*2025 International Conference on Intelligent Computing*
*July 26-29, Ningbo, China*
https://www.ic-icc.cn/2025/index.php

# Geological Layer-Aware Cross-Modal Learning for Multi-Label Drill Cuttings Classification

Shuying Cui[1], Jianwei Niu[1], Xuefeng Liu[1(✉)], Yan Ding[2]

[1] School of Computer Science, Beihang University, Beijing 100191, China.
liu_xuefeng@buaa.edu.cn
[2] CNPC Engineering Technology R&D Company Limited,
Beijing 102206, China

**Abstract.** Multi-label drill cuttings classification reveals the current lithologies during drilling, which is crucial for guiding oil drilling operations. Many existing methods rely on annotated images for feature extraction, making their accuracy highly dependent on dataset size. However, due to equipment and manpower constraints, datasets in this field are generally small, posing a significant challenge for improving traditional methods for accurate multi-label classification. Notably, we observed that geological layers influence the distribution of drill cuttings, highlighting the importance of effectively leveraging geological priors for classification. In this paper, we propose a geological layer-aware cross-modal learning framework, which explicitly leverages local layer-wise information and global label co-occurrence patterns for multi-label drill cuttings classification. Unlike conventional end-to-end models, our framework first estimates the geological layer of a given image and derives a corresponding cuttings proportion vector. These priors are then employed to guide the alignment between visual and textual features, leading to more precise visual representations. Furthermore, we introduce a global co-occurrence matrix that captures label dependencies and enhances the learning of visual representations through a graph convolutional network (GCN), resulting in more accurate label predictions. Experiments on our dataset demonstrate that our approach significantly outperforms state-of-the-art methods, achieving a mean average precision (mAP) of 98.8%.

**Keywords:** Multi-label Image Classification, Drill Cuttings, Layer-Wise Proportion, Co-Occurrence Matrix.

## 1 Introduction

Multi-label image classification (MLIC) aims to predict multiple labels for an image, serving a wide range of practical applications, such as image retrieval [1] and scene

understanding [2]. During petroleum drilling, particles and debris collected from the bottom of the well are processed and photographed to obtain drill cuttings images. Typically, a single cuttings image contains a mixture of various rock fragments, such as sandstone, mudstone, limestone, shale, and volcanic rock. However, many existing methods focus on performing single-label classification for rock fragments [3, 4], which may lead to incomplete results when applied to real-world multi-lithology cuttings images. MLIC can effectively identify the lithology of drill cuttings, revealing the geological composition and aiding both resource exploration [4] and drilling operations.

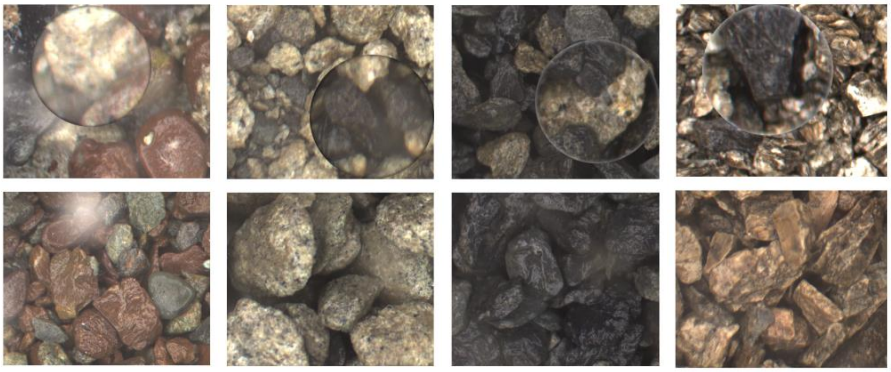


**Fig. 1.** Example Images of Drill Cuttings: Mixed vs. Pure Lithologies. The first row shows dominant lithologies mixed with a small amount of other rock types: mudstone with sparse sandstone, sandstone with sparse shale, shale with sparse sandstone, and limestone with sparse shale. The second row shows pure samples of the dominant lithologies: mudstone, sandstone, shale, and limestone.

Many existing methods for multi-label drill cuttings classification rely on directly extracting features from annotated images. While these approaches can achieve promising results, their accuracy largely depends on the size and quality of the dataset. However, collecting high-quality color images of drill cuttings remains a major bottleneck during drilling due to equipment constraints and limited human resources, making accurate classification even more challenging with conventional methods. To address this limitation, it is essential to explore the potential of leveraging the unique prior knowledge of drill cuttings datasets to enhance classification. As shown in Fig. 1, each geological layer of our dataset corresponds to a dominant lithology, which constitutes the majority of the images collected from that layer, accompanied by a small amount of other lithologies. This observation suggests that the lithologies in mixed drill cuttings images are strongly influenced by the geological layer from which they originate. Moreover, images collected from the same layer share nearly identical dominant lithology, resulting in highly similar visual features that facilitate layer division. Thus, we consider that identifying the geological layer of an image and incorporating layer-wise prior knowledge into the model can effectively enhance classification accuracy. It is

important to clarify that the estimated layers here do not contain lithological information. Instead, we group images from the same geological layer mainly to collect the cuttings proportion for each layer.

