原创论文

多尺度通道增强Transformer在岩石薄片识别与序列一致性优化中的应用

Xiaoyao Guo[1,2]· Yan Chen[1,2]· Shipeng He[3]· Xingpeng Zhang[1,2]· Jing Zhou[1]· Xucheng Bao[1,2]

摘要

岩石薄片鉴定在地质勘探中具有关键作用，其能揭示岩石基本特性与组成的核心信息。然而矿物颗粒的精确识别面临三大挑战：数据分布固有失衡、特征相似性导致的误分类，以及不同正交偏光角度下显著的特征变异。这些复杂性使得传统单一结构的深度学习模型难以实现精准矿物颗粒识别。为此，本研究提出融合多尺度通道增强Transformer（MSCET）与序列一致性优化（SCO）策略的岩石薄片图像分类新方法。该集成方案能有效提取矿物颗粒的鉴别性特征，并充分挖掘偏光角度变化的影响。MSCET架构通过协同整合卷积神经网络（CNN）、压缩激励网络（SENet）和Transformer机制，增强网络特征表征能力：采用差异化卷积运算提取矿物颗粒的粗/细粒度特征，继而利用SENet与Transformer结构聚合通道与空间维度的全局信息。此外，引入SCO策略优化低置信度预测，从而缓解多角度正交偏光图像中特征变异的影响。综合实验表明该方法测试集分类准确率达92.35%，召回率、精确率和F1分数等关键指标均有显著提升，证实其在地质应用中实现稳健岩石薄片识别的潜力。

关键词 岩石薄片 - 多尺度融合 - SENet网络 - Transformer模型 - 序列一致性优化

数学学科分类(2010) 68T07 - 86A60 - 68T20

1 引言

岩石储层是油气的重要载体，其特性直接影响石油地质勘探的成效。岩石薄片鉴定作为关键环节，能够揭示

何世鹏、张兴鹏、周静与鲍旭成对本研究贡献均等。

ORIGINAL PAPER

Multi-scale channel enhanced transformer for rock thin sections identification and sequence consistency optimization

Xiaoyao Guo[1,2] · Yan Chen[1,2] · Shipeng He[3] · Xingpeng Zhang[1,2] · Jing Zhou[1] · Xucheng Bao[1,2]

Abstract

The identification of rock thin sections plays a pivotal role in geological exploration, as it provides critical insights into the fundamental properties and composition of rocks. However, the accurate identification of mineral particles presents significant challenges due to three primary factors: the inherent imbalance in data distribution, misclassification caused by feature similarity, and substantial feature variations observed under different cross-polarized angles. These complexities render conventional deep-learning models with single-structure architectures inadequate for precise mineral particle identification. Therefore, this study proposes a novel rock thin section image classification methodology that combines a Multi-Scale Channel Enhanced Transformer (MSCET) with a Sequence Consistency Optimization (SCO) strategy. This integrated approach is designed to effectively extract distinctive features of mineral particles while fully exploiting the influence of polarization angle variations. The MSCET architecture synergistically combines Convolutional Neural Networks (CNN), Squeeze-and-Excitation Networks (SENet), and Transformer mechanisms to enhance the network's feature representation capabilities. Specifically, it employs distinct convolutional operations to extract both coarse- and fine-grained features of mineral particles. The SENet and Transformer structures are then utilized to aggregate global information across both channel and spatial dimensions. Furthermore, we introduce the SCO strategy to refine low-confidence predictions, thereby mitigating the impact of feature variations in multi-angle cross-polarized images. Comprehensive experimental evaluations demonstrate the efficacy of our proposed method, achieving a classification accuracy of 92.35% on the test set. The method also shows significant improvements in key performance metrics, including recall, precision, and F1 score, substantiating its potential for robust rock thin section identification in geological applications.

Keywords Rock thin sections · Multi-scale fusion · SENet · Transformer · Sequence consistency optimization

Mathematics Subject Classification (2010) 68T07 · 86A60 · 68T20

1 Introduction

Rock reservoirs are crucial carriers of oil and gas, and their characteristics significantly affect the success of petroleum geological exploration. Rock thin section identification is an essential step in this process since itreveals

Shipeng He, Xingpeng Zhang, Jing Zhou, and Xucheng Bao contributed equally to this work.

矿物成分、颗粒特性及岩石类型是油气勘探与储层评价的关键信息。长期以来，传统鉴定方法依赖经验丰富的地质学家对岩石薄片图像中的矿物颗粒进行观察分析。然而该方法固有的主观性和冗长流程，在快速发展的勘探领域面临重大挑战[1]。这凸显了在该关键领域引入更客观、高效且可扩展方法的必要性。通过自动化实现岩石薄片图像中矿物颗粒的分类，可显著提升地质调查工作质量。

近年来，人工智能技术逐步解决了传统方法的局限性，推动了岩石薄片图像自动分类技术的发展。机器学习技术在该研究的早期阶段起到了关键作用。文献[2]提出了一种先进的分类器组合方法，通过对每个描述符进行k近邻分类来实现最终决策，从而从这些图像中获取高维视觉特征。另有研究采用数据挖掘技术评估了关联规则、决策树和支持向量机在火成岩岩性识别中的有效性[3]。尽管这些技术取得了有益成果，但由于依赖人工特征确定，它们仅适用于小规模数据集，且泛化能力较差。

为克服这些局限，研究者们日益转向卷积神经网络(CNNs)以实现岩石薄片图像的自动化精准识别[4]。例如文献[5]采用CNN通过自动提取图像特征，实现了高效精确的岩石分类。类似地，文献[6]和[7]分别运用Inception-v3与ResNet18网络，结合图像增强技术来提升不同类型岩石的分类效果。尽管当前深度学习模型在岩石薄片图像分类领域取得重大进展，但其仍受限于网络设计在复杂矿物颗粒特征提取方面的不足。需特别指出的是，矿物颗粒的岩石薄片图像在不同正交偏光角度下会呈现显著差异特征。如图1所示，五种常见矿物颗粒在五个不同角度会呈现颜色与对比度的明显变化。然而现有方法因未能考虑矿物颗粒特性在变化正交偏光角度下的改变，不仅降低了模型整体效能，还导致了分类误差。

为解决岩石薄片图像细粒度特征提取难题，本研究提出一种创新的图像分类框架。该方案采用多尺度通道增强Transformer（MSCET）模型进行初始图像分类，通过将多尺度融合与Transformer架构结合，有效捕捉各类矿物颗粒的关键特征。多尺度融合组件采用3×3和7×7卷积核分别提取并整合细/粗粒度特征，引入SENet形成的通道注意力机制则能强化关键特征通道的识别优先级，从而提升整体特征表征能力。针对同一矿物颗粒在不同偏光角度下可能呈现差异化特征导致的误分类问题，本研究进一步提出序列一致性优化（SCO）策略：通过比对该颗粒多角度图像的预测结果，依据类别一致性调整低置信度分类预测。置信度量化采用模型预测倾向清晰度（CPI）指标，该值由最高与次高预测概率计算得出。SCO策略通过该机制显著提升模型预测的一致性与准确性，为岩石薄片矿物颗粒自动化分类提供可靠支撑。

主要贡献可归纳如下：

1. MSCET模型通过融合多尺度特征与Transformer技术，显著提升了岩石薄片图像分类精度，为该领域提供了创新性应用方案。

the mineral composition, particle properties, and type of rock, vital information for oil and gas exploration and reservoir evaluation. For a considerable amount of time, traditional identification methods rely on experienced geologists observing and analyzing mineral particles in rock thin section images. However, the method's inherent subjectivity and lengthy process present significant challenges in the rapidly evolving field of exploration [1]. This underscores the necessity for more objective, efficient, and scalable methods in this critical area. Consequently, the quality of geological investigations can be significantly improved by automating the classification of mineral particles in rock thin section images.

The limitations of traditional methods have been steadily addressed by artificial intelligence techniques in recent years, driving the development of automated classification techniques for rock thin section images. And machine learning techniques were crucial in the early phases of the study. Using k-nearest neighbor classification on each descriptor for ultimate decision-making, [2] presented an advanced mix of classifiers for obtaining high-dimensional visual features from these images. A different approach evaluates the effectiveness of association rules, decision trees, and support vector machines in identifying the lithology of igneous rocks by using data mining techniques [3]. Despite the beneficial outcomes of these techniques, the dependency on manual feature determination made them limited to tiny datasets and resulted in poor generalization.

To overcome these limitations, researchers have increasingly turned to Convolutional Neural Networks (CNNs) for automated and accurate identification of rock thin section images [4]. For instance, [5] employed a CNN for efficient and accurate rock classification by automatically extracting image features. Comparably, Inception-v3 and ResNet18 have been used, respectively, by [6] and [7] in conjunction with image enhancement techniques to improve the classification of different types of rocks. Although current deep learning models have made significant advancements in rock thin section image classification, they remain limited by the inadequacy of network designs to effectively extract features from complex mineral particles. Furthermore, it is important to note that the rock thin section images of mineral particles will exhibit distinct features at various cross-polarized angles. The five common kinds of mineral particles exhibit varying colors and contrasts at five distinct angles, as depicted in Fig. 1. Unfortunately, the limitations of the current methodologies reduce the overall effectiveness of the model and cause classification errors due to the failure to account for the alteration of mineral particle properties under varying cross-polarized angles.

To address the challenges in extracting fine-grained features from rock thin section images, this study introduces an innovative image classification framework. The proposed approach utilizes the Multi-Scale Channel Enhanced Transformer (MSCET) model for initial image classification. By integrating Multi-Scale Fusion with Transformer architectures, the MSCET model captures essential features of various mineral particles. The Multi-Scale Fusion component employs convolutional kernels of sizes 3 and 7 to extract and integrate both fine- and coarse-grained features. Additionally, the incorporation of SENet introduces a channel attention mechanism, enhancing the model's ability to identify and prioritize critical feature channels, thereby improving overall feature representation. Furthermore, a Sequence Consistency Optimization (SCO) strategy is proposed to improve the classification of cross-polarized images of the same mineral particle captured from different angles. As the cross-polarized angle changes, the same particle may display distinct features, which can lead to potential misclassifications. The SCO strategy addresses this issue by comparing predictions from multi-angle images of the same particle and adjusting low-confidence classification predictions based on category consistency. The SCO strategy addresses this issue by comparing predictions from multi-angle images of the same particle and adjusting low-confidence classification predictions based on category consistency. Confidence levels are quantified using the model's Clarity of Predictive Inclination (CPI), which is calculated from the highest and second-highest prediction probabilities. By leveraging this approach, SCO enhances the consistency and accuracy of the model's predictions, providing robust support for the automated classification of mineral particles in rock thin sections.

The key contributions can be summarized as follows:

1. The MSCET model's integration of Multi-Scale Fusion with Transformer technology significantly enhances the precision of rock thin section image classification, presenting a novel application in this field.

