Application of IMVR Convolutional Neural Networks to Classification of Land Use Remote Sensing Datasets

1st Yuanzhen Shuai
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Yuanzhen.Shuai19@student.xjtlu.edu.cn

2nd Ning Xin
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Ning.Xin21@student.xjtlu.edu.cn

3rd Md Maruf Hasan
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
MdMaruf.Hasan@xjtlu.edu.cn

4th Bintao.hu
School of Internet of Things
Xi'an Jiaotong-Liverpool University
Suzhou, China
Bintao.hu@xjtlu.edu.cn

5th Tianhong.Dai
School of Natural and Computing Sciences
The University of Aberdeen
Aberdeen,UK
tianhong.dai@abdn.ac.uk

6th Hengyan Liu*
School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Hengyan.Liu@xjtlu.edu.cn

***Abstract*—Despite extensive research, remote sensing image classification remains a challenging issue within the field of remote sensing image analysis. Achieving a balance between classification accuracy and computational efficiency remains challenging, as traditional methods often face difficulties in attaining both high speed and precision simultaneously. To tackle this dilemma, we propose a method named IMVR which significantly reduces the computational burden while maintaining validity. This method enhances the richness and accuracy of high-dimensional feature representations through its output. Extensive experiments are conducted on the UC Merced Land-Use Dataset to demonstrate that our method can substantially improve classification performance and efficiency in comparison to traditional methods.**

***Index Terms*—Feature classification, image classification, deep learning, Remote sensing image classification**

I. Introduction

Remote sensing image classification plays a crucial role in analysing remote sensing images, enabling accurate identification and classification of different feature categories in these images. It has significant implications in various fields, such as environmental monitoring, urban planning, and agricultural management. With the continuous advancement of satellite technology, remote sensing images' acquisition and processing capabilities have been greatly improved [1]. Satellites now provide higher spatial resolution, which means that they can capture more detailed surface information. In addition, the availability of multispectral and hyperspectral sensors makes it possible to capture images in multiple bands, allowing for more comprehensive and precise information extraction Advances in image processing algorithms and techniques have enabled more efficient and accurate extraction of valuable information from remotely sensed images.

Remote sensing image classification involves the task of assigning predefined labels to pixels or regions in a remote sensing image based on their spectral, spatial, and contextual characteristics. The goal is to accurately classify different land cover types or objects present in the image. This task is challenging due to the complexity and variability of remote sensing data, including variations in illumination, scale, and spatial distribution [2].

Traditionally, remote sensing image classification methods have been categorized into three main approaches based on the level of visual features used: low-level, mid-level, and deep learning-based methods.

Low-level visual feature-based methods focus on extracting features from the low-level visual attributes of high-resolution remote sensing images. Commonly used methods include color histograms [3] and scale-invariant feature transform [4]. These traditional methods demonstrate good classification performance for high-resolution remote sensing images with uniform spatial distribution and structural patterns. However, they often fail to perform well in scenes with non-uniform spatial distribution.

Mid-level visual representation-based methods aim to encode the low-level local visual features of high-resolution remote sensing images to form a global feature representation of the scene. Common encoding models include bag-of-visual-words [5], spatial pyramid matching [6], local constrained linear coding [7], probabilistic latent semantic analysis [8]. Compared to low-level visual feature-based methods, mid-level visual representation-based methods have Mid-level visual representation-based methods aim to encode the low-level

We sincerely thank the Climatic Data Centre, part of the National Meteorological Information Centre (CMA Meteorological Data Centre), for their invaluable assistance and cooperation in providing us with the meteorological data used in this study.

local visual features of high-resolution remote sensing images to form a global feature representation of the scene. Common encoding models include bag-of-visual-words [5], spatial pyramid matching [6], local constrained linear coding [7], probabilistic latent semantic analysis [8]. Compared to low-level visual feature-based methods, mid-level visual representation-based methods have shown significant improvements in classification accuracy. However, they are still limited by the low-level visual features and encoding methods, which prevent them from achieving optimal classification performance and classification accuracy.

In recent years, deep learning-based methods have attracted extensive attention due to their remarkable achievements across various fields. These methods utilize deep neural networks to automatically learn image features and improve classification performance. Commonly used deep learning models for remote sensing image classification include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders. Subsequently, ResNet-50, a CNN with skip connections, has proven highly effective for computer vision tasks [9]. Additionally, Inception V3 improves the neural network structure by decomposing the original large convolution kernel into small convolution kernels with equivalent operations [10], performing spatial decomposition of asymmetric convolution, and using auxiliary filters while further reducing the feature map and computation amount. This enables more effective preservation of image features, extracting remote sensing image features well while maintaining excellent training speed. These methods have achieved remarkable success in classifying high-resolution remote sensing images, effectively handling complex spatial distribution and structural patterns, and improving classification accuracy and robustness.