In this paper, we propose a novel framework leveraging prior knowledge within images for multi-label drill cuttings classification. Unlike conventional end-to-end models that primarily rely on visual features, our approach incorporates geological layer information and label co-occurrence patterns to refine feature learning and improve classification performance. Specifically, we first estimate the geological layer of each image and construct a proportion vector to represent the distribution of various lithologies within that layer. This vector is initialized as zero and iteratively updated during training by mapping the visual embeddings of images to their corresponding layers and accumulating the proportion annotations. In other words, our approach is similar to biological sampling, where we infer the cuttings proportions of different layers based on images obtained during drilling. As a prior, the proportion vector guides the alignment between visual and semantic features, resulting in more precise visual representations. Furthermore, we develop a global co-occurrence matrix to capture label correlations among different lithologies. By integrating both visual representations and the co-occurrence matrix into a graph convolutional network, our model explores the interactions among visual features corresponding to different labels, leading to more robust label predictions.

The key contributions of this paper are as follows:

- We propose a layer-aware proportion-guided fusion module that incorporates geological priors to guide the alignment of multiple modalities.
- We introduce a global co-occurrence matrix to model label dependencies and improve prediction robustness through interactions among visual representations.
- Extensive experimental results demonstrate that our method achieves state-of-the-art performance in multi-label drill cuttings classification.

## 2 Previous Work

Multi-label image classification methods for natural images have reached a high level of maturity. Early approaches leveraged problem transformation techniques, breaking down multi-label image classification into multiple single-label classification tasks [9]. Some methods [10, 11] employ convolutional neural networks to extract visual features and treat MLIC as a series of binary classification problems.

Transformer-based models have also emerged, such as Q2L [14], which utilizes a transformer decoder to facilitate interactions between labels and image features. C-Tran [8] introduces a transformer framework to model the complex dependencies between visual features and labels. TDRG [20] proposes a transformer-based dual relation learning framework that constructs structural and semantic relation graphs to enhance multi-label image recognition.

Graph-based approaches have also been investigated for MLIC. SSGRL [7] enhances classification by leveraging semantic decoupling and graph-based semantic interactions. GKGNet [23], the first fully graph convolutional model for MLIC, dynamically constructs graphs to efficiently capture semantic and spatial relationships.

Some methods integrate multi-modal information to enhance classification accuracy. TSFormer [6] employs a two-stream transformer framework with cross-modal interactions between visual and semantic features through a multi-shot attention mechanism. Moreover, the M3TR [22] integrates visual and linguistic modalities through semantic cross-attention and linguistic-guided enhancement to improve multi-label image recognition. Other methods [12, 15, 18] focus on modeling the relationships between image regions or labels to further improve multi-label classification performance.

Although existing methods effectively leverage deep learning architectures for MLIC, they often overlook the geological priors and co-occurrence relationships inherent in drill cuttings images. Our approach addresses this gap by explicitly incorporating geological layer information and global label dependencies, resulting in more accurate classifications.

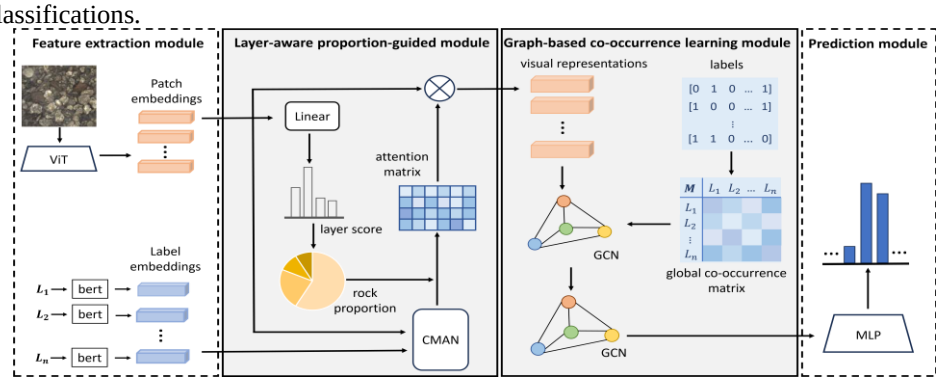



**Fig. 2.** Overall framework of our approach.

## 3 Methodology

Our framework consists of four main components: the feature extraction module, layer-aware proportion-guided fusion module, graph-based co-occurrence learning module, and prediction module. As illustrated in Fig. 2, we first encode the input drill cuttings image to extract visual feature maps, while semantic labels are processed to generate semantic embeddings. The two modalities are then fed into the layer-aware proportion-guided fusion module, where they undergo cross-modal fusion guided by layer-wise cuttings proportion to obtain visual representations. Next, the graph-based co-occurrence learning module leverages a global co-occurrence matrix to model label dependencies, which then guides interactions among visual representations via a graph convolutional network. Finally, the refined label-specific visual representations are fed into a set of binary classifiers to determine the presence of each label in the image. The following sections provide a detailed introduction to four components of our framework.

