

RESEARCH

Open Access



Remote sensing scene classification based on rotation-invariant feature learning and joint decision making

Yong Zhou[†], Xuning Liu, Jiaqi Zhao^{*†} , Ding Ma, Rui Yao, Bing Liu and Yi Zheng

Abstract

With the popular use of high-resolution satellite images, remote sensing scene classification has always been a hot research topic in its related areas. However, limited to the issues of remote sensing datasets including the small scale of scene classes, the lack of rich label information and so on, it is quite challenging for deep learning methods to learn powerful feature representation. To overcome this problem, we propose a rotation-invariant feature learning and joint decision-making method based on Siamese convolutional neural networks with the combination of identification and verification models. Firstly, a novel data augmentation strategy is proposed specially for the Siamese model to learning rotation-invariant features. Secondly, a joint decision mechanism is introduced in our method, which is realized by the identification and verification model to better improve the classification performance. The proposed method can not only suppress problems caused by lack of rich label samples but also improve the robustness of Siamese convolutional neural networks. Experimental results demonstrate that the proposed method is effective and efficient for remote sensing scene classification.

Keywords: Remote sensing scene classification, Identification and verification models, Rotation-invariant features, Joint decision mechanism

1 Introduction

With the development of remote sensing techniques, a large collection of high-resolution remote sensing images is becoming available. These images have been widely applied to many fields [1–3], such as urban planning, natural hazard detection, and environment monitoring. For this reason, more and more research efforts have been put into developing methods for remote sensing scene classification which is a hot research topic in the remote sensing field to better interpret the images [4–6].

Feature extraction is a critical step for remote sensing scene classification which can significantly affect the performance. According to the used features, the existing remote sensing classification methods can generally be divided into three main categories [7, 8]: handcrafted-feature-based method, unsupervised-feature-

learning-based methods, and deep-feature-learning-based methods.

Handcrafted features contain color, shape, and texture information which are primary characteristics of the remote sensing images and carry useful information for scene classification. The early scene classification methods are mainly based on these features [9–11]. To be specific, there are several representative handcrafted features such as color histograms [12], GIST [13], scale-invariant feature transform (SIFT) [14], and HOG [15]. The color histograms and GIST descriptors are the global features used to describe the statistical properties of an entire image in the perspective of color, texture, and spatial structure information, whereas SIFT descriptor and HOG feature are local features that represent local structure and shape information. Multiple types of feature can convey scene information of an entire image, which cannot be done by one single type of feature. Hence, many methods based on the combination of various complementary features for scene classification have been proposed to boost the performance. These

* Correspondence: jiaqizhao@cumt.edu.cn

[†]Yong Zhou and Jiaqi Zhao contributed equally to this work.
School of Computer Science and Technology, China University of Mining and Technology, No1, Daxue Road, 221116 Xuzhou, Jiangsu, People's Republic of China

human-engineering features are manually designed and require domain expertise, which greatly limits the representation capability when the scene images become challenging.

To overcome the problems of handcrafted features for scene classification, unsupervised feature learning is reckoned as the potential strategy. It can automatically learn features from unlabeled input data and has made astonishing progress in remote sensing scene classification [16–18]. The unsupervised-learning-based features are more discriminative and better suited to the classification problem. PCA, K-means clustering, sparse coding, and autoencoder are typical unsupervised learning methods. These methods and their variants have achieved great success in the scene classification field.

In recent years, various deep learning methods have shown their powerful feature representation in the field of machine learning [19]. Convolutional neural network (CNN) is one of the most successful methods and has acquired various applications in remote sensing community [20–22]. Despite the success of CNN, there are several challenging problems of remote sensing classification. First, the small scale of remote sensing datasets has severely hindered the development of deep-learning-based methods for scene classification because deep learning models need large datasets to be trained on in order to avoid overfitting. Besides they could not learn robust feature representation without abundant and diverse images. Second, unlike natural image datasets, remote sensing datasets have their own characteristics. For example, NWPU-RESISC45 dataset carefully selects images under various real-world conditions including illuminations, seasons, and weather. For images of the same category, they are very different in terms of object pose, appearance, spatial resolution, and background. These practical factors significantly affect the performance of useful feature extraction which is a critical step for scene classification. In brief, it is vital to learn discriminative feature representation of remote sensing images.

In this paper, we present a robust tool for classifying the remote sensing images which is accomplished by adopting a Siamese CNN architecture better adapted to the characteristics of remote sensing datasets. Our architecture has two identical CNN channels that combine the identification and verification models. The identification model accepts a single image as input and utilizes the CNN to extract the useful features and the final convolutional layer to predict its label. Meanwhile, the verification model compares the feature vectors of the two images extracted by their respective identification models and calculate their distance in feature space. These two models are complementary and their combination allows the method to learn discriminative feature

representation. Considering the two CNN channels of our architecture, one channel input adopts data augmentation like random rotation while the other remains unchanged. This operation can let our network learn rotation-invariant features which positively affects the classification results. To make the most of this combination, a joint decision making is introduced in our work. That is, we encode the probabilistic relationships drawn by identification and verification models to mine the valuable information of input data.

In general, the main contributions of this paper are listed below:

- 1) Data augmentation, especially random rotation, is adopted in one of the CNN channel input, which allows the identification model to learn rotation-invariant features to enhance the classification performance.
- 2) To take full advantage of the identification model and verification model, we introduce the joint decision made by these two models to the classification task.

The remaining part of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the details of our method. Section 4 shows the experiments and discussion. Section 5 makes a conclusion about this paper and describes the possible future work.

2 Related work

2.1 Convolutional neural networks

Convolutional neural network (CNN) is a typical deep learning model which consists of the input layer, the convolutional layers, the pooling layers, the fully connected layers, and the output layer. Convolutional layers are used for extracting multi-level features of data according to convolution kernels of different sizes. Pooling layers aim at reducing the dimensions of feature representation and making the feature invariant from the location through a pooling function. Fully connected layers combine the outputs of all the previous layers into high-level features. Generally, CNN can automatically learn high-level features and has proved its powerful classification capability.

There are three successful CNN models used in our paper: AlexNet [23], VGG16 [24], and ResNet50 [25]. AlexNet contains five convolutional layers, three pooling layers, and three fully connected layers. The output of the last fully connected layer is fed to a softmax function to produce a distribution over 1000 class labels. VGG16 contains five convolutional layers, five max-pooling layers, and three fully connected layers. The first two fully connected layers have 4096 channels each, and the last has 1000 channels for each class. The final layer is a

softmax layer. It is important to point out that the convolutional layers of VGG16 use very small 3×3 receptive fields (stride 1) instead of relative larger ones and max-pooling is performed on 2×2 pixel window (stride 2) so as to simplify the network structure. ResNets insert short connections that perform identity mapping based on the philosophy of VGG Nets and have shown significant performance on various computer vision tasks compared with state-of-the-arts.