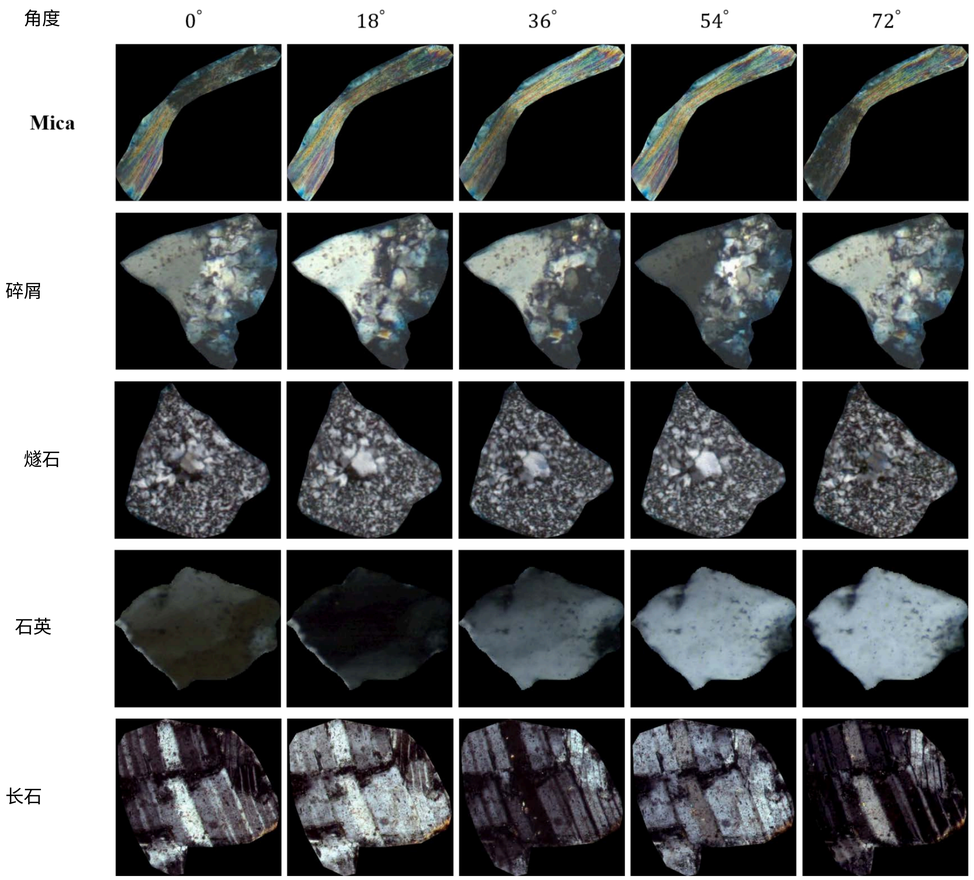


图1 五种矿物颗粒类型的岩石薄片图像

2. MSCET模型采用多尺度融合策略，通过不同尺寸的卷积核增强从岩石薄片图像中提取和表征多样化地质特征的能力。
3. SCO模块的引入有效解决了不同偏光角度带来的分类挑战，确保模型预测的准确性与一致性。

2 方法

如图1所示，矿物颗粒呈现高度不规则形态，且在不同偏光角度下表现出差异性特征。这类特征通常需要长距离分析，仅具局部感受野的卷积神经网络(CNNs)难以有效捕捉[8]。相比之下，Transformer架构凭借自注意力机制擅长处理全局信息和长程依赖，非常适合此类任务。因此，我们提出一种新方法来提升矿物颗粒在不同偏光角度下的识别鲁棒性。如图2所示，该方法首先使用MSCET进行初始分类，再通过SCO策略提升精度。

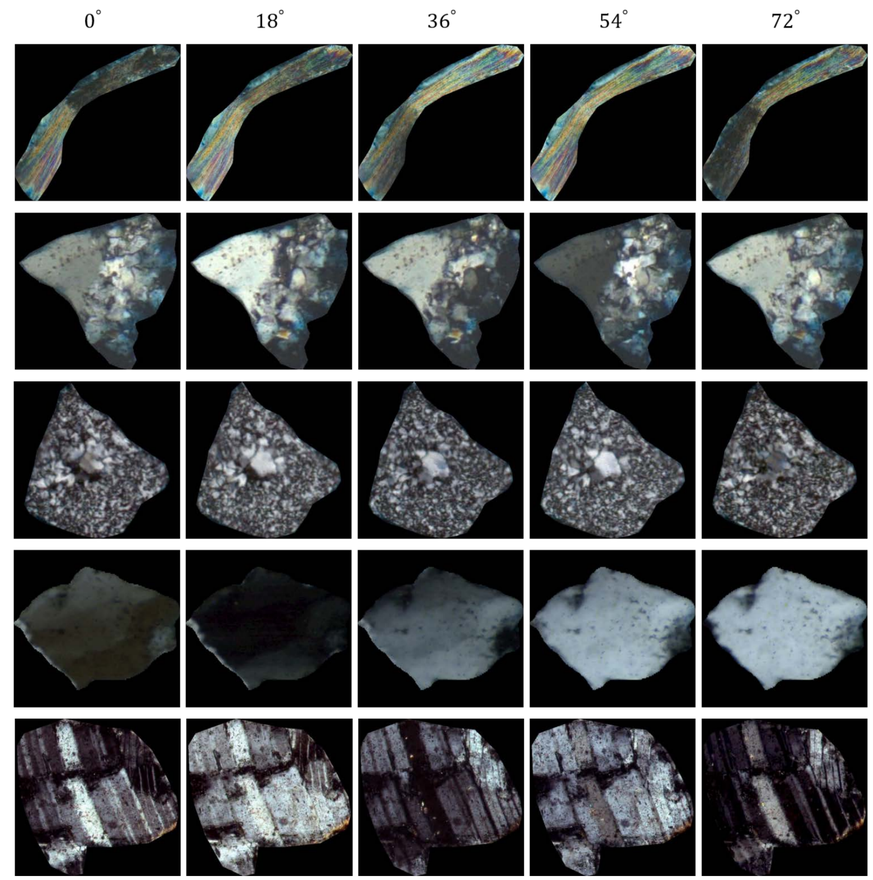


Fig. 1 Rock thin section images of five mineral particle types

2. Leveraging a Multi-Scale Fusion strategy, the MSCET model utilizes convolutions with varied kernel sizes to enhance its ability to extract and represent diverse geological features from rock thin section images.
3. The introduction of SCO effectively addresses the classification challenges presented by different cross-polarized angles, ensuring the accuracy and consistency of the model's predictions.

2 Method

Mineral particles exhibit highly irregular shapes and display varying characteristics at different cross-polarized angles, as illustrated in Fig. 1. These features typically require analysis over long distances, making convolutional neural networks (CNNs) with only local receptive fields inadequate for effectively capturing them [8]. In contrast, the Transformer architecture, with its self-attention mechanism, excels at processing global information and long-range dependencies, making it well-suited for such tasks. Consequently, we propose a novel approach to improve the robustness of mineral particle identification across various cross-polarized angles. As illustrated in Fig. 2, our method begins with the MSCET for initial classification, followed by the SCO strategy to enhance accuracy.

https://doi.org/10.1007/s10596-025-10356-8

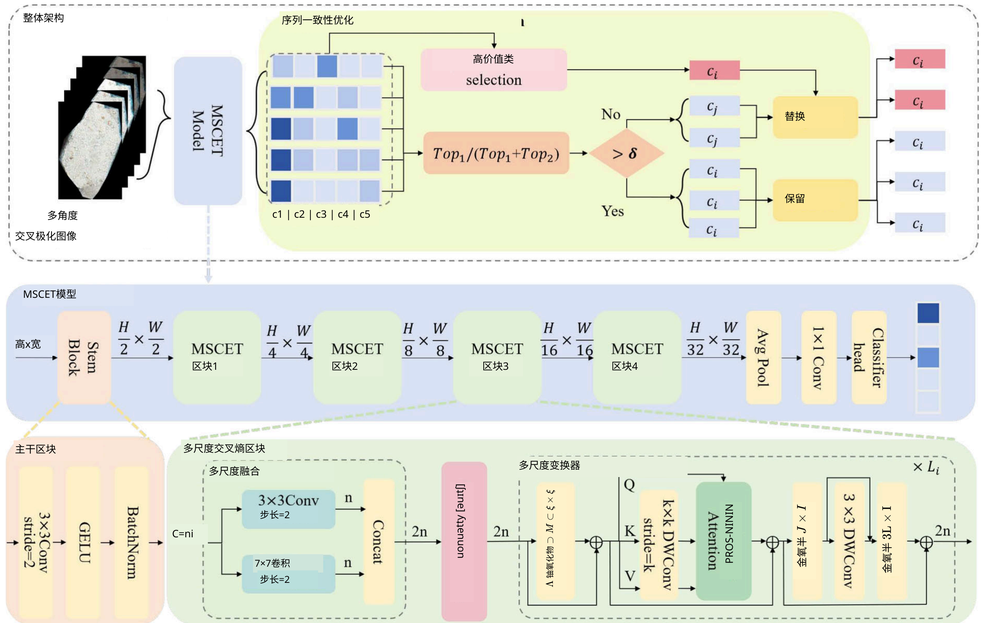


图2 提出的岩石薄片图像识别方法

MSCET模型通过集成多尺度融合模块和Transformer模块，有效提取矿物颗粒的局部与全局特征。通道注意力机制的引入进一步强化了纹理特征，实现了矿物颗粒复杂特征的自动化提取。该初始阶段为后续优化过程提供了必要的预测类别概率分布。

SCO技术采用基于序列的优化方法，通过不同角度偏光图像修正分类结果，旨在提升各类成像条件下的识别一致性，从而全面提升模型性能。

2.1 MSCET模型

由于矿物颗粒具有更复杂且视觉相似的特征，需整合全局与局部细节以提高分类精度。Transformer擅长处理全局信息，但在局部细节处理效率上不及CNN[8]。提出的MSCET模型通过结合Transformer的优势与CNN的局部特征提取能力解决这一问题。如图2所示，该模型由茎块(Stem Block)和四个MSCET块组成，最终通过全局平均池化、1×1 卷积投影和全连接层完成分类。

本方法采用卷积茎(convolution stem)这一Transformer新创新，通过利用图像内在空间关系 $[9, 10]$ 来改进局部特征表示。对于尺寸为 $H \times W \times 3$ 的输入图像，依次执行步长为2的 3×3 卷积、GELU激活函数和批量归一化层，最终生成尺寸为 $H/2 \times W/2 \times C$ 的输出。

特征图随后通过四个MSCET模块进行进一步细化，每个模块包含多尺度融合、通道注意力及堆叠的多尺度Transformer层，其层深度分别为3、3、16和3。

https://doi.org/10.1007/s10596-025-10356-8

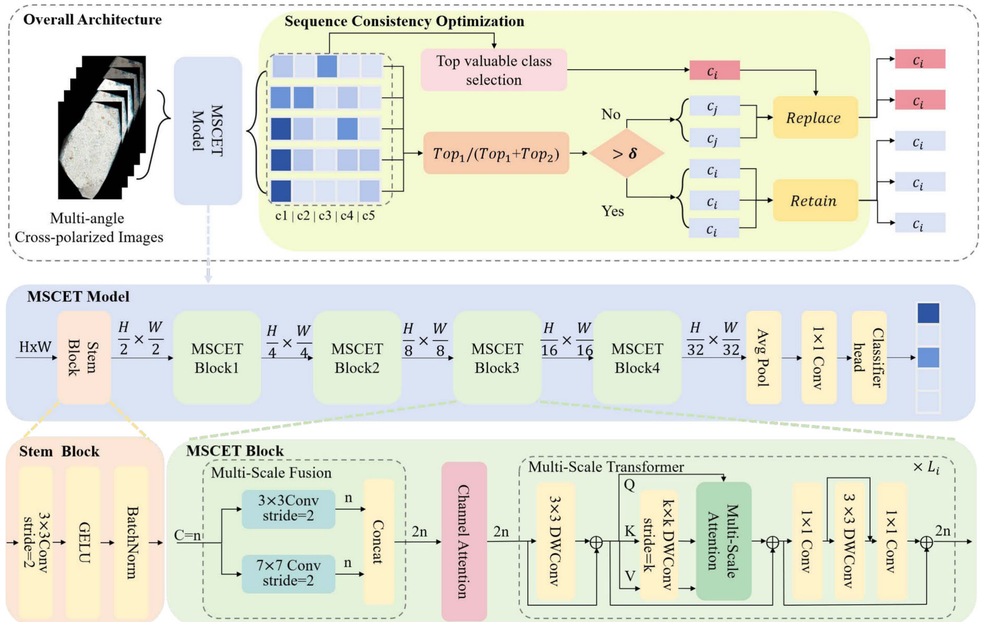


Fig. 2 The proposed rock thin section image identification method

The MSCET model effectively extracts both local and global features of mineral particles by integrating Multi-Scale Fusion and Transformer modules. The introduction of a channel attention mechanism further strengthens the textural features, automating the extraction of complex features of mineral particles. This initial stage provides the predicted category probability distribution necessary for the subsequent optimization process.

Through a sequence-based optimization approach that modifies classification results based on cross-polarized images at different angles, the SCO technique seeks to improve recognition consistency under various imaging situations, hence enhancing the overall performance of the model.

