

Received January 9, 2020, accepted February 21, 2020, date of publication February 27, 2020, date of current version March 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976484

Multi-Structure Joint Decision-Making Approach for Land Use Classification of High-Resolution Remote Sensing Images Based on CNNs

LU XU¹, YIYUN CHEN^{1,2}, JIAWEI PAN¹, AND AJI GAO¹

¹School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China

²Centre of Development Studies, University of Cambridge, Cambridge CB3 9DT, U.K.

Corresponding author: Yiyun Chen (cheny@whu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFD1100801-01.

ABSTRACT Land use classification of high-resolution remote sensing (HRRS) images is a challenging and prominent problem in which pretrained convolutional neural networks (CNNs) have made amazing achievements. However, single-structured pretrained CNNs have limitations to obtain high classification accuracy. Besides, each pretrained CNNs has different classification ability to classify land use. Therefore, taking advantages of different pretrained CNNs is essential for land use classification. In this study, we propose a novel classification approach based on multi-structure joint decision-making strategy and pretrained CNNs. The basic idea is to apply three CNNs to classify land use separately with the final classification results achieved by joint decision-making strategy. The proposed approach comprises of three steps. First, we create a new fully connected layer and Softmax classification layer. We combine them with the convolutional layers of AlexNet, Inception-v3, and ResNet18. AlexNet also includes the first two layers of fully connected layers. Secondly, we train these designed CNNs to converge by momentum-driven stochastic gradient descent. Thirdly, we utilize joint decision-making strategy to obtain the final prediction results by combining the prediction results of these designed CNNs. The performance of the proposed approach is evaluated on the UC Merced land use, AID, NWPU-45, OPTIMAL-31 datasets and further compared with the state-of-the-art methods. Results demonstrate that the proposed approach outperforms other methods. The benefits of the proposed approach are threefold. First, the multi-structure network maximizes different pretrained CNN structures to extract rich convolution features. Secondly, it could remarkably improve the classification accuracy of indistinguishable land use types of the HRRS images. Thirdly, it has great potential on small dataset with more land use types. The proposed CNN based on multi-structure joint decision approach achieves accurate and reliable land use classification with these benefits.

INDEX TERMS Land use classification, convolutional neural network, transfer learning, high-resolution remote sensing images, multi-structure.

I. INTRODUCTION

Land is an indispensable component of human production and life [1]. Precise and refined land use classification is essential for land resource management [2], [3], urban planning [4]–[6], precision agriculture [7], [8], environmental protection [9], and sustainable development [10]. The same land use types have complex spatial and structural patterns, whereas different land use types may have similar reflectance

spectra and textures [11], [12]. The precise classification of land use types has become a difficult issue due to high intraclass heterogeneity and low interclass diversity [13]–[15]. With the recent advancement of remote sensing technology and the increase in the number of satellites, numerous high-resolution remote sensing (HRRS) images are being obtained. HRRS images cover substantial land use information, which not only provides opportunities for land use classification but also brings new challenges.

Currently, HRRS image land use classification strategies can be categorized into three classes: (1) low-level visual

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico¹.

feature-based, (2) midlevel visual representation-based, and (3) high-level vision information-based methods [14]. The first classification strategy is based on low-level visual attributes of remote sensing images. Such attributes include spectrum, texture, and structure, which rely on the color histogram (CH), local binary pattern (LBP), and scale invariant feature transform (SIFT) to extract features for classification [14], [16]–[19]. However, this strategy relies heavily on low-level features and generally performs poorly in generalization, especially when the scene is complex and diverse. To compensate, scholars have proposed the second classification strategy based on mid-level visual representation. They designed the bag-of-visual-words model (BOVW), spatial pyramid co-occurrence kernel (SPM), probabilistic latent semantic analysis (pLSA), and other approaches to encode low-level features locally and improve classification accuracy [20]–[22]. This strategy requires rich expert experience and manual design encoding rules, which are difficult and subjective. Given these challenges, the classification strategy based on high-level vision information is proposed. The advantage of this strategy is considered to acquire abstract and discriminative semantic from HRRS images features automatically and convert the classification task into an end-to-end problem. It could avoid the complexity of data reconstruction in feature extraction and classification. Typical methods based on high-level vision information include restricted boltzmann machine (RBN) [23], deep belief networks (DBN) [24], stacked autoencoder (SAE) [25], [26], convolutional neural networks (CNNs) [27]–[33], etc.

Recently, CNNs have achieved remarkable achievements in the field of HRRS images land use classification [34]. Most of the recent works are accustomed to applying advanced mathematical methods to enhance the feature representation ability of a single pretrained CNN and improve overall classification accuracy. Nevertheless, the ability of a single-structured pretrained CNN to express image features and classifications is limited. This limitation brings unbalanced classification accuracy of different land use types and “short board effect”, which is responsible for an unsatisfactory overall classification accuracy. It can be solved by finding alternative ways to promote the classification performance.

In this paper, we propose a multi-structure joint decision approach based on CNN (MJDCNN) and evaluate its performance in four HRRS image datasets of land use. Specifically, we aim to (1) establish MJDCNN for HRRS image classification based on three different pretrained CNNs (i.e. AlexNet, Inception-v3 and ResNet18) and joint decision-making strategy; (2) explore the effect of network training iterations on the MJDCNN’s classification performance; (3) evaluate the performance of MJDCNN and the three pretrained CNNs in terms of the overall classification accuracy, F1-score, and the classification accuracy of single land use type; and (4) compare the overall classification accuracies of MJDCNN and other state-of-the-art classification methods.

The remainder of this paper is organized as follows: Section II summarizes related works on CNN-based

methods. Section III introduces experimental methods and strategies, including pretrained CNNs and the proposed MJDCNN. Section IV shows the experimental results and analysis. Section V discusses the classification performance of MJDCNN. The last section draws the conclusions.

II. RELATED WORKS

CNNs utilize convolution kernels to convolute calculations on images and extract simple to complex features for classification. Training a capable and well-learned CNN requires substantial high-quality training sample data with labels, high computational performance configuration, and an immense amount of time. With the progress of technology, problems related to computing performance and training time can be mitigated. However, in the case of the explosive growth of HRRS images, high-quality training sample data with labels remains scarce and difficult to obtain. The existing land use HRRS image datasets mostly include UC Merced land use (21 categories with 2100 images) [22], WHU-RS19 (19 categories with 1,005 images) [35], RSSCN7 (7 categories with 2,800 images) [24], AID (30 types with 10,000 images) [14], NWPU-45 (45 categories with 31,500 images) [25] and OPTIMAL-31 (31 categories with 1,860 images) [36] datasets, which are all small. Training a new CNN with a small dataset is prone to overfitting, and the training effect and practicability are poor.

To overcome this problem, transfer learning can utilize the similarity between tasks, adapt the parameters of pretrained CNNs from the original task to land use classification tasks, and effectively solve the data scarcity of land use images [37]. Researchers have trained several CNNs [38]–[41]. The refined classification of HRRS images of land use based on pretrained CNNs has become an important trend in the current research in accordance with the similarity of image structure and classification tasks. Currently, two mainstream transfer learning strategies are available. The first is to depend on the pretrained CNNs to extract the feature vectors of HRRS images and then input them into other classifiers for classification [42]–[46]. Boualleg *et al.* used the well-known VGGNet-16 to extract deep features and then fed to deep forest classifier for classification, which saved a considerable amount of training time [43]. Weng *et al.* obtained image feature vectors through AlexNet, then classified images by extreme learning machine. They further compared its performance with that of Bayes, KNN, and SVM classifiers [44]. Zhou *et al.* explored the effects of feature representations from different fully connected layers of pretrained CNNs for SVM classifiers [45]. The second strategy is to input HRRS images of land use in fine-tuning pretrained CNNs for classification [47]–[53]. Liang *et al.* examined the performance of different pretrained CNNs under different fine-tuning strategies [47]. Liu *et al.* and Alias *et al.* shared low-level parameters and fine-tuned the upper-level parameters of VGGNet-16 to achieve land use classification, and remarkably improved the accuracy [48], [49].

Moreover, researchers improved fine-tuning strategies with other novel methods. For example, Scott *et al.* combined fine-tuning DCNNs and data augmentation to increase the overall accuracies [50]. Yang *et al.* presented the Drop-Band technique to prevent fine-tuning CNNs overfitting [53]. Liu *et al.* first proposed attention mechanism based on CNN to obtain outstanding classification results by integrating local and global features [51]. Zhang *et al.* designed a gradient boosting random method based on CNN that overcome the limitation of training dataset and reduce the computation complexity [52]. Korshunova added a fuzzy self organization layer in the last convolution layer of a pretrained CNN and named it convolutional fuzzy neural network. This method brought fuzzy logic in CNN, thereby improving the capacity of handling uncertainty and impreciseness [54].

Despite the above achievement, the ability of a single pretrained CNN to express image features and perform image classifications is limited. Thus, mutli-CNNs are proposed to compensate for this limitation. For instance, Alhichri *et al.* proposed a three-branch deep CNN based on SqueezeNet, which could accept multi-scale image sizes and improve classification performance [55]. Liu *et al.* proposed a random-scale stretched CNN to classify each image under multiple times and obtain the final results by voting. In this manner, multiscale images could be obtained and CNN could learn additional robust features. This method solved the influence of image scale variation on classification performance [56]. Liu *et al.* also combined a fixed-scale net and a varied-scale net to solve the scale variation for remote sensing image classification [57]. Sun *et al.* proposed a gated bidirectional network to merge semantic-assist and appearance-assist information to perform classification, which could enhance important representation and reduce interference information [58]. Zheng *et al.* utilized multi-scale pooling to enrich image features that extracted from the last convolutional layer [59]. Lu *et al.* aggregated different features from different convolution layers through an end-to-end feature aggregation CNN [60]. Yang *et al.* fused different scale image features that extracted from the convolutional and fully-connected layers to multi-kernel SVM classifier, which achieved good classification [61]. Du *et al.* and Li *et al.* proposed a multiscale improved Fisher kernel coding approach to integrate multilayer features from multiple CNNs [62], [63]. Chaib *et al.* utilized discriminant correlation analysis to fuse image features that extracted from fully-connected layers of VGG-Net [46]. They proved that this method had the advantages of convenient operation, low cost, less training time and good accuracies. However, these methods only consider multi-scale and multi-layer visual representations of HRRS images, with little attention being paid to the influence of different CNN structures.

CNNs considerably vary in their ability to capture features from different land-use types. Simple CNNs are likely to extract features of simple land-use types, such as farmlands and forests, and cannot express abstractive features, such as dense residential and sparse residential. Complicated CNNs

TABLE 1. The details of three pretrained CNNs.