2.2 Data augmentation

In many classification problems, especially remote sensing scene classification, the available data is inadequate to train accurate and robust classifiers. To alleviate the impact of this problem, one popular approach is data augmentation. Data augmentation is the process of generating new similar samples to the training set by using label-preserving transformations. The common image transformations are as follows:

- 1) Flipping. Flip the image horizontally.
- 2) Rotation. Rotate the image at random orientation.
- 3) Cropping. Crop a part from the original image and resize the cropped image.
- 4) Shifting. Shift the image to the left or right, and the translation range and step length can be specified in order to change the location of the image content.
- 5) Noise. The image is added random noise to RGB channels of each pixel.
- 6) Color jittering. Change the random factors of color saturation, brightness, and contrast in the image color space.

- 7) PCA jittering [23]. PCA is performed on the image to get the principal component, which is added to the original image with a Gaussian disturbance of (0,0.1) to generate the new image.

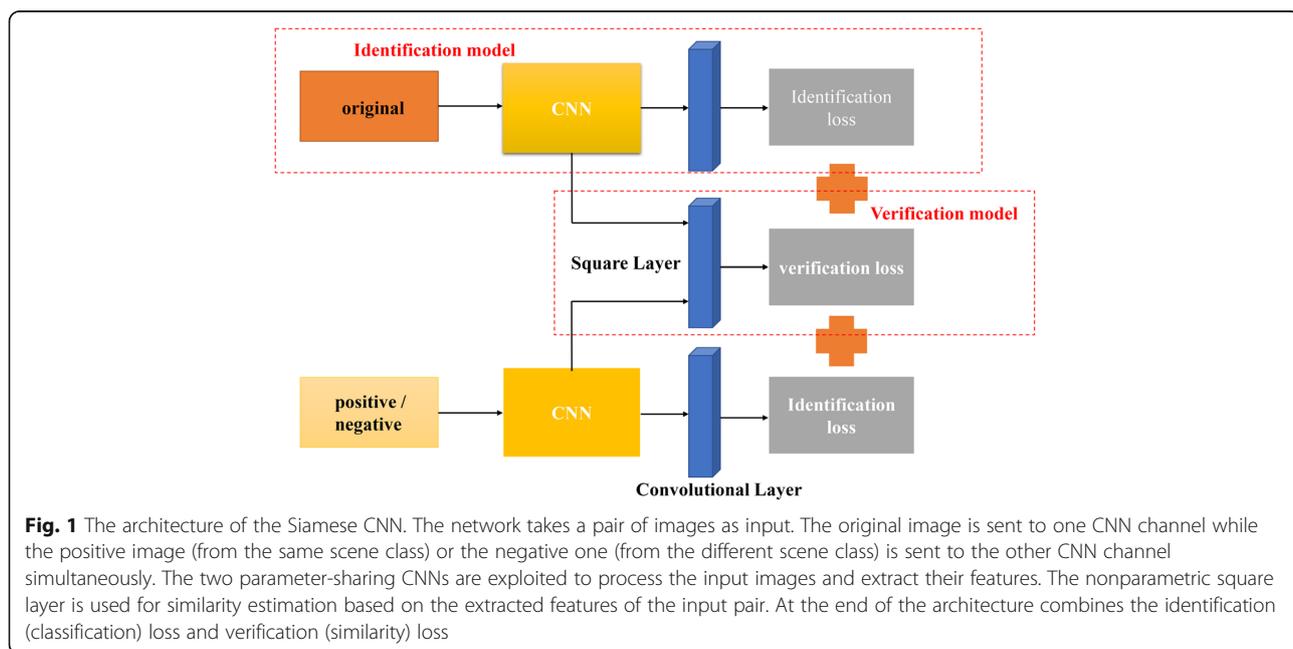
Jia et al. [26] had carried out comprehensive experiments which indicate that flipping, rotation, and cropping not only outperform the other enhancement methods but also are more effective on the smaller scale of data sets.

3 The proposed method

3.1 The architecture of the Siamese CNNs

Based on the work of [27], we utilize the Siamese CNNs which combines identification and verification models for remote sensing scene classification. Figure 1 presents the overall architecture. The network consists of two ImageNet [28] pre-trained CNN models, three additional convolutional layers, and a square layer. AlexNet [23], VGG16 [24], and ResNet50 [25] are employed as the pre-trained CNN models. To better adapt to the characteristics of remote sensing datasets, we replace the final fully connected layer of three CNN models with a convolutional layer whose number of kernels is identical with the scene classes of datasets. The whole network is trained to minimize three cross-entropy losses and one distance loss jointly.

The whole architecture takes a pair of images as input. The input pair can be divided into a positive one (from the same scene class) and a negative one (from different scene classes). Given a pair of images, the network predicts their labels and the similarity simultaneously. For



the identification model, it uses the CNN to extract image features and then predicts the label of the image based on the extracted features. For the verification model, it accepts the two feature vectors produced by the two CNNs as input and utilizes the square layer to calculate the similarity score of two images.

There are two identical CNN models in our architecture, which share weights and predict the two scene class labels of the input image pair simultaneously. The identification model is composed of each CNN channel with an additional convolutional layer. To fine tune the whole network on remote sensing datasets, the final fully connected layer of the pre-trained CNN model is replaced with a convolutional layer. NWPU dataset has 45 scene classes. Accordingly, this convolutional layer has 45 kernels of size $1 \times 1 \times 4096$ connected to the output of the CNN model. And a softmax unit is added to normalize the final output. We use the cross-entropy loss as the identification (classification) loss, which is expressed as:

$$\hat{p} = \text{softmax}(\theta_I * f) \tag{1}$$

$$L_{id}(f, t, \theta_I) = \sum_{i=1}^K -p_i \log(\hat{p}_i) \tag{2}$$

Here, asterisk denotes the convolution operation. f is defined as the output of the CNN model and is a 4096-dim feature vector. θ_I is the parameters of the additional convolutional layer. Based on the work of [27], t is the target scene class, p is the target possibility, and accordingly \hat{p} is the possibility of the predicted scene class. So for the target class t , $p_t = 1$ and $p_i = 0$ ($i \neq t$).