2.1 MSCET model

Because of the more complex and visually similar characteristics of mineral particles, it is necessary to integrate global and local details improve classification accuracy. The Transformer is particularly good at handling global information, but it is less efficient than CNN in processing local details [8]. The proposed MSCET model addresses this by combining the strengths of Transformers with the local feature extraction capabilities of CNNs. As depicted in Fig. 2, the model consists of a Stem Block and four MSCET Blocks, concluding with global average pooling, a 1×1 convolution projection, and a fully connected layer for classification.

Our method makes use of convolution stem, a new Transformer innovation that improves local feature representation by utilizing the intrinsic spatial relationships in images [9, 10]. For an input image of dimensions $H \times W \times 3$, a 3×3 convolution with a stride of 2, GELU activation function, batch normalization layer are performed successively, producing outputs of size $H/2 \times W/2 \times C$.

The feature map then undergoes further refinement through four MSCET blocks, each consisting of Multi-Scale Fusion, Channel Attention, and stacked Multi-Scale Transformer layers, with layer depths of 3, 3, 16, and 3 respectively.

2.1.1 多尺度融合

矿物颗粒尺寸的差异导致其岩石薄片图像呈现不同维度，这对传统单卷积技术提出了挑战。这些技术往往难以跨尺度提取特征，限制了模型的表征能力。尤其在燧石类别中，内部石英晶体尺寸的变化凸显了实施多尺度融合策略的必要性，以全面表征这些复杂特征。

受文献[11]采用多并行空洞卷积进行多尺度特征提取的启发，我们开发了多尺度融合方法来增强岩石薄片图像分析。具体而言，前一层输入的 $X_{in}^{i} \in R^{H_{i-1}\times W_{i-1}\times C_{i-1}}$ （其中 i 取值1至4，4代表堆叠MSCET模块的数量）会通过两种不同核尺寸的卷积并行处理后再进行拼接。

基于所述方法，我们的实验采用了核尺寸为 3×3 和 7×7 的卷积层。 7×7 核擅长覆盖更广的空间范围，有助于提取较大尺度的特征。其与 3×3 核的结合实现了局部纹理细节捕捉与更广阔全局模式提取之间的平衡。为确保输出特征图尺寸一致，我们为 3×3 和 7×7 核分别设置了步长为2、填充为1和3的参数。最终输出为 $X_{m}^{i} \in \mathbb{R}^{H_i\times W_i\times C_i}$ ，其中 $C_i = 2\times C_{i-1}, H_i = H_{i-1}/2$ ，且 $W_i = W_{i-1}/2$ 。

多尺度融合是本模型的核心创新，能有效处理颗粒尺寸的变异性。它整合了不同尺度的信息，突破了单尺度提取方法的局限，从而提升特征表征能力并改善分类性能。

2.1.2 通道注意力机制

传统CNN常难以区分岩石薄片图像中相似的纹理特征（如碎屑与长石虽属不同类别但视觉相似），这种相似性会阻碍模型识别细微特征差异以实现准确分类。为增强特征辨别力，我们引入挤压-激励网络(SENet)[12]，该通道注意力机制显著提升了模型识别细微特征变化的能力。

对于来自多尺度融合的输入 $X_{m}^{i} \in \mathbb{R}^{H\times W\times C}$ ，挤压操作将 X_{m}^{i} 的每个通道全局池化为单一值，以获取其全局感受野。

$$z_c = \frac{1}{H\times W}\sum_{j=1}^{H}\sum_{k=1}^{W} X_m^i(j,k) \tag{1}$$

该公式表示对每个通道进行全局平均池化，将空间维度压缩为每个通道的单一标量。

随后，激励阶段采用两个线性层，分别使用ReLU和Sigmoid激活函数，生成通道特异性权重。

$$s = \sigma(W_2\delta(W_1 z)) \tag{2}$$

其中 δ 表示ReLU激活函数， σ 表示Sigmoid激活函数，生成的权重集 s 用于对通道输出进行重新校准。

最后一步是将这些权重 s 按通道逐一对原始特征 X_{m}^{i} 进行加权，从而在通道维度上完成对原始特征的重新校准。

$$X_c^i = s_c \cdot X_m^i \tag{3}$$

https://doi.org/10.1007/s10596-025-10356-8 Page 5 of 19 19

2.1.1 Multi-scale fusion

The varying sizes of mineral particles lead to different dimensions in their rock-thin section images, posing challenges for conventional single-convolution techniques. These techniques often struggle to extract features across all scales, which limits the model's representational power. Particularly in the Flint category, the variation in internal quartz crystal size accentuates thenecessity of implementing a Multi-Scale Fusion strategy to comprehensively represent these complex features.

Drawing inspiration from [11], which utilizes multiple parallel atrous convolutions for multi-scale feature extraction, we develop a Multi-Scale Fusion approach to enhance the analysis of rock thin section images. Specifically, an input $X_{in}^{i} \in R^{H_{i-1}\times W_{i-1}\times C_{i-1}}$ from the previous layer (where i ranges from 1 to 4, with 4 representing the number of stacked MSCET Blocks) is processed in parallel through convolutions with two distinct kernel sizes and then concatenated.

Building on the approach described, our experiments utilized convolution layers with kernel sizes of 3×3 and 7×7. The 7×7 kernel is adept at covering a broader spatial range, which aids in the extraction of larger-scale features. Its combination with the 3×3 kernel achieves a balance between the detailed capture of local textures and the extraction of more expansive global patterns. To ensure consistency in the output feature map size, we set a stride of 2 and padding of 1 and 3 for the 3×3 and 7×7 kernels, respectively. This resulted in an output $X_{m}^{i} \in \mathbb{R}^{H_i\times W_i\times C_i}$, where $C_i = 2\times C_{i-1}$, $H_i = H_{i-1}/2$, and $W_i = W_{i-1}/2$.

The Multi-Scale Fusion is a key innovation in our model, adeptly handling the variability in particle sizes. It effectively integrates disparate scale information and transcends the constraints of single-scale extraction methods, leading to enhanced feature representation and improved classification performance.

2.1.2 Channel attention

Traditional CNNs often struggle to differentiate between similar texture features in rock thin section images, such as those between Detritus and Feldspar, despite their different categories. Such visual resemblances can impede the model's ability to discern subtle feature differences for accurate classification. To enhance feature discrimination, we integrate the Squeeze-and-Excitation Network (SENet) [12], a channel attention mechanism that substantially enhances the model's ability to identify nuanced feature variations.

For an input $X_{m}^{i} \in \mathbb{R}^{H\times W\times C}$ from the Multi-Scale Fusion, the Squeeze operation globally pools each channel of X_{m}^{i} into a single value to obtain its global receptive field.

$$z_c = \frac{1}{H\times W}\sum_{j=1}^{H}\sum_{k=1}^{W} X_m^i(j,k) \tag{1}$$

This equation represents the global average pooling of each channel, reducing spatial dimensions to a single scalar per channel.

Following this, the Excitation stage employs two linear layers, with ReLU and sigmoid activation functions, to generate channel-specific weights.

$$s = \sigma(W_2\delta(W_1 z)) \tag{2}$$

Here, δ denotes the ReLU activation function, and σ the sigmoid activation, producing a set of weights s that are used to recalibrate the channel outputs.

The final step involves applying these weights s to the original features X_{m}^{i} on a per-channel basis, thereby recalibrating the original features in the channel dimension.

$$X_c^i = s_c \cdot X_m^i \tag{3}$$

Comput Geosci (2025) 29:19 Springer

SENet通过聚合各通道的全局信息并计算自适应权重来重新校准通道。这种重校准机制使模型能聚焦于信息量更丰富的特征，从而增强网络的判别能力。这种针对性校准显著提升了模型对矿物颗粒固有纹理特征的辨识与表征能力，进而改善整体分类性能。

2.1.3 多尺度变换器

卷积运算能高效提取矿物颗粒的局部特征（如细粒度纹理和结构细节），但其提取全局特征（如消光特性）的能力受限于卷积核的局部性。虽然池化操作或更深的CNN架构可聚合更广的上下文信息，但这些方法常导致细粒度空间细节的丢失——这些细节对准确识别正交偏光条件下的矿物颗粒至关重要。相比之下，Transformer的复杂注意力机制能精确建模长程依赖关系并有效整合全局上下文信息，从而与卷积运算的局部特征提取能力形成互补。

我们采用如图2所示的多尺度Transformer架构。与传统绝对位置编码不同，该架构首先采用 3×3 深度卷积进行位置编码。这种深度卷积能隐式学习局部区域的位置信息，灵活适应不同分辨率的输入，并更高效地捕捉图像局部特征[13]。随后运用多尺度注意力机制提取粒子间的全局信息与上下文关系。为降低计算复杂度，我们采用轻量级注意力(LightAttn)[14]。

LightAttn通过深度卷积与线性变换的策略性整合，提升了自注意力机制的效率。给定输入 $X_c^i \in R^{N \times C}$ ，其中 N 表示token数量（其高度与宽度的乘积）， C 为特征通道数，该模型首先生成查询向量 $Q \in R^{N \times d_k}$ 。同时，对键值和值向量采用 $k \times k$ 深度卷积进行降维处理，得到 $K' \in R^{\frac{N}{k^2} \times d_k}$ 和 $V' \in R^{\frac{N}{k^2} \times d_v}$ 。此外，通过双三次插值适配不同维度的相对位置偏置 B ，增强了注意力的适应性与计算效率。LightAttn的计算过程如公式4所示。

$$LightAttn(Q, K, V) = \text{Softmax}\left(\frac{QK'^T}{\sqrt{d_k}} + B\right)V' \tag{4}$$

深度卷积的应用显著降低了计算负荷，时间复杂度缩减至 $O\left(NC + \frac{N^2}{k^2}C\right)$ 。这种降低使得MSCET模块能够根据不同 k 值（8、 $4, 2, 1$ ）进行可扩展处理，优化了细节保留与处理效率之间的平衡。

与传统前馈神经网络[15]不同，我们采用纯卷积前馈网络结构，包含两个 1×1 卷积层，中间插入 3×3 深度卷积层以增强局部特征表征能力。

多尺度Transformer扩展了对岩石薄片图像不同尺度的理解，提升了模型识别关键特征的能力。此外，LightAttn模块降低了计算需求，使其能更高效地处理海量图像数据，并确保在资源有限环境下仍保持可靠性能。

2.2 序列一致性优化

由不同成岩环境导致的矿物外观多样性及消光特性，对岩石薄片鉴定中矿物颗粒的有效分类提出了挑战。特别是当从多个正交偏光角度观察时，后者会导致颗粒出现显著变化

SENet operates by aggregating global information from each channel and computing adaptive weights to recalibrate channels. This recalibration allows the model to focus on more informative features, thus enhancing the discriminative power of the network. Such targeted recalibration significantly boosts the model's ability to discern and represent the diverse textural characteristics inherent in mineral particles, improving overall classification performance.

2.1.3 Multi-scale transformer

Convolution operations are highly effective at extracting local features of mineral particles, such as fine-grained textures and structural details. However, their capacity to extract global features, such as extinction characteristics, is constrained by the localized nature of convolutional kernels. Although pooling operations or deeper CNN architectures can aggregate broader context, these approaches often lead to a loss of fine-grained spatial details, which are essential for accurately identifying mineral particles under cross-polarized conditions. In contrast, the sophisticated attention mechanism of Transformers allows for precise modeling of long-range dependencies and effective integration of global contextual information, thereby complementing the local feature extraction capabilities of convolution operations.

We adopt a Multi-Scale Transformer architecture, depicted in Fig. 2. Diverging from traditional absolute position encoding, it begins with a 3×3 depthwise convolution for position encoding. This depthwise convolution implicitly learns positional information in local areas, providing the flexibility to adapt to inputs of various resolutions and capturing local image features more efficiently [13]. We then apply a multi-scale attention mechanism to extract global information and contextual relationships among the particles. To mitigate the computational demands, we employ Lightweight Attention (LightAttn) [14].