Pretrained CNNs	Number of Layers	Parameters (Millions)	Images Input Size	Structure Characteristics
AlexNet	8	60	227×227	-
Inception-v3	48	23.9	299×299	Inception Module
ResNet18	18	11.7	224×224	Shortcut Connections

are better than simple CNNs at capturing complicated features, but suffering from negative effects of redundant features. Therefore, it is beneficial to take advantages of different CNNs through jointing decision-making.

III. PROPOSED METHODS

A. CNNs

CNNs are one of the most widely used methods in deep learning. They learn the image general features in shallow convolutional layers and combine these features to describe images in a bottom-up manner. The entire process implements mapping from low-level features to high-level semantics. It consists of a multi-layer structure, which is divided into convolution, pooling, activation, fully connected, and classification layers [64].

The convolution layers are used to convolve images to extract their features. They can reduce the number of network weights through local connection and weight sharing. In the convolutional layer structure, the obtained features become increasingly inclined to the images' specific characteristics as the convolution layers continue to deepen. The pooling layers are used to reduce the dimension of the convolutional layers. Thus, they reduce the size of input data, improve computational efficiency, and prevent dimensionality disasters. The activation layers are a nonlinear operation of the results of the convolutional layers or the pooling layers. The fully connected layers summarize feature vectors obtained from the convolutional layers, forming image global representation. The classification layer is used to classify the feature vectors formed by the fully connected layers into the classifier to obtain the image class.

B. PRETRAINED CNNs

A mass of differently structured pretrained CNNs with good classification performance are available. This study selects three pretrained CNNs with completely different structures, including AlexNet, Inception-v3, and ResNet18, to create MJDCNN. The details of three pretrained CNNs are shown in Table 1.

AlexNet, with a total of 60 million parameters and 650,000 neurons, belongs to the traditional CNN. The weighted layer composes of five convolutional layers and three fully connected layers [41].

Inception-v3, with a total of 23.9 million parameters and a depth of 48 layers, is a sparsely connected CNN. It is characterized by the inclusion of the inception module. This module has multiple convolution kernels of different scales in the same convolutional layer, which balances the width and depth of the CNN [40].

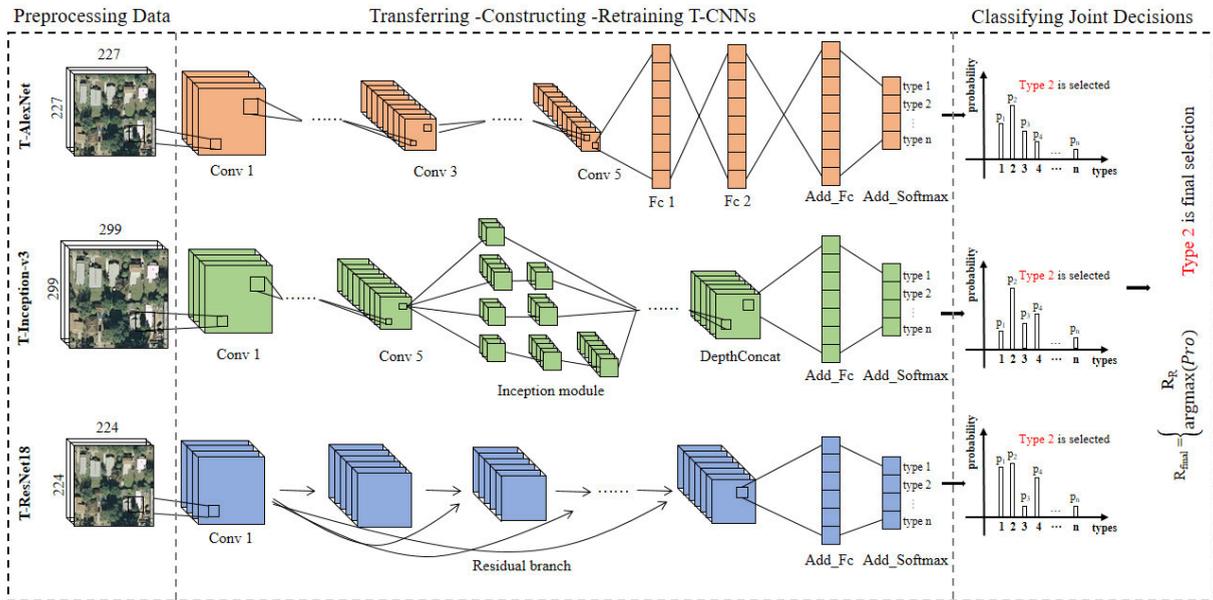


FIGURE 1. Flowchart of MJDCNN classification.

ResNet18, with a total of 11.7 million parameters and a depth of 18 layers, is a CNN of residual structure. It can solve accuracy degradation, which is caused by the increase in network depth. A forward neural network with “short-cut connections” is adopted to train this CNN. Shortcut connection skips one or more layers. Such strategy does not add additional parameters or increase computational complexity [42].

AlexNet has the least number of hidden layers in the champion networks of ImageNet LSVRC over the years. Its structure, which only deepens in the longitudinal, is the simplest. The structure of Inception-v3 is the most complex in terms of vertical deepening and horizontal widening. The structure of ResNet18 is moderate in complexity, and its depth is between the first two CNNs. In addition, the pretrained CNNs strictly define the size of the input data. The three selected pretrained CNNs require different sizes of input data. We can ensure the diversity of the input data and eliminate the lack of image information by reshaping the image size. In accordance with the structural characteristics of the three CNNs, Multi-structure Joint Decision Convolutional Neural Network (MJDCNN) is developed to obtain final classification results.

C. MJDCNN

MJDCNN is proposed through the following five steps: pre-processing data, transferring pretrained CNN, constructing T-CNN, retraining T-CNN, and classifying joint decision (Figure 1).

1) PREPROCESSING DATA

The input size of each pretrained CNN is strictly regulated. We randomly reshape the HRRS images of land use into regulated size.

2) TRANSFERRING PRETRAINED CNNs

All convolutional layer parameters of the three pretrained CNNs are transferred. AlexNet also includes the first two layers of the fully connected layers. When a CNN implements the classification task, the starting convolution layer extracts simple image features, such as textures, edges, and colors. As the number of convolution layers deepens, the image features extracted by the convolutional layers gradually become increasingly inclined from the general features to the features of the original dataset. Thus, the parameters of all convolutional layers are transferred [42]. The structure of AlexNet is relatively simple with less convolution layers. In contrast to the other two networks, AlexNet has three fully connected layers. Reducing any layer may result in the degradation of network classification performance. Therefore, we transfer the first two fully connected layers from AlexNet to ensure optimal feature extraction.

3) CONSTRUCTING T-CNNs

We reconstruct a new fully connected layer and a Softmax classification layer after transferring the pretrained CNNs. The fully connected layer aggregates the feature vectors obtained by the convolutional layer and then forms a global representation of HRRS images. HRRS images of land use and natural optical images have different global features. The parameters of the fully connected layers in the pretrained CNNs are specific to the natural optical images. Therefore, reconstructing the fully connected layer is necessary. The Softmax layer outputs the classification results of the CNN. In this study, we name the redesigned CNNs as T-AlexNet, T-Inception-v3, and T-ResNet18.

4) RETRAINING T-CNNs

We train the redesigned networks using momentum-driven stochastic gradient descent algorithms. The transferred lay-

ers use the initial parameters, whereas the fully connected layer parameters are randomly initialized. Then we input the training dataset and the corresponding land use type into the T-CNNs. Afterward, we calculate the entropy loss as the cost function to update the weights of T-CNNs. The formula is as follows:

$$Loss = -\frac{1}{K} \sum_{k=1}^K \left[y^k \log(t^k) + (1 - y^k) \log(1 - t^k) \right] \quad (1)$$

where t^k denotes the k th sample that corresponds to the predicted land use type, y^k denotes the k th sample that corresponds to the actual land use type, K is the number of training samples, and $Loss$ is the cross entropy of the K samples.

5) CLASSIFYING JOINT DECISIONS

The Softmax classifier is widely used given its has good application effect in ImageNet classification. In addition, its classification accuracy and running time have shown satisfactory results. This classification principle uses the Softmax function to output the probability of all types of images according to the feature vectors of the inputted images, and the image belongs to the category with the highest probability.

The Softmax classifier and joint decision-making approach are used to achieve the joint decision classification function. According to three classification results from T-AlexNet, T-Inceptionv3, and T-ResNet18 based on Softmax classifier, we combine the three classification results to determine the final classification through joint decision-making approach. The details of joint decision-making approach are as follows:

(i) if the classification results of T-AlexNet, T-Inception-v3, and T-ResNet18 are the same, the final classification result is the classification result of T-ResNet18;

(ii) if the classification results of T-AlexNet, T-Inception-v3, and T-ResNet18 are not the same, the final classification result is the land use type which has the maximum weighted average of prediction probability. The weighted average of prediction probability is calculated by prediction probability of T-AlexNet, T-Inception-v3, and T-ResNet18. According to the performance of these three T-CNNs, we define the weight ratio of the prediction probability of T-AlexNet, T-Inception-v3, and T-ResNet18 is 1:2:3. The formulas are as follows:

$$R_{final_i} = \begin{cases} R_{R_i}, & R_{A_i}, R_{I_i}, \text{ and } R_{R_i} \text{ are the same} \\ \text{argmax}(Pro_{ij}), & \text{otherwise} \end{cases} \quad (2)$$

$$Pro_{ij} = \frac{1}{6} \left(\frac{\exp(\theta_{A_j}^T x_i)}{\sum_{m=1}^M \exp(\theta_{Am}^T x_i)} + 2 \frac{\exp(\theta_{I_j}^T x_i)}{\sum_{m=1}^M \exp(\theta_{Im}^T x_i)} + 3 \frac{\exp(\theta_{R_j}^T x_i)}{\sum_{m=1}^M \exp(\theta_{Rm}^T x_i)} \right), \quad i \in 1, \dots, K; j \in 1, \dots, M \quad (3)$$

where R_{final_i} denotes the final classification result of MJDCNN of the i th sample; R_{A_i} , R_{I_i} , and R_{R_i} denote the classification results of T-AlexNet, T-Inception-v3, and T-ResNet18 of the i th sample, respectively; Pro_{ij} represents

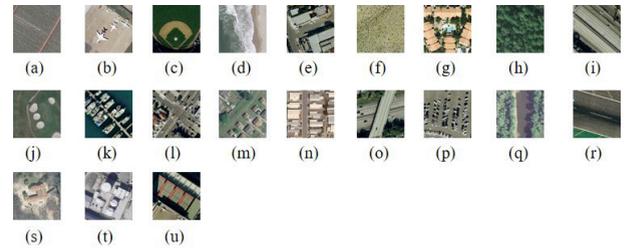


FIGURE 2. Examples of 21 land use categories in the UC Merced land use dataset: (a) agriculture; (b) airplane; (c) baseball diamond; (d) beach; (e) buildings; (f) chaparral; (g) dense residential; (h) forest; (i) freeway; (j) golf course; (k) harbor; (l) intersection; (m) medium density residential; (n) mobile home park; (o) overpass; (p) parking lot; (q) river; (r) runway; (s) sparse residential; (t) storage tanks; (u) tennis courts.

