

# Classification of Limestone Thin-Section Images Based on an Improved MobileNetV3 Incorporating RTEM and MFCM

**Tingting Wang**

Northeast Petroleum University

**Jianwei Lu**

Northeast Petroleum University

**Wanchun Zhao**

[zhaowanchun@nepu.edu.cn](mailto:zhaowanchun@nepu.edu.cn)

Northeast Petroleum University

**Ranjith P.G**

Monash University

---

## Research Article

**Keywords:** rock image classification, MobileNetV3, rock-texture feature enhancement, multi-source feature complementation, attention mechanism, deep supervision

**Posted Date:** October 14th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11041815/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

1 Classification of Limestone Thin-Section Images Based on  
2 an Improved MobileNetV3 Incorporating RTEM and  
3 MFCM

4 Tingting Wang <sup>a,b</sup>, Jianwei Lu <sup>a</sup>, Wanchun Zhao<sup>c,d,e</sup>, Ranjith P.G<sup>f</sup>

5 \* Corresponding Author :Wanchun Zhao, zhaowanchun@nepu.edu.cn

6 <sup>a</sup>School of Electrical Engineering & Information,Northeast Petroleum University, Daqing ,163318  
7 ,China.

8 <sup>b</sup>Heilongjiang Provincial Key Laboratory of Networking and Intelligent Control, Daqing,163318,China

9 <sup>c</sup>Northeast Petroleum University, Daqing ,163318 ,China.

10 <sup>d</sup>Innovation Team of Leading Scientific and Technological Innovation Talent Team: Shale Oil Geology  
11 and Engineering Integrated Fracturing and Prevention Technology, Daqing 163318, China;

12 <sup>e</sup>State Key Laboratory of Continental Shale Oil, Daqing 163318, China;

13 <sup>f</sup>Deep Earth Energy Laboratory, Department of Civil Engineering, Monash University, Building 60,  
14 Melbourne, Victoria 3800, Australia.

15

16 Keywords:

17 rock image classification

18 MobileNetV3

19 rock-texture feature enhancement

20 multi-source feature complementation

21 attention mechanism

22 deep supervision

23

24 Authorship contribution statement

Tingting Wang: Supervision, Resources, Methodology, Writing – review & editing. Jianwei Lu:  
Writing – original draft, Visualization, Validation, Methodology, Data curation. Wanchun Zhao:  
Supervision, Methodology, Investigation. Ranjith P.G<sup>f</sup>: Supervision, Investigation, Data curation,  
Writing – review & editing.

## Abstract

To address the challenges of substantial intra-class variation, highly coupled inter-class features, and extremely complex texture structures in limestone rock images, this paper proposes a limestone thin section image classification framework based on an improved MobileNetV3. The architecture adopts the MobileNetV3 inverted residual structure as its backbone and embeds a Rock-Texture Enhancement Module (RTEM) at different feature stages. Through an edge branch, a texture branch, and a lightweight attention mechanism, RTEM enhances grain contours, mineral boundaries, and local texture information in rock images from multiple perspectives. Meanwhile, a Multi-source Feature Complementation Module (MFCM) is designed to map shallow texture representations, intermediate-level structural descriptors, and deep semantic vectors into a common-scale space for channel-wise deep fusion, thereby supplementing the detail-preserving capability of the final high-level representation. To further strengthen the discriminative capability of intermediate representations, an auxiliary classification head is introduced during training for deep supervision. In addition, data augmentation, MixUp/CutMix, label smoothing, the AdamW optimizer, exponential moving average, multi-scale test-time augmentation, and weighted ensembling across multiple random seeds are jointly employed to improve generalization performance. Experimental results demonstrate that the proposed method achieves an accuracy of 91.11% on a six-class rock image recognition task, with macro-

averaged precision, recall, and F1-score reaching 91.28%, 91.11%, and 91.12%, respectively.

Compared with the original MobileNetV3, the improved framework significantly enhances rock classification performance while maintaining an extremely low parameter count.

## 1. Introduction

Lithology identification from petrographic thin sections is fundamental to geological surveying, hydrocarbon exploration, and reservoir evaluation. Thin sections preserve diagnostic evidence of mineral composition, grain texture, cementation, pore architecture, and diagenetic alteration, forming the primary basis for refined carbonate classification. Conventional petrographic analysis, however, relies on manual microscopic examination and subjective expert judgment, which is labor-intensive and sensitive to sample preparation, field-of-view selection, and inter-observer variability. These limitations are particularly acute for fine-grained carbonate subclasses (e.g., mudstone, wackestone, packstone, and grainstone), where ambiguous grain boundaries and gradational matrix content produce pronounced inter-class similarity and intra-class variability, making reliable automated discrimination a persistent challenge.

Intelligent lithology identification has been explored across multiple data modalities. For rock thin sections, Dong et al. (2025) integrated SENet with ConvNeXt and focal loss to strengthen feature representation and mitigate class imbalance. Fan et al. (2026) developed RoImAI, a vision foundation model that combines particle segmentation, hierarchical identification, and automated lithology report generation across multi-center datasets. Wei et al. (2026) proposed HFANet, coupling a DenseNet-based local branch with a Swin Transformer-based global branch and bidirectional cross-attention to fuse fine textures with contextual semantics. Zhang and Lin (2026) introduced the CG-AE-VAE data augmentation framework to increase intra-class diversity and alleviate data scarcity in small-sample classification. Beyond thin-section images, neural architecture search has been applied to lithology identification (Zhang et

al., 2025), gravity-magnetic data fusion has improved igneous lithology reliability (Xu et al., 2025), and multi-light polarized image fusion has demonstrated the value of complementary optical modalities for geological image segmentation (Zeng et al., 2026). Nevertheless, geophysical and multimodal data suffer from acquisition fluctuations, interpretive non-uniqueness, and demands for additional sensors and complex preprocessing. For thin-section-only fine-grained classification, deep convolutional networks offer a more direct pathway. Xu et al. (2021) established an end-to-end deep learning framework demonstrating that convolutional features can effectively discriminate lithologies from rock images. Zhao et al. (2026) proposed DEAE, a semi-supervised denoising-enhanced adaptive encoder that suppresses noise and improves robustness under limited annotation. Zhu et al. (2024) proposed RockS2Net, which combines a spatial localization mechanism with a Siamese architecture to jointly extract global and local discriminative features. Wang et al. (2025) proposed FSDS, a FasterNet variant embedding DualSE attention and ScConv to enhance spatial-channel features and suppress redundancy for efficient limestone photomicrograph classification. Despite these advances, extracting stable discriminative representations from textures, grain boundaries, and multi-scale structures within a single efficient architecture remains insufficiently addressed.

Lightweight networks provide a viable route toward practical deployment, and the MobileNet family has been widely adopted for thin-section identification owing to its small parameter count and fast inference. Liu et al. (2025) incorporated an attention mechanism into MobileNetV3 to strengthen global feature fusion for rock thin-section classification. However, the original MobileNetV3 (Howard et al., 2019) propagates features through a predominantly single-path topology, which gives rise to three deficiencies for fine-grained carbonate classification: (1) successive downsampling irreversibly degrades delicate textural cues such as grain boundaries, micritic matrix, pores, and cements, aggravating inter-class co

nfusion; (2) shallow texture features and deep semantic features lack explicit complementarity, limiting the representation of multi-scale geological structures; and (3) intermediate-layer features receive no direct classification supervision, which is especially problematic under limited samples and strong inter-class similarity.

To address these issues, this paper proposes an improved MobileNetV3 for limestone thin-section classification. First, a Rock-Texture Enhancement Module (RTEM) is designed with parallel edge and texture branches coupled through a lightweight attention mechanism, strengthening sensitivity to grain boundaries and textural patterns without sacrificing efficiency. Second, a multi-source feature complementarity module (MFCM) adaptively fuses mid- and shallow-level structural features with high-level semantics through a gated aggregation strategy, preserving fine-grained information across scales. Third, auxiliary classification heads with efficient channel attention are attached to intermediate layers to provide deep supervision, while multi-strategy data augmentation enhances generalization. Experiments demonstrate that the proposed method achieves a favorable balance between classification accuracy and lightweight efficiency for automated carbonate thin-section identification.

