IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 September 2023, accepted 21 September 2023, date of publication 29 September 2023,
date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320792

RESEARCH ARTICLE

High Resolution Remote Sensing Image Classification Based on Deep Transfer Learning and Multi Feature Network

XINYAN HUANG
School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China
e-mail: 15953130256@163.com

This work was supported by the Research and Analysis on the Big Data Laboratory Platform Construction System of the Industry-University Cooperative Education Project of Ministry of Education, in 2022, under Grant 220506090201041.

ABSTRACT To improve the automatic classification accuracy of remote sensing images, this study raises a high-resolution remote sensing image classification model that combines deep transfer learning and multi-feature network. In this paper, deep transfer learning is the core technology of remote sensing image classification model, and VGG16, Inception V3, ResNet50 and MobileNet are used to build a fusion classification model through serial fusion. By testing the fusion model, the Transfer Learning ResNet50-MobileNet (TL-RM) model with the best performance was obtained. Finally, experimental analysis verified its significant stability: the average accuracy of TL-RM on a small sample high-resolution remote sensing image dataset was 96.8%, and the Kappa coefficient was 0.964, both of which were the highest values among all models. The accuracy of this model shows a slight upward trend and then stabilizes as the iterations increases. The training and testing sets accuracy ultimately stabilizes at around 100% and 98%, and the loss value ultimately stabilizes at around 1%. Moreover, TL-RM only has a low classification accuracy for residential areas in remote sensing images, with a classification accuracy of over 97% for other categories. The experiment shows that the TL-RM model has significant accuracy and stability, providing a reliable theoretical and experimental basis for remote sensing image classification research.

INDEX TERMS Transfer learning, multiple features, CNN, remote sensing images, classification.

I. INTRODUCTION

As a result of the unique nature of the image acquisition method for HiR-RSI, numerous works required during the image acquisition. Therefore, the speed of information acquisition is often slow, and the resulting high-resolution remote sensing image samples are also limited. This poses certain difficulties for automated high-resolution remote sensing image classification (RSIC) [1], [2], [3]. The insufficient number of samples makes it impossible for machine learning methods to have sufficient training set data for model learning, which also makes automated classification difficult and severely inaccurate. Although there are currently some small sample machine learning algorithms suitable for HiR-RSI, these all require manual extraction of information features from HiR-RSI. Artificial information feature extraction is time and labor consuming, as well as often has significant limitations. The extracted information features cannot be used for high-precision computer classification [4], [5], [6]. Simultaneously, a small sample set can easily cause over-fitting problems in the model during training, leading to insufficient performance of the classification model. Transfer learning can deal with the over fitting caused by insufficient sample size. Deep transfer learning is widely used in the application of small training samples, and has practicability [7], [8], [9]. To address the issue of insufficient feature extraction, a multi feature network can be used to solve it. The multi feature network itself can perform varying degrees of deep learning on the same image and extract image information features with different focuses during the learning process. These image information features can become information features that carry the model's judgment

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

basis after information fusion, providing a foundation for high-precision classification of the model. The study employs techniques that rely on deep convolutional neural networks (CNN) and transfer learning. This network not only mitigates the over-fitting problem associated with model training due to small sample sizes but also solves the issue of limited feature extraction inherent in a single deep CNN. The study will utilize four deep CNN: VGG16, InceptionV3, ResNet50, and MobileNet. They will be combined with deep transfer learning to produce a classification model perfect for high-resolution remote sensing images. This will help mitigate model over-fitting during training in small sample situations. After the selection process, three high-performing models will serve as the foundation for multi-feature network fusion. Based on this research, a deep multi-feature network fusion model is proposed with the aim of achieving improved classification performance. The model employs a fusion technique that combines characteristics extracted from two deep transfer CNN to address the issue of classifying high-resolution remote sensing images when only a small sample is available. This work is essential to advance research and application in the field of remote sensing images.

II. RELATED WORKS

For the past few years, applying deep transfer learning (DTL) in various fields has gradually deepened. Jaiswal A's team applied DTL to the automatic analysis and diagnosis of radioactive images of COVID-19 and developed an optimized DTL model based on DensdNetde. This model can divide COVID-19 patients into different symptom categories through automated data analysis based on chest CT radioactive images, so as to achieve more efficient automatic diagnosis [10]. Ahuja S et al. combined the DTL model with CT image scanning technology to construct a three-phase detection mode on the ground of transfer learning method. It was separated into three main operating intervals: in the first interval, the model utilizes stationary wavelet technology to perform data augmentation operations; The second interval requires the use of CNN pre training to achieve clinical symptom detection and classification; In the third interval, the model performs anomaly localization on CT images. The testing accuracy of this model reached 99.45, and it had good application results [11]. Liu C et al. developed a monitoring framework that combines DTL technology, which can extract information from implicit channel features while detecting and replying to label signals. The framework is mainly composed of 3 parts: the 1st part is offline learning, the 2nd is transfer learning, and the 3rd is online detection. At the same time, this study used CNN to analyze the covariance matrix features of data samples. The results showed that the error rate indicator performance of this method currently belongs to the best among the same type of models [12]. Phan H's team had developed a deep migration small queue sleep staging model to address data variability issues. The model took end-to-end DTL as the main architecture, and on this basis, two parallel networks were used as the main means of transfer learning. In the SD of the model, pre training of large databases is carried out, while in the target domain, the network of small queues is slightly adjusted to achieve knowledge transfer. This model is effective [13]. Naseer et al. proposed a neural network model for deep apology learning and data augmentation techniques. Its needle can provide automated early diagnosis of early clinical symptoms of Parkinson's disease. The conclusion of this study is that the accuracy of this model has advantages over similar models [14].

The research of HiR-RSI is gradually deepening. Peng D et al. designed a HiR-RSI change detection method that introduces semi supervised convolutional networks, which has reliability and superiority [15]. This method uses two different types of discrimination to network segment labeled and unlabeled data, and achieves feature distribution consistency through data augmentation. Li H's research team designed a semantic segmentation technique for HiR-RSI, and experiments have shown that the model is effective. This technology uses end-to-end segmented networks as the infrastructure and combines them with spatial attention models to achieve automated adaptation [16]. Guo H and other scholars have designed a building classification and extraction model for HiR-RSI, and also adopted a parallel training method to achieve classification results through end-to-end transfer learning. This model is based on the semantic differences between different types of buildings and performs attention based scenario based parallel judgment [17]. In addition, the application of multi feature networks in image analysis is gradually deepening. Qiu S's team has proposed a multi feature network model for remote sensing images. This model mainly addresses the issues of poor image quality and insufficient feature utilization in the transmission process of remote sensing images. It combines image enhancement algorithms with dual attention networks for decision fusion analysis. Research data shows that this model can improve the accuracy of visual analysis [18]. Wang Y et al. designed a multi-feature and backpropagation network for the problem of low recognition rate of remote sensing images under single feature recognition. It obtains binary images by threshold segmentation of the image, and on this basis, uses grayscale co-occurrence matrix and binary pattern method to obtain texture features of weeds, thereby detecting and classifying weed images [19]. Sheykhmousa M et al. conducted a comparative analysis on centralized application algorithms in RSIC. Their research focused on the comprehensive utilization of random forest algorithm (RF) and support vector machine (SVM) in RSIC. The research analyzed the parameter performance, data application types and spatial resolution of RF and SVM respectively, and explored the ability of feature engineering to extract image features on this basis. The study also constructed a research database containing 40 quantitative and qualitative fields, and conducted a summary analysis of the characteristics of literature related to RSIC, including the time, frequency, and geographical distribution of the research [20]. To obtain the basic data information of narrow spectral bands in hyper-spectral remote sensing images, Uddin MP