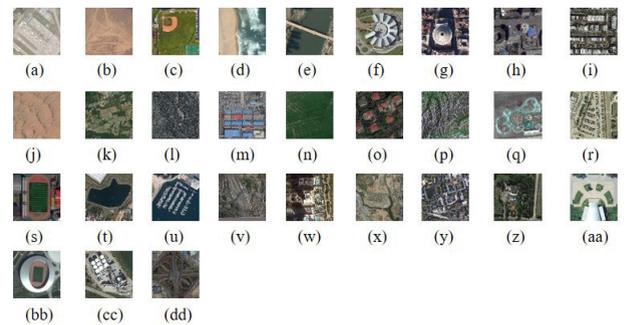


FIGURE 3. Examples of 30 land use categories in the AID dataset: (a) airport; (b) bareland; (c) baseball field; (d) beach; (e) bridge; (f) center; (g) church; (h) commercial; (i) dense residential; (j) desert; (k) farmland; (l) forest; (m) industrial; (n) meadow; (o) medium residential; (p) mountain; (q) park; (r) parking; (s) playground; (t) pond; (u) port; (v) railway station; (w) resort; (x) river; (y) school; (z) sparse residential; (aa) square; (bb) stadium; (cc) storage tanks; (dd) viaduct.

the weighted average prediction probability of the j th land use type of the i th sample; θ_{A_j} , θ_{I_j} , and θ_{R_j} are the weights of the j th land use type of the i th sample under T-AlexNet, T-Inception-v3, and T-ResNet18, respectively; x_i is the output of the last fully-connected layer of the i th sample; M is the number of land use types; K is the number of training samples.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DATASET

The experiment trains and tests MJDCNN using UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets.

The UC Merced land use dataset, which is composed of 2100 aerial images, is collected from the map downloaded by the US Geological Survey. A total of 21 land use types, wherein each type is composed of 100 256 × 256 pixel images with a spatial resolution of 0.3 m, are available. The 21 land use types are shown in Figure 2 [22].

The AID dataset, which is composed of 10000 aerial images, is collected from Google Earth. A total of 30 land use types, wherein the number of each type changes from 220 to 420, are available. The images have the size of 600 × 600 pixels with a resolution from 8m to 0.5m. The 30 land use types are shown in Figure 3 [14].

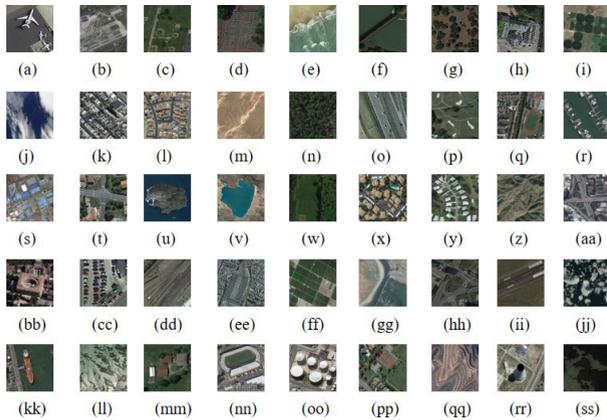


FIGURE 4. Examples of 45 land use categories in the NWPU-45 dataset: (a) airplane; (b) airport; (c) baseball diamond; (d) basketball court; (e) beach; (f) bridge; (g) chaparral; (h) church; (i) circular farmland; (j) cloud; (k) commercial area; (l) dense residential; (m) desert; (n) forest; (o) freeway; (p) golf course; (q) ground track field; (r) harbor; (s) industrial area; (t) intersection; (u) island; (v) lake; (w) meadow; (x) medium residential; (y) mobile home park; (z) mountain; (aa) overpass; (bb) palace; (cc) parking lot; (dd) railway; (ee) railway station; (ff) rectangular farmland; (gg) river; (hh) roundabout; (ii) runway; (jj) sea ice; (kk) ship; (ll) snowberg; (mm) sparse residential; (nn) stadium; (oo) storage tank; (pp) tennis court; (qq) terrace; (rr) thermal power station; (ss) wetland.

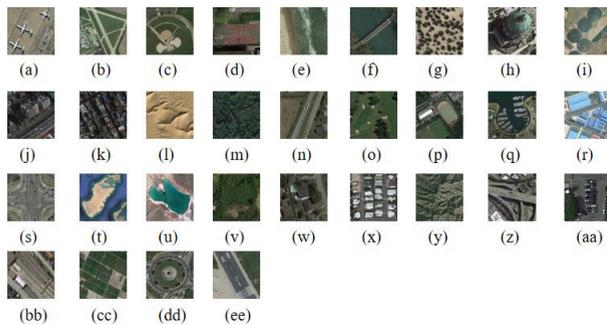


FIGURE 5. Examples of 31 land use categories in the OPTIMAL-31 dataset: (a) airplane; (b) airport; (c) baseball field; (d) basketball court; (e) beach; (f) bridge; (g) chaparral; (h) church; (i) round farmland; (j) business district; (k) dense houses; (l) desert; (m) forest; (n) freeway; (o) golf field; (p) playground; (q) harbor; (r) factory; (s) crossroads; (t) island; (u) lake; (v) meadow; (w) medium houses; (x) mobile house area; (y) mountain; (z) overpass; (aa) parking lot; (bb) railway; (cc) square farmland; (dd) roundabout; (ee) runway.

The NWPU-45 dataset, which is composed of 31500 aerial images, is collected from Google Earth. A total of 45 land use types, wherein each type is composed of 700 256×256 pixel images with a spatial resolution that varies from about 30 m to 0.2 m, are available. The 45 land use types are shown in Figure 4 [25].

The OPTIMAL-31 dataset, which is composed of 1860 aerial images, is collected from Google Earth. A total of 31 land use types, wherein each type is composed of 60 256×256 pixel images with a spatial resolution of 0.3 m, are available. The 31 land use types are shown in Figure 5 [36].

B. PARAMETERS SETTING

The experiment uses the MATLAB deep learning toolbox as the deep learning framework. The experimental environment

TABLE 2. Selected values for the training parameters of the CNNs.

Parameters	Value
Iterations	5/10/20
Learning rate	0.001
Mini-batch size	32
Validation frequency	10
Number of repeated experiments	10

is Linux system and uses an NVIDIA Quadro P4000 GPU to accelerate the calculation. The ratio of train set, validation set, and test set ratio is 4:2:4. The learning rate is 0.001 and the mini-batch size is 32. The validation set verifies the accuracy and loss rate every 10 steps. The numbers of iterations are 5, 10, and 20. The experiment is repeated 10 times. The mean results of the experiment repeated for 10 times are as the final results. These parameters are provided in Table 2.

C. RESULTS AND ANALYSIS

1) OVERALL CLASSIFICATION ACCURACIES (OAS)

OAs reflect the overall performance of the CNNs. The OAs of the proposed MJDCNN and three single-structured CNNs vary on different training iterations for different datasets, shown in Figure 6.

a: UC MERCED LAND USE DATASET

The proposed MJDCNN has the highest OA among the single-structured CNNs in the same iterations, as shown in Figure 6(a). Among these three single-structured CNNs, T-ResNet18 has the best OA, whereas T-AlexNet has the lowest. When the number of iterations is 20, MJDCNN's OA reaches 95.79%, whereas those of T-AlexNet, T-Inception-v3, and T-ResNet18 are 90.02%, 91.89%, and 93.95%, respectively. MJDCNN's OA increases by 5.77%, 3.90%, and 1.84% compared with that of the three single-structured networks. In the different iterations, the OA of each CNN has similar variation trend. The OAs of MJDCNN and T-AlexNet increase gradually as the number of iterations increases. The OAs of T-Inception-v3 and T-ResNet18 increase slowly as the number of iterations increases. These CNNs have the best OAs when the iterations are 20, whereas the worst OAs are obtained when the iterations are 5. The OA of MJDCNN increases from 94.85% to 95.79% when the number of iterations is from 5 to 20.

b: AID DATASET

Similar to the UC Merced land use dataset, MJDCNN has the best OA and T-AlexNet has the lowest OA under the same iterations, as shown in Figure 6(b). When the number of iterations is 20, MJDCNN's OA reaches 94.18%, and those of T-AlexNet, T-Inception-v3, and T-ResNet18 are 87.91%, 92.05%, and 92.58%, respectively. MJDCNN's OA increases by 6.27%, 2.13%, and 1.60% compared with that of the three single-structured networks. Moreover, the OA of each CNN has a similar increasing trend when the number of iterations is from 5 to 20. The OAs under 20 iterations are always higher than under 5 and 10 iterations for four CNNs. For example, the OAs of MJDCNN are 93.19%, 93.94%, 94.18% in 5, 10, 20 iterations, respectively.

