Graphite Ore Grade Classification Algorithm Based on Multi-scale Fused Image Features

Jionghui Wang[1(✉)], Yaokun Liu[2], Xueyu Huang[2,3], and Shaopeng Chang[2]

[1] Minmetals Exploration & Development Co. Ltd., Beijing 100010, People's Republic of China
wangjh@minmetals.com
[2] School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330013, People's Republic of China
[3] Nanchang Key Laboratory of Virtual Digital Factory and Cultural Communications, Nanchang 330013, People's Republic of China

Abstract. Aiming at the problems of complex pre-processing and expensive equipment in chemical detection of graphite ore grade, a graphite ore identification and classification method based on fusing multi-scale image features is proposed. In the feature extraction stage, a deep convolutional neural network and a residual network model based on spatial attention mechanism are constructed to improve the learning ability of local and global features of graphite ore images; in the feature aggregation stage, a global response normalization technique is introduced to achieve more accurate graphite ore grade recognition, and the accuracy of the model reaches 93.401% and the macro F1 reaches 93.086%, which is better than the single The accuracy of the model reaches 93.401% and the macro F1 reaches 93.086%, which is better than the traditional machine learning methods with single feature. The experimental results show that the features extracted by different methods can describe the texture and edge information of graphite ore, and the proposed method has better extraction ability in terms of local features and global features of graphite ore images, and achieves more accurate graphite ore grade recognition with good robustness.

Keywords: feature aggregation · texture features · depthwise convolution · residual network · attention mechanism · global response normalization (GRN) · graphite ore

1 Introduction

Timely prediction of ore grade is crucial in the mining and processing of graphite ore. It not only improves production efficiency but also helps enterprises save costs in subsequent intelligent blending and scientific ore selection processes. Currently, two traditional methods are commonly used to determine the grade of graphite ore: sulfur-carbon

This work is supported by the National Key Research and Development Program of China 2020YFC1909602.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
C. Wu et al. (Eds.): MONAMI 2023, LNICST 559, pp. 178–198, 2024.
https://doi.org/10.1007/978-3-031-55471-1_14

analysis and high-frequency infrared methods. However, these methods have complex pre-processing requirements, expensive equipment, and are difficult to operate in mining environments. Moreover, delays in determining the grade of graphite ore can occur when faced with a heavy workload. Resolving this time delay issue is crucial for improving production efficiency and aligns with the trend of industrial production becoming more intelligent. Therefore, constructing an automated ore identification model is beneficial for unleashing productivity and assisting industrial enterprises in enhancing the accuracy and efficiency of classification, thereby holding significant practical implications.

In recent years, computer vision techniques have gained significant attention in the industrial field and mineral classification due to the advancements in deep learning and image processing technologies. Su et al. [1] optimized the network structure of LeNet-5, one of the earliest convolutional neural network-based image classification algorithms. They trained the model on a dataset of 20,000 images to successfully perform binary classification of coal and gangue. The experimental results exhibited an impressive accuracy of 95.88% on the validation set. For smaller datasets comprising 240 images each for coal and gangue, Pu et al. [2] utilized transfer learning by freezing the convolutional layers of VGG16 and customizing the fully connected layers. Their trained model achieved a classification accuracy of 82.5% on the test set. Wang et al. adopted the Wu-VGG19 transfer learning network structure to accomplish binary classification of surrounding rocks and black tungsten ore, resulting in an outstanding recognition rate of 97.51%. These findings highlight the effectiveness of utilizing advanced deep learning techniques for accurate and efficient mineral classification in various scenarios.

Efficiently identifying the grade of graphite ore through image recognition, by building a relationship model between mineral texture features, color features, and mineral types, is a valuable research topic. Deep learning has contributed to significant advancements in various fields, including computer vision, in recent years. In the domain of mineral image classification and recognition, researchers from both domestic and international backgrounds have gradually incorporated deep learning techniques. Notably, Zhang et al. [3] successfully employed the Inception-v3 network to intelligently classify granite, quartz diorite, and gabbro. Similarly, Baraboshkin et al. and Bai et al. utilized the same network to classify 5 and 7 different types of minerals, respectively. Li et al. [4] further developed a comprehensive intelligent coupled classification method for minerals using Inception-v3. Their approach effectively differentiated 19 distinct minerals by leveraging texture and color features derived from mineral images through K-means. Building on this foundation, Liu et al. [5] validated the effectiveness of combining deep learning with clustering algorithms. Zeng et al. employed a two-layer fully connected neural network to enhance the scalar Mohs hardness. They utilized EfficientNet-b4 [6] for extracting mineral image features and successfully achieved the classification of 36 different types of minerals by integrating the results and feeding them into a fully connected layer. To address overfitting, Liang et al. [7] employed CutMix and image cutting as data augmentation methods. Notably, they pioneered the use of ViT, an evolution of Transformer, for classifying 7 different types of minerals.

Expanding problem-solving approaches and flexibly applying deep learning techniques are valuable for mineral type recognition tasks. These tasks extend beyond natural

scene images and can incorporate additional data types, such as microscopic images [8–11] and spectral images [12]. Polarized microscopic images extract fine-grained features of minerals. Iglesias et al. [8] utilized the ResNet18 model to classify polarized microscopic images of five minerals: biotite, quartz, garnet, muscovite, and olivine, achieving an accuracy of 89%. Spectral images also play a significant role in mineral classification tasks. Han et al. [13] acquired spectral images of minerals using a visible-infrared reflectance spectrometer and trained a custom hollow convolutional neural network with these images. This approach successfully classified hematite, magnetite, granite, quartz diorite, and greenstone.

Comparing neural network characteristics, developing deep learning models suitable for mineral type classification, and optimizing conventional algorithms are important topics among researchers in the field. One technique of interest is the dual-task processing capability of object detection networks, which effectively classify multiple mineral blocks within the same image. In a study referenced as [14], an object detection dataset with around 800 mineral samples was constructed. Faster R-CNN was trained on this dataset to classify eight types of minerals, including olivine, basalt, marble, slate, conglomerate, limestone, granite, and magnetite quartzite. In contrast, another study referenced as [15] revealed that Faster R-CNN outperformed YOLOv4 [16] in classifying three categories (volcanic rock, sedimentary rock, and metamorphic rock) and 32 subcategories of minerals. Additionally, in a study mentioned as [17], Faster R-CNN was optimized using multiscale feature fusion techniques and the particle swarm optimization algorithm, achieving an impressive 98% classification accuracy for minerals such as biotite, hematite, turquoise, and quartz.

Additionally, semantic segmentation networks have the capability to classify multiple mineral blocks within the same image at the pixel level. In Reference [18], an improved U-Net was utilized to segment minerals in images, resulting in the classification of red rocks, green hematite, yellow siderite, blue greenstone, and purple pyrite. In Reference, the instance segmentation network Mask RCNN [19] was employed, which combines the functionalities of object detection and semantic segmentation. This approach achieved a comprehensive accuracy of 97.6% in both mineral identification and localization.

