Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient Residual Learning With Multi-scale Attention for Robust Coal-rock Image Classification

Tongshuai Yu
Shandong Huayu Institute of Technology
Sheng Wang
Shandong Huayu Institute of Technology
Yongli Chen
Linyi University
Dong Zhu
zhudong0511@163.com

Chengdu Ginkgo Hotel Management College

Research Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11031992/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Efficient Residual Learning With Multi-scale Attention for Robust Coal-rock Image Classification

Tongshuai Yu[1], Sheng Wang[1], Yongli Chen[2], and Dong Zhu[3,*]

[1]Shandong Huayu Institute of Technology, School of Information Engineering, Shandong, 253000, China
[2]Linyi University, School of Information Science and Engineering, Linyi, 276000, China
[3]Chengdu Ginkgo Hotel Management College, School of Information Engineering Management, Chengdu, 611743, China
[*]zhudong0511@163.com

ABSTRACT

Robust coal-rock image classification is challenging because visual differences among coal, gangue, limestone and sandstone are often subtle and can be obscured by lighting variation, dust, humidity, sensor noise and complex texture similarity in mining environments. This study proposes an efficient multi-scale attention residual learning framework for image classification, aiming to improve the balance among classification accuracy, parameter efficiency, convergence speed and cross-domain robustness. The proposed network introduces a residual convolutional block that integrates efficient multi-scale attention for spatial and channel-wise feature enhancement, while replacing conventional convolutions with depthwise over-parameterised convolution during training to improve convergence without increasing inference complexity. The model is evaluated on three public natural image benchmarks, CIFAR-10, CIFAR-100 and CINIC-10, and on a proprietary four-class coal-rock dataset. Experimental results show that the proposed method achieves competitive performance on standard benchmarks while maintaining a moderate parameter scale. On CIFAR-100, it improves classification accuracy over several attention-enhanced residual baselines, and on COALSet-4 it obtains a substantial improvement compared with the corresponding residual baseline, indicating stronger adaptability to fine-grained industrial visual patterns. Ablation experiments further suggest that both the multi-scale attention component and the modified convolutional operation contribute to accuracy gains and faster convergence. Visual inspection of attention regions indicates that the model can focus on discriminative local texture structures. The results demonstrate the potential of efficient multi-scale residual learning for coal-rock recognition and related industrial image classification tasks. Our source code is available at: 10.5281/zenodo.20841285.

Introduction

Image classification[1] is a fundamental task in computer vision, aimed at assigning input images to specific categories. Recent advancements in deep learning have significantly advanced image classification, enabling applications across various domains[2–4]. Convolutional neural networks (CNNs) have proven to be powerful tools for feature extraction, capable of autonomously learning multi-level image representations, thereby achieving exceptional performance in classification tasks across diverse fields. Deep learning techniques, in particular, have demonstrated remarkable effectiveness in handling tasks that involve complex features and large datasets, such as natural scene classification[5], medical image analysis[6], and remote sensing image classification[7].

In the field of image classification, a clear positive correlation exists between a model's computational complexity and its classification accuracy. High-precision models typically require longer convergence times, while models with fewer parameters often exhibit concerning classification performance. Constructing models that combine low computational complexity with outstanding classification performance remains a significant challenge. Furthermore, existing models are optimised only for specific image domains. Another objective of this research is to enhance model robustness by improving performance on coal rock data while maintaining high classification accuracy on natural images. We comprehensively validated the model's robustness by applying it to both natural and coal rock images[8].

Unlike the more distinct visual differences observed in natural images, the distinctions between coal and rock can be subtle. Although the materials have notable differences in their optical and physical properties, these variations are often obscured in practical images due to environmental factors such as lighting, dust, and humidity. Additionally, the textures and structures of coal and rock are complex, with considerable potential for textural similarity between the two materials. For example, high-resolution images may reveal microscopic structures that closely resemble each other, making traditional image processing techniques insufficient for effective differentiation. This challenge is further exacerbated by significant noise in coal and rock images, often caused by equipment vibrations in mining environments, limitations in image capture devices, and external

environmental influences.

In recent years, deep learning-based image classification methods have proliferated, aiming to extract rich features from images. As noted in references[9–11], some approaches focus on increasing model depth and complexity to enhance their ability to distinguish specific features in coal rock images. However, such expansions often lead to increased model parameters and computational demands. Whilst deploying such models on classification devices for natural images poses no significant difficulty, implementing these complex deep learning models within coal mining environments may be constrained by the computational capabilities of underground mobile equipment. Furthermore, models may exhibit performance degradation when confronted with the variable conditions of coal mines, exposing deficiencies in their generalisation capabilities. Similarly, existing models demonstrate marked discrepancies in distinguishing between coal and rock images versus natural images, underscoring the urgent need to enhance model robustness.

To address these challenges, we propose a deformable residual block named DoMaBlock. This innovation significantly enhances local feature extraction capabilities, accelerates convergence, substantially improves accuracy, and bolsters model robustness, all while requiring only a modest increase in baseline model parameters. Furthermore, we have developed a novel convolutional neural network architecture, DoMaResNet, whose experimental results demonstrate superior performance over existing methods.

The principal contributions of this paper are outlined as follows:

- This study introduces a novel residual convolution module, termed DoMaBlock, which combines multi-scale processing with computational efficiency and rapid convergence. Additionally, its design allows it to be seamlessly integrated as an insertable block into other residual network architectures.
- Based on our proposed multi-scale grouped attention for texture-sensitive classification, over-parameterised training for faster convergence, we present a new residual network-DoMaResNet.
- Extensive experiments were conducted to evaluate the performance of DoMaResNet against other existing CNN models in image classification. The results demonstrate that DoMaResNet outperforms the others, with superior accuracy, computational efficiency, robustness, and rapid convergence.

Related works

Visual Attention

Convolutional neural networks (CNNs) have made significant strides in visual tasks within deep learning. However, excessively deep networks are prone to partial feature loss during feature propagation and come with substantial computational costs. To address these challenges, various visual attention mechanisms have been proposed and successfully integrated into CNNs. Hu et al.[12] introduced the Squeeze-and-Excitation (SE) network, which incorporates global contextual information to enhance inter-channel correlations. The SE module recalibrates channel responses by mapping input features to global channel weights, allowing the network to assign varying weights to different channels and thereby improving the representation of specific features. Woo et al.[13] proposed a more flexible attention mechanism, the Convolutional Block Attention Module (CBAM), which considers both channel and spatial attention. By capturing feature importance across both dimensions, CBAM further enhances the expressive power of the network. Coordinate Attention (CA)[14] combines coordinate information with channel attention to capture cross-spatial dependencies more precisely. Unlike SE and CBAM, which rely on global pooling operations, CA introduces relative spatial positional information while preserving global context, enabling more effective long-range dependency modeling. More recently, Ouyang et al.[15] argued that modeling cross-channel relationships through channel dimensionality reduction may introduce side effects when extracting deep visual representations. To address this, they proposed a novel Efficient Multi-scale Attention (EMA) module. EMA reconfigures part of the channel dimension into a batch dimension and groups the channel dimension into multiple sub-features, ensuring that spatial semantic features are uniformly distributed within each feature group. In 2025, Zu et al.[61] introduced the RESwinT model for pollen image classification. This model combines parallel window Transformer modules with a coordinate attention mechanism to capture both global contextual features and salient features of pollen.