LightAttn enhances the efficiency of the self-attention mechanism through a strategic integration of depthwise convolution and linear transformations. Given an input $X_c^i \in R^{N \times C}$, where N is the number of tokens (the product of its height and width) and C is the number of feature channels, the model initially applies a linear transformation to generate queries $Q \in R^{N \times d_k}$. Simultaneously, it employs a $k \times k$ depthwise convolution on the keys and values to reduce their dimensions, resulting in $K' \in R^{\frac{N}{k^2} \times d_k}$ and $V' \in R^{\frac{N}{k^2} \times d_v}$. Additionally, a relative positional bias B, adapted through bicubic interpolation to fit various dimensions, enhances the attention's adaptability and efficiency. The computation of LightAttn is as shown in Eq. 4.

$$LightAttn(Q, K, V) = \text{Softmax}\left(\frac{QK'^T}{\sqrt{d_k}} + B\right)V' \tag{4}$$

The depthwise convolution applied significantly decreases the computational load, with time complexity reduced to $O(NC + \frac{N^2}{k^2}C)$. This reduction allows for scalable processing across MSCET blocks with varying k values (8, 4, 2, 1), optimizing the balance between detail retention and processing efficiency.

Differing from the traditional feedforward neural network [15], we use a purely convolutional feedforward network, comprising two 1×1 convolutional layers with a 3×3 depthwise convolutional layer in between to bolster local feature representation.

The Multi-Scale Transformer extends the comprehension of different scales in rock thin section images and improves the model's capacity to identify important features. Furthermore, the LightAttn module reduces computing needs, making it easier to handle huge amounts of image data efficiently and guaranteeing reliable performance even in contexts with limited resources.

2.2 Sequence consistency optimization

The variety in appearance caused by different lithogenic circumstances and the extinction qualities of minerals poses a challenge to the effective classification of mineral particles in rock thin sections identification. When examined from multiple cross-polarized angles, the latter in particular can cause notable changes in a particle's

亮度与纹理特征的差异会导致特征不一致，从而降低基于单角度图像的识别方法的可靠性。

针对这些挑战，我们开发了序列一致性优化（SCO）方法。与传统图像分类方法直接选择最高概率类别作为最终预测不同，我们的方法通过五幅图像的预测结果迭代分析，将具有最高 Top_1 得分的类别确定为最具价值类别。

鉴于MSCET模型的输出能明确指示特定类别，我们将输出中的最高概率和次高概率分别视为 Top_1 和 Top_2。较高的 Top_1 与 Top_2 比值表明模型在类别区分上具有显著效果，而较低比值则预示预测可能存在模糊性。为量化模型的预测倾向明确度（CPI），我们采用公式5进行计算。

$$\frac{Top_1}{Top_1 + Top_2} \geq \delta \tag{5}$$

通过设定阈值 δ，当CPI超过该值时，MSCET模型的预测被视为可靠，结果予以保留；若CPI低于 δ，则认为预测置信度不足，此时将选取最具价值类别作为图像的最终预测结果。

SCO在保持MSCET模型固有分类能力的同时，有效应对预测偏差。该策略显著降低了数据变异导致的识别错误率，从而提升分类结果的整体稳定性和准确性。

3 实验

3.1 数据集

本研究所用图像源自全岩薄片在偏光显微镜下的拍摄样本。这些岩石薄片样品采自不同地质区块，涵盖多种成岩环境。图3展示了同一薄片样品在同一视域下拍摄的图像，其中图3(a)为单偏光图像，图3(b-f)则是在 $0°, 18°, 36°, 54°, 72°$ 角度下拍摄的交叉偏光图像。这些角度的选择基于岩相学分析的经验知识与标准流程，能有效揭示矿物颗粒的光学特性变化，从而辅助实现更精确的矿物颗粒分类。

在单偏光图像中，矿物颗粒通常呈现无色透明状态，这限制了其在矿物鉴定中的应用价值。而正交偏光图像则能展现矿物独特的消光特性所产生的一系列鉴别特征，如条纹状结构和明暗交替的颗粒分布，从而有效区分不同种类的矿物颗粒。为此，本研究从全岩薄片不同角度拍摄的正交偏光图像中提取单颗粒图像及对应标签，采用黑色填充物作为背景以消除背景噪声并增强颗粒特征的可见度。

由于岩薄片中矿物颗粒形态与尺寸各异，数据集中颗粒图像分辨率存在显著差异。同时为提升模型对复杂现实环境的适应能力，数据集保留了薄片制备过程中可能产生的、影响颗粒特征的噪声数据，例如切片操作不当引入的气泡或裂隙。图1展示了数据集中不同矿物颗粒在多种正交偏光角度下的图像，涵盖云母、岩屑、燧石、石英及长石等类型，这些矿物合计占地质勘探开发常见矿物的90%以上。

此外，为确保数据集的多样性，我们检查了同一样本的交叉极化图像，剔除了过度相似的样本。经数据处理后，该数据集共包含58,238张图像。

brightness and texture, resulting in inconsistent features and diminishing the reliability of identification methods based on single-angle images.

In response to these challenges, we develop the Sequence Consistency Optimization (SCO) method. Unlike traditional image classification methods that select the highest probability category as the final prediction, our approach iterates through the five images' predictions and identifies the category with the highest Top_1 score as the Top valuable class.

Given that the MSCET model outputs decisively indicate a specific category, we consider the highest and second-highest probabilities in the output as Top_1 and Top_2, respectively. A high Top_1 to Top_2 ratio shows the model's effectiveness in distinguishing a clear category. In contrast, a lower ratio suggests potential ambiguity in the predictions. To quantify the model's Clarity of Predictive Inclination (CPI), we apply Eq. 5.

$$\frac{Top_1}{Top_1 + Top_2} \geq \delta \tag{5}$$

By setting a threshold δ, the MSCET model's prediction is deemed reliable when the CPI exceeds this value, and the results are retained. If the CPI is below δ, the prediction is considered insufficiently confident, prompting the selection of the Top valuable class as the final prediction for the image.

SCO effectively counters prediction inaccuracies while maintaining the MSCET model's inherent classification abilities. This strategy significantly reduces the recognition error rate due to data variability, thereby boosting the overall stability and accuracy of the classification results.

3 Experiments

3.1 Datasets

The images used in this study were obtained from complete rock thin sections photographed using a polarizing microscope. These rock thin section samples were collected from different geological blocks, encompassing diverse diagenetic environments. Figure 3 shows images of the same thin section sample taken in the same field of view, where Fig. 3(a) represents the single-polarized image, and Fig. 3(b-f) are cross-polarized images taken at angles of 0°, 18°, 36°, 54°, 72°. These angles were selected based on empirical knowledge and standard practices in petrographic analysis, effectively revealing the optical characteristic variations of mineral particles, thus aiding in more accurate mineral particle classification.

In single-polarized images, mineral particles typically appear colorless and transparent, which limits their usefulness for mineral identification. In contrast, cross-polarized images reveal a range of distinctive features due to the unique extinction properties of the minerals, such as stripes and alternating bright and dark grains. This enables effective differentiation among various types of mineral particles. Consequently, this study extracted single-particle images and their corresponding labels from cross-polarized images of complete rock thin sections captured at different angles. A black filler was employed as the background to eliminate background noise and enhance the visibility of the particle features.

Due to the varying shapes and sizes of mineral particles in rock thin sections, the particle images in the dataset exhibit diverse resolutions. Additionally, to enhance the model's adaptability to complex real-world environments, the dataset retains noise data that affects the particles, which may arise during thin section preparation. This noise can include bubbles or cracks introduced by improper slicing operations. Figure 1 illustrates images of different mineral particles in the dataset under various cross-polarized angles. The mineral types represented include Mica, Detritus, Flint, Quartz, and Feldspar, which collectively account for over 90% of the minerals commonly found in geological exploration and development.

Further, to ensure that the dataset is diverse, we examined cross-polarized images from the same sample, eliminating any that were unduly similar. The dataset consists of 58,238 images after data processing. It was made

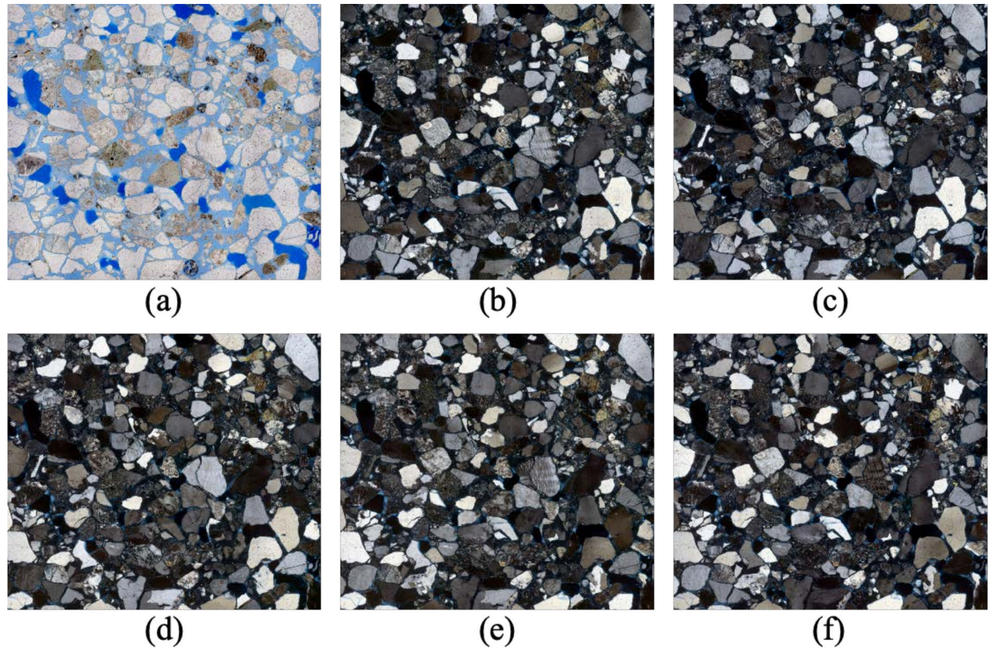


图3 不同角度下的单偏光与正交偏光岩石薄片图像

确保同一样本的所有图像仅被分配到训练集或测试集，避免任何重叠。因此，数据集按80%和20%的比例划分为训练集（46,600张图像）和测试集（11,638张图像）。表1详细展示了矿物类型的分布情况，其样本量差异显著反映了矿物分布的自然变异。为保持自然环境岩石薄片分类的真实性，我们对样本较少的类别未进行数据增强。通过这种方法，我们旨在为理解和分析岩石薄片图像分类问题提供更准确、更现实的视角。

3.2 评估指标

本文采用准确率(ACC)、精确率(P)、召回率(R)和F1值(F1)四项指标对岩石薄片分类性能进行综合评估。

表1 岩石数据集
薄片图像分类

矿物类型	图像数量	训练数据	测试数据
碎屑	24640	19715	4925
石英	19139	15313	3826
长石	12580	10067	2513
燧石	1048	839	209
云母	831	666	165

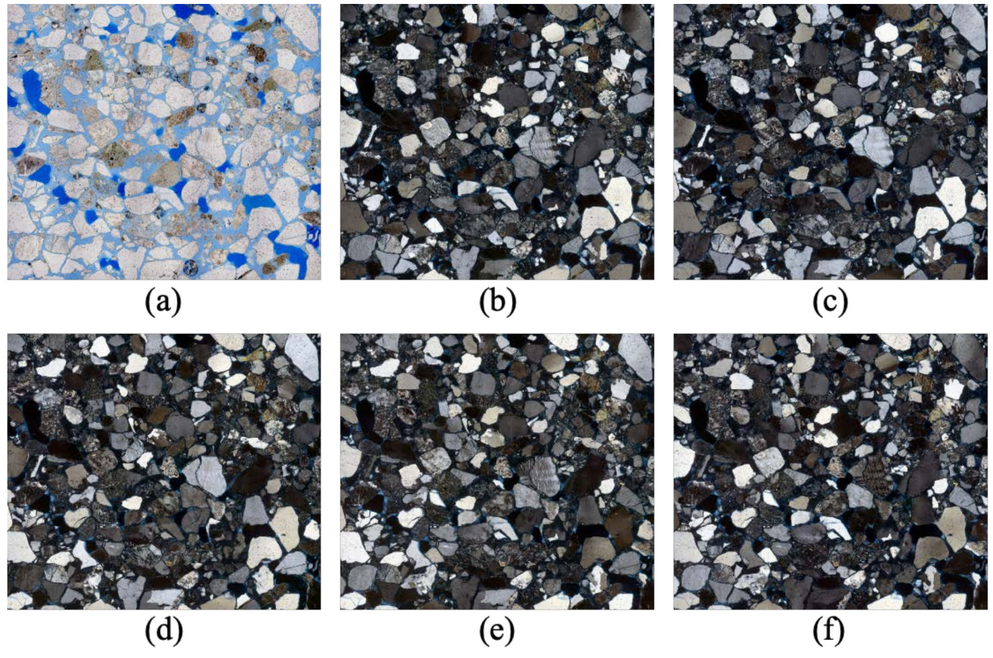


Fig. 3 Single-polarized and cross-polarized rock thin section images at different angles

sure that each image from the same sample was exclusively assigned to either the training or test set, preventing any overlap. Accordingly, the dataset was divided into training and testing sets at proportions of 80% and 20%, containing 46,600 and 11,638 images, respectively. Table 1 details the distribution and shows significant sample size disparities among the mineral types reflecting natural variance in mineral distribution. To maintain the realities of rock thin section classification in natural environments, we avoided data augmentation for categories with fewer samples. Through this approach, we seek to offer a more accurate and realistic perspective for understanding and analyzing the classification issues of rock thin section images.

