IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Deep Learning Based Mineral Image Classification Combined with Visual Attention Mechanism

Yang Liu2#, Zelin Zhang1,2#, Xiang Liu1#, Lei Wang1*, Xuhui Xia1*

[1] Key Laboratory of Metallurgical Equipment and Control Technology, Wuhan University of Science and Technology, Wuhan 430081, China;
[2] Hubei Key Laboratory for Efficient Utilization and Agglomeration of metallurgic Mineral Resources, School of Resource and Environmental Engineering, Wuhan University of Science and Technology, Wuhan 430081, China.

Yang Liu (clearliu777@163.com), Zelin Zhang (zhangzelin@wust.edu.cn) and Xiang Liu (liuxiang@wust.edu.cn) contributes equally to this work.

Corresponding author: candywang@wust.edu.cn (Lei Wang); xiaxuhui@wust.edu.cn (Xuhui Xia)

This work was supported in part by the National Natural Science Foundation of China (No.51604196 and 51805385), and the Key Research and Development project of Hubei Province (No. 2020BAA024 and 2020BAB047).

ABSTRACT Mineral image classification technology based on machine vision is an efficient system for ore sorting. With the development of artificial intelligence and computer technology, the deep learning-based mineral image classification system is gradually applied to ore sorting. However, there is a bottleneck in improving classification accuracy, and the feature extraction ability of the CNNs model is relatively limited for multi-category mineral image classification tasks. Therefore, four visual attention blocks are designed and embedded in the existing CNNs model, and new mineral image classification models based on the visual attention mechanism and CNNs are proposed. Then, referring to the building strategies of the different depth ResNet, we build various CNNs model embedding with attention blocks for mineral image classification and visualize the models by Grad-CAM to observe the change in classification weight distributions and classification weight values. Finally, by using the confusion matrices, this experiment systematically evaluates the classification performance of the proposed models and analyzes the misjudgment rate.

INDEX TERMS Deep Learning; Visual Attention mechanism; Mineral image classification; Grad-CAM;

Section 1. Introduction

At this stage, the exploitation and application of mineral resources have entered a new era since the inventory of their mineral resources has declined rapidly with the growth of industrial development, which raises new demands for ore mining and application technology. Recently, intelligent ore sorting has become one of the crucial factors for mineral processing and mining enterprises, which not only saves workforce and material consumption, increases mining safety factors but also lays the foundation for sustainable development. For example, intelligent ore sorting technology can quickly realize the gangue discharge or pre-separation of underground or concentrator feed and effectively reduce the energy consumption of lump ore crushing, grinding, and other processes.

When exploring intelligent ore sorting equipment, scholars first applied it based on high-tech sensors, which effectively replaces the manual sorting process, improves particle separation efficiency, and reduces pollution treatment costs [1]. At present, the intelligent ore sorting equipment put into production is mainly based on ray sensors and used in large-grain particle identification and separation, including XRT and XRF, which has a high classification accuracy and fast classification speed [2]–[5]. However, the problems such as high cost and high radiation still limit their further application and development.

With the development of computer technology and digital image acquisition equipment, the intelligent ore sorting equipment with digital images as processing objects has been gradually applied to industrial practice. In contrast to ray sensor-based sorting equipment, machine vision-based ore sorting equipment extracts the ore feature information from the images collected through optical components and completes the image classification task in static or dynamic scenes. Therefore, it has the advantages of low cost, high efficiency, no radiation, and easy installation. At this stage, there are two central technical cores of ore sorting equipment based on machine vision: the machine learning-based image classification technology and the deep learning-based image classification technology. Firstly, there are two main streams of the machine learning-based image classification technology, including the supervised learning algorithm and unsupervised learning algorithm, among which supervised learning algorithm has a better performance. Specifically, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

supervised learning algorithms mainly include Decision Tree [6], [7], Naive Bayesian [8], K-Nearest Neighbors [9], Support Vector Machine (SVM) [10], which all have been experimented, tested, and applied in mineral image classification tasks [11]–[19]. However, the applications of machine learning-based mineral image classification models need to be supported with higher resolution images. Due to harsh image acquisition environments and complex working conditions (rainy weather or dust), the stable acquisition of high-resolution images is relatively difficult, which increases the workload and difficulty of image acquisition. Additionally, in the machine learning-based ore image classification models, the process of feature selection requires a series of experience and knowledge, which increases the threshold of its application and limits its development prospects.

On the other hand, with the development of artificial intelligence and further exploration of computer technology, the deep learning-based image classification technology matures gradually and has achieved excellent performances in many image-classification tasks [20]–[22]. Specifically, it replaces the feature selection process with convolution neural networks (CNNs) to automatically extract image features and filter the extracted feature maps. Additionally, the deep learning-based image classification model reduces the dependence on high-resolution images, improving the model classification efficiency and accuracy. In the field of mineral image classification, scholars have explored the relative application potentials of deep learning-based image classification systems. For example, combining the deep learning technology and CNNs, Fu and Aldrich used VGGNet to classify mineral images in South Africa and compared its performance with traditional machine learning-based mineral image classification systems [23]. Similarly, using VGGNet, Zhu et al. classified ten different ore slice images, the classification accuracy reached 98.1%, and the time consuming of the single image classification is only 1.5s [24]. Combined with VGG16 and Principal Component Analysis (PCA), Sudakov et al. classified core slice images and achieve a preferred classification accuracy [25]. In the exploration of the complex CNNs models, based on the InceptionV3, LP et al. proposed a four-classes rock image classification system, which has an higher classification accuracy than machine learning-based image classification models [26]. Zhang Y et al. efficiently completed the classification task of potassium feldspar, perlite, plagioclase, and quartz images by combining Transfer Learning technology with the InceptionV3 [27]. In the comparison of the different CNNs-based image classification models, Baraboshkin et al. used AlexNet, VGGNet, and Inception to classify 20,000 rock images collected from different regions and strata [28]. Additionally, because of the excellent classification performance of CNNs models in mineral image classification tasks, it is also used for coal gangue discharge [29], [30] and iron ore image classification [31].

The performances of those mentioned above deep learning-based ore image classification models on the corresponding tasks have proved that it will as the mainstream of the intelligent ore sorting equipment. However, there still have bottlenecks in its classification performance for the multi-category (> 2) mineral image classification tasks. Meanwhile, it is widely known that the feature extraction and operation processes of the CNNs models are automatic. Therefore, some irrelevant information in the ore images will interfere with the model feature extraction ability during the model training phase, such as reflect light, dust, and noise points, which will result in loss of model classification accuracy. As a result, the above problems will limit the application potentials and development prospects of the deep learning-based ore image classification systems.

Nowadays, CNNs models that incorporate visual attention mechanisms have become a popular area in deep learning-based image classification research, which is inspired by the physiological perception of the human eyes for environments. Precisely, the CNNs-based image classification model incorporating visual attention can extract image feature information at key locations with a lower extra computational cost, thus improving the classification performance [32]. Therefore, in order to solve the above difficulties and improve the application potentials of the deep learning-based ore sorting equipment, this paper takes the multi-category ore image classification task as the research aspect and proposes to embed the visual attention mechanism in the deep learning-based mineral image classification model. Specifically, taking gas coal, coking coal and anthracite as experimental objects, referring to the building strategy of the ResNet, this experiment firstly builds four ResNet mineral image classification models with different depths for four-classes ($<1.4 g/cm^3$, $1.4\text{-}1.6 g/cm^3$, $1.6\text{-}1.8 g/cm^3$, and $>1.8 g/cm^3$) mineral image classification, including ResNet18, ResNet34, ResNet50, and ResNet101. After that, we embedded four visual attention modules into the ResNet and compared the performance of different visual attention modules. Finally, through various performance metrics and classification result visualization, this experiment measures the increase of model complexity and the distribution change of the classification weight after adding the visual attention modules.

In summary, this paper focuses on the following four aspects of the deep learning-based mineral image classification systems:

(1) How to build and embed visual attention modules for mineral image classification models?

(2) Comparing the performance of the visual attention modules in multi-category mineral image classification tasks.

(3) Comparing the classification performance of the general CNNs models and CNNs model embedded with visual attention modules.

(4) How the visual attention modules influence the distributions and values of model classification weight?

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

Section 2. Methodology

This section mainly introduces the strategies and methodologies of building the deep learning-based mineral image classification system that incorporates visual attention mechanism, mainly including dataset preparation and the building of CNNs models embedded with attention blocks, as shown in FIG. 1. Specifically, in the data set preparation stage, this experiment uses Data Augmentation (DA) technology to solve the problem of insufficient image data during training, and when building the CNNs classification model embedded with visual attention modules, this experiment mainly considers three aspects: Pooling strategy selection, attention block construction, and CNNs model settings.

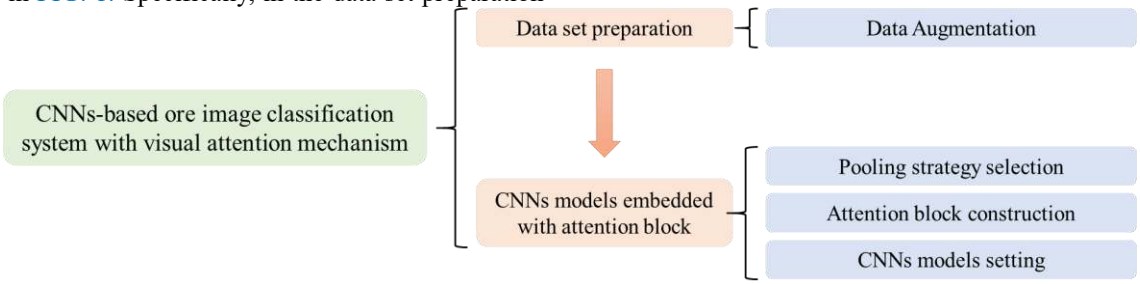

FIGURE 1. **Flow chart of CNNs-based classification system with visual attention mechanism**

2.1 Data Augmentation

Sufficient image data is the application basis of the deep learning-based image classification technology. However, with the gradual exploration of CNNs models with deeper layers and more complex structures, scholars often encounter problems such as insufficient training data and unbalanced data quantity between the different categories. Specifically, for the deep learning-based mineral image classification tasks, researchers often encounter problems from three aspects: Firstly, the harsh working conditions of industrial applications reduce the stability and efficiency of the high-quality mineral image collection. Secondly, mineral image classification is different from general image classification tasks, so no uniform and large mineral image datasets have been established. Therefore, the inadequate amount of training images can lead to over-fitting problems in the training phase. Additionally, in the process of the dataset preparation, the imbalance quantity of ore images between different categories will lead to the imbalance of the extracted features, which will influence the model classification accuracy.

