MA-ViT: An Ore Classification Method Based on Attention Mechanism and Class-Balanced Learning

Ziyang Qin[1], Wanwan Wang[2]

[1]*School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China*

[2]*iFLYTEK Co., Ltd., Hefei, Anhui, China*

Abstract: Fine-grained classification of ores is of significant importance for geological exploration and mineral processing, whereas traditional manual identification methods are inefficient and heavily reliant on expert experience. Existing deep learning-based identification methods often face challenges such as insufficient texture feature extraction and high miss rates for minority classes when dealing with complex field backgrounds and severely imbalanced sample distributions. To address these issues, this paper proposes an ore classification model based on a mixed attention mechanism and class-balanced learning—MA-ViT (Mixed-Attention Vision Transformer). This method utilizes the Vision Transformer (ViT) as the backbone network. First, it introduces the Convolutional Block Attention Module (CBAM) to effectively suppress background noise and focus on key ore texture features. Second, it designs a class-weighted loss function to mitigate model bias caused by data imbalance. Experimental results on an image dataset containing seven types of ores show that, when comparing MA-ViT with mainstream models such as ResNet-50 and the original ViT, MA-ViT achieves an overall accuracy of 95.61% and a macro-averaged F1 score of 0.9435, outperforming current mainstream models. In particular, for the sample-scarce "Muscovite" category, the recall rate increased significantly from 80.00% in the baseline model to 91.43%, achieving a balance between high precision and high recall. The method proposed in this paper demonstrates strong robustness and generalization ability, providing an effective reference for automatic ore identification in complex environments.

Keywords: Fine-grained Classification; Attention Mechanism; Class-balanced Learning; Ore Identification; Vision Transformer

1. Introduction

Mineral resources serve as the cornerstone and lifeblood of the modern industrial system, and the efficiency of their exploration and utilization directly concerns the lifeline of the national economy. For a long time, ore identification work at geological exploration sites has often had to rely on experienced geological experts holding magnifying glasses to discriminate minerals one by one based on color, luster, and crystal structure. However, this traditional operational mode, which is highly dependent on individual expert experience and sensory judgment, is facing multiple bottlenecks: low operational efficiency, heavy personnel burden, and, in particular, judgment accuracy is significantly interfered with by environmental factors and subjective states. This has become a core challenge restricting mines from achieving large-scale, high-throughput, and intelligent sorting.

With the development of science and technology, computer vision technology has begun to intervene in this traditional field, attempting to liberate human experts from heavy repetitive labor. However, when researchers attempt to transfer existing general vision models directly to real wild ore images, unlike object recognition in laboratories with clean backgrounds, wild-collected ore images are often accompanied by strong background noise. For example, soil, weeds, and even shadows during shooting often seriously interfere with the subjectivity of the foreground target (ore), causing the model to be easily led astray by this irrelevant information when extracting features, unable to precisely focus on the subtle textures of the ore itself.

In real mining scenarios, common minerals like Quartz and Feldspar are abundant, while high-value rare minerals such as Muscovite are few and far between. The existence of extreme imbalance in data distribution further exacerbates the difficulty of analysis, making models tend to improve overall accuracy during training by catering to majority class samples, while selectively ignoring those key minority classes, causing severe risks of missed detection. Existing mainstream models, whether Convolutional Neural Networks (CNNs) skilled in local features or Transformers skilled in global modeling, often appear stretched when facing this dual challenge of needing to resist interference while maintaining balance.

How to precisely lock onto targets in complicated backgrounds while simultaneously accounting for identification sensitivity to scarce samples has become the core challenge for the landing of current intelligent geological identification technology. Therefore, this paper proposes an enhanced solution fusing a mixed attention mechanism and a class balance strategy—MA-ViT. It attempts to equip the model with a filter that automatically filters noise by introducing the CBAM module, allowing it to focus on ore textures in noisy backgrounds; meanwhile, it utilizes a cost-sensitive learning strategy to penalize the model's bias towards majority classes. This is not only a deep reconstruction of existing algorithm architectures but also a solid step towards automated, high-precision geological identification in complex environments.