3.2 Evaluation metrics

To evaluate the performance of rock thin section classification, this paper adopts four metrics for comprehensive evaluation: Accuracy (ACC), Precision (P), Recall (R), and F1 score ($F1$).

Table 1 Dataset for rock thin section image classification

Mineral type	Number of images	Training data	Testing data
Detritus	24640	19715	4925
Quartz	19139	15313	3826
Feldspar	12580	10067	2513
Flint	1048	839	209
Mica	831	666	165

各项指标的计算公式如下：

$$
\begin{aligned}
ACC &= \frac{TP + TN}{TP + TN + FP + FN} \\
P &= \frac{TP}{TP + FP} \\
R &= \frac{TP}{TP + FN} \\
F1 &= 2 \times \frac{P \times R}{P + R}
\end{aligned} \tag{6}
$$

其中TP（真正例）和FP（假正例）分别表示对某类别的正确预测与错误预测，TN（真负例）和FN（假负例）则分别表示对不属于某类别的正确预测与错误预测。

ACC 表示模型准确分类的样本占总样本集的比例。 P 指被预测为某类别的样本中确实属于该类别的比例。 R 代表实际属于某类别的样本中被正确识别的比例。 $F1$ 是精确率与召回率的调和平均数，可综合评估这两个指标。

3.3 实验描述

模型训练在Nvidia A100 GPU上使用PyTorch 1.12.0完成。我们设置批次大小为128，进行400轮训练，采用AdamW算法优化模型。该算法动量设为0.9，权重衰减为 5×10^{-2} 。初始学习率为 5×10^{-4} ，并采用每30轮衰减0.1的策略逐步优化学习过程。选用软目标交叉熵作为损失函数，通过标签平滑有效降低噪声敏感性。为抑制过拟合，当验证损失不再显著下降时即终止训练。

针对序列一致性优化方法，经多次测试后，我们确定将阈值 δ 设为0.6时，能在分类精度与整体性能间取得最佳平衡。

3.4 实验结果

我们将数据集载入MSCET模型进行迭代训练。图4展示了测试集上损失函数与类别准确率的变化趋势：初期损失值显著下降预示预测准确率提升，后期趋于平稳表明模型收敛；准确率持续上升并最终稳定在较高水平，凸显该模型在多矿物类型分类中的有效性。

表2展示了我们MSCET模型与ResNet[16]、VggNet[17]、DenseNet[18]、InceptionNet[19]、Vision Transformer[15]及Swin Transformer[10]等主流模型的对比评估。本方法在测试集上表现出卓越的准确率（92.35%），同时在精确率（89.38%）、召回率（90.33%）和F1分数（89.84%）等指标上全面领先。值得注意的是，这一优异性能仅需 $30.20\mathrm{M}$ 个相对精简的参数即可实现，体现了其高效的参数利用率。MSCET模型成功的关键在于融合了CNN的局部特征提取能力与Transformer的全局信息处理优势，这种混合架构能更全面地理解矿物颗粒岩石薄片图像的复杂特征，相较传统CNN或纯Transformer模型具有显著优势。

The formulas for these metrics are as follows:

$$
\begin{aligned}
ACC &= \frac{TP + TN}{TP + TN + FP + FN} \\
P &= \frac{TP}{TP + FP} \\
R &= \frac{TP}{TP + FN} \\
F1 &= 2 \times \frac{P \times R}{P + R}
\end{aligned} \tag{6}
$$

Where TP (True Positive) and FP (False Positive) refer to correct and incorrect predictions of a category, respectively. TN (True Negative) and FN (False Negative) indicate correct and incorrect predictions of non-belonging to a category, respectively.

ACC denotes the fraction of samples accurately classified by the model out of the total sample set. P signifies the ratio of samples correctly identified as a specific category among all samples predicted to belong to that category. R represents the proportion of samples correctly recognized as a specific category out of the total samples belonging to that category. $F1$ denotes the harmonic mean of precision and recall, offering a comprehensive assessment of both precision and recall aspects.

3.3 Experimental description

The training of our model was executed using PyTorch 1.12.0 on an Nvidia A100 GPU. We set the batch size to 128 and conducted training over 400 epochs, optimizing the model with the AdamW algorithm. This algorithm included a momentum of 0.9 and a weight decay of 5×10^{-2}. The initial learning rate was 5×10^{-4}, with a decay strategy reducing it by 0.1 every 30 epochs to refine learning over time. We chose soft target cross-entropy as the loss function, effective for reducing noise sensitivity through label smoothing. To combat overfitting, training was halted upon observing no significant reduction in validation loss.

For the Sequential Consistency Optimization method, after conducting multiple tests, we determined that setting the threshold δ at 0.6 provides the best balance between classification accuracy and overall performance for our dataset.

3.4 Results

We loaded the dataset into the MSCET model for iterative training. Figure 4 illustrates the trends in the loss function and category accuracy on the test set. Initially, a significant loss reduction indicated improved predictive accuracy, which stabilized over time, signaling model convergence. The accuracy exhibited a consistent upward trend, eventually stabilizing at a high-performance level, highlighting the model's effectiveness in classifying various mineral types.

Table 2 provides a comparative evaluation of our MSCET model against leading models such as ResNet [16], VggNet [17], DenseNet [18], InceptionNet [19], Vision Transformer [15], and Swin Transformer [10]. Our method demonstrated superior accuracy, achieving 92.35% on the test set. Furthermore, it excelled in precision, recall, and F1 scores, reaching 89.38%, 90.33%, and 89.84%, respectively. Notably, this superior performance was attained with a relatively modest parameter count of 30.20M, highlighting its efficient utilization of parameters. The success of the MSCET model can be attributed to its integration of CNN's local feature extraction capabilities with the Transformer's global information processing strengths. This hybrid approach facilitates a more comprehensive understanding of the complex characteristics of rock thin section images of mineral particles, offering significant advantages over traditional CNN models or purely Transformer-based models.

 https://doi.org/10.1007/s10596-025-10356-8

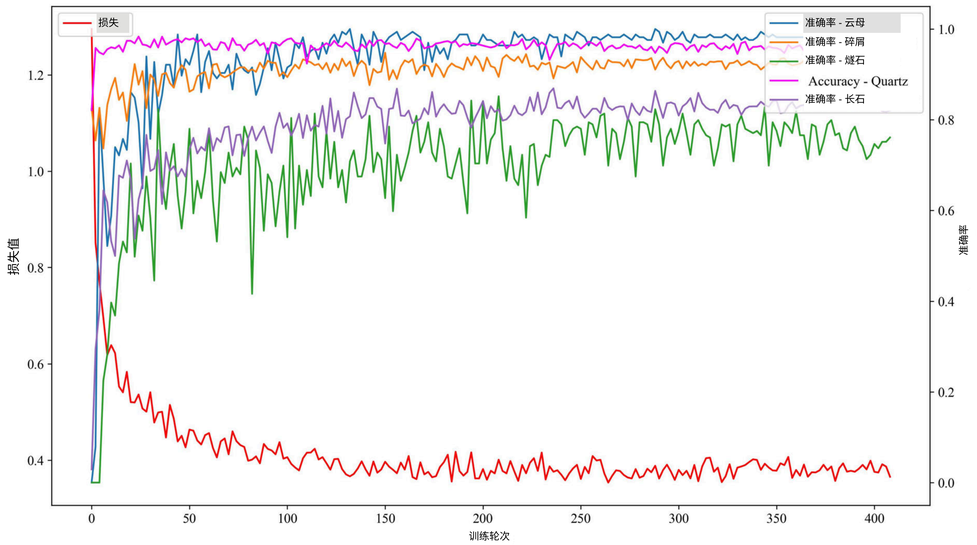


图4 MSCET模型的测试损失值与准确率

此外，我们评估了所提方法在五种不同矿物颗粒分类中的表现。如表3所示，该模型展现出卓越的准确率，特别是在识别云母（98.18%）和石英（95.95%）方面。这些类别的精确率、召回率和F1分数同样出色，凸显了模型在区分其关键特征上的优势。

对于由石英和长石等矿物组成的复杂类别碎屑岩，模型表现出强劲性能，所有指标均保持在92%以上。这表明模型擅长处理混合矿物成分。相比之下，长石分类因绢云母化样本导致特征模糊而面临挑战。尽管如此，模型仍达到86.63%的准确率，证明其识别细微特征差异的能力。燧石因视觉上与石英岩相似且样本有限，带来独特困难，但模型仍取得77.99%的准确率，展现了其在处理复杂且小样本类别时的有效性。

总体而言，MSCET模型在岩石薄片图像分类中的效率和精确度表现显著，尽管在低效类别中仍有改进空间。这些结果为地质学和矿物学专业人员提供了岩石薄片图像自动化分类的重要参考。

表2 岩石薄片图像分类方法对比

方法	准确率(%)	精确率(%)	召回率(%)	F1分数(%)	参数量(百万)
残差网络50	88.64	88.92	81.14	83.97	24
VGG16	89.45	88.68	81.45	84.53	134
密集网络121	88.87	86.40	84.10	85.05	**7**
初始网络V3	88.70	86.58	86.07	86.13	25
视觉变换器	87.18	84.83	79.64	81.67	86
Swin Transformer	85.99	83.42	77.27	79.57	87
我们的方法	**92.35**	**89.38**	**90.33**	**89.84**	30

加粗条目表示最优方法

 https://doi.org/10.1007/s10596-025-10356-8

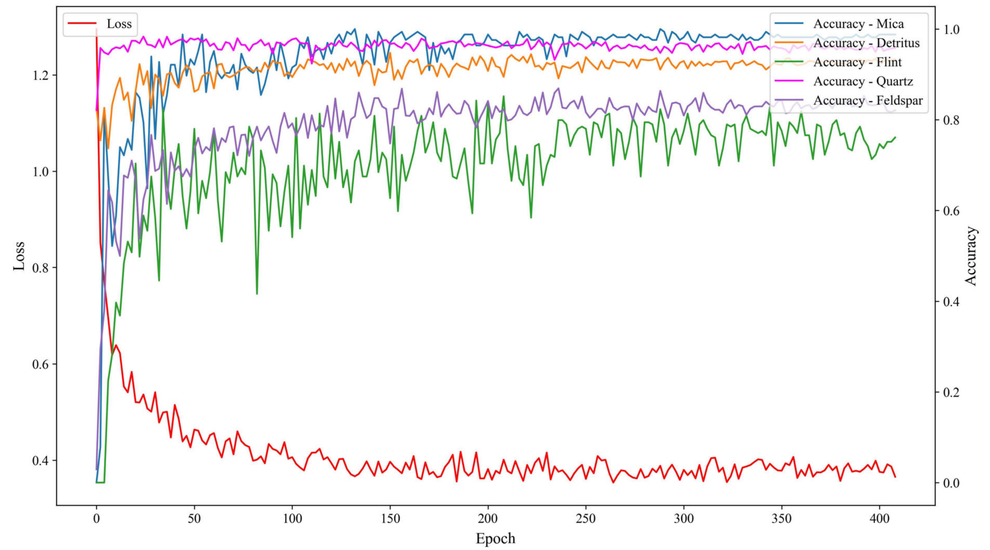


Fig. 4 The test loss and accuracy of the MSCET model.