A practical and effective strategy to solve the above bottlenecks is DA technology [33]. Specifically, DA technology expands existing image data sets based on small data sets already obtained by the applicants, and the classic image DA methods include flipping, rotating, scaling, clipping, color dithering, and adding Gaussian noise. Besides, the DCGAN has been gradually applied to the preparation of mineral image data sets, which uses limited mineral images to generate more new mineral images for specified tasks automatically, and the generated images will not lose the feature information. Consequently, making full use of the DA technology will improve the robustness of the model and reduce the possibility of over-fitting in model training.

Due to the mentioned advantages of the DA technology, in the task of mineral image recognition and classification, scholars always use DA for mineral image dataset preparation, which effectively solves the problem of insufficient data and unbalanced data quantity between different category in the training phase, promoting the further application of deep learning-based ore sorting equipment [28], [34], [35].

2.2 Convolutional Neural Networks (CNNs)

CNNs is a feed-forward neural network with deep structure and convolution calculation, and it is one of the representative algorithms in deep learning. It has strong feature-learning abilities and uses convolution layers to extract features from input images by hierarchical structure automatically. Specifically, the basic CNNs are composed of convolution layers, activation layers, normalization layers, pooling layers, and fully connected layers, as shown in FIG. 2. For image classification tasks, the convolution layers use convolution kernel filters to calculate the pixel information in the input image and output it as matrices. Normalization layers reduce the dimension of the convolution layer outputs, enhance the model convergence and improve the model training efficiency. Activation layers process the output feature information of the normalization layers to determine whether valid image features have been captured. Pooling layers operate a down-sampling process that preserves some of the representative features in specific ways, thereby reducing the dimensions of the feature space. The fully connected layers link the front convolution sections and combine the extracted features non-linearly to get the output. At the same time, the output value of the last fully connected layer is an N-dimensional vector, which is the number of classification categories.

Additionally, it is worth knowing that deep learning-based image classification technology has shown excellent performance in many fields, such as agriculture image classification [36] and medical image classification [37], and the typical networks include AlexNet [33], VGGNet [38], Inception [39] and ResNet [40].

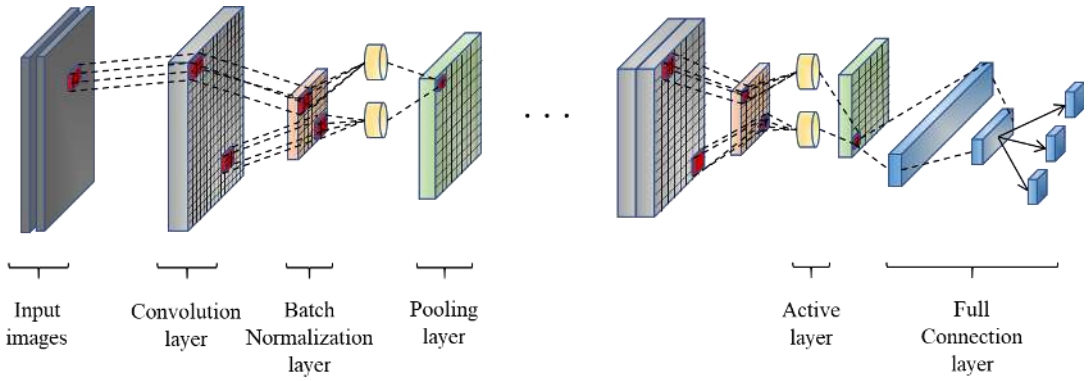

FIGURE 2. General Convolutional Neural Networks framework including inputs, convolution layers, batch normalization layers, pooling layers, active layers, and fully connected layers

2.3 Pooling strategy in CNNs

Pooling calculation is one of the common processes in CNNs, which mimics the human visual system and reduces the data dimension, often referred to as sub-sampling or down-sampling. When building a CNNs model, the position of the pooling layer is behind the convolution layer to reduce the dimension of the convolution layer output, effectively reducing network parameters and preventing over-fitting. Additionally, pooling calculation will suppress noise, reduce information redundancy, and improve detection scale and rotation invariance. Pooling calculations in the CNNs model include various pooling strategies, mainly divided into Max Pooling, Average Pooling, and Stochastic Pooling, as shown in FIG. 3.

Firstly, for Max Pooling, it selects the maximum value in the pooling kernel area as the output value of the pooling operation, and the gradient of the pooling value at other locations is 0. Secondly, for Average Pooling, it adds and averages the eigenvalues of each location in the corresponding pooling kernel area, using the average as the output value of the pooling operation. Additionally, for Stochastic Pooling, the probability of being selected is first determined by comparing each value in the pooling kernel (the darker the color in FIG. 3., the higher the probability of being selected). Then, it randomly selects the representative value of the pooling kernel based on the selecting probability from each location.

For image classification tasks, different pooling strategies focus on and preserve different information in the input images. Specifically, Max Pooling, which chooses the maximum value, will pay more attention to the more specific information in the input image to better preserve the texture information in the input image. Average Pooling tends to preserve the feature information of the overall input image, highlighting the background information and contour information of the target objects better. Stochastic Pooling has no specific direction of interest in preserving feature information in images, but it normalizes the input image feature information by random selection, which improves the robustness of the CNNs model and averages the attention to different feature information.

However, there are still some drawbacks to using only one pooling strategy for down sampling. Firstly, using Average Pooling or Max Pooling alone can lead to loss of useful information. Since the Average Pooling picks the average of the activation values for all pixels, the higher positive activation values may offset the lower negative activation values, resulting in a loss of discriminatory feature information, while the Max Pooling discards all non-maximum values, this will directly result in the loss of helpful feature information. Additionally, Stoical Pooling has a lower tendency to the specific feature preservation, like texture information or contour information.

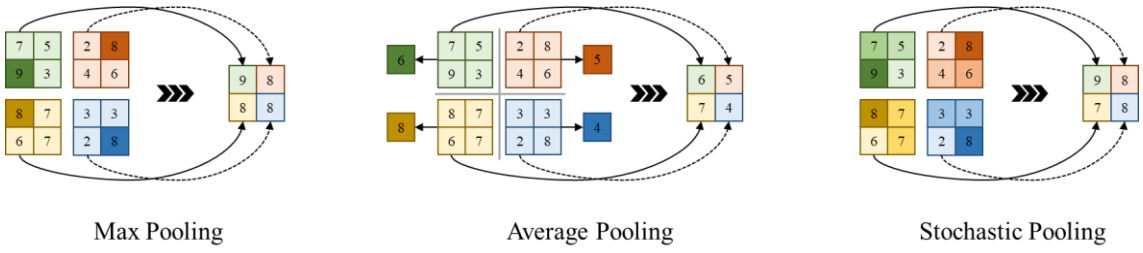

FIGURE 3. Different pooling strategies in CNNs including Max Pooling, Mean Pooling, and Stochastic Pooling.

2.4 Attention Mechanism in CNNs

In order to solve the problem of the loss of helpful and specific feature information caused by using a single pooling strategy, this experiment proposes four attention blocks that incorporate visual attention mechanisms, mainly including Squeeze and Excitation (SE) block, Channel Attention (CA) block, Spatial Attention (SA) block, and Mixed Attention (MA) block. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

construction details of various visual attention blocks are as follow.

2.4.1 Squeeze and Excitation block

SE block originated from SE Net proposed by Hu et al. in the Image Net 2017 competition, which emphasizes the information relationship along the channel direction in the CNNs model, and the basic design method of SE block is shown in FIG. 4., mainly consisting of three steps, Squeeze, Excitation and Scale [41].

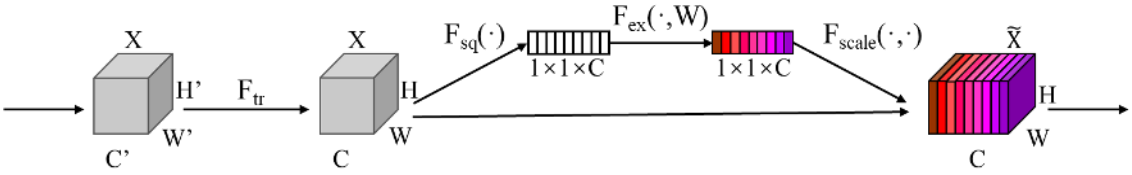

FIGURE 4. Squeeze and Excitation block framework in CNNs, where H, W, and C are the dimension of the feature map, H is height, W is width, C is channel, Ftr is the convolution calculation, Fsq is the Squeeze operation, Fex is the Excitation operation, Fscale is the Reweight operation, $\tilde{X}$ is the feature map after attention.

The implementation of the SE block requires the input of previous feature maps, so in the front-end of the SE block, the input image needs to go through a standard convolution operation to get the feature maps (Ftr), as shown in EQ (1).

$$u_c = v_c * X = \sum_{S=1}^{c'} v_c^s * x^s \quad (1)$$

v_c——The c-th filter
X——The input images
$*$——The convolution calculation operation
u_c——The output feature map
v_c^s——A 2D spatial kernel representing a single channel of v_c that acts on the corresponding channel of x^s

After the feature maps enter the SE block, it first performs the Squeeze operation, which a simple cluster technique, like Global Average pooling (EQ (2).). Specifically, Squeeze processing uses global average pooling operations to squeeze each feature map, turning each two-dimensional channel feature into a real number, and it will have a global perception field, representing the global feature distribution of the channel response, which makes the global information available to the CNNs lower layers. Eventually, c feature maps will become a $1 \times 1 \times c$ real number sequence.

$$Z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \quad (2)$$

H——The height of the input feature map
W——The width of the input feature map
$u_c(i,j)$——The feature map at (i,j)

After the Squeeze operation, the output feature map is processed by Excitation operation, which is an adaptive recalibration, and the SE block accomplishes this by using a fully connected layer (EQ (3).). During the processing, the dimensions of W_1 are $C/r * C$, where r is a scale parameter to reduce the number of channels and the amount of computation. Then, it will be calculated by a Rectified Linear Unit (ReLU) function and multiplied by W_2, which is also a fully connected layer operation, and the dimensions of W_2 are $C * C/r$. The bottleneck structure of the two fully connected layers effectively reduces the model complexity, improves the model generalization ability, and makes the SE block more non-linear to fit the complex relations between channels better. Among them, the first fully connected layer plays the role of dimension reduction, and the second fully connected layer is used for dimension restoring. Finally, a normalized weight between 0 and 1, the output value s, is obtained by the Sigmoid gating. The Sigmoid function effectively learns about the non-linear, non-mutually exclusive correlations between channels and ensures visual attentions for multiple channels.