Computational Complexity

The performance of CNNs is heavily dependent on their architecture. Over the course of deep learning's evolution, several influential architectures have been proposed, such as AlexNet[16], VGG[17], GoogLeNet[18], ResNet[19], and MobileNets[20]. The goal of this study is to integrate our approach with these established architectures to achieve superior performance without altering their overall framework. Like many contemporary approaches, our work enhances feature extraction capabilities by adjusting the network's depth and breadth, rather than modifying the architecture itself. This adjustment leads to improved task performance, as seen in[21,22]. However, as demonstrated in[23], increasing network depth leads to exponential growth in computational

cost, which reduces training efficiency. In recent years, numerous approaches have explored variations of conventional convolutions to improve model performance[24–27]. Our work shares a similar objective—achieving substantial performance gains with minimal architectural changes. Additionally, many recent approaches have focused on model sparsification to reduce computational load and accelerate training. Huang et al.[28] introduced a pioneering 'drop-and-grow' strategy, which progressively augments non-zero weights to maintain both high sparsity and accuracy. While such approaches offer favorable sparsity, reduced parameter counts, and faster training speeds, maintaining classification accuracy remains a challenge. In 2025, Zhang et al.[60] designed an improved residual network that combines pyramid-style and depth-wise overparameterised convolutions for fine-grained hardware classification. This research achieved exceptional performance without increasing computational complexity. Targeting fine-grained visual classification difficulties caused by subtle inter-class gaps and large intra-class variations, Wang et al.[63] proposes CoDeR, a parameter-free regularization module. It leverages class prototypes to decouple feature representations and enhance inter-class distinction, and parallel multi-module deployment fixes delayed shallow representation updates. Validated on natural, medical and semi-supervised datasets, it shows robust cross-domain generalization. This study focuses on enhancing task performance by controlling model parameters and accelerating training—key areas of ongoing research.

The Development of Deep Learning in the Field of Coal and Rock Image Processing

In recent years, research on coal gangue classification has advanced rapidly. Current classification methods primarily focus on X-ray, gamma-ray, image monitoring, radar detection, and vibration detection techniques[52]. The classification principles of the first two methods rely heavily on the primary constituent elements of coal and gangue. However, both radiographic detection methods pose radiation risks, presenting potential hazards to human health[53]. With advances in computer vision technology, coal-rock image classification methods utilising deep learning have been extensively explored.[54,55] developed a coal gangue recognition model employing support vector machines, random forests, and feedforward neural networks to extract multidimensional features from the grey-scale textures of coal gangue images. Furthermore, the literature[56] employed an enhanced LeNet-5 deep neural network to train on coal gangue samples, yielding a deep learning model for coal gangue classification. However, direct acquisition of coal gangue images is susceptible to environmental interference, resulting in relatively low classification accuracy for deep learning models. To address these challenges, work in[57] proposed a CNN-based coal-rock identification method. This approach utilises hyperspectral data, pre-processing the reflectance spectral data of coal-rock samples. Through a one-dimensional convolutional neural network model, it achieved high classification performance. In 2022, Dong et al.[58] proposed a coal identification model integrating convolutional neural networks with extreme learning machines. This provided a low-cost, convenient, and effective method for rapid on-site coal identification. To tackle complex lighting, target occlusion and small target missed detection in underground coal mines, Liu et al.[64] proposes YOLOv8-HAC. It integrates HAC-Net, AGC-Block and a tiny target detection layer. The model achieves 94.8% mAP and 90.4% recall, outperforming mainstream detectors for underground helmet detection. Dong et al.[62] proposes an integrated aboveground-underground modeling approach for open pit mines. It adopts oblique photography for surface reconstruction and a novel surface-stitching algorithm for underground geological modeling. Comparative experiments verify its better model integrity, stability and efficiency over ball-pivoting and Poisson algorithms. The integrated model further supports mining planning systems to boost mine management efficiency. Although previous studies have investigated recognition and classification algorithms, none achieved the desired accuracy.

Method

In this section, we revisit the fundamental principles of EMA and Do-conv, and present the design of a novel convolutional residual block, termed DoMaBlock, which is constructed based on both EMA and Do-conv. Compared to the PreActBlock[29], the DoMaBlock demonstrates several advantages, including faster convergence, improved generalization capabilities, and the ability to capture multi-scale features. Additionally, recognizing that Do-conv[30] accelerates network convergence without increasing the number of parameters, we replace all standard convolutions within the network with Do-conv.

EMA

As demonstrated in[23], the receptive field of convolutional neural networks expands in proportion to increasing network depth. However, excessively deep networks have been shown to require a significant investment of time and labour during the training process. EMA employs a parallel word structure with a view to circumventing the issue under discussion. The overall architecture of EMA is illustrated in Fig 1(left). For input X, the EMA partitions X into G sub-feature groups, with each group acquiring distinct semantic representations. Furthermore, EMA employs three sub-parallel structures to extract attention weights for grouped feature maps (two 1×1 branches and one 3×3 branch). The two 1×1 branches execute average pooling operations across two spatial dimensions to acquire channel information, while the 3×3 branch extracts multi-scale

features. This approach facilitates enhanced adjustment of channel weights by EMA, thereby ensuring the preservation of spatial information deemed crucial within the channels.

Do-conv

Convolutional layers represent a fundamental component of convolutional neural networks, playing a pivotal role in the extraction of image features. In the course of the development of deep learning, a considerable number of alternative approaches to convolutional layers have been proposed with a view to enhancing a network's capacity for feature extraction[31,32]. One such method is Do-conv. As demonstrated in Fig 1(right), DO-Conv is a deep convolutional layer designed to enhance the performance of convolutional layers by employing distinct 2D convolutional kernels on each input channel. This combination of deep convolutional layers constitutes overparameterisation, as it increases learnable parameters whilst the resulting linear operation can be represented by a single convolutional layer. DO-Conv replaces conventional convolutional layers during training, thereby enhancing CNN performance. At inference time, the deep convolutional layer folds into a traditional convolutional layer, maintaining identical computational complexity. DO-Conv thus introduces performance gains without increasing inference computational complexity.