The high-level features f_1 and f_2 from the fine-tuned CNNs in the identification model are directly used by the verification model for similarity estimation. From Fig. 1, our work uses a square layer which is nonparametric to compare these high features. It takes two feature tensors f_1, f_2 as inputs and outputs tensor f_s , where f_1, f_2 are 4096-dim feature embeddings and $f_s = (f_1 - f_2)^2$. After the square layer, we add a convolutional layer and a softmax function to embed f_s to a 2-dim vector (\hat{q}_1, \hat{q}_2) that denotes the predicted probability of the input image pair belonging to the same scene class. To be more specific, the convolutional layer filters the input f_s with two kernels of size $1 \times 1 \times 4096$. The softmax function is used to normalize the output. Similar to the identification loss, we also use the cross-entropy loss as the verification loss, which is expressed as:

$$\hat{q} = \text{softmax}(\theta_s * f_s) \tag{3}$$

$$L_{ver}(f_1, f_2, s, \theta_s) = \sum_{i=1}^2 -q_i \log(\hat{q}_i) \tag{4}$$

Here, f_1 and f_2 are the 4906-dim feature tensors from the finetuned CNN in the identification model. θ_s

denotes the parameters of the added convolutional layer. s means whether the two images are from the same scene class or not. \hat{q} is the predicted probability. If the two input images are from the same scene class, $q_1 = 1$ and $q_2 = 0$; otherwise, $q_1 = 0$ and $q_2 = 1$.

3.2 Rotation-invariant feature training

We will perform random-rotation-based data augmentation on the training sets which feed the network to learn rotation-invariant features. For remote sensing datasets, images of the same category have many variants in terms of different directions, appearances, backgrounds, and so on, which hamper the improvement of scene classification. As our architecture takes a positive/negative pair as input, the random rotation is only used for the positive pair of images which come from the same category. There are two CNN channels in our architecture as shown in Fig. 1. One CNN channel accepts the original images, and meanwhile, we randomly rotate the images if the other CNN channel accepts the positive (from the same class) one. Ideally, the two CNN channels will learn nearly the same features of two images. Towards this end, we add a metric learning regularization term on these two features learned by the CNNs, which enforces the training samples with and without random rotation to share the similar features, hence achieving rotation invariance.

In the influential work of [29], we intend to use the high-level feature tensors f_1, f_2 to calculate the Euclidean feature distance of the input pair (x_i, x_j) , which comes from the same scene class. The distance formulation is expressed as follows:

$$D(x_i, x_j) = \|f_1 - f_2\|_2^2 \tag{5}$$

$$\begin{cases} D(x_i, x_j) < \tau, y_i = y_j \\ D(x_i, x_j) > \tau, y_i \neq y_j \end{cases}$$

Of which, given a training sample (x_i, x_j) , their target identity is (y_i, y_j) . The margin τ is a threshold to define the feature distance between the similar pairs. That is, if (x_i, x_j) is from the same scene class, their feature distance should be smaller than τ ; otherwise, they are from the different scene class. It is necessary to point out that our model only focuses on minimizing the feature distance between the similar input pairs (positive pairs) through this margin τ . Accordingly, the metric learning regularization term is defined as below:

$$\text{Dist}(f_1, f_2) = \sum_{i,j} D(x_i, x_j), D(x_i, x_j) < \tau \tag{6}$$

Moreover, we set a ratio of random rotation on the positive training images. The ratio set between positive training images with and without randomly rotating is

{1:1, 1:2, 1:3, 1:4}. To figure out which ratio performs best, we train the models based on the ratios separately. We will discuss the results in Section 4.

3.3 Joint decision making

To make the most of the combination of identification and verification models, we introduce the joint decision-making mechanism to the classification work. This mechanism intends to encode the probabilistic relationships given by these two models. Considering the whole architecture, the identification model extracts the features of the image to predict its class label, which means this model outputs a soft-max probability P_i . And the verification model outputs a prediction probability P_s of whether the two images belong to the same scene class. First of all, we randomly select the same number of typical images for each scene class from the training set as the comparison set. Then, we sample the image x from the test set and the image y from the comparison set to generate an input pair (x, y) . The whole network will output two kinds of probabilities with regard to the input data. This probabilistic relationship can be encoded by the following eq.:

$$P(x) = P_s(x, y_j, \theta_s) * P_i(y_j|x, \theta_I), j = 1, 2, \dots, K \quad (7)$$

Of which, x and y are random variables from the dataset. They are totally independent of each other before fed through the network. K is the image number of the

comparison set. Figure 2 visualizes the joint decision-making mechanism. Every shape in Fig. 2 represents one class. One test image along with the images in the comparison set is fed through the whole network. The identification model outputs the class probability $P_i(j)$ ($j = 1, 2, \dots, K$) of images in the comparison set. Meanwhile, the verification model outputs the similarity measurement $P_v(j)$ which is the relationship score between the test image and every single image in the comparison set. Based on Eq. (7), we can get a set of $P(x)(P_1(x), P_2(x), \dots, P_K(x))$ and then use the highest score as the class probability of the test image.

4 Experimental study

To evaluate the classification performance of the proposed model, we conduct plenty of experiments on three remote sensing datasets including SIRI-WHU dataset [30], UC Merced Land-Use dataset [31], and NWPU-RESISC45 dataset [7]. In this section, we will present the experimental results and corresponding discussions.

4.1 Datasets

SIRI-WHU [30] is a Google image dataset and mainly covers urban areas in China. It contains 12 scene classes with 200 images per class. Each image measures 200×200 pixels with a 2-m spatial resolution in the red-green-blue (RGB) color space. Figure 3 shows some examples.