In this paper, we propose the IMVR model, which integrates the respective advantages of Inception, MobileNet, VGG, and ResNet50 through transfer learning. Transfer learning is utilized to perform initial preprocessing on the large natural image dataset ImageNet. To evaluate the performance of the proposed method, we compare IMVR against previous classical models using large benchmark datasets. The ensemble model harnessing the strengths of specialized neural networks demonstrates advanced performance for remote sensing image classification. The critical contributions of this work are as follows:

1. A new remote sensing image classification neural network is proposed, IMVR, which demonstrates improved performance compared to traditional remote sensing classification methods.
2. The model can monitor specific characteristics of a given area in real-time, enabling supervisors and decision-makers to obtain the latest information in real-time to make more informed decisions.
3. This work advances interdisciplinary research across machine learning, geography, climate science, etc., opening avenues for applying the proposed techniques in diverse fields such as geography and pollutant distribution analysis.

II. METHODOLOGY

A. Basic models

1) Inception v3: Inception V3 has optimised the structure of the Inception Module, and there are now more varieties of Inception Modules as shown below, and the practice of splitting a larger two-dimensional convolution into two smaller one-dimensional convolutions has also been introduced in Inception V3 [11]. For example, a 7 × 7 convolution can be split into a 1 × 7 convolution and a 7 × l convolution. This kind of asymmetric convolutional structure splitting is better than symmetric convolutional structure splitting in terms of handling more and richer spatial features and increasing feature diversity, and at the same time, it can reduce the amount of computation.

2) Resnet50: The ResNet-50 [12] network structure comprises two fundamental blocks: the Conv Block and the Identity Block, with a total of four blocks in this connection module. The complete model is depicted below.

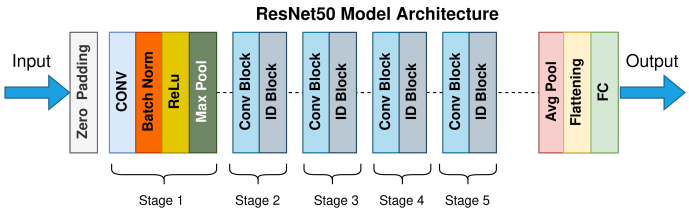

Fig. 1. ResNet50 Module [12]:From stage1-stage3 there will be two kinds of Bottleneck, two kinds of Bottleneck corresponds to two kinds of situations: the same number of input and output channels (BTNK2), the number of input and output channels are different (BTNK1), stage1 first use BTNK1 and then add two BTNK2, stage2 use a BTNK1 and then add three BTNK2, stage3 use BTNK1 and then add two BTNK2, stage4 use BTNK1 and then add five BTNK2

3) Mobilenet: Mobilenet replaces ordinary convolution with deep separable convolution [13], the convolution formula for deep convolution is as Eq.1. In Eq.1, Output(i,j) denotes the value of the output feature map at position (i,j), Input$(i+m,j+n,k)$ denotes the value of the input feature map at position $(i+m,j+n)$ and channel k, and Kernel(m,n,k) denotes the value of the convolution kernel at position(m,n). In Eq.1, the variables of channel k, M, N and K denote the height, width and number of channels of the convolution kernel, respectively.

$$\text{Output}(i,j) = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} \text{Input}(i+m,j+n,k) \times \text{Kernel}(m,n,k) \tag{1}$$

The computation of depth separable convolution is also composed of two parts: the convolution kernel size of depth convolution is $D_k * D_k * M$, and a total of $D_w * D_h$ multiplication and addition operations have to be done; the convolution kernel size of point-by-point convolution is $1 * 1 * M$, and there are N of them, and a total of $D_w * D_h$ multiplication and addition operations have to be done so

that the calculation amount of depth separable convolution is: $D_K * D_K * M * D_w * D_h + M * N * D_w * D_h$.

4) VGG: In VGG [14](Visual Geometry Group), three 3x3 convolutional kernels are used instead of the 7x7 convolutional kernels in AlexNet [15], and two 3x3 convolutional kernels are used instead of the 5*5 convolutional kernels, and the main purpose of this is to enhance the depth of the network under the condition of ensuring that it has the same perceptual field, which enhances the effect of neural network to some extent.