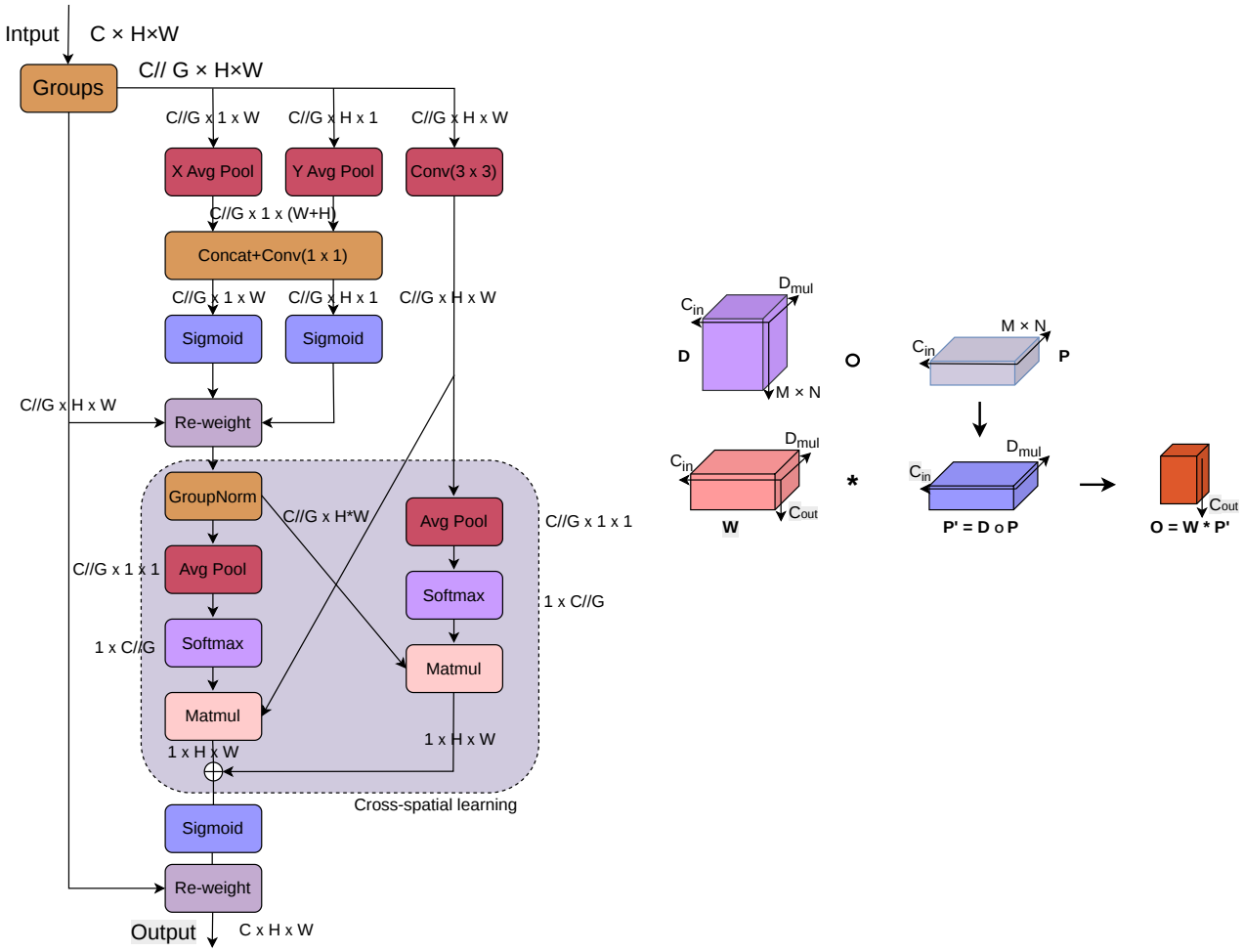


Figure 1. The overall architecture of EMA (left) and Do-conv (right). In the proposed EMA module, "g" means the divided groups, "X Avg Pool" represent the 1D horizontal global pooling and "Y Avg Pool" indicates the 1D vertical global pooling, respectively. In Do-conv, $M \times N$ denotes the spatial dimension of input tensor P, D_{mul} stands for the learnable kernel of depth-wise convolution, and W represents the learnable kernel of standard convolution.

DoMaBlock

The objective of this study was to establish a convolutional residual module that is computationally faster, converges more readily, extracts more comprehensive features, and exhibits greater robustness. To this end, DoMaBlock was created, which is based on EMA and Do-conv, as illustrated in the figure below. DoMaBlock is primarily implemented in four parts. Initially, the input feature map X traverses the first Do-conv, thereby yielding a feature map. The operational formula for Do-conv is as Eq. (1) shows:

$$y_1 = W * (D \circ X) \tag{1}$$

Given an input feature map, a convolutional layer processes it in a sliding window fashion, applying a set of convolution kernels to a patch of corresponding size, at each window position. For the purposes of our exposition, it is convenient to think of a patch as a 2D tensor $P \in R^{(h \times w) \times C_{in}}$, where h and w are the spatial dimensions of the patch. Where $W \in R^{C_{out} \times D_{mul} \times C_{in}}$ represents the trainable kernel of a conventional convolution, and $D \in R^{h \times w \times D_{mul} \times C_{in}}$ denotes a trainable kernel for depth-wise convolutions. h and w denote the spatial dimensions of X, D_{mul} denotes the depth multiplier in depthwise convolution, $*$ denotes the standard convolution operator, and $\circ$ denotes the depthwise convolution operator. Subsequently, y_1 undergoes batch normalisation, an activation function, and a Dropout layer before inputting to the second Do-conv layer to produce the output feature map y_2. y_2 captures pixel-level pairing relationships via the EMA mechanism while emphasising the global context of all pixels. Specifically, the feature map $y_2 \in R^{C \times h \times w}$ first passes through a feature grouping module. Within this module, EMA partitions y_2 along the channel dimension into G sub-features to learn distinct semantic information. The partitioned features can be represented as $y_2 = [y_{2_0}, y_{2_1}, y_{2_2}, \ldots, y_{2_{G-2}}, y_{2_{G-1}}]$, where $y_{2_i} \in R^{\frac{C}{G} \times h \times w}$. Subsequently, y_{2_i} is fed into three parallel sub-networks to extract attention weights for grouped feature maps, comprising two 1×1 branches and one 3×3 branch. To capture inter-channel dependencies while reducing computational burden, EMA performs $1D$ global average pooling across both directions of the feature map on each 1×1 branch. The two pooled feature encodings are concatenated and decomposed into two feature vectors via a shared 1×1 convolution, fitting a two-dimensional binomial distribution after linear convolution using two Sigmoid functions. Subsequently, the two channel attention maps within each group are aggregated. Concurrently, the 3×3 branch captures local cross-channel interaction features via 3×3 convolutions. After obtaining outputs from both branches, the 1×1 branch's output undergoes global average pooling. Prior to the Softmax activation function, the pooled outputs are reshaped to their respective dimensionality. The formula for $global pooling$ is expressed as Eq. (2) shows:

$$Z_c = \frac{1}{h \times w} \sum_{j}^{h} \sum_{i}^{w} y_{2_i} \tag{2}$$

The initial spatial attention map is derived by multiplying the matrix dot product of the $Softmax$ activation function with the 3×3 branch following the 3×3 convolution. In the 3×3 branch, the Exponential Moving Average (ExpMA) is subject to global average pooling and Softmax activation, following which it performs matrix dot product multiplication with the channel attention map that has been aggregated from the two 1×1 branches. This process yields the second spatial attention map. Finally, the output features within each group are mapped to a set comprising two spatial attention weights. The sigmoid function captures pixel-level pairwise relationships while highlighting the global context of all pixels. It is important to note that the feature map processed by the ExpMA retains the same dimensions as the original feature map.

DoMaResNet

Building upon ParaLkResNet[23], the original PreActBlock was replaced by DoMaBlock, creating a new convolutional neural network framework named DoMaResNet. Its overall architecture is shown in Fig 2. Similar to ParaLkResNet, DoMaResNet also has two variants: DoMaResNet-18 and DoMaResNet-34, differing in the number of DoMaBlocks used. To evaluate network performance while balancing parameter count and computational complexity, we statistically analyzed the parameter growth and floating-point computation growth of the baseline network and the proposed method. The comparison results are shown in Tab 1. Integrating the Do-conv and EMA modules into the baseline model resulted in a slight increase in parameter count and floating-point computation. However, experimental results confirmed the effectiveness of the proposed method, showing improved accuracy in image classification tasks compared to the baseline.

Experiment

In this section, we compare our approach with existing image classification methods using the CIFAR-10[33], CIFAR-100[33], CINIC-10[34] and COALSet-4 datasets. In all the classification experiments on the datasets in this chapter, we have used the average of five model training runs as the final classification accuracy.