2. Related Work

The dominance of deep learning in the field of image classification is largely attributed to the continuous iteration and evolution of neural network architectures. Early exploration mainly focused on the field of Convolutional Neural Networks (CNN). ResNet, proposed by He et al. [1], successfully solved the gradient vanishing problem in deep networks through the ingenious design of residual connections, establishing the hegemony of CNN in vision tasks. Deng and Yu [2] et al. improved the convergence speed of the Faster R-CNN algorithm and used a multi-scale feature fusion method to achieve recognition of specific ore image samples. Xiao et al. [3] proposed a comprehensively improved Mask R-CNN ore type detection algorithm, using ResNetV1d-50 to extract feature maps of ore images at various stages, and improving FPN and RPN to further improve the recognition accuracy of irregular ores. Zou et al. [4] proposed an ore image recognition model based on interactive feature fusion, providing a novel image recognition method for the field of intelligent ore sorting. Gao et al. [5] proposed a lightweight ore image classification algorithm based on cross-channel fine-grained feature fusion. This approach addresses issues such as low accuracy when processing fine-grained texture features, high computational resource demands, and difficulties in mobile deployment encountered by deep learning algorithms. However, the inherent local receptive field of CNNs limits their ability to capture long-distance pixel correlations, making them struggle when processing global complex textures.

To break through this limitation, Dosovitskiy et al. [6] proposed the Vision Transformer (ViT), transforming image processing into sequence modeling. ViT achieves direct modeling of global features by slicing images into patches and utilizing the self-attention mechanism, performing excellently on large datasets. The high computational cost of ViT and its relative lack of local detail capture capabilities have also triggered subsequent related optimization research. For example, the Cu-ViT model proposed by Jin et al. [7] systematically improved the recognition capability of the ViT model for image capture by substituting patch embeddings with compression units; the RegionViT model proposed by Chen et al. [8] introduced a region-to-local (R2L) attention mechanism, further optimizing the feature extraction process, aiming to improve the comprehensive performance of vision tasks.

Although the above studies have made significant progress in their respective dimensions, there is still a clear missing piece in existing solutions for the specific proposition of ore identification. Most studies either focus on the lightweighting of model structures or on texture classification in ideal environments, but few works can simultaneously account for the anti-interference requirements of complex backgrounds and the class balance problem under long-tail data distributions. Existing open-source datasets are mostly general natural scenes, lacking high-quality annotated data specifically for ore features, which directly leads to high miss rates of existing models when facing scarce ore species. Therefore, constructing a dedicated architecture integrating attention enhancement and cost-sensitive learning has become the key to filling this academic gap.

3. Experimental Principles and Methods

3.1 Vision Transformer Backbone Network

The Vision Transformer (ViT) is a visual recognition model entirely based on the self-attention mechanism, discarding traditional convolution operations and transforming the image understanding problem into a sequence modeling task. This paper selects the Vision Transformer as the backbone network for feature extraction, directly utilizing the self-attention mechanism to model the global context information of the image. This characteristic is particularly critical for ore image recognition with complex texture distribution and long-distance dependency features.

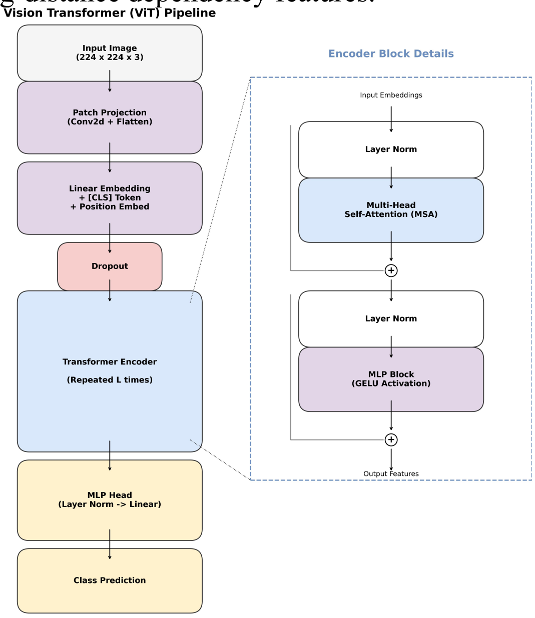
Vision Transformer (ViT) Pipeline
Input Image (224 x 224 x 3)
Patch Projection (Conv2d + Flatten)
Linear Embedding + [CLS] Token + Position Embed
Dropout
Transformer Encoder
(Repeated L times)
MLP Head (Layer Norm -> Linear)
Class Prediction
Encoder Block Details
Input Embeddings
Layer Norm
Multi-Head Self-Attention (MSA)
+
Layer Norm
MLP Block (GELU Activation)
+
Output Features