B. IMVR

We chose four classical CNN models as feature extractors, NASNetMobile, ResNet50, VGG16, and InceptionV3. These models were pre-trained on ImageNet on large-scale datasets and have good feature extraction capabilities. We constructed classifiers by concatenating their outputs and adding a fully connected layer and softmax layer. During training, we used data enhancement and preprocessing techniques, including operations such as image rotation, translation, cropping, scaling and horizontal flipping to increase the diversity and generalisation of the training data. According to the transfer learning technique, the pre-trained model weights are frozen, and only the classifier weights are updated.

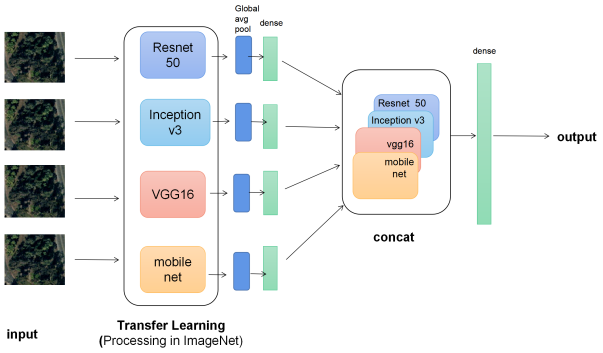

Fig. 2. IMVR:Firstly, the 256*256 remote image data will be pre-processed and passed into the pre-trained Resnet50, Inception v3, Mobilenet and Vgg16, respectively, and extracted to 1*256 high-dimensional features by GlobalAveragePooling2D and Dense to concatenate the high-dimensional features of the four models, and then carry out the multi-classification task of 21 classes by Dense.

1) Transfer Learning: Transfer learning [16] can effectively solve the information silo problem by transferring effective information from the original domain to improve the learning and training efficiency of another domain (target domain), which can effectively solve the information silo problem. Using the powerful functions of deep neural networks and imagenet datasets, the knowledge learnt from natural image processing models applicable to large data volumes can be transferred to remote sensing image datasets applicable to small data volumes to achieve effective migration.

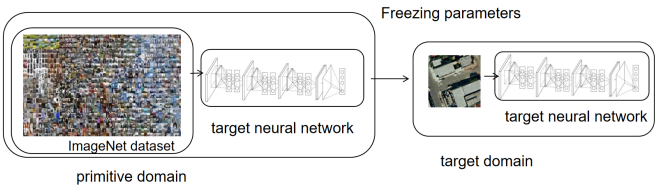

Fig. 3. Transfer Learning: the target network is trained with ImageNet parameters in the original domain, the corresponding parameters are frozen after training, and remote sensing images are passed into the already pre-trained model in the target domain for the multi-classification task .

2) Convolutional neural network: Convolutional neural networks have the ability of representation learning, able to shift-invariant classification of the input information according to its hierarchical structure, convolutional layer, there are two key operations, local correlation and window sliding, each convolutional neuron serves as a filter through the corresponding parameter to carry out the sliding to carry out the calculation of the local data, to get the high-dimensional features of the image.
The full for convolution is:

$$z(u,v) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} x_{i,j} \cdot k_{u-i,v-j} \tag{2}$$

The defining equation for the convolution is:

$$z(u,v) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} x_{i+u,j+v} \cdot k_{roti,j} \cdot \chi(i,j) \tag{3}$$

$$\chi(i,j) = \begin{cases} 1, & 0 \leqslant i,j \leqslant n \\ 0, & \text{others} \end{cases} \tag{4}$$

Backpropagation calculates the residuals (error term) for gradient descent: The detailed derivation of backpropagation is in Eqs. 5-7.

3) Concatenate: Our model chose to use Concatenate to extract high-dimensional features as this retains more information. Cascading would connect the outputs of the model sequentially to form a longer feature vector, which would lose the information interaction between the models. Concatenate helps to improve the expressiveness and generalisation of the model. And it can be more flexible to concatenate in different dimensions without the limitation of dimension matching, which provides more flexibility and freedom.

III. EXPERIMENTS

A. Dataset Description

The UC Merced Land-Use Dataset [17] used in this paper is a 21-class remote sensing dataset of land-use imagery for research purposes, with a total of 100 classes of imagery extracted from the USGS National Map Urban Area Imagery series, which is used in urban areas across the country. This dataset of public domain images has a pixel resolution of 1 ft, an image pixel size of 256*256, and contains a total of 2100 scene images in 21 classes, of which 100 are in each class.