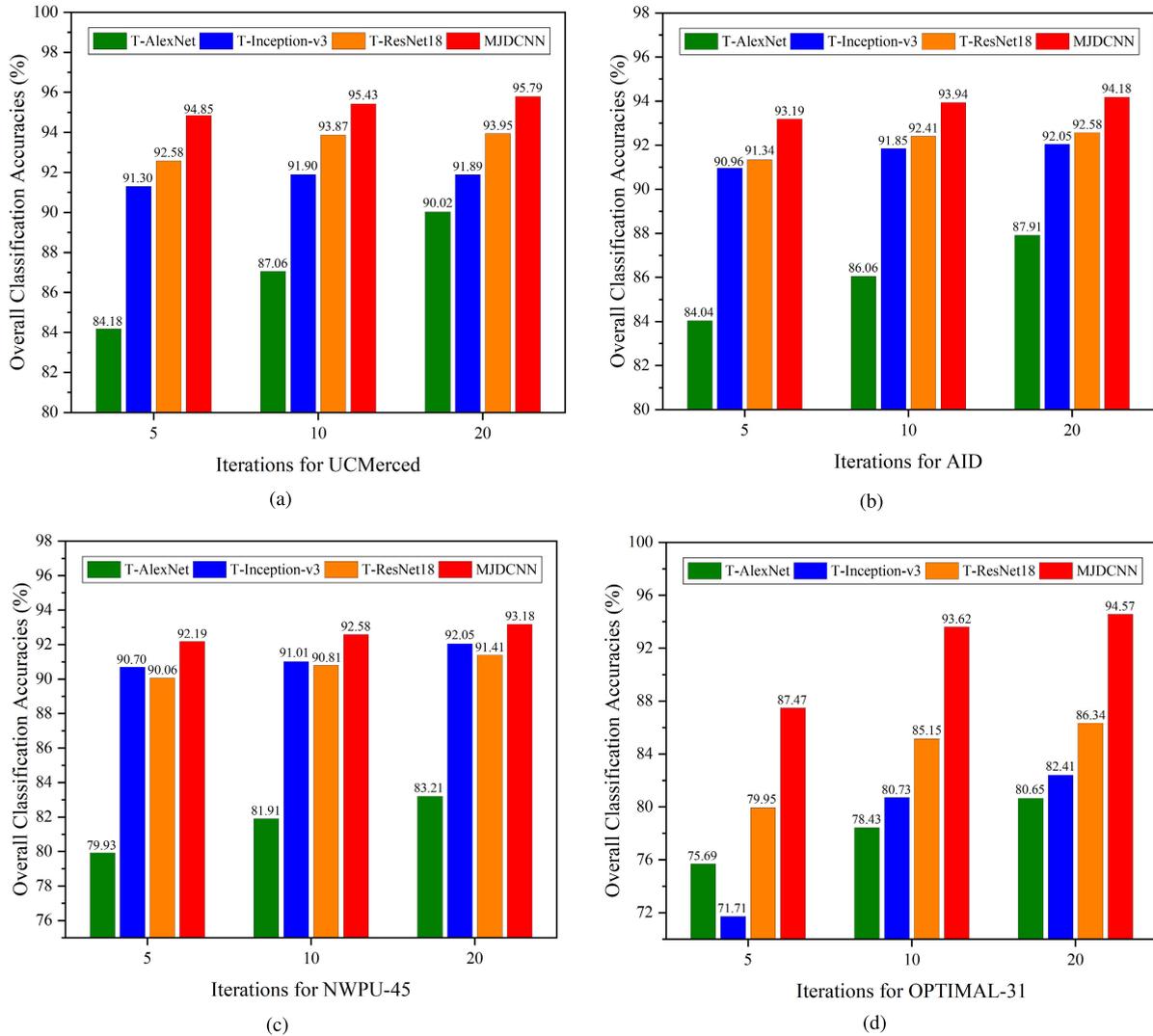


FIGURE 6. Overall classification accuracies of T-CNNs in different iterations for UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets. The overall classification accuracies of T-AlexNet, T-Inception-v3, T-ResNet18, and MJDCNN when the number of iterations is 5, 10 and 20 are shown from left to right.

c: NWPU-45 DATASET

Similarly, the performance of the MJDCNN is the best and the performance of T-AlexNet is the worst under the same iterations, as shown in Figure 6(c). When the number of iterations is 20, MJDCNN’s OA reaches 93.18%, and those of T-AlexNet, T-Inception-v3, and T-ResNet18 are 83.21%, 92.05%, and 91.41%, respectively. MJDCNN’s OA increases by 9.97%, 1.13%, and 1.77% compared with other CNNs. As the number of iterations increases, the OAs of four CNNs increase gradually as well. The OAs of MJDCNN have risen from 92.19% to 93.18%. Moreover, the OAs of T-ResNet18 are always poorer than that of T-inception-v3.

d: OPTIMAL-31 DATASET

As shown in Figure 6(d), the proposed MJDCNN remains to have the best results. Among these three single-structured CNNs, T-ResNet18 has the best OA.

Under 20 iterations, MJDCNN’s OA reaches 94.57%, whereas those of T-AlexNet, T-Inception-v3, and T-ResNet18 are 80.65%, 82.41%, and 86.34%, respectively. MJDCNN’s OA increases by 13.92%, 12.16%, and 8.23% compared with that of the other CNNs, exhibiting a significant extent of increase. In the different iterations, the OA of each CNN has slightly different trend. When the number of iterations is 5, the OA of T-AlexNet is better than that of T-Inception-v3. By contrast, the result is inverted when the number of iterations is 10 and 20. When the number of iterations is from 5 to 20, the OA of MJDCNN increases from 87.47% to 94.57%.

2) F1-SCORE

F1-score comprehensively evaluates the classification ability of the CNNs. It is obtained by dividing the arithmetic mean by the geometric mean of recall and precision. Precision and

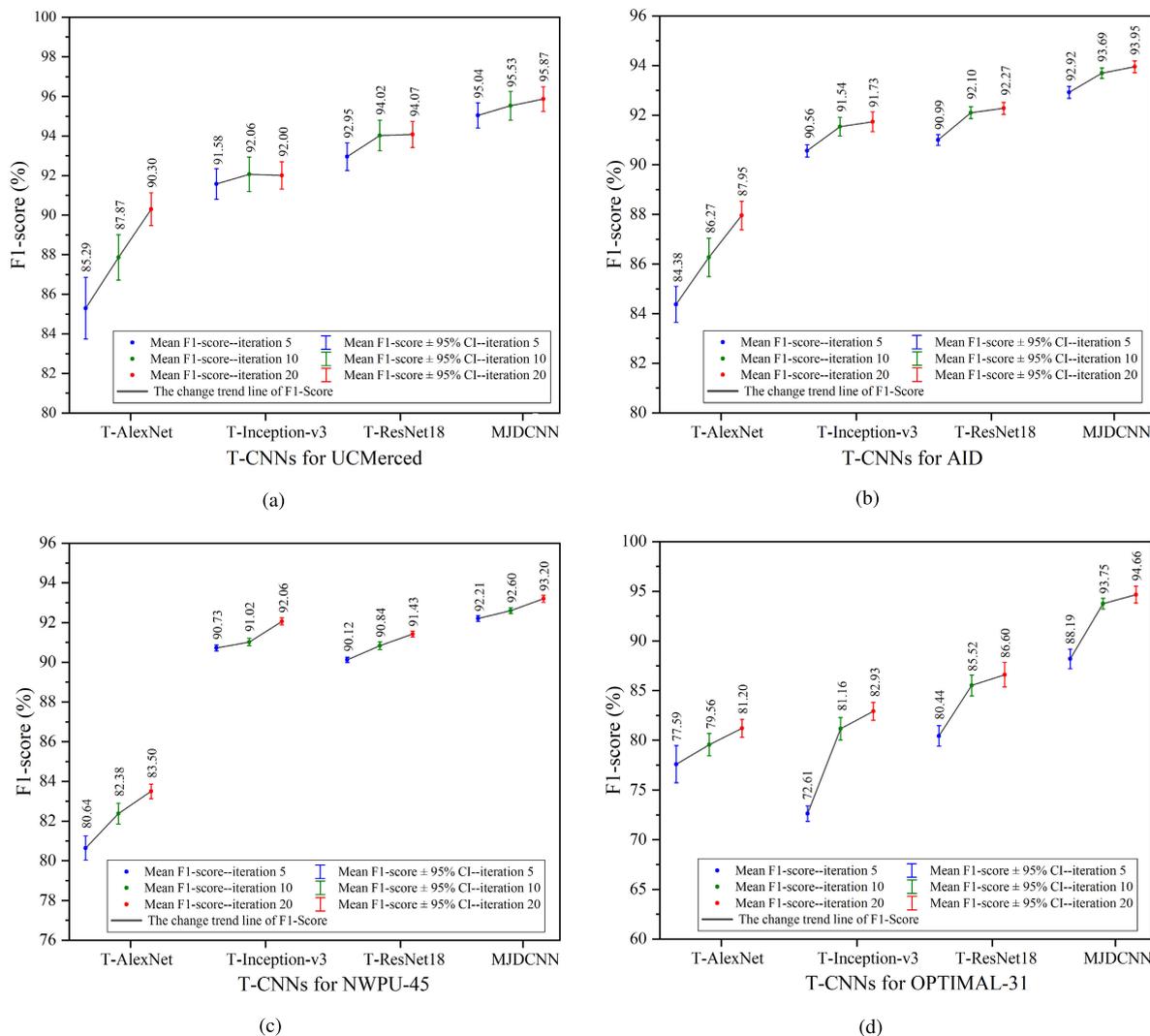


FIGURE 7. F1-Score of CNNs under different iterations for the UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets. The F1-score of T-AlexNet, T-Inception-v3, T-ResNet18, and MJDCNN when the number of iterations is 5, 10 and 20 are shown. The blue circle, the green circle and the red circle respectively represent F1-score when the number of iterations is 5, 10 and 20. The gray-colored line represents the change trend of CNN’s F1-score.

recall are two indicators that measure the performance of the classification from two aspects. Precision is the ratio that is accurately predicted to be positive for all forecasts. Recall is the proportion that is accurately predicted to be positive for all in reality. The performance of CNNs is stronger when the F1-score is higher.

Figure 7 shows the comparison of the F1-score of each CNN in different iterations for the UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets. The three points from left to right on each fold line represent the F1-score when the number of iterations is 5, 10, and 20. As the number of iterations increases, the F1-score shows a growing trend as well.

a: UC MERCED LAND USE DATASET

The F1-score of T-AlexNet increases rapidly, and that of the other CNNs increase slowly. MJDCNN’s F1-score is always larger than the other three single-structured CNNs.

In the single-structured CNNs, T-ResNet18 has the highest F1-score, followed by T-Inception-v3, and T-AlexNet successively (Figure 7(a)).

b: AID DATASET

The growth trends of F1-scores of the four CNNs are similar to that under UC Merced land use dataset. The performance of MJDCNN is the best in all scenarios. When the number of iterations is 20, the MJDCNN’s F1-score reaches 93.95%. In the single-structured CNNs, T-ResNet18 still performs excellently compared with others (Figure 7(b)).

c: NWPU-45 DATASET

The four CNNs have approximate changes of the F1-score. The F1-score of MJDCNN remains the best, reaching a value of 93.20%. For the single-structured CNNs, the T-Inception-v3’s F1-score is slightly larger than that of T-ResNet18 (Figure 7(c)).