applied a frequency band reduction feature extraction and data analysis strategy. The study used principal component analysis to extract feature information and utilizes variance accumulation to classify the top feature data. On this basis, he used SVMs to classify the features of remote sensing image data in specific regions. The research concluded that this method has more efficient classification performance and can effectively cope with the spatial and temporal complexity in feature extraction [8]. Hong D's research team has developed a multimodal deep learning framework that can effectively classify ground materials in remote sensing images. Meanwhile, they also designed a cross modal model with multimodal learning capabilities. In the design, five fusion frameworks were introduced and effective functional integration was achieved. To expand the functionality of the model on the basis of pixel classification tasks, the study also introduced CNN and established a spatial information model. The model in this study has its own advantages and is superior to other models under the same dataset [21].

Due to the small amount of data in remote sensing image datasets, traditional manual feature extraction methods can effectively avoid overfitting problems caused by small sample sizes. However, their classification performance is limited due to the significant impact of feature extraction. Previous studies have shown that due to the relatively small sample size of the training dataset, the overall resolution was poor. The use of deep transfer learning methods can to some extent compensate for the poor classification performance caused by a small number of training samples. In addition, feature extraction is particularly important for classification tasks. The method of feature fusion can improve the classification performance of the model to a certain extent. In summary, transfer learning and feature fusion have excellent classification performance in high-resolution remote sensing images. This article further combines the two technologies and proposes a high-resolution remote sensing image classification model suitable for small samples based on deep transfer learning and multi feature network fusion.

III. RSIC MODEL BASED ON DTL AND MULTI FEATURE NETWORK

A. DTL AND CNN ARCHITECTURE

Researchers have found that in convolutional neural networks, the features extracted by shallow convolutional layers are universal, including color, texture, shape, etc., known as low-level semantic features. And deeper convolutional layers can extract features closely related to tasks or datasets, called high-level semantic features, such as nose, mouth, eyes, etc. in facial images. These high-level semantic features are more suitable for subsequent classification tasks.

The deep transfer learning network model is constructed using the characteristics of convolutional neural networks at different levels. Firstly, train a high-performance classification network on a large dataset. Then, extract the convolutional layer structure and trained weights of the network. Next, combine the convolutional layer structure with the fully connected layer structure of the new task to construct a deep transfer learning network. During the training process of a new task, load the previously trained convolutional weights and perform fine-tuning or weight freezing training.

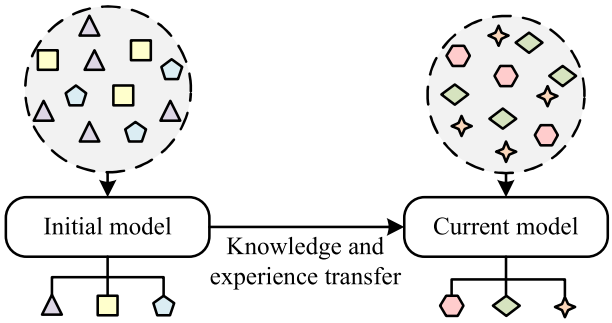


FIGURE 1. Schematic diagram of transfer learning.

Deep learning learns feature expressions with stronger generalization ability through strong data fitting ability, while transfer learning can learn feature expressions that are domain independent. Combining deep learning with transfer learning can fully utilize neural networks to learn common feature representations across different domains. This combination can leverage the advantages of deep learning and transfer learning, further improving model performance and applicability.

Transfer learning proves to be a highly effective approach when dealing with limited training specimens. CNN training sample size is too small, resulting in over-fitting phenomenon, and transfer learning can alleviate this situation. The core idea is to obtain useful knowledge from one or more source domain (SD) tasks and put this knowledge into the use of new target tasks. The crux lies in the transfer and application of knowledge, utilizing the principle of similarity between the SD and the target for transfer. Fig. 1 is the transfer learning diagram.

Assuming the two basic concepts of domain R and task T are defined, the domain generally consists of 2 parts: feature space X and edge probability distribution $p(x)$, as demonstrated in eq. (1).

$$p(x) = \{x = x_1, x_2, \cdots, x_n\} \in X \tag{1}$$

If two domains are different, then the probability distributions of the two feature spaces or edges are also different. Therefore, one of the domains is represented by equation (2).

$$D = \{X, p(x)\} \tag{2}$$

There are 2 parts of T, namely label space Y and prediction function $f(x)$, where $f(x)$ is obtained from model training. From the perspective of probability theory, $f(x)$ is basically equal to $p(y|x)$, that is, if the value of x is fixed, the probability of the output is y. In the classification task, y represents the set of all labels. The domain is able to be further separated into SDR_s and target domain R_t. The SD refers to a large

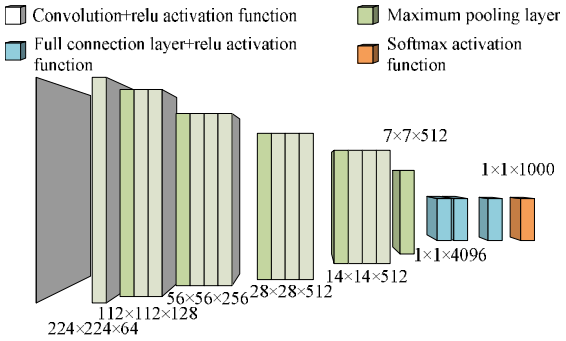


FIGURE 2. Structure diagram of VGG16.

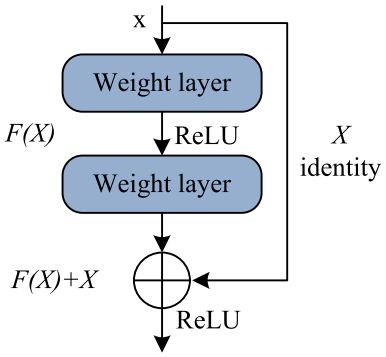


FIGURE 3. Structure of residual blocks.

