f(RGB)$ denotes a RGB pixel values mapping to another color-space representation. Deep color models are trained separately, from scratch, using different color spaces. We extract activations from FC6 and FC7 layers of each deep color network, respectively. These activations are then concatenated (indicated by a ‘plus’) to be used as image features and input to a classifier. See text for more details. (Color figure online)

are acquired from the web, the issue of retrieving noisy images is addressed by employing PLSA approach [70]. Figure 2 shows the proposed color fusion in a deep convolutional neural network architecture. We use same architecture to train all deep color convolutional neural networks. Each deep color network is trained separately. The details of the underlying network architecture is provided in Sect. 4.1.

3.2 Compact Deep Color Features

As discussed earlier, the final layers of the deep convolutional neural network (FC layers) are typically employed to extract deep features since they are generic and previously used for a variety of vision applications [5], including remote sensing scene classification [54,73]. Here, we extract 4096 dimensional activations from the FC7 (second last) and FC6 (third last) layers of each deep color network respectively. These activations are then concatenated and used as image features, $D = [d_{c1}, d_{c2}, d_{c3}, \dots, d_{cn}]$. However, the combination of these activations from multiple deep color convolutional neural networks has the disadvantage of being high-dimensional (more than 57k in size) for each image. Here, we propose to use an information theoretic compression approach (DITC) [22] to compress the high-dimensional multi-color deep representation. The DITC algorithm works by discriminatively learning a pre-determined compact representation by minimizing the loss in mutual information between clusters and the class labels of training samples. The approach operates on the class-conditional distributions over deep multi-color image representations. The class-conditional estimation is measured by the probability distributions $p(R|d)$, where $R = \{r_1, r_2, \dots, r_{CL}\}$ is the set of CL classes. The approach then estimates the drop in mutual information MI between the combined deep color representation D and the category labels R . The high-dimensional deep multi-color image representation is then transformed to a compact representation $D^R = \{D_1, D_2, \dots, D_J\}$ (where each D_j represents a collection of bins in the original uncompressed high-dimensional representation) as

$$\begin{aligned} \Delta MI &= MI(R; D) - MI(R; D^R) \\ &= \sum_{j=1}^J \sum_{d \in D_j} p(d) KL(p(R|d), p(R|D_j)), \end{aligned} \tag{4}$$

where KL is the Kullback-Leibler (KL) divergence. The Kullback-Leibler divergence between the two distributions is defined by

$$KL(p_1, p_2) = \sum_{y \in Y} p_1(y) \log \frac{p_1(y)}{p_2(y)}. \quad (5)$$

It is worth mentioning that the category information is exploited using only the training samples. The high-dimensional deep multi-color image representation is compressed by merging the bins, over the classes, with similar discriminative powers. We refer to [22] for additional details of the DITC algorithm.

4 Experimental Results

4.1 Experimental Setup

We first describe the underlying deep convolutional neural network architecture employed to obtain our deep color models. The deep convolutional neural network is based on the VGG architecture and is similar to [43]. The deep convolutional neural network consists of 5 convolutional layers ($C1, C2, C3, C4,$ and $C5$) and 3 fully-connected (FC) layers ($FC6, FC7$ and $FC8$). The deep convolutional neural network takes as input an image of size 224×224 . Throughout our experiments, images are resized to 224×224 pixels and then input to the network. The first convolutional layer $C1$ contains 64 convolutional filters with a filter size of 11×11 . The convolution stride is set to 4 and a max-pooling downsampling factor of 2. The second convolutional layer $C2$ comprises of 256 convolutional filters with a filter size of 5×5 . The convolution stride is set to 1, spatial padding is set to 2, and a max-pooling downsampling factor of 2. For the third, fourth and fifth convolutional layer $C3, C4$ and $C5$, the number of convolutional filters is 256, filter size is 3×3 , and the convolution stride and spatial padding are 1. For the fifth convolutional layer, a max-pooling downsampling factor of 2 is employed. Furthermore, the first two FC layers ($FC1$ and $FC2$) are regularised using dropout [43] with dropout ratio set to 0.5. Consequently, the last FC layer $FC3$ is a multi-class soft-max classifier. Other than the FC3 layer, the activation function for the rest of the weight layers is the Rectified Linear Unit (ReLU) [35,43,50].

We train all deep color convolutional neural networks, described in Sect. 3.1, from scratch on the ImageNet 2012 training set. We use the same set of hyper-parameters as in [43] during network training in our experiments. For all CNNs training, the learning rate is set to 0.001 and momentum is set to 0.9. The initial learning rate is decreased by a factor of 10, in case the validation error stops to decrease further. We initialize the layers from a Gaussian distribution with a zero mean and variance equal to 10^{-2} . A similar data augmentation, as in [8], in the form of random crops, horizontal flips, and RGB color jittering is employed during training. For a fair comparison, we train the baseline standard RGB by increasing the depth of the network architecture with a factor of seven resulting in same number of network parameters as our color fusion. Furthermore, pre-trained deep color convolutional neural networks are employed as feature extractors by extracting 4096 dimensional activations from the FC7 and FC6 layers as image features. All the image features are L_2 -normalized and input to a one-versus-all linear SVM classifier.