$$\delta_{g,h}^{(l)} = \frac{\partial J(W,b;x,y)}{\partial z_{g,h}^{(l)}} = \sum_{i=0}^{r-1}\sum_{j=0}^{r-1} \frac{\partial J(W,b;x,y)}{\partial z_{i,j}^{(l+1)}} \frac{\partial z_{i,j}^{(l+1)}}{\partial z_{g,h}^{(l)}} \tag{5}$$

$$= \sum_{i=0}^{r-1}\sum_{j=0}^{r-1} \frac{\partial J(W,b;x,y)}{\partial z_{i,j}^{(l+1)}} \frac{\partial \beta^{(l+1)} \sum_{u=ir}^{(i+1)r-1} \sum_{v=jr}^{(j+1)r-1} f\left(z_{u,v}^{(l)}\right) + b^{(l+1)}}{\partial z_{g,h}^{(l)}} \tag{6}$$

$$= \beta^{(l+1)} \delta_{i+pr,j+qr}^{(l+1)} f'(z_{g,h}^{(l)}) \tag{7}$$

B. Evaluation Methods

The choice of accuracy rate as the evaluation criterion rate is one of the most intuitive and commonly used evaluation metrics, which provides a simple measure of how correct the model is in its predictions. The accuracy rate can intuitively reflect the model's prediction accuracy, i.e., the proportion of correctly predicted samples. Calculating the accuracy rate is very simple, need to count the number of correctly predicted samples and the total number of samples.

The accuracy rate is very widely used: accuracy rate is one of the most commonly used evaluation metrics in machine learning and deep learning and is widely used in various tasks and fields.

C. Remote Sensing Image Recognition

1) Preprocessing: The raw data is first normalised by scaling the image's pixel values to between 0 and 1 by dividing the pixel values by 255. It is convenient for model training and optimisation to unify the pixel values of the image into a smaller range. The data enhancement is carried out randomly by rotating the image by a certain range of angles can increase the diversity of data so that the model has a certain degree of invariance for different angles of the image, and then randomly translating the position of the image can simulate the changes of the image under different positions, increase the diversity of data, and improve the generalisation ability of the model.

2) Trend analysis: Vgg's Validation Accuracy does trend upwards along with the Train Accuracy trend, but the climb is slow. Resnet50 Validation Accuracy does not compare well with the Train Accuracy trend and remains low, mobilenet Validation Accuracy and Train Accuracy trends are both increasing but with low initial accuracy, and inceptionV3 Validation Accuracy and Train Accuracy trends are both increasing but with low initial accuracy. The inceptionV3 Validation Accuracy and Train Accuracy trends are both increasing, but the initial accuracy is low and fluctuates greatly.
IMVR validation Accuracy and Train Accuracy trends are both increasing, with high initial accuracy of 0.85 and little fluctuation.

IV. Result and Discussion

According to the 100 epoch training accuracy graph comparison can be seen that 10 epoch has reached the optimal value of the model. To prevent overfitting, choose to use 10 epoch accuracy comparison.

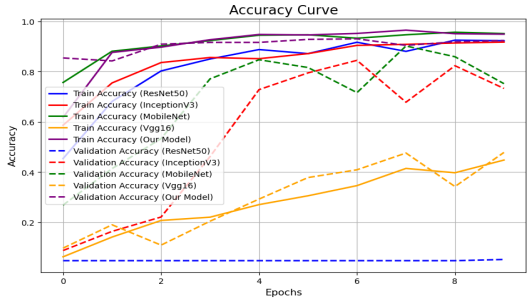

Fig. 4. Training results of 10 epochs

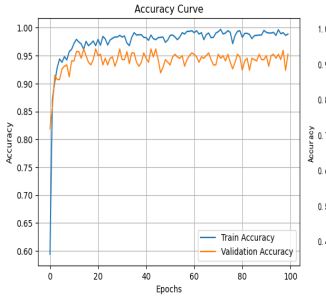

Fig. 5. IMVR:Training performance of 100 epochs

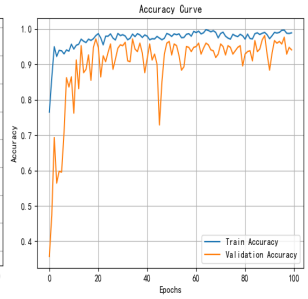

Fig. 6. Mobilenet:Training performance of 100 epochs

The main results are shown in Table 2. From the table, it can be seen that IMVR successfully outperforms other methods with a test accuracy of 44.52% over Vgg, 17.14% over Mobilenet, and 19.05% over Inception, which is strong proof of the effectiveness of our proposed method. From the performance comparison based on Mobilenet and Resnet, it can be concluded with certainty that different effective features can be extracted from different models, and more high dimensional features can be extracted from the fusion of multiple models, which can be used to judge the categories more accurately in classification. And IMVR has higher accuracy in the same epoch through the comparison graph of accuracy, which indicates that the advantage of IMVR in training efficiency is obvious.