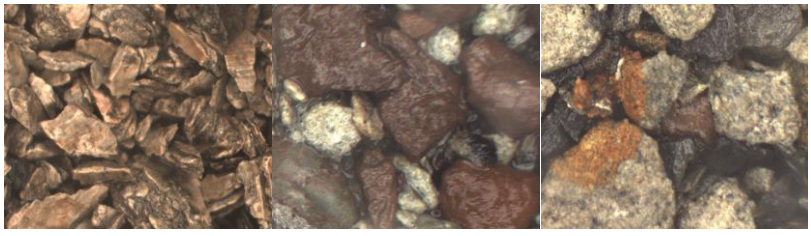
### 3.1 Feature extraction module

**Visual feature extraction.** Since each image in our dataset is densely populated with various cuttings, we employ Vision Transformer (ViT) [17] to capture global relationships and extract more discriminative visual features. Given a drill cuttings image $I$, we divide it into m non-overlapping 2D patches, which serve as the input to the encoder. The Positional embeddings are incorporated after the patch embedding layer to enhance the input sequence. Finally the ViT encoder with $L$ layers processes this sequence via self-attention mechanisms, generating the visual features $V \in R^{h \times w \times d}$. Here, $h$ and $w$ denote the spatial dimensions of the feature map, while $d$ represents the fixed dimensionality of the transformer's embedding space.

**Semantic feature extraction.** For a candidate label set, a pretrained language model BERT [19] is employed to generate the semantic embeddings, resulting in a set of representations $T = \{t_1, t_2, \ldots, t_n\}$, where $t_i \in R^d$, $n$ represents the number of labels, and $d$ is the dimension of the label embeddings.

### 3.2 Layer-Aware Proportion-Guided Fusion Module

To capture relationships between cuttings and labels more effectively, we introduce a Layer-Aware Proportion-Guided Fusion Module. This module fuses visual features $V$ and semantic embeddings $T$ while integrating prior proportion knowledge into the attention mechanism. In this section, we first construct the layer-wise proportion vectors, followed by a detailed explanation of our cross-modal attention network.



$s_{limetone} = 2$ | $s_{mudstone} = 2$, $s_{sandtone} = 1$ | $s_{sandstone} = 2$, $s_{mudstone} = 1$, $s_{shale} = 1$

**Fig. 3.** Example images of drill cuttings with proportion annotations.

**Proportions Initialization.** Our drill cuttings dataset differs significantly from natural images, each image contains cuttings filling every corner of the image, with varying

proportions of different lithologies. In most images, only one lithology dominates, while others are scattered sparsely throughout. Therefore, considering the nature of drill cuttings, we introduce a new attribute state s in the annotation of each object within the image. The state $s$ is set to 2 if the image contains only one lithology and to 0 if the lithology is absent. If multiple lithologies are present, the dominant one is assigned $s = 2$, and the others are labeled with $s = 1$, as shown in Fig. 3. Thus, we can simplify the drill cuttings proportion states of each image with the vector $S = \{s_0, s_1, \dots, s_n\}$, where $s_i \in \{0,1,2\}$ for $i \in \{0, 1, \dots, n-1\}$.

**Layer Detection.** The dominant lithology varies across layers and covers the majority of the image, resulting in layer-level distinct image features. Therefore, following [5], we employ a softmax classifier, leveraging contextual image features to identify the geological layer of the input image $I$. It's important to clarify that the geological layer dection of input images here is purely clustering-based, without any lithology labels. The layers are simply designated as $Layer1$, $Layer2$, etc. We first apply global average pooling to the visual features extracted from the last block of the ViT:

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m} V_i \tag{1}$$

where m refers to the number of non-overlapping 2D patches. Given a set of $C$ layers, we use a linear mapping to compute the score of cth layer $f_c$ as follows:

$$f_c = W_c^T \bar{V} \tag{2}$$

where $W_c \in R^d$ is a learnable parameter vector. Based on the computed probability distribution, the input image $I$ is assigned to the geological layer $g$ with the highest probability, formulated as:

$$g = arg \max_{c \in \{0,1,\dots,C-1\}} \frac{exp(f_c)}{\sum_{j=0}^{C-1} exp(f_j)} \tag{3}$$

**Proportions calculation.** The cuttings proportion of each geological layer is obtained by aggregating the proportion state vectors of input images from the same layer, as shown in Fig. 4. The accumulated cuttings composition for geological layer $g$ is represented as:

$$P_g = P_g + S_i \tag{4}$$

where $P_g$ is iteratively updated as more images are processed, and $S_i$ represents the proportion state vector of image $i$ belonging to geological layer $g$.