$$s = F_{ex}(z, W) = \sigma\big(g(z, W)\big) = \sigma\big(W_2 \delta(W_1 z)\big) \quad (3)$$

z——The output of the Squeeze operation
W_1——The first fully connected operation
δ——The ReLu calculation
W_2——The second fully connected operation
σ——The sigmoid function calculation

After an Excitation operation, the feature will be reweighted, which is accomplished by concatenating the normalized weights onto feature maps of each channel, also known as Scale operation. Specifically, the Scale operation regards the Excitation output as the importance of each channel and weights it to the previous feature by multiplying the weight coefficients (EQ (4).).

$$\widetilde{X_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \quad (4)$$

$\widetilde{X_c}$——The channel-wise multiplication between the scalar s_c and the feature map u_c

The SE block can be directly and flexibly applied to the existing CNNs models, and its embedding strategy in ResNet is shown in FIG. 5. Specifically, for ResNet, since it contains residual modules, the SE block can be directly embedded in its residual learning branch.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

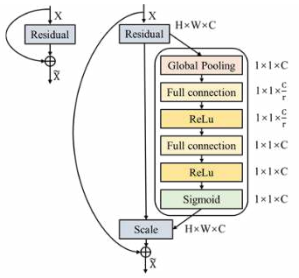

FIGURE 5. Squeeze and Excitation block setting in ResNet

Firstly, for feature maps with the size of $H \times W \times C$, the SE block first squeezes each feature map through global average pooling calculation to get a $1 \times 1 \times C$ real number column, which is the Squeeze operation; Two fully connected layers are then introduced for Excitation operation, with a $1 \times 1 \times C/r$ size first fully connected layer and a $1 \times 1 \times C$ size second fully connected layer. Next, through the Sigmoid gating, the SE block normalizes the front output values between 0-1, which represents the weight of each channel. After that, the channel is recalibrated by multiplying the attention weights with the original input feature map by Scale operation. Consequently, through a set of input feature map processing by SE block, the CNNs model will obtain a feature map that incorporates the visual attention mechanism.

Although the embedding of the SE block will increase the model training parameters and calculation complexity, the increase is usually less than 1% of the original calculation parameters when r is set reasonably, and the calculation formula of the parameter increment is shown in EQ (5). As a result, embedding the SE blocks to CNNs models is an efficient, fast, and less-cost way to apply visual attention mechanisms.

$$Parameter\ increment = n\frac{2C^2}{r} \tag{5}$$

n——The numbers of SE block in CNNs model
r——The value of dimensionality reduction
C——The numbers of channels in SE block

2.4.2 Channel Attention block and Spatial Attention block

In order to further promote the application of CNNs model that incorporates visual attention mechanism in image classification tasks, based on SE block, Woo proposed CA block and SA block [42]. The setting strategies of the CA block and SA block are shown in FIG. 6.

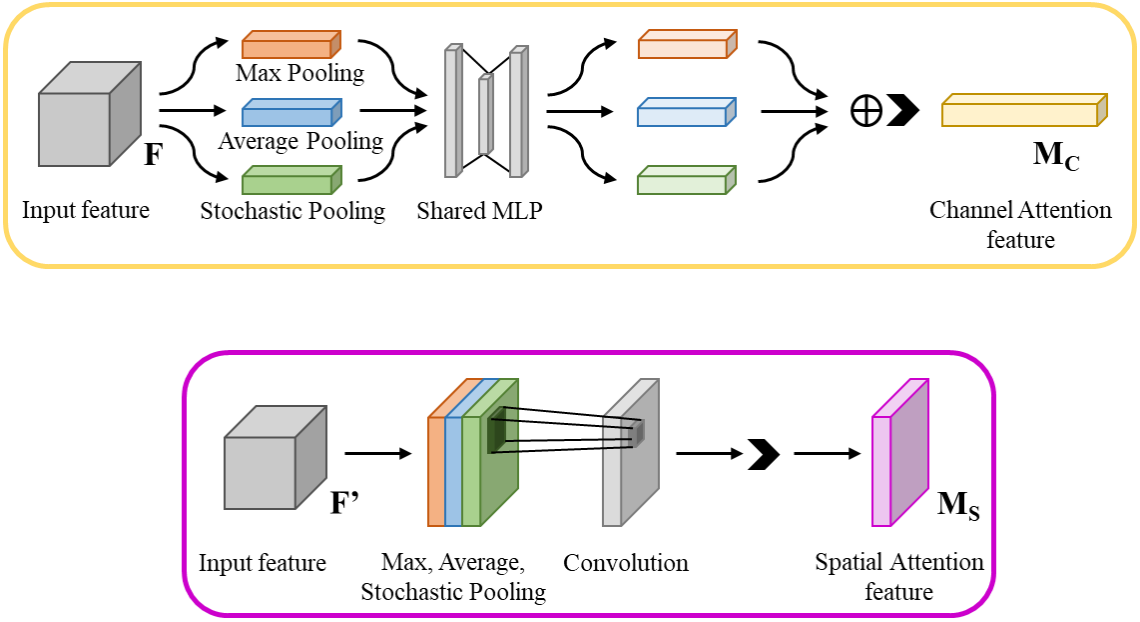

FIGURE 6. Channel attention block and spatial attention block frameworks

For CA block, it focuses on the problem of "what is meaningful?" in the input feature map and calculates the internal relationship between different channels. Specifically, CA block uses Max Pooling, Average Pooling, and Stochastic Pooling to calculate input feature map F, respectively. Among them, Max Pooling obtains more detailed texture features in the input image, Average Pooling integrates spatial information on each channel, and Stochastic Pooling increases the generalization and robustness of the CNNs models. After that, the three output feature maps F^c_{max}, F^c_{avg}, and F^c_{sto}, which are processing by different pooling strategies, will enter a shared network consisting of a Multilayer Perceptron (MLP) with only one hidden layer. Next, to reduce the training parameters of the visual attention block, the middle layer size of MLP will be set to $R^{c/r*1*1}$ (r is the decrement rate). Then, the three new feature maps after Shared MLP are element-wise summation, which adds up the corresponding elements to get the channel attention feature maps. Therefore, the channel attention feature maps represent the intrinsic relationship between different channels, solving the problems of which channels are important and which channels should be ignored, and the formula of the whole process is expressed as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

EQ (6). Additionally, the parameter increment of embedding CA block is same as embedding SE block, as shown in EQ (5).

$$M_c(F) = \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big) + MLP\big(StoPool(F)\big)\Big) = \sigma\Big(W_1\Big(W_0\big(F^c_{avg}\big)\Big) + W_1\Big(W_0\big(F^c_{avg}\big)\Big) + W_1\big(W_0(F^c_{sto})\big)\Big) \quad (6)$$

σ——The sigmoid function calculation
M_c——The weight coefficient of channel attention

For SA blocks, the main focus is the intrinsic relationship of feature maps at the spatial level, solving the problems of which regions are important and which are secondary. As a complementary block of the CA block, the spatial attention process is relatively simple and convenient. Specifically, to get channel information for feature maps at spatial space, SA block makes Max Pooling, Average Pooling, and Stochastic Pooling along channel axis, respectively, to get $F^S_{max} \in \mathbb{R}^{1*H*W}$, $F^S_{avg} \in \mathbb{R}^{1*H*W}$, and $F^S_{sto} \in \mathbb{R}^{1*H*W}$. Then, the SA block will concatenate the three output feature maps and uses a standard convolution layer to get a spatial attention feature map; the formula for the entire process is expressed as EQ (7). Additionally, the parameter increment brought about by the SA block is shown in EQ (8).

$$M_S(F) = \sigma\big(f^{7\times7}([AvgPool(F);\ MaxPool(F);\ StoPool(F)])\big) = \sigma\Big(f^{7\times7}\big([F^s_{avg}; F^s_{max};\ F^s_{sto}]\big)\Big) \quad (7)$$

σ——The sigmoid function calculation
$f^{7\times7}$——The convolution operation with the filter size of 7×7
M_S——The weight coefficient of spatial attention

$$Parameters\ increment = 2nk^2 \quad (8)$$

n——The number of embedding SA block in CNNs model
k——The size of convolution kernel

It is worth noting that the CA block and SA block can be flexibly embedded in existing CNNs models independently to achieve channel or spatial feature attention, and their embedding strategies are similar to SE block.

2.4.3 Mixed Attention block (Channel Attention& Spatial Attention)

During the exploration of CNNs models that incorporate visual attention mechanisms, researchers have found that using a single type of attention block (spatial or channel) may not meet the requirements of feature extraction.

Therefore, to address this problem, Woo proposed a mixed attention block [42], which is the simultaneous application of the CA block and SA block., as shown in FIG. 7. At the same time, they pointed out that the combination, which sets the CA block first and the SA block later, can improve the feature-extraction performance and classification accuracy better than the combination positions were interchanged.

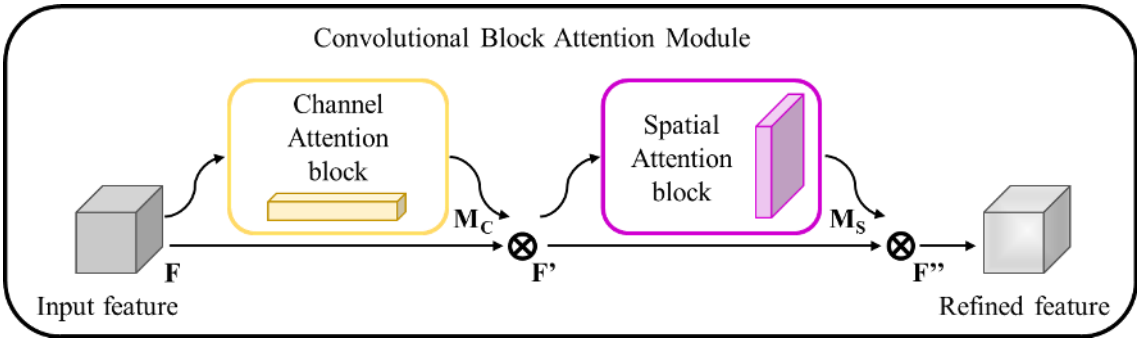

FIGURE 7. Mixed attention block combining with channel attention block and spatial attention block, F represents the input feature map, M_C refers to the weight coefficient of channel attention, F' refers to the feature map after channel attention block, M_S refers to the weight coefficient of spatial attention, F'' refers to the feature map after the spatial attention block, which is the final refined feature map, $\otimes$ refers to the element-wise multiplication