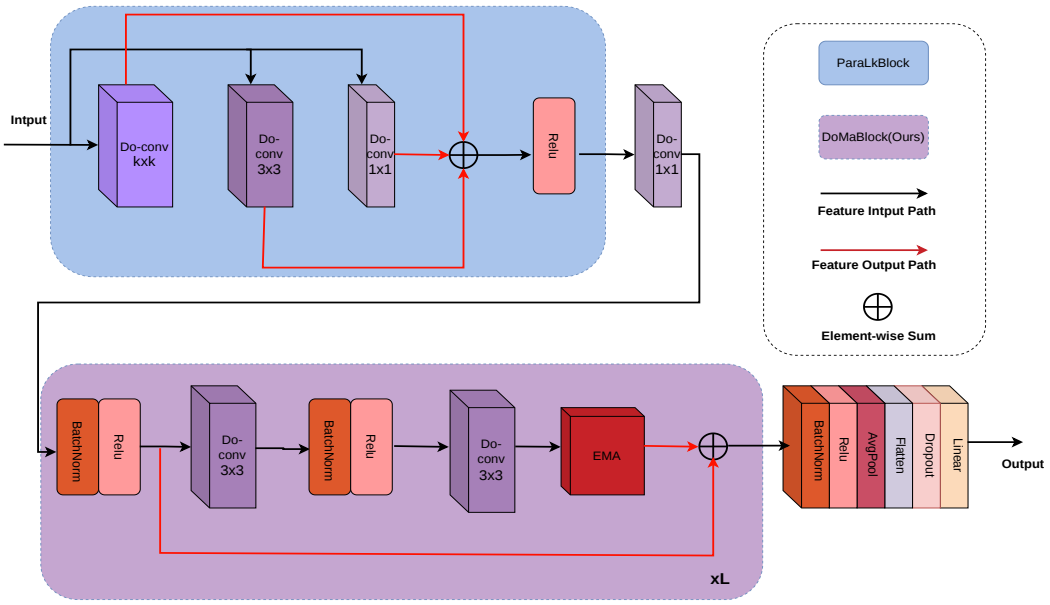


Figure 2. DoMaResNet overall architecture. In the DoMaBlock module, the input feature maps are first preprocessed by BatchNorm and ReLU and then split into two branches. The main branch utilizes two successive 3×3 Do-conv layers to hierarchically extract local multi-scale spatial features. Batch normalization and activation functions are embedded between the two convolution layers to strengthen nonlinear representation. The convolutional outputs are fed into the efficient multi-scale EMA attention module to implement adaptive feature weighting. The other branch is a residual shortcut that directly delivers the original preprocessed features, and the features from the two branches are fused via element-wise summation. This architecture adopts the design philosophy of "dual depthwise convolution for front-end feature extraction and EMA attention for rear-end feature enhancement". It first lightweightly excavates abundant low-level texture features, then leverages multi-scale attention to filter and highlight crucial informative features. Combined with residual connections to ensure stable gradient propagation, the module improves the model's feature representation capability with controllable computational overhead.

DataSet

We evaluated the effectiveness of classification networks using three publicly available datasets (CIFAR-10, CIFAR-100 and CINIC-10) and one private coal rock dataset (COALSet-4).

The CIFAR-10 dataset comprises 60,000 colour images measuring 32×32 pixels, of which 50,000 are used for training and 10,000 for testing. The dataset contains ten categories: aeroplane, car, bird, cat, deer, dog, frog, horse, ship and truck.

CIFAR-100 comprises 100 categories, each containing 600 colour images sized 32×32 pixels, which are further subdivided into 500 for training and 100 for testing. This dataset totals 60,000 images, with 600 images per category.

The CINIC-10 dataset comprises 270,000 images of size 32×32, divided into 90,000 for training, 90,000 for validation, and 90,000 for testing. We combined the original validation and training sets to create a large new training set for our model. CINIC-10 comprises 270,000 32×32 images divided into 90,000 for training, 90,000 for validation, and 90,000 for testing. We combined the original validation and training sets to create a large training set for our model.

COALSet-4 comprises image data collected through field surveys and from major websites. In May 2025, we visited three mines, including the Burtay Coal Mine in Ulanmuren Town, Yijinhuoluo Banner, Ordos City, Inner Mongolia, to collect coal rock image data. The raw images were collected from real scene in the mine, using high-speed industrial cameras. The acquisition conditions closely mirror real-world mining environments, capturing variations in illumination, dust occlusion, and humidity. The original images have a resolution distribution ranging from 800×600 to 1920×1080 pixels.After quality screening of all the data, we selected approximately 80 high-quality images to form the dataset. Combined with coal and rock images from major websites, this constitutes the COALSet-4 dataset(The various image features are shown in 3). COALSet-4 comprises four image categories: coal, coal gangue, limestone and sandstone (where coal and coal gangue data originate from

Method	Do-conv	EMA	#.Param.growth	Flops growth
ParaLkResNet18	✓	-	+0.28M	+0.10G
	-	✓	+0.29M	+0.53G
	✓	✓	+0.57M	+0.63G

Table 1. ParaLkResNet18 parameter changes after adding two methods

Figure	Category	Color	Characteristic
	Black coal	Black	Lumpy, Granular
	Coal gangue	Dark gray	Nodules, Granular, Massive
	Limestone	Gray	Fragmentation, Grain structure
	Sandstone	Hazel	Granular

Figure 3. Example images of various rock images in COALSet-4

both field collection and websites, whilst limestone and sandstone data are sourced exclusively from websites). Each class contains 130 images, bringing the total number of images in the dataset to 520. Due to the varying sources of the image data, there are some differences in image dimensions; we have standardised all images to a size of 224×224. Furthermore, as the dataset is relatively small, Given the relatively small scale of the dataset, explicit measures were taken to mitigate data leakage and overfitting risks. First, a rigorous duplicate removal procedure was implemented using Structural Similarity Index (SSIM)] to eliminate highly identical frames caused by high-speed continuous shooting. Most crucially, we manually verified the dataset to ensure that images containing the same physical object or captured from the exact same temporal scene clip strictly do not appear across different data splits. We split the dataset into a training set and a test set in a 7:3 ratio. To avoid the random effects that small-sample datasets can have on model training, as shown in 4, We then applied six types of data augmentation to the training set, these strategies include random horizontal and vertical flipping, random rotation , color jittering (randomly adjusting brightness, contrast, and saturation by 20% to simulate lighting variations in mines), random cropping and so on, resulting in a processed training set of 2,184 images, It is worth noting that the division between the training and test sets is an even split, with a $1:1$ ratio of each image type in both sets. Our test set is kept in a pristine state at all times. Furthermore, during model training, we monitor the validation set loss in real time; if the values do not decrease for five consecutive rounds, model training is halted to mitigate the risk of overfitting associated with small datasets.

Implementation details

For the classification tasks on CIFAR-10/100 and CINIC-10, colour images measuring $32 \times 32 \times 3$ pixels are used as input, with the training set being augmented through random cropping and flipping. Current data augmentation techniques can act as a form of regularisation, improving model generalisation by introducing extra learning and constraints beyond the model parameters. The training framework is implemented in PyTorch, and the network is trained using the SGD optimiser. The learning rate is adjusted using the MultiStepLR strategy with an initial learning rate of 0.1 and a momentum value of 0.9. To simulate the resource constraints of underground mining equipment, we used two RTX 3090s for both training and inference of all models.

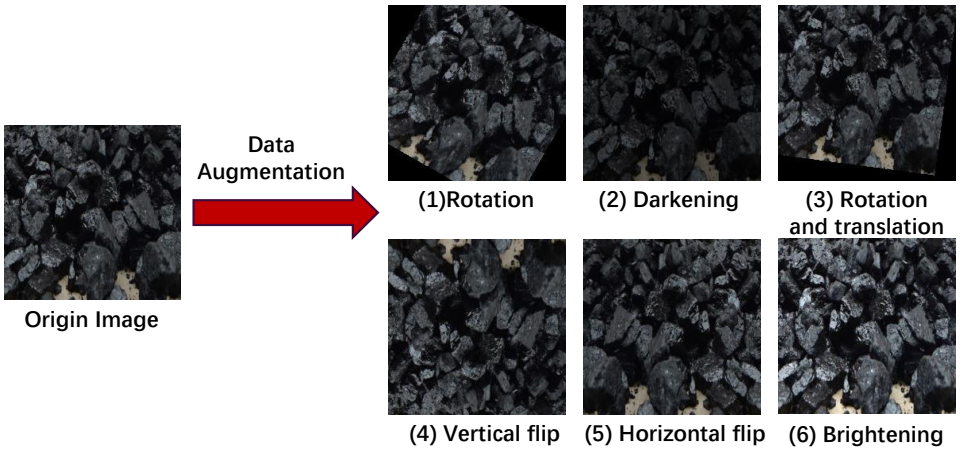


Figure 4. Example of COALSet-4 original image after image enhancement

We set the batch size to 6 for COALSet-4, and to 50 and 128 for the lower-resolution CIFAR and CINIC datasets respectively. CIFAR-10/100 and CINIC-10 were trained for 300 epochs, whereas COALSet-4, due to its smaller dataset size, was trained for only 60 epochs. Under these configuration settings, our method took approximately 10 hours to train on the CIFAR dataset and approximately 0.7 hours on the COALSet-4 dataset. Whilst our method does not offer a significant advantage in terms of training duration compared to baselines and other models, it far surpasses them in terms of convergence.