Throughout all experiments, the classification results are reported in terms of average recognition accuracy over all scene categories in a remote sensing scene classification dataset. From the classifier, the scene class label providing the highest confidence is assigned to the test image. The overall recognition results are obtained by computing the average classification score over all scene categories in a remote sensing scene classification dataset. As in [16,73], each dataset is randomly split into training and test sets for performance evaluation. For



Fig. 3 Example images from the five high resolution remote sensing datasets used in our experiments. The top row, second row, third row, fourth row and bottom row contains example images from UC-Merced, WHU-RS19, RSSCN7, AID and NWPU-RESISC45, respectively

all datasets, the ratio of training to test images is set to 50:50, where images are randomly selected from each aerial scene category. To obtain a reliable performance comparison, we repeat the evaluation procedure ten times. The final classification results are then reported as the mean over these ten runs together with the standard deviation.

4.2 Datasets

We conduct experiments on multiple datasets (see Fig. 3).

UC-Merced is a commonly used remote sensing dataset [76] that is publicly available and comprises of 2100 images. There are 21 classes in this dataset. Some of the scene categories in this dataset are: agriculture, golf course, baseball diamond, dense residential, medium density residential, forest, river, sparse residential, overpass, parking lot, storage tanks, and tennis courts. The images in the dataset are cropped to 256×256 pixels and are collected from 20 regions across the USA.

WHU-RS19 is a public dataset [62] with 950 aerial images acquired from Google Earth imagery. There are 50 samples per scene class in the dataset whereas the images are of size 600×600 pixels. There are 19 aerial scene categories in this dataset. Some of the scene categories in the dataset are: airport, meadow, pond, parking, port, beach, bridge, river, railway station, viaduct, commercial area, desert, farmland, industrial area, and park. The dataset poses several challenges due to scale and illumination variations.

RSSCN7 is a dataset [84] with seven aerial scene categories: farmland, grassland forest, industrial region, lake, parking lot, residential region, and river. The dataset was released in 2015 and is publicly available. Each aerial scene category contains 400 images. The images in the dataset are of size 400×400 pixels with sampling performed at varying scales (four).

AID is a large scale public dataset [73] with 30 classes and 10,000 images. The dataset consists of 30 aerial scene categories. Some of the scene categories in the dataset are: playground, sparse residential, medium residential, bare land, center, desert, farmland, mountain, park, parking, forest, resort, school, church, square, river, storage tanks, and viaduct. The

images in the AID dataset are collected from different countries, including China, USA, Germany, France, UK, and Italy.

NWPU-RESISC45 is another large scale public dataset [16] with 31500 images having 700 images per category. The images in the dataset are of size 256×256 pixels. The dataset comprises 45 aerial scene categories where some of the classes in the dataset are: airplane, railway, railway station, bridge, church, stadium, sparse residential, forest, ship, terrace, freeway, storage tank, golf course, lake, ground track field, baseball diamond, mountain, parking lot, wetland, river, and roundabout.

Here, we evaluate different deep color features, described in Sect. 3.1, on a variety of five datasets. In all cases, we employ activations from the FC7 layer of the CNNs as deep color features. For the color fusion, we concatenate all the deep color features resulting in a 28672 dimensional feature vector. Table 1 shows the comparison of deep color features on the UC-Merced, WHU-RS19, RSSCN7, AID, and NWPU-RESISC45 datasets. On the UC-Merced dataset, the baseline approach provides a mean recognition rate of 94.7%. Image features from the color names and HSV based CNNs achieve mean classification scores of 93.7% and 93.8%, respectively. Deep features from the C and Lab color space based CNNs provide mean recognition scores of 93.6% and 93.9%, respectively. Deep features from the opponent color space-based deep convolutional neural network provides an average classification accuracy of 94.5%. Furthermore, the proposed deep color feature fusion significantly improves the classification performance, achieving a mean recognition score of 96.3%. The proposed deep color feature fusion provides an absolute gain of +1.6% in terms of classification performance compared to the baseline standard RGB deep features.

On the WHU-RS19 dataset, the baseline (RGB) network provides an average recognition rate of 96.0%. Deep features from the color names, C and HSV color space based CNNs achieve mean recognition scores of 95.1%, 94.7% and 94.4%, respectively. Furthermore, image features from the YCbCr and Lab color space based CNNs provide mean recognition rates of 95.4% and 95.0%, respectively. On this dataset, deep features from the opponent color space based deep convolutional neural network achieves similar performance with a mean classification score of 96.0%, compared to the baseline RGB features (96.0%). Moreover, the combined set of deep color features improves the classification performance with an absolute gain of +1.4%, compared the baseline standard RGB deep features. Similarly on the RSSCN7 dataset, deep features from the opponent color space-based deep convolutional neural network provide similar classification results with a mean recognition rate of 89.4%, compared to the baseline RGB features (89.5%). Furthermore, the classification performance is improved by employing the combined set of deep color features, which obtains an average recognition accuracy of 92.3%.