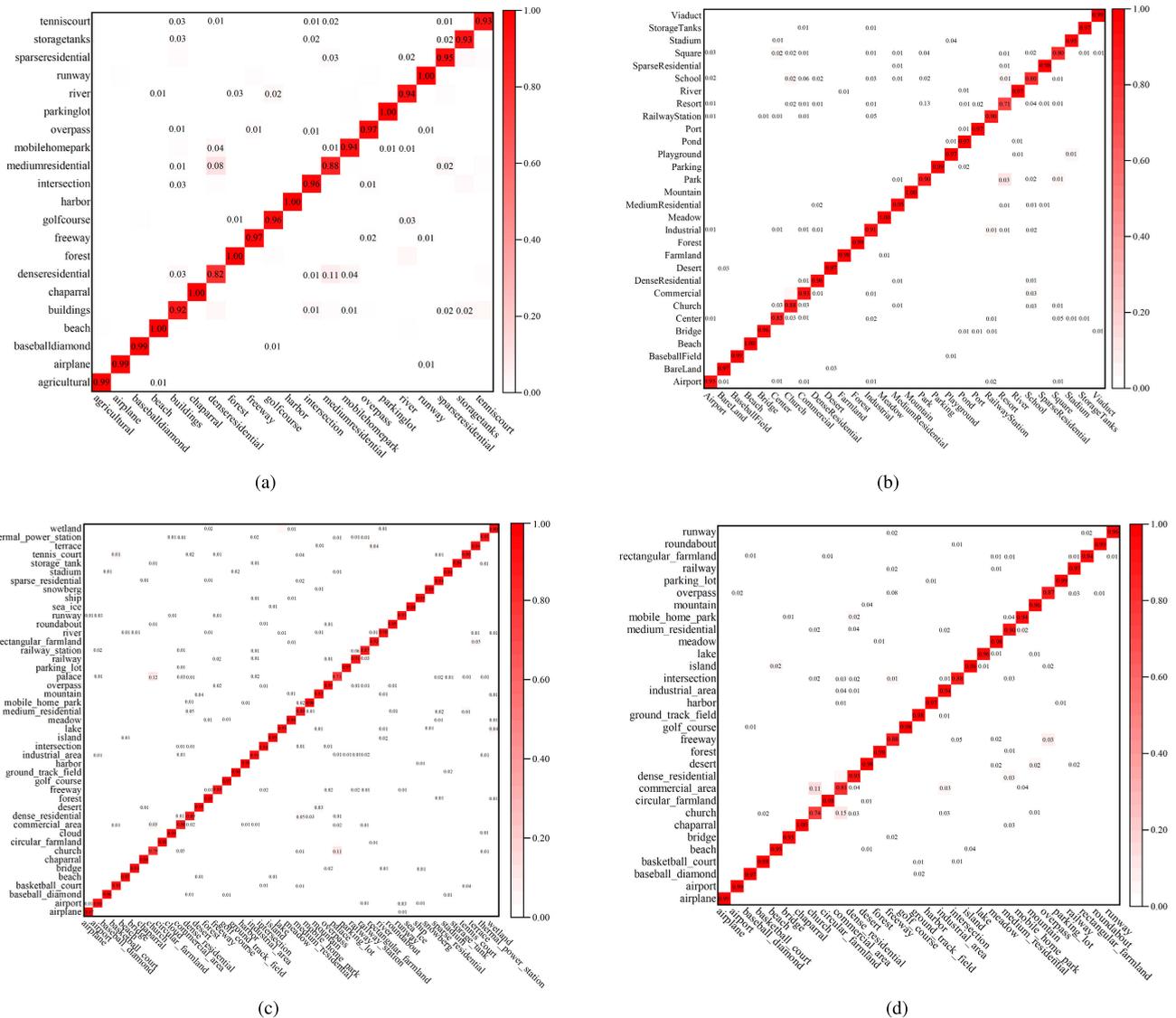


FIGURE 8. Confusion matrices of MJDCNN for UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets when the number of iterations is 20 ((a) Confusion matrix of UC Merced land use dataset; (b) Confusion matrix of AID dataset; (c) Confusion matrix of NWPU-45 dataset; (d) Confusion matrix of OPTIMAL-31 dataset).

d: OPTIMAL-31 DATASET

The F1-score of MJDCNN is always the best among the three single-structured CNNs, reaching a value of 94.66%. It grows instantly from 5 iterations to 10 iterations but slowly from 10 iterations to 20 iterations. Some differences are observed among the three single-structured CNNs. T-Inception-v3 performs unsteadily. Its performance is worse than T-AlexNet’s performance when the number of iterations is 5 and 20 but better than the T-AlexNet’s performance when the number of iterations is 10. The performance of T-ResNet18 is only second to that of MJDCNN (Figure 7(d)).

3) CLASSIFICATION ACCURACY OF SINGLE LAND USE TYPE

An effective classifier should have not only high OA but also a uniform distribution of single land use type classification accuracy.

We calculate and illustrate the mean classification accuracy of each land use type of CNNs based on MJDCNN and the three single-structured CNNs under the iteration number of 20, shown in Table 3. Confusion matrices of MJDCNN for UCMerced land use, AID, NWPU-45 and OPTIMAL-31 datasets under the iteration number of 20 are showed in Figure 8. Figure 9 shows that the increase degree of classification accuracy of single land use type for MJDCNN under the iteration number of 20 compared with the best classification accuracy of single land use type among T-AlexNet, T-Inception-v3, and T-ResNet18.

a: UC MERCED LAND USE DATASET

As shown in Table 3, the maximum accuracy of the three single-structured CNNs is similar, whereas the minimum accuracy gap is large. The difference between T-Alexnet’s

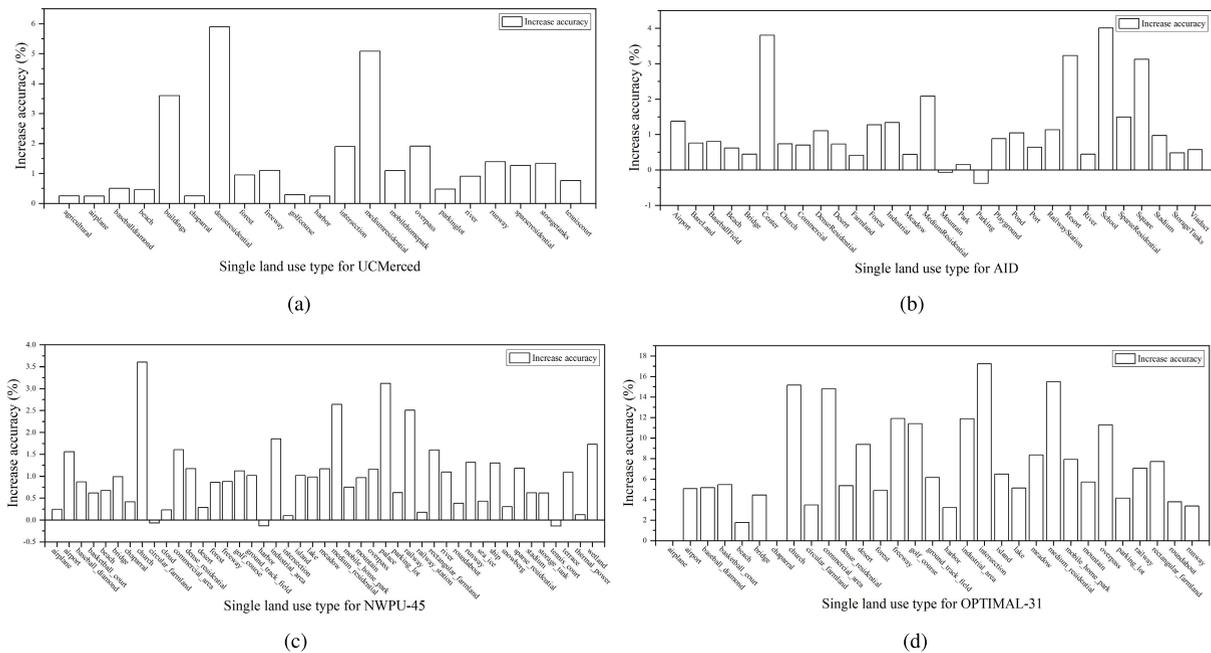


FIGURE 9. Increase classification accuracy of single land use type under MJDCNN for UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets compared with the best classification accuracy of the single land use type among T-AlexNet, T-Inception-v3, and T-ResNet18 (the number of iterations is 20).

maximum and minimum accuracy is 25.05%, and that of T-ResNet18 is the smallest at 20.71%. The maximum accuracy of MJDCNN is 100%, the minimum accuracy is 84.26%, and the difference is 15.74%. The difference is 9.31%, 7.47%, and 4.97% lower than that in T-AlexNet, T-Inception-v3, and T-ResNet18, respectively. In Figure 8(a), land use types with low classification accuracy are represented by buildings, dense residential, and medium density residential. Classification accuracy of all land use types is improved through MJDCNN. Specifically in buildings, dense residential, and medium density residential, these accuracies respectively increase by 3.60%, 5.90%, and 5.08% compared with the best performance among the three single-structured CNNs Figure 9(a).

b: AID DATASET

In Table 3, the maximum and minimum accuracy are similar to those of the three single-structured CNNs. The difference between the T-AlexNet’s maximum accuracy and minimum accuracy is 29.79%, and that of T-ResNet18 is the smallest at 23.26%.The maximum accuracy of MJDCNN is 99.51%, whereas the minimum accuracy is 80.02%. The difference is 19.49%, which is 10.30%, 6.21%, and 3.77% lower than the differences in T-AlexNet, T-Inception-v3, and T-ResNet18, respectively. In Figure 8(b), parking lot and school have low classification accuracies. Except for parking and mountain, the classification accuracy of land use types is improved by MJDCNN. Specifically, in center, medium residential, resort, school and square, the accuracies respectively increase by 3.80%, 2.09%, 3.23%, 4.01% and 3.12% compared with the

TABLE 3. Classification accuracy of single land use type of CNNs for the UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets. Max_acc and Min_acc respectively denote the maximum and minimum classification accuracy of single land use type of CNNs. Max-Min denotes the difference between the maximum and the minimum. T-A, T-I, T-R and MJD respectively denote T-AlexNet, T-Inception-v3, T-ResNet18 and MJDCNN.