Specifically, in the MA block, the former CA block solves the problem of which parts of the input feature maps have greater classification weight, and the latter SA block solves the problem of which regions are more important in the input feature maps. In the block, after the input feature map F ($F \in \mathbb{R}^{C*H*W}$) pass through the CA block, the feature map and weight coefficient of channel attention (M_C) will be multiplied to obtain F', which is the input of the SA block, as shown in EQ (9). Then, when the F' pass through the SA block, the weight coefficient of spatial attention (M_S) is multiplied with feature map to obtain the final refined feature F'', as shown in EQ (10).

$$F' = M_C(F) \otimes F \quad (9)$$
$$F'' = M_S(F') \otimes F' \quad (10)$$

F——The input feature map
M_C——The weight coefficient of channel attention
F'——The feature map after CA block
M_S——The weight coefficient of spatial attention

F''——The feature map after the SA block, which is the final refined feature map

$\otimes$——The element-wise multiplication

Besides, the MA block can carry out end-to-end training and can be embedded in any position in CNNs models while only adding a few amounts parameters, as shown in EQ (11). Additionally, taking the embedding strategy of the MA block in ResNet as an example, it is obvious knowing that the embedding strategy of the MA block is similar to that of the SE block, which can be embedded in the residual branch directly, as shown in FIG. 8.

$$Parameters\ increment = n\left(\frac{2C^2}{r} + 2k^2\right) \quad (11)$$

n——The number of the embedded MA blocks

C——The number of channels

r——The decrement rate

k——The size of convolution kernel

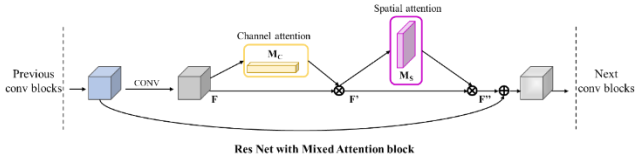

FIGURE 8. Mixed Attention block setting in ResNet

Section 3. CASE STUDY

3.1 Experimental settings

3.1.1 Material preparation

In order to explore the application potentials of CNNs models incorporating visual attention mechanism in mineral image classification tasks, this experiment takes three types of ore particles in China as experimental objects for mineral image classification tasks, including gas coal, coking coal, and anthracite coal. Specifically, three types of coal with 13-25 mm granularity are selected by manual screening, and each type of coal is 20 kg. In this experiment, to simulate industrial separation, the four classes mineral particles are divided into <1.4g/cm^3, 1.4g/cm^3-1.6g/cm^3, 1.6g/cm^3-1.8g/cm^3 and >1.8g/cm^3 according to the different density level. Before the experiment, we determined the ash content and macerals of the ore samples. The mean ash content of each type of coal sample is shown in TAB 1., and the maceral analysis is shown in TAB 2.

The results show that among the three types of coal samples, the ash content increases with the increase of density level, that is, the ash content of each kind of <1.4g/cm^3 coal is the lowest, and that of each kind of >1.8g/cm^3 coal is the highest. In the comparison of the ash content in the three kinds of coal samples, the ash content of anthracite in each density set is relatively low, followed by gas coal and coking coal.

The results of maceral analysis show that under the same density level of three types of coal, the organic matter of each type of coal sample mainly concentrates on the vitrinite. With the increase of density level, the content of organic matters decreases, and the content of mineral matters increases gradually, which means the particle of each <1.4g/cm^3 coal contains higher organic matters, and the particle of each >1.8g/cm^3 coal contains more mineral matters (inorganic matters). In comparing three types of coal samples, the anthracite vitrinite has higher organic matters, coke coal is next, and gas coal is lowest under each density level. Secondly, anthracite and gas coal have relatively more mineral matters (inorganic matters) in each density of coal samples, while gas coal has relatively fewer mineral matters in each density. It is well known that differences in rock composition between different density levels will influence their apparent characteristics. Therefore, the differences in apparent characteristics will also affect the feature extraction and classification performance of the CNNs-based ore image classification model incorporating visual attention mechanism.

TABLE 1. Mass percentage and mean ash content of four density level gas coal, cooking coal and anthracite samples

Coal type	Coal property	Density level			
		<1.4 g/cm³	1.4-1.6g/cm³	1.6-1.8g/cm³	>1.8g/cm³
Gas Coal	Ash Content	7.5%	22.8%	46.3%	85.7%
	Mass Percentage	33.3%	14.4%	20.3%	32.0%
Cooking Coal	Ash Content	9.3%	24.3%	41.3%	87.4%
	Mass Percentage	27.6%	27.8%	11.0%	33.6%
Anthracite	Ash Content	7.1%	20.6%	40.3%	83.6%
	Mass Percentage	36.0%	23.1%	10.5%	30.5%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

TABLE 2. Maceral analysis of four density level gas coal, cooking coal and anthracite samples

Coal type	Coal property	Density level			
		<1.4g/cm³	1.4-1.6g/cm³	1.6-1.8g/cm³	>1.8g/cm³
Gas Coal	Vitrinite	67.3%	64.0%	52.6%	12.6%
	Exinite	16.4%	19.8%	22.8%	5.6%
	Inertinite	12.4%	10.5%	11.9%	4.1%
	Minerals	2.0%	5.7%	12.7%	77.7%
Cooking Coal	Vitrinite	82.1%	74.3%	54.0%	11.0%
	Exinite	0.6%	3.7%	2.3%	0.4%
	Inertinite	13.1%	13.7%	21.8%	2.0%
	Minerals	4.2%	8.3%	21.9%	86.6%
Anthracite	Vitrinite	97.2%	88.0%	73.8%	0.0%
	Exinite	0.0%	0.0%	0.0%	15.6%
	Inertinite	0.9%	1.5%	2.6%	0.0%
	Minerals	1.9%	10.5%	23.6%	84.4%

3.1.2 Image acquisition & Dataset preparation

The dynamic mineral image acquisition system mainly consists of six parts: vibration feeder, conveyor belt, linear array industrial camera, linear lighting source, computer, and tail collector, as shown in FIG. 9. In the overall operation processes, mineral sample particles meeting the experimental requirements are evenly scattered in the front side of the conveyor belt after vibration screening. Then, the conveyor belt transmits the mineral particles to the bottom position of the industrial camera for dynamic shooting and storing. Finally, the mineral particles will be collected at the tail of the conveyor belt. Additionally, the sensor on the rotating shaft will convert the speed of the belt conveyor into a digital signal and transmit it to the industrial camera to make adaptive adjustments, which prevents frame loss and deformation of the collected images.

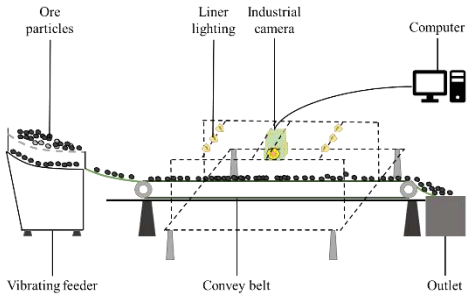

FIGURE 9. Ore particle image acquisition system consisted of vibrating feeder, convey belt, liner lighting, industrial camera, computer and outlet

Specifically, the industrial camera used in the experiment is the 4K color 3CCD linear array camera (JAI 3CCD Datasheet_LT-400CL), enabling continuous shooting and under constant speed; The linear lighting source is 500 mm that can provide uniform illumination; The color temperature is 5800-7000 k, and the surface brightness during the shooting process is about 250 Klux.

In this experiment, the color threshold segmentation algorithm is used to prepare image data set of three types of minerals, including internal particle image segmentation and edge particle image segmentation, as shown in FIG. 10.

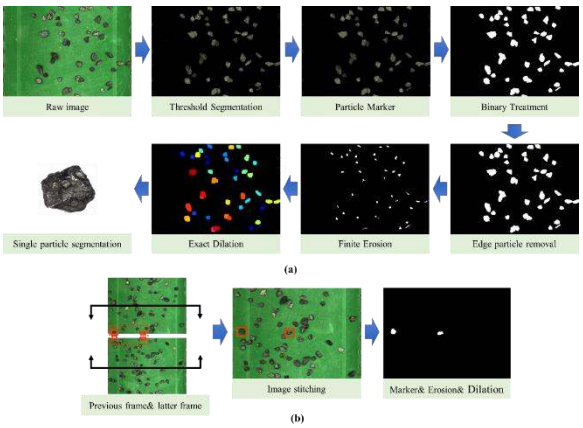

FIGURE 10. Ore particle image segmentation by color threshold segmentation algorithm: (a) inner particle image segmentation; (b) edge particle image segmentation.

In the internal particle segmentation process, the raw images are first segmented by pre-threshold value into the target areas and the background areas (EQ (12).). Secondly, the raw mineral images are processed by particle marker, binary treatment, and edge particle removal in turn. Then, the binary images without edge particles will be processed by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

Finite Erection & Exact Dilation (FEED) to resolve the adhesion and overlap problems between the adjacent particles. Specifically, Finite Erosion (FD) will erode each target area in the binary image inward with square structure elements to no connected regions between particles and record the number of FD processing. Next, Exact Dilation (ED) will restore each target area independently according to the number of FD. Finally, the segmentation system will intercept the minimum bounding rectangle of each target area after ED processing in the raw images.

In the edge particle segmentation process, for the raw images that contain edge particles, the lower half of the previous frame image and the upper half of the latter frame image will be stitched to synthesize the image. After that, the stitched raw images will be processed by particle marker, binary treatment, FEED, and the images of edge particles will be intercepted from the stitched image.

$$f(x,y,z)=\begin{cases}[0,0,0] & y-x>M\cup x<M\\ [x,y,z] & other\end{cases} \quad (12)$$

$f(x,y,z)$——The pixel value of any point in RGB images
M——The value of preset threshold

After segmentation by the above-mentioned segmentation system, we collected a total of 85,529 coal grain images from four density sets, containing 28,292 gas coal images, 29,050 coking coal images, and 28,187 anthracite images, as shown in TAB 3.

In order to thoroughly test the application potentials and classification performance of the CNNs image classification model embedded with visual attention blocks in multi-category mineral image classification tasks (four classifications) and to avoid the problems of insufficient and unbalanced image data in the CNNs model training phase, this experiment used DA technology to expand the obtained training set images to twice the original quantity, and 14,000 images from each density sets of each type of coal were randomly selected to make the data sets used for the experiment. Additionally, the proportion of the training set, valid set, and valid set is 7:2:1, as shown in TAB 4.