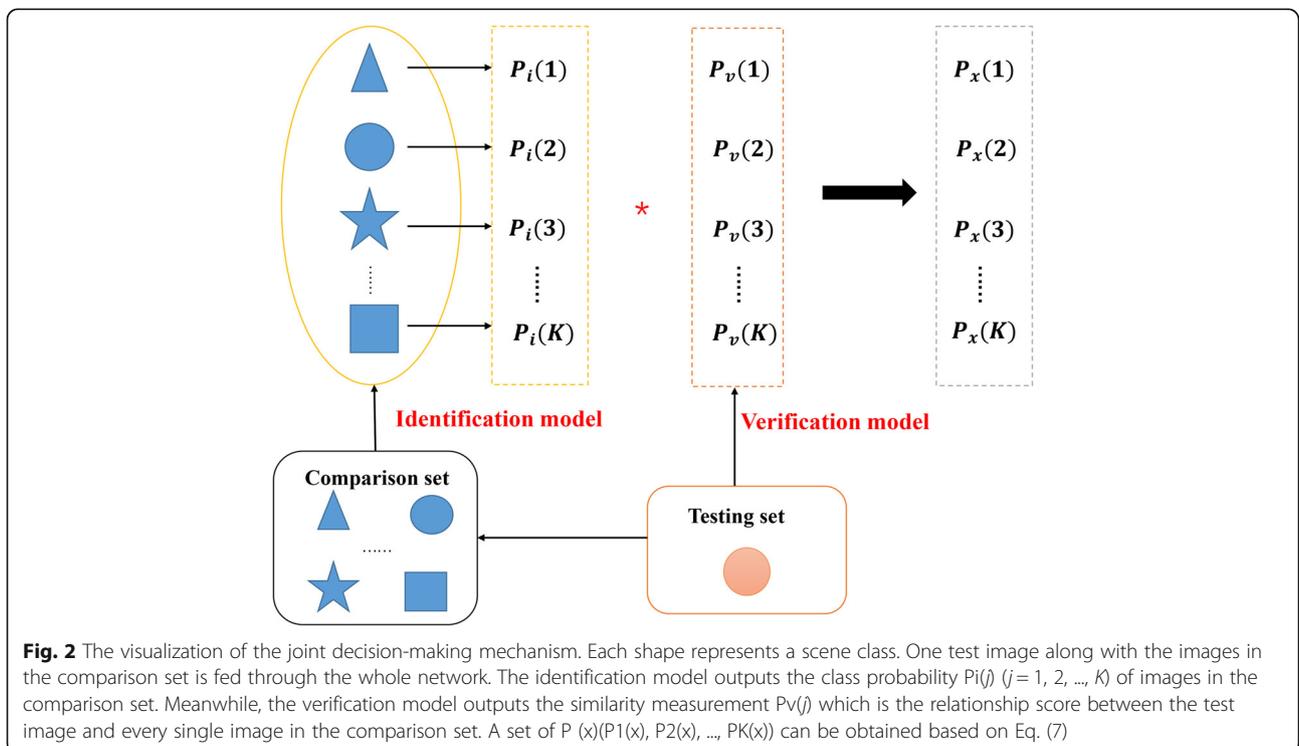


Fig. 2 The visualization of the joint decision-making mechanism. Each shape represents a scene class. One test image along with the images in the comparison set is fed through the whole network. The identification model outputs the class probability $P_i(j)$ ($j = 1, 2, \dots, K$) of images in the comparison set. Meanwhile, the verification model outputs the similarity measurement $P_v(j)$ which is the relationship score between the test image and every single image in the comparison set. A set of $P(x)(P_1(x), P_2(x), \dots, P_K(x))$ can be obtained based on Eq. (7)

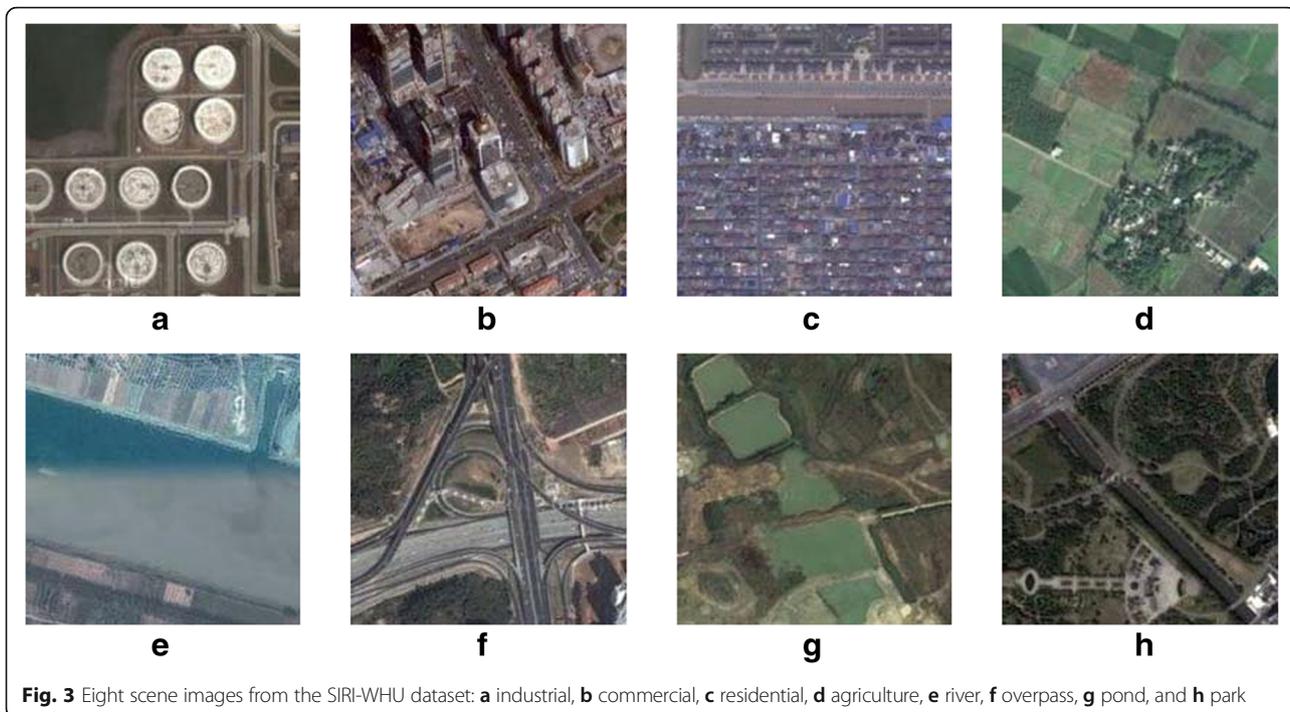


Fig. 3 Eight scene images from the SIRI-WHU dataset: **a** industrial, **b** commercial, **c** residential, **d** agriculture, **e** river, **f** overpass, **g** pond, and **h** park

UC Merced Land-Use dataset [31] covers 21 land-use scene classes with 100 images per class. Each aerial image measures 256×256 pixels, with a spatial resolution of 0.3 m per pixel in the RGB color space. Figure 4 presents the examples of this dataset.

NWPU-RESISC45 dataset [7] is a publicly available benchmark for remote sensing scene classification (RESISC). It contains 31,500 images divided into 45 scene classes. Each class has 700 images measuring 256×256 pixels in the RGB color space. Figure 5 presents the examples of this dataset.

4.2 Implementation setup

We utilize AlexNet [23], VGG16 [24], and ResNet50 [25] pre-trained on ImageNet dataset as the CNN architecture of our model and then fine tune them to better adapt to the remote sensing datasets.