Figure 1 Vision Transformer Network Architecture

As shown in Figure 1, the overall structure of the Vision Transformer mainly includes four parts: input processing, feature embedding, Transformer encoder, and output prediction. The model first divides the input image into patches of consistent size and no overlap, and flattens them into vectors, thereby converting the two-dimensional image into a sequence form. Subsequently, uniform-dimension Patch Embeddings are obtained through linear mapping, and learnable position encodings are introduced to retain spatial position information while modeling globally. In the core Transformer encoder, multiple layers of Transformer Blocks establish global dependencies between different patches through the Multi-Head Self-Attention mechanism and combine with Feed-Forward Neural Networks for non-linear enhancement of features, effectively compensating for the limited receptive field of convolutional networks. Finally, the model uses the [CLS] Token to aggregate the global features of the entire image and outputs the recognition result through the prediction head.

3.2 CBAM Mixed Attention Mechanism

To mitigate the interference of background noise on subject features, the CBAM module proposed by Woo et al. [9] innovatively combines channel attention and spatial attention, capable of adaptively refining feature maps through weighting. To cope with the background noise of wild-collected ore images, this paper introduces the CBAM module, enabling the model to automatically filter background noise when facing complex images.

As shown in Figure 2, the CBAM architecture is essentially a dual-stage filtering mechanism composed of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). Although structurally independent, the two can work synergistically, aiming to optimize the extraction capability of the network through refined feature reconstruction, thereby significantly improving the model's detection accuracy.

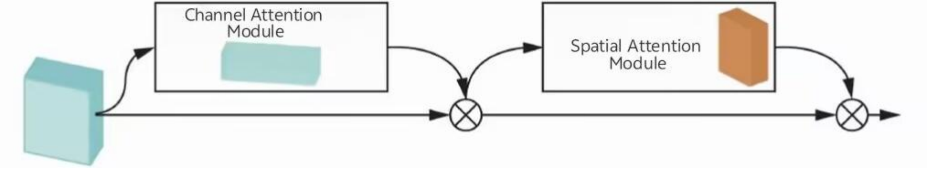

Figure 2. CBAM Structure

The design logic of the Channel Attention Module is shown in Figure 3: it does not rush to process complicated spatial information but first uses a dual approach of max pooling and average pooling to separately extract the extreme value response and background mean of the input feature map in the channel dimension. Subsequently, these feature vectors aggregating key information are sent into a Shared Multilayer Perceptron (Shared MLP). This design effectively reduces computational redundancy while accurately capturing dependencies between channels. After non-linear mapping through superposition and activation functions, the generated channel weights are broadcast back to the original feature map, completing the preliminary enhancement of key semantic features through channel-wise multiplication, which serves as the input for the next stage.

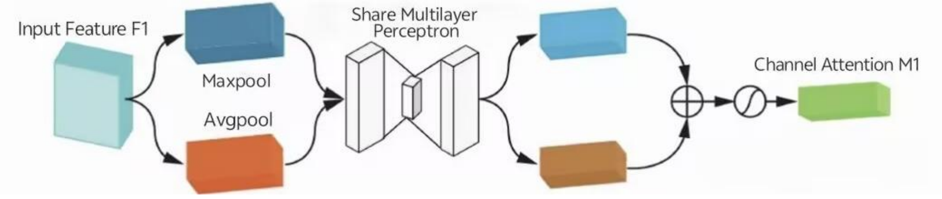

Figure 3. Channel Attention Module

Following closely is the Spatial Attention Module as shown in Figure 4. It receives the features weighted by the channel, performs max pooling and average pooling again along the channel axis, and the generated two two-dimensional feature maps are concatenated to form a dual-channel map containing rich spatial information. Subsequently, a convolutional layer is used for deep fusion and dimensionality reduction of spatial features, and the final spatial attention map is generated through an activation function. This weight matrix is multiplied point-wise with the previous feature map, finally outputting an enhanced feature map that contains both key semantics and precise location information.