Datasets	Values	T-CNNs			
		T-A	T-I	T-R	MJD
UC Merced	Max_acc	99.28%	99.76%	99.76%	100%
	Min_acc	74.23%	76.54%	79.04%	84.26%
	Max-Min	25.05%	23.21%	20.71%	15.74%
AID	Max_acc	98.60%	99.42%	99.27%	99.51%
	Min_acc	68.81%	73.73%	76.01%	80.02%
	Max-Min	29.79%	25.70%	23.26%	19.49%
NWPU-45	Max_acc	95.15%	99.14%	98.85%	99.35%
	Min_acc	66.08%	75.28%	73.49%	78.40%
	Max-Min	29.07%	23.86%	25.35%	20.95%
OPTIMAL-31	Max_acc	94.92%	99.20%	100%	100%
	Min_acc	63.84%	48.20%	64.00%	78.79%
	Max-Min	31.08%	51.00%	36.00%	21.21%

best performance among the three single-structured CNNs Figure 9(b).

c: NWPU-45 DATASET

According to Table 3, T-AlexNet has the worst maximum and minimum accuracy. MJDCNN has the best maximum accuracy and the smallest difference of maximum and minimum accuracy. The difference between the maximum and minimum accuracy of MJDCNN decreases to 20.95%, which is 8.12%, 2.91%, and 4.40% lower than the difference in T-AlexNet, T-Inception-v3, and T-ResNet18, respectively. In Figure 8(c), land use types with low classification accuracy are represented by church, commercial area,

dense residential, industrial area, medium residential, palace, railway, and railway station. The accuracies of circular farmland, harbor and tennis court exhibit a slight decrease through MJDCNN, wherein the range of decline remains within 0.14%. The classification accuracy of other land use types improves through MJDCNN. Specifically, in church, medium residential, palace, and railway, these accuracies increase by 3.61%, 2.64%, 3.12%, and 2.50% compared with the best performance among the three single-structured CNNs Figure 9(c).

d: OPTIMAL-31 DATASET

For this dataset, the strategy of MJDCNN for classification has a remarkable effect. As shown in Table 3, T-Inception-v3 has the worst minimum accuracy (48.20%) and the difference between maximum accuracy and minimum accuracy (51.00%). For MJDCNN, the maximum accuracy is 100%, the minimum accuracy is 78.79%, and the difference is 21.21%. Compared with T-AlexNet, T-Inception-v3, and T-ResNet18, the value of the difference is reduced by 9.87%, 29.79%, and 14.79%. In Figure 8(d), land use types with low classification accuracy are represented by church, commercial area, industrial area and medium residential. All land use types improve their classification accuracies by MJDCNN. Specifically in church, commercial area, intersection and medium residential, these accuracies respectively increase by 15.16%, 14.79%, 17.23%, and 15.49% compared with the best performance among the three single-structure CNNs Figure 9(d).

V. DISCUSSION

A. COMPARISON WITH OTHER CLASSIFICATION METHODS

To evaluate the classification performance of MJDCNN further, we select commonly state-of-the-art methods for comparison, shown in Table 4, Table 5, Table 6, and Table 7. We apply the MJDCNN under 20 iterations as the final models.

1) UC MERCED LAND USE DATASET

This dataset is the first public dataset for remote sensing classification, which is widely used. We select some representative methods to compare with our proposed MJDCNN. Given that we train MJDCNN under the 40% training ratio, we compare with the results of state-of-the-art methods under the 50% training ratio or the similar training ratio. As shown in Table 4, MJDCNN achieves a $95.79\% \pm 0.91\%$ accuracy under the 40% training ratio. The performance of MJDCNN performs better than the listed methods except for GBNet+global feature. Our MJDCNN makes an increase of 1.58% over SalM3LBPLM [71]. Multiscale ADPM, Multi-scale CNN, and Multiview deep learning have similar network structure to our proposed MJDCNN [55], [67], [72]. Compared with their OAs, the OA of our proposed MJDCNN increases by 0.93%, 6.09%, and 2.31%, respectively. GBNet

TABLE 4. OAs on the UC Merced land use dataset compared with other classification methods. MJDCNN represents the OAs of MJDCNN when the number of iterations is 20.

Methods	OAs(%)
CaffeNet [14]	93.98 \pm 0.67
VGG-VD-16 [14]	94.14 \pm 0.69
GoogLeNet [14]	92.70 \pm 0.60
ConvNet [65]	89.39 \pm 1.10
SPM+deconvolution networks [66]	95.71 \pm 1.00
Plain model without weight [66]	91.43 \pm 0.67
Multiscale ADPM [67]	94.86
LPCNN [68]	89.9
GBRCN [52]	94.53
CNN-S [57]	92.84
CNN-M [57]	93.51 \pm 1.00
MCNN-NS [57]	94.12 \pm 1.00
Multi-scale CNN [55]	89.70
Res-CapsNet [69]	92.26
MS-CLBP+FV [70]	88.76 \pm 0.76
SalM3LBPLM [71]	94.21 \pm 0.75
Multiview deep learning [72]	93.48 \pm 0.82
OverFeat+Trainable CNN [73]	92.4
Random-scale stretched CNN [56]	95.1
FSSTM [74]	95.71 \pm 1.01
GBNet [58]	95.71 \pm 0.19
GBNet+global feature [58]	97.05 \pm 0.19
MJDCNN	95.79 \pm 0.91

TABLE 5. OAs on the AID dataset compared with other classification methods. MJDCNN represents the OAs of MJDCNN when the number of iterations is 20.

Methods	OAs(%)
CaffeNet [14]	89.53 \pm 0.31
VGG-VD-16 [14]	89.64 \pm 0.36
GoogLeNet [14]	86.39 \pm 0.55
Fusion by addition [75]	91.87 \pm 0.36
Fusion by concatenation [75]	91.86 \pm 0.28
DCA with concatenation [75]	89.71 \pm 0.33
CNN-S [57]	89.70 \pm 0.53
CNN-M [57]	90.10 \pm 0.34
MCNN-NS [57]	90.40 \pm 0.41
MCNN [57]	91.80 \pm 0.22
Multi-scale CNN [55]	88.26
ARCNet-VGG16 [36]	93.10 \pm 0.55
VGG-16(fine-tuning) [58]	93.60 \pm 0.64
GBNet [58]	93.72 \pm 0.34
MKL-avg [76]	90.76 \pm 0.32
GBNet+global feature [58]	95.48 \pm 0.12
MJDCNN	94.18 \pm 0.31

is one of the latest proposed method for remote sensing classification [58]. Our proposed MJDCNN achieves 0.09% accuracy higher than GBNet. Besides, our proposed MJDCNN has a slightly lower accuracy than GBNet+global feature [58]. This is partly due to the different training ratio: MJDCNN has a 40% training ratio, a 20% validation ratio, and a 40% test ratio, while GBNet+global feature has a 50% training ratio, and a 50% test ratio. As the increase of training ratio, the accuracy of classification networks increases. In fact, our proposed MJDCNN misclassifies a total of 36 images, and GBNet+global feature misclassifies a total of 31 images. Such facts indicate that the proposed MJDCNN is competent for UC Merced dataset classification with less number of training remote sensing images.

TABLE 6. OAs on the NWPU-45 dataset compared with other classification methods. MJDCNN represents the OAs of MJDCNN when the number of iterations is 20.

Methods	OAs(%)
AlexNet [25]	79.85 ± 0.13
VGG-VD-16 [25]	79.79 ± 0.15
Googlenet [25]	78.48 ± 0.26
Fine-tuning AlexNet [25]	85.16 ± 0.18
Fine-tuning VGGNet16 [25]	90.36 ± 0.18
Fine-tuning Googlenet [25]	86.02 ± 0.18
VGG16 on CAE [77]	0.90
BoCF+VGGNet-16 [25]	84.32 ± 0.17
D-CNN with VGGNet-16 [78]	91.89 ± 0.22
DCNN [78]	87.06 ± 0.47
TrAdaBoost based on modified PSO [79]	87.13 ± 0.58
MKL-avg [76]	89.39 ± 0.35
IOR4-VGG16 [80]	91.30 ± 0.17
SE-ResNeXt [81]	92.18
VGG-16+interlayer fusion [82]	92.73
MJDCNN	93.18 ± 0.24

2) AID DATASET

This dataset has more complex images with more land use types than UC Merced dataset. As shown in Table 5, MJDCNN achieves a 94.18% ± 0.31% accuracy under the 40% training ratio. The performance of MJDCNN performs better than the listed methods except for GBNet+global feature. Our MJDCNN makes an increase of 5.92% over Multi-scale CNN that has similar network structure [55]. Fusion by concatenation is one of the most effective method for AID dataset classification, and our MJDCNN makes an increase of 2.32% than this method [75]. MCNN, GBNet, and ARCNet-VGG16 are the latest methods for AID dataset classification [36], [57], [58]. In comparison of these latest methods, our proposed MJDCNN increases by 2.38%, 0.42%, and 1.08%. The performance of GBNet+global feature outperforms the proposed MJDCNN with the cost of more number of training remote sensing images [58].

3) NWPU-45 DATASET

This dataset has the largest number of land use types and HRRS images among public HRRS image datasets. As shown in Table 6, MJDCNN achieves 93.18% ± 0.24% accuracy. Our proposed MJDCNN has the best performance compared with the other state-of-the-art methods. Both IOR4-VGG16 and VGG-16+interlayer fusion are the latest proposed method for NWPU-45 dataset classification [80], [82]. Our proposed MJDCNN achieves 1.88 percentage and 0.45 percentage points higher than these two methods. Besides, AlexNet is one of the based structures for our proposed MJDCNN. Therefore, we compare the performance of the proposed MJDCNN with that of AlexNet. AlexNet obtains 79.85% ± 0.13% accuracy, and fine-tuning AlexNet gets 85.16% ± 0.18% accuracy, wherein our proposed MJDCNN achieves 13.33 percentage and 8.02 percentage points higher than these two methods [25]. SE-ResNeXt based on residual block obtains 92.18% accuracy [81]. Our proposed MJDCNN also has residual structure and achieves

TABLE 7. OAs on the OPTIMAL-31 dataset compared with other classification methods. MJDCNN represents the OAs of MJDCNN when the number of iterations is 20.

Methods	OAs(%)
Fine-tuning AlexNet [36]	81.22 ± 0.19
Fine-tuning VGGNet16 [36]	87.45 ± 0.45
Fine-tuning Googlenet [36]	82.57 ± 0.12
VGG-VD-16 [36]	89.12 ± 0.35
ARCNet-ResNet34 [36]	91.28 ± 0.45
ARCNet-AlexNet [36]	85.75 ± 0.35
ARCNet-VGG16 [36]	92.70 ± 0.35
VGG-16(fine-tuning) [58]	89.52 ± 0.26
GBNet [58]	91.40 ± 0.27
GBNet+global feature [58]	93.28 ± 0.27
MJDCNN	94.57 ± 1.25

1.00% improvement compared with SE-ResNeXt. It confirms the efficiency of the proposed MJDCNN.