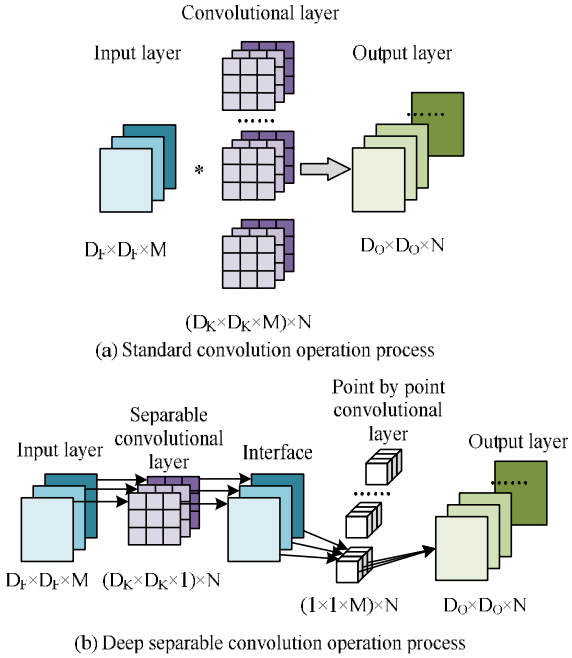


FIGURE 4. The process of standard convolution operation and deep separable convolution operation.

sample dataset used for pre training; The target domain refers to HiR-RSI with small samples that require classification. The task domain can also be correspondingly segmented into source task T_s and target task T_t. Source task refers to completing classification on a large-scale dataset; The target task refers to completing the HiR-RSI classification task in small sample situations. Therefore, transfer learning can also be denoted as that, given a SD and a source task, a target domain and a target task, the knowledge obtained from training in the SD and source task can be utilized to help the target domain learn the target prediction function. Equation (3) is the objective prediction function $T(t)$.

$$T(t) = f(R_t) \tag{3}$$

Eq. (3) satisfies conditions $R_s \neq R_t$ and $T_s \neq T_t$. For the network construction of DTL, the research uses VGG16, InceptionV3, ResNet50, and MobileNet as the basic CNN to design image classification models. The VGG series model is a classic and widely used CNN model, and Figure 2 shows the structural diagram of VGG16.

The structure of VGG16 is relatively simple, and the model depth is appropriately considered without lifting the total number of parameters. The convolutional kernel parameters used in VGG16 convolutional layers are the same, with 3 kernel size in width and height, 1step size, and a uniform filling method. This can maintain the same width and height for the convolution's input and output. The pooling layer of VGG16 also uses the same pooling parameters and adopts a maximum pooling strategy. Set the size of the pooling core to 2×2, and the step size is 2, resulting in an output size that is half the input size. Inception V3 is a new CNN model, which not only has a certain depth and width, but also introduces the idea of factorization into convolution operations.

As the depth of the convolutional layer continues to increase, deeper convolutional layers can extract advanced semantic features, which can improve the image classification accuracy. Nevertheless, deep convolutional layers also own defects such as vanishing or exploding gradients, leading to network convergence problems. The ResNet residual network structure can solve such problems in convolutional layers, and Fig.3 displays the structure of residual blocks.

ResNet50 is composed of stacked residual blocks, and the principle of residual blocks is displayed in equations (4) and (5).

$$y_l = h(x_l) + F(x_l, \omega_l) \tag{4}$$

$$X_{l+1} = f(y_l) \tag{5}$$

In Eq. (4) and (5), x_l and x_{l+1} are the input and output of the l-th residual unit. Where each residual unit typically contains a multi-layer convolutional structure. F means the residual function, representing the residual learned by the network; $h(x_l) = x_l$ is the identity mapping, and f represents the activation function of the rectifier linear unit (ReLU). This structure helps solve the problems of gradient dispersion and network performance degradation by adding a shortcut of the network. MobileNet is a lightweight deep CNN, and its biggest advantage is the use of deep separable convolution, a feature convolution structure with

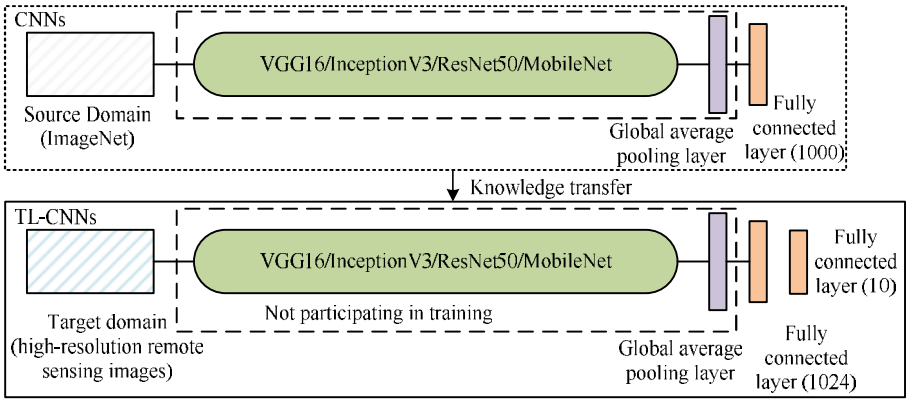


FIGURE 5. Deep transfer learning network model.

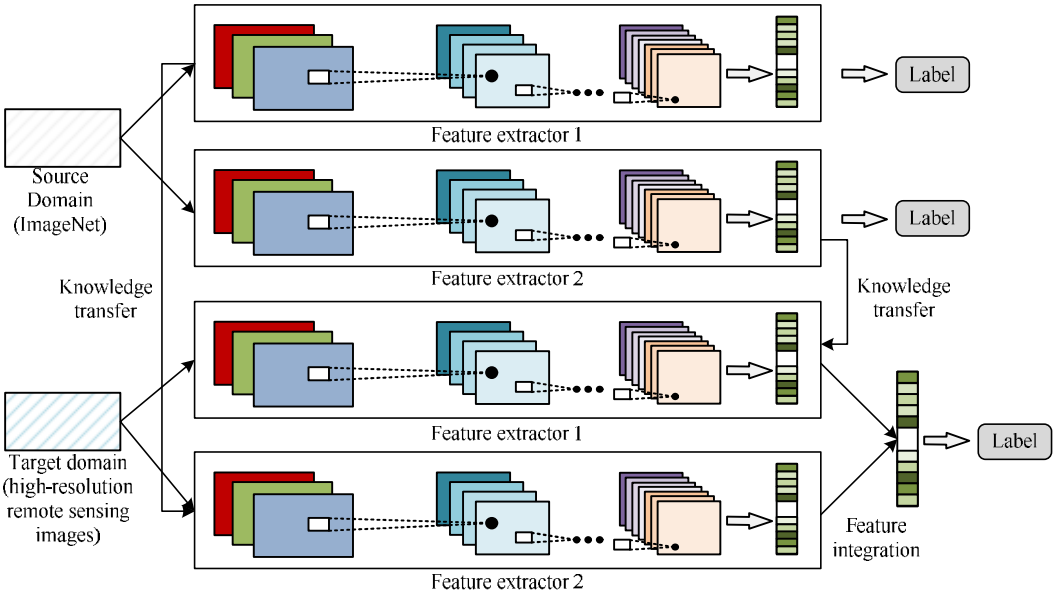


FIGURE 6. Fig.6 Basic model of deep multi feature fusion framework.