Comparative experiment

CIFAR 10/100

Tab 2 shows the image classification results for our approach and several ResNet-based variants on the CIFAR-100 dataset. As shown in Tab 2, our method maintains a minimal accuracy gap compared to other approaches despite halving the number of parameters. After incorporating EMA, ResNet50 achieves a Top-1 accuracy of 80.69% and a Top-5 accuracy of 95.59%. With half the number of parameters, our approach achieves a Top-1 accuracy that is only 0.11% lower than that of ResNet50, while its Top-5 accuracy is 0.6% higher. These results conclusively demonstrate the superiority of our approach over traditional ResNet variants. Similarly, ResNet101 utilises 101 residual blocks to achieve a Top-1 accuracy that is 0.28 percentage points higher than ours; however, its Top-5 accuracy remains below that of our method. Notably, these results were attained using only 18 residual blocks.

Furthermore, comparative experiments were conducted on the CIFAR-10 dataset. As illustrated in Tab 3, a comparative analysis is presented between DoMaResNet and several recent methods that have been proposed for ResNet+ visual attention. It is evident that our approach exhibits a substantial degree of superiority in comparison to these alternatives. Employing DoMaBlock convolutional residual blocks, the proposed method attains an accuracy that is 0.3 percentage points higher than that of ResNet-18 when the number of residual blocks is held constant. Compared to ResNet-18+LCA (1D adaptive convolution + batch operations, ultra-lightweight channel attention), the classification accuracy improved by 1.69%. Furthermore, with a reduced number of residual blocks, it surpasses ResNet-50 (DST-EE) by 1.2 percentage points, robustly validating the efficacy of the methodology employed.

CINIC-10

In order to validate the effectiveness of our approach, comparative experiments were conducted against existing methods on the CINIC-10 dataset. The experimental methods incorporated state-of-the-art approaches from the preceding two years, with the results presented in Tab 4. DoMaResNet32 achieved a classification accuracy of 91.75%, representing an improvement of

Method	Backbone	#.Param	Top-1(%)	Top-5(%)
Baseline[19]		23.71M	77.26	93.63
+CBAM[35]		26.24M	80.56	95.34
+SA		23.71M	79.92	95.00
+ECA	ResNet50	23.71M	79.68	95.05
+NAM[35]		23.71M	80.62	95.28
+CA		25.57M	80.17	94.94
+EMA		23.85M	80.69	95.59
+AttZoom[44]	ResNext	-	77.53	95.34
VIPAug[49]	-	-	77.2	-
Baseline[19]		42.70M	77.78	94.39
+CA	ResNet101	46.22M	80.01	94.78
+EMA		42.96M	80.86	95.76
Deformable Attention[45]	Transformer-base	-	79.23(+0.39)	-
+STAtten[46]	QKFormer-4-384	-	80.20	-
MobileNetV3-Large[65]	MobileNetV3	5.4M	69.70	90.25
Baseline[29]		11.22M	78.49	95.74
ParaLkResNet18	PreActResNet18	11.23M	79.56	95.84
DoMaResNet18		11.79M	80.58	96.19

Table 2. Comparison of Training Parameters and Classification Accuracy between DoMaResNet18 and Existing Methods on CIFAR-100

approximately 0.46% compared to the second-highest accuracy achieved by InceptionV1 (91.17% ± 0.12%). The discrepancy in accuracy is more pronounced when compared to traditional ResNet56-based variants (e.g. ResNet56+SE and ResNet56+CBAM), with improvements of approximately 3.42% and 2.72%, respectively. In the field of ResNet-based approaches incorporating visual attention mechanisms, DoMaResNet32 has been demonstrated to exhibit superior feature extraction capabilities and classification performance in comparison to traditional ResNet56 and its variants (+SE, +CBAM, +SCMA).

COALSet-4

Finally, the performance of the proposed DoMaResNet method was validated on the COALSet-4 dataset. Experiments were conducted under identical experimental conditions, with all models trained for 60 epochs to evaluate classification accuracy and network convergence speed. Classification accuracy comparisons are presented in Tab 5. Compared to PreActResNet18 and ParaLkResNet18, DoMaResNet achieves a significant improvement in coal rock image classification accuracy while maintaining comparable parameter counts and computational demands. As analysed in Tab 2, DoMaResNet also outperforms PreActResNet18 and ParaLkResNet18 on the CIFAR-100 dataset. While the classification accuracy difference between ParaLkResNet18 and DoMaResNet is approximately 2% and 1%, respectively, the gap exceeds 10% on the COALSet-4 dataset. The proposed method exhibits markedly smaller performance variations across different datasets than existing models, validating DoMaResNet's high generalisation capability across diverse datasets. Compared to the baseline model, ParaLkResNet18, the proposed DoMaResNet18 achieves an 8.32% improvement in Top-1 accuracy whilst adding only 0.63 GFLOPs (an approximate 2.3% increase in computational load) and 0.56 million parameters. The accuracy of DoMaResNet18 (93.85%) significantly outperforms other models, exceeding that of ParaLkResNet18+CA by 5.5 percentage points. In terms of model parameter count, the lightweight advantages of CA and CBAM are even more pronounced: CA introduces only 0.02–0.07M additional parameters, whilst CBAM adds virtually no additional parameters; The incorporation of the EMA module results in a slight increase in parameters, with the ResNet18, Epsanet and PreActResNet18 backbone models seeing increases of 0.67M, 0.56M and 0.57M respectively; the overall increase in parameters remains manageable. In terms of overall accuracy gains, EMA

Method	Classification accuracy(%)
Tri-Pod ResNet-20[36]	92.47
ResNet56[37]	93.83
BiPer(ResNet-18)[50]	93.75
ResNet164-v2[45]	94.54(-0.09)
ResNet-18[38]	96.04
Efficient ResNets[39]	96.04
VGG-19(DST-EE)[40]	94.13±0.09
ResNet-18+LCA[59]	94.68
VIPAug[49]	95.8
STAtten + QKFormer-4-384[46]	95.35
ResNet-50(DST-EE)[40]	95.01±0.16
DeepTAGE(ResNet-18)[47]	95.86
DoMaResNet18	96.37

Table 3. Comparison of Classification Accuracy Between DoMaResNet18 and Existing Methods on CIFAR-10

Method	Backbone	Classification accuracy(%)	Batches/ s.
VGG-16		87.77±0.16	32
VGG-19_bn[41]	VGG	90.28	-
VGG-19_bn(Spinal FC) [41]		91.00	-
FedPAC+CALM[51]	FedPAC	88.29	10
ResNet-56+SE[42]		87.288(+1.037)	64
ResNet-56+CBAM[42]	ResNet56	87.64(+1.389)	64
ResNet-56+SCMA[42]		87.15(+0.899)	64
Xception	-	87.21±0.08	7
Supernet-Only[48]	-	88.09±0.81	-
ParaLkResNet34	PreActResNet34	91.01	64
InceptionV1[43]	-	91.17±0.12	6
DoMaResNet34	ParaLkResNet34	91.75	128

Table 4. Comparison of Classification Accuracy Between DoMaResNet34 and Existing Methods on CINIC-10

achieves a leap in accuracy with only a modest expansion of the model size, offering better value for money in terms of parameters than CA and CBAM. In terms of computational overhead, the incremental floating-point operations for the three approaches increase progressively: CA adds only 0.2 G FLOPs, CBAM adds 0.3 G FLOPs, whilst the EMA module introduces an additional 0.63 G FLOPs, resulting in higher inference computational costs. However, this computational cost is offset by a classification accuracy far exceeding that of the former two. The EMA attention module is better suited to industrial vision tasks where accuracy is prioritised and computational resources are abundant. It addresses the issues of insufficient accuracy gains and difficulty in surpassing baseline recognition limits encountered with CA and CBAM, thereby achieving an exceptionally high 'accuracy-to-computational-cost' cost-benefit ratio.