4) OPTIMAL-31 DATASET

This dataset is a newly challenging dataset and contains more complicated remote sensing images. As shown in Table 7, our proposed MJDCNN gets 94.75% ± 1.25% accuracy, which performs the obvious superiority. ARCNet-VGG16 and GBNet+global feature are the latest methods for this dataset classification [36], [58]. Our proposed MJDCNN makes the increase of 1.87% and 1.29% than these two methods. AlexNet and ResNet18 are two basic network structures of MJDCNN. We could find that our proposed MJDCNN increases by 13.35%, 8.82%, and 3.29% compared with fine-tuning AlexNet, ARCNet-AlexNet, and ARCNet-ResNet34 [36]. Once again, it demonstrates the superiority of our proposed MJDCNN.

B. ADVANTAGE OF THE PROPOSED MJDCNN

The proposed MJDCNN is utilized to perform classification tasks, wherein it exhibits excellent performance. MJDCNN can obtain optimal classification results of three single-structured CNNs to improve the overall classification accuracy and ability. Convolutional layers extract features of HRRS images of land use via convolution operation. As the convolutional layer deepens and widens, the extracted features become increasingly complex, enabling easy image classification. T-AlexNet, T-Inceptionv3, and T-ResNet18 have different convolutional layers in terms of depth and width. Thus, they can extract different features to describe these images and accomplish classification. Joint decision-making approach can select the best results. Regardless of the number of iterations, the proposed MJDCNN is superior to the three single-structured CNNs in terms of OAs, F1-score, and classification accuracy of single land use type.

In addition, the performance of the proposed MJDCNN is enhanced as the number of iterations increases due to nonconvergent single-structured CNNs. The three single-structured CNNs decide the classification accuracies of the proposed MJDCNN. To illustrate the effect of the non-convergent single-structured CNNs, we draw their accuracy and loss

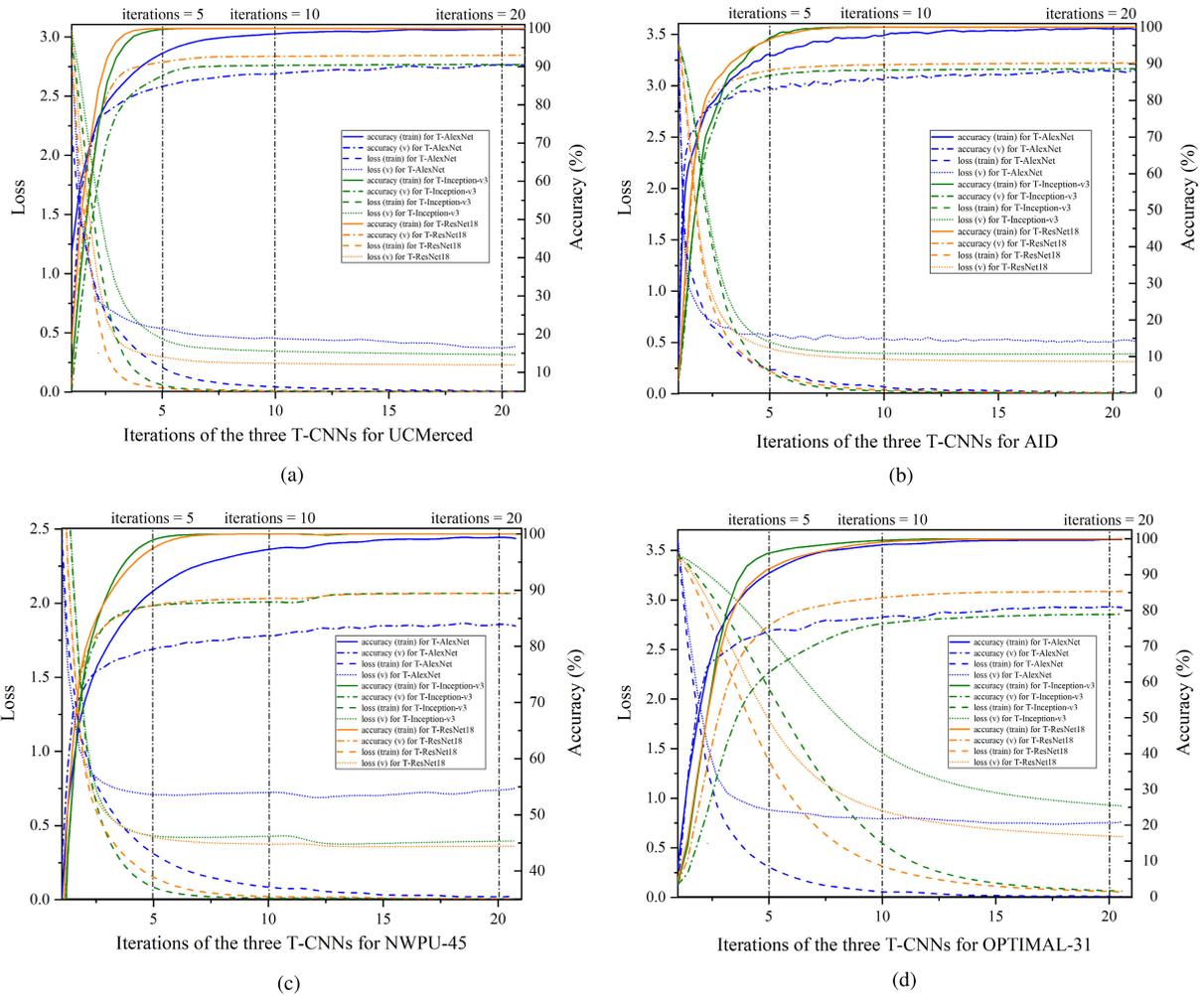


FIGURE 10. Trend curve of accuracy and loss values of T-AlexNet (blue), T-Inception-v3 (green), and T-ResNet18 (orange) for UC Merced land use, AID, NWPU-45 and OPTIMAL-31 datasets. The polyline and dashed line respectively represent the accuracy and loss of the T-CNNs for train dataset. The dash-dotted line and dotted line respectively represent the accuracy and loss of the T-CNNs for validation dataset. Three vertical dash-dotted lines respectively represent different iterations.

during training under four datasets (Figure 10). For the UC Merced land use, AID, and NWPU-45 datasets, when the number of iterations is 5, T-AlexNet, T-Inceptionv3, and T-ResNet18 remain in the training state. T-AlexNet reaches the convergence when the number of iterations is 20, and T-Inceptionv3 and T-ResNet18 reach convergence when the number of iterations is 10. For the OPTIMAL-31 dataset, T-AlexNet, T-Inceptionv3 and T-ResNet18 reach convergence when the number of iterations is 20. If these single-structured CNNs converge, then the classification accuracies of the proposed MJDCNN are stable. By contrast, if these single-structured CNNs do not converge, then the classification accuracies of the proposed MJDCNN change. Hence, the classification accuracies of the proposed MJDCNN continuously increase from 5 to 20 iterations.

Moreover, we also confirm that the size of datasets has some effect on the performance of MJDCNN. More than 100 images in each land use type are used for UC Merced land

use, AID and NWPU-45 datasets. The difference between the train and validation datasets of accuracy and loss has an approximate value when the number of iterations is 20 (Figure 10 (a), (b), and (c)). In the OPTIMAL-31 dataset, only 60 images are used in each land use type. A large difference between the train and validation datasets of accuracy and loss is observed (Figure 10 (d)). This finding may be caused overfitting for T-AlexNet, T-Inception-v3, and T-ResNet18 given the small dataset. However, MJDCNN can alleviate the overfitting effect and greatly increase the accuracies. This finding shows a clear benefit of utilizing the approach based on joint decision-making strategy and three single-structured CNNs for small HRRS images.

C. IMPROVEMENT OF UNBALANCED CLASSIFICATION ACCURACY

The proposed MJDCNN improves the unbalanced classification accuracy of single land use type. Although the

single-structured CNNs can achieve high-precision overall classification for HRRS images of land use, the classification accuracy of single land use type varies greatly. For four HRRS image datasets, the classification accuracies of buildings, dense residential, medium residential, commercial area, industrial area, and church have the worst performance under the classification of T-Alexnet, T-Inception-v3, and T-ResNet18. This result is due to the similarities of the HRRS images of these land use types in terms of texture and color (Figure 2(e), (g), (m), Figure 3(g), (h), (i), (m), (o), Figure 4(h), (k), (l), (s), (x), and Figure 5(h), (j), (k), (w)). During CNN classification, the convolutional layer extracts similar features. Thus, these land use types cannot be accurately distinguished. Fortunately, the proposed MJDCNN can make joint decision based on the classification results of three single-structure CNNs. It increases the classification accuracies of these land use types to a certain extent, thereby ameliorating the overall imbalance of classification accuracy.

VI. CONCLUSION

This study proposes MJDCNN to classify land use types using HRRS images. The proposed approach applies three pretrained CNNs to achieve land use classification and obtain the final classification results by joint decision-making strategy. The MJDCNN classification effect is crucial, especially for OPTIMAL-31 dataset. As the number of iterations increases, the overall classification accuracy and the F1-score of the proposed MJDCNN improve as well. The classification performance of the proposed MJDCNN is better than that of the single-structured pretrained CNNs as the benchmark. Compared with different classification methods for HRRS images of land use, the proposed MJDCNN is proven to be a promising approach. The highlights of this study are listed as follows:

(1) On the structure of the proposed MJDCNN. This multi-structure network is designed using different single-structured networks. We focus on the influence of different CNN structures. The proposed MJDCNN does not integrate features that are extracted from differently structured CNNs. It relies on the different structures of CNNs to classify and then joint decision-making based on the classification results. This strategy directly uses the original feature vectors obtained by CNNs to classify, without fusing feature vectors. It can prevent the feature from being deformed or losing information during the fusion process.

(2) On the selection of the single pretrained network. AlexNet, Inception-v3, and ResNet18 are chose to build the MJDCNN. AlexNet belongs to traditionally structured CNNs, Inception-v3 belongs to the CNNs with inception module, and ResNet18 belongs to residual CNNs. These three CNNs have different network depths and widths, size of input images, and convolutional kernel structures. Thus, they can ensure the diversity of input data and the richness of features extracted from HRRS images. As such, the proposed MJDCNN exhibits an excellent performance.

(3) On the size of the dataset. The proposed MJDCNN has great potential on small dataset. For the above four datasets, the proposed MJDCNN has the most significant improvement effect on the OPTIMAL-31 dataset. Compared with other datasets, the OPTIMAL-31 dataset has more land use types and fewer HRRS images in each land use type. CNNs consumes data and requires numerous labeled HRRS images for classification. This characteristic may be caused by bad performance of CNNs. The proposed MJDCNN can mitigate this knotty problem.