In the classification of graphite ore images, the texture features and global image correlations present in the ore images receive more attention compared to other image classification tasks. Additionally, the recognition of ore itself is highly sensitive to positional information. Traditional convolutional neural networks (CNNs) gradually capture image characteristics by extracting features through multiple convolutional layers [20]. CNNs inherently possess a strong inductive bias due to their design characteristics of locality and weight sharing mechanisms. Moreover, CNNs demonstrate sample and parameter efficiency due to their translational equivariance properties. On the other hand, Visual Transformers excel in modeling long-range dependencies through self-attention mechanisms, making them dominant in natural language processing (NLP) research. Transformers have recently been successfully applied to various computer vision tasks, showcasing impressive performance. Consequently, some researchers have explored the direct incorporation of convolutional operations into visual Transformers to introduce the

inductive bias. However, forcibly modifying the structure may compromise the integrity of the Transformer and reduce model capacity.

To address these issues, this paper presents a novel pretraining-driven approach for graphite ore classification and recognition, utilizing multiscale image feature fusion. In the feature extraction stage, deep convolutional layers are employed to extract local texture features of the image in the channel dimension. Additionally, a spatial attention mechanism is introduced to extract global contextual feature information. The two sets of extracted features are then normalized and fused together in a cascaded manner. The resulting fused features provide a comprehensive description of the texture information present in graphite ore images, encompassing both channel and spatial dimensions. This progressive fusion approach enables the model to capture the correlations between different spatial regions and channels, effectively guiding the network to focus on the target region. As a result, the recognition accuracy of graphite ore images is significantly improved.

The main contributions of this paper are as follows:

(1) In the classification of graphite ore images, transfer learning methods are employed to pre-train the backbone network using existing large-scale publicly available image datasets. This approach ensures full optimization of the model parameters and effectively addresses the problem of limited training data.
(2) To enhance the recognition and classification of graphite ore images, this paper proposes a residual network structure with a spatial attention mechanism. This novel approach improves the model's capability to learn long-range dependencies within the images.
(3) To enhance the accuracy of recognition and classification tasks, we propose the addition of a global response normalization layer to each convolutional neural network module. This layer normalizes the features in the channel dimension, allowing for better control of feature proportions during training. As a result, the model exhibits improved generalization and performance.

In the second section of this paper, we present the overall structure and specific details of the graphite ore recognition and classification method that integrates multiscale image features. The third section covers the experimental process and results. By comparing the macro precision, macro recall, macro F1 score, accuracy, and confusion matrix with other methods, we analyze the performance of the proposed method. Finally, in the fourth section, we provide the main conclusions of this article.

2 Graphite Ore Image Classification and Recognition Method by Fusing Multi-scale Features

This paper proposes a graphite ore image classification and recognition method based on fused multiscale features, firstly, the input image is preprocessed by better initialization improvement; secondly, a deep convolutional layer performs feature extraction on the incoming preprocessed data, and the deep convolutional will run on a per-channel basis and mix the information of spatial dimensions; after that, an independent downsampling layer integrates the data; finally, the fused feature data is global level pooling to obtain a

highly informative feature subset, and complete the classification by the classifier. The overall model has strong generalization ability, high accuracy and robustness. To achieve classification and recognition of different types of graphite ores. The framework of the classification method is shown in Fig. 1 below.

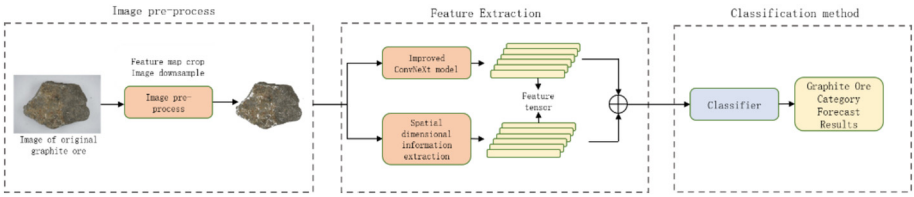

Fig. 1. Graphite ore image classification method framework based on multi-scale feature fusion

2.1 Migration Learning and Graphite Ore Image Data Situation

Image Pre-training. In addition to designing the model architecture, the effective utilization of large-scale datasets plays a crucial role in training an excellent network model with high accuracy and strong robustness. However, the availability of publicly accessible data for graphite ore image classification is limited. On the other hand, ImageNet, a comprehensive dataset containing diverse images with rich colors and textures, is extensively employed for image classification tasks. Transfer learning, widely adopted in the field of deep learning, has become a conventional strategy for classification tasks. Convolutional neural networks have achieved remarkable advancements in recognition accuracy, with the emergence of deep models such as ResNet, EfficientNet, and PatchConvnet, which comprise millions or even tens of millions of parameters. To address the challenge of limited training data, this paper proposes an enhanced ConvNeXt model that pretrains parameters on ImageNet and initializes the backbone network. This approach optimizes model parameters and alleviates the adverse impact of insufficient training data.

Experimental Dataset Description. The availability of publicly accessible graphite ore image data is limited. To address this, rock samples were meticulously collected by professional personnel to obtain comprehensive image data, which were then carefully labeled. Furthermore, highly accurate chemical methods were employed to determine the carbon content in the ore, and these results were annotated using a carbon-sulfur analyzer. Consequently, a dataset of graphite ore images was constructed. There are four categories of graphite ores included in the dataset, from low to high according to the carbon grade values contained in the minerals: waste rock (for comparison reference only, not involved in training 0%~1%) 1126 sheets of low grade ores (1%~5%), 2170 sheets of lower grade ores (6%~10%), 3476 sheets of higher grade ores (11%~15%) and 1567 sheets of high grade ores (16%~20%), as shown in Fig. 2.

In order to mitigate the impact of image quality on recognition performance, we employed professional equipment to capture high-quality images with a resolution of 6000×4000 pixels during the data collection process. To ensure optimal feature learning