## 2. Methodology

The technical challenge of limestone image classification lies in the extremely subtle textural differences between classes; types such as mudstone, wackestone, and packstone exhibit high similarity in color distribution, grain boundaries, and local structures(Li et al.,2026). Although standard lightweight networks incur low computational cost, they tend to irreversibly attenuate fine-grained textural information during successive downsampling(Liu et al.,2021). To address this issue, this study adopts MobileNetV3 as the baseline and proposes an improved lightweight architecture for limestone image classification, which enhances the representational capacity for local textures and multi-scale semantic information

by simplifying the inverted residual structure, introducing a Rock-Texture Enhancement Module (RTE M), fusing multi-level representations, and optimizing the classification attention mechanism.

### 2.1. Overall Architecture of the Improved MobileNetV3

The overall framework of the improved architecture is illustrated in Figure 1, which comprises three components: feature extraction, feature fusion, and feature classification. The feature extraction component is built upon the inverted residual structure, with the Rock-Texture Enhancement Module (RTE M) embedded at different channel stages; the feature fusion component adopts the MFCM to aggregate multi-level representations; and the feature classification component employs auxiliary classification heads together with a main classification head to produce the final prediction. The framework is collaboratively designed to address three challenges: the susceptibility of fine details to loss, the high inter-class similarity, and the stringent lightweight constraints.

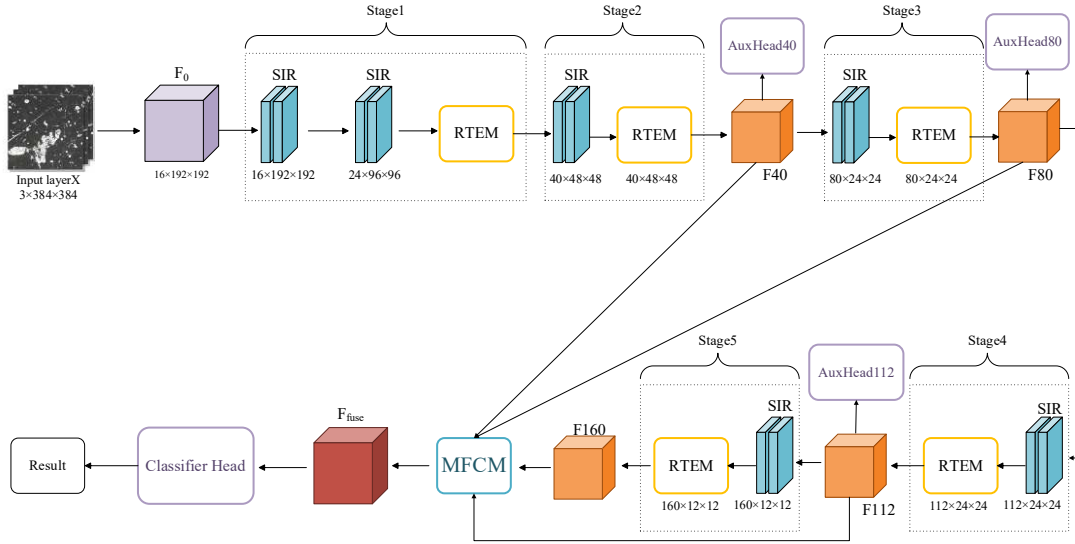


Figure 1: Overall framework of the improved model

(1) Feature extraction component. The input limestone image  $X$  passes through the initial convolutional layer, the inverted residual blocks, and the RTEM to yield multi-stage features:

$$\{F_{40}, F_{80}, F_{112}, F_{160}\} = f_{\theta}(X) \quad (1)$$

where  $f_{\theta}$  denotes the improved MobileNetV3 feature extraction network and  $\theta$  represents the set

t of learnable parameters.  $F_{40}$ ,  $F_{80}$ ,  $F_{112}$ , and  $F_{160}$  denote the feature maps at different channel stages, respectively.

(2) Feature fusion component. To avoid the loss of shallow textural information caused by relying solely on a single high-level feature, the multi-stage features are fed into the MFCM for fusion:

$$F_{fuse} = MFCM(F_{40}, F_{80}, F_{112}, F_{160}) \quad (2)$$

where  $F_{fuse}$  denotes the fused multi-scale features.

(3) Feature classification component. The fused features are passed through the classification head to obtain the limestone class prediction:

$$\hat{y} = \operatorname{argmax} \left( \operatorname{Softmax} \left( \operatorname{Head} \left( F_{fuse} \right) \right) \right) \quad (3)$$

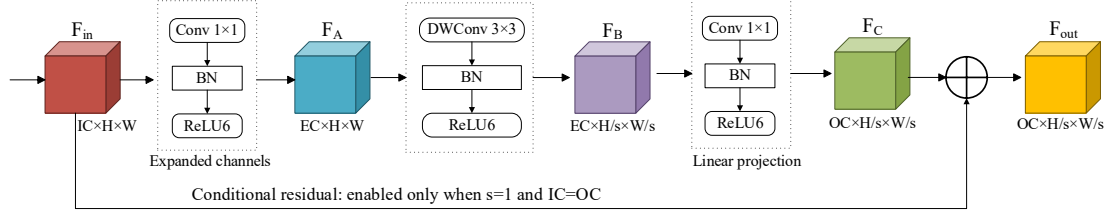
where Head denotes the final classification head, Softmax denotes the probability normalization function, and  $\hat{y}$  represents the final predicted class.

## 2.2. Simplified Inverted Residual Block (SIR)

The standard MobileNetV3-Large stacks 15 inverted residual blocks with 8 SE attention modules, yielding strong general feature extraction but excessive parameters and computation. For fine-grained limestone classification, this heavy backbone constrains the design space for subsequent RTEM and MFCM modules. Moreover, the layer-by-layer progressive representation lacks explicit cross-scale fusion, limiting synergistic use of grain-scale texture and fabric-scale features.

We therefore propose the Simplified Inverted Residual (SIR) block for basic feature extraction, channel transformation, and spatial downsampling, delegating texture enhancement and multi-level fusion to RTEM and MFCM respectively.

As shown in Figure 2, the SIR block comprises expansion convolution, depthwise convolution, and linear projection. It removes the internal SE attention and  $1 \times 1$  depthwise convolutions, uniformly adopting  $3 \times 3$  depthwise convolutions with ReLU6 to reduce backbone complexity.



Input channel count = IC; the first  $1 \times 1$  convolution expands the channels from IC to EC, and the last  $1 \times 1$  convolution performs linear projection to OC.

Figure 2: Architecture of the SIR block

For an input feature  $F_{in}$ , the SIR block first expands channels via  $1 \times 1$  convolution, then applies stride- $s$  depthwise convolution for spatial feature extraction, followed by a linear  $1 \times 1$  projection. A residual connection is added when stride=1 and input/output channels match:

$$F_A = \text{ReLU6}[\text{BN}(\text{Conv}1 \times 1(F_{in}))] \quad (4)$$

$$F_B = \text{ReLU6}[\text{BN}(\text{DWConv}3 \times 3(F_A))] \quad (5)$$

$$F_C = \text{BN}(\text{Conv}1 \times 1(F_B)) \quad (6)$$

$$F_{out} = \begin{cases} F_C + F_{in}, & s=1 \text{ and } C_{in} = C_{out} \\ F_C, & \text{otherwise} \end{cases} \quad (7)$$

where  $F_{out}$  denotes the SIR output;  $C_{in}$ ,  $C_{exp}$ , and  $C_{out}$  denote input, expansion, and output channels; and  $\text{DWConv}$  denotes depthwise convolution.

Six SIR blocks are stacked, with only the first using a residual connection. Channel dimensions progress as  $16 \rightarrow 24 \rightarrow 40 \rightarrow 80 \rightarrow 112 \rightarrow 160$ .

Table1. Complexity comparison between the SIR stack and the standard MobileNetV3-Large inverted

residual stack

| Item                                 | Proposed SIR stack | Standard MobileNetV3-Large inverted residual stack |
|--------------------------------------|--------------------|----------------------------------------------------|
| Number of modules                    | 6                  | 15                                                 |
| Param/M                              | 0.332              | 2.816                                              |
| FLOPs/G                              | 0.510              | 1.176                                              |
| Number of SE modules                 | 0                  | 8                                                  |
| Number of residual-connected modules | 1                  | 10                                                 |

Note: Param denotes the parameter count, FLOPs denotes the computational cost, and 1 MAC = 2 FLOPs.