To ensure effective guidance for the subsequent cross-modal fusion, we normalize $P_g$ to obtain the final proportion distribution $\widehat{P_g}$ as follows:

$$\widehat{P_g} = \frac{P_g}{\sqrt{\sum_{i=0}^{n-1} p_i^2}} \tag{5}$$

where $p_i$ represents the element of vector $P_g$. Then the proportion vectors $\hat{P} \in R^{1\times n}$ for all layers will be computed and sent for guiding subsequent cross-modal fusion.

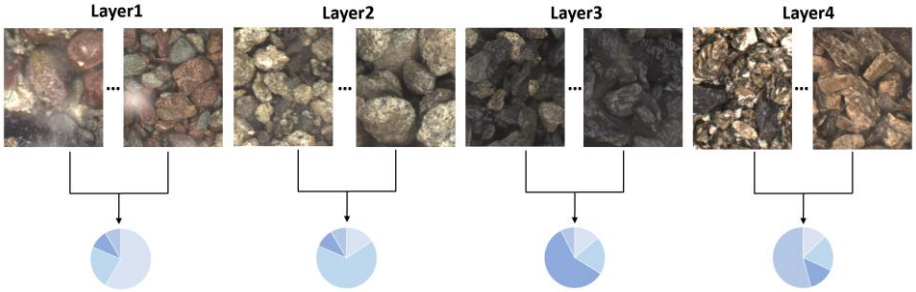



**Fig. 4.** Proportion calculation of each layer.

**Cross Modal Attention Network(CMAN).** In this work, we modify [6] to design our cross modal attention network. At the $l$-th layer, visual features from the corresponding ViT block $\mathcal{F}_{\mathcal{V}}^{\ell} \in R^{m\times d}$ and semantic embeddings $\mathcal{F}_{\mathcal{T}}^{\ell-1} \in R^{n\times d}$ are fused via multi-head self-attention, generating an updated visual representation $\mathcal{F}_{\mathcal{T}}^{\ell}$. Specifically, we use semantic embeddings as the query while visual features as the key and value:

$$\mathcal{Q}_{\mathcal{T}}^{\ell} = \mathcal{F}_{\mathcal{T}}^{\ell-1} W_Q \tag{6}$$

$$\mathcal{K}_{\mathcal{V}}^{\ell} = \mathcal{F}_{\mathcal{V}}^{\ell} W_K \tag{7}$$

$$\mathcal{V}_{\mathcal{V}}^{\ell} = \mathcal{F}_{\mathcal{V}}^{\ell} W_V \tag{8}$$

where$W_Q \in R^{d\times d}$, $W_K \in R^{d\times d}$, and$W_V \in R^{d\times d}$ are parameter matrices to be learned.

Then, we modify the computation of traditional alignment matrix by introducing the drill cuttings proportion vectors. To prevent the attention values from becoming too small after the proportion vector is multiplied with $\mathcal{Q}_{\mathcal{T}}^{\ell}$ and $\mathcal{K}_{\mathcal{V}}^{\ell}$, resulting in the loss of important information, we first add a bias to all elements of the vector, ensuring numerical stability during subsequent normalization, preventing division errors:

$$\hat{P} = \hat{P} + \varepsilon \tag{9}$$

Next, we reshaped the layer-wise proportion vector by expanding its dimensions and repeating its values to match $\mathcal{Q}_{\mathcal{T}}^{\ell}$ and $\mathcal{K}_{\mathcal{V}}^{\ell}$. Finally, we incorporate the new layer-wise proportion vector $\widehat{P'}$ into the calculation of the attention matrix:

$$\mathcal{F}_{\mathcal{A}}^{\ell} = softmax\left(\frac{\mathcal{Q}_{\mathcal{T}}^{\ell}\left(\mathcal{K}_{\mathcal{V}}^{\ell}\right)^{T} \odot \widehat{P'}}{\sqrt{d}}\right) \tag{10}$$

The attention matrix $\mathcal{F}_{\mathcal{A}}^{\ell} \in R^{n \times m}$ captures the correlations between labels and image patches. We then apply a residual connection to refine the semantic representation at the $i$-th layer, formulated as:

$$\mathcal{F}_{\mathcal{T}}^{\ell} = \mathcal{F}_{\mathcal{T}}^{\ell-1} + \mathcal{F}_{\mathcal{A}}^{\ell} \mathcal{V}_{\mathcal{V}}^{\ell} \tag{11}$$

In this way, after passing through L layers of fusion and interaction, we obtain the final-layer visual representation $\mathcal{F}_{\mathcal{T}}$ that integrates all visual information, which is then used for subsequent semantic interactions.