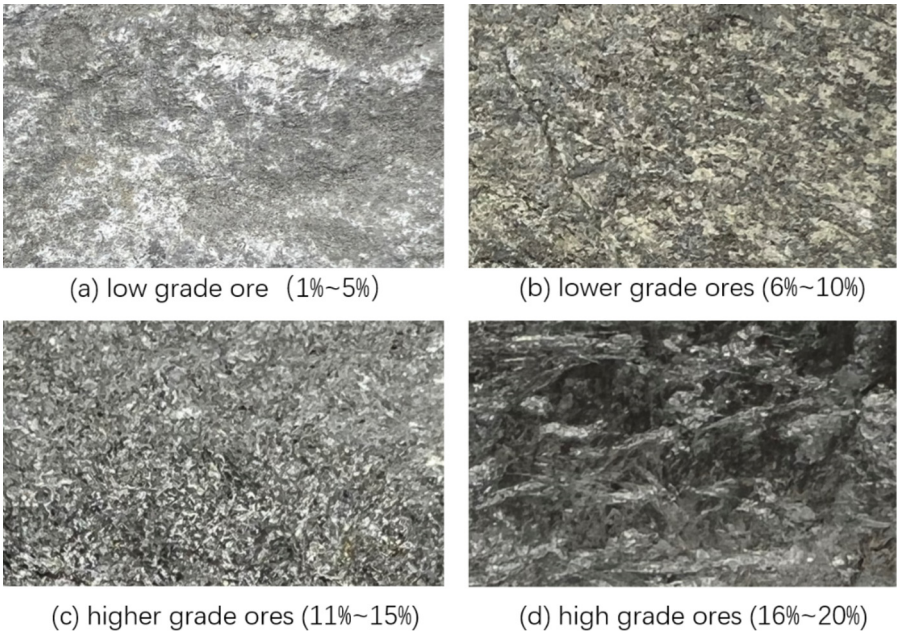

Fig. 2. Image samples of ore with different grades in the dataset

during training and reliable evaluation, the image data for each category was randomly divided into training, validation, and test sets using a ratio of 6:2:2. The table below illustrates the categories of graphite ore rocks as an example (Table 1).

Table 1. Distribution of graphite ore image dataset after preprocessing

Data set type	Different grades of graphite ore			
	1%~5%	6%~10%	11%~15%	16%~20%
train set	676	1302	2086	941
validation set	225	434	695	313
test set	225	434	695	313

Preprocessing the Dataset. In convolutional neural networks, the stem layer plays a crucial role in processing input images at the network's outset. In natural environments, redundant information is present, and the stem layer efficiently downsamples input images to meet the feature map size requirements of standard convolutional networks and visual Transformers. In the standard ResNet, the stem layer comprises a 7 × 7 convolutional layer with a stride of 2 and a MaxPooling layer, resulting in a 4x downsampling of input images. Visual Transformers employ a more direct "Patchify" strategy in their stem layer, utilizing larger kernel sizes and non-overlapping convolutions. To

accommodate the multi-stage design of the Swin Transformer, a similar "Patchify" layer with smaller convolutions is employed. Inspired by visual Transformers, we replace the stem layer in ResNet with a Patchify layer implemented using a 4×4 convolutional layer with a stride of 1. This modification yields a slight improvement in accuracy while reducing computational complexity. Experimental results affirm the positive impact of enhanced initialization on graphite ore image classification tasks.

2.2 Graphite Ore Image Feature Extraction

During the feature extraction stage, we addressed the challenge of capturing long-range dependencies in global features extracted from graphite ore images by enhancing the ResNet-50 model. Inspired by the capabilities of visual Transformers in modeling long-range dependencies, we developed a hierarchical convolutional neural network similar to Swin Transformer. Figure 3 illustrates the architecture. Initially, we trained the base model using a training strategy similar to that of visual Transformers, resulting in improved performance compared to the original ResNet-50. Subsequently, we made several enhancements, including: 1) improved initialization, 2) deeper convolutions, 3) preactivation bottleneck blocks, and 4) normalization and activation functions. In the following sections, we will delve into these enhancements made to the convolutional neural network for graphite ore image classification. Given the close relationship between network complexity and model performance, we carefully controlled the size of floating-point operations (FLOPs) during the enhancement process.

Feature Extraction Based on Deep Convolution in the Channel Dimension. The texture features of graphite ore images exhibit a random spatial distribution on the image, and they are typically composed of multiple layers and angles of textures, resulting in high complexity. Conventional convolutional networks often lack sufficient attention to the same region and fail to achieve sufficient depth in feature extraction.In this part, we attempt to draw inspiration from the design principles of ResNeXt [21]. Compared to conventional ResNet, ResNeXt offers better depth in feature extraction and can achieve higher accuracy. Its core component is the grouped convolution, where the convolutional filters are divided into different groups. From a macro perspective, ResNeXt aims to increase the number of groups to expand the convolutional width. This is reflected in the specific ResNeXt module, where the 3×3 convolutional layer adopts the grouped convolution method.By incorporating the ResNeXt design, we can effectively capture the complex texture features of graphite ore images. The grouped convolution allows for the exploration of different perspectives and variations within the texture, leading to more comprehensive feature representations. The experimental results demonstrate the effectiveness of this approach in improving accuracy and enhancing the model's ability to capture fine-grained texture information.

Drawing inspiration from the idea of grouped convolution, in our structural design, we employ a form of deep grouped convolution, as illustrated in Fig. 4. It is achieved by dividing the channels into equal-sized groups. The deep convolution applies a separate convolutional kernel to each input channel (input depth), followed by a simple 1 × 1 convolutional layer for pointwise convolution to create a linear combination of the outputs from the deep convolutional layer.

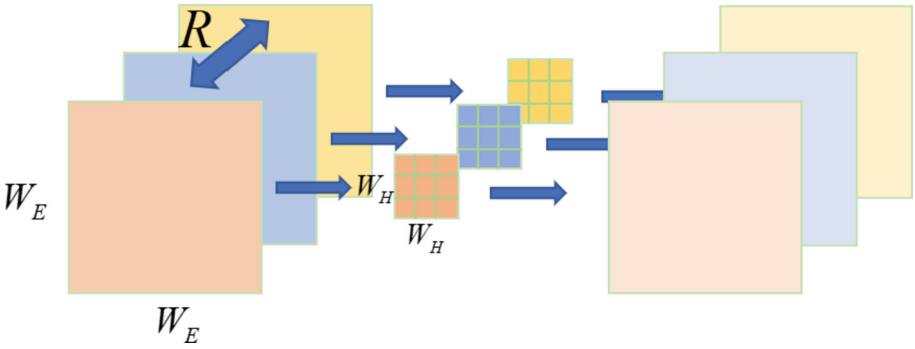

Fig. 3. Depthwise convolution with separate kernels applied

The formula expression for each input channel in the deep convolution is as follows:

$$\hat{Z}_{k,l,p} = \sum_{i,j} \hat{H}_{i,j,q} \cdot E_{k+i-1,l+j-1,p} \tag{1}$$

In the above formula, $\hat{H}$ is a depth convolutional kernel with a size of $W_H \cdot W_H \cdot R$. Each of the p th convolutional kernels in $\hat{H}$ is applied to the p th channel of the feature map E, resulting in the p th channel of the output feature map $\hat{Z}$ corresponding to that convolutional kernel.

The computational cost of deep convolution is:

$$C_{H1} = W_H \cdot W_H \cdot R \cdot W_E \cdot W_E \tag{2}$$

The computational cost in the above formula depends on the product of the number of input channels M, the size of the convolutional kernel $W_H \cdot W_H$, and the size of the feature map $W_E \cdot W_E$.