We also evaluate different deep color features on two recently introduced large scale AID and NWPU-RESISC45 datasets. On the AID dataset, the baseline standard RGB color space-based deep convolutional neural network achieves an average classification score of 90.3%. The deep features from most other color spaces provide slightly inferior results compared to the standard RGB. However, deep features from the opponent color space-based deep convolutional neural network again provide similar performance, with an average classification accuracy of 89.9%, compared to the baseline RGB features. Furthermore, the proposed deep color feature fusion significantly improves the classification performance, with an absolute gain of +3.1% in terms of classification performance, compared to the baseline standard RGB deep features. Finally on the NWPU-RESISC45 dataset, the baseline standard RGB deep network provides a mean recognition rate of 85.7%. Image features from the color names, HSV and C based deep convolutional neural networks (CNNs) obtain mean classification scores of 83.2%, 83.1% and 82.7%, respectively. Deep features from the YCbCr

Table 1 Comparison of different deep color features (overall accuracy (OA) in %) on the UC-Merced, WHU-RS19, RSSCNT7, AID, and NWPU-RESISC45 datasets

Approach	UC-Merced (50%)	WHU-RS19 (50%)	RSSCNT7 (50%)	AID (50%)	NWPU-RESISC45 (50%)
Standard RGB (Baseline)	94.7 ±0.68	96.0 ±0.97	89.5 ±0.52	90.3 ±0.23	85.7 ±0.13
Color names	93.7 ±0.66	95.1 ±1.05	87.6 ±0.64	88.7 ±0.51	83.2 ±0.22
HSV	93.8 ±0.80	94.4 ±0.92	87.9 ±0.72	88.5 ±0.49	83.1 ±0.21
Opponent	94.5 ±0.67	96.0 ±0.75	89.4 ±0.54	89.9 ±0.29	84.7 ±0.19
C	93.6 ±0.59	94.7 ±0.75	87.6 ±0.61	88.2 ±0.46	82.7 ±0.18
YCbCr	94.1 ±0.63	95.4 ±0.81	88.2 ±0.70	89.6 ±0.35	84.3 ±0.26
Lab	93.9 ±0.70	95.0 ±0.70	87.9 ±0.81	88.6 ±0.33	83.0 ±0.21
Color Fusion	96.3 ±0.61	97.4 ±0.58	92.3 ±0.64	93.4 ±0.21	90.0 ±0.20

For each dataset, 50% of the images are used for training and the rest for testing. The best result for each dataset is shown in bold value respectively. For all deep models, we use the same network architecture with same set of parameters. Overall, our late fusion based color fusion *always* outperforms the standard baseline RGB color space-based deep convolutional neural network. On the large-scale NWPU-RESISC45 dataset, the combined set of deep color features provides significant improvement in classification performance, with an absolute gain of 4.3% over baseline RGB deep features

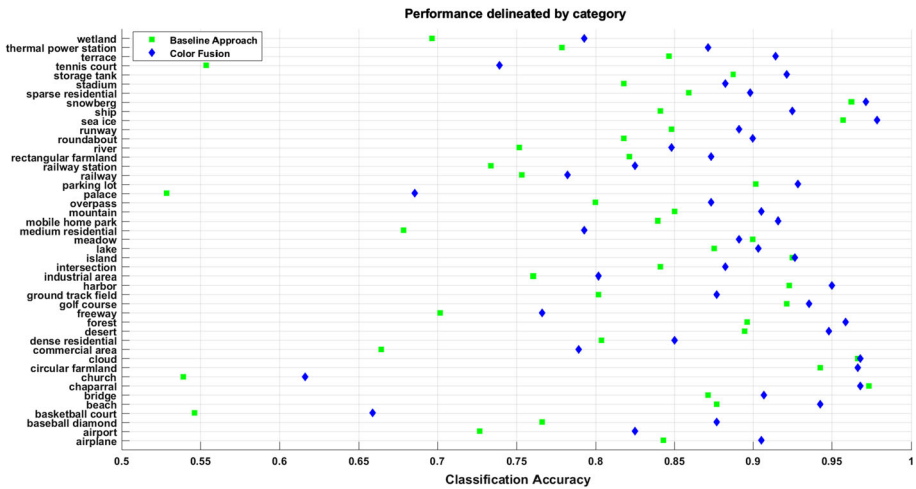


Fig. 4 Category-specific classification performance comparison (accuracy in %) of the proposed method, compared to the baseline RGB deep features, on the large scale NWPU-RESISC45 dataset. Our deep color feature fusion improves the classification performance on 43 out of 45 aerial scene categories. Notably, a significant gain in classification accuracy is achieved for tennis court (+18%), palace (+15%), commercial area (+12%), medium residential (+11%), and basketball court (+11%) aerial scene classes, compared to the baseline (standard RGB). (Color figure online)

and Lab color space based deep convolutional neural networks (CNNs) achieve average classification scores of 84.3% and 83.0%, respectively. The proposed deep color feature fusion provides significant improvement in classification performance with an absolute gain of +4.3%, compared to the baseline standard RGB deep features.

Figure 4 shows a per-category recognition performance comparison between the deep color feature fusion and the baseline RGB deep convolutional neural network on the NWPU-RESISC45 dataset. The combined set of deep color features provides consistent improvements on 43 out of 45 aerial scene categories compared to the baseline RGB features. Particularly significant gains in classification performance are achieved for the tennis court (+18%), palace (+15%), commercial area (+12%), medium residential (+11%), and basketball court (+11%) aerial scene categories.