**Semantic interaction module.** After integrating layer-level prior knowledge to guide feature fusion, we identified an issue. In layer dominated by lithology A, lithology B occupies a small proportion, and vice versa in layer dominated by lithology B. Despite the low correlation indicated by the layer-wise proportions, the overall label co-occurrence probability between these two lithologies remains high. Therefore, focusing solely on the layer-level proportion while ignoring the overall label co-occurrence may result in the loss of valuable prior information.

In this module, we introduce a global co-occurrence matrix to further explore the interaction between label representations. We represent the multi-labels of an image using the vector $v = [v_0, v_1, \ldots, v_{n-1}]^T$, where $v_i = 1$ indicates that label $t_i$ exists, and $v_i = 0$ otherwise. The overall dependency matrix $M \in R^{n \times n}$ is computed as the sum of the outer products of the label vectors across all images. This matrix is symmetric, where $M_{ii}$ represents the total occurrences of label $t_i$, and $M_{ij}$ records the number of times labels $t_i$ and $t_j$ co-occur in the dataset.We normalize each row by dividing its elements by the corresponding diagonal value, ensuring that the diagonal elements become 1, while off-diagonal elements represent the conditional co-occurrence probability:

$$M'_{ij} = \frac{M_{ij}}{M_{ii}} \tag{12}$$

Next, the visual representations of labels interact with each other in the multi-layer graph convolution network under the guidance of the global label co-occurrence probability matrix as follows:

$$\mathcal{H}_{\mathcal{T}}^{(\ell+1)} = ReLU\left(M' \mathcal{H}_{\mathcal{T}}^{(\ell)} W^{(l)} + b^{(l)}\right) \tag{13}$$

where $\mathcal{H}_{\mathcal{T}}^{(\ell)} \in R^{n \times d}$ serves as the input to the $(l+1)$-th layer of the network, and initialized as $\mathcal{H}_{\mathcal{T}}^{(0)} = \mathcal{F}_{\mathcal{T}}$, $W^{(l)} \in R^{d \times d'}$ is the learnable weight matrix and $b^{(l)} \in R^{d'}$ is the bias term. Finally, the updated visual representations $\mathcal{H}_{\mathcal{T}}$ are obtained, where each representation integrates both its intrinsic label characteristics and contextual information from other labels.

**Prediction module.** The ground truth labels of an image I are represented as $v \in R^n$, where $v_i = 1$ indicates indicates the presence of label $t_i$, and $v_i = 0$ otherwise. With

the output $\mathcal{H}_{\mathcal{T}}$ from the Semantic interaction module, we compute the predicted probability vector $\mathrm{p} = [\mathrm{p}_0, \mathrm{p}_1, \ldots, \mathrm{p}_{n-1}]^{\mathrm{T}}$ for image I using a set of binary classifiers:

$$\mathrm{p} = \sigma\big(\mathrm{f}(\mathcal{H}_{\mathcal{T}} \odot \mathrm{W}_{\mathrm{C}})\big) \tag{14}$$

where $W_C \in R^{n \times d}$ is a learnable matrix, with each row serving as the weight vector of a binary classifier. $\odot$ denotes the element-wise multiplication and $\mathrm{f}(\cdot)$ performs row-wise summation. The sigmoid function $\sigma(\cdot)$ converts values into probabilities.

The model is trained using the binary cross-entropy loss, defined as:

$$\mathcal{L} = -\sum_{i=0}^{n-1}[v_i \, log \, p_i + (1 - v_i) \, log(1 - p_i)] \tag{15}$$

# 4 Experiments

## 4.1 Datasets

Our dataset consists of 3,984 RGB multi-label images of drill cuttings, collected from a development well on the southern flank of the Zhongao Top Structure in Weiyuan, Sichuan Basin, with samples taken from depths ranging from 20m to 2,825m underground. Each image has a resolution of 448×448 pixels and has been professionally annotated.

The dataset comprises four lithology labels: mudstone, sandstone, shale, and limestone, where each image is labeled with one to three lithologies. We define four geological layers, with each layer dominated by one of the four lithologies. The distribution of dominant lithologies are as follows: limestone accounts for 31%, mudstone for 42%, sandstone for 11%, and shale for 16%. Sandstone and mudstone are dominant at shallower depths, while shale and limestone are more prevalent above 1,000 meters.

## 4.2 Implementation Details

We leverage ViT-B16, pretrained on ImageNet21k, to extract visual features, while BERT, pretrained on Wikipedia data, is used to obtain textual embeddings. The feature dimension $d$ is set to 768, and the graph convolutional network consists of five layers. The model is optimized using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 16. The initial learning rate is set to $1\mathrm{e}^{-5}$ and decays by a factor of 10 when the loss stabilizes. To enhance training robustness, we apply random horizontal flipping and random resized cropping as data augmentation techniques. All experiments are conducted on a server equipped with two NVIDIA 4090 GPUs.