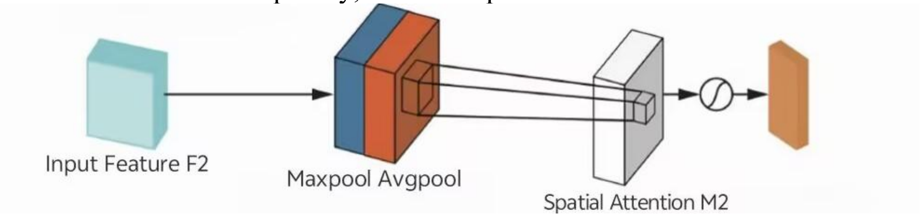

Figure 4. Spatial Attention Module

3.3 Class Imbalance Handling Strategy

To break the unfairness brought about by data distribution, this paper does not adopt traditional oversampling or undersampling techniques. The former easily introduces redundant noise leading to overfitting, while the latter may lose valuable feature information. Instead, a cost-sensitive learning mechanism [10] is introduced to achieve the model's attention redirection by reconstructing the penalty intensity of the loss function [11]. The core idea of this strategy is inverse weighting, that is, assigning inversely proportional penalty weights based on the frequency of appearance of various samples in the training set. Specifically, for the i-th class of ore, the calculation formula for its weight Wi is defined as:

$$W_I = \frac{N}{C \cdot n_i} \tag{1}$$

Where N represents the total number of samples in the training set, C is the total number of classes, and ni is the number of samples for the i-th class of ore. Through this formula, the extremely scarce "Muscovite" category will be assigned a larger weight coefficient, while categories with redundant samples will be assigned smaller weights.

On this basis, we formulate the improved weighted loss function Lweighted as：

$$L_{\text{weighted}} =-\frac{1}{N}\sum_{j=1}^{N} w_{y_j} \cdot \log\left(p_{y_j}\right) \tag{2}$$

Where yj is the true label of the j-th sample, and pyj is the probability that the model predicts the sample belongs to the true category. During the backpropagation process, the larger weight wyj will significantly amplify the gradient signal generated by minority class samples, forcing the optimizer to pay more attention to these difficult-to-classify samples when updating parameters. This mechanism effectively balances the contribution of different categories to model updates, enabling MA-ViT to significantly improve the recall rate for imbalanced data while maintaining high overall accuracy.

3.4 MA-ViT Overall Network Architecture

When facing original ore images filled with soil coverage, vegetation occlusion, and uneven lighting, conventional models often struggle to accurately extract effective features. Therefore, this paper proposes a technical path spanning

"Feature Enhancement—Global Modeling—Cost-Sensitive Optimization". As shown in Figure 5:

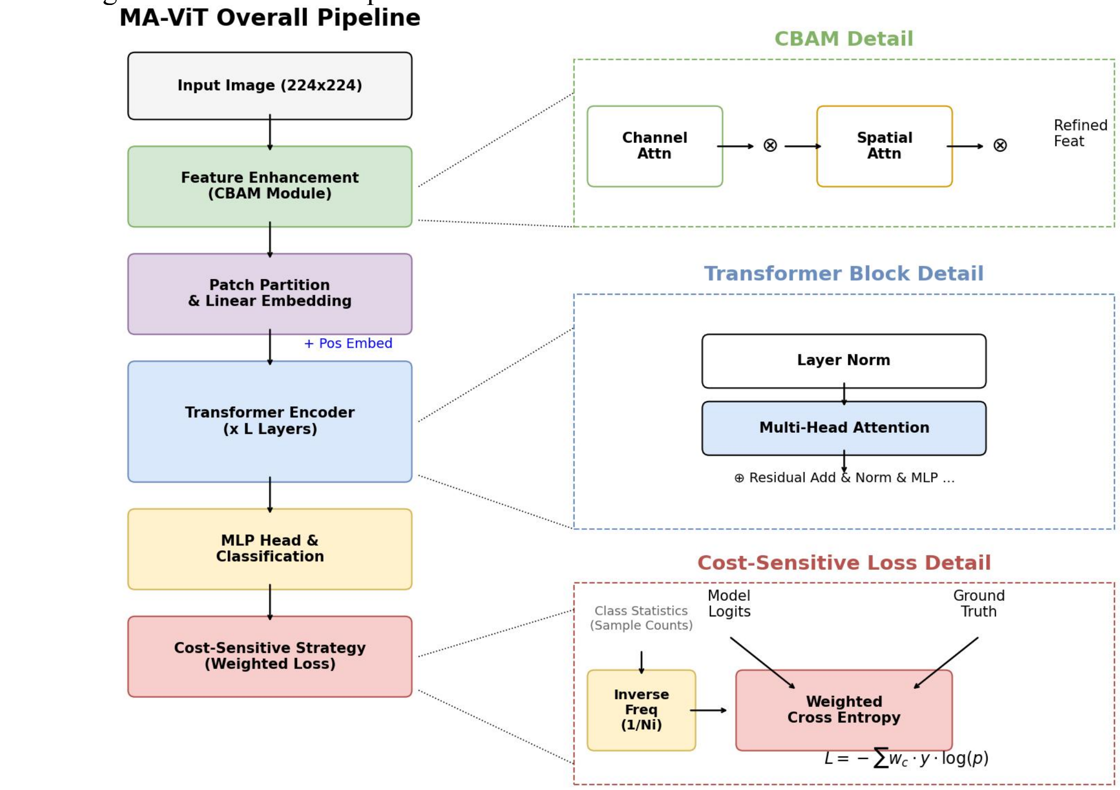
MA-ViT Overall Pipeline
Input Image (224x224)
Feature Enhancement (CBAM Module)
Patch Partition & Linear Embedding
+ Pos Embed
Transformer Encoder (x L Layers)
MLP Head & Classification
Cost-Sensitive Strategy (Weighted Loss)
CBAM Detail
Channel Attn
⊗
Spatial Attn
⊗
Refined Feat
Transformer Block Detail
Layer Norm
Multi-Head Attention
⊕ Residual Add & Norm & MLP ...
Cost-Sensitive Loss Detail
Class Statistics (Sample Counts)
Model Logits
Ground Truth
Inverse Freq (1/Ni)
Weighted Cross Entropy
$L = -\sum w_c \cdot y \cdot \log(p)$