It is noticeable that the proposed MJDCNN obtains excellent results under the opening HRRS image datasets. In future research, we will focus on how to use the proposed MJDCNN to cope with the diversity of a real-world image, to make MJDCNN more useful in practical applications.

REFERENCES

- [1] L. H. Nguyen, D. R. Joshi, D. E. Clay, and G. M. Henebry, "Characterizing land cover/land use from multiple years of landsat and MODIS time series: A novel approach using land surface phenology modeling and random forest classifier," *Remote Sens. Environ.*, vol. 238, pp. 589–599, Mar. 2020.
- [2] J. Wang, T. He, Q. Zhou, and X. Guo, "Developing land use/cover classification system based on remote sensing data in China," in *Proc. Remote Sens. Environ. Monitor., GIS Appl., Geol. IV*, Oct. 2004, pp. 52–61.
- [3] Q. Liao, M. Su, G. Luo, and X. Wei, "Research on land use classification of land change survey for the integration of urban planning and land resource administration," in *Proc. 17th Int. Symp. Adv. Construct. Manage. Real Estate*. Berlin, Germany: Springer, 2014, pp. 197–205.
- [4] H. Thunig, N. Wolf, S. Naumann, A. Siegmund, C. Jurgens, C. Uysal, and D. Maktav, "Land use/land cover classification for applied urban planning—the challenge of automation," in *Proc. Joint Urban Remote Sens. Event*, Apr. 2011, pp. 229–232.
- [5] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 113–123, Mar. 2013.
- [6] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 9, pp. 1988–2007, May 2014.
- [7] D. Punter, T. W. Kellenberger, Y. Buehler, and B. Lennartz, "Mapping diversified peri-urban agriculture—Potential of object-based versus per-field land cover/land use classification," *Geocarto Int.*, vol. 25, no. 3, pp. 171–186, Jun. 2010.
- [8] M. Pun, D. Mutiibwa, and R. Li, "Land use classification: A surface energy balance and vegetation index application to map and monitor irrigated lands," *Remote Sens.*, vol. 9, no. 12, p. 1256, Dec. 2017.
- [9] S. Avelar and P. Tokarczyk, "Analysis of land use and land cover change in a coastal area of rio de janeiro using high-resolution remotely sensed data," *J. Appl. Remote Sens.*, vol. 8, no. 1, May 2014, Art. no. 083631.
- [10] A. Upadhyay, A. Shetty, S. K. Singh, and Z. Siddiqui, "Land use and land cover classification of LISS-III satellite image using KNN and decision tree," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIA-Com)*, 2016, pp. 1277–1280.
- [11] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.
- [12] C. Zhang, I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, and P. M. Atkinson, "An object-based convolutional neural network (OCNN) for urban land use classification," *Remote Sens. Environ.*, vol. 216, pp. 57–70, Oct. 2018.
- [13] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.
- [14] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

- [15] M. Carranza-García, J. García-Gutiérrez, and J. Riquelme, "A framework for evaluating land use and land cover classification using convolutional neural networks," *Remote Sens.*, vol. 11, no. 3, p. 274, Jan. 2019.
- [16] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
- [17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
- [18] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.
- [21] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop Multimedia Inf. Retr. (MIR)*, 2007, pp. 197–206.
- [22] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.
- [23] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.
- [24] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
- [25] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
- [26] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, vol. 5, pp. 9021–9031, 2017.
- [27] Q. Zhang, R. Qin, X. Huang, Y. Fang, and L. Liu, "Classification of ultra-high resolution orthophotos combined with DSM using a dual morphological top hat profile," *Remote Sens.*, vol. 7, no. 12, pp. 16422–16440, Dec. 2015.
- [28] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, no. 7, p. 666, Jun. 2017.
- [29] R. Cao, J. Zhu, W. Tu, Q. Li, J. Cao, B. Liu, Q. Zhang, and G. Qiu, "Integrating aerial and street view images for urban land use classification," *Remote Sens.*, vol. 10, no. 10, p. 1553, Sep. 2018.
- [30] J. Chen, C. Wang, Z. Ma, J. Chen, D. He, and S. Ackland, "Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters," *Remote Sens.*, vol. 10, no. 2, p. 290, Feb. 2018.
- [31] X. Lv, D. Ming, T. Lu, K. Zhou, M. Wang, and H. Bao, "A new method for region-based majority voting CNNs for very high resolution image classification," *Remote Sens.*, vol. 10, no. 12, p. 1946, Dec. 2018.
- [32] Z. Gong, P. Zhong, W. Hu, and Y. Hua, "Joint learning of the center points and deep metrics for land-use classification in remote sensing," *Remote Sens.*, vol. 11, no. 1, p. 76, Jan. 2019.
- [33] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, p. 494, Feb. 2019.
- [34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
- [35] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. TC VII Symp.-100 Years (ISPRS)*, vol. 38, 2010, pp. 298–303.
- [36] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
- [37] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
- [38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [42] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.
- [43] Y. Boualleg, M. Farah, and I. R. Farah, "Remote sensing scene classification using convolutional features and deep forest classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1944–1948, Dec. 2019.
- [44] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 704–708, May 2017.
- [45] W. Zhou, Z. Shao, and Q. Cheng, "Deep feature representations for high-resolution remote sensing scene classification," in *Proc. 4th Int. Workshop Earth Observ. Remote Sens. Appl. (EORS)*, Jul. 2016, pp. 338–342.
- [46] S. Chaib, Y. Gu, H. Yao, and K. Belkadi, "Very high resolution image scene classification with semantic Fisher vectors," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6844–6847.
- [47] Y. Liang, S. T. Monteiro, and E. S. Saber, "Transfer learning for high resolution aerial image classification," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2016, pp. 1–8.
- [48] X. Liu, M. Chi, Y. Zhang, and Y. Qin, "Classifying high resolution remote sensing images by fine-tuned VGG deep networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7137–7140.
- [49] B. Alias, R. Karthika, and L. Parameswaran, "Classification of high resolution remote sensing images using deep learning techniques," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 1196–1202.
- [50] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land-cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.
- [51] S. Liu, Q. Wang, and X. Li, "Attention based network for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4740–4743.
- [52] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.
- [53] N. Yang, H. Tang, H. Sun, and X. Yang, "DropBand: A simple and effective method for promoting the scene classification accuracy of convolutional neural networks for VHR remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 257–261, Feb. 2018.
- [54] K. P. Korshunova, "A convolutional fuzzy neural network for image classification," in *Proc. 3rd Russian-Pacific Conf. Comput. Technol. Appl. (RPC)*, Aug. 2018, pp. 1–4.
- [55] H. Alhichri, N. Alajlan, Y. Bazi, and T. Rabczuk, "Multi-scale convolutional neural network for remote sensing scene classification," in *Proc. IEEE Int. Conf. ElectroInf. Technol. (EIT)*, May 2018, pp. 1–5.
- [56] Y. Liu, Y. Zhong, F. Fei, and L. Zhang, "Scene semantic classification based on random-scale stretched convolutional neural network for high-spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 763–766.
- [57] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.
- [58] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
- [59] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.
- [60] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.
- [61] Z. Yang, X.-D. Mu, and F.-A. Zhao, "Scene classification of remote sensing image based on deep network and multi-scale features fusion," *Optik*, vol. 171, pp. 287–293, Oct. 2018.

- [62] P. Du, E. Li, J. Xia, A. Samat, and X. Bai, "Feature and model level fusion of pretrained CNN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2600–2611, Aug. 2019.
- [63] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.
- [64] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*. [Online]. Available: <http://arxiv.org/abs/1508.00092>
- [65] K. Nogueira, W. O. Miranda, and J. A. D. Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in *Proc. 28th SIBGRAPI Conf. Graph., Patterns Images*, Aug. 2015, pp. 289–296.
- [66] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.
- [67] Q. Liu, R. Hang, H. Song, F. Zhu, J. Plaza, and A. Plaza, "Adaptive deep pyramid matching for remote sensing scene classification," 2016, *arXiv:1611.03589*. [Online]. Available: <http://arxiv.org/abs/1611.03589>
- [68] Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, Apr. 2016, Art. no. 025006.
- [69] T. Tian, X. Liu, and L. Wang, "Remote sensing scene classification based on res-capsnet," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 525–528.
- [70] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and Fisher vectors," *Remote Sens.*, vol. 8, no. 6, p. 483, Jun. 2016.
- [71] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.
- [72] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.
- [73] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
- [74] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous-heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.
- [75] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [76] X. Bian, Y. Sheng, Y. Xu, and Q. Du, "A novel multiple kernel learning framework for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4717–4720.
- [77] H. Wang, Y. Miao, H. Wang, and B. Zhang, "Convolutional attention in ensemble with knowledge transferred for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 643–647, Apr. 2019.
- [78] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [79] L. Yan, R. Zhu, Y. Liu, and N. Mo, "TrAdaBoost based on improved particle swarm optimization for cross-domain scene classification with limited samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3235–3251, Sep. 2018.
- [80] J. Wang, W. Liu, L. Ma, H. Chen, and L. Chen, "IORN: An effective remote sensing image scene classification framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1695–1699, Nov. 2018.
- [81] W. Han, R. Feng, L. Wang, and L. Gao, "Adaptive spatial-scale-aware deep convolutional neural network for high-resolution remote sensing imagery scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4736–4739.
- [82] C. Ma, X. Mu, and D. Sha, "Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing," *IEEE Access*, vol. 7, pp. 121685–121694, 2019.



LU XU received the B.E. degree from the China University of Geosciences, China, in 2018. She is currently pursuing the M.S. degree with the School of Resource and Environmental Sciences, Wuhan University. Her research interests include deep learning, remote sensing, and land use classification.



YIYUN CHEN received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2006, 2008, and 2011, respectively, and the M.Sc. degree from ITC, The Netherlands, in 2008. He was an Academic Visitor with the University of Cambridge, from 2014 to 2015. Since 2019, he has been an Academic Visitor with the University of Cambridge. He serves as an Associate Professor with Wuhan University. He has more than 60 research articles, and he serves as a Reviewer of over 20 science citation index magazines. His research interest includes remote sensing of resource and environment.



JIAWEI PAN received the B.S. degree from Hubei Normal University, China, in 2016. She is currently pursuing the M.S. degree with the School of Resource and Environmental Sciences, Wuhan University. Her research interest includes urban remote sensing.



AJI GAO received the B.E. degree from the China University of Geosciences, China, in 2018. He is currently pursuing the M.S. degree with the School of Resource and Environmental Sciences, Wuhan University. His research interests include machine learning and remote sensing.