Data testing

To evaluate the predictive efficacy of this method on the COALSet-4 classification dataset, we randomly selected four images from different categories and input them into the trained model. The model then outputs predicted probability values for each category, with the final classification result corresponding to the category with the highest predicted probability (see Fig 5). The model achieved excellent classification performance across all rock types, with prediction accuracy exceeding 97% for each of the four categories. Specifically, the classification accuracy for limestone and sandstone approached 100%. These results highlight that training the model on the augmented COALSet-4 data enhances its robustness and generalization capabilities.

Method	Backbone	Input Image Size	Top-1 Accuracy(%)	#.Param	Flops growth
MobileNetV3-Large	MobileNetV3	224×224	89.36	5.4M	-
FasterNet-T2[66]	FasterNet	224×224	90.72	15.0M	-
MobileNetV2 1.1 ×	MobileNetV2[67]	224×224	86.91	4.0M	-
MobileNetV2 1.1 × + CBAM		224×224	87.58	4.1M	+0.04G
MobileNetV2 1.1 × + CA		224×224	88.27	4.1M	+0.05G
MobileNetV2 1.0 × + DFC[68]		224×224	91.48	4.30M	+0.01G
Baseline	PreActResNet18	224×224	85.10	11.22M	-
ParaLkResNet18		224×224	85.53	11.23M	+0.2G
ParaLkResNet18+CA		224×224	88.35	11.25M	+0.2G
Baseline	Epsanet	224×224	88.86	12.30M	-
Epsanet+CA		224×224	90.25	12.34M	+0.2G
Epsanet+DoMaBlock		224×224	91.73	12.86M	+0.63G
Baseline	ResNet18	224×224	88.35	11.18M	-
ResNet18+CA		224×224	88.73	11.25M	+0.2G
ResNet18+CBAM		224×224	90.14	11.18M	+0.3G
ResNet18+DoMaBlock		224×224	91.49	11.85M	+0.63G
DoMaResNet18	ParaLkResNet18	224×224	93.85	11.79M	+0.63G

Table 5. Comparison of Classification Accuracy Between DoMaResNet18 and Existing Methods on COALSet-4

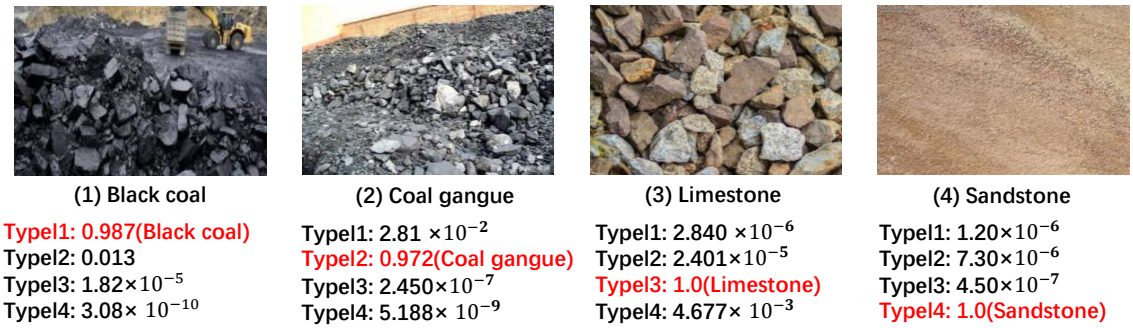


Figure 5. The predicted probability for each rock type sample, as predicted by DoMaResNet18.

Convergence analysis

The accuracy increase scatter plots for the proposed method, baseline models, and PreActResNet18 were plotted over the initial 60 training rounds on the COALSet-4 dataset. Accuracy measurements were taken at intervals of two training rounds. As demonstrated in Fig 6, our proposed method exhibits a substantially superior convergence speed and classification accuracy during the training phase in comparison to the other two models.

Ablation study

This ablation study adhered to the single-variable principle, keeping the network architecture, Parallel Large-Core Convolution Blocks (PLB), Do-conv, EMA and data augmentation schemes fixed, whilst analysing model training results by independently adding or removing a single module under uniform training conditions. The results of the step-by-step ablation experiments demonstrate that the EMA module has a decisive positive impact on the model's feature representation capabilities and classification accuracy. In the experimental group where only the EMA module was added, the feature extraction system of the baseline network was substantially optimised, with the model's Top-1 and Top-5 classification accuracies on the CIFAR-

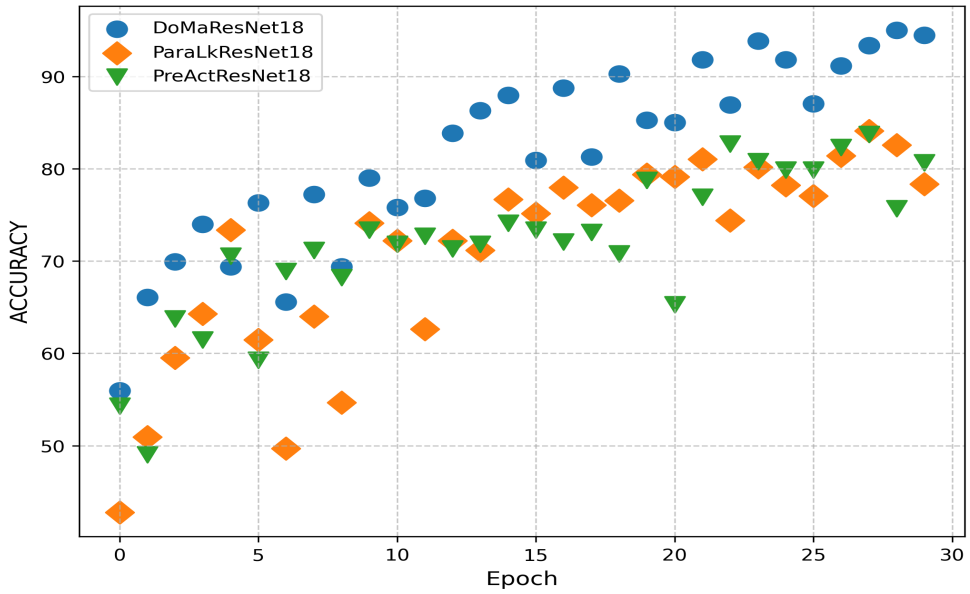


Figure 6. Scatter plot of classification accuracy for DoMaResNet18, PreActResNet18, and the baseline model over the first 60 training rounds on the COALSet-4 dataset.



Figure 7. Heatmap visualization of experimental results of DoMaResNet34 on the CINIC-10 dataset

100 dataset both showing significant improvements of 2.16% and 1.01% respectively. Even in classification results on the Coalset-4 dataset without data augmentation, accuracy increased by 4.25%, demonstrating that this module can independently deliver stable performance gains to the model. Compared to optimisation strategies such as the PLB convolutional structure, Do-conv and data augmentation, the performance gains achieved by the EMA module are more pronounced, making it the core component that contributes most significantly to classification results among all the proposed improvements.