The computational setup utilized for this analysis consisted of CPU 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz 2.30 GHz. Consequently, hardware limitations may exist when attempting to optimize model parameters further. With more advanced hardware devices and larger data availability, clearer and more precise classification results can be achieved.

TABLE I
EXPERIMENT RESULT

	test loss	test accuracy
Vgg	1.5709	0.4786
Inception V3	1.2600	0.7333
MobileNet	1.2020	0.7524
Our Approach	**0.2314**	**0.9238**

V. CONCLUSION

In this thesis, we propose an integrated model-based approach. In order to improve the training efficiency, we applied migration learning to pre-train on Imagenet large dataset. Compared with the classical deep learning model alone, the accuracy was chosen as the evaluation criterion for the test. Experimental results are demonstrated with comparisons of the performance, i.e., accuracy.

Comparison results on the UC Merced land use dataset show that our method successfully outperforms other individual methods on remote sensing classification tasks and outperforms other models in terms of training efficiency.

However, there are some limitations to this study, firstly the choice of a single evaluation criterion may have some evaluation error and the choice of a single dataset may limit the ability to generalise the results. In the future, we will test our method b on other benchmark datasets to evaluate its performance in remote sensing classification tasks.

REFERENCES

[1] P. G. , et al., "Finer resolution observation and monitoring of global land cover: first mapping results with landsat tm and etm+ data," *International Journal of Remote Sensing*, vol. 34, no. 7, pp. 2607–2654, 2013.

[2] H. Zhu, X. Wang, and R. Chen, "Deep learning algorithm based remote sensing image classification research.," *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Electronic Technology, Communication and Information (ICETCI), 2023 IEEE 3rd International Conference on*, pp. 1368 – 1373, 2023.

[3] X. Chen and C. Lu, "An end-to-end adversarial hashing method for unsupervised multispectral remote sensing image retrieval.," *2020 IEEE International Conference on Image Processing (ICIP), Image Processing (ICIP), 2020 IEEE International Conference on*, pp. 1536 – 1540, 2020.

[4] S. Chen, S. Zhong, B. Xue, X. Li, L. Zhao, and C. Chang, "Iterative scale-invariant feature transform for remote sensing image registration.," *IEEE Transactions on Geoscience and Remote Sensing, Geoscience and Remote Sensing, IEEE Transactions on, IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 4, pp. 3244 – 3265, 2021.

[5] X. Chen, G. Zhu, and M. Liu, "Bag-of-visual-words scene classifier for remote sensing image based on region covariance.," *IEEE Geoscience and Remote Sensing Letters, Geoscience and Remote Sensing Letters, IEEE, IEEE Geosci. Remote Sensing Lett*, vol. 19, pp. 1 – 5, 2022.

[6] *Techniques and Applications of UAV-Based Photogrammetric 3D Mapping.* 2022.

[7] J. Lipeng, H. Xiaohui, and W. Mingye, "Saliency preprocessing locality-constrained linear coding for remote sensing scene classification.," *Electronics*, vol. 7, no. 9, p. 169, 2018.

[8] Z. Xiong, X. Chen, J. Luo, C. Shen, and Z. Xu, "scsagan: A scrna-seq data imputation method based on semi-supervised learning and probabilistic latent semantic analysis.," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Bioinformatics and Biomedicine (BIBM), 2022 IEEE International Conference on*, pp. 178 – 181, 2022.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[11] N. Mishra, I. Jahan, M. R. Nadeem, and V. Sharma, "A comparative study of resnet50, efficientnetb7, inceptionv3, vgg16 models in crop and weed classification.," *2023 4th International Conference on Intelligent Engineering and Management (ICIEM), Intelligent Engineering and Management (ICIEM), 2023 4th International Conference on*, pp. 1 – 5, 2023.

[12] B. Kumar, A. K. Singh, and P. Banerjee, "A deep learning approach for product recommendation using resnet-50 cnn model.," *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Sustainable Computing and Smart Systems (ICSCSS), 2023 International Conference on*, pp. 604 – 610, 2023.

[13] E. Landi, F. Spinelli, M. Intravaia, M. Mugnaini, A. Fort, M. Bianchini, B. T. Corradini, F. Scarselli, and M. Tanfoni, "A mobilenet neural network model for fault diagnosis in roller bearings.," *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Instrumentation and Measurement Technology Conference (I2MTC), 2023 IEEE International*, pp. 01 – 06, 2023.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[16] Q. Yang, *Transfer learning.* Cambridge University Press, 2020.

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