Figure 5. MA-ViT Network Architecture

At the source of data input, this experiment did not rush to send the original images directly into the Transformer's encoding sequence. Considering that wild-collected ore images are often filled with irrelevant noise such as soil, weeds, and even shooting shadows, the model first introduces a feature enhancement module as a "pre-filter". The original 224*224 image first passes through the concatenated Channel Attention and Spatial Attention units, suppressing the response value of background noise to the minimum. This preprocessing mechanism ensures that what subsequently enters the backbone network is no longer noisy raw pixels, but purified high signal-to-noise ratio features.

Afterward, the processed feature maps enter the Vision Transformer backbone network. Here, continuous image signals are sliced into fixed-size patches, and position encodings are superimposed to retain spatial structure information. Unlike convolutional neural networks which are limited to local receptive fields, the stacked Transformer Encoders utilize the self-attention mechanism to span the physical distance of image blocks and capture long-distance semantic dependencies in ore textures—whether it is the conchoidal fracture on the surface of quartz or the tabular crystal form of feldspar. These subtle clues distributed in different corners of the image are integrated into a unified global feature representation and finally aggregated in the Class Token, which is mapped to a preliminary prediction through the MLP Head.

Targeting the objectively existing data imbalance problem in the training data, a conventional loss calculation method was not adopted; instead, a weighted cross-entropy loss function was designed. As shown on the right side of Figure 5, this module receives true labels from the dataset and pre-calculated class weights. During the backpropagation process: when the model faces scarce samples such as Muscovite, the loss function dynamically amplifies the gradient signal based on the inverse of its frequency, forcing the optimizer to give higher attention to these difficult-to-classify samples, thereby achieving high-precision and high-recall identification of ore samples in complex environments.

4. Experimental Process

4.1 Experimental Preparation

This experiment used an ore classification dataset containing 5640 valid samples, covering seven typical rock-forming minerals: Biotite, Bornite, Chrysocolla, Malachite, Muscovite, Pyrite, and Quartz. The dataset exhibits distinct data imbalance characteristics. As shown in Figure 6, Quartz (1185 images) and Pyrite (1086 images), which have the most abundant sample sizes, dominate, while Muscovite, as a rare category, has only 344 samples.