### 4.3 Comparison with State-of-the-Art Methods

We split the annotated dataset into training, validation, and test sets with a ratio of 70%, 10%, and 20%, respectively. The models selected for comparison included MCAR [16], TDRG [20], TSFormer [6], MLD-TResNetL-AAM [15], MSRN [13], IDA [21], and M3TR [22]. These methods represent state-of-the-art multi-label classification approaches on public datasets in recent years and can be replicated on our dataset due to their compatible data formats. The experimental results are presented in Table 1 and Table 2, with the best scores highlighted in bold.

**Table 1.** Experiments results on the drill cuttings dataset in terms of class-wise precision (AP in %) and mean average precision (mAP in %)

| Method | Limestone | Mudstone | Sandstone | Shale | mAP |
|---|---|---|---|---|---|
| MCAR [16] | 99.8 | 98.4 | 94.1 | 90.9 | 95.8 |
| TDRG [20] | 99.9 | 99.3 | 94.7 | 91.1 | 96.3 |
| TSFormer [6] | 99.9 | 98.7 | 93.8 | 94.5 | 96.7 |
| MLD-TResNetL-AAM [15] | 99.5 | 97.3 | 96.3 | 94.8 | 96.9 |
| MSRN [13] | **100** | 98.9 | 96.6 | 94.7 | 97.5 |
| IDA [21] | 99.3 | 98.4 | 95.3 | 94.9 | 96.9 |
| M3TR [22] | **100** | 99.0 | 97.7 | 96.1 | 98.2 |
| Ours | **100** | **99.5** | **98.0** | **97.9** | **98.8** |

To evaluate model performance, we first calculate the average precision for each of the four lithologies, along with the mean average precision (mAP) across all categories. As shown in the Table 1, our framework achieves the best overall performance, significantly surpassing all previous state-of-the-art methods in terms of mean average precision (mAP), with a score of 98.8%. Notably, it outperforms the second-highest scores by 0.3% and 1.8% in distinguishing sandstone and shale classification, two of the most challenging lithologies, as they often appear as minor components in images where other lithologies are dominant. Moreover, our framework achieves 100% precision for limestone classification, ensuring accurate recognition of limestone formations during drilling operations.

**Table 2.** Experiments results on the drill cuttings dataset in terms of precision, recall and F1 scores (in %)

| Method | CP | CR | $CF_1$ | OP | OR | $OF_1$ |
|---|---|---|---|---|---|---|
| MCAR [16] | 90.7 | 86.2 | 88.4 | 91.9 | 86.5 | 89.1 |
| TDRG [20] | 95.6 | 82.8 | 88.7 | 93.2 | 85.1 | 88.9 |
| TSFormer [6] | 89.4 | 92.1 | 90.7 | 88.6 | 93.6 | 91.0 |
| MLD-TResNetL-AAM [15] | 93.6 | **96.0** | 93.3 | 90.5 | 94.7 | 92.6 |
| MSRN [13] | 91.8 | **96.0** | 93.8 | 91.7 | **96.7** | 94.1 |
| IDA [21] | 93.3 | 93.9 | 93.6 | 93.6 | 94.4 | 94.0 |
| M3TR [22] | 95.0 | 93.3 | 94.1 | 94.9 | 93.8 | 94.3 |
| Ours | **96.8** | 94.2 | **95.5** | **96.3** | 95.1 | **95.7** |

To further assess model effectiveness, we calculated the precision, recall, and F1-measure for each method. As shown in Table 2, our approach achieves the highest scores across all key evaluation metrics, including $CF_1$ and $OF_1$, and outperforms all other methods in most remaining metrics. However, we observed that the recall values were slightly lower than precision. This could be attributed to the imbalanced sample distribution, where shale and sandstone are underrepresented compared to mudstone and limestone. Additionally, the morphological complexity of these lithologies may further impact the recall performance.

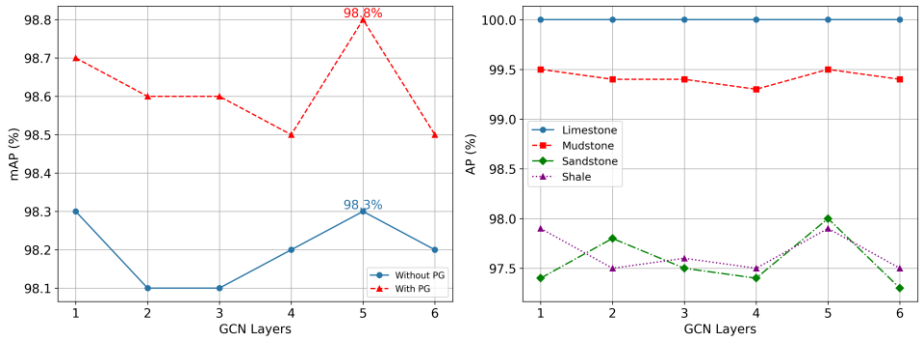
### 4.4 Ablation Study

In this section, we conduct ablation experiments to evaluate the impact of key points in our model and the number of layers in the graph convolutional network on prediction probabilities.