During training, we randomly crop images into 227×227 for AlexNet and 224×224 for VGG16 and ResNet50. To alleviate the prediction bias, we first shuffle all the datasets and use the random order of the images. Then, we sample another image from the same/different scene class to form a positive/negative pair. The ratio between

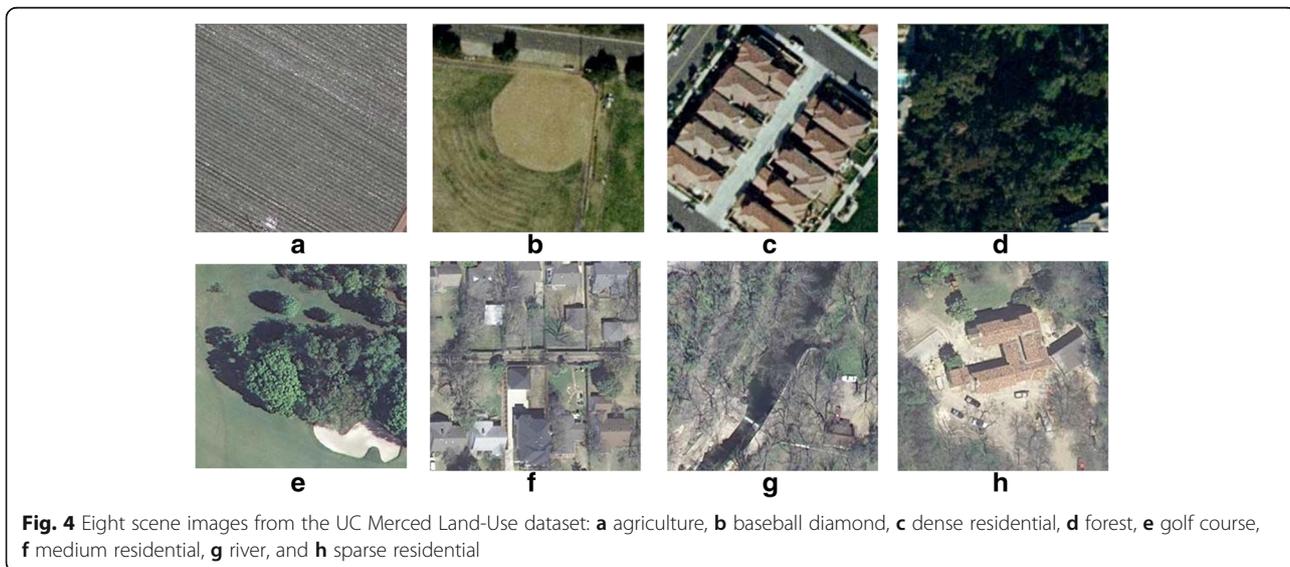
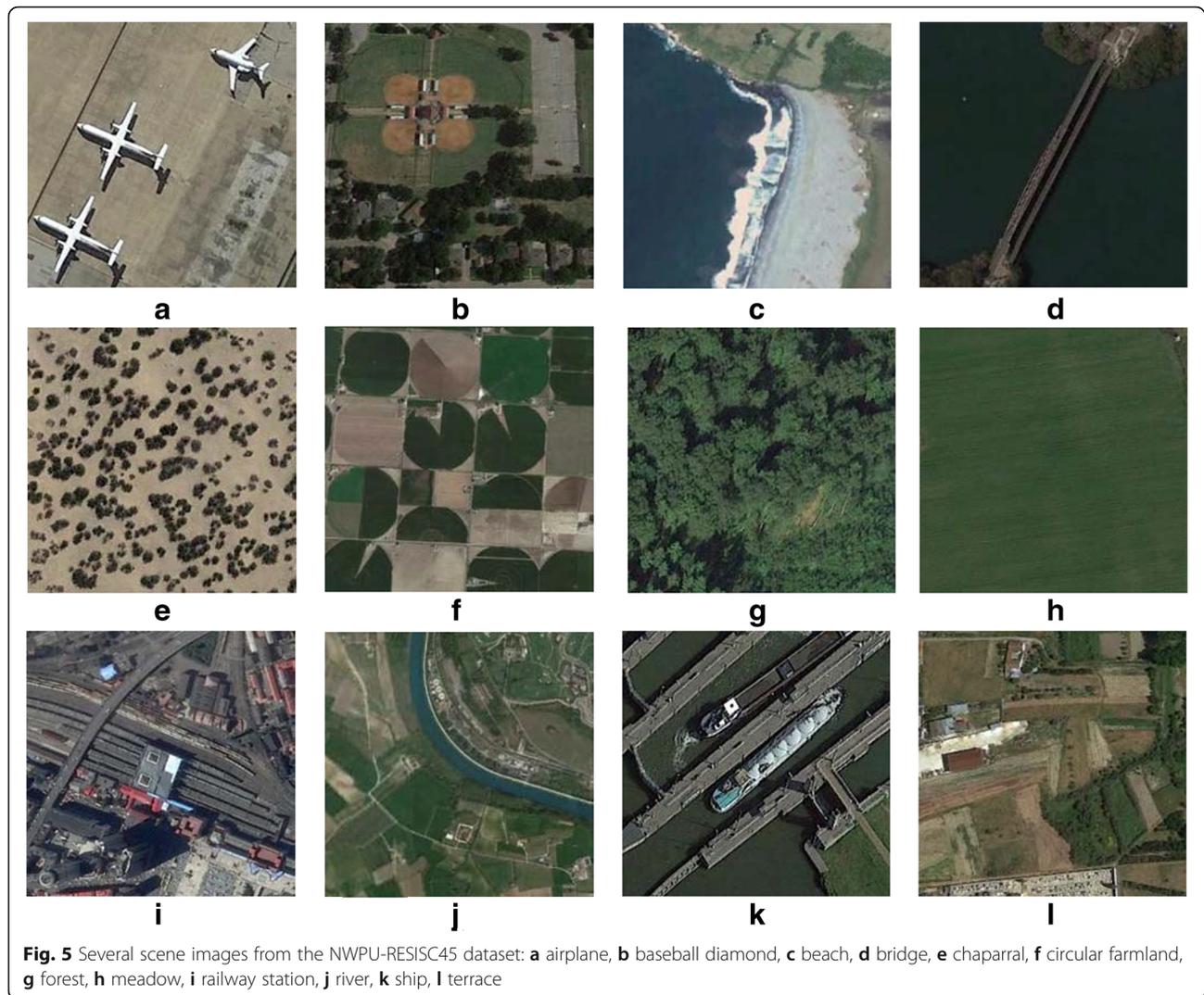


Fig. 4 Eight scene images from the UC Merced Land-Use dataset: **a** agriculture, **b** baseball diamond, **c** dense residential, **d** forest, **e** golf course, **f** medium residential, **g** river, and **h** sparse residential



positive and negative pairs is initially set to 1:1 and then is multiplied by a factor of 1.01 until it reaches 4:1.