Compared to standard convolution operations, depth-wise convolution is highly effective in improving overall performance. However, it only extracts features from input data without combining the learned features to create new ones. Therefore, an additional layer is needed to combine the output of depth-wise convolution through a 1 × 1 convolution, resulting in the generation of new features. At this point, the computational cost of depth-wise convolution is given by:

$$C_{H2} = W_H \cdot W_H \cdot R \cdot W_E \cdot W_E + R \cdot U \cdot W_E \cdot W_E \tag{3}$$

where U is the number of output channels, the computational cost of standard convolution is given by:

$$C_n = W_H \cdot W_H \cdot R \cdot U \cdot W_E \cdot W_E \tag{4}$$

By decomposing the standard convolution into feature extraction and feature fusion parts, the overall computational cost is reduced from the formula (4) of standard convolution to formula (3). Therefore, the total computational cost is approximately one-eighth of that of a standard convolution.

$$\frac{C_{H2}}{C_n} = \frac{W_H \cdot W_H \cdot R \cdot W_E \cdot W_E + R \cdot U \cdot W_E \cdot W_E}{W_H \cdot W_H \cdot R \cdot U \cdot W_E \cdot W_E} = \frac{1}{U} + \frac{1}{W_H^2} \tag{5}$$

Depthwise convolution has been popularized by MobileNet and Xception. It can be observed that operating on each channel is a characteristic of weighted operations in self-attention. This similarity to depthwise convolution lies in the fact that each convolution kernel individually processes a channel, just like the self-attention mechanism that performs spatial information fusion and weighting within a single channel. In this regard, we refer to the design logic of Swin Transformer and increase the spatial network width from 64 to 96.

Spatial Attention Mechanism. In convolutional neural networks, the spatial attention mechanism enhances the network's ability to extract features by selectively focusing on specific spatial positions of the input [22]. This mechanism allows the net-work to allocate its attention resources more effectively, leading to improved feature extraction performance.

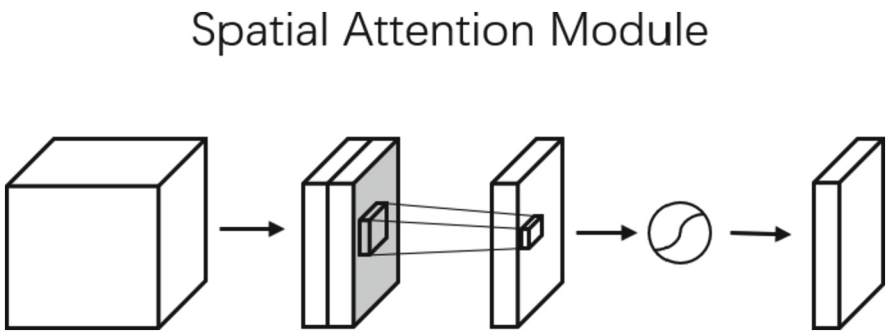

Fig. 4. Spatial Attention Module

In the spatial attention mechanism, a 2D attention weight matrix is commonly employed to indicate the importance of different positions. Each element of this matrix corresponds to a specific position in the input feature map, reflecting its significance.

During each convolutional operation, the neural network calculates a weighted average of the input feature map, with the weight for each position determined by the attention weight matrix. This enables the neural network to assign distinct weights to various spatial positions, facilitating more effective feature extraction.

Bottleneck Pre-activation. In standard ResNet, the bottleneck structure is employed as (large-dimension-small-dimension-large-dimension) to minimize computational complexity. Subsequently, the inverted bottleneck structure was introduced in Mo-bileNetV2, which follows the pattern of (small-dimension-large-dimension-small-dimension). This design enables seamless information transfer between different-dimensional feature spaces, avoiding information loss caused by dimension com-pression. A similar structure was also adopted in the MLP (Multi-Layer Perceptron) of Transformers, where the dimensionality of the middle layer and the fully connected layer is four times that of the two endpoints (Fig. 5).

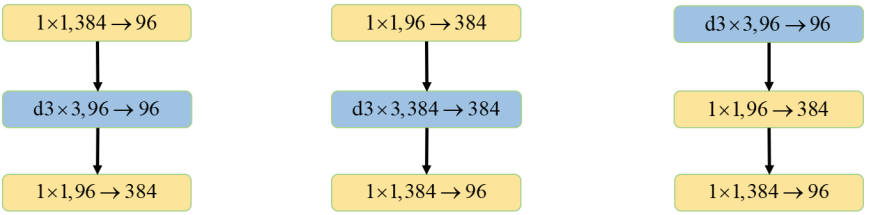

Fig. 5. Inversion and front improvement based on ResNeXt

The design order of convolutional blocks plays a crucial role in the sequence of targeted feature maps. In the context of mineral rock classification, it is advantageous to place the bottleneck layer in the front as it enables earlier learning of feature maps by the convolutional layers. Inspired by a significant design aspect of each Transformer block, an inverted residual bottleneck module was developed. In this module, the hidden dimension of the MLP block is expanded by a factor of four compared to the input dimension. Remarkably, this Transformer design exhibits a connection to the inverted residual bottleneck module utilized in ConvNets with an expansion rate of 4. This concept gained prominence through MobileNetV2 and has since been further adopted in various advanced convolutional architectures.

Therefore, in networks utilizing the inverted residual module, the FLOPs (floating-point operations) of deep convolutional layers demonstrate a notable increase. However, this is offset by a substantial reduction in FLOPs within the shortcut layers of subsequent downsampling residual blocks. Consequently, the overall network experiences a slight decrease in FLOPs while concurrently achieving improved performance.

Activation Function and Normalization Layer Optimization. To achieve performance on par with Vision Transformers in graphite ore image classification, we leverage the advantages of microarchitecture design. An important consideration that distinguishes natural language processing from vision architectures is the choice of activation

function. In the field of computer vision, several activation functions have been proposed. While Rectified Linear Unit (ReLU) has been widely used in convolu-tional neural networks due to its simplicity and efficiency, recent models have in-creasingly adopted Gaussian Error Linear Unit (GELU) as the activation function. Notably, GELU has been employed in models such as BERT (by Google), GPT-2 (by OpenAI), and Transformers. GELU can be viewed as a smooth variant of ReLU. In this context, for the sake of alignment with other metrics, we have selected GELU as our activation function.

In standard ResNet, downsampling of spatial dimensions is typically achieved through 3×3 convolutions with a stride of 2. For convolutional blocks with residual connections, downsampling is performed in the shortcut connection using 1×1 convolutions with a stride of 2. This ensures that the downsampling layers in the CNN maintain a similar computational strategy as the other layers. However, Swin Transformer introduces an additional dedicated downsampling layer between stages. To align with the design principles of Swin Transformer, we experimented with a different approach, employing a separate 2×2 convolutional layer with a stride of 2 for spatial downsampling. Furthermore, to ensure stable training during changes in spatial resolution, we incorporated a normalization layer. The commonly used normalization layers in neural networks include Local Response Normalization (LRN), Batch Normalization (BN), Layer Normalization (LN), and Global Response Normalization (GRN). After comparing these options, we ultimately selected GRN normalization [23] due to its superior effectiveness.