4.3 Deep Color Features Evaluation

We also perform a comparison between convolutional features (Conv1, Conv2, Conv3, Conv4 and Conv5) and FC features (FC6 and FC7). Table 2 shows the comparison for both the baseline RGB and our color fusion. In all cases, superior classification results are obtained using features from FC layers. Note that no significant improvement in performance is observed when combining convolutional and FC layer features.

To summarize, the deep color feature fusion provides consistent improvements on all five datasets, compared to the baseline RGB features. It is worth mentioning that the most considerable gains in performance are obtained on large-scale AID and NWPU-RESISC45 datasets. These results suggest that different deep color features possess complementary information as their combination leads to a significant performance boost for remote sensing

Table 2 Comparison (overall accuracy (OA) in %) of the baseline RGB and deep color feature fusion when using convolutional layers (Conv) and FC layers (FC)

Approach	UC-Merced (50%)	WHU-RS19 (50%)	RSSCN7 (50%)	AID (50%)	NWPU-RESISC45 (50%)
Baseline RGB (Conv)	92.4 ±0.71	93.7 ±0.98	86.8 ±0.59	87.9 ±0.27	82.3 ±0.16
Baseline RGB (FC)	94.7 ±0.68	96.0 ±0.97	89.5 ±0.52	90.3 ±0.23	85.7 ±0.13
Color Fusion (Conv)	94.9 ±0.63	95.8 ±0.64	91.1 ±0.67	91.7 ±0.27	88.1 ±0.23
Color Fusion (FC)	96.3 ±0.61	97.4 ±0.58	92.3 ±0.64	93.4 ±0.21	90.0 ±0.20

In case of 'Conv', we use features extracted from the five convolutional layers. In case of 'FC', we use features from the two FC layers (Fc6 and Fc7). For both the baseline and our color fusion, features from the FC layers provide superior performance compared to the convolutional layer features

Bold values highlight the results

scene classification. Furthermore, superior results are obtained using features from the FC layers, compared to convolutional layer features.

4.4 Compact Deep Color Features

As demonstrated above, the combined set of deep color features always improves the classification performance compared to the baseline RGB. However, this gain in classification performance comes at the cost of high-dimensionality. When fusing deep color features from the FC6 and FC7 layers of the networks, the resulting dimensionality becomes significantly higher (57K). To tackle this issue, we evaluate the compression of deep color feature fusion using the approach described in Sect. 3.2.

Table 3 shows the results obtained when compressing the combined set of deep color features using the DITC approach. The final dimensions of the compact deep color fusion image representation are fixed to 8k so that it is similar to the dimensionality of the standard RGB deep features commonly employed for classification. The DITC compression approach compresses the combined set of deep color features from 57k to 8k without any substantial deterioration in classification accuracy for all datasets. In the case of UC-Merced and WHU-RS19, there is even a slight improvement in performance when compressing the combined set of deep color features indicating an increase in discriminative power by removing the redundancy. In the case of the RSSCN7, AID and NWPU-RESISC45 datasets, there is a marginal reduction in accuracy compared to the original combined set of deep color features. In all cases, the compact deep color feature fusion significantly reduces the dimensionality without sacrificing the classification accuracy.

We additionally analyze the extreme compression of the deep color feature fusion and compare it with several commonly used dimensionality reduction techniques: principle component analysis (PCA), partial least squares (PLS) and diffusion maps (DM). Among these existing approaches, PLS is a category-aware dimensionality reduction statistical technique that models relations between sets of observations by means of latent variables. We perform the comparison to obtain very low-dimensional (100 to 500 dimensional) deep color feature fusion based image representations. Figure 5 shows the results of extreme compression (even to 100 dimensions) on the UC-Merced dataset. The DITC compression technique provides superior classification results even in the case of extreme compression of the deep color feature fusion based image representation. Figure 6 shows the results on the NWPU-RESISC45 dataset.

4.5 State-of-the-art Performance Comparison

Finally, we compare our compact deep color feature fusion representation with state-of-the-art methods in the literature. Table 4 shows the results on the five remote sensing scene classification datasets. For fair comparison, we adopt the same sampling setting as [32,62,76], taking 80 images per class for training for the UC-Merced dataset. In the case of the WHU-RS19 dataset, 30 images per aerial scene category are used for training and the rest for testing. For the RSSCN7 and AID datasets, 50 images per aerial scene category are employed for training. Furthermore, 20 images per class are employed for training in the case of the NWPU-RESISC45 dataset.

On the UC-Merced dataset, the work of [76] proposed an approach that extends the bag-of-visual-words (BOVW) framework with the spatial co-occurrence kernel, achieving an average classification accuracy of 77.7%. In their work, the integration of color features within a

Table 3 Comparison (overall accuracy (OA) in %) between the deep color feature fusion and the compact color fusion representation

Approach	Dimensions	UC-Merced(50%)	WHU-RS19 (50%)	RSSCN7 (50%)	AID (50%)	NWPU-RESISDC45 (50%)
Color Fusion	57k	97.1 ±0.58	98.4 ±0.57	93.1 ±0.62	94.3 ±0.24	91.8 ±0.18
Compact Color Fusion	8k	97.4 ±0.62	98.8 ±0.56	92.9 ±0.61	94.0 ±0.22	91.1 ±0.19