low computational complexity and low memory consumption for weight parameters. Deep separable convolution decomposes conventional standard convolution operations into two parts: deep convolution and point-by-point convolution. Deep convolution involves using different convolution kernels for different input channels, stacking the output results together, and finally further calculating the results through point by point convolution. Point-by-point convolution is a standard convolution method that uses 1×1 convolutional kernel. Eq. (6) is the calculation formula for standard convolution operations.

$$D_o = \frac{D_F + 2 \times p - D_K}{s} + 1 \tag{6}$$

p in equation (6) represents padding, and s is step size. The total parameter quantity of convolution operation is $D_K \times D_K \times M \times N$; The calculation cost of standard convolution operation is $D_K \times D_K \times M \times N \times D_F \times D_F$ while the output feature map space size remains unchanged. The process of standard convolution operation and deep separable convolution operation is listed in Fig.4.

B. RSIC MODEL BASED ON TRANSFER LEARNING AND MULTI FEATURE NETWORK

In accordance with four kinds of deep CNN, the research combines transfer learning to realize the construction of classification model. The ordinary CNN model is represented as CNNs, while the DTL network model is represented as TL-CNNs. The DTL network model is Fig.5.

The CNNs in Fig. 5 are ordinary CNNs, and the TL CNNs are DTL network models constructed by combining CNNs with transfer learning. Both share the weight parameters of convolution part obtained by CNN training. After completing the construction of the DTL network model, it is trained. The ImageNet dataset is used as the SD of transfer learning, and the CNN model is constructed under the framework of keras,

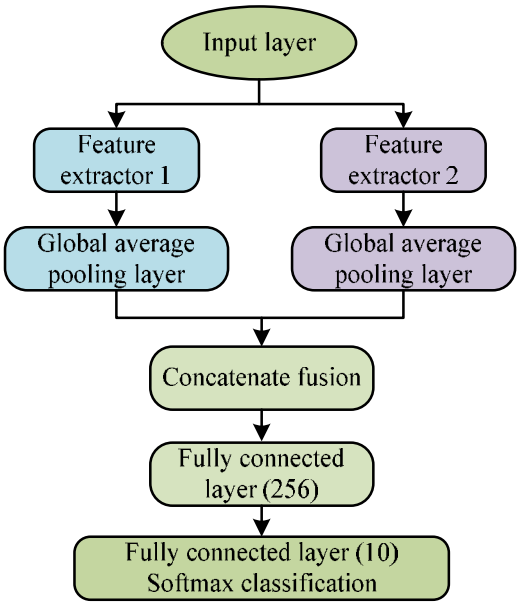


FIGURE 7. Fig.7 Structure diagram of CNN-based serial feature fusion.

and the weight parameters that pass the training are saved. The model is represented by equation (7).

$$F_s = f_{CNNs}\left(\omega, \omega_{fc}; x_i\right) \tag{7}$$

F_s in equation (7) represents the output of the model; ω is the weight parameter of the convolutional part in deep CNN; ω_{fc} refers to the weight parameters obtained through training in the fully connected layer of deep CNN; x_i means the training data of the SD. The HiR-RSI dataset of small samples is used as the target domain of transfer learning. From Figure 5 that CNNs and TL CNNs have the same convolutional part structure. Therefore, the weights of the convolutional part trained in CNNs are extracted and loaded into the convolutional part of TL-CNNs. During model training, only the fully connected layers require weight application, resulting in a reduction of training parameters and an acceleration of network training speed. The model is defined as equation (8).

$$F_T = f_{TL-CNNs}\left(\omega, \omega_{n-fc}; x_i\right) \tag{8}$$

F_T in equation (8) represents the output of the model; ω refers to the weight parameter of the convolutional part in deep migration CNN; ω_{n-fc} is the weight parameters obtained from the fully connected layer training of the model; x_i means the training data of the target domain. Through training among the four network structures, the TL-Inception V3, TL-Res Net50, and TL-Mobile Net models have relatively good classification performance on small sample HiR-RSI datasets. Thus, the study selected these three deep CNN models and constructed the basic model of a deep multi feature fusion framework through pair wise fusion. The basic model of the deep multi feature fusion framework is Fig. 6.

The methods of feature fusion are mainly segmented into 2 types: serial fusion and parallel fusion. From the analysis, it can be concluded that the feature fusion method of serial fusion has a better effect on classification accuracy. Therefore, the study chose the serial fusion method to fuse the model. The structural diagram of serial feature fusion based on CNN is exhibited in Figure 7.

TABLE 1. Classification of HiR-RSI.

CLASSIFICATION	TRAINING SET	TEST SET
STORAGE WAREHOUSE	10	100
RESIDENCE COMMUNITY	12	100
PARKING LOT	10	100
WHARF	15	100
ROAD	10	100
ROADSIDE TREES	20	100
BUILDING	10	100
BRIDGE	20	100
BOULEVARD	10	100
AIRPORT RUNWAY	15	100
OVERALL	132	1000

For the training of fusion models, ImageNet was selected as the SD dataset, and the target domain dataset is a small sample HiR-RSI dataset for classification. ResNet50, InceptionV3, and MobileNet are three types of CNNs used for training on the SD ImageNet dataset, and the weights of the trained convolutional parts are extracted and saved. Due to the use of pairwise fusion in the model, the two basic models are represented by m_1 and m_2, and formula (9) is the mathematical expression of m_1.

$$F_{m_1}(i) = f_{m_1}\left(\omega_{m_1-conv}, \omega_{m_1-fc}; x_i\right) \tag{9}$$

In equation (9), x_i is the i -th input of the SD dataset; ω_{m_1-conv} refers to the weight parameters obtained from model m_1 training; ω_{m_1-fc} is the weight parameters of the fully connected layer obtained from model m_1 training. Following the training on the SD, the features extracted by the convolutional layer are denoted as formula (10).