The maximum number of training epochs is set to 120. The batch size is 48 for AlexNet and VGG16 and 36 for ResNet50. The learning rate is 0.001 and 0.0001 for the final ten epochs. And we set the distance margin τ to 1. The mini-batch stochastic gradient descent (SGD) is adopted to update all the parameters of our network. In our model, the network has three kinds of losses including identification loss, verification loss, and distance loss to minimize. In the training of our model, the trade-off parameter λ_1 for identification loss, λ_2 for verification loss, and λ_3 for the regularization term are three important factors that can exert a noticeable impact on the performance of scene classification. Hence, we set $\lambda_1 = \{1, 0.5\}$, $\lambda_2 = \{0.5, 0.05, 0.01\}$, and $\lambda_3 = \{0.01, 0.001\}$, respectively, in our work. After extensive experiments, the set $\{\lambda_1 = 1, \lambda_2 = 0.05, \lambda_3 = 0.001\}$ has achieved the best results. In Section 3, we set the ratio between

positive training images with and without randomly rotating is $\{1:1, 1:2, 1:3, 1:4\}$. To figure out which ratio performs best, we conduct experiments on SIRI-WHU dataset. Table 1 shows that the ratio 1:1 for random rotation performs best. Hence, we use this parameter set in our subsequent training. We compare the proposed model with the CNN classifier including AlexNet, VGG16, and ResNet50 and also with the exact same architecture as our model without random rotation and joint decision making which is represented by R.D in following tables. We use the overall accuracy and Kappa coefficient to evaluate the scene classification performance of the proposed method. Overall accuracy is the percentage of the correctly classified images among all the testing set. Kappa coefficient is another widely used evaluation standard, which is based on the confusion matrix to assess the precision of remote sensing classification. Our model was trained on a PC with a 3.7-GHz 7-core CPUs and 16-GB memory. We use NVIDIA GTX

Table 1 The experiments of random rotation ratio on SIRI-WHU dataset

Method	Ratio	OA (%)	Kappa (%)
Siamese ResNet50	1:1	82.50	81.08
	1:2	81.04	79.32
	1:3	79.89	78.07
	1:4	78.59	76.75

1080 GPUs for acceleration. The whole training takes about 14 h, 20 h, and 75 h, respectively, for our models based on AlexNet, VGG16, and ResNet50.

5 Results and discussion

The training of the proposed model uses three training percentages (TP): 20%, 50%, and 80%. Three tables present the classification results on NWPU-RESISC45, UC Merced Land-Use, and SIRI-WHU datasets. Generally, from these tables, we can draw that the classification performance of our model with data augmentation and Bayesian theory is superior to the CNN classifiers and the plain model. All the methods adopted in our work have obtained the best results on the NWPU-RESISC45 dataset, which demonstrates that deep-learning-based models can achieve better performance using relatively larger and abundant training samples. The statistics in Tables 2, 3, and 4 show that the classification accuracy on SIRI-WHU data set is increased by a larger margin compared with the other two datasets using our method, which indicates that data augmentation can achieve better enhancement on the smaller scale of the original training set [26]. Compared with UC Merced Land-Use and NWPU-RESISC45 dataset, SIRI-WHU contains fewer scene classes and may lead to overfitting in some way. Our method has

Table 2 Comparison of OA (%) and Kappa (%) on NWPU-RESISC45 data set

Methods	Epoch	TP(%)					
		20		50		80	
		OA	Kappa	OA	Kappa	OA	Kappa
AlexNet	120	70.00	69.22	78.37	77.85	82.67	82.27
Siamese AlexNet	120	71.31	70.66	78.65	78.14	83.81	83.44
Siamese AlexNet+R.D	120	72.79	71.11	80.45	80.01	84.35	83.99
VGG16	120	89.75	89.51	92.91	92.74	94.59	94.46
Siamese VGG16	120	90.06	89.83	93.45	93.30	94.70	94.58
Siamese VGG16+R.D	120	91.03	90.82	94.03	93.89	95.24	95.13
ResNet50	120	90.90	90.69	94.87	93.69	95.71	95.62
Siamese ResNet50	120	92.28	92.11	94.94	94.82	95.95	95.68
Siamese ResNet50+R.D	120	92.67	92.50	95.00	94.88	96.11	96.02

Table 3 Comparison of OA (%) and Kappa (%) on UC Merced Land-Use data set

Methods	Epoch	TP(%)					
		20		50		80	
		OA	Kappa	OA	Kappa	OA	Kappa
AlexNet	120	36.01	32.82	55.05	52.80	67.70	66.00
Siamese AlexNet	120	37.56	34.44	56.76	54.60	69.52	68.00
Siamese AlexNet+R.D	120	38.89	35.44	58.29	56.20	71.62	69.00
VGG16	120	76.53	75.56	85.05	84.30	90.24	89.75
Siamese VGG16	120	76.90	75.75	85.14	84.50	92.38	92.00
Siamese VGG16+R.D	120	78.13	76.94	88.10	87.50	93.33	93.00
ResNet50	120	74.11	72.81	89.43	89.00	91.90	91.50
Siamese ResNet50	120	76.52	75.06	90.95	90.50	94.29	94.00
Siamese ResNet50+R.D	120	78.87	77.81	91.71	91.20	94.76	94.50

enhanced the classification accuracy on the NWPU-RESISC45 dataset, but still has misclassifications. This is because this dataset has not only rich image variations in terms of viewpoint, object pose, appearance, spatial resolution, background, etc., but high within-class diversity and between-class similarity. As Fig. 6 shows, it covers high-semantic overlapping scene classes such as commercial area and industrial area, circular farmland and rectangular farmland, and railway and railway station. Hence, NWPU-RESISC45 is a challenging dataset for our method.

6 Conclusion

In this paper, we propose a rotation-invariant feature learning and joint decision-making method based on Siamese convolutional neural networks which combine identification and verification models. A large number of

Table 4 Comparison of OA (%) and Kappa (%) on SIRI-WHU data set

Methods	Epoch	TP(%)					
		20		50		80	
		OA	Kappa	OA	Kappa	OA	Kappa
AlexNet	120	46.72	41.88	82.50	80.91	88.33	87.27
Siamese AlexNet	120	54.38	50.23	83.25	81.73	88.96	87.95
Siamese AlexNet+R.D	120	58.28	54.49	86.71	85.20	89.12	88.18
VGG16	120	83.65	82.16	94.42	93.91	96.25	95.90
Siamese VGG16	120	84.43	83.01	94.50	94.00	97.30	97.05
Siamese VGG16+R.D	120	86.98	85.80	95.25	94.82	98.46	97.14
ResNet50	120	58.13	54.32	94.67	94.18	95.63	95.23
Siamese ResNet50	120	63.02	59.60	95.75	95.36	97.50	97.27
Siamese ResNet50+R.D	120	82.67	81.08	96.88	96.01	98.85	98.23