The best result for each dataset is shown in bold value. In all cases, the compact deep color feature fusion achieves favorable performance while being significantly low-dimensional

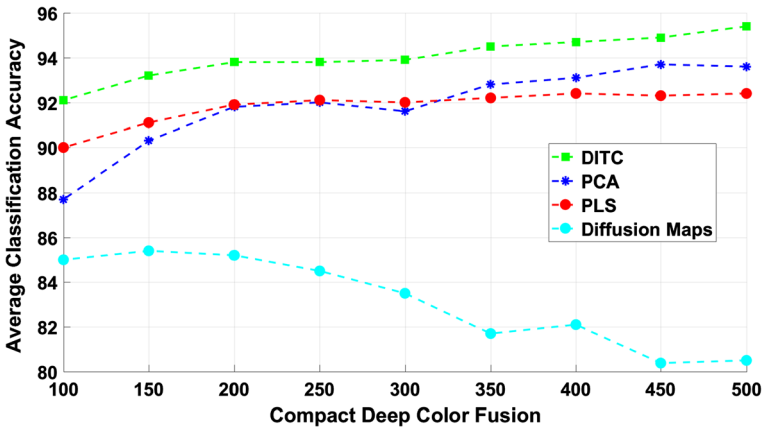


Fig. 5 Comparison of different compression techniques in terms of classification performance (overall accuracy (OA) in %) on the UC-Merced dataset. The DITC based compression technique achieves the best results when performing extreme compression (100 dimensions from the original 57K) of the combined set of deep color features

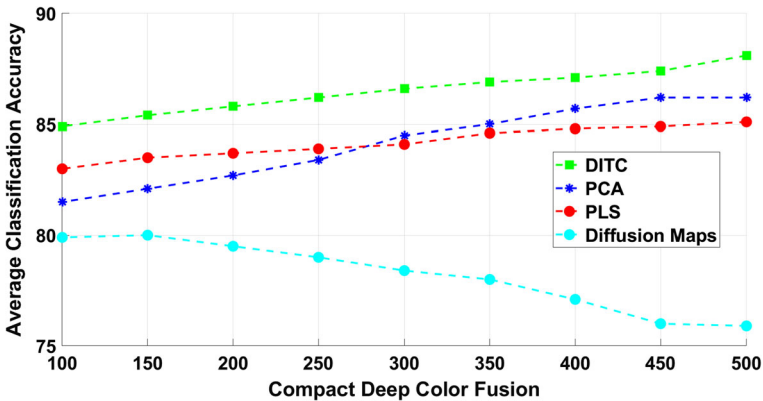


Fig. 6 Comparison of different compression techniques in terms of classification performance (overall accuracy (OA) in %) on the NWPU-RESISC45 dataset. The DITC based compression technique obtains favorable results when employing extreme compression (100 dimensions from the original 57K) of the combined deep color features

Gabor representation was also investigated, leading to a mean recognition rate of 80.5%. The impact of texture information on remote sensing scene classification has been investigated by previous works [9,58,82]. One such texture description based on multi-scale completed LBP features achieved an average classification accuracy of 90.6%. A pyramidal co-occurrence feature representation, accounting for both photometric and geometric aspects of an image, was proposed by [75] achieving a classification accuracy of 77.4%. With the recent advent of deep features, a considerable jump in classification performance has been observed. The work of [72] proposed deep filter banks based on CNNs and obtained a classification accuracy of 92.7%. Previous works have also investigated transferring pre-trained deep features from both the FC and the convolutional (conv) layers of the CNNs. The work of [72] investigated transferability of deep CNNs with respect to both FC and convolutional layers. In case of FC layers (Case I: FC features), their approach achieved an accuracy of 96.8% whereas a mean

Table 4 State-of-the-art performance comparison (overall recognition rate in %) on five datasets used in our evaluation. Our approach, while being compact, provides favorable results compared to the existing works in literature

Approach	UC-Merced	WHU-RS19	RSSCN7	AID	NWPU-RESISC45
Color Gabor [76]	80.5	–	–	–	–
BOVW + spatial co-occurrence kernel [76]	77.7	–	–	–	–
SPCK + SPM [75]	77.4	–	–	–	–
Unsupervised feature learning [18]	81.1	–	–	–	–
Saliency-based feature learning [80]	82.7	–	–	–	–
Circle-structured Concentric BOVW [83]	86.6	–	–	–	–
Wavelet BOVW [82]	87.4	–	–	–	–
Structural texture similarity [58]	86.0	–	–	–	–
Multifeature concatenation [61]	89.5	–	–	–	–
Pyramid-of-spatial-relations [12]	89.1	–	–	–	–
CLBP [9]	85.5	–	–	–	–
MS-CLBP [9]	90.6	–	–	–	–
HHCV [71]	91.8	–	86.4	–	–
DBN based feature selection [84]	–	–	77.0	–	–
Dirichlet [42]	92.8	–	–	–	–
VLAT [52]	94.3	–	–	–	–
Deep Filter Banks [72]	92.7	–	90.4	–	–
Transferring CNNs (Case I: FC features) [32]	96.8	96.7	–	–	–
Transferring CNNs (Case II: Conv features) [32]	96.9	98.6	–	–	–
Category-Specific Vocabulary + Two-Step Categorisation [74]	93.8	93.7	–	–	–
CaffeNet [73]	95.0	94.8	88.2	89.5	–
GoogleNet [73]	94.3	92.9	85.8	86.3	–
VGG-VD-16 [73]	95.2	95.1	87.1	89.6	–
MDDC [55]	96.9	98.3	–	–	–
SSF-AlexNet [10]	–	–	–	88.7	–
BAFF [36]	–	–	–	93.6	–