Furthermore, we evaluated the performance of our proposed method in classifying five distinct mineral particle categories. As shown in Table 3, the model demonstrated remarkable accuracy, especially in identifying Mica (98.18%) and Quartz (95.95%). The precision, recall, and F1 scores in these categories were equally impressive, highlighting the model's proficiency in distinguishing their defining characteristics.

For Detritus, a complex category comprising minerals like Quartz and Feldspar, the model demonstrated robust performance, maintaining over 92% across all metrics. This emphasizes the model's adeptness in handling mixed mineral compositions. In contrast, classifying Feldspar faced challenges due to sericitized samples that obscured distinct features. Despite these difficulties, the model achieved 86.63% accuracy, demonstrating its capability to discern subtle differences in features. Flint, often visually similar to quartzite and with limited samples, presented unique obstacles. Nevertheless, the model managed a 77.99% accuracy, showcasing its effectiveness in dealing with complex and limited-sample categories.

Overall, the MSCET model's efficiency and precision in rock thin section image classification are evident, though there is room for improvement in categories with lower performance. These results offer valuable insights for geology and mineralogy professionals in automated rock thin section image classification.

Table 2 Comparison of classification methods for rock thin section images

Method	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)	Params(M)
ResNet50	88.64	88.92	81.14	83.97	24
VGG16	89.45	88.68	81.45	84.53	134
DenseNet121	88.87	86.40	84.10	85.05	**7**
Inception V3	88.70	86.58	86.07	86.13	25
Vision Transformer	87.18	84.83	79.64	81.67	86
Swin Transformer	85.99	83.42	77.27	79.57	87
Our Method	**92.35**	**89.38**	**90.33**	**89.84**	30

Entries in bold indicate the best methods

表3 性能指标
本方法各分类指标

类别	准确率(%)	精确率(%)	召回率(%)	F1分数(%)
云母	98.18	97.59	98.18	97.89
碎屑岩	92.89	92.95	92.89	92.92
燧石	77.99	73.76	77.99	75.81
石英	95.95	96.89	95.95	96.41
长石	86.63	85.71	86.63	86.17

3.5 消融实验

3.5.1 关键模块

我们通过消融实验分别评估策略各组成部分的独立贡献。"原始模型"基于CMT架构[14]，该架构结合了CNN与Transformer。随后我们通过重新设计茎部结构并用多尺度融合替代补丁嵌入，构建了"改进版原始模型"。如表4所示，在改进版基础上进一步加入通道注意力和序列一致性优化等模块后，分类准确率逐步提升，验证了各模块的有效性。

Origin模型在捕捉复杂特性方面的缺陷，通过其对云母和石英检测的卓越表现与燧石识别效果不佳的对比得以揭示。我们通过定制化茎干结构和多尺度融合等改进措施，显著提升了燧石类别的识别精度，证实了精细化多尺度特征提取的必要性。引入SENet通道注意力机制后，碎屑岩与燧石的分类准确率得到进一步提升。该改进还显著提高了长石识别精度，凸显了选择性特征增强对矿物精准鉴别的重要性。最终，序列一致性优化模块的集成有效提升了模型在多变条件下（尤其是长石分类）的稳定性，降低了复杂纹理误判的概率。

模型整体精度的提升验证了新增模块的贡献。每个组件均针对特定挑战设计，最终为矿物颗粒岩石薄片图像的自动分类提供了更精准可靠的解决方案。

3.5.2 多尺度融合中的卷积组合

在消融实验中，我们探究了多尺度融合模块中不同卷积核组合对分类精度的影响。测试了多种组合方案：包括小尺寸核 $(1\times1$ 与 3×3 的组合、中尺寸核 $(3\times3$with$5\times5)$、大尺寸核 $(3\times3$ 与 $9\times9, 5\times5$ 及 9×9 的组合、以及 5×5 与 7×7 的组合，同时对比了我们提出的 3×3 与 7×7 创新组合方案。实验结果汇总于表5。

在这些卷积核组合中，云母分类准确率随核尺寸增大而变化。例如，采用较小核组合（如 1ux1y 与 3ex3s 或 3×3 与 7×7）时，云母分类准确率可达 97.58%。然而当核尺寸增至

表4 消融分析
各模块的

方法	准确率(%) 云母	碎屑	燧石	石英	长石	总体
起源	97.58	91.76	66.03	96.26	83.49	91.072
修正起源	95.76	92.81	75.12	**96.96**	83.09	91.682
+ 通道注意力机制(SENet)	97.58	92.57	**78.95**	95.64	85.2	91.811
+ SCO	**98.18**	**92.89**	77.99	95.95	**86.63**	**92.35**

加粗条目表示最优方法

Table 3 Performance metrics of our method for each category

Category	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
Mica	98.18	97.59	98.18	97.89
Detritus	92.89	92.95	92.89	92.92
Flint	77.99	73.76	77.99	75.81
Quartz	95.95	96.89	95.95	96.41
Feldspar	86.63	85.71	86.63	86.17

3.5 Ablation study

3.5.1 Key modules

We carried out ablation studies to assess the distinct contributions of the various parts of our strategy individually. The 'Origin' model is based on the CMT architecture [14], which combines CNNs and Transformers. We then created the 'Modified Origin' by redesigning the stem and replacing the patch embedding with a Multi-Scale Fusion. As shown in Table 4, further enhancements such as Channel Attention and Sequence Consistency Optimization were added to the 'Modified Origin'. The stepwise integration of these modules consistently improved classification accuracy, validating the effectiveness of each component.

The Origin model's shortcomings in capturing complex traits were brought to light by its remarkable performance in detecting Mica and Quartz but poor performance with Flint. Our modifications, including a custom stem and Multi-Scale Fusion, substantially improved the accuracy in the Flint category, demonstrating the necessity for nuanced multi-scale feature extraction. Adding Channel Attention via SENet further improved the classification accuracy for Detritus and Flint. It resulted in a notable improvement in Feldspar accuracy, emphasizing the importance of selective feature enhancement for precise mineral discrimination. Finally, the integration of Sequence Consistency Optimization led to a notable enhancement in maintaining consistent classification across variable conditions, particularly for Feldspar, reducing the likelihood of misclassifying complex textures.

The overall improvement in the model's accuracy validates the contributions of the added modules. Each component was designed to address distinct challenges, resulting in a more accurate and reliable solution for the automated classification of rock thin section images of mineral particles.

3.5.2 Convolutional combinations in multi-scale fusion

In our ablation study, we explored the impact of different convolutional kernel combinations within the Multi-Scale Fusion module on classification accuracy. Various combinations were tested, including smaller kernels (1×1 with 3×3), medium-sized kernels (3×3 with 5×5), larger kernels (3×3 with 9×9, 5×5 with 9×9, and 5×5 with 7×7), as well as our proposed 3×3 with 7×7. The results are summarized in Table 5.

Among these convolutional kernel combinations, the classification accuracy for Mica gradually decreases as the kernel size increases. For example, when using smaller kernel combinations (such as 1×1 with 3×3 and 3×3 with 7×7), the classification accuracy for Mica can reach 97.58%. However, when the kernel size increases to

Table 4 Ablation analysis of each module.

Method	Accuracy (%) Mica	Detritus	Flint	Quartz	Feldspar	Overall
Origin	97.58	91.76	66.03	96.26	83.49	91.072
Modified Origin	95.76	92.81	75.12	**96.96**	83.09	91.682
+ Channel Attention(SENet)	97.58	92.57	**78.95**	95.64	85.2	91.811
+ SCO	**98.18**	**92.89**	77.99	95.95	**86.63**	**92.35**

Entries in bold indicate the best methods

https://doi.org/10.1007/s10596-025-10356-8

表5 消融分析
多尺度融合中的卷积组合

核组合	准确率(%) 云母	碎屑	燧石	石英	长石	总体
内核1与3	97.58	92.2	72.75	96.96	83.8	91.61
内核3与5	97.58	**92.99**	69.86	96.79	81.1	91.52
内核3与9	96.97	91.94	77.99	96.24	83.96	91.44
内核5与7	**98.79**	92.83	71.77	**97.02**	82.61	91.71
内核5与9	97.58	92.85	76.08	96.68	81.89	91.51
内核3与7（我们的方案）	97.58	92.57	**78.95**	95.64	**85.2**	**91.81**

加粗条目表示最优方法

3×3 与 9×9 结合时，分类准确率降至 96.97% 。该现象表明较大内核更倾向于捕捉全局特征而忽略局部细节，这对云母等简单纹理的分类性能产生负面影响，削弱了模型区分此类别的能力。

与其他核组合相比， 3×3 与 7×7 的组合在长石和燧石的分类任务中表现最优，展现了均衡的特征提取能力。对于长石，该组合的分类准确率达 85.20% ，显著优于 5×5 与 7×7 及 5×5 与 9×9 的组合。这表明 3×3 核能有效捕捉局部细节特征，而 7×7 核提供更广的上下文信息，从而更好区分长石的细微模式。同样对于燧石， 3×3 与 7×7 组合的准确率达 78.95% ，明显高于 3×3 与 5×5 及 1×1 与 3×3 的组合。这说明 3×3 与 7×7 的组合通过保留关键细粒度特征同时提取粗粒度特征，实现了最优平衡，因而获得卓越性能。

这些结果凸显了在模型中为特定岩类选择最优卷积核尺寸的重要性，尤其对于纹理复杂的岩石，表明较大核尺寸能提供更全面的特征捕获能力。因此，我们设计的 3×3 与 7×7 卷积核组合在岩石薄片图像分类中表现卓越，证实了其利用多尺度特征精确分类复杂纹理的优势。

3.5.3 多尺度变换器中的注意力机制

为评估多尺度变换器中不同注意力机制对MSCET模型性能的影响，我们将传统注意力与两种轻量级变体进行对比：一种采用递减核尺寸序列(8, $4, 2, 1$))，另一种采用更大序列(16,8,4,2)。表6详细展示了这些对比结果，证明(8,4,2,1)序列的轻量级注意力机制在多数岩类（尤其是燧石和长石）分类中显著优于传统方法，这得益于其对细节特征和上下文信息更高效的处理能力。

(16,8,4,2)序列相较传统注意力机制有所改进，但效果不及(8,4,2,1)序列，这凸显了其在细节捕捉与上下文理解间的平衡性。此外，我们的方法

表6 多尺度Transformer中注意力的消融分析

注意力类型	准确率(%) 云母	碎屑	燧石	石英	长石	总体	参数量(百万)
注意	95.76	**92.72**	73.68	95.16	80.3	91.24	52.48
轻量注意力机制(k=8,4,2,1)	**97.58**	92.57	**78.95**	95.64	**85.2**	**91.81**	30.2
轻量注意力机制(k=16,8,4,2)	96.36	92.04	75.6	**96.59**	82.49	91.37	**26**

加粗条目表示最优方法

https://doi.org/10.1007/s10596-025-10356-8

Table 5 Ablation analysis of convolutional combinations in Multi-Scale Fusion

Kernel Combination	Accuracy (%) Mica	Detritus	Flint	Quartz	Feldspar	Overall
Kernel 1 and 3	97.58	92.2	72.75	96.96	83.8	91.61
Kernel 3 and 5	97.58	**92.99**	69.86	96.79	81.1	91.52
Kernel 3 and 9	96.97	91.94	77.99	96.24	83.96	91.44
Kernel 5 and 7	**98.79**	92.83	71.77	**97.02**	82.61	91.71
Kernel 5 and 9	97.58	92.85	76.08	96.68	81.89	91.51
Kernel 3 and 7 (Ours)	97.58	92.57	**78.95**	95.64	**85.2**	**91.81**

Entries in bold indicate the best methods

3×3 with 9×9, the classification accuracy drops to 96.97%. This phenomenon indicates that larger kernels tend to focus more on capturing global features while overlooking local details, which negatively impacts the classification performance for simpler textures like Mica, weakening the model's ability to distinguish such categories.