We first examine the effectiveness of the proportion vector in layer-aware proportion-guided fusion module and the graph-based co-occurrence learning (GCM) module. As shown in Table 3, adding the proportion guidance (PG) and GCM individually to the baseline increases the mAP from 96.7 to 98.2 and 98.3, respectively. When both modules are incorporated, the mAP further improves to 98.8. These significant gains clearly demonstrates the crucial role of geological prior knowledge in extracting accurate features and enhancing classification performance.

**Table 3.** The effect of key contributions on the performance of our framework (mAP in %)

| Method | mAP(%) |
|---|---|
| Baseline | 96.7 |
| Baseline + PG | 98.2 |
| Baseline + GCM | 98.3 |
| Baseline + PG + GCM (Ours) | 98.8 |




(a) (b)

**Fig. 5.** The effect of GCN layers on the performance of our framework. (a) mAP (%) with and without proportion guidance. (b) AP (%) of different lithologies at different GCN layers under whole framework.

We also explored the impact of the number of layers in the graph convolutional network (GCN) within the GCM module on mAP. Fig. 5 (a) illustrates the overall mAP values across different GCN layers. Regardless of the presence of the proportion guidance, the mAP reaches its peak when $L' = 5$, achieving 98.3% and 98.8% for the cases without and with proportion guidance, respectively. In the case with only GCM, the mAP also achieves 98.3% when $L' = 1$.

Fig. 5 (b) presents the prediction accuracy of each lithology at different GCN layers under our complete framework. Limestone is consistently predicted with 100% accuracy across all layers. This suggests that limestone has distinctive visual characteristics that make it easier to classify, in contrast to shale and sandstone, which exhibit more intra-class variability. Shale and mudstone achieve their highest accuracy when $L' = 1$ and $L' = 5$, reaching 97.9% and 99.5%, respectively. Sandstone achieves its peak accuracy of 98.0% when $L' = 5$. Considering both classification accuracy and computational efficiency, we set the number of GCN layers to 5 in our final experimental setup.

### 4.5 Visualization

In this section, we first visualize the distribution of dominant lithology across different layers. As shown in Fig. 6, the horizontal axis represents different lithologies, while the vertical axis indicates the number of images where a specific lithology serves as the dominant lithology. The results show that the true dominant lithology varies across the four layers, which aligns with our expectations. Due to the mixed distribution of rock fragments, some images may have a different lithology appearing as dominant. However, in each layer, the true dominant lithology consistently appears in more than 50% of the images, significantly outnumbering other lithologies. This confirms the distinct lithological variations across geological layers in our dataset, and suggests that the Softmax classifier effectively identifies the geological layer of a given image, providing a solid foundation for collecting layer-wise cuttings proportions.

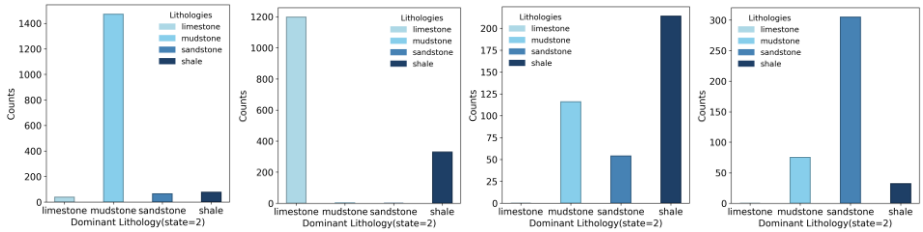



**Fig. 6.** Histogram of dominant lithology distribution of each layers.

Subsequently, we utilize cross-attention maps to evaluate the model's classification performance across the four labels. Due to the overlapping nature of rock fragments in our dataset, the CAM heatmaps may not always capture entire fragments. Nevertheless, as shown in Fig. 7, the model successfully attends to the key regions associated with each lithology, demonstrating its effectiveness in precise classification.