TABLE 3. The number of images of each density level of three types ore particles

Coal type	<1.4g/cm^3	1.4-1.6g/cm^3	1.6-1.8g/cm^3	>1.8g/cm^3	Total
Gas Coal	7074	7071	7066	7081	28292
Cooking Coal	7204	7361	7213	7272	29050
Anthracite	7063	7089	7028	7007	28187

TABLE 4. Dataset setting of each class ore particle images include training set, valid set, and test set

	Training set	Valid set	Test Set
Numbers	9800	2800	1400

3.2 Model development

3.2.1 Model building details

Referring to the ResNet building strategy, this experiment first builds four ResNet mineral image classification models with different depths, including ResNet18, ResNet34, ResNet50, and ResNet101, and adds the attention blocks (SE block, CA block, SA block, and MA block) for the residual module after each convolution section, respectively, as shown in TAB 5. Therefore, visual attention will be applied to the whole model, promoting the transmission of useful information in the network. The final layer of the constructed models is a fully connected layer, which is the Softmax classifier used to perform the four classes classification tasks (<1.4g/cm^3, 1.4-1.6g/cm^3, 1.6-1.8g/cm^3, and >1.8g/cm^3) for experimental mineral images. Additionally, this embedding strategy has little amount of increase in the model complexity and calculation parameters compared with the original network.

According to the parameter increment formulas in Section 2.3, the training parameters of four models embedding with four different attention blocks are shown in TAB 6. It can be seen from the comparison that after embedding different attention blocks to the different depths ResNet, the increased parameters of each model are similar, and it is relatively fewer compared to the original training parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

TABLE 5. Experiment CNNs models setting details

Layer name	Output size	ResNet18	ResNet34	ResNet50	ResNet101
Conv_1	112×112	7×7, 64, stride 2			
		Attention block			
Conv_2	56×56	3×3 max pool, stride 2			
		$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
		Attention block			
Conv_3	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
		Attention block			
Conv_4	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$
		Attention block			
Conv_5	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
		Attention block			
FC	1×1	Average pool, softmax (four-classes)			

TABLE 6. Parameters of different CNNs models including different depth general ResNet and ResNet embedded with different attention blocks

	General	SE block	CA block	SA block	MA block
ResNet18	11.73M	11.81M	11.81M	11.81M	11.81M
ResNet34	21.85M	22.01M	22.01M	22.01M	22.02M
ResNet50	25.68M	28.21M	28.21M	28.21M	28.22M
ResNet101	45.13M	49.91M	49.91M	49.91M	49.91M

3.2.2 Implementation setting details

This experiment is based on Python 3.6 environment, uses the Pytorch toolbox to build the models, and the detailed model training parameters are shown in TAB 7. Specifically, the model optimizer is SGD, the Learning rate is 1 × 10-4, momentum is 0.9, the Dropout rate is 0.5, the Loss function is categorical_crossentropy, and the Decay rate r of the attention blocks is 16. When model fitting, the batch size is 32, the number of epochs is 200, and the ReduceLROnPlatea (factor=0.5, patience=3) is used as the training callback, which monitors valid loss. Additionally, to improve the model training efficiency and obtain the optimal classification model quickly and accurately, this experiment uses the Early Stopping (Min_delta=0, patience=10) to monitor the valid loss in the training phase, and the models are trained in NVIDIA RTX 2080ti, cuda 10.1, and cudnn 7.3.1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

TABLE 7. Model training hyper parameters setting details

	Parameter name	Selected value
Basic setting	Optimization name	SGD
	Learning rate	1×10-4
	Momentum	0.9
	Dropout rate	0.5
	Loss function	categorical_crossentropy
Attention block setting	r	16
	Kernel size	7
Fit setting	Batch size	32
	Epoch	200
ReduceLROnPlateau	Monitor	Valid loss
	Factor	0.5
	Patience	3
EarlyStopping	Monitor	Valid loss
	Min_delta	0
	Patience	10
Environment	GPU	Nvidia RTX 2080Ti
	Platform	Python 3.6
	Tool box	Pytorch

3.3 Result analysis

3.3.1 Model evaluation and comparison

After the training of different depth ResNet and ResNet embedded with attention blocks, we recorded the training accuracy, training loss, valid accuracy, valid loss, training time (per epoch), and convergent epoch of the models to evaluate and compare the classification performance of different models in multi-category mineral image classification tasks, as shown in TAB 8., TAB 9., and TAB 10. Train accuracy and train loss are used to evaluate the training performance of the models, valid accuracy and valid loss are used to evaluate the classification performance of the models, training time (per epoch) and convergent epoch are used to evaluate the training difficulty of the models, and all of the data are copied from the Pytorch toolbox. The specific evaluation and analysis of each model in different mineral image classification tasks are as follows.

TABLE 8. Different depth ResNet and ResNet with attention block evaluation in gas coal dataset including train accuracy, train loss, valid accuracy, valid loss, training time (per epoch), and convergent epoch

	Train Accuracy	Train Loss	Valid Accuracy	Valid Loss	Training Time	Convergent Epoch
ResNet18	85.38%	0.9348	78.93%	1.3481	303s	175
ResNet18_SE	86.50%	0.9101	79.12%	1.3052	306s	164
ResNet18_CA	86.55%	0.9084	79.19%	1.3011	306s	171
ResNet18_SA	85.21%	0.9245	78.94%	1.3195	306s	173
ResNet18_MA	86.98%	0.8994	**79.29%**	**1.2884**	308s	**168**
ResNet34	87.48%	0.8471	81.47%	1.0492	427s	155
ResNet34_SE	89.74%	0.5824	81.80%	1.0384	431s	147
ResNet34_CA	89.69%	0.6385	81.75%	1.0399	431s	143
ResNet34_SA	89.41%	0.7148	81.51%	1.0498	430s	139
ResNet34_MA	90.95%	0.5215	**81.95%**	**1.0062**	438s	**131**
ResNet50	92.19%	0.4728	83.59%	0.9717	586s	126
ResNet50_SE	92.58%	0.4597	83.92%	0.9673	591s	**121**
ResNet50_CA	92.51%	0.4637	83.95%	0.9611	591s	125
ResNet50_SA	92.47%	0.4683	83.39%	0.9973	589s	**121**
ResNet50_MA	93.13%	0.4297	**84.89%**	**0.8571**	594s	127
ResNet101	93.97%	0.3887	85.18%	0.8395	945s	141
ResNet101_SE	94.48%	0.3417	86.21%	0.8274	959s	124
ResNet101_CA	93.94%	0.3899	86.23%	0.8215	948s	136
ResNet101_SA	93.57%	0.4067	85.83%	0.8304	952s	131
ResNet101_MA	94.51%	0.3985	**86.41%**	**0.8148**	961s	**120**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

From the performance metrics of the gas coal dataset, firstly, when comparing the ResNet with different depths without attention block, the results show that with the increase of network depth, the classification performance of the model for gas coal images is gradually improved, and the accuracy of the deepest-layer ResNet101 is 6.25% higher than that of the shallowest-layer ResNet18. Secondly, when comparing the ResNet with the same depth (general ResNet and ResNet embedded with attention blocks), the results show that embedding attention block to the models can effectively improve gas coal image classification accuracy. Notably, in the overall trend, embedding MA block to ResNet can maximize the classification accuracy of the models. For example, the classification accuracy of ResNet18_MA, ResNet34_MA, ResNet50_MA, and ResNet101_MA is 79.29%, 81.95%, 84.89%, and 86.41%, respectively. Compared with the same-depth ResNet without attention block, the classification accuracy of four ResNet_MA increase by 0.36%, 0.48%, 1.3%, and 1.23%, respectively. Additionally, the models embedded with SE block, CA block, and SA block also have higher classification accuracy than those without attention block. Therefore, the above experimental results show that embedding attention blocks to CNNs models can effectively improve the classification accuracy for the multi-category mineral image classification task.

TABLE 9. Different depth ResNet and ResNet with attention block evaluation in cocking coal dataset including train accuracy, train loss, valid accuracy, valid loss, training time (per epoch), and convergent epoch

	Train Accuracy	Train Loss	Valid Accuracy	Valid Loss	Training Time	Convergent Epoch
ResNet18	83.54%	0.9736	76.17%	1.5319	303s	184
ResNet18_SE	82.74%	0.9964	76.35%	1.5289	306s	181
ResNet18_CA	83.83%	0.9621	76.33%	1.5297	306s	178
ResNet18_SA	83.19%	0.9996	76.21%	1.5112	306s	168
ResNet18_MA	84.35%	0.9574	**76.42%**	**1.5174**	307s	**173**
ResNet34	84.21%	0.9598	77.98%	1.4627	427s	145
ResNet34_SE	84.35%	0.9571	78.48%	1.3704	431s	142
ResNet34_CA	84.29%	0.9592	78.51%	1.3579	431s	144
ResNet34_SA	83.89%	0.9690	78.46%	1.3586	431s	144
ResNet34_MA	84.90%	0.9503	**78.55%**	**1.3573**	436s	**136**
ResNet50	84.97%	0.9486	79.79%	1.1751	586s	137
ResNet50_SE	86.03%	0.8303	80.27%	1.1382	591s	**122**
ResNet50_CA	85.73%	0.8342	80.33%	1.1334	591s	128
ResNet50_SA	85.39%	0.8366	80.25%	1.1380	589s	130
ResNet50_MA	87.51%	0.8075	**81.19%**	**1.0574**	595s	134
ResNet101	86.63%	0.8084	81.71%	1.0231	945s	136
ResNet101_SE	87.83%	0.8059	82.52%	0.9957	959s	124
ResNet101_CA	87.28%	0.8163	82.49%	0.9963	948s	132
ResNet101_SA	87.22%	0.8168	82.43%	0.9968	952s	136
ResNet101_MA	90.27%	0.5378	**82.95%**	**0.9787**	960s	**109**

In the coking coal image data set, different depth ResNet shows similar classification performance as in the gas coal image data set, and the ResNet embedded with MA block has the highest classification accuracy compared with the same-depth general and variant models. When analyzing the effect of embedding attention block on the training difficulty of different depth models, the results indicate that embedding attention block to the CNNs models only slightly improves the training time (per epoch). Specifically, in models (ResNet18, ResNet34, ResNet50, and ResNet101), embedding attention block only increases the training time (per epoch) of 1-2s, 4-9s, 3-9s, and 3-15s, respectively, which is relatively little compared with the training time of same-depth ResNet without attention block. At the same time, we notice that different attention blocks increase the training time differently, but in the overall trend, the MA block brings the highest training time (per epoch) increment. Therefore, the above analysis of the training time increases after embedding the attention block can provide guidance for the CNNs-based ore image classification model setting, which will help researchers apply the visual attention block into the existing model according to specific tasks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