Table 1 shows the SIR stack uses 11.80% of the parameters and 43.38% of the FLOPs of the standard MobileNetV3-Large inverted residual stack, markedly compressing the backbone and reserving capacity for RTEM and MFCM.

### 2.3. Rock-Texture Enhancement Module (RTEM)

Inter-class differences in limestone are typically manifested as variations in grain boundaries, texture orientation, matrix distribution, and local structural configurations. Although the inverted residual block can accomplish feature extraction and channel transformation through pointwise and depthwise convolutions, its primary objective is to construct lightweight feature representations; it lacks targeted enhancement of rock edge and texture information, which tends to result in inadequate representation of fine-grained discriminative features. To address this issue, this study introduces the Rock-Texture Enhancement Module (RTEM) after the feature stages with 24, 40, 80, 112, and 160 channels, respectively.

The architecture of the RTEM is illustrated in Figure 3.

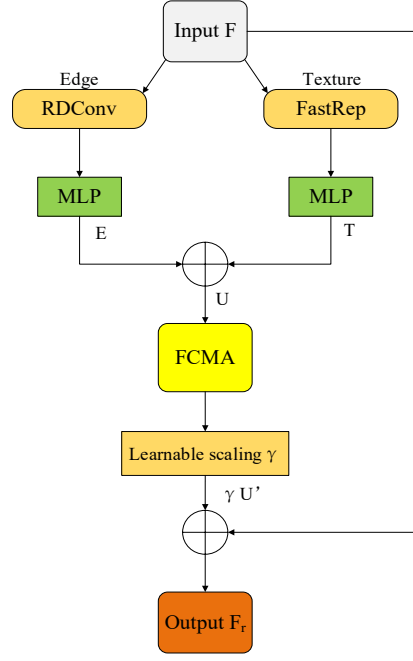


Figure 3: Architecture of the RTEM

The RTEM consists of an edge branch, a texture branch, a Full-Channel Mapping Attention (FCMA) module, and a residual scaling structure. Its design collectively draws upon the principles of lightweight convolution, residual learning, and feature response normalization. The residual connection proposed by He et al.(2016) preserves original information during feature transformation; group normalization introduced by Wu and He(2018) is independent of batch statistics and can enhance the stability of feature distributions; and global response normalization proposed by Woo et al.(2023) in ConvNeXt V2 strengthens feature representation through inter-channel response competition. Building upon these studies and targeting the coexistence of grain boundaries and textural structures in limestone images, an edge branch and a texture branch are constructed separately.

For an input feature  $F$ , the edge branch employs Rock Depthwise Convolution (RDConv) for local spatial feature extraction. RDConv consists of a  $3 \times 3$  depthwise convolution, group normalization, and the ReLU6 activation function. The computation of the edge branch is expressed as:

$$\begin{aligned} E &= MLP[RDConv(F)] \\ &= MLP[ReLU6(GN_G(DWConv_{3 \times 3}(F)))] \end{aligned} \quad (8)$$

where  $DWConv_{3 \times 3}$  denotes the depthwise convolution,  $GN_G$  denotes group normalization with  $G$  groups,  $E_0$  denotes the RDConv output, and  $E$  denotes the edge branch output. The depthwise convolution extracts spatial information independently within each channel; serving as the local feature

operator of the edge branch, it primarily enhances the responses of grain contours, mineral boundaries, and grayscale abrupt-change regions. Group normalization is adopted to reduce the dependence of feature statistics on batch size and to improve the stability of feature distributions across different channel groups.

A lightweight MLP is placed after RDConv. The MLP first compresses the channel dimension  $C$  via a  $1 \times 1$  convolution, and after passing through batch normalization, ReLU6, and global response normalization, restores the channel dimension to  $C$  via another  $1 \times 1$  convolution. Global response normalization serves to establish a competitive relationship among channel responses, suppressing channels with strong redundant responses while enhancing discriminative edge features.

The texture branch adopts a Fast Residual Representation (FastRep) unit to extract texture orientation, grain arrangement, and local structural information. Its computation is expressed as:

$$T_d = \text{ReLU6}[\text{BN}(\text{DWConv}3 \times 3(F))] \quad (9)$$

$$\begin{aligned} T &= \text{MLP}[\text{FastRep}(F)] \\ &= \text{MLP}[\text{ReLU6}(\text{BN}(\text{PWConv}1 \times 1(T_d + F)))] \end{aligned} \quad (10)$$

where BN denotes batch normalization,  $\text{PWConv}1 \times 1$  denotes pointwise convolution,  $T_d$  denotes the local textural features extracted by depthwise convolution,  $T_0$  denotes the FastRep output, and  $T$  denotes the texture branch output. FastRep first applies a depthwise convolution to extract local textural patterns within each channel, then performs element-wise addition with the input features to preserve the original textural information; a  $1 \times 1$  convolution is subsequently employed for channel information fusion. This structure combines the low computational cost of depthwise convolution with the information-preserving capability of residual connections, without involving structural re-parameterization.

The edge branch emphasizes grain contours and grayscale abrupt changes, whereas the texture branch emphasizes textural arrangement and local structures; the two branches are preliminarily fused through element-wise addition:

$$U = E + T \quad (11)$$

where  $U$  denotes the intermediate feature obtained by fusing the edge and textural features. Since different channels carry different textural information and vary in their discriminative contribution

to classification, FCMA is subsequently applied to perform channel-wise recalibration on the fused features, enhancing effective textural responses and suppressing redundant information.

#### 2.4. Full-Channel Mapping Attention (FCMA)

Wang et al.(2020) observed that channel dimensionality reduction in SE(Hu et al.2020) weakens the correspondence between input channels and their attention weights, and that a single-layer  $C \rightarrow C$  mapping (SE-Var3) outperforms the reduction-expansion bottleneck of original SE. Motivated by this finding and the channel recalibration principle of SE, we design a Full-Channel Mapping Attention (FCMA) module within RTEM (Figure 4).

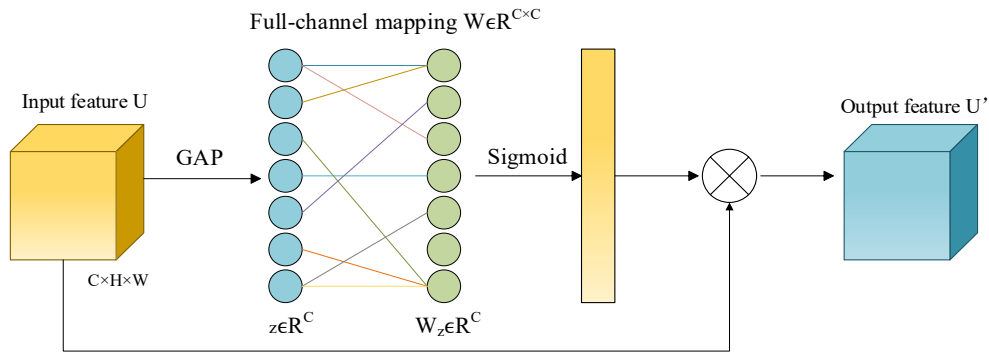


Figure 4: Full-Channel Mapping Attention module

Given the fused feature  $U$  from the edge and texture branches, FCMA first applies global average pooling to obtain a channel descriptor vector  $Z \in R^C$ . It then employs a single-layer  $1 \times 1$  convolution for direct  $C \rightarrow C$  channel mapping, bypassing the reduction-expansion bottleneck of SE:

$$a = \text{Sigmoid}(\text{Conv}1 \times 1(z)) = \text{Sigmoid}(W_z) \quad (12)$$

$$a_c = \text{Sigmoid}\left(\sum_{j=1}^C \omega_{cj} z_j\right) \quad (13)$$

where  $W \in R^{C \times C}$  is the channel mapping matrix and  $a \in R^C$  denotes the generated attention weights. Since the pooled feature has spatial size  $1 \times 1$ , this convolution operates as a full linear mapping over channels. The  $c$ -th attention weight is further expressed as:

where  $\omega_{cj}$  denotes the influence of the  $j$ -th input channel on the  $c$ -th output channel. In FCMA, each channel's attention weight is determined by all input channels, enabling global dependency modeling on the edge-texture joint features.