$$T_{m_1}(i) = f_{m_1}\left(\omega_{m_1-conv}; x_i\right) \tag{10}$$

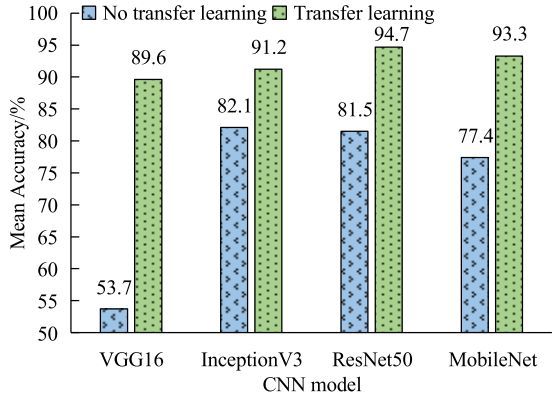
In the same way, the mathematical expression of m_2 is defined as formula (11).

$$F_{m_2}(i) = f_{m_2}\left(\omega_{m_2-conv}, \omega_{m_2-fc}; x_i\right) \tag{11}$$

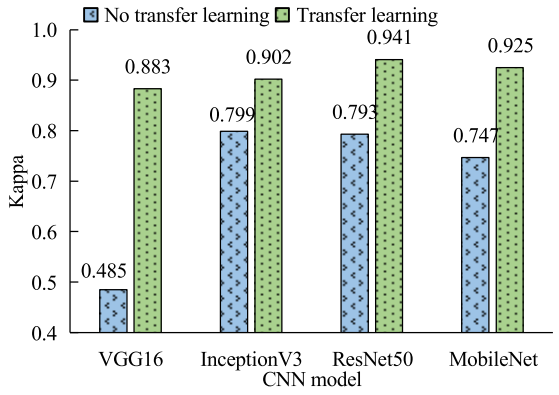
ω_{m_2-conv} in Eq. (11) means the weight parameters of the convolutional part obtained from model m_2 training; ω_{m_2-fc} refers to the weight parameters of the other layer obtained from model m_2 training. Eq. (12) can be derived as features through training on the SD.

$$T_{m_2}(i) = f_{m_2}\left(\omega_{m_2-conv}; x_i\right) \tag{12}$$

The feature vectors extracted from m_1 and m_2 are fused using a serial fusion method, and the fused features are

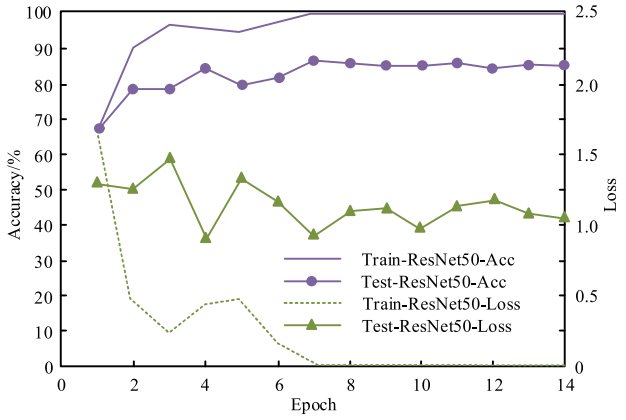


(a) Mean accuracy of different CNN models in two scenarios

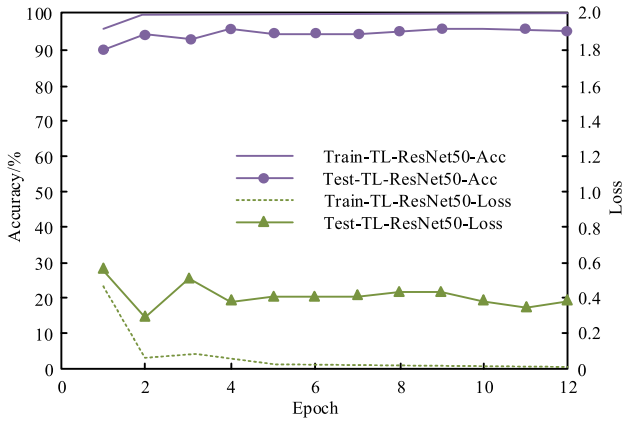


(b) Kappa of Different CNN Models in Two Cases

FIGURE 8. Classification performance of different CNN models without transfer learning and transfer learning.



(a) Accuracy and loss value of the model without transfer learning



(b) Accuracy and loss value of the model with transfer learning

FIGURE 9. Accuracy and loss values of the model in two scenarios.

expressed as Eq. (13).

$$T_M(i) = T_{m_1}(i) + T_{m_2}(i) \tag{13}$$

Fused characteristics are fed into a fully connected layer for subsequent classification tasks, and finally, the deep multi feature network fusion model is shown in equation (14).

$$F_M(i) = f\left(T_M(i), \omega_{M-fc}; x_i\right) \tag{14}$$

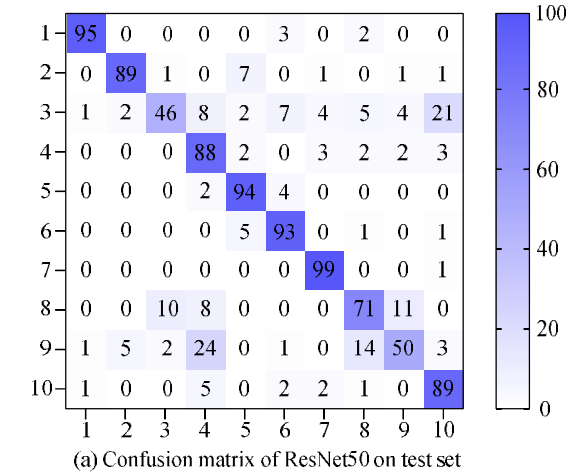
In equation (14), ω_{M-fc} represents the weight parameter of the fully connected layer. Save the weight parameters of the convolutional part of the new model and train the parameters of the fully connected layer separately. For the convenience of representation, the deep multi feature network fusion model constructed based on TL ResNet50 and TL MobileNet networks is referred to as the Transfer Learning-ResNet50-MobileNet model (TL-RM). The same applies to other deep multi feature network fusion models. In traditional multi feature network fusion models, the weight adoption number of the convolutional part and the weight parameters of the fully connected layer are only initialized before training begins. During the training process after initialization, these parameters will be iteratively updated as the constructed fusion model continues to train. The traditional multi feature network fusion model constructed using ResNet50 and MobileNet as the basic models is referred to as RM, and other traditional multi feature network fusion models are the same.