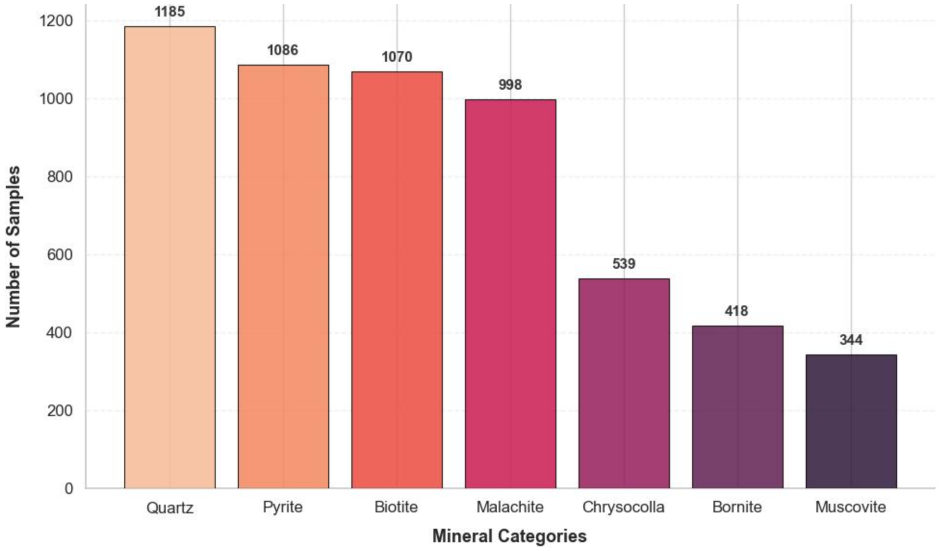

Figure 6. Category Distribution of the Ore Dataset

In the data preprocessing stage, considering the input sequence standardization requirements of the Vision Transformer architecture, all original images were first uniformly cropped and scaled to 224*224 pixels. In mineralogical identification, the color and luster of ores are key physical attributes for distinguishing mineral species (such as the unique emerald green of Malachite); arbitrarily changing the image hue can easily destroy this core semantic information. Therefore, this experiment only introduced geometric transformations such as random horizontal flipping, vertical flipping, and 90-degree rotation, aimed at letting the model learn the rotational invariance of ore textures in space while maximizing the preservation of its original optical features. Finally, the dataset was divided into a training set (4512 images), a validation set (564 images), and a test set (564 images) in a ratio of 8:1:1.

4.2 Model Training

The model training and inference processes were deployed on a high-performance computing platform. The hardware environment was equipped with an NVIDIA Tesla V100 (16GB) graphics acceleration card, utilizing its powerful parallel floating-point computing capability to accelerate gradient backpropagation; the software architecture was based on the Python 3.11 environment and used the PyTorch deep learning framework to build the computational graph, cooperating with the CUDA operation library to achieve underlying hardware acceleration.

Given that the Vision Transformer architecture lacks the inductive bias inherent in convolutional neural networks and usually requires massive amounts of data for pre-training to converge, while the scale of the self-built ore dataset in this experiment is relatively limited, a transfer learning strategy was adopted to overcome this "data hunger" challenge. Specifically, the backbone network of MA-ViT loaded vit_base_patch16_224 weight parameters pre-trained on the ImageNet-21k dataset, enabling the model to possess general visual capabilities for identifying basic edges and textures from the beginning of training. To suppress overfitting, a stochastic depth strategy was introduced in the Transformer Encoder, setting the DropPath rate to 0.1, i.e., randomly dropping part of the residual paths during training, enhancing the model's robustness.

In terms of specific hyperparameter configuration and optimization strategy, this experiment selected the AdamW optimizer. Compared to traditional SGD, AdamW demonstrates better regularization effects when dealing with models with huge parameters like Transformers by decoupling weight decay (Weight Decay, set to 0.01) and gradient updates. We set the Batch Size to 32, which fully utilized the 16GB video memory space and ensured the stability of BN layer statistics. The learning rate adjustment strategy adopted cosine annealing: the initial learning rate was set to a relatively conservative 3*10-5 to adapt to the fine-tuning needs of pre-trained weights. Throughout the training cycle, the learning rate gradually decayed to 0 according to the cosine function curve. The entire training process was set to 30 Epochs. To prevent the model from rote memorizing on the training set, a best model saving mechanism was adopted: after each round of training, the model performance was evaluated on the validation set, and the current weight file was updated and saved only when the validation set accuracy surpassed the historical best record. The final test evaluation was based on these optimal weights to ensure the objectivity and fairness of the indicators.