The attention weights are then applied channel-wise to the fused feature:

$$U' = U \otimes a \quad (14)$$

where  $\otimes$  denotes element-wise multiplication and  $U'$  is the recalibrated feature. FCMA thus enhances discriminative edge and texture responses while suppressing redundant channels.

Both FCMA and ECA avoid channel dimensionality reduction but adopt distinct cross-channel interaction mechanisms. FCMA uses a dense all-to-all mapping matrix, whereas ECA employs a parameter-shared 1D convolution for local neighboring-channel interaction. Their parameter counts are respectively:

$$P_{FCMA} = C^2 \quad (15)$$

$$P_{ECA} = k \quad (16)$$

Thus, FCMA models global channel dependencies at the cost of quadratically growing parameters, whereas ECA captures only local neighboring-channel interactions with higher parameter efficiency. In this work, FCMA is applied to edge-texture joint features for complete channel dependency modeling, while ECA is used for lightweight channel selection in high-dimensional fused features.

The five RTEM stages have channel dimensions of 24, 40, 80, 112, and 160, respectively; the total FCMA overhead is approximately 0.047M parameters (2.23% of the 2.095M model). Given these moderate channel dimensions, FCMA incurs acceptable overhead, and its parameter count is independent of the feature map spatial size.

Finally, RTEM injects the enhanced feature via a learnable scaling factor  $\gamma$  with a residual connection:

$$F_r = F + \gamma U' \quad (17)$$

where  $F_\gamma$  is the RTEM output and  $\gamma$  is a learnable scaling factor initialized to 0.1. The small initial  $\gamma$  preserves the original feature distribution during early training, while allowing adaptive adjustment of the enhancement contribution as training progresses.

## 2.5. Multi-source Feature Complementation Module (MFCM)

The multi-level feature extraction network produces features at four stages: F40, F80, F112, and F160. As depth increases, spatial resolution decreases while channel dimension and semantic abstraction increase: F40 retains fine grain boundaries and textural details; F80 and F112 capture grain arrangement and local structural relations; F160 encodes abstract class semantics with a larger receptive field.

Using only F160 discards fine-grained shallow details, while direct concatenation across stages suffers from spatial and channel misalignment and introduces redundancy. We therefore propose the Multi-source Feature Complementation Module(MFCM), which performs size alignment, channel projection, concatenation, channel selection, and gated fusion (Figure 5).

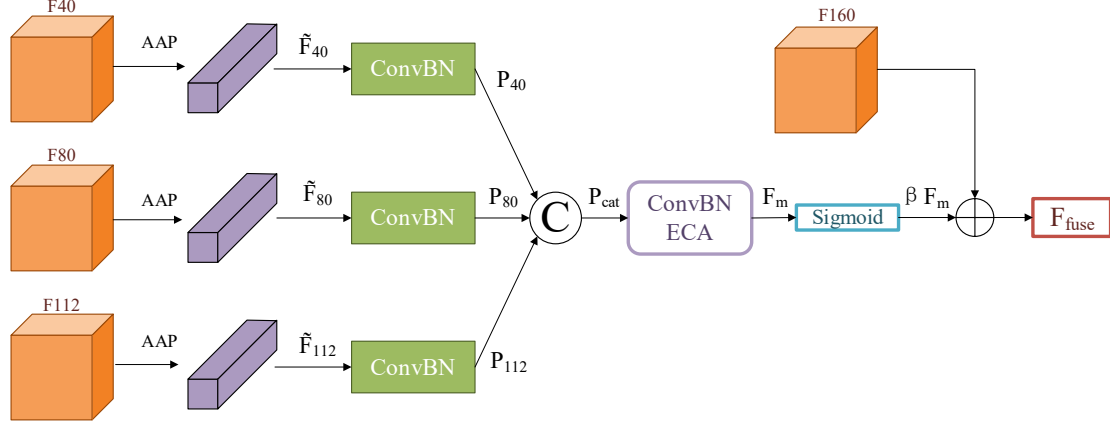


Figure 5: Multi-source Feature Complementation Module

MFCM takes F160 as the semantic backbone and F40, F80, F112 as supplementary inputs. For a  $384 \times 384$  input, the four stages yield feature maps of approximately  $40 \times 48 \times 48$ ,  $80 \times 24 \times 24$ ,  $112 \times 24 \times 24$ , and  $160 \times 12 \times 12$  ( $C \times H \times W$ ). The preprocessing pipeline first aligns spatial dimensions via adaptive average pooling, then projects channels, concatenates, and compresses:

$$\tilde{F}_i = AAP(F_i, H_{160}, W_{160}), \quad i \in \{40, 80, 112\} \quad (18)$$

$$P_i = ReLU6[BN(Conv1 \times 1(\tilde{F}_i))], \quad i \in \{40, 80, 112\} \quad (19)$$

$$P_{cat} = Concat(P_{40}, P_{80}, P_{112}) \quad (20)$$

$$V = HSwish[BN(Conv1 \times 1(P_{cat}))] \quad (21)$$

where AAP denotes adaptive average pooling to the spatial size of F160 ( $12 \times 12$ ),  $Conv1 \times 1$  projects each stage to 160 channels with BN and ReLU6, Concat aggregates the three projected features into a 480-channel tensor, and the final  $Conv1 \times 1$  with Hardswish compresses channels back to 160 while fusing multi-stage information.

where P40, P80, and P112 are all projected features of size  $160 \times 12 \times 12$ . The  $1 \times 1$  convolutions perform channel transformation without altering the spatial dimensions, endowing features from different stages with a consistent channel dimension; batch normalization and ReLU6 are employed to stabilize feature distributions and introduce nonlinearity.

278 The compressed feature then undergoes ECA-based channel selection. For an input feature  $V$ , a  
 279 global average pooling step produces a channel descriptor, followed by 1D convolution and channel-  
 280 wise reweighting:

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W V_c(i, j) \quad (22)$$

$$a = \text{Sigmoid}(\text{Conv1D}_3(z)) \quad (23)$$

$$F_m = V \otimes a \quad (24)$$

281 where  $z = \text{GAP}(V)$  is the channel descriptor,  $a = \text{sigmoid}(\text{Conv1D}_k(z))$  are the attention weights  
 282 determined by each channel and its neighbors, and  $V' = V \otimes a$  is the reweighted feature. The parameter-  
 283 shared 1D convolution introduces only  $k$  parameters. Unlike FCMA in RTE, which models global  
 284 channel dependencies, ECA here performs local cross-channel interaction to suppress redundancy in the  
 285 already-concatenated multi-level features at low cost.

286 Finally, the selected multi-level feature is injected into the semantic backbone via a learnable gating  
 287 coefficient:

$$\beta = \text{Sigmoid}(\alpha) \quad (25)$$

$$F_{fuse} = F_{160} + \beta F_m \quad (26)$$

288 where  $\alpha$  is a learnable gating parameter. A small initial  $\alpha$  lets the model rely on stable high-level  
 289 semantics during early training, while gradually increasing the contribution of multi-level details as  
 290 training progresses.

291 MFCM thus preserves the class semantics of  $F_{160}$  while incorporating grain boundary, textural,  
 292 and structural details from shallower stages, yielding a fused feature that combines fine-grained  
 293 representation with discriminative power.

## 294 2.6. Main Classification Head

295 After MFCM-based feature fusion, the model obtains a 160-channel fused feature map  $F_{fuse}$ . The  
 296 lithology classification component comprises a main classification head and three auxiliary  
 297 classification heads. The main head is responsible for outputting the final classification results during  
 298 validation and inference; the auxiliary heads are connected to the intermediate features  $F_{40}$ ,  $F_{80}$ , and

299  $F_{112}$ , respectively, and output intermediate predictions only during training.