TABLE 10. Different depth ResNet and ResNet with attention block evaluation in anthracite dataset including train accuracy, train loss, valid accuracy, valid loss, training time (per epoch), and convergent epoch

	Train Accuracy	Train Loss	Valid Accuracy	Valid Loss	Training Time	Convergent Epoch
ResNet18	94.76%	0.3745	86.44%	0.9162	303s	171
ResNet18_SE	93.83%	0.3684	86.58%	0.9064	306s	167
ResNet18_CA	93.37%	0.3728	86.55%	0.9022	306s	**165**
ResNet18_SA	94.97%	0.3581	86.42%	0.9158	306s	169
ResNet18_MA	94.65%	0.3597	**86.68%**	**0.9046**	307s	170
ResNet34	95.83%	0.3388	88.63%	0.6054	426s	141
ResNet34_SE	96.26%	0.3229	89.16%	0.5895	431s	**138**
ResNet34_CA	96.74%	0.3185	89.17%	0.5899	431s	142
ResNet34_SA	95.94%	0.3341	88.74%	0.5634	431s	139
ResNet34_MA	95.99%	0.3328	**89.28%**	**0.5242**	438s	**138**
ResNet50	96.25%	0.3257	90.65%	0.5242	586s	127
ResNet50_SE	96.85%	0.2978	91.17%	0.5029	590s	**123**
ResNet50_CA	96.17%	0.3048	91.03%	0.5086	591s	126
ResNet50_SA	95.79%	0.3404	90.85 %	0.5199	589s	124
ResNet50_MA	96.05%	0.3286	**91.74%**	**0.4895**	595s	127
ResNet101	96.63%	0.3249	92.68%	0.4751	945s	128
ResNet101_SE	97.18%	0.2684	93.50%	0.4626	959s	114
ResNet101_CA	97.48%	0.2481	93.47%	0.4637	949s	122
ResNet101_SA	95.89%	0.3689	92.65%	0.4765	952s	126
ResNet101_MA	96.82%	0.3057	**93.92%**	**0.3916**	962s	**113**

According to the evaluation results of ResNet in the anthracite data set, we find that the improvement of classification performance and the change of training time caused by embedding attention block are similar to that in gas coal and coking coal data sets. However, it is worth mentioning that the model classification accuracy in anthracite images is higher than that in gas coal and coking coal image datasets, which is determined by the apparent characteristics of anthracite particles, the feature extraction ability of ResNet, and the gain effect of attention blocks. Therefore, in this part, we focus on analyzing the effect of embedding attention block on the model convergent rate. Specifically, by observing the convergent epoch of different ResNet in anthracite dataset and combining with the model convergent rate in gas coal and coking coal image datasets, we find that embedding attention block to CNNs model can improve the model convergence rate to a certain extent. For instance, the ResNet18_ CA converges six epochs earlier than ResNet18, ResNet34_ SE converges three epochs earlier than ResNet34, ResNet50_ SE converges four epochs earlier than ResNet50, and ResNet101_MA converges 15 epochs earlier than ResNet101. At the same time, other ResNet that contain attention block also have faster convergence speed than those same-depth models without attention block. As a result, we conclude that the above results are related to the improvement of feature extraction ability caused by embedding attention block, which means the attention block can improve the CNNs model to extract better feature maps in mineral images with faster speed.

In summary, the classification performance of the different-depth ResNet and its variant models embedding with attention block for three types of coal images indicate that embedding attention block to the CNNs-based ore image classification models has the following three main influences. First of all, embedding attention block can effectively improve the model classification accuracy for mineral images, and the MA block has the highest improvement on the classification performance of the CNNs models, followed by SE block, CA block, and SA block. Secondly, the training time increment caused by embedding attention block is relatively low compared with the original training time. Finally, embedding attention block to the CNNs model can improve the model convergence speed to some extent.

3.3.2 Classification visualization with Grad-CAM

In order to better observe the classification weight distribution of the CNNs model embedding with attention blocks in the mineral image classification task and analyze the regions with higher classification weight, representing the sensitive regions of the input image, this experiment introduces the Gradient-weighted Class Activation Mapping (Grad-CAM) for model visualization [43]. In short, Grad-CAM technology uses the gradient information of the last convolution layer of the CNN model to assign importance to each neuron for specific attention decisions. Therefore, the primary purpose of Grad-CAM is to display the key parts that affect the classification decision, and the calculation formula of the class activation mapping is shown in EQ (13).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

$$L^c_{Grad-CAM} = ReLU\left(\sum_I \alpha_i^c A^i\right), \quad \alpha_i^c = \frac{1}{Z}\sum_{k=1}^{c_1}\sum_{j=1}^{c_2}\frac{\partial S_c}{\partial A^i_{kj}} \tag{13}$$

S_c——The Softmax score of c category
α_i^c——The classification weight of global pooling layer
c_1——The length of the input feature map
c_2——The width of the input feature map
A^i_{kj}——The pixel value of row k and column j of the i feature map

Taking the ResNet50 embedding with different attention blocks as examples, the Grad-CAM maps of the three types of coal samples are shown in FIG. 11., FIG. 12., and FIG. 13. In the Grad-CAM maps, red represents the high classification weight areas, and blue represents the low classification weight areas.

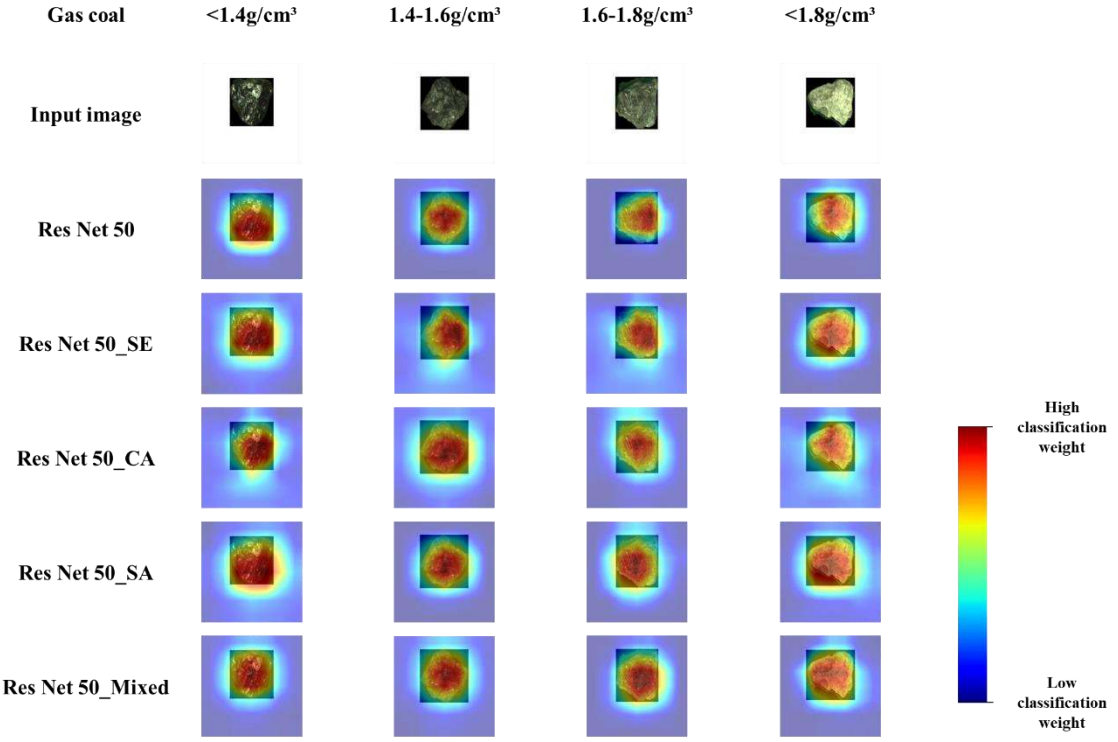

FIGURE 11. Grad-CAM maps of ResNet50 with different attention blocks for four-classes gas coal image classification

The visualization results indicate that on the whole trend, ResNet embedding different attention blocks have different classification weight distribution in the gas coal images, which means the attention blocks will affect the feature extraction process of the CNNs model. Firstly, after embedding the attention blocks to the models, the areas that contain classification weight in the image become relatively larger, which better covers the surface of the gas coal particle. Specifically, in the ResNet50_MA, the areas containing classification weight cover the surface of gas coal particles uniformly, and the weight of texture features inside gas coal particles is higher than that in general ResNet50. Meanwhile, ResNet50_SE and ResNet50_CA also show the same distribution change, but its coverage is relatively smaller than that in the ResNet50_MA. Secondly, in the ResNet50_SA, it can be observed that the distribution of classification weight is relatively uneven. As a result, combining the model evaluation results of different ResNet50 in the gas coal dataset, we can conclude that the classification weight distributions in Grad-CAM maps will reflect the feature extraction ability and classification performance of the different ResNet: that is, the ResNet50_MA has the best feature extraction ability, followed by ResNet50_SE, ResNet50_CA, ResNet50_SA, and general ResNet50.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