IV. EXPERIMENTAL ANALYSIS OF RSIC MODEL COMBINING DTL AND MULTI-FEATURE NETWORK

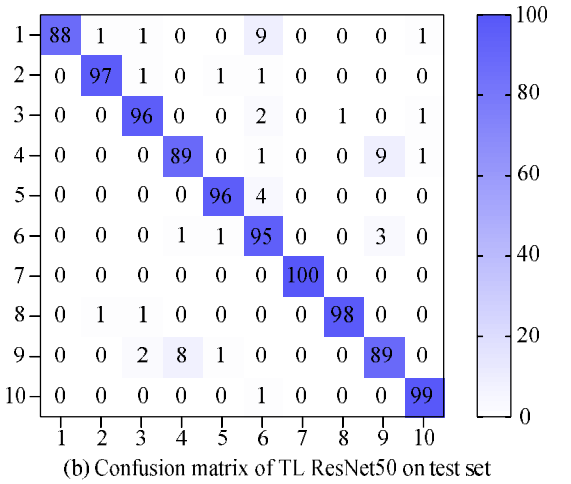
A. PERFORMANCE ANALYSIS OF RSIC MODEL BUILT ON DTL

To evaluate the classification efficacy of the proposed fusion-DTL and multi-feature network model on HiR-RSI, an experimental analysis was conducted to assess the performance of the model. The Python language is adopted for code programming, and the Keras deep learning framework is used as the experimental environment framework. The UC land-use dataset and the RSIC benchmark dataset RSI-CB were selected as experimental datasets for the study. UC dataset includes 21 different categories of HiR-RSI, with each image containing three RGB channels and an image size of 256 × 256. RSI-CB can be divided into 256 × 256 and 128 × 128 datasets of two sizes, each containing 35 and 45 categories. This study extracted 10 difficult to classify HiR-RSI categories from two datasets (Table 1).

The study used VGG16, InceptionV3, ResNet50, and MobileNet as the basic network models to construct deep

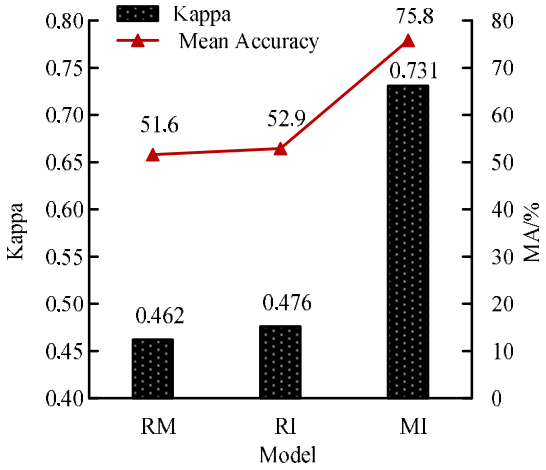


(a) Confusion matrix of ResNet50 on test set



(b) Confusion matrix of TL ResNet50 on test set

FIGURE 10. Confusion matrix of ResNet50 and TL-ResNet classified on test set.



(a) Classification performance of traditional multi feature network fusion models

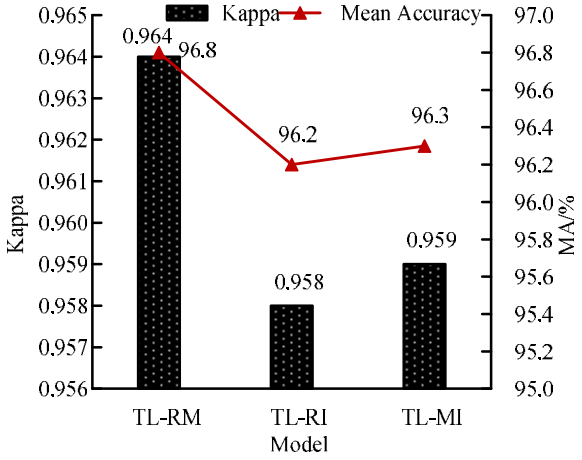


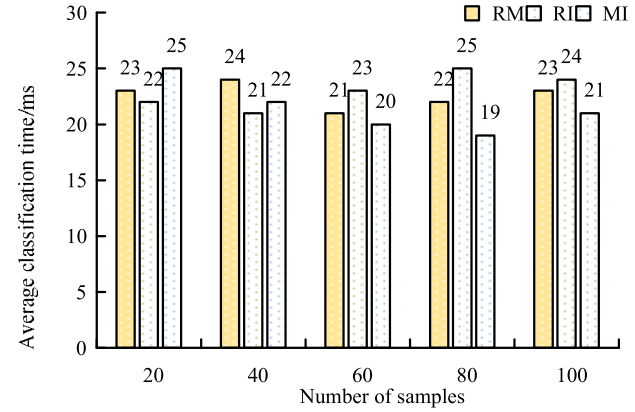
FIGURE 11. Classification results of traditional multi feature network fusion models and deep multi feature network fusion models.

migration CNN based on four types of networks. Without transfer learning and transfer learning, the deep CNN model is used to classify image data. Fig. 7 shows the experimental results.

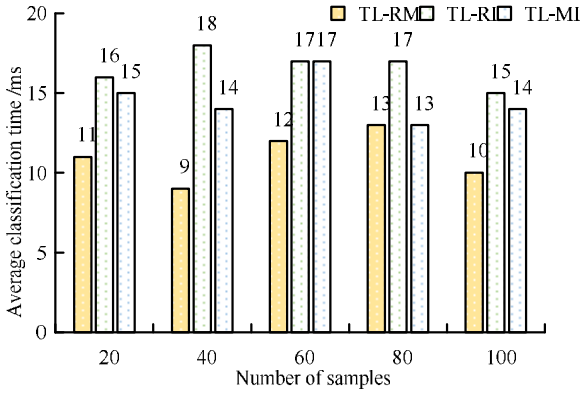
From Fig. 8 (a), the deep CNN models based on the four networks have higher average accuracy with DTL compared to those without DTL. The average accuracy of VGG16, Inception V3, ResNet50 and MobileNet without transfer learning was 53.7%, 82.1%, 81.5% and 77.4% respectively; In the case of transfer learning, it was 89.6%, 91.2%, 94.7% and 93.3% respectively. In Fig. 8 (b), the Kappa coefficients of VGG16, Inception V3, ResNet50 and MobileNet without transfer learning are 0.485, 0799, 0.793 and 0.747 respectively; In contrast, it is 0.883, 0.902, 0.941 and 0.925 respectively. From this, DTL has a significant improvement effect on the classification performance of CNN models.

The training curves of the four classification models are basically consistent, and ResNet50 was selected as an example for analysis. Under the conditions of no transfer learning and transfer learning, as the iterations in the training process lifts, the accuracy and loss trend of the model on both sets is demonstrated in Fig. 9.