300 The main classification head consists of a  $1 \times 1$  ConvBN module (convolution + batch  
301 normalization + Hardswish), ECA channel attention, global average pooling, and two fully connected  
302 layers with Hardswish activation after the first FC layer and Dropout regularization. The entire feature  
303 transformation pipeline is merged into one composite function:

$$y = W_2 \text{Dropout}(\text{HSwish}(W_1(\text{GAP}(\text{ECA}(\text{ConvBN}_{1 \times 1}(F_{fuse})))) + b_1)) + b_2 \quad (27)$$

304 where ConvBN  $1 \times 1$  expands the channel dimension from 160 to 960 with Hardswish activation;  
305 ECA with kernel size 3 recalibrates channels via sigmoid-weighted multiplication using only 3  
306 convolution parameters, avoiding the heavy cost of full channel-to-channel mapping; GAP compresses  
307 spatial dimensions into a 960-dimensional global feature vector;  $W_1 \in R^{\{1280 \times 960\}}$  and  $b_1 \in R^{\{1280\}}$   
308 are the first fully connected layer parameters, followed by Hardswish activation and Dropout with a  
309 rate of 0.3;  $W_2 \in R^{\{6 \times 1280\}}$  and  $b_2 \in R^6$  map the high-dimensional features to six-dimensional logits  
310  $z$  corresponding to the six limestone classes.

311 The final predicted class is obtained by applying Softmax and argmax:

$$\hat{y} = \arg \max \frac{\exp(y_i)}{\sum_{j=1}^6 \exp(y_j)} \quad (28)$$

312 where the fraction inside the brackets is the Softmax probability  $p_i$  of the  $i$ -th limestone  
313 class, and  $\hat{y}$  denotes the class with the highest probability, that is, the final classification result of the  
314 model.

## 315 2.7. Auxiliary Classification Heads

316 To enhance the discriminative capability of intermediate-layer representations, this study draws on  
317 the deep supervision idea (Lee et al., 2015) and derives three auxiliary classification branches from  
318  $F_{40}$ ,  $F_{80}$ , and  $F_{112}$ , respectively, to obtain independent classification outputs from features at different  
319 depths. The auxiliary branches are connected in parallel with the backbone network and do not alter the  
320 original feature propagation path: while the three intermediate features are fed into the auxiliary heads,

they continue to participate in subsequent feature extraction and MFCM fusion.

The three auxiliary classification heads are denoted as  $AuxHead_{40}$ ,  $AuxHead_{80}$ , and  $AuxHead_{112}$ , and their outputs are expressed as:

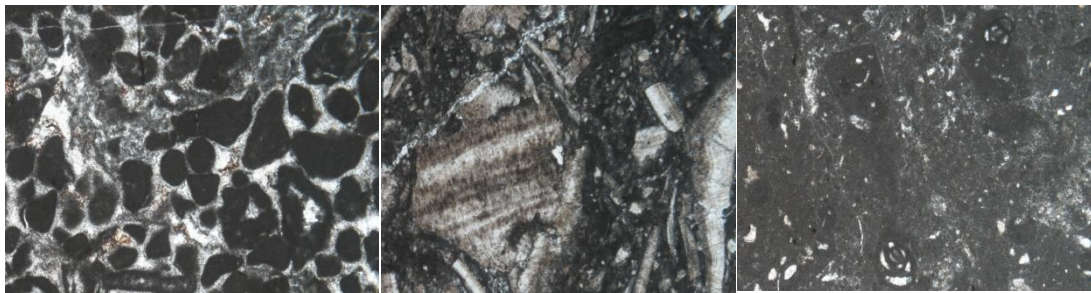
$$y_s^{aux} = AuxHead_s(F_s), \quad s \in \{40, 80, 112\} \quad (29)$$

where  $y_s^{aux} \in R^6$  is the six-dimensional logits vector output by the auxiliary head at stages.

The three auxiliary heads share the same basic structure, each consisting of a  $1 \times 1$  convolution, batch normalization, HSwish, ECA channel attention, global average pooling, a flatten operation, Dropout(Srivastava et al., 2014), and a fully connected layer in sequence; however, they differ in channel projection dimensions and Dropout rates.

## 2.8. Data Source and Class Composition

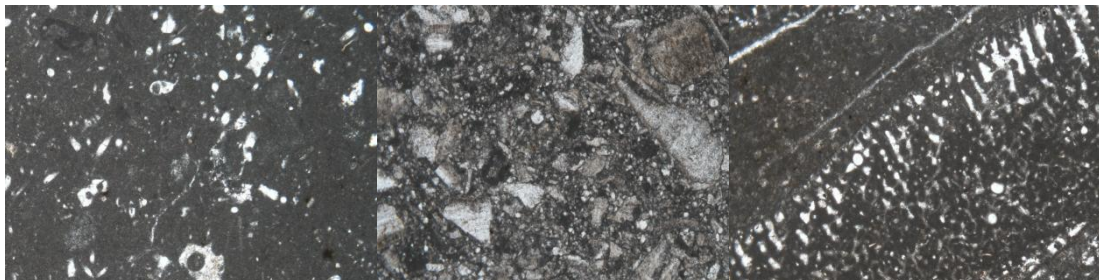
The limestone microscopic image samples used in this study were collected from January to April 2024 in regions including Ordos in Inner Mongolia, Xinjiang, Tibet, and the Lesser Khingan Mountains in Heilongjiang. After sampling, representative limestone specimens were selected, and the original images were acquired through cutting, polishing, and microscopic imaging. Based on rock fabric, grain content, and microscopic textural features, the dataset is divided into six classes: grainstone, intraclast limestone, packstone, mudstone, wackestone, and floatstone (Figure 6), with the original data comprising a total of 1,286 microscopic images.



(a) Grainstone

(b) Intraclast limestone

(c) Wackestone



340 (d)Mudstone (e)Packstone (f)Floatstone

341 Figure 6: Images of different types of rock thin sections

342 The original sample counts for each class are listed in Table 2. Among the original data, mudstone  
 343 has the largest number of samples and intraclast limestone the smallest, with a ratio of approximately 8  
 344 .60 between the largest and smallest classes, indicating a pronounced class imbalance. To alleviate the  
 345 impact of class size disparity on classifier training and evaluation, this study performs class balancing b  
 346 ased on region cropping of the original images, ultimately constructing an experimental dataset of 3,60  
 347 0 images, with 600 images per class.

348 Table2. Class distribution of the original limestone microscopic images

| Class               | Grainstone | Intraclast<br>limestone | Packstone | Mudstone | Wackestone | Floatstone | Total |
|---------------------|------------|-------------------------|-----------|----------|------------|------------|-------|
| Number of<br>images | 204        | 53                      | 338       | 456      | 157        | 78         | 1286  |

349

## 350 2.9. Dataset Splitting

351 The experimental data are divided into a training set and a validation set at an 8:2 ratio, comprisin  
 352 g 2,880 and 720 images, respectively. All six classes maintain balanced counts in both sets, with 480 tr  
 353 aining images and 120 validation images per class. The training set is used for model parameter learnin  
 354 g, and the validation set for model selection and performance evaluation. To ensure comparability acro  
 355 ss different models, the same data split is adopted in all subsequent comparative and ablation experime  
 356 nts.

## 357 2.10. Image Preprocessing and Data Augmentation

358 Training samples undergo a pipeline of single-image transforms and batch-level mixed augmentati  
 359 on. For spatial augmentation, the short side is resized to 400 pixels followed by a random 384×384 cro  
 360 p; horizontal/vertical flips ( $p=0.5$ ) and random rotation within  $[-15^\circ, 15^\circ]$  are applied. For color transfor  
 361 ms, random perturbations with a magnitude of 0.1 are applied to the brightness, contrast, and saturation  
 362 of the image (Krizhevsky et al., 2014). Images are then normalized using ImageNet channel statistics an  
 363 d subjected to random erasing ( $p=0.15$ , area ratio 0.02–0.12) (Zhong et al., 2020) to suppress over-relian

ce on localized grain-boundary cues.

Batch-level regularization comprises label smoothing(Szegedy et al.,2016), MixUp(Zhang et al.,2018), and CutMix(Yun et al.,2019). Label smoothing reformulates the one-hot target as:

$$t_{i,c}^{LS} = (1 - \varepsilon)I(c = y_i) + \frac{\varepsilon}{C} \quad c=1,2,...,C \quad (30)$$

where  $\varepsilon$  is the smoothing coefficient,  $C$  the number of classes, and  $y_i$  the ground-truth index of sample  $i$ . Mixed augmentation is activated with probability 0.5 per batch, selecting MixUp or CutMix equiprobably. MixUp performs pixel-wise linear interpolation of images and labels with  $\lambda \sim \text{Beta}(0.2, 0.2)$ ; CutMix replaces a rectangular region sampled via  $\lambda \sim \text{Beta}(1.0, 1.0)$  and recomputes the label ratio from the actual retained area. At validation, only resize-to-400, center-crop, and normalization are applied deterministically.