4.3 Analysis of Experimental Results

4.3.1 Training process convergence analysis

Figure 7 shows the loss decline curve and accuracy rise curve of the MA-ViT model during the training process. Thanks to the introduction of ImageNet pre-trained weights, the model showed extremely fast convergence speed in the early stage of training: within less than 3 Epochs, the training set loss dropped rapidly from 1.2 to below 0.2, and the validation set accuracy broke through 90%. As training advanced to around the 10th Epoch, the curve gradually flattened, entering the fine-tuning stage. The curves of the training set and validation set always maintained a close accompanying relationship, and there was no significant sign of overfitting where the training set Loss continued to decline while the validation set Loss rose conversely. This indicates that the DropPath stochastic depth strategy and the AdamW weight decay mechanism played a key role, successfully constraining the complexity of the model and maintaining high fitting ability while balancing good generalization performance.

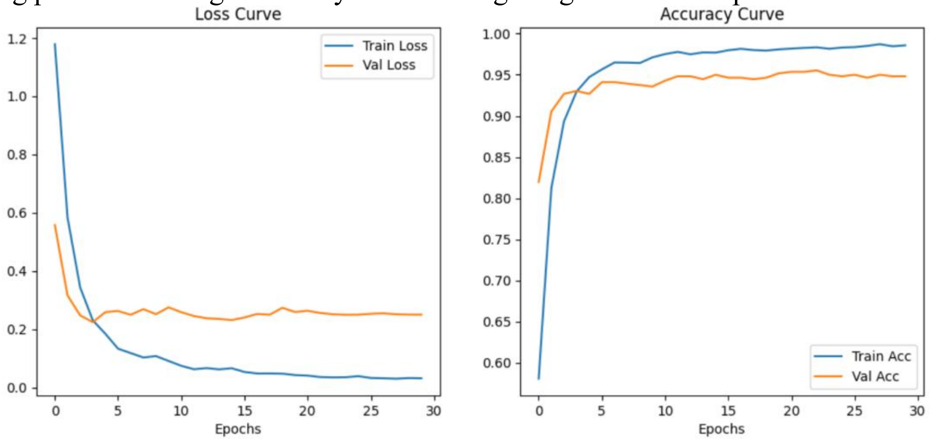

Figure 7. Training Process Curves

4.3.2 Confusion matrix and misjudgment analysis

To further explore the specific classification details of the model, we plotted the confusion matrix on the test set, as shown in Figure 8.

From Figure 8, it can be intuitively seen that for Quartz and Pyrite with sufficient sample sizes, the model correctly identified 113 and 108 images respectively, with very few misjudgments. For the Muscovite category with insufficient samples, out of 35 Muscovite samples in the test set, the model correctly identified 32, and only 3 were confused (misclassified as 1 Biotite, 1 Bornite, and 1 Quartz respectively). MA-ViT held the recognition baseline of over 90% under sample imbalance, proving its robustness in actual ore classification applications.

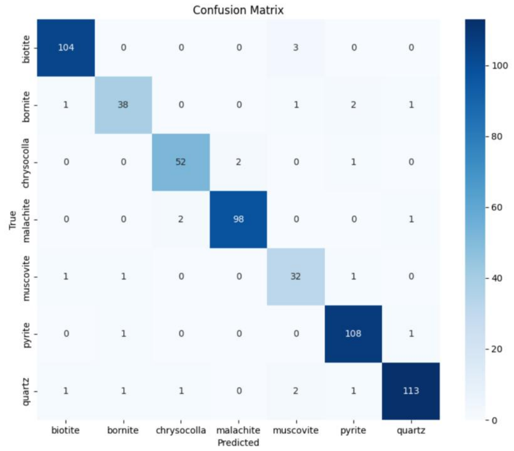

Figure 8. Confusion Matrix of the Test Set

4.3.3 Overall performance and ablation experiment analysis

To quantify the contribution of each module of MA-ViT, we compared this method horizontally with ResNet-50 and the original ViT. Detailed data is shown in Table 1.