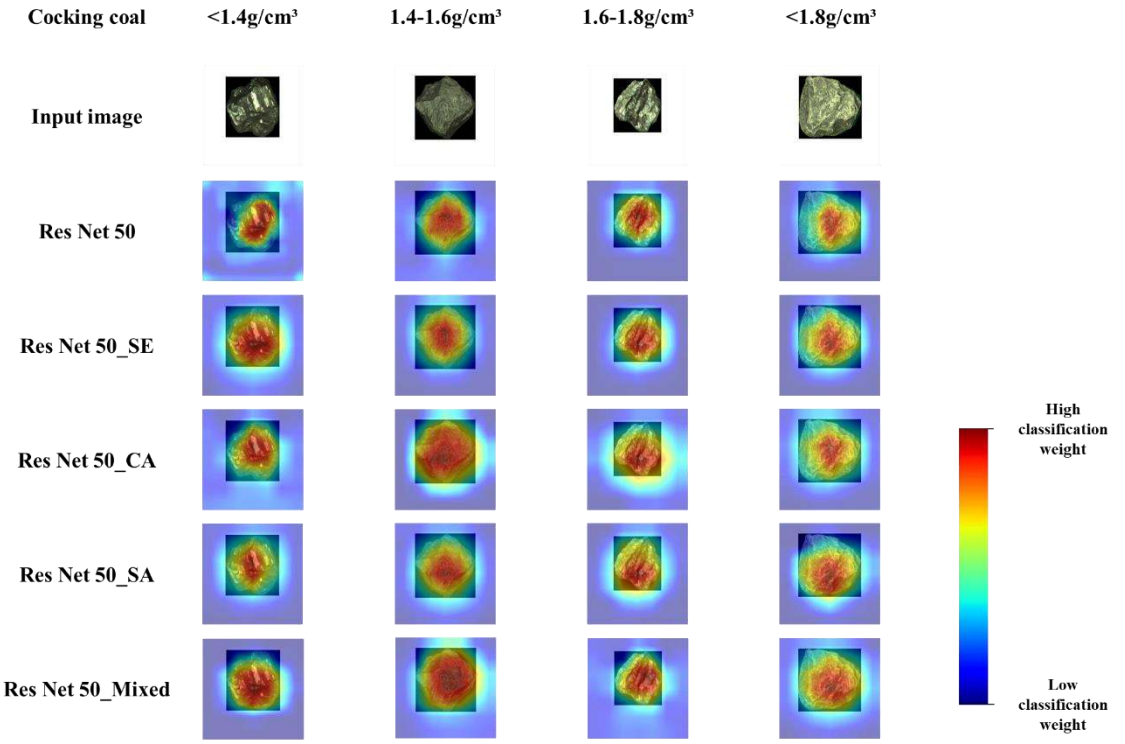

FIGURE 12. Grad-CAM maps of ResNet50 with different attention blocks for four-classes cocking coal image classification

In the visualization results of ResNet50 embedding with different attention blocks for four density coking coal particles, we found that embedding attention block to the CNNs models also affects the value of classification weight. Specifically, after embedding the attention blocks, the color depth of the red areas (high classification weight) in the general ResNet50 gradually becomes deeper, and the distribution range is larger, indicating that the classification weight of this area is relatively increased. Taking the Grad-CAM maps of <1.4g/cm^3 and 1.4-1.6g/cm^3 coking coal particles as examples, the color depth of red areas in the ResNet50_MA is deeper than that in general ResNet50. Therefore, the above results point out that embedding the attention block to CNNs model will affect the value of classification weight.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

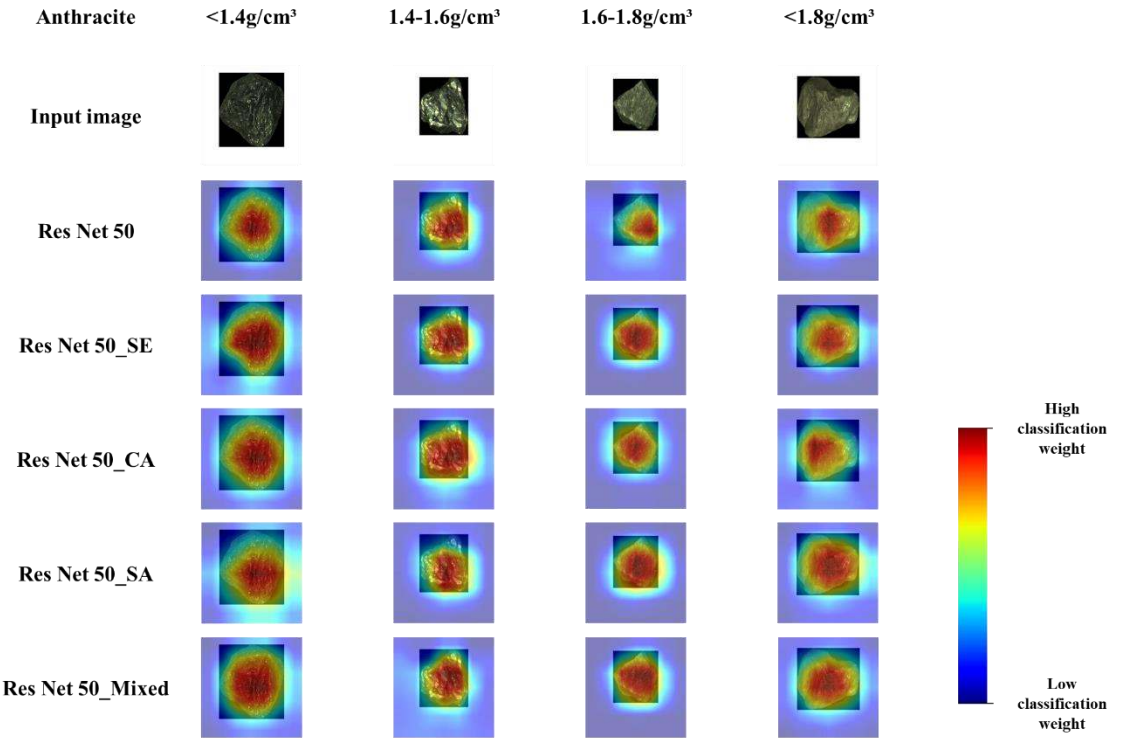

FIGURE 13. Grad-CAM maps of ResNet50 with different attention blocks for four-classes anthracite image classification

The Grad-CAM maps of ResNet50 embedding with different attention blocks in the anthracite dataset show the same trend as that in gas coal and coking coal datasets. In ResNet50_MA, ResNet50_SE, and ResNet50_CA, the regions that have classification weight relatively cover the whole particle surface. At the same time, the texture areas in the anthracite particle have a higher classification weight (dark red), the edge areas have a lower weight (green, cyan), and the background areas do not have classification weight (blue).

In summary, the Grad-CAM visualization results of the above-mentioned models indicate that embedding attention blocks to the CNNs model will affect the classification weight in two aspects: 1) Embedding the attention blocks will enlarge the areas that have classification weight in the input images; 2) Embedding the attention block will increase the value of classification weight.

3.3.3 Classification performance evaluation with confusion matrices

In order to better evaluate and compare the performance of attention blocks in the deep learning-based mineral image classification system, this experiment introduces the confusion matrices to reveal the classification performance of each model and the discrimination results of ResNet101 embedding with different attention blocks in four density anthracite image test set are selected, as shown in FIG. 14.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3095368, IEEE Access

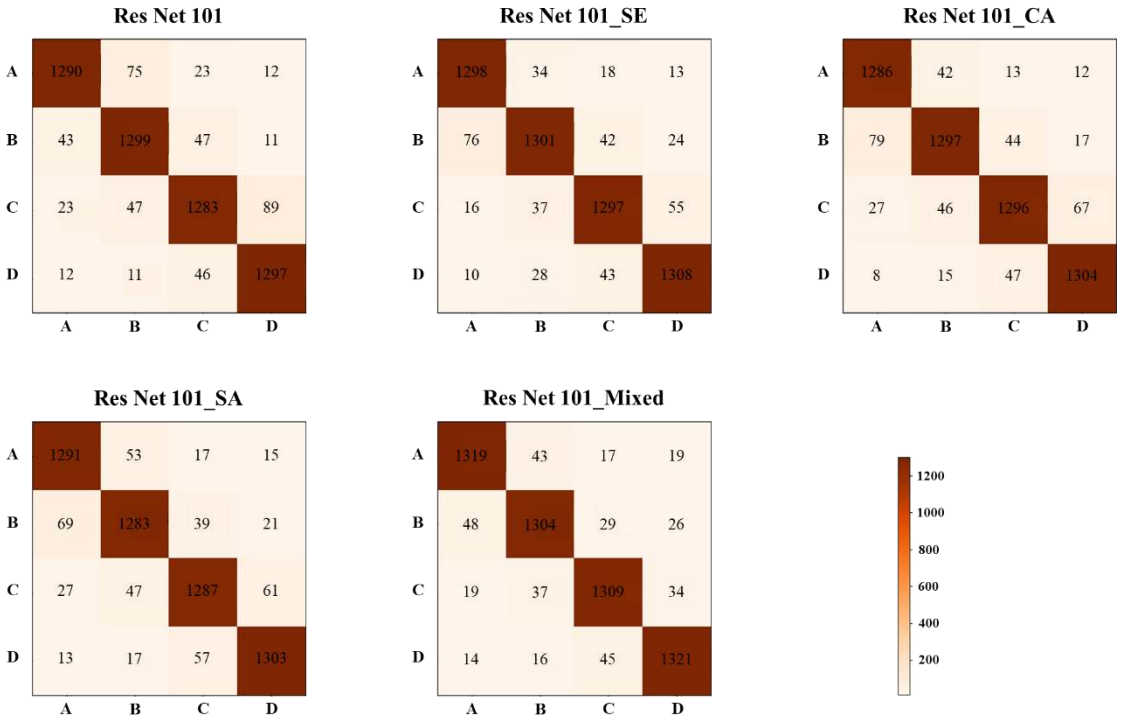

FIGURE 14. Confusion matrices of ResNet101 and ResNet101 embedded with different attention blocks in anthracite image classification task

Firstly, the discrimination results of each model indicate that embedding attention block to CNNs model can reduce the misjudgment to a certain extent. Specifically, in confusion matrices of the ResNet101, 439 anthracite images were misjudged, and the 1.6-1.8g/cm^3 category has the highest misjudgment rate, followed by >1.8g/cm^3 category and 1.4-1.6g/cm^3 category, and the lowest misjudgment was 85 for <1.4g/cm^3 gas coal images. In contrast, the misclassification rate of ResNet101 embedding with Attention blocks is lower, and the amounts of misjudged images of ResNet101_SE, ResNet101_CA, ResNet101_SA, and ResNet101_MA were 396, 417, 436, and 347, respectively. Secondly, in the comparison of different ResNet101, the results show that the performance of embedding MA block or SE block is better than that of embedding CA block or SA block to the model. Additionally, when analyzing the confusion matrices of ResNet101_MA, we find that the ResNet101_MA has the lowest misjudgment rate for >1.8g/cm^3 anthracite images, only 75, followed by 1.6-1.8g/cm^3 and <1.4g/cm^3 anthracite images, and the highest misjudgment rate for 1.4-1.6g/cm^3 anthracite images, which is 103. Therefore, the above result indicates that although the misjudgment rate of ResNet101 embedding with attention block is reduced, the misjudgment objects will change, which to some extent reflects the influence of the attention block to feature extraction.

In summary, the confusion matrices of general ResNet101 and ResNet101 embedding with different attention blocks indicate that embedding attention block to CNNs model can effectively reduce the misjudgment rate, that is, improve the model classification performance for mineral images, but the misjudgment objects of CNNs model will change accordingly.