#### 2.11. Training Parameter Settings

The model is trained for 240 epochs with  $384 \times 384$  inputs. A gradient accumulation strategy(Ott et al.,2018) is employed during training,a mini-batch size of 16 with gradient accumulation over 2 steps yields an effective batch size of 32. Optimization employs AdamW (initial learning rate  $1 \times 10^{-3}$ , weight decay  $5 \times 10^{-3}$ ) with a linear warmup of 8 epochs followed by cosine annealing. Gradient clipping by L2 norm is applied before each update. Three independent runs with seeds 42, 2024, and 3407 are conducted, with automatic mixed precision enabled under CUDA.

#### 2.12. Joint Loss and Auxiliary Supervision

Both the main and auxiliary heads are supervised by identical soft labels. The soft-label cross-entropy loss is defined as:

$$L_{CE}^{\text{soft}}(\mathbf{q}, \mathbf{t}) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C t_{i,c} \log[\text{Softmax}(\mathbf{q}_i)_c] \quad (31)$$

where  $\mathbf{q}$  and  $\mathbf{t}$  denote the prediction score matrix and soft-label matrix, respectively. The three auxiliary branches attached to stages F40, F80, and F112 produce an averaged auxiliary loss:

$$L_{\text{aux}} = \frac{L_{\text{aux}}^{(40)} + L_{\text{aux}}^{(80)} + L_{\text{aux}}^{(112)}}{3} \quad (32)$$

The auxiliary weight decays linearly so that supervision concentrates in early epochs, when shallow features lack discriminative structure; as training progresses, the main head and MFCM form stable high-level representations and the auxiliary influence is phased out. The total training objective is:

$$L_{\text{total}} = L_{\text{CE}}^{\text{soft}}(\mathbf{y}, \mathbf{t}) + \lambda_e L_{\text{aux}} \quad (33)$$

$$\lambda_e = 0.3 \cdot \max\left(0, 1 - \frac{e}{0.7E}\right) \quad (34)$$

where  $E$  is the total number of epochs and  $e$  the current epoch. After epoch  $0.7E$ , auxiliary supervision ceases entirely. Auxiliary heads are active only during training; validation and inference rely solely on the main head.

### 2.13. Test-Time Augmentation and Model Ensemble

Model selection for each seed uses three-view TTA (original, horizontal flip, vertical flip). Final evaluation employs multi-scale TTA at short sides 400/416/432, with center and four-corner  $384 \times 384$  crops and four views per crop (original, H-flip, V-flip, both). Predictions from the three seed models are fused by weighted averaging:

$$\mathbf{z} = \sum_{m=1}^M w_m \left( \frac{1}{V} \sum_{v=1}^V \mathbf{z}_{m,v} \right), \quad \sum_{m=1}^M w_m = 1 \quad (35)$$

where  $w_m$  denotes the ensemble weight of model  $m$  and  $V$  the number of test views. The final class is assigned by argmax over the fused score vector.

### 2.14. Evaluation Metrics

Classification performance is assessed by Accuracy, Precision, Recall, and F1-score; model complexity by parameter count (M) and FLOPs (G). Per-class metrics are macro-averaged over the six limestone classes.

Let the evaluation dataset contain  $N$  samples, among which  $N_{\text{correct}}$  are correctly classified; the accuracy is defined as:

$$ACC = \frac{N_{\text{correct}}}{N} \quad (36)$$

Precision, recall, and F1-score are respectively defined as:

$$P = \frac{TP}{TP + FP} \quad (37)$$

$$R = \frac{TP}{TP + FN} \quad (38)$$

$$F_1 = \frac{2PR}{P + R} \quad (39)$$

where TP/FP/FN/TN denote true positive, false positive, false negative, and true negative in a one-versus-rest confusion matrix. FLOPs are profiled via THOP on a single-view 384×384 input along the main classification path, excluding auxiliary heads and TTA overhead.

### 3. Results

#### 3.1. Training Process Analysis

Figure 7 shows training loss and validation accuracy across three seeds. Loss decreases monotonically while accuracy stabilizes near convergence, indicating effective learning of discriminative textural and structural features. Inter-seed fluctuations arise from random initialization, sample ordering, and augmentation stochasticity.

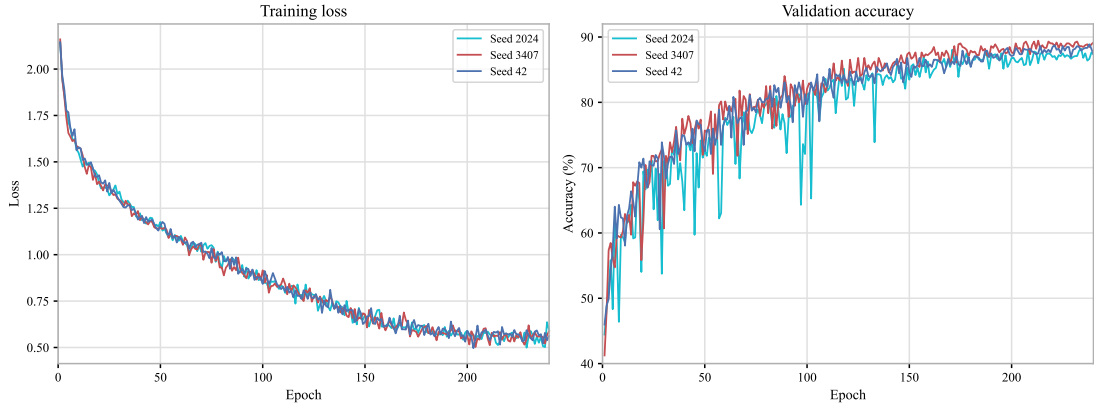


Figure 7: Training loss and validation accuracy curves for the three random seeds

Table 3 summarizes the best single-model results. Accuracy ranges from 88.19% to 89.44%, with seed 3407 achieving the highest 89.44%. Given the observed seed-dependent variability, a multi-model ensemble is employed to exploit predictive complementarity and mitigate single-run fluctuations.

Table3. Best validation results for different random seeds

| Random seed | Best epoch | ACC/% |
|-------------|------------|-------|
| 42          | 238        | 88.89 |
| 2024        | 234        | 88.19 |
| 3407        | 213        | 89.44 |

### 3.2. Overall Performance and Class-level Results

On the 720-image validation set, the equal-weight three-model ensemble yields 90.28% accuracy; logits-weighted ensemble (0.30, 0.05, 0.65) improves to 91.11% (656/720 correct). The final model achieves macro-averaged precision, recall, and F1-score of 91.28%, 91.11%, and 91.12%, respectively (Table 4).

Table4. Class-level recognition results of the proposed method

| Class                | Precision/% | Recall/% | F1/%  |
|----------------------|-------------|----------|-------|
| Mudstone             | 91.15       | 85.83    | 88.41 |
| Packstone            | 89.19       | 82.50    | 85.71 |
| Floatstone           | 97.50       | 97.50    | 97.50 |
| Intraclast limestone | 94.49       | 100.00   | 97.17 |
| Wackestone           | 78.79       | 86.67    | 82.54 |
| Grainstone           | 96.58       | 94.17    | 95.36 |
| Macro-average        | 91.28       | 91.11    | 91.12 |

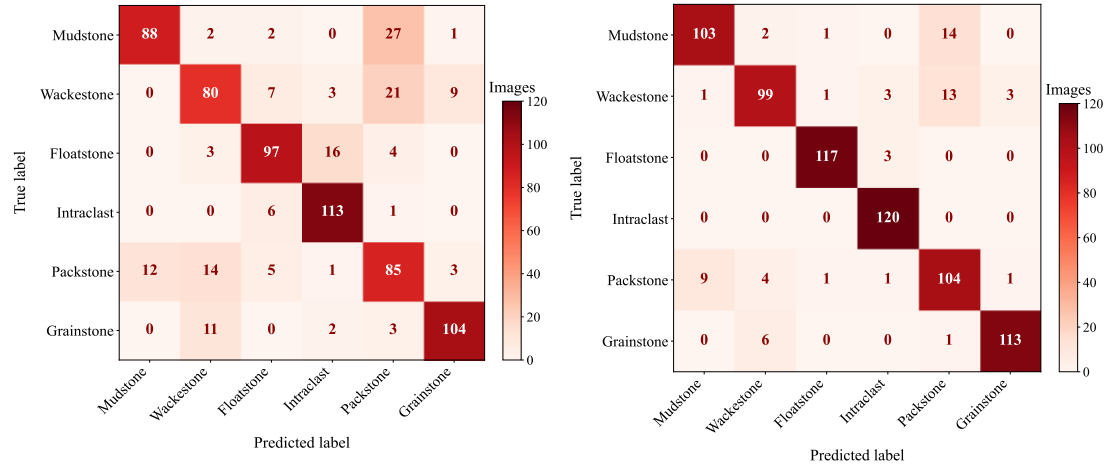
Intraclast limestone reaches 100% recall; floatstone achieves 97.50% across all metrics; grainstone attains 95.36% F1-score, reflecting strong discrimination for classes with prominent grain contours. Wackestone has the lowest precision (78.79%), constituting the primary bottleneck for macro-averaged performance.