Table 1. Comparison of Results of Different Algorithms

Model	Strategy	Overall Accuracy (ACC)	Macro F1	Muscovite Recall (Minority Class)	Muscovite Precision
ResNet-50	Baseline	94.04%	0.9212	80.00%	0.9032
ViT	Baseline	95.09%	0.9369	80.00%	0.9333
ViT+weighted	Weighted Loss	94.91%	0.9343	88.57%	0.7750
MA-ViT	CBDA+ Weighted Loss	95.61%	0.9345	91.43%	0.8421

From the perspective of baseline models, although ResNet-50 can also achieve an accuracy of 94.04%, it is slightly inferior to the original ViT (95.09%) in various indicators. This result to some extent breaks the stereotype that CNNs are more suitable for small samples, indicating that after appropriate pre-training, the Transformer architecture, with its global receptive field, indeed has advantages over local convolutions when capturing objects like ores with complex textures and non-fixed shapes.

As shown in Table 1, the pure ViT did not solve the long-tail problem; the recall rate of Vanilla ViT on the small sample category Muscovite was only 80.00%. When the class-weighted loss strategy was introduced (i.e., ViT + Weighted model), the recall rate of Muscovite increased significantly by 8.6% (reaching 88.57%), proving that the model began to pay attention to minority classes. However, the cost was that the model's overall accuracy dropped slightly to 94.91%, and the precision for Muscovite experienced a precipitous drop (falling to 0.7750). This indicates that although pure weighting forced the model to predict Muscovite more often, it also caused the model to become overly sensitive, easily misclassifying some similar ores as Muscovite, bringing more false positives.

By introducing the CBAM attention mechanism, we endowed the model with stronger feature discrimination capabilities. Data shows that MA-ViT not only pushed the overall accuracy to the highest 95.61% but also, while maintaining a rebound in Muscovite precision (0.8421), further pulled its recall rate up to 91.43%. Compared with the baseline model, a huge improvement of 11.4% was achieved in the small sample category.

5. Conclusion

This study is based on the cross-application of computer vision technology in the field of geology, focusing on overcoming the difficult problem of fine-grained classification of ore images under complex backgrounds. Addressing the current status of a lack of high-quality annotated data in existing research, this paper used an image dataset containing seven typical rock-forming minerals exhibiting significant data imbalance, providing key data support for subsequent algorithm research. On this basis, this paper innovatively proposed an MA-ViT (Mixed-Attention Vision Transformer) classification model fusing a mixed attention mechanism and cost-sensitive optimization, and verified its effectiveness in solving core problems such as difficulty in ore feature extraction and uneven sample distribution through systematic comparative experiments and ablation analysis. Experimental data shows that MA-ViT outperforms mainstream baseline models such as ResNet-50 and the original Vision Transformer in various classification indicators. In particular, benefiting from the keen capture of mineral optical texture features by the CBAM module and the gradient correction for scarce samples by the class-weighted cross-entropy loss, the model successfully broke the recognition dilemma of "majority class dominance, minority class submersion" commonly faced by long-tail data. The research results not only provide a high-precision algorithm scheme for intelligent auxiliary classification of ores but also offer referential technical ideas for processing other types of long-tail data.

References

[1] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. CoRR, 2015, abs/1512.03385.

[2] Deng T, Yu Y. Research on Ore Recognition and Classification Based on Improved PSO-Faster R-CNN Algorithm. Mining Research and Development, 2021, 41 (02): 178-182.

[3] Xiao C, Li Q, Li H, et al.Ore Type Detection Algorithm Based on Improved Mask R-CNN. Sintering and Pelletizing, 2024, 49 (02): 65-73+106.

[4] Zou Y, He J, Xia F, et al. Ore image

classification method of dual-energy X-ray based on interactive feature fusion. Nonferrous Metals (Extractive Metallurgy), 2025, 15(06): 990-998.

[5] Gao Y, Lv F, Feng Y. Ore image classification algorithm based on cross-channel fine-grained feature fusion. Computer Engineering and Applications, 2025, 61(10): 214-227.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[7] Jin Y, Sun J, Luo J, et al. Development and application of grading recognition model for tobacco weather fleck based on Cu-ViT deep learning. Plant Medicine, 2025, 4(06):65-76.

[8] Chen C F, Panda R, Fan Q. Regionvit: Regional-to-local attention for vision transformers. arXiv preprint arXiv:2106.02689, 2021.

[9] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[10]Batista A P A E G, Prati C R, Monard C M. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 2004, 6 (1): 20-29.

[11]Cui Y, Jia M, Lin T, et al. Class-Balanced Loss Based on Effective Number of Samples. CoRR, 2019, abs/1901.05555.