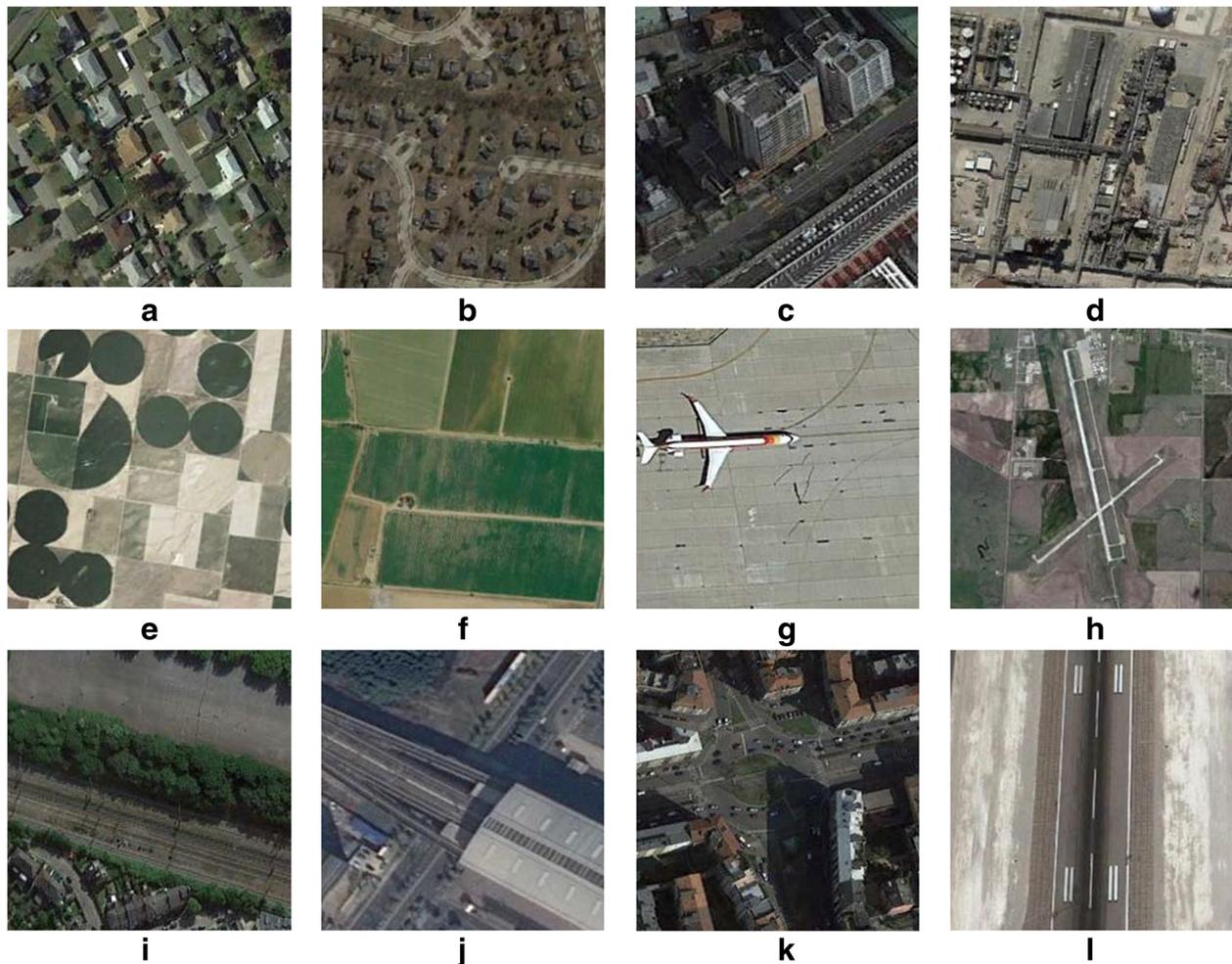


Fig. 6 Several examples from the NWPU-RESISC45 dataset, which were carefully selected under various real-world conditions including illuminations, seasons, and weather. These images have variations in object pose, appearance, spatial resolution, and background. **a** dense residential, **b** medium residential, **c** commercial area, **d** industrial area, **e** circular farmland, **f** rectangular farmland, **g** airplane, **h** airport, **i** railway, **j** railway station, **k** roundabout, and **l** runway

experiments have proved the effectiveness of the proposed method for the classification task on three widely used datasets.

In general, we have the following contributions. First of all, the random rotation is adopted in one CNN channel input while the other CNN channel remains unchanged. This kind of data augmentation not only can expand the insufficient training samples of remote sensing datasets, which makes our CNN-based method develop its full potential, but also makes the identification model learn rotation-invariant features to strengthen the classification power of our method. Secondly, a joint decision making is introduced by encoding the probabilistic relationships output by identification and verification models to classify the remote sensing images for better precision. In the future, we will try to optimize the structure of deep neural networks by means of multi-objective optimization

methods [32] to explore a more robust way to classify the challenging NWPU-RESISC45 dataset.

The datasets analyzed during the current study are available in the NWPU-RESISC45 repository (<http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.tml>), UC Merced Land-Use repository (<http://weegeevision.ucmerced.edu/datasets/landuse.html>), and SIRI-WHU repository (http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html).

Abbreviations

CNN: Convolutional neural network; RESISC: Remote sensing scene classification; RGB: Red-green-blue; SGD: Stochastic gradient descent; SIFT: Scale-invariant feature transform; TP: Training percentages

Acknowledgements

The authors would like to thank the editor and reviewers for providing valuable comments and suggestions that greatly improve the quality of this paper.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61572505, 61772530, 61806206), the State's Key Project of Research and Development Plan of China (No. 2016YFC0600900) and the Six Talent Peaks Project in Jiangsu Province (No. 2015-DZXX-010), Natural Science Foundation of Jiangsu Province (No. BK20180639), and the China Postdoctoral Science Foundation (2018 M642359).

Availability of data and materials

The datasets analyzed during the current study are available in the NWPU-RESISC45 repository (<http://www.esience.cn/people/JunweiHan/NWPU-RESISC45.html>), UC Merced Land-Use repository (<http://weegee.vision.ucmerced.edu/datasets/landuse.html>), and SIRI-WHU repository (http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html).

Authors' contributions

All the authors took part in the whole research described in this paper. Y. Zhou and JZ put forward the main idea. XL did the experiments of this research. Y. Zhou wrote the first version of this paper. The rest of the authors revised the paper in different versions. All the authors read and approved the final manuscript.

Authors' information

1. Yong Zhou. He is now a professor in China University of Mining and Technology. His research mainly focuses on data mining, machine learning, and artificial intelligence.
2. Xuning Liu. She is now a student of School of Computer and Technology in China University of Mining and Technology. Her research is mainly about machine learning and image processing.
3. Jiaqi Zhao. He received the B.Eng. degrees in intelligence science and technology in 2010, the Ph.D. degree in circuits and systems in 2017 from Xidian University, Xian, China. Between 2013 and 2014, he was an exchange Ph.D. student with the Leiden Institute for Advanced Computer Science (LIACS), University of Leiden, the Netherlands. He is currently with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current research interests include Multiobjective optimization, machine learning deep learning, and image processing.
4. Ding Ma. He is now a student of School of Computer and Technology in China University of Mining and Technology. His research is mainly about machine learning and video processing.
5. Rui Yao. He received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2013. From September 2011 to September 2012, he was a Visiting Student with the University of Adelaide, Adelaide, SA, Australia. He is currently with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current research interests include computer vision and machine learning.
6. Bing Liu. He is now an associate professor in China University of Mining and Technology. His research interests include data mining, machine learning, and artificial intelligence.
7. Yi Zheng. He is now a student of School of Computer and Technology in China University of Mining and Technology. His research interests include machine learning and person re-identification.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Received: 31 October 2018 Accepted: 19 December 2018