Compared to other kernel combinations, 3×3 with 7×7 achieves the best performance in the classification tasks for Feldspar and Flint, demonstrating a well-balanced feature extraction capability. For Feldspar, the classification accuracy of this combination reaches 85.20%, significantly outperforming 5×5 with 7×7 and 5×5 with 9×9 combinations. This suggests that the 3×3 kernel effectively captures local detail features, while the 7×7 kernel provides broader contextual information, enabling better differentiation of subtle patterns in Feldspar. Similarly, for Flint, the classification accuracy of 3×3 with 7×7 reaches 78.95%, which is notably higher than 3×3 with 5×5 and 1×1 with 3×3. This indicates that the 3×3 with 7×7 combination strikes an optimal balance by preserving critical fine-grained features while simultaneously extracting coarse-grained features, thereby achieving superior performance.

These results underscore the importance of selecting optimal kernel sizes for specific categories in our model, particularly for rocks with complex textures, demonstrating that larger kernels can provide a more comprehensive feature capture capability. Thus, our 3×3 and 7×7 convolutional kernel combination stands out as highly effective for classifying rock thin section images, affirming its strength in leveraging multi-scale features for accurate categorization of complex textures.

3.5.3 Attention in multi-scale transformer

To assess the impact of various attention mechanisms within the Multi-Scale Transformer on the MSCET model's performance, we compare traditional attention with two lightweight variants: one with decreasing kernel sizes (8, 4, 2, 1) and another with a larger sequence (16, 8, 4, 2). Table 6 details these comparisons and shows that the lightweight attention mechanism with the (8, 4, 2, 1) sequence significantly outperforms the traditional approach in most categories, particularly for Flint and Feldspar, due to its more effective processing of detailed and contextual features.

The (16, 8, 4, 2) sequence also improved over traditional attention but was less effective than the (8, 4, 2, 1) sequence, highlighting its balance between detail capture and contextual understanding. Additionally, our approach

Table 6 Ablation analysis of attention in Multi-Scale Transformer

Attention Type	Accuracy (%) Mica	Detritus	Flint	Quartz	Feldspar	Overall	Params(M)
Attn	95.76	**92.72**	73.68	95.16	80.3	91.24	52.48
LightAttn(k=8,4,2,1)	**97.58**	92.57	**78.95**	95.64	**85.2**	**91.81**	30.2
LightAttn(k=16,8,4,2)	96.36	92.04	75.6	**96.59**	82.49	91.37	**26**

Entries in bold indicate the best methods

参数更少的设计提升了计算效率，使其适用于需要快速响应的实际应用场景。

总体而言，采用(8,4,2,1)序列的轻量级注意力机制在岩石薄片图像分类中表现出色，在资源高效的框架下实现了细节特征提取与宏观上下文洞察的完美结合。

3.5.4 δ 在序列一致性优化中的影响

在序列一致性优化方法中，阈值 δ 是关键参数，它控制模型确保同一颗粒在不同角度下预测类别一致性的能力。本文研究了不同 δ 值对分类性能的影响，表7展示了不同 δ 设置下的分类性能。

总体而言，较高的 δ 值（如0.9或0.85）能提升整体准确率，但会显著降低燧石类别的分类精度。反之，较低的 δ 值（如0.55）可能削弱预测一致性，导致某些类别的分类性能波动。跨类别分析表明，当 $\delta = 0.6$ 时，模型性能最为均衡——该阈值在保持云母、石英等简单类别高精度的同时，有效减少了燧石类别的误判。

相较于其他阈值，$\delta = 0.6$ 在特定类别性能与模型稳定性间取得了更佳平衡。当 δ 设为0.6时，燧石分类结果趋于稳定，且其他类别的准确率未受优化过程影响。因此，本文最终选定0.6作为序列一致性优化方法的 δ 阈值。

3.6 序列一致性优化结果分析

为利用交叉偏振图像中颗粒特性的变化，我们采用序列一致性优化方法，显著提升了MSCET模型的分类精度。图5展示了从 0° 到 72° 的四组交叉偏振图像，表8则详细列出了对应的预测倾向清晰度(CPI)、MSCET模型的预测结果及序列一致性优化后的效果。

交叉偏振图像序列在MSCET模型中的预测清晰度因亮度和纹理变化呈现显著波动。例如案例1中，18° 与 54° 位置的图像显示出类似碎屑岩的暗色特征，其CPI值超过0.8；而 $0^\circ, 36^\circ$ 和 72° 位置的图像亮度较高，导致预测稳定性下降。尤其在 72° 位置，高亮度使CPI值显著降低，造成石英误判，后通过修正为碎屑岩分类与暗角观测结果保持一致。

表7 分类结果
不同 δ 值的性能表现

δ 数值	准确率(%)					
	云母	碎屑岩	燧石	石英	长石	总体
0.55	98.18	92.77	77.99	95.97	86.19	92.22
0.6	98.18	92.89	77.99	95.95	86.63	92.35
0.65	98.18	93.02	76.08	96.18	86.39	92.40
0.7	98.18	93.10	75.12	96.18	86.55	92.45
0.75	98.79	93.18	74.16	96.34	86.95	92.61
0.8	98.79	93.30	73.68	96.39	86.95	92.67
0.85	98.49	93.25	75.31	96.65	86.77	92.72
0.9	98.79	93.34	73.68	96.47	86.79	92.68

with fewer parameters enhances computational efficiency, making it suitable for practical applications requiring rapid response.

Overall, the lightweight attention mechanism with the (8, 4, 2, 1) sequence proved highly effective for classifying rock thin section images, combining detailed feature extraction with broad contextual insight in a resource-efficient framework.

3.5.4 Impact of δ in sequence consistency optimization

In the Sequence Consistency Optimization method, the threshold δ is a crucial parameter that governs the model's ability to ensure consistent prediction categories for the same particle at different angles. This paper examines the effect of varying δ values on classification performance. Table 7 presents the classification performance for different δ settings.

Overall, higher δ values, such as 0.9 or 0.85, achieved better overall accuracy but significantly reduced the accuracy for the Flint category. On the other hand, lower δ values (such as 0.55) may reduce prediction consistency, causing fluctuations in the classification performance of certain categories. Analyzing classification performance across categories reveals that when $\delta = 0.6$, the performance is well-balanced. This threshold reduced misclassifications in the Flint category while maintaining high classification accuracy for simpler categories like Mica and Quartz.

Compared with other thresholds, $\delta = 0.6$ strikes a better balance between category-specific performance and model stability. With δ set at 0.6, the classification results for Flint became more stable, and the accuracy of other categories remained unaffected by the optimization process. Consequently, this paper selects 0.6 as the threshold δ for the Sequence Consistency Optimization method.

3.6 Analysis of sequence consistency optimization results

To harness variations in particle characteristics in cross-polarized images, we implemented the Sequence Consistency Optimization method, significantly enhancing the MSCET model's classification accuracy. Figure 5 displays four sets of cross-polarized images spanning from 0° to 72°, while Table 8 details the corresponding Clarity of Predictive Inclination (CPI), predictions by the MSCET model, and outcomes following the application of Sequence Consistency Optimization.

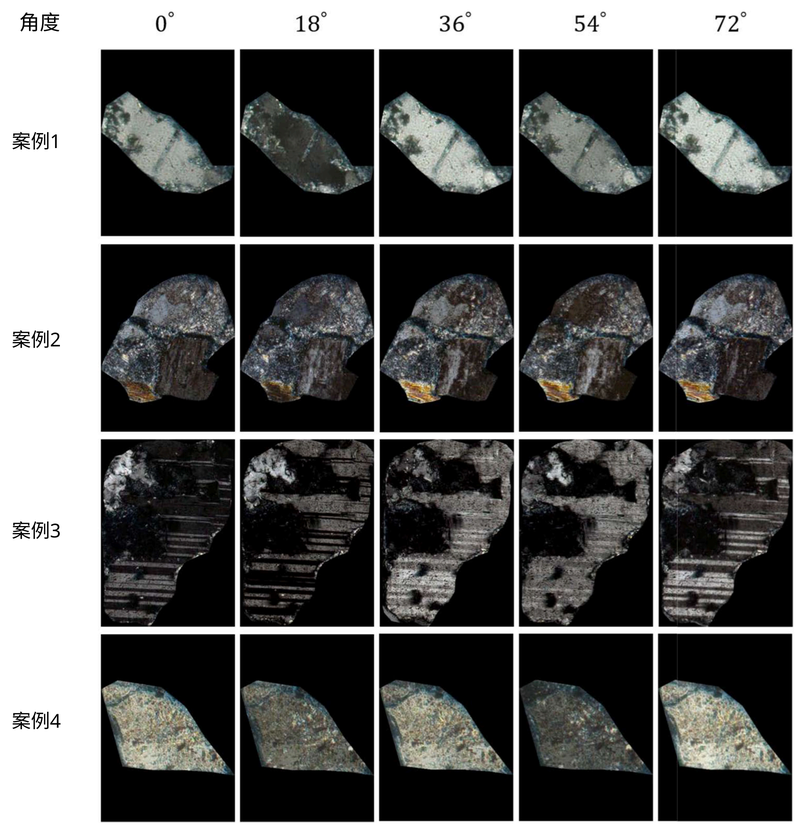
The sequence of cross-polarized images exhibits notable fluctuations in the MSCET model's predictive clarity due to changes in brightness and texture. For instance, in Case 1, images at 18° and 54° displayed darker features similar to Detritus with a CPI exceeding 0.8. Conversely, images at 0°, 36°, and 72° were brighter, resulting in less stable predictions. Particularly, at 72°, the high brightness significantly lowered the CPI, leading to misclassification as Quartz, which was corrected by revising the classification to Detritus, aligning it with the consistent Detritus observations at darker angles.

Table 7 Classification Performance for Different δ Values

Value of δ	Accuracy (%)					
	Mica	Detritus	Flint	Quartz	Feldspar	Overall
0.55	98.18	92.77	77.99	95.97	86.19	92.22
0.6	98.18	92.89	77.99	95.95	86.63	92.35
0.65	98.18	93.02	76.08	96.18	86.39	92.40
0.7	98.18	93.10	75.12	96.18	86.55	92.45
0.75	98.79	93.18	74.16	96.34	86.95	92.61
0.8	98.79	93.30	73.68	96.39	86.95	92.67
0.85	98.49	93.25	75.31	96.65	86.77	92.72
0.9	98.79	93.34	73.68	96.47	86.79	92.68

https://doi.org/10.1007/s10596-025-10356-8

图5 四种示例
不同角度的交叉偏振图像组



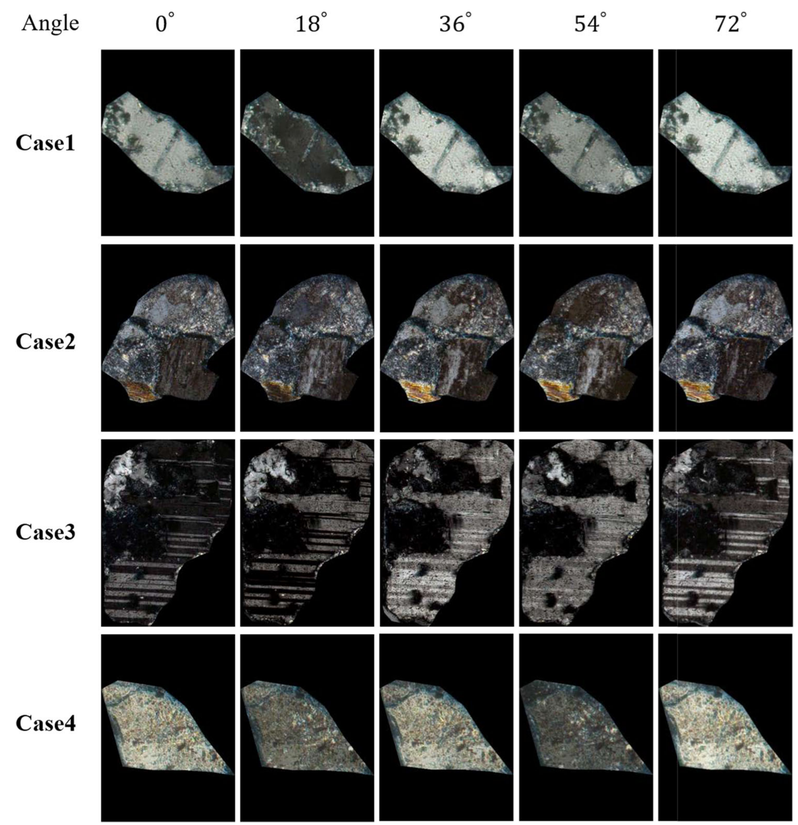
在案例2中，虽然 0° 至 72° 的交叉偏振图像均显示为碎屑岩，但 18° 处观察到的特征呈现类似长石的性质，导致获得0.5324的CPI指数并错误预测为长石。该现象源于此颗粒中长石晶体在该角度的纹理特别清晰，而与其他区域的对比度降低。这种增强效应使得该颗粒的长石特征相较其他角度更为明显。通过采用置信度更高的 36° 图像预测（其CPI为0.9009），将 18° 的预测结果修正为碎屑岩，与多数角度的观测结果保持一致。