Visual inspection

In order to further demonstrate the method's capability in the extraction of features from natural images, the trained model was employed to conduct visual analysis of the regions of interest within images from the CINIC-10 dataset. The selection of images was conducted at random. As illustrated in Fig 7, the results of feature extraction and region of interest information

Method	PLB	Do-conv	EMA	Data augmentation	Coalset-4(%)	CIFAR-100	
						Top-1(%)	Top-5(%)
Baseline(PreActResNet18)	-	-	-	-	35.59	77.27	94.58
+PLB(ParaLkResNet18)	✓	-	-	-	36.78	78.95	94.91
+Do-conv	-	✓	-	-	36.01	78.23	94.62
+EMA	-	-	✓	-	39.84	79.43	95.59
+Data augmentation	-	-	-	✓	85.10	78.20	94.63
DoMaResNet18	✓	✓	✓	✓	93.85	80.58	96.19

Table 6. The incremental classification accuracy achieved by our approach when applied to ParaLkResNet.

obtained using the DoMaResNet32 model are presented. The figure demonstrates that the DoMaResNet32 model places significant emphasis on local features, accurately covering key feature information regions across various categories. This provides robust support for achieving high classification accuracy. To investigate the basis of the model's decisions, Grad-CAM was used to generate a heatmap of the regions of interest for DoMaResNet18 (Fig 8). In the figure, the blue regions represent areas of high model response, corresponding to core discriminative features; the yellow and red regions represent areas of low response in the background. The visualisation results indicate that the model's attention is consistently and precisely focused on the coal and rock itself, whilst assigning extremely low response values to irrelevant background elements such as hands, shovels and tabletops in the image. In both uncluttered coal pile scenarios and complex scenes involving hands and tools, the model consistently identifies key features of coal and rock without being affected by background noise, demonstrating excellent robustness and feature-focusing capability. This provides intuitive validation that the EMA multi-scale attention module has guided the model to learn discriminative features strongly correlated with the coal-rock classification task, offering an explanatory basis for the model's superior performance.

Deployment analysis

The comparative experimental results on the COALSet-4 dataset demonstrate that DoMaBlock, as an improved residual unit incorporating EMA multi-scale attention, exhibits excellent general adaptability and can be integrated into various convolutional residual backbone networks such as ResNet18, Epsanet, and ParaLkResNet18. In an 18-layer residual network, this module introduces only a modest 0.63G FLOPs increment and a small number of additional parameters, enabling more significant classification accuracy improvements with lower computational costs compared to traditional attention modules like CA and CBAM. Across different residual baselines, it consistently achieves Top-1 accuracy gains of 2.87% to 8.32%. The complete model DoMaResNet18 further achieves optimal recognition accuracy of 93.85% with just 11.79M parameters, surpassing mainstream lightweight industrial vision models such as MobileNet series, FasterNet, and DFC. From the perspective of underground edge and mobile device deployment, the 0.63G floating-point computation increment accounts for a minimal fraction of the power consumption in mining-oriented embedded NPUs and low-power terminal computing resources. With model parameter growth under 6%, post-quantization storage is compact, avoiding significant inference latency and storage pressure. Meanwhile, leveraging the module's built-in multi-scale attention to capture fine-grained texture features, it effectively mitigates complex imaging interferences like mine dust and weak illumination. Without requiring hardware upgrades in existing underground systems, model replacement alone can substantially reduce coal-rock identification misclassification and omission rates. Considering computational efficiency, storage, and recognition performance, residual networks incorporating DoMaBlock are fully suitable for real-time detection in underground mobile inspection terminals and edge monitoring equipment.

Conclusion

This paper proposes an innovative convolutional residual module named DoMaBlock, which realizes efficient feature extraction for coal-rock recognition and natural image scenarios to serve industrial visual inspection and robust image classification tasks. Based on a lightweight residual learning architecture, the proposed method integrates a multi-scale attention mechanism to comprehensively learn cross-spatial, multi-scale features as well as features for fine-grained texture classification, while accelerating network convergence to a certain extent. Experiments are conducted on three public image classification datasets (CIFAR-10, CIFAR-100 and CINIC-10) and one dedicated coal-rock dataset (COALSet-4) to verify the effectiveness of the proposed method in robust image classification, coal-rock recognition and fine-grained texture classification tasks. Experimental results demonstrate that embedding this module into baseline models brings significant improvements in classification performance. Furthermore, guided by the concept of lightweight residual learning, the proposed method features a concise structure and strong practicability. It can strengthen feature representation via the multi-scale attention mechanism, and

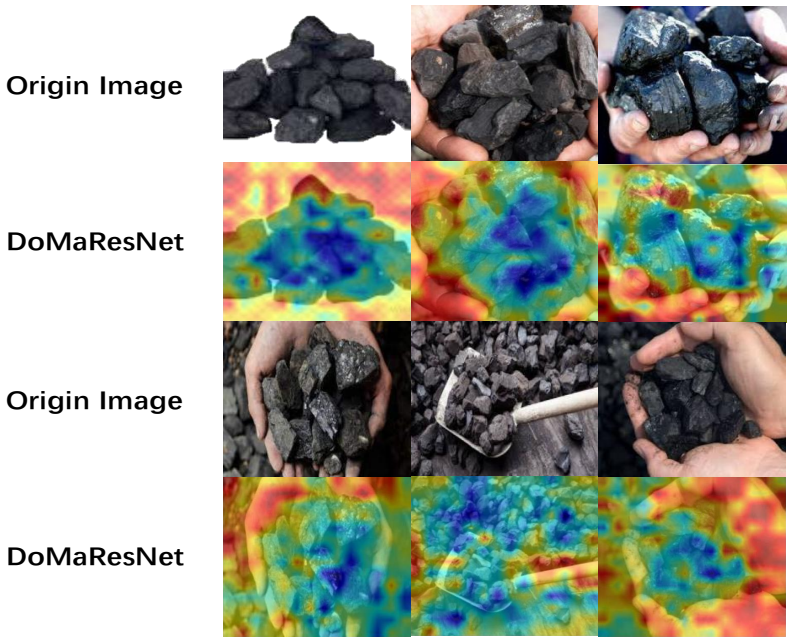


Figure 8. Heatmap visualization of experimental results of DoMaResNet18 on the CoalSet-4 dataset

is applicable to all types of ResNet and its variants, providing solid support for practical industrial visual inspection applications. Future research will further reduce the number of module parameters to achieve better performance with minor structural modifications. It will continuously boost the accuracy of robust image classification and coal-rock recognition, and deliver more efficient lightweight residual learning solutions for industrial visual inspection and fine-grained texture classification.

Consent to Publish declaration

Not applicable

Ethics and Consent to Participate declarations

Not applicable

Acknowledgements

This research was funded by the Dezhou Engineering Research Centre for Big Data and Intelligent Perception Technology (Platform Number: PT2025KJT002).

Data and code availability

The CIFAR dataset supporting the results of this study is publicly available at https://www.cs.toronto.edu/~kriz/cifar.html. The CINIC-10 dataset is openly available at http://dx.doi.org/10.7488/ds/2448. Readers who wish to access the proprietary COALSet-4 dataset may contact the corresponding author for relevant data materials.

Author contributions statement

T.S.Y. conceived and designed the research. Y.L.C. and S.W. carried out all experiments. T.S.Y. analyzed the experimental data. T.S.Y., D.Z. and S.W. generated all figures and tables. Y.L.C. and T.S.Y. jointly analyzed the results and drafted the manuscript. T.S.Y. supervised the whole manuscript writing process. Y.L.C., D.Z. and S.W. revised and polished the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.Z.