This study utilizes ResNet-50 as the main network for extracting features from graphite ore images. The parameters of ResNet-50 are initialized through pretraining techniques. To capture long-range spatial context dependencies during the feature extraction stage in the convolutional layers, we incorporate the design logic of Swin Transformer. Additionally, the output of the local-global features is further normalized and serves as the output of the convolutional operation, leading to the generation of a multi-scale feature map.

The ConvNeXt-G model comprises five stages. The initial stage, known as "stem," employs a single convolutional layer for input data preprocessing. Subsequently, four bottleneck pre-activated deep convolutional layers follow, characterized by progressively increasing parameters. To ensure stable feature extraction within the convolutional layers, Layer Normalization (LN) is incorporated into each downsampling layer, enhancing training stability. Additionally, the two 1×1 convolutions within each block are implemented using fully connected layers, offering a slight speed advantage over convolutional layers. Normalization layers are applied before downsampling, after the stem, and after the Global Average Pooling (GAP) layer. Finally, the hyperparameters related to data augmentation, preprocessing, and optimizer are harmonized to achieve the enhanced network structure of ConvNeXt-G (Figs. 6 and 7, Table 2).

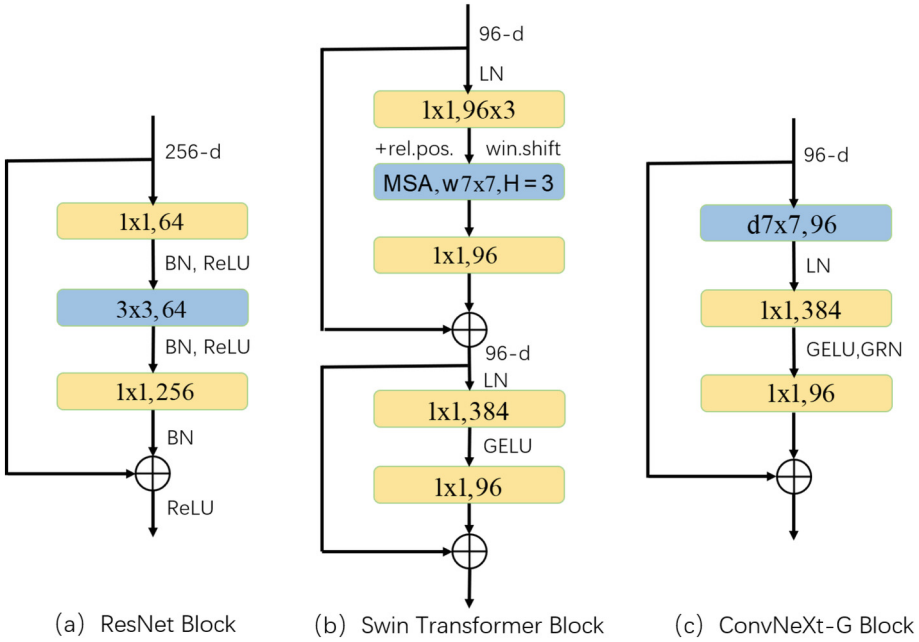

Fig. 6. Comparison between ConvNeXt-G based on ResNet fusion Swin Transformer design logic and the former two

Table 2. Swin Transformer, ResNet-50 and ConvNeXt-G network parameters

	output size	Swin-T	ResNet-50	ConvNeXt-G
Stem	*56 × 56*	*4 × 4, 96, stride 4*	*7 × 7, 64, stride 2*	*4 × 4, 96, stride 4*
Stage1	*56 × 56*	$\begin{bmatrix} 1\times1, 96\times3 \\ \text{MSA, w7}\times7, \text{H}=3, \text{rel.pos.} \\ 1\times1, 96 \end{bmatrix}\times2$ $\begin{bmatrix} 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} \text{d}7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}\times3$
Stage2	*28 × 28*	$\begin{bmatrix} 1\times1, 192\times3 \\ \text{MSA, w}7\times7, \text{H}=6, \text{rel.pos.} \\ 1\times1, 192 \end{bmatrix}\times2$ $\begin{bmatrix} 1\times1, 768 \\ 1\times1, 192 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} \text{d}7\times7, 192 \\ 1\times1, 768 \\ 1\times1, 192 \end{bmatrix}\times3$

(*continued*)

Table 2. (*continued*)

	output size	Swin-T	ResNet-50	ConvNeXt-G
Stage3	*14 × 14*	$\begin{bmatrix} 1\times1, 384\times3 \\ \text{MSA, w7}\times7, \text{H}=12, \text{rel.pos.} \\ 1\times1, 384 \end{bmatrix} \times 6 \begin{bmatrix} 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{d7}\times7, 384 \\ 1\times1, 1536 \\ 1\times1, 384 \end{bmatrix} \times 9$
Stage4	*7 × 7*	$\begin{bmatrix} 1\times1, 768\times3 \\ \text{MSA, w7}\times7, \text{H}=24, \text{rel.pos.} \\ 1\times1, 768 \end{bmatrix} \times 2 \begin{bmatrix} 1\times1, 3072 \\ 1\times1, 768 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{d7}\times7, 768 \\ 1\times1, 3072 \\ 1\times1, 768 \end{bmatrix} \times 3$
FLOPs		4.5×10^9	4.1×10^9	4.5×10^9
#params.		28.3×10^6	25.6×10^6	28.6×10^6

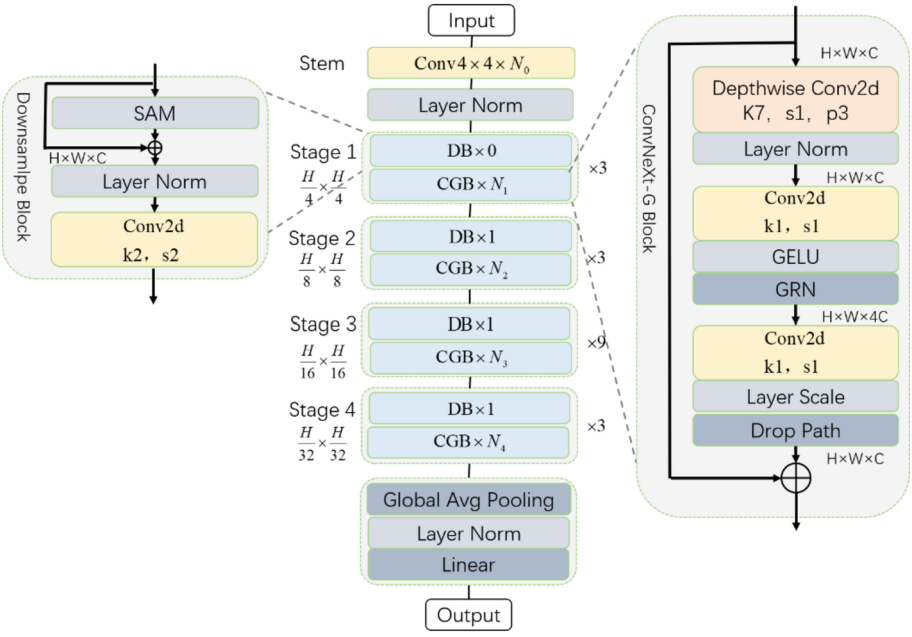

Fig. 7. ConvNeXt-G network model overall structure and module details

3 Experimental Analysis

3.1 Experimental Setup

The hardware environment used in this study includes: GPU: RTX 3080*1 with 10 GB VRAM, CPU: 12-core Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz. The software environment consists of Python 3.8, PyTorch 1.8, Torchvision 0.9, CUDA 11.1, and others.