### 3.3. Confusion Matrix Analysis

Figure 8 compares confusion matrices of the baseline and proposed model. Correct classifications increase from 567 to 656, reducing misclassifications by 58.17% (153→64) and improving accuracy from 78.75% to 91.11% (+12.36 pp). All six classes show diagonal gains, indicating consistent improvement rather than class-specific gains.

Correct identifications increase notably for packstone (80→99), floatstone (97→117), wackestone (85→104), and grainstone (104→113); intraclast limestone reaches full correctness (113→120). Bidirectional confusion between floatstone and intraclast limestone decreases from 16/6 to 3/0 misclassifications, indicating enhanced contour-boundary discrimination.

Residual errors concentrate among mudstone, wackestone, and packstone. Wackestone-related misclassifications decrease from 74 to 40, yet wackestone remains the primary confusion target for mudstone and packstone. This reflects gradual transitions in micritic matrix content and grain-support relationships among these fine-grained classes, highlighting a direction for future optimization.



(a) Confusion matrix of the baseline model

(b) Confusion matrix of the proposed model

Figure 8: Confusion matrix of the proposed method on the validation set

### 3.4. Comparative Experiments

Ten models are selected for comparison, including ResNet-101(He et al.,2016), FasterNet-L (Chen et al.,2023), MaxViT-L(Tu et al.,2022), EfficientNet-B7(Tan et al.,2019), Swin V2-Base (Liu et al.,2022), EfficientNetV2-XL(Tan et al.,2021), ConvNeXt-Base(Liu et al.,2022), ConvMixer-1536/20(Trockman et al.,2023), and MobileNetV4-Conv-Large(Qin et al.,2024). All are evaluated on the same six-class validation set (Table 5).

Table5. Comparison of classification performance and model complexity across different models

| Model      | ACC/% | P/%   | R/%   | F1/%  | Param/M | FLOPs/G |
|------------|-------|-------|-------|-------|---------|---------|
| ResNet-101 | 79.58 | 79.77 | 79.58 | 79.34 | 42.512  | 15.728  |

| Model                  | ACC/% | P/%   | R/%   | F1/%  | Param/M | FLOPs/G |
|------------------------|-------|-------|-------|-------|---------|---------|
| FasterNet-L            | 80.00 | 80.35 | 80.00 | 79.85 | 92.192  | 31.090  |
| ConvNeXt-Base          | 79.72 | 80.11 | 79.72 | 79.77 | 87.573  | 30.708  |
| EfficientNet-B7        | 84.31 | 84.52 | 84.31 | 84.25 | 63.802  | 10.656  |
| SwinTransformerV2-Base | 80.42 | 80.33 | 80.42 | 80.33 | 86.900  | 30.476  |
| EfficientNetV2-XL      | 81.39 | 81.26 | 81.39 | 81.18 | 206.846 | 36.210  |
| MaxViT-L               | 90.14 | 90.17 | 90.14 | 90.12 | 164.862 | 68.318  |
| ConvMixer-1536/20      | 83.06 | 82.88 | 83.06 | 82.94 | 50.098  | 102.714 |
| MobileNetV4-Conv-Large | 81.67 | 81.57 | 81.67 | 81.56 | 31.318  | 4.414   |
| Proposed method        | 91.11 | 91.28 | 91.11 | 91.12 | 2.095   | 0.868   |

The proposed method achieves 91.11% accuracy and 91.12% macro-F1, ranking first among all models. It outperforms the closest competitor MaxViT-L by 0.97 pp, and ConvMixer-1536/20 by 8.05 pp while reducing parameters and FLOPs by 95.82% and 99.15%, respectively. This validates that texture-specific RTEM and MFCM designs deliver strong performance at a compact model scale.

### 3.5. Ablation Experiments

Ablation experiments progressively add training augmentation, classification-head ECA, RTEM with auxiliary supervision, MFCM, test-time augmentation, and multi-seed ensemble to the MobileNetV3-Large baseline (Table 6).

Table6. Ablation experiment results

| No. | Experimental configuration | Param/M | FLOPs/G | ACC/% |
|-----|----------------------------|---------|---------|-------|
| 0   | Baseline                   | 4.210   | 1.358   | 78.75 |
| 1   | +DataAug+AdamW             | 4.210   | 1.358   | 82.36 |
| 2   | +HeadECA                   | 4.210   | 1.358   | 87.08 |
| 3   | +SIR+RTEM+AuxMix           | 1.980   | 0.834   | 88.75 |
| 4   | +MFCM                      | 2.095   | 0.868   | 89.03 |
| 5   | +AdvancedTTA               | 2.095   | 0.868   | 90.28 |
| 6   | Proposed method            | 2.095   | 0.868   | 91.11 |

The complete pipeline improves accuracy from 78.75% to 91.11% (+12.50 pp cumulative). Data augmentation and optimization strategies contribute +3.75 pp; classification-head ECA adds +4.72 pp.

Adding SIR, RTEM, auxiliary supervision, and MixUp/CutMix reaches 88.75%; MFCM further improves to 89.03%. Single-model 60-view TTA contributes +1.25 pp, and weighted three-seed ensemble adds +0.83 pp, yielding the final 91.11%.

#### 4. Conclusion

This paper proposes a lightweight limestone thin-section recognition method based on improved MobileNetV3. RTEM enhances edge and textural features; MFCM fuses shallow texture, mid-level structure, and deep semantics; ECA provides efficient channel selection. Training incorporates auxiliary supervision, multi-level augmentation, AdamW, and multi-seed weighted ensemble. On 3,600 images across six limestone classes, the model achieves 91.11% accuracy, 91.28%/91.11%/91.12% macro precision/recall/F1, with only 2.095M parameters. Ablation validates each module's contribution, with a cumulative 12.5 pp improvement over baseline.

Limitations include data scale, patch-level partitioning, and deployment overhead from multi-model multi-view inference. Future work will expand data across regions and imaging conditions, adopt original thin-section grouping with independent test sets, and apply knowledge distillation, pruning, and quantization to reduce computational cost for geological exploration and lithology analysis applications.

#### 5. Acknowledgments

This research was funded by National Natural Science Foundation of China (grant number:52474036), Heilongjiang Provincial Natural Science Foundation of China(grant number:LH2024E008),2024 Leading Scientific and Technological Innovation Talent Team: Shale Oil Geology and Engineering Integrated Fracturing and Prevention Technology Innovation Team"(grant number:CYCX24015).The authors also gratefully acknowledge the helpful comments and suggestions from the reviewers,which have significantly improved the paper.