From Figure 9 (a), as the quantity of model iterations rises, the accuracy of the training and testing sets gradually increases. When the iterations reaches 7, the accuracy of the model stabilizes at a high value, with the training set accuracy basically stable at 100% and the other stable at around 84%. The loss function of the model in the training set shows a trend of rapid decline first and then tends to be stable; The loss function on the test set fluctuates up and down and gradually decreases. In Fig. 9 (b), the trend of model accuracy is slowly increasing and gradually stabilizing, with training set's accuracy reaching 100% earlier and test set' accuracy stabilizing at around 95%. The loss function of the model tends to decline first and then to stabilize. From the figure that the model's accuracy is superior and the loss function value is smaller when DTL is performed. The research still takes ResNet50 network as an example, and classifies the results

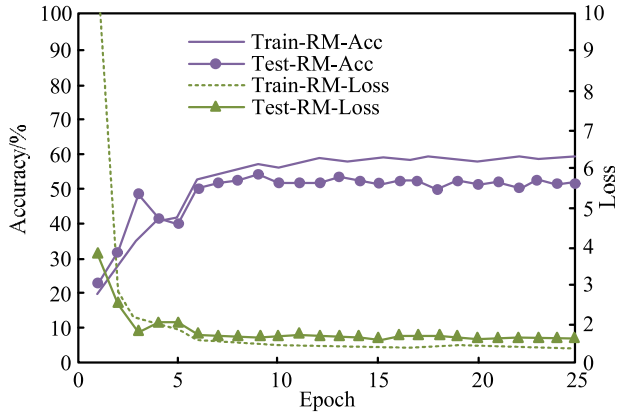


(a) The average classification time of traditional multi feature network fusion models

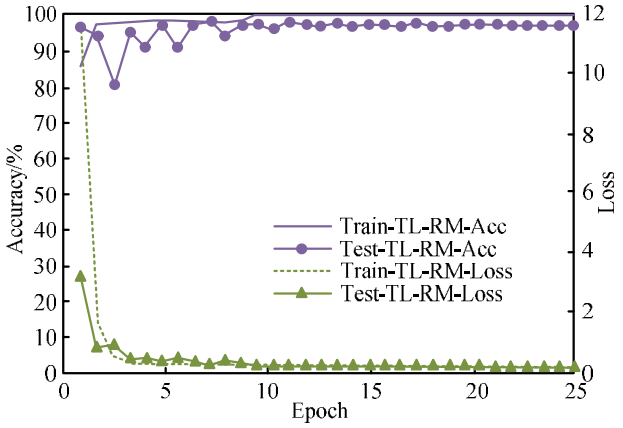


(b) Average classification time of deep multi feature network fusion model

FIGURE 12. Comparison of average classification time between traditional multi feature network fusion models and deep multi feature network fusion models.



(a) The accuracy and loss trend of RM model training process



(b) The accuracy and loss trend of TL-RM model training process

FIGURE 13. The accuracy and loss trend of the training process of RM model and TL-RM model.

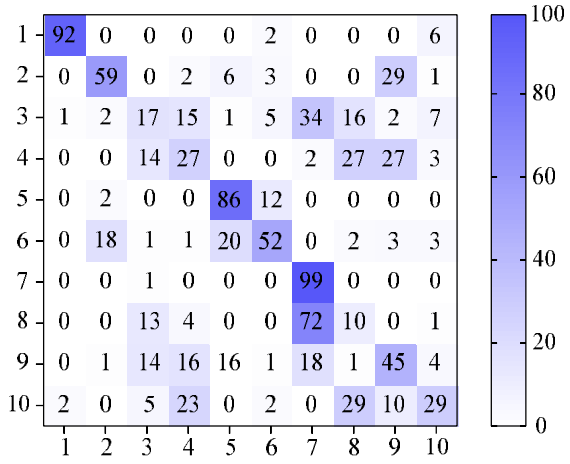
on the test set to obtain the confusion matrix without transfer learning and transfer learning (Fig. 10).

In Fig.10, the 1-10 horizontal and vertical axes correspond to airport runways, boulevards, bridges, buildings, roadside trees, roads, docks, parking lots, residential areas, and storage warehouses in remote sensing images. Without transfer learning, the classification accuracy of remote sensing images of airport runways, roadside trees, roads and docks reached 95%, 94%, 93% and 99% respectively, and the data reached more than 90%. The classification accuracy of remote sensing images of docks is the highest, followed by boulevards, buildings, and storage warehouses, with a classification accuracy of over 80%. The classification accuracy of remote sensing images of boulevards, buildings, and storage warehouses has reached 89%, 88%, and 89%. The main categories that have not achieved an accuracy rate of 80% are bridges, parking lots, and residential areas. In the case of transfer learning, the classification accuracy rate of tree lined boulevards, bridges, roadside trees, roads, docks, parking lots and storage warehouses has reached more than 90%. However, only the classification accuracy of airport runways, buildings, and residential areas has not reached 90%. Among them, the classification accuracy of remote sensing images of airport runways, buildings, and residential areas is only 88%, 89%, and 89%. The accuracy of the three types of remote sensing images that did not achieve 90% classification accuracy is still not lower than 88%, and the overall accuracy remains around 90%. Hence, the model after transfer learning has more obvious advantages in classification accuracy, and the classification accuracy on specific remote sensing images can reach 100%. Transfer learning can significantly lift the accuracy, but this effect is unlike in different remote sensing image types.

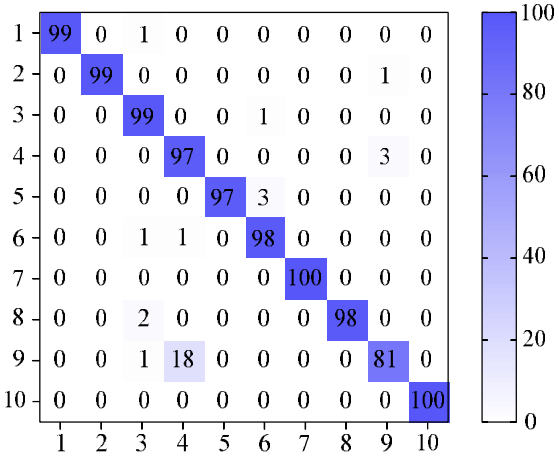
B. PERFORMANCE ANALYSIS OF FUSION RSIC MODEL

Aiming at the performance of the deep multiple feature network fusion model, the research will conduct experimental analysis on it under the premise of transfer learning. The classification results of three traditional multi feature network fusion models and three deep multi feature network fusion models on a small sample HiR-RSI dataset are demonstrated in Figure 11.

In Figure 11, the accuracy and Kappa coefficient of the three deep multi feature network fusion models are higher



(a) Confusion matrix of RM model on test set



(b) Confusion matrix of TL-RM model on test set

FIGURE 14. Confusion matrix of RM model and TL-RM model on test set.

than those of the three traditional multi feature network fusion models, and their advantages are obvious. The maximum average accuracy of traditional multi feature network fusion models on small sample HiR-RSI datasets is 75.8%, and the maximum Kappa coefficient is 0.731; The maximum average accuracy of the deep multi feature network fusion model is 96.8%, and the maximum Kappa coefficient is 0.964. Fig. 12 shows the average classification time comparison of three traditional multi feature network fusion models and three deep multi feature network fusion models under different sample sizes.