Table 4 continued

Approach	UC-Merced	WHU-RS19	RSSCN7	AID	NWPU-RESISC45
BoVW+SPM [16]	–	–	–	–	33.0
CNN [16]	–	–	–	–	79.8
LASC-CNN (single-scale) [79]	–	–	–	–	83.6
LASC-CNN (multiscale) [79]	–	–	–	–	84.3
BOCF [17]	–	–	–	–	84.3
This paper	98.1	96.6	92.9	94.0	87.5

Bold values highlight the results

Table 5 Absolute gain (overall accuracy in %) in classification performance obtained using the proposed compact deep color feature fusion compared to deep features using the standard RGB representation

Dataset	UC-Merced	WHU-RS19	RSSCN7	AID	NWPU-RESISC45
Gain	+2.7	+2.7	+3.5	+3.8	+5.4

recognition rate of 96.9% is obtained when using features from convolutional layers (Case II: Conv features) in conjunction with VLAD encoding strategy. The work of [55] proposed a multi-scale deeply described correlations-based model and achieved an accuracy of 96.9%. Our proposed approach, while being compact, achieves an average classification accuracy of 98.1%.

On the WHU-RS19 dataset, the work of [74] based on class-specific vocabulary employing kernel collaborative representation obtained an average classification accuracy of 93.7%. Among the deep learning approaches, CaffeNet model achieved an average classification accuracy of 94.8%. Our compact deep color feature fusion approach also employing FC layer features obtains an average classification accuracy of 96.6%. The best result (98.6%) on this dataset is obtained by transferring deep features from the conv layers together with VLAD encoding technique. Such an encoding of conv features is complementary to our approach using FC features. In the case of RSSCN7 dataset, the work of [71] based on hierarchical coding vectors based classification obtained a mean recognition accuracy of 86.4%. The work of [72] based on deep filter banks achieved a mean recognition rate of 90.4%. Our approach obtains outperforms state-of-the-art methods on this dataset with a mean recognition rate of 92.9%.

For the AID dataset, the work of [10] proposed a deep convolutional neural network pre-training approach (SSF-AlexNet) and achieved a mean recognition accuracy of 88.7%. The work of [36] proposed a fusion approach (BAFF) to integrate SIFT and deep features. Their approach achieved a mean recognition rate of 93.6%. Our approach provides superior results with an average classification accuracy of 94.0%. Finally, on the NWPU-RESISC45 dataset, the work of [79] based on single-scale deep features achieved a mean recognition rate of 83.6%. The multi-scale variant of their approach obtained an average classification score of 84.3%. The bag-of-convolutional feature approach of [17] achieved a mean recognition rate of 84.3%. Our approach again provides superior classification performance by achieving an average classification accuracy of 87.5%.

5 Conclusions

In this paper, we investigated the contribution of color within a deep learning framework (CNNs) for the problem of remote sensing scene classification. We demonstrated that different deep color features possess complementary information and combining them leads to a significant performance boost for the remote sensing scene classification task. Additionally, we addressed the high-dimensionality of deep color feature fusion and compressed them to obtain a compact final image representation without a significant deterioration in classification performance. To validate our approach, we perform comprehensive experiments on five challenging remote sensing scene classification datasets. The results from our experiments clearly demonstrated the effectiveness of the proposed approach. Table 5 shows the gain in classification performance obtained using the proposed compact deep color feature fusion, compared to the standard RGB deep features, on the five remote sensing scene classification datasets.

A potential future research direction is to investigate the fusion of other available spectral bands (, near infrared) besides RGB in the form of different color transformations. Additionally, integrating other visual cues, such as texture features with color features in a deep learning framework may improve remote sensing scene classification performance and is therefore a promising research direction. Another future research direction is to investigate the impact of integrating multiple deep color features for other remote sensing image analysis tasks, such as object detection (simultaneous classification and localization).

Acknowledgements This work has been funded by the Academy of Finland in project AIROBEST (317388) and supported by the Aalto Science-IT project.

Funding Open access funding provided by Aalto University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Alvarez J, Gevers T, Lopez A (2010) Learning photometric invariance for object detection. *IJCV* 10(1):45–61
2. Anwer RM, Khan FS, van de Weijer J, Molinier M, Laaksonen J (2018) Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J Photogramm Remote Sens* 138:74–85
3. Anwer RM, Vazquez D, Lopez A (2011) Color contribution to part-based person detection in different types of scenarios. In: CAIP
4. Anwer RM, Vazquez D, Lopez A (2011) Opponent colors for human detection. In: IbPRIA
5. Azizpour H, Sullivan J, Carlsson, S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW
6. Berlin B, Kay P (1969) Basic color terms: their universality and evolution. University of California Press, Berkeley
7. Bosch A, Zisserman A, Munoz X (2008) Scene classification using a hybrid generative/discriminative approach. *PAMI* 30(4):712–727

8. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. In: *BMVC*
9. Chen C, Zhang B, Su H, Li W, Wang L (2016) Land-use scene classification using multi-scale completed local binary patterns. *SIVP* 4:745–752
10. Chen J, Wang C, Ma Z, Chen J, He D, Ackland S (2018) Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters. *Remote Sens* 10(12):1–16
11. Chen L, Yang W, Xu K, Xu T (2011) Evaluation of local features for scene classification using vhr satellite images. In: *JURSE*
12. Chen S, Tian Y (2015) Pyramid of spatial relations for scene-level land use classification. *TGRS* 53(4):1947–1957
13. Chen X, Fang T, Huo H, Li D (2015) Measuring the effectiveness of various features for thematic information extraction from very high resolution remote sensing imagery. *TGRS* 53(9):4837–4851
14. Chen Y, Lin Z, Zhao X, Wang G, Gu Y (2014) Deep learning-based classification of hyperspectral data. *JSTARS* 7(6):2094–2107
15. Chen Y, Zhao X, Jia X (2015) Spectralspatial classification of hyperspectral data based on deep belief network. *JSTARS* 8(6):2381–2392
16. Cheng G, Han J, Lu X (2010) Remote sensing image scene classification: benchmark and state of the art. *JPROC* 105(10):1865–1883
17. Cheng G, Li Z, Yao X (2017) Remote sensing image scene classification using bag of convolutional features. *LGRS* 14(10):1735–1739
18. Cheriyyadath A (2014) Unsupervised feature learning for aerial scene classification. *TGRS* 52(1):439–451
19. Cholakkal H, Sun G, Khan FS, Shao L (2019) Object counting and instance segmentation with image-level supervision. In: *CVPR*
20. Cimpoi M, Maji S, Kokkinos I, Vedaldi A (2016) Deep filter banks for texture recognition, description, and segmentation. *IJCV* 118(1):65–94
21. Danelljan M, Meneghetti G, Khan FS, Felsberg M (2016) A probabilistic framework for color-based point set registration. In: *CVPR*
22. Dhillon I, Mallela S, Kumar R (2003) A divisive information-theoretic feature clustering algorithm for text classification. *JMLR* 3:1265–1287
23. Dong C, Loy CC, He K, Tang X (2016) Image super-resolution using deep convolutional networks. *PAMI* 38(2):295–307
24. dos Santos JA, Penatti OAB, da Silva Torres R (2010) Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: *VISAPP*
25. Elfiky N, Khan FS, van de Weijer J, Gonzalez J (2012) Discriminative compact pyramids for object and scene recognition. *PR* 45(4):1627–1636
26. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: *CVPR*
27. Geusebroek JM, van den Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. *PAMI* 23(12):1338–1350
28. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *CVPR* (2014)
29. Gong Z, Zhong P, Hu W, Hua Y (2019) Joint learning of the center points and deep metrics for land-use classification in remote sensing. *Remote Sens* 11(1):1–21
30. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-CNN. In: *ICCV*
31. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *CVPR*
32. Hu F, Xia GS, Hu J, Zhang L (2015) Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens* 7(11):680–707
33. Hu F, Xia GS, Wang Z, Huang X, Zhang L, Sun H (2015) Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *JSTARS* 8(5):2015–2030
34. Huang L, Chen C, Li W, Du Q (2016) Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors. *Remote Sens* 8(6):483–490
35. Jarrett K, Kavukcuoglu K, Ranzato MA, LeCun Y (2009) What is the best multi-stage architecture for object recognition. In: *ICCV*
36. Ji W, Li X, Lu X (2017) Bidirectional adaptive feature fusion for remote sensing scene classification. In: *CCCV* (2017)
37. Khan FS, Anwer RM, van de Weijer J, Felsberg M, Laaksonen J (2015) Compact color-texture description for texture classification. *PRL* 51:16–22
38. Khan FS, van de Weijer J, Anwer RM, Bagdanov A, Felsberg M, Laaksonen J (2018) Scale coding bag of deep features for human attribute and action recognition. *MVA*