3.2 Evaluation Metrics

The implemented task in this paper is graphite ore image classification, which is a multi-class problem. Common evaluation metrics used in image classification tasks are employed to assess the algorithm's performance. These metrics include confusion matrix, macro-precision, macro-recall, macro-F1 score, and accuracy.

3.3 Network Model Effectiveness Ablation Experiment

This study primarily conducted ablation experiments on a dataset of graphite ore images collected by professionals working in graphite mines. The purpose was to validate the effectiveness of the pretraining strategy and various module components used in this study. We evaluated the contribution and detection capabilities of each module by constructing different combinations of modules. The specific combinations are described as follows:

Baseline: The baseline model in this study consists of a ResNet50 backbone network based on the RPG image modality and an image recognition classification module. The main component is a mask generation block with a sequential structure of Conv-ReLU-Conv-Sigmoid. Additionally, this model does not include any pretraining strategy.

Baseline + P: This combination builds upon the baseline model by incorporating a ResNet50 backbone encoding network that is pretrained (P) on the ImageNet dataset. By replacing the traditional random initialization approach with a pretrained model, it allows for better learning of image feature information.

Baseline + P + PAT: This combination builds upon the Baseline + P model by introducing an improved initialization method. It replaces the downsampling layer (stem) after the original input layer with a patchify layer (PAT), which ensures non-overlapping sliding windows and processes information from one patch window at a time. This approach allows for better learning and capturing of image feature information from a more optimal initialization perspective.

Baseline + P + PAT + DEPTH: This combination extends the Baseline + P + PAT model by introducing the concept of grouped convolution. It replaces the original residual convolutional layers with depthwise convolution (DEPTH) layers. By increasing the base channel number and performing operations on a per-channel basis, this approach enhances the computational speed of the model while further extracting spatial context dependencies.

Baseline + P + PAT + DEPTH + IB: This combination builds upon the Baseline + P + PAT + DEPTH model by repositioning the bottleneck layer within the convolutional

feature extraction block (Inverted Bottleneck, IB). By placing the bottleneck layer earlier in the feature extraction stage, the model is enhanced to focus more on capturing long-range dependencies in the graphite ore images.

Baseline + P + PAT + DEPTH + IB + SAM: This combination extends the Baseline + P + PAT + DEPTH + IB model by incorporating micro-level adjustments to the architecture for the task of graphite ore image classification. Firstly, a spatial attention mechanism (SAM) is introduced in the downsampling layers. This mechanism pools the feature maps along the channel dimension, compressing the channel size and facilitating the learning of spatial features. The compressed feature maps are then concatenated along the channel dimension and subjected to convolutional operations and activation functions. Additionally, the more advanced and stable GELU activation function is used instead of ReLU to enhance the overall robustness of the model.

Baseline + P + PAT + DEPTH + IB + SAM + GRN: This combination represents the proposed network architecture in this paper, designed for the task of graphite ore image classification. Firstly, a Global Response Normalization (GRN) is introduced between the two 1×1 convolutional layers in the feature extraction module. This includes three steps: global feature aggregation, feature normalization, and feature recalibration, which enhances the contrast and selectivity among channels. Additionally, a separate downsampling layer is added between each stage. The purpose of this layer is to apply Layer Normalization (LN) to normalize all data in the samples, effectively alleviating the issue of feature collapse and preserving the diversity of graphite ore image features during the propagation of the network layers (Table 3).

3.4 Quantitative Experimental Results Compared with Other Methods Are Presented in This Section

This subsection provides a comparative analysis of popular convolutional neural network architectures for graphite ore image classification. The evaluated models include the ResNet series (ResNet-50, ResNet-101, ResNet-152), the EfficientNet series (EfficientNet-B0, EfficientNet-B2), and the Vision Transformer series (ViT-S, ViT-B, ViT-L). Their performance in the task of graphite ore image classification and recognition is analyzed.

During model training, a consistent set of hyperparameters was employed to ensure a fair and unbiased comparison among different convolutional neural network experiments. Images underwent preprocessing, with their resolution adjusted to 224×224 before being fed into the network. The participating convolutional neural networks were initialized with an initial learning rate of 5e-5, utilizing AdamW as the chosen optimizer. Each epoch involved a batch size of 512, indicating that 512 images were used for training in each iteration. The maximum number of epochs was set to 200. To decay the learning rate, a cosine decay learning rate schedule was implemented. This strategy gradually reduces the learning rate as the model approaches the global minimum of the loss function, facilitating closer convergence to this point. The cosine decay training method follows a pattern of initially slow descent in the cosine function value, followed by accelerated descent, and finally, decelerated descent. Hence, cosine decay was chosen to effectively reduce the learning rate. Figure 10 depicts the distribution of the confusion matrix on the test set, while Fig. 8 displays the accuracy curve of the four models on the

Table 3. Comparison results of ablation experiments of each module combination on the graphite ore dataset

Model Variants	Modules						Accuracy (%)	Convergent Epoch
	Pre-trained	PAT	DEPTH	IB	SAM	GRN		
Baseline							73.965	203
Baseline + P	√						86.683	179
Baseline + P + PAT	√	√					86.803	172
Baseline + P + PAT + DEPTH	√	√	√				91.542	174
Baseline + P + PAT + DEPTH + IB	√	√	√	√			91.662	171
Baseline + P + PAT + DEPTH + IB + SAM	√	√	√	√	√		92.442	145
Baseline + P + PAT + DEPTH + IB + SD + GRN	√	√	√	√	√	√	93.401	136

validation set. Furthermore, Fig. 9 illustrates the loss curve in relation to the number of epochs.