From Figure 12, in the comparison of traditional models, when the quantity of samples is relatively small, with 20 and 40 cases respectively, the RI model has a more obvious classification time advantage, with classification times of 22ms and 21ms, respectively. When the number of samples reaches 60 or more, MI has a better classification time advantage, with classification times of 20ms, 19ms, and 21ms, respectively. Overall, RM does not have the advantage of classification time. After deep improvement, TL-RM has classification time advantages at all sample size levels. When the sample size was 20, 40, 60, 80, and 100, the classification time of TL-RM was 11ms, 9ms, 12ms, 13ms, and 10ms respectively, showing an overall trend of fluctuation around 10ms. This has significant advantages in classification time compared to other improved models. This indicates that RM has a better foundation for improvement compared to other models. Compared to RI and MI models, the foundation for improvement is relatively poor. Taking the RM model and TL-RM model as examples, the trend of accuracy and loss values during the training process of the two models is listed in Figure 13.

In Figure 13, the accuracy of RM on the training and testing sets exhibits an initial increase followed by stabilization, ultimately stabilizing at around 60% and 52%. The accuracy trend of TL-RM on the training and testing sets shows a slight increase and tends to stabilize, ultimately stabilizing at around 100% and 98%, demonstrating its superior accuracy and stability. The loss function of RM and TL-RM both showed a trend of sharp decline first and then stability. The loss value of RM finally stabilized at about 7%, and the loss value of TL-RM finally stabilized at about 1%. This indicates that TL-RM has significant stability and reliability. The specific classification confusion matrix on the test set after the training of the two models is Figure 14.

As Fig. 14 (a), the RM model has high classification accuracy for only airport runways and docks in remote sensing images, reaching 92% and 99% respectively. In Fig. 14 (b), the TL-RM model only has a low classification accuracy of 81% for residential areas in remote sensing images; The classification accuracy of other categories has reached over 97%, demonstrating significant accuracy, stability, and reliability.

V. CONCLUSION

As the result of the particularity and complexity of HiR-RSI acquisition method, the classification model for remote sensing images often has a over-fitting problem caused by insufficient training samples. A RSIC model built on DTL and multi-feature network fusion is proposed to address such issues. Based on DTL, the research uses serial fusion to fuse basic CNN to build a classification model, and experiments verify the effectiveness of transfer learning and the reliability of TL-RM model. The average accuracy of VGG16, Inception V3, ResNet50 and MobileNet in transfer learning was 89.6%, 91.2%, 94.7% and 93.3% respectively; Kappa coefficients are 0.883, 0.902, 0.941 and 0.925, respectively, which are superior to the performance without transfer learning. Taking ResNet50 as an example, under the condition of transfer learning, the model's accuracy on the training set has reached 100% earlier, and on the test set is stable at about 95%. The average accuracy of TL-RM on the remote sensing image dataset is 96.8%, and the Kappa coefficient on the remote sensing image dataset is 0.964, which has significant advantages compared to other models. The accuracy of the training and testing sets is ultimately stable at around 100% and 98%. And the TL-RM model has achieved a classification accuracy of over 97% for all categories

in remote sensing images except for residential areas. The results demonstrate the superior classification accuracy and reliability of the TL-RM model. There may be significant differences in the classification performance for different categories in the research, so ensemble learning can be considered to further improve the model.

REFERENCES

[1] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu, M. Weinmann, S. Hinz, C. Wang, and K. Fu, "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 116–130, Feb. 2022.

[2] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

[3] F. Özyurt, "Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures," *J. Supercomput.*, vol. 76, no. 11, pp. 8413–8431, Dec. 2019.

[4] Y. Xie, J. Zhu, Y. Cao, D. Feng, M. Hu, W. Li, Y. Zhang, and L. Fu, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, 2020.

[5] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[6] J. Gong, C. Liu, and X. Huang, "Advances in urban information extraction from high-resolution remote sensing imagery," *Sci. China Earth Sci.*, vol. 63, no. 4, pp. 463–475, Dec. 2019.

[7] X. Li, B. Liu, G. Zheng, Y. Ren, S. Zhang, Y. Liu, L. Gao, Y. Liu, B. Zhang, and F. Wang, "Deep-learning-based information mining from ocean remote-sensing imagery," *Nat. Sci. Rev.*, vol. 7, no. 10, pp. 1584–1605, Oct. 2020.

[8] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Tech. Rev.*, vol. 38, no. 4, pp. 377–396, Mar. 2020.

[9] W. Li, Z. Wang, Y. Wang, J. Wu, J. Wang, Y. Jia, and G. Gui, "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1986–1995, 2020.

[10] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J. Biomolecular Struct. Dyn.*, vol. 39, no. 15, pp. 5682–5689, Jul. 2020.

[11] S. Ahuja, B. K. Panigrahi, N. Dey, V. Rajinikanth, and T. K. Gandhi, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," *Appl. Intell.*, vol. 51, pp. 571–585, Jun. 2020.

[12] C. Liu, Z. Wei, D. W. K. Ng, J. Yuan, and Y. C. Liang, "Deep transfer learning-assisted signal detection for ambient backscatter communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Nov. 2020, vol. 20, no. 3, pp. 1624–1638.

[13] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards more accurate automatic sleep staging via deep transfer learning," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 1787–1798, Jun. 2021.

[14] A. Naseer, M. Rani, S. Naz, M. I. Razzak, M. Imran, and G. Xu, "Refining Parkinson's neurological disorder identification through deep transfer learning," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 839–854, Feb. 2019.

[15] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[16] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[17] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[18] S. Qiu, M. Wang, Y. Yang, G. Yu, J. Wang, Z. Yan, C. Domeniconi, and M. Guo, "Meta multi-instance multi-label learning by heterogeneous network fusion," *Inf. Fusion*, vol. 94, pp. 272–283, Jun. 2023.

[19] Y. Wang, X. Zhang, G. Ma, X. Du, N. Shaheen, and H. Mao, "Recognition of weeds at asparagus fields using multi-feature fusion and backpropagation neural network," *Int. J. Agricult. Biol. Eng.*, vol. 14, no. 3, pp. 190–198, 2021.

[20] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.

[21] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

XINYAN HUANG was born in Shandong, China, in 1978. She received the B.S. degree from the Shandong University of Finance and Economics, in 2001, the M.S. degree from the Ocean University of China, in 2006, and the Ph.D. degree from Shandong University, in 2016. Since 2006, she has been an Instructor with the School of Computer Science and Technology, Shandong University of Finance and Economics. She is the author of more than six articles. Her research interests include big data analysis, big data mining, and machine learning.