Section 4. CONCLUSION& OUTLOOK

In order to solve the problems of low classification accuracy in multi-category mineral image classification tasks and low efficiency of mineral image feature extraction in CNNs models, combining with the visual attention mechanism, four construction strategies of the visual attention module are proposed, including the SE block, CA block, SA block, and MA attention block, and all of them can be flexibly embedded into the existing general CNNs models. Then, taking gas coal, coking coal, and anthracite as the experimental objects, and referring to different depth ResNet, this experiment builds various CNNs mineral image classification models embedding with different attention blocks and tests the improvement of classification accuracy and feature extraction ability. Then, the model classification weight distribution and classification ability are visualized by Grad-CAM and confusion matrices, respectively. The detailed conclusions are as follows.

(1) Firstly, embedding the attention blocks to the different-depth CNNs models can effectively improve the classification accuracy for mineral images, and the improvement value is 0.2%-1.3% (ResNet). Secondly, the MA block has the highest classification accuracy increment for each depth ResNet, followed by SE block and CA block, and SA block. Besides, embedding attention block to the different-depth ResNet will lead to 1-15s training time increment, which is relatively little compared to the original training time. Finally, embedding the

attention block to CNNs models can also improve the convergence speed in the training phase.

(2) The Grad-CAM visualization results of different ResNet for different types of mineral images show that embedding attention block to the CNNs model will affect the classification weight distribution and the value of classification weight, which means the attention blocks can effectively improve the feature extraction ability. In other words, the attention blocks can enhance the extraction of useful feature information while suppressing the feature information with less contribution, which will not only save computing power but also bring stable classification performance improvement.

(3) When predicting the test set images of anthracite particles, the confusion matrices of different ResNet101 point out that embedding attention block to CNNs model can effectively reduce the misjudgment rate, but the misjudgment objects will also change. Specifically, compared with the general ResNet101, the number of misjudgment images of ResNet101_SE, ResNet101_CA, ResNet101_SA, and ResNet101_MA by 43, 22, 3, and 92, respectively, but the models embedding with attention blocks show a higher misjudgment rate for 1.4-1.6g/cm^3 anthracite images.

In future experiments and research, we will further explore the application of visual attention mechanism in mineral image classification tasks and explore its potential application with other mineral image processing tasks, such as image segmentation, particle size estimation, and component prediction. Additionally, the construction and embedding strategies of visual attention modules are still one of our research centers.

REFERENCES

[1] H. Knapp, K. Neubert, C. Schropp, and H. Wotruba, "Viable Applications of Sensor-Based Sorting for the Processing of Mineral Resources," *ChemBioEng Rev.*, vol. 1, no. 3, pp. 86–95, 2014, doi: 10.1002/cben.201400011.

[2] C. Robben, P. Condori, A. Pinto, R. Machaca, and A. Takala, "X-ray-transmission based ore sorting at the San Rafael tin mine," *Miner. Eng.*, vol. 145, p. 105870, 2020, doi: 10.1016/j.mineng.2019.105870.

[3] C. Robben and H. Wotruba, "Sensor-based ore sorting technology in mining—past, present and future," *Minerals*, vol. 9, no. 9, p. 523, 2019, doi: 10.3390/min9090523.

[4] C. Robben, J. De Korte, H. Wotruba, and M. Robben, "Experiences in dry coarse coal separation using X-ray-transmission-based sorting," *Int. J. Coal Prep. Util.*, vol. 34, no. 3–4, pp. 210–219, 2014, doi: 10.1080/19392699.2014.869938.

[5] J. Kolacz, "Advanced separation technologies for pre-concentration of metal ores and the additional process control," *E3S Web Conf.*, vol. 18, p. 1001, 2017, doi: 10.1051/e3sconf/201712301001.

[6] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/bf00058655.

[7] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.*, vol. 40, no. 3. Belmont, Calif.: Wadsworth International Group, 1984.

[8] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, 2007, doi: 10.1128/AEM.00062-07.

[9] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[10] C. Corinna and V. Vladimir, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[11] Z. Zelin, Z. Zhiwei, W. Lei, L. Yang, X. Xiangdong, and H. Qi, "Deep Learning-based Image Classification of Gas Coal," *Int. J. Glob. Energy Issues*, vol. In Press, 2020.

[12] Z. Zhang, Y. Liu, Q. Hu, Z. Zhang, and Y. Liu, "Competitive Voting-based Multi-class Prediction for Ore Selection," in *IEEE International Conference on Automation Science and Engineering*, 2020, vol. 2020-Augus, pp. 514–519, doi: 10.1109/CASE48305.2020.9217017.

[13] Z. Zhang *et al.*, "Multi-information online detection of coal quality based on machinevision," *Powder Technol.*, vol. 374, pp. 250–262, 2020, doi: 10.1016/j.powtec.2020.07.040.

[14] A. K. Patel, S. Chatterjee, and A. K. Gorai, "Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades," *Earth Sci. Informatics*, vol. 12, no. 2, pp. 197–210, 2019, doi: 10.1007/s12145-018-0370-6.

[15] M. Massinaei, A. Jahedsaravani, E. Taheri, and J. Khalilpour, "Machine vision based monitoring and analysis of a coal column flotation circuit," *Powder Technol.*, vol. 343, no. 5, pp. 330–341, 2019, doi: 10.1016/j.powtec.2018.11.056.

[16] F. Khorram, A. H. Morshedy, H. Memarian, B. Tokhmechi, and H. S. Zadeh, "Lithological classification and chemical component estimation based on the visual features of crushed rock samples," *Arab. J. Geosci.*, vol. 10, no. 15, pp. 1–9, 2017, doi: 10.1007/s12517-017-3116-8.

[17] F. J. Galdames, C. A. Perez, P. A. Estévez, and M. Adams, "Rock lithological classification by hyperspectral, range 3D and color images," *Chemom. Intell. Lab. Syst.*, vol. 189, pp. 138–148, 2019, doi: 10.1016/j.chemolab.2019.04.006.

[18] K. Itano, K. Ueki, T. Iizuka, and T. Kuwatani, "Geochemical discrimination of monazite source rock based on machine learning techniques and multinomial logistic regression analysis," *Geosci.*, vol. 10, no. 2, p. 63, 2020, doi: 10.3390/geosciences10020063.

[19] D. Hasterok, M. Gard, C. M. B. Bishop, and D. Kelsey, "Chemical identification of metamorphic protoliths using machine learning methods," *Comput. Geosci.*, vol. 132, pp. 56–68, 2019, doi: 10.1016/j.cageo.2019.07.004.

[20] X. Chen, S. Wang, C. Shi, H. Wu, J. Zhao, and J. Fu, "Robust ship tracking via multi-view learning and sparse representation," *J. Navig.*, vol. 72, no. 1, pp. 176–192, 2019, doi: 10.1017/S0373463318000504.

[21] J. Tang, F. Gao, F. Liu, and X. Chen, "A Denoising Scheme-Based Traffic Flow Prediction Model: Combination of Ensemble Empirical Mode Decomposition and Fuzzy C-Means Neural Network," *IEEE Access*, vol. 8, pp. 11546–11559, 2020, doi: 10.1109/ACCESS.2020.2964070.

[22] X. Chen *et al.*, "Ship type recognition via a coarse-to-fine cascaded convolution neural network," *J. Navig.*, vol. 73, no. 4, pp. 813–832, 2020, doi: 10.1017/S0373463319000900.

[23] Y. Fu and C. Aldrich, "Quantitative Ore Texture Analysis with Convolutional Neural Networks," in *IFAC-PapersOnLine*, 2019, vol. 52, no. 14, pp. 99–104, doi: 10.1016/j.ifacol.2019.09.171.

[24] S. Zhu, W. Yang, G. Hou, B. Lu, and S. Wei, "An intelligent classification and recognition method of rock thin section.," *Acta Petrol. Sin.*, vol. 40, no. 1, p. 106, 2020.

[25] O. Sudakov, E. Burnaev, and D. Koroteev, "Driving digital

rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks," *Comput. Geosci.*, vol. 127, pp. 91–98, 2019, doi: 10.1016/j.cageo.2019.02.002.

[26] B. LP, L. RPD, B. A, C. DD, and N. C, "Deep convolutional neural networks as a geological image classification tool," *Sediment. Rec*, vol. 17, pp. 4–9, 2019, doi: 10.2110/sedred.2019.2.4.

[27] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, "Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms," *Sensors (Switzerland)*, 2019, doi: 10.3390/s19183914.

[28] E. E. Baraboshkin *et al.*, "Deep convolutions for in-depth automated rock typing," *arXiv*, vol. 135, p. 104330, 2019, doi: 10.1016/j.cageo.2019.104330.

[29] L. Si, X. Xiong, Z. Wang, and C. Tan, "A Deep Convolutional Neural Network Model for Intelligent Discrimination between Coal and Rocks in Coal Mining Face," *Math. Probl. Eng.*, vol. 2020, pp. 1–12, 2020, doi: 10.1155/2020/2616510.

[30] H. Hong, L. Zheng, J. Zhu, S. Pan, and K. Zhou, "Automatic recognition of coal and gangue based on convolution neural network," *arXiv*, 2017.

[31] J. C. Á. Iglesias, R. B. M. Santos, and S. Paciornik, "Deep learning discrimination of quartz and resin in optical microscopy images of minerals," *Miner. Eng.*, vol. 138, pp. 79–85, 2019, doi: 10.1016/j.mineng.2019.04.032.

[32] W. Fei *et al.*, "Residual attention network for image classification," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, vol. 3156–3164, 2017.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[34] S. Xu and Y. Zhou, "Artificial intelligence identification of ore minerals under microscope based on deep learning algorithm.," *Acta Petrol. Sin.*, vol. 34, no. 11, pp. 3244–3252, 2018, doi: CNKI:SUN:YSXB.0.2018-11-010.

[35] Z. C. Horn, L. Auret, J. T. McCoy, C. Aldrich, and B. M. Herbst, "Performance of Convolutional Neural Networks for Feature Extraction in Froth Flotation Sensing," *IFAC-PapersOnLine*, vol. 50, no. 2, pp. 13–18, 2017, doi: 10.1016/j.ifacol.2017.12.003.

[36] L. Hashemi-Beni and A. Gebrehiwot, "Deep Learning for Remote Sensing Image Classification for Agriculture Applications," *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLIV-M-2–2, pp. 51–54, 2020, doi: 10.5194/isprs-archives-xliv-m-2-2020-51-2020.

[37] N. Dey, *Classification Techniques for Medical Image Analysis and Computer Aided Diagnosis*, 1st editio. 2019.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, 2015.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[41] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2017, doi: 10.1109/TPAMI.2019.2913372.

[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11211 LNCS, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.