Based on Fig. 8, the ResNet-101 model, a benchmark in the ResNet series of convolutional neural networks, achieves convergence at approximately 180 epochs. Following convergence, the accuracy remains steady around 88%. The Vision Transformer, originally developed for natural language processing and later adapted for image classification, demonstrates convergence within roughly 60 epochs, aided by transfer learning, and maintains an accuracy of around 88.6%. Despite its greater computational complexity and resource requirements, the EfficientNet-B3 model from the acclaimed EfficientNet series converges in 120 epochs, with the final accuracy fluctuating around 92.8%. In contrast, the ConvNeXt-G model reaches convergence at approximately 160 epochs, positioning it between the ResNet and Vision Transformer models, with an accuracy fluctuating around 93.4%. These findings highlight the similar performance of these four models in the multi-classification recognition of graphite mineral images.

Based on the comparison of loss values in Fig. 9, several key findings emerge. When using the same batch size, the Vision Transformer demonstrates slower convergence within the first 50 epochs compared to neural network models with residual connections. Its loss value fluctuates between 0.7 and 1.5, gradually stabilizing around 1.25 after 60 epochs. In contrast, ResNet-101 exhibits faster convergence, with the loss value approaching stability after 50 epochs and fluctuating around 0.8. However, larger fluctuations occur around 80, 120, and 160 epochs, eventually converging to approximately 0.8.ConvNeXt-G, a hybrid model combining ResNet-50 and Vision Transformer concepts, shows a gradual reduction in loss values from an initial 2.5. Due to appropriate initialization, the loss value stabilizes after approximately 45 epochs, gradually converging around 0.7. EfficientNet demonstrates a gradual decrease in loss values from an initial 3.5. By the 40th epoch, the loss value reaches 1.0 and further diminishes to around 0.6 at the 120th epoch, where convergence is achieved.In summary, these results indicate that ConvNeXt-G achieves significantly faster convergence compared to Vision Transformer and deeper EfficientNet models, while performing at a similar speed to ResNet models utilizing residual connections.

Based on the confusion matrix distribution shown in Fig. 10, the attention-based improved convolutional neural network model demonstrates high accuracy in recognizing mineral images when the actual grade distribution is within the ranges of 6%~10% and 11%~15%. By integrating spatial information and image features, the model effectively extracts and classifies features in graphite mineral images. Additionally, the introduction of the Global Response Normalization (GRN) layer enhances the error adjacency property of the improved neural network model during confusion matrix performance evaluation. This implies that misclassified samples often have grades that are close to each other. Leveraging this property can help minimize resource wastage resulting from recognition errors.

Based on the hyperparameter settings described in Sect. 2.4, the quantitative analysis of the four models and their variants, following the same training strategy, is presented in Table 4. ConvNeXt-G achieves an accuracy of 93.401%, macro precision of 93.317%, macro recall of 92.856%, and macro F1 score of 93.086%. Notably, ConvNeXt-G demonstrates significant improvement compared to ResNet-50 and slightly outperforms Vision Transformer across all evaluated parameters. EfficientNet-B3 attains an accuracy of 92.718%, macro precision of 92.261%, macro recall of 92.488%, and macro F1 score of 92.802%, placing its performance on par with that of the ConvNeXt-G model in terms of evaluation metrics.

Table 4. Quantitative analysis and comparison results of evaluation indexes of different convolutional neural networks on the test set

Model	macro-P	macro-R	macro-F1	accuracy/%
ResNet-50	86.784	86.355	86.569	86.862
ResNet-101	86.845	88.957	87.888	88.002
ResNet-152	89.002	88.562	88.782	89.082
EfficientNet-B0	87.623	87.191	87.406	87.702
EfficientNet-B3	**93.001**	**91.105**	**92.043**	**92.802**
Vision Transformer-S	88.762	88.324	88.542	88.842
Vision Transformer-B	87.951	88.903	88.424	88.602
Vision Transformer-L	86.905	86.475	86.691	86.983
ConvNeXt-G	**93.317**	**92.856**	**93.086**	**93.401**

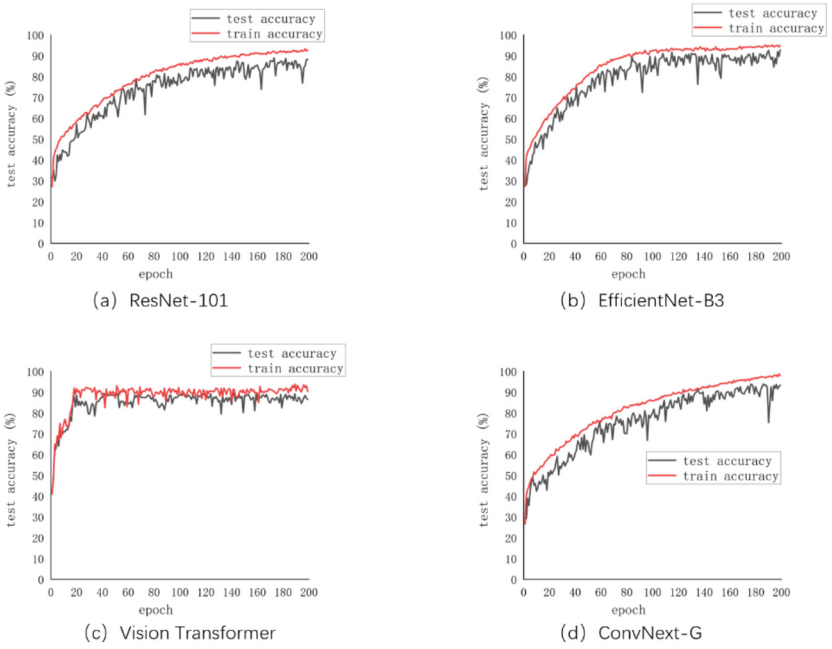

Fig. 8. The accuracy comparison of four main neural networks in the training process