Published online: 07 January 2019

References

1. X. Liu, L. Jiao, J. Zhao, J. Zhao, D. Zhang, F. Liu, S. Yang, X. Tang, IEEE Transactions on Geoscience and Remote Sensing, **56**(1), 461–473 (2018)
2. J. Zhao, L. Jiao, S. Xia, V.B. Fernandes, I. Yevseyeva, Y. Zhou, M.T.M. Emmerich, Multiobjective sparse ensemble learning by means of evolutionary algorithms. *Decis. Support. Syst.* **111**, 86–100 (2018)
3. L. Li, Y. Zhao, J. Sun, R. Stolkin, Q. Pan, J.C. Chan, S.G. Kong, Z. Liu, Deformable dictionary learning for SAR image change detection. *IEEE Trans. Geosci. Remote Sens.* **56**(8), 4605–4617 (2018). <https://doi.org/10.1109/TGRS.2018.2829630>
4. L. Zhang, W. Wei, C. Bai, Y. Gao, Y. Zhang, Exploiting clustering manifold structure for hyperspectral imagery super-resolution. *IEEE Trans. Image Process.* **27**(12), 5969–5982 (2018)
5. Z. Ren, B. Hou, Z. Wen, L. Jiao, Patch-sorted deep feature learning for high resolution SAR image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**(9), 3113–3126 (2018). <https://doi.org/10.1109/JSTARS.2018.2851023>
6. J. Gu, L. Jiao, F. Liu, S. Yang, R. Wang, P. Chen, Y. Cui, J. Xie, Y. Zhang, Random subspace based ensemble sparse representation. *Pattern Recogn.* **74**, 544–555 (2018). <https://doi.org/10.1016/j.patcog.2017.09.016>
7. G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **105**(10), 1865–1883 (2017)
8. L. Li, Y. Zhou, K. Gu, W. Lin, S. Wang, Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness. *IEEE Transactions on Multimedia* **20**(4), 914–926 (2018). <https://doi.org/10.1109/TMM.2017.2760062>
9. G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, J. Ren, Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images. *IEEE Transactions on Geoscience & Remote Sensing* **53**(8), 4238–4249 (2015)
10. G. Cheng, J. Han, P. Zhou, L. Guo, Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *Isprs Journal of Photogrammetry & Remote Sensing* **98**(1), 119–132 (2014)
11. E. Aptoula, Remote sensing image retrieval with global morphological texture descriptors. *IEEE Transactions on Geoscience & Remote Sensing* **52**(5), 3023–3034 (2014)
12. J. Swain Michael, H. Ballard Dana, Color indexing. *Int. J. Comput. Vis.* **7**(1), 11–32 (1991)
13. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**(3), 145–175 (2001)
14. D.G. Lowe, in *International Journal of Computer Vision*. Distinctive image features from scale-invariant keypoints (2004), pp. 91–110
15. Dalal, Navneet, Triggs, Bill, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Histograms of oriented gradients for human detection (2005), pp. 886–893
16. Y. Guo, L. Jiao, S. Wang, S. Wang, F. Liu, W. Hua, Fuzzy superpixels for polarimetric SAR images classification. *IEEE Trans. Fuzzy Syst.* **26**(5), 2846–2860 (2018). <https://doi.org/10.1109/TFUZZ.2018.2814591>
17. Risojević, V., Babić, Z, Unsupervised quaternion feature learning for remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* **9**(4), 1521–1531 (2016)
18. P. Chen, L. Jiao, F. Liu, S. Gou, J. Zhao, Z. Zhao, Dimensionality reduction of hyperspectral imagery using sparse graph learning. *IEEE Journal of Selected Topics in Applied Earth Observations. Remote Sens.* **10**(3), 1165–1181 (2017)
19. Jian, M., Zhang, S., Wang, X., He, Y., Wu, L.: Deep key frame extraction for sport training. In: Yang, J., Hu, Q., Cheng, M.-M., Wang, L., Liu, Q., Bai, X., Meng, D. (eds.) *Computer Vision*, pp. 607–616. Springer, Singapore (2017)
20. X. Zhang, Y. Sun, J. Zhang, P. Wu, L. Jiao, Hyperspectral unmixing via deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **15**(11), 1755–1759 (2018). <https://doi.org/10.1109/LGRS.2018.2857804>
21. M.E. Paoletti, J.M. Haut, J. Plaza, A. Plaza, A new deep convolutional neural network for fast hyperspectral image classification, *ISPRS Journal of Photogrammetry and Remote Sensing.* **145**, 120–147 (2018)
22. L. Zhang, W. Wei, Y. Zhang, C. Shen, A.V.D. Hengel, Q. Shi, Cluster sparsity field: an internal hyperspectral imagery prior for reconstruction. *Int. J. Comput. Vis.* **11**, 1–25 (2018)
23. A. Krizhevsky, I. Sutskever, G.E. Hinton, in *International Conference on Neural Information Processing Systems*. Imagenet classification with deep convolutional neural networks (2012), pp. 1097–1105

24. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint, <http://arxiv.org/abs/1409.1556>, (2014)
25. K. He, X. Zhang, S. Ren, J. Sun, in *Computer Vision and Pattern Recognition*. Deep residual learning for image recognition (2016), pp. 770–778
26. Jia, S., Wang, P., Jia, P., Hu, S.: Research on data augmentation for image classification based on convolution neural networks. In: Chinese Automation Congress, pp. 4165–4170 (2018)
27. Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person re-identification. *ACM Transactions on Multimedia Computing Communications and Applications* (2017). <https://doi.org/10.1145/3159171>
28. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
29. G. Cheng, C. Yang, X. Yao, L. Guo, J. Han, When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*. **56**(5), 2811–2821, (2018)
30. B. Zhao, Y. Zhong, G. Xia, L. Zhang, Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery, in *IEEE Transactions on Geoscience and Remote Sensing*. **54**(4), 2108–2123, (2016)
31. Y. Yang, S. Newsam, in *Sigspatial International Conference on Advances in Geographic Information Systems*. Bag-of-visual-words and spatial extensions for land-use classification (2010), pp. 270–279
32. J. Zhao, L. Jiao, F. Liu, V.B. Fernandes, I. Yevseyeva, S. Xia, M.T.M. Emmerich, 3d fast convex-hull-based evolutionary multiobjective optimization algorithm. *Appl. Soft Comput.* **67**, 322–336 (2018). <https://doi.org/10.1016/j.asoc.2018.03.005>

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