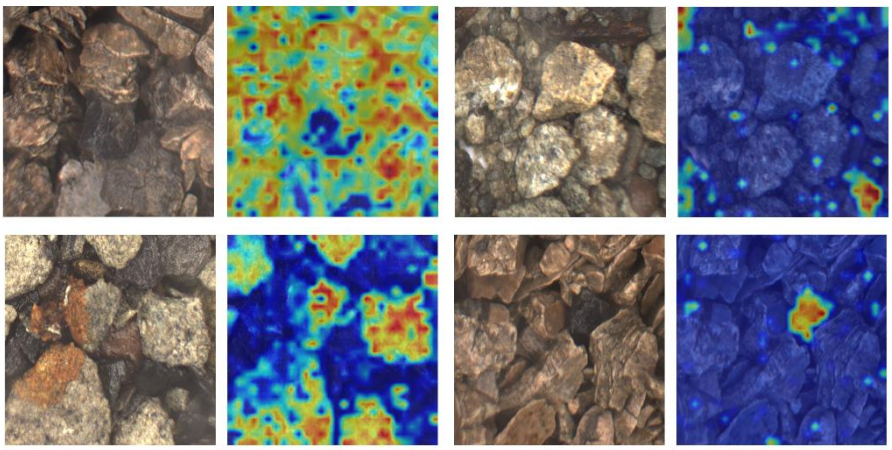


**Fig. 7.** Cross-attention maps for all labels.

## 5 Conclusion

We introduce a geological layer-aware cross-modal learning framework for multi-label drill cuttings classification, which effectively integrates geological priors to enhance feature learning. By leveraging local layer-wise information and global label co-occurrence patterns, our approach significantly improves classification accuracy, achieving a new state-of-the-art mAP of 98.8%. The experimental results demonstrate its superiority over conventional methods. Furthermore, our approach highlights the importance

of incorporating domain-specific knowledge in image classification, providing valuable insights for applications in geological analysis and other specialized scientific fields.

**Acknowledgments.** This study was funded by the National Science and Technology Major Project for New Oil and Gas Exploration and Development (grant number: 2024ZD1401806), and the Fundamental and Forward-looking Science and Technology Project of CNPC Intelligent Program (grant number: 2023ZZ06).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Wei, S., Liao, L., Li, J., Zheng, Q., Yang, F., Zhao, Y.: Saliency inside: Learning attentive CNNs for content-based image retrieval. IEEE Trans. Image Process. **28**(9), 4580--4593 (2019)
2. Shao, J., Kang, K., Change Loy, C., Wang, X.: Deeply learned attributes for crowded scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4657--4666. IEEE (2015)
3. Qiqi, Z., Sai, W., Shun, T., Liangbin, Y., Kunlun, Q., Qingfeng, G.: RockS2Net: Rock image classification via a spatial localization siamese network. Comput. Geosci. **185**, 105560 (2024)
4. Hao, H., Jiang, Z., Ge, S., Wang, C., Gu, Q.: Siamese adversarial network for image classification of heavy mineral grains. Comput. Geosci. **159**, 105016 (2022)
5. Zhu, X., Liu, J., Liu, W., Ge, J., Liu, B., Cao, J.: Scene-aware label graph learning for multi-label image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1473--1482. IEEE (2023)
6. Zhu, X., Cao, J., Ge, J., Liu, W., Liu, B.: Two-stream transformer for multi-label image classification. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 3598--3607. ACM (2022)
7. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 522--531. IEEE (2019)
8. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16478--16488. IEEE (2021)
9. Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Yan, S.: HCP: A flexible CNN framework for multi-label image classification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1901--1907 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770--778. IEEE (2016)
11. Jia, J., Chen, X., Huang, K.: Spatial and semantic consistency regularizations for pedestrian attribute recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 962--971. IEEE (2021)
12. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: A unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285--2294. IEEE (2016)

13. Qu, X., Che, H., Huang, J., Xu, L., Zheng, X.: Multi-layered semantic representation network for multi-label image classification. Int. J. Mach. Learn. Cybern. **14**(10), 3427--3435 (2023)
14. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint https://arxiv.org/abs/2107.10834 (2021)
15. Sovrasov, V.: Combining metric learning and attention heads for accurate and efficient multilabel image classification. arXiv preprint https://arxiv.org/abs/2209.06585 (2022)
16. Gao, B.B., Zhou, H.Y.: Learning to discover multi-class attentional regions for multi-label image recognition. IEEE Trans. Image Process. **30**, 5920--5932 (2021)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint https://arxiv.org/abs/2010.11929 (2020)
18. Chen, T., Wang, Z., Li, G., Lin, L.: Recurrent attentional reinforcement learning for multi-label image recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, pp. 1--8. AAAI (2018)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171--4186. ACL (2019)
20. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 163--172. IEEE (2021)
21. Liu, R., Huang, J., Li, T.H., Li, G.: Causality compensated attention for contextual biased visual recognition. In: The Eleventh International Conference on Learning Representations (2022)
22. Zhao, J., Zhao, Y., Li, J.: M3TR: Multi-modal multi-label recognition with transformer. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 469--477. ACM (2021)
23. Yao, R., Jin, S., Xu, L., Zeng, W., Liu, W., Qian, C., Wu, J.: GKGNet: Group k-nearest neighbor based graph convolutional network for multi-label image recognition. In: European Conference on Computer Vision, pp. 91--107. Springer, Cham (2024)