39. Khan FS, van de Weijer J, Vanrell M (2012) Modulating shape features by color attention for object recognition. *IJCV* 98(1):49–64
40. Khan FS, Xu J, van de Weijer J, Bagdanov A, Anwer RM, Lopez A (2015) Recognizing actions through action-specific person detection. *TIP* 24(11):4422–4432
41. Kim H, Ro YM (2016) Collaborative facial color feature learning of multiple color spaces for face recognition. In: *ICIP*
42. Kobayashi T (2014) Dirichlet-based histogram feature transform for image classification. In: *CVPR*
43. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *NIPS*
44. Kusumaningrum R, Wei H, Manurung R, Murni A (2014) Integrated visual vocabulary in latent Dirichlet allocation-based scene classification for IKONOS image. *JARS* 8(1):083690
45. Lenz R, Bui TH, Hernandez-Andres J (2005) Group theoretical structure of spectral spaces. *J Math Imaging Vis* 23(3):297–313
46. Liu H, Li J, He L, Wang Y (2019) Superpixel-guided layer-wise embedding CNN for remote sensing image classification. *Remote Sens* 11(2):1–25
47. Luo B, Jiang S, Zhang L (2013) Indexing of remote sensing images with different resolutions by multiple features. *JSTARS* 6(4):1899–1912
48. Mandal D, Narayan S, Dwivedi SK, Gupta V, Ahmed S, Khan FS, Shao L (2019) Out-of-distribution detection for generalized zero-shot action recognition. In: *CVPR*
49. Marmanis D, Datcu M, Esch T, Stilla U (2016) Deep learning earth observation classification using imagenet pretrained networks. *LGRS* 13(1):105–109
50. Nair V, Hinton G (2010) Rectified linear units improve restricted Boltzmann machines. In: *ICML*
51. Narayan S, Cholakkal H, Khan FS, Shao L (2019) 3c-net: Category count and center loss for weakly-supervised action localization. In: *ICCV*
52. Negrel R, Picard D, Gosselin PH (2014) Evaluation of second-order visual features for land-use classification. In: *CBMIW*
53. Pang Y, Wang T, Anwer RM, Khan FS (2019) Efficient featured image pyramid network for single shot detector. In: *CVPR*
54. Penatti O, Nogueira K, Santos J (2015) Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: *CVPRW*
55. Qi K, Yang C, Guan Q, Wu H, Gong J (2017) A multiscale deeply described correlatons-based model for land-use scene classification. *Remote Sens* 9(9):1–17
56. Rachmadi RF, Purnama IKE (2015) Vehicle color recognition using convolutional neural network. *arXiv preprint arXiv:1510.07391*
57. Ren S, He K, Girshick R, Sun J (2015) Faster r-CNN: towards real-time object detection with region proposal networks. In: *NIPS*
58. Risojevic V, Babic Z (2011) Aerial image classification using structural texture similarity. In: *ISSPIT*
59. Risojevic V, Babic Z (2016) Unsupervised quaternion feature learning for remote sensing image classification. *JSTARS* 9(4):1521–1531
60. Rojas D, Khan FS, van de Weijer J, Gevers T (2010) The impact of color on bag-of-words based object recognition. In: *ICPR*
61. Shao W, Yang W, Xia GS, Liu G (2013) A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization. In: *ICCVS*
62. Sheng G, Yang W, Xu T, Sun H (2012) High-resolution satellite scene classification using a sparse coding based multiple feature combination. *IJRS* 33(8):2395–2412
63. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: *NIPS*
64. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *ICLR*
65. Tokarczyk P, Wegner JD, Walk S, Schindler K (2015) Features, color spaces, and boosting: new insights on semantic classification of remote sensing images. *IEEE Trans Geosci Remote Sens* 53(1):280–295
66. Tuia D, Flamary R, Courty N (2015) Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions. *JPRS* 105:272–285
67. van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. *PAMI* 32(9):1582–1596
68. Wang T, Anwer RM, Cholakkal H, Khan FS, Pang Y, Shao L (2019) Learning rich features at high-speed for single-shot object detection. In: *ICCV*
69. Wang T, Anwer RM, Khan MH, Khan FS, Pang Y, Shao L, Laaksonen J (2019) Deep contextual attention for human-object interaction detection. In: *ICCV*

70. van de Weijer J, Schmid C, Verbeek JJ, Larlus D (2009) Learning color names for real-world applications. *TIP* 18(7):1512–1524
71. Wu H, Liu B, Su W, Sun J (2016) Hierarchical coding vectors for scene level land-use classification. *Remote Sens* 8(5):436–453
72. Wu H, Liu B, Su W, Zhang W, Sun J (2016) Deep filter banks for land-use scene classification. *LGRS* 13(12):1895–1899
73. Xia GS, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L (2017) Aid: A benchmark dataset for performance evaluation of aerial scene classification. *TGRS*
74. Yan L, Zhu R, Mo N, Liu Y (2017) Improved class-specific codebook with two-step classification for scene-level classification of high resolution remote sensing images. *Remote Sens* 9(3):223–247
75. Yi Yang, Newsam S (2011) Spatial pyramid co-occurrence for image classification. In: *ICCV*
76. Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: *GIS*
77. Yang Y, Newsam S (2013) Geographic image retrieval using local invariant features. *TGRS* 51(2):818–832
78. Yu H, Yang W, Xia GS, Liu G (2016) A color-texture-structure descriptor for high-resolution satellite image classification. *Remote Sens* 8(3):1–24
79. Yuan B, Li S, Li N (2018) Multiscale deep features learning for land-use scene recognition. *JARS* 12(1):1–12
80. Zhang F, Du B, Zhang L (2015) Saliency-guided unsupervised feature learning for scene classification. *TGRS* 53(4):2175–2184
81. Zhang L, Gonzalez-Garcia A, van de Weijer J, Danelljan M, Khan FS (2019) Synthetic data generation for end-to-end thermal infrared tracking. *TIP* 28(4):1837–1850
82. Zhao L, Tang P, Huo L (2014) A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification. *IJRS* 35:2296–2310
83. Zhao L, Tang P, Huo L (2014) Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *JSTARS* 7(12):4620–4631
84. Zou Q, Ni L, Zhang T, Wang Q (2015) Deep learning based feature selection for remote sensing scene classification. *LGRS* 12(11):2321–2325

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.