References

1. Tripathy, S. and R. Singh (2022). Convolutional neural network: an overview and application in image classification. Proceedings of Third International Conference on Sustainable Computing: SUSCOM 2021, Springer.
2. Cai, L., et al. (2020). "A review of the application of deep learning in medical image classification and segmentation." 8(11).
3. Parkhi, O., et al. (2015). Deep face recognition. BMVC 2015-Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association.
4. Yurtsever, E., et al. (2020). "A survey of autonomous driving: Common practices and emerging technologies." 8: 58443-58469.
5. Rajagopal, A., et al. (2020). "A deep learning model based on multi-objective particle swarm optimization for scene classification in unmanned aerial vehicles." 8: 135383-135393.
6. Suganyadevi, S., et al. (2022). "A review on deep learning in medical image analysis." 11(1): 19-38.
7. Wang, X., et al. (2021). "Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification." 59(9): 7918-7932.
8. Sun, C., et al. (2022). "Coal rock image recognition method based on improved CLBP and receptive field theory." 1(2): 165-173.
9. Woo, S., et al. (2023). Convnext v2: Co-designing and scaling convnets with masked autoencoders. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
10. Tan, M. and Q. Le (2021). Efficientnetv2: Smaller models and faster training. International conference on machine learning, PMLR.
11. Han, K., et al. (2021). "Transformer in transformer." 34: 15908-15919.
12. Hu, J., et al. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition.
13. Woo, S., et al. (2018). Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV).
14. Hou, Q., et al. (2021). Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
15. Ouyang, D., et al. (2023). Efficient multi-scale attention module with cross-spatial learning. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.
16. Krizhevsky, A., et al. (2012). "Imagenet classification with deep convolutional neural networks." 25.
17. Simonyan, K. and A. J. a. p. a. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition."
18. Szegedy, C., et al. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition.
19. He, K., et al. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.
20. Howard, A. G. J. a. p. a. (2017). "Mobilenets: Efficient convolutional neural networks for mobile vision applications."
21. Huang, G., et al. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.
22. Huang, G., et al. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.
23. Yu, T., et al. (2024). "ParaLkResNet: an efficient multi-scale image classification network." 1-10.

24. Dai, J., et al. (2017). Deformable convolutional networks. Proceedings of the IEEE international conference on computer vision.

25. Zhu, X., et al. (2019). Deformable convnets v2: More deformable, better results. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

26. Li, X., et al. (2019). Selective kernel networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

27. Chen, Y., et al. (2019). Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. Proceedings of the IEEE/CVF international conference on computer vision.

28. Huang, S., et al. (2023). Neurogenesis dynamics-inspired spiking neural network training acceleration. 2023 60th ACM/IEEE Design Automation Conference (DAC), IEEE.

29. He, K., et al. (2016). Identity mappings in deep residual networks. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer.

30. Cao, J., et al. (2022). "Do-conv: Depthwise over-parameterized convolutional layer." 31: 3726-3736.

31. Ding, X., et al. (2021). Repvgg: Making vgg-style convnets great again. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

32. Duta, I. C., et al. (2020). "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition."

33. Krizhevsky, A. and G. Hinton (2009). "Learning multiple layers of features from tiny images."

34. Darlow, L. N., et al. (2018). "Cinic-10 is not imagenet or cifar-10."

35. Liu, Y., et al. (2021). "NAM: Normalization-based attention module."

36. Pan, H., et al. (2022). "Multipod convolutional network."

37. Tang, H., et al. (2023). Sr-init: An interpretable layer pruning method. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.

38. Bu, T., et al. (2022). Optimized potential initialization for low-latency spiking neural networks. Proceedings of the AAAI conference on artificial intelligence.

39. Thakur, A., et al. (2023). "Efficient ResNets: Residual Network Design."

40. Huang, S., et al. (2023). Dynamic sparse training via balancing the exploration-exploitation trade-off. In 2023 60th ACM/IEEE Design Automation Conference (DAC), IEEE.

41. Kabir, H. D., et al. (2022). "Spinalnet: Deep neural network with gradual input." 4(5): 1165-1177.

42. Samad, S. A. and J. J. I. A. Gitanjali (2023). "SCMA: exploring dual-module attention with multi-scale kernels for effective feature extraction."

43. Yuan, H., et al. (2022). "Low-res MobileNet: An efficient lightweight network for low-resolution image classification in resource-constrained scenarios." 81(27): 38513-38530.

44. DeAlcala, D., et al. (2025). AttZoom: Attention Zoom for Better Visual Features. Proceedings of the IEEE/CVF International Conference on Computer Vision.

45. Alfatih, M. I. and S. A. Wibowo (2025). "Star Classifier Head on Deformable Attention Vision Transformer for Small Datasets." IEEE Access.

46. Lee, D., et al. (2025). Spiking transformer with spatial-temporal attention. Proceedings of the Computer Vision and Pattern Recognition Conference.

47. Liu, W., et al. (2025). Deeptage: Deep temporal-aligned gradient enhancement for optimizing spiking neural networks. The Thirteenth International Conference on Learning Representations.

48. Chen, Y., et al. (2025). "A Supernet-Only Framework for Federated Learning in Computationally Heterogeneous Scenarios." Applied Sciences 15(10): 5666.

49. Lee, I., et al. (2024). Domain generalization with vital phase augmentation. Proceedings of the AAAI conference on artificial intelligence.

50. Vargas, E., et al. (2024). BiPer: Binary neural networks using a periodic function. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

51. Zheng, H., et al. (2025). FedCALM: Conflict-aware Layer-wise Mitigation for Selective Aggregation in Deeper Personalized Federated Learning. Proceedings of the Computer Vision and Pattern Recognition Conference.

52. TIAN, L., et al. (2020). "Coal-rock identification method based on multi-strain data fusion of shearer rocker pin shaft." Journal of China Coal Society 45(3).

53. Zhang, N. and C. Liu (2018). "Radiation characteristics of natural gamma-ray from coal and gangue for recognition in top coal caving." Scientific reports 8(1): 190.

54. Hou, W. (2019). "Identification of coal and gangue by feed-forward neural network based on data analysis." International Journal of Coal Preparation and Utilization 39(1): 33-43.

55. Xue, G., et al. (2020). "Random forest-based image recognition of coal gangue at heaving workface." Ind. Min. Autom 46(05): 57-62.

56. Su, L., et al. (2018). Research on coal gangue identification by using convolutional neural network. 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE.

57. Yang, J., et al. (2021). "Research on CNN Coal and Rock Recognition Method Based on Hyperspectral Data."

58. Xiao, D., et al. (2022). "Coal identification based on a deep network and reflectance spectroscopy." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 270: 120859.

59. Kanaparthi, P. B. and T. V. S. V. Padamata (2026). "Lightweight Channel Attention for Efficient CNNs." arXiv preprint arXiv:2601.01002.

60. Zhang, Z., et al. (2025). "An improved residual learning model and its application to hardware image classification." The Visual Computer 41(1): 451-464.

61. Zu, B., et al. (2025). "RESwinT: enhanced pollen image classification with parallel window transformer and coordinate attention." The Visual Computer 41(7): 4975-4990.

62. Dong, B., et al. (2025). "Integrating models of real aboveground scene and underground geological structures at an open pit mine." Virtual Reality & Intelligent Hardware 7(4): 406-420.

63. Wang, Z., et al. (2026). "Contrastive decoupling: Dynamic regularization for enhanced fine-grained image classification." IEEE Transactions on Circuits and Systems for Video Technology.

64. Liu, R., et al. (2025). "Yolov8-HAC: Safety Helmet Detection Model for Complex Underground Coal Mine Scene." Computer Animation and Virtual Worlds 36(4): e70051.

65. Howard, A., et al. (2019). Searching for mobilenetv3. Proceedings of the IEEE/CVF international conference on computer vision.

66. Chen, J., et al. (2023). Run, don't walk: chasing higher FLOPS for faster neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

67. Sandler, M., et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition.

68. Tang, Y., et al. (2022). "GhostNetv2: Enhance cheap operation with long-range attention." Advances in Neural Information Processing Systems 35: 9969-9982.