## References

- Dong, F.J., Li, K.K., Sun, Z., et al., 2025. Classification of rock thin section images based on SENet-ConvNeXt-FL. *Carbonates and Evaporites* 40, 107. <https://doi.org/10.1007/s13146-025-01140-x>
- Fan, J.S., Yu, X.L., Di, Y.C., et al., 2026. A foundation model for rock thin-section images analysis. *Communications Engineering* 5, 9. <https://doi.org/10.1038/s44172-025-00565-5>
- Wei, P.Y., Fan, C.Y., Yang, X.W., et al., 2026. A hybrid local-global feature attention network for thin section rock image classification. *Scientific Reports* 16, 6446. <https://doi.org/10.1038/s41598-026-36669-x>
- Zhang, C., Lin, K., 2026. Data augmentation framework for rock thin section identification and classification under the condition of small samples. *Earth Science Informatics* 19, 144. <https://doi.org/10.1007/s12145-026-02194-8>
- He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Howard, A., Sandler, M., Chu, G., et al., 2019. Searching for MobileNetV3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Piscataway, NJ, pp. 1314–1324.
- Hu, J., Shen, L., Albanie, S., et al., 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2011–2023.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. Vol. 25.
- Lee, C.Y., Xie, S., Gallagher, P., et al., 2015. Deeply-supervised nets. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. pp. 562–570.
- Li, K.R., Song, J.M., Zhang, Z.Y., et al., 2026. A dataset and benchmark of carbonate thin-section images for deep learning. *Scientific Data* 13, 340.
- Liu, C., Ding, H.H., Jiang, X.D., 2021. Towards enhancing fine-grained details for image matting. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 385–393.
- Liu, Q.N., Yan, J., Liu, C., 2025. Research on rock thin section classification based on improved MobileNetV3. *Computer Knowledge and Technology* 21, 26–28.
- Liu, S.W., Ma, Z.W., Wei, S.Q., et al., 2026. A lithology identification method for rock images based on lightweight convolutional neural networks. *Bulletin of Geological Science and Technology* 45, 360–370. <https://doi.org/10.19509/j.cnki.dzkq.tb20240348>

527 Liu, Z., Hu, H., Lin, Y., et al., 2022. Swin Transformer V2: Scaling up capacity and resolution. In: Proceedings of the  
 528 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

529 Liu, Z., Mao, H., Wu, C.Y., et al., 2022. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on  
 530 Computer Vision and Pattern Recognition. pp. 11976–11986.

531 Ott, M., Edunov, S., Granger, D., et al., 2018. Scaling neural machine translation. In: Proceedings of the Third  
 532 Conference on Machine Translation: Research Papers. pp. 1–9.

533 Qin, D., Leichner, C., Delakis, M., et al., 2024. MobileNetV4: Universal models for the mobile ecosystem. In:  
 534 Proceedings of the European Conference on Computer Vision. pp. 78–96.

535 Qin, S., Wang, Q., Zeng, Q., Ye, M., Fu, A., Chen, G., 2025. Automatic recognition of debris rock lithology based on  
 536 unsupervised semantic segmentation. Computers & Geosciences 196, 105790.  
 537 <https://doi.org/10.1016/j.cageo.2024.105790>

538 Shi, Y., Guo, H.M., Deng, R., 2026. Carbonate lithology identification based on LC-PCA combined with reverse-  
 539 driven cuckoo search optimized neural network. Geophysical Prospecting for Petroleum 65, 533–547.

540 Song, S.Q., Wei, Y.S., Yan, H.J., et al., 2026. Lithology identification and sedimentary model of intra-platform  
 541 carbonate rocks based on GA-XGBoost: a case study of the Well Gaoshi 18 area in the Anyue Gas Field, Sichuan  
 542 Basin. Natural Gas Geoscience, 1–17. (advance online publication)

543 Srivastava, N., Hinton, G., Krizhevsky, A., et al., 2014. Dropout: a simple way to prevent neural networks from  
 544 overfitting. The Journal of Machine Learning Research 15, 1929–1958.

545 Szegedy, C., Vanhoucke, V., Ioffe, S., et al., 2016. Rethinking the Inception architecture for computer vision. In:  
 546 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

547 Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings  
 548 of the 36th International Conference on Machine Learning. pp. 6105–6114.

549 Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller models and faster training. In: Proceedings of the 38th International  
 550 Conference on Machine Learning. pp. 10096–10106.

551 Trockman, A., Kolter, J.Z., 2023. Patches are all you need? Transactions on Machine Learning Research.

552 Tu, Z., Talebi, H., Zhang, H., et al., 2022. MaxViT: Multi-axis vision transformer. In: Proceedings of the European  
 553 Conference on Computer Vision. pp. 459–479.

554 Wang, Q., Wu, B., Zhu, P., et al., 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks.  
555 In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534–11542.

556 Wang, T., Niu, Y., Zhao, W., et al., 2025. Research on intelligent classification of limestone photomicrographs based  
557 on the improved FasterNet architecture. *Earth Science Informatics* 18, 538.

558 Woo, S., Debnath, S., Hu, R., et al., 2023. ConvNeXt V2: Co-designing and scaling ConvNets with masked  
559 autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp.  
560 16133–16142.

561 Wu, Y., He, K., 2018. Group normalization. In: Proceedings of the European Conference on Computer Vision. pp. 3–  
562 19.

563 Xu, K.J., Liu, X., Yu, H.Z., et al., 2025. Igneous lithology identification in the Qinggelidi block based on gravity–  
564 magnetic data fusion. *Oil Geophysical Prospecting* 60, 225–233, 272.

565 Xu, Z.H., Ma, W., Lin, P., Shi, H., Pan, D., Liu, T., 2021. Deep learning of rock images for intelligent lithology  
566 identification. *Computers & Geosciences* 154, 104799. <https://doi.org/10.1016/j.cageo.2021.104799>

567 Yun, S., Han, D., Oh, S.J., et al., 2019. CutMix: Regularization strategy to train strong classifiers with localizable  
568 features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

569 Zhang, H., Cisse, M., Dauphin, Y.N., et al., 2018. mixup: Beyond empirical risk minimization. In: International  
570 Conference on Learning Representations.

571 Zhang, Y., Xiao, H., Du, M., et al., 2025. Application of neural architecture search in lithology identification. *Journal*  
572 *of Petroleum Exploration and Production Technology* 15, 128.

573 Zhao, F., Zhou, Z., Zhai, H., Zhang, P., Li, X., 2026. DEAE: A semi-supervised denoising enhanced adaptive encoder  
574 network for inter-well lithology identification in continental sandstone-igneous formations of Daqing Oilfield,  
575 China. *Computers & Geosciences* 212, 106168. <https://doi.org/10.1016/j.cageo.2026.106168>

576 Zhong, Z., Zheng, L., Kang, G., et al., 2020. Random erasing data augmentation. In: Proceedings of the AAAI  
577 Conference on Artificial Intelligence. Vol. 34, No. 7, pp. 13001–13008.

578 Zhou, C.Y., Liu, W., Wu, T.R., et al., 2024. Rock thin section image classification based on a mixture-of-experts  
579 model. *Journal of Jilin University (Science Edition)* 62, 905–914.

580 Zhu, Q.Q., Wang, S., Tong, S., et al., 2024. RockS2Net: Rock image classification via a spatial localization siamese  
581 network. *Computers & Geosciences* 185, 105560.

## 582 List of Figures

- 583 1. Figure 1: Overall framework of the improved model
- 584 2. Figure 2: Architecture of the SIR block
- 585 3. Figure 3: Architecture of the RTEM
- 586 4. Figure 4: Full-Channel Mapping Attention module
- 587 5. Figure 5: Multi-source Feature Complementation Module
- 588 6. Figure 6: Images of different types of rock thin sections
- 589 7. Figure 7: Training loss and validation accuracy curves for the three random seeds
- 590 8. Figure 8: Confusion matrix of the proposed method on the validation set

591

## 592 Declarations

593 Ethics approval and consent to participate

594 Not applicable. This study did not involve human participants, human data, or animals.

595 Consent for publication

596 Not applicable.

597 Availability of data and materials

598 Non-public dataset

599 Competing interests

600 The authors declare that they have no competing interests.

601 Funding

602 Tingting Wang, National Natural Science Foundation of China, 52474036

603 Tingting Wang, Natural Science Foundation of Heilongjiang Province, ZL2024E008

604 Wanchun Zhao, National Natural Science Foundation of China, 52174022

605 Wanchun Zhao, Natural Science Foundation of Heilongjiang Province, ZL2024E008

606 Authors' contributions

607 Tingting Wang: Supervision, Resources, Methodology, Writing – review & editing. Jianwei Lu: Writing – original

608 draft, Visualization, Validation, Methodology, Data curation. Wanchun Zhao: Supervision, Methodology,

609 Investigation. Ranjith P.G<sup>f</sup>: Supervision, Investigation, Data curation, Writing – review & editing.

610 Acknowledgements

611 Not applicable.