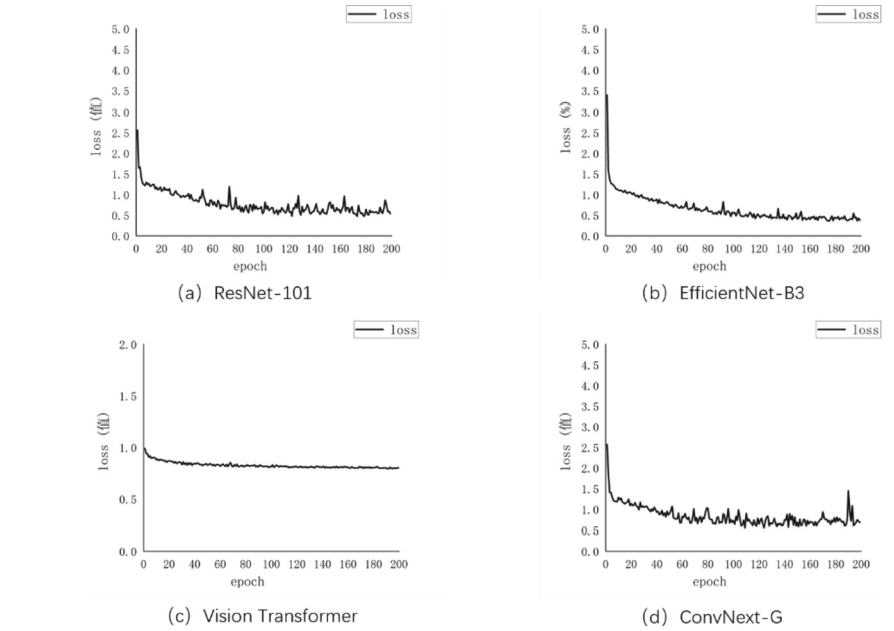

Fig. 9. Comparison of loss values of four main neural networks during training

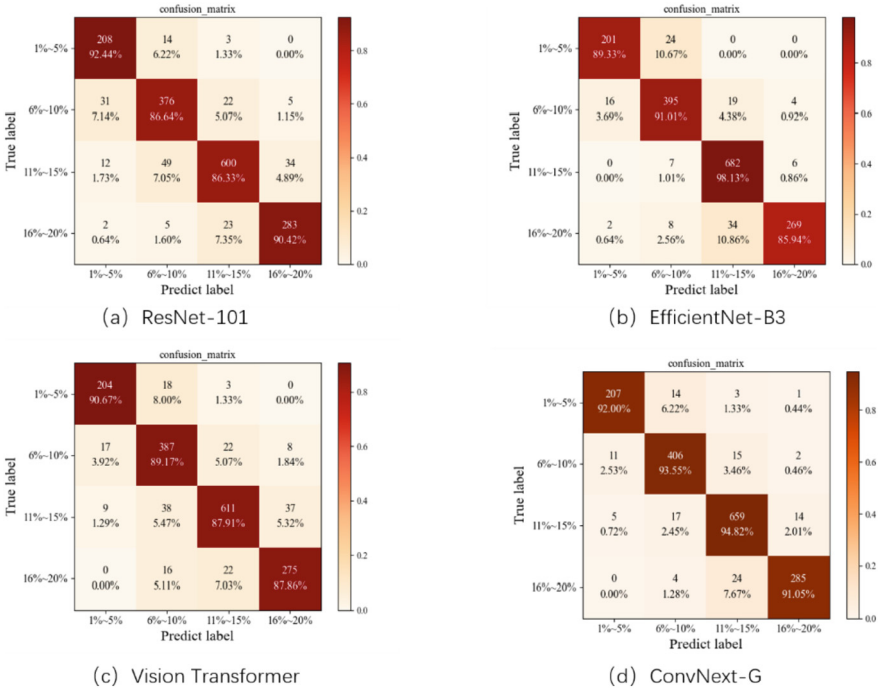

Fig. 10. The confusion matrix comparison of four main neural networks in the training process

4 Conclusion

This study proposes a convolutional neural network model that combines transfer learning and multi-scale fused image features to automatically recognize and classify graphite ore images, addressing the recognition and classification problem in this domain. Through experimental comparative analysis, we explore the application of deep learning techniques in classifying graphite ore grades, achieving improved accuracy by incorporating the ResNet-50 backbone network. To further enhance classification accuracy, we introduce spatial attention mechanisms to capture long-range dependencies more effectively. Additionally, Global Response Normalization (GRN) techniques are incorporated to enhance the precision of recognizing graphite ore grades. Our method demonstrates superior classification accuracy compared to other approaches through practical experiments. This demonstrates the effectiveness of our deep learning-based classification approach, utilizing the ResNet backbone network, in graphite ore grade classification tasks. Furthermore, compared to carbon-sulfur analysis methods that suffer from significant time delays, our approach of utilizing texture features from graphite ore images for ore grade recognition offers advantages in terms of speed, accuracy, and portability, making it a feasible solution. Through our experiments, we illustrate the complementary nature of long-range dependency learning and global response normalization in the context of graphite ore image classification. By integrating multi-scale features, we effectively improve the accuracy of graphite ore image recognition.

References

1. Su, L., Cao, X., Ma, H., et al.: Research on coal gangue identification by using convolutional neural network. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 810–814. IEEE (2018)
2. Pu, Y., Apel, D.B., Szmigiel, A., et al.: Image recognition of coal and coal gangue using a convolutional neural network and transfer learning. Energies **12**(9), 1735 (2019)
3. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
4. Baraboshkin, E.E., Ismailova, L.S., Orlov, D.M., et al.: Deep convolutions for in-depth automated rock typing. Comput. Geosci. **135**, 104330 (2020)
5. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
6. Liu, C., Li, M., Zhang, Y., et al.: An enhanced rock mineral recognition method integrating a deep learning model and clustering algorithm. Minerals **9**(9), 516 (2019)
7. Zeng, X., Xiao, Y., Ji, X., et al.: Mineral identification based on deep learning that combines image and mohs hardness. Minerals **11**(5), 506 (2021)
8. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
9. Yun, S., Han, D., Oh, S.J., et al.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
10. Liang, Y., Cui, Q., Luo, X., et al.: Research on classification of fine-grained rock images based on deep learning. Comput. Intell. Neurosci. **2021** (2021)
11. Iglesias, J.C.Á., Santos, R.B.M., Paciornik, S.: Deep learning discrimination of quartz and resin in optical microscopy images of minerals. Miner. Eng. **138**, 79–85 (2019)

12. Han, S., Li, H., Li, M., et al.: Measuring rock surface strength based on spectrograms with deep convolutional networks. Comput. Geosci. **133**, 104312 (2019)
13. Ran, X., Xue, L., Zhang, Y., et al.: Rock classification from field image patches analyzed using a deep convolutional neural network. Mathematics **7**(8), 755 (2019)
14. de Lima, R.P., Bonar, A., Coronado, D.D., et al.: Deep convolutional neural networks as a geological image classification tool. Sediment. Rec. **17**(2), 4–9 (2019)
15. Xiao, D., Le, B.T., Ha, T.T.L.: Iron ore identification method using reflectance spectrometer and a deep neural network framework. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. **248**, 119168 (2021)
16. Liu, X., Wang, H., Jing, H., et al.: Research on intelligent identification of rock types based on faster R-CNN method. IEEE Access **8**, 21804–21812 (2020)
17. Xu, Z., Ma, W., Lin, P., et al.: Deep learning of rock images for intelligent lithology identification. Comput. Geosci. **154**, 104799 (2021)
18. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOV4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
19. He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
20. Chen, Z., Yang, J., Chen, L., et al.: Garbage classification system based on improved ShuffleNet v2. Resources Conserv. Recycl. **178**, 106090 (2022)
21. Xie, S., Girshick, R., Dollár, P., et al.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
22. Woo, S., Park, J., Lee, J.Y., et al.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
23. Woo, S., Debnath, S., Hu, R., et al.: ConvNeXt V2: co-designing and Scaling convnets with masked autoencoders. arXiv preprint arXiv:2301.00808 (2023)