案例3中， 54° 图像因纹理特征暗淡模糊，导致CPI值降至0.5412，被误判为碎屑岩。这一误差可归因于该角度下的颗粒未明显呈现交错的长石特征，反而间接强化了碎屑岩特有的明暗交替特征。然而，CPI值为0.8911的 36° 图像提供了可靠参照。通过采用SCO方法优先采纳 36° 图像中置信度更高的预测结果， 54° 图像最终被正确识别为长石，验证了该优化策略的鲁棒性。

最后，在案例4中， 54° 图像因颜色较暗且纹理清晰度降低，其CPI值为0.5770，被误判为碎屑。该角度下颗粒整体呈现暗色特征，凸显了碎屑典型的明暗交替特性。通过SCO优化，基于<b1></b0>图像（CPI值0.8993）的可靠分类结果对 54° 的预测进行了调整。这一修正不仅纠正了错误，更验证了优化方法在不同角度下的可靠性。

https://doi.org/10.1007/s10596-025-10356-8

Fig. 5 Examples of four sets of cross-polarized images at different angles



In Case 2, while cross-polarized images from 0° to 72° all depicted Detritus, the features observed at 18° exhibited characteristics resembling Feldspar, leading to a CPI of 0.5324 and an erroneous prediction of Feldspar. This phenomenon occurred because the texture of the feldspar crystals in this particle is particularly distinct at this angle, and the contrast with other areas is diminished. This enhancement makes the feldspar characteristics of this particle more apparent compared to other angles. By leveraging the higher-confidence prediction from the 36° image, which has a CPI of 0.9009, the prediction for 18° was revised to Detritus, aligning with the majority of the observations across other angles.

Case 3 presented a scenario where the 54° image, due to its dim and less distinct texture features, caused the CPI to drop to 0.5412, leading to a misclassification as Detritus. This error can be attributed to the particle at this angle not clearly exhibiting the interlaced feldspar characteristics, which indirectly emphasized the alternating light and dark features typical of Detritus. However, the 36° image, with a CPI of 0.8911, served as a reliable reference. By utilizing the SCO method to prioritize the more confident prediction from 36° image, the 54° image was correctly identified as Feldspar, showcasing the robustness of the optimization strategy.

Finally, in Case 4, the 54° image, characterized by its dark coloration and reduced texture clarity, resulted in a CPI of 0.5770 and was misclassified as Detritus. At this angle, the overall characteristics of the particle appear darker, highlighting the alternating light and dark features typical of detritus. Using SCO, the prediction for 54° was adjusted based on the confident classification at 0° image, which has a CPI of 0.8993. This adjustment not only corrected the error but also demonstrated the reliability of the optimization approach across varying angles.

表8 不同角度下的优化结果

案例	角度	CPI	MSCET	优化
案例1	0°	0.7880	碎屑	碎屑
	18°	0.8920	碎屑	碎屑
	36°	0.7900	碎屑	碎屑
	54°	0.8260	碎屑	碎屑
	72°	**0.5217**	石英	碎屑
案例2	0°	0.7619	碎屑	碎屑
	18°	**0.5324**	长石	碎屑
	36°	0.9009	碎屑	碎屑
	54°	0.7874	碎屑	碎屑
	72°	0.7828	碎屑	碎屑
案例3	0°	0.8004	长石	长石
	18°	0.8413	长石	长石
	36°	0.8911	长石	长石
	54°	**0.5412**	碎屑	长石
	72°	0.8569	长石	长石
案例4	0°	0.8993	长石	长石
	18°	0.7954	长石	长石
	36°	0.8842	长石	长石
	54°	**0.5770**	碎屑	长石
	72°	0.8739	长石	长石

加粗条目表示最优方法

这些发现凸显了序列一致性优化(SCO)方法在提升MSCET模型准确性与可靠性中的关键作用。具体而言，在观测案例中，某些角度可能对特定类别的分类更为重要。例如， 18° 和 54° 角度的颗粒图像能更有效地保留碎屑特征，而 72° 角度可能更有利于识别石英。通过利用CPI和跨角度一致性，SCO有效缓解了由亮度和纹理变化引起的分类误差，从而显著提升了MSCET模型在复杂地质数据集中的适应性和分类性能。

3.7 SENet分析

我们通过可视化特征图变换分析了SENet对MSCET模型的影响，旨在评估其在突出相关特征方面的有效性。图6展示了特征表示通过SENet的转变过程：首行为原始输入图像，第二行显示应用SENet前生成的特征图，第三行则聚焦经SENet处理后各输入最高权重通道的特征图变化。

SENet使模型能评估并调整各特征通道的重要性，从而提升任务特定性能。图6(c)清晰呈现了最高权重通道在SENet应用后的显著激活变化，各特征图上数值标明了其重要性。这种前后对比揭示了网络通过特征重校准机制，有效增强了图像纹理细节的表征能力。

将SENet整合到MSCET中增强了其捕获关键图像信息的能力，凸显了通道注意力机制对特征表征的优化作用。总体而言，SENet在MSCET中的集成被证明是提升通道注意力的有效方法，显著提高了岩石薄片图像分类精度。

Table 8 Optimization results across different angles

Case	Angle	CPI	MSCET	Optimize
Case 1	0°	0.7880	Detritus	Detritus
	18°	0.8920	Detritus	Detritus
	36°	0.7900	Detritus	Detritus
	54°	0.8260	Detritus	Detritus
	72°	**0.5217**	**Quartz**	**Detritus**
Case 2	0°	0.7619	Detritus	Detritus
	18°	**0.5324**	**Feldspar**	**Detritus**
	36°	0.9009	Detritus	Detritus
	54°	0.7874	Detritus	Detritus
	72°	0.7828	Detritus	Detritus
Case 3	0°	0.8004	Feldspar	Feldspar
	18°	0.8413	Feldspar	Feldspar
	36°	0.8911	Feldspar	Feldspar
	54°	**0.5412**	**Detritus**	**Feldspar**
	72°	0.8569	Feldspar	Feldspar
Case 4	0°	0.8993	Feldspar	Feldspar
	18°	0.7954	Feldspar	Feldspar
	36°	0.8842	Feldspar	Feldspar
	54°	**0.5770**	**Detritus**	**Feldspar**
	72°	0.8739	Feldspar	Feldspar

Entries in bold indicate the best methods

These findings highlight the critical role of the Sequence Consistency Optimization (SCO) method in enhancing the accuracy and reliability of the MSCET model. Specifically, in the observed cases, certain angles may be more important for the classification of specific categories. For example, particle images at 18° and 54° are more effective in preserving Detritus features, while those at 72° may be more advantageous for identifying Quartz. By leveraging CPI and cross-angle consistency, SCO effectively mitigates classification errors caused by brightness and textural variations, thereby significantly improving the adaptability and classification performance of the MSCET model in complex geological datasets.

3.7 Analysis of SENet

We analyzed SENet's impact on the MSCET model by visualizing feature map transformations, aiming to assess its effectiveness in accentuating pertinent features. Figure 6 illustrates the transformation of feature representation through the SENet. The first row displays the original input images. The second row shows the feature maps generated before the application of the SENet. The third row highlights the feature maps after the SENet processing, specifically focusing on the single highest-weighted channel for each input.

SENet enables the model to assess and adjust the significance of each feature channel, enhancing task-specific performance. This is particularly evident in Fig. 6(c), where channels with the highest weights show notable activation changes after the SENet application, with their importance indicated numerically on each map. The transition from before to after the application of SENet demonstrates the network's impact on feature recalibration, enhancing the representation of textural details in the images.

Incorporating SENet into MSCET enriches its ability to capture and utilize key image information, which underscores the channel attention mechanism's role in enhancing feature representation. Overall, SENet's integration within MSCET emerges as an effective approach for channel attention enhancement, markedly advancing rock thin section image classification precision.

 https://doi.org/10.1007/s10596-025-10356-8

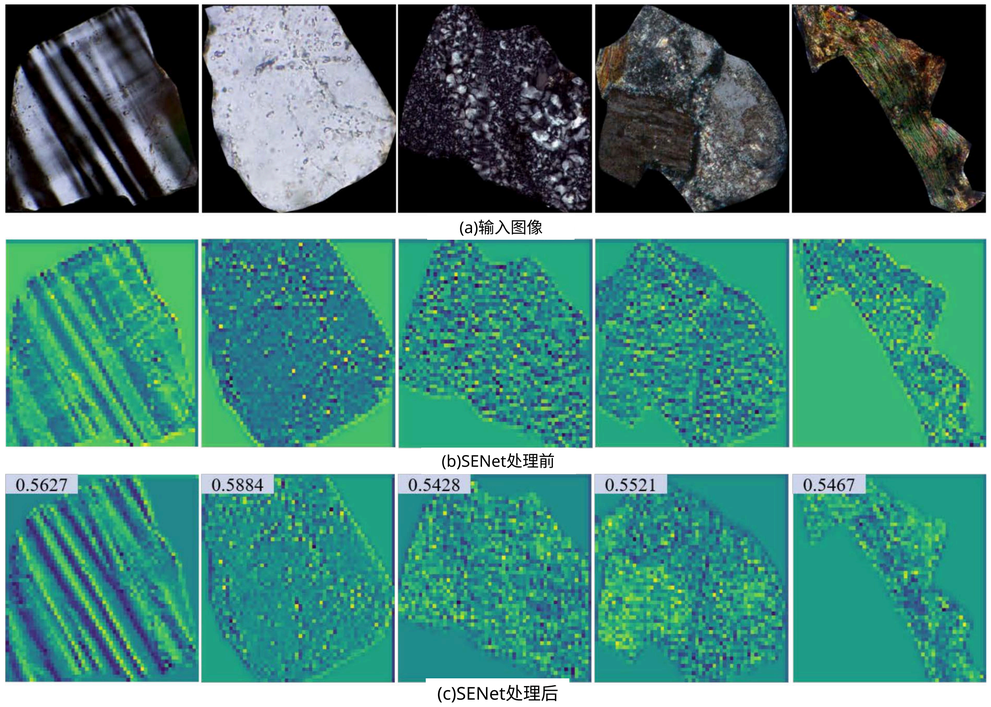


图6 SENet应用前后特征图可视化

3.8 数据不平衡分析

如表1所示，数据集的类别数量存在显著不平衡。为降低模型分类性能向大样本类别偏倚的风险，我们采用了数据增强和重新加权技术。

数据增强：如表1所示，Flint和Mica类别的样本量显著少于其他类别。为改善这种不平衡，我们采用四种数据增强技术（调整图像亮度、对比度、旋转和饱和度[20]）对这两个类别的训练集进行了扩充。扩展后的训练集与验证集分布如表9所示。

权重重置：针对数据集中类别数量的不平衡问题，我们尝试通过权重重置技术进行改善。具体采用Focal Loss函数，根据各类别样本量分配权重，使模型更关注样本量较少的类别，从而有效缓解数据不平衡的影响。

表9 数据增强后的类别分布

矿物类型	训练数据	测试数据
碎屑	19715	4925
石英	15313	3826
长石	10067	2513
燧石	4195	209
云母	3330	165

 https://doi.org/10.1007/s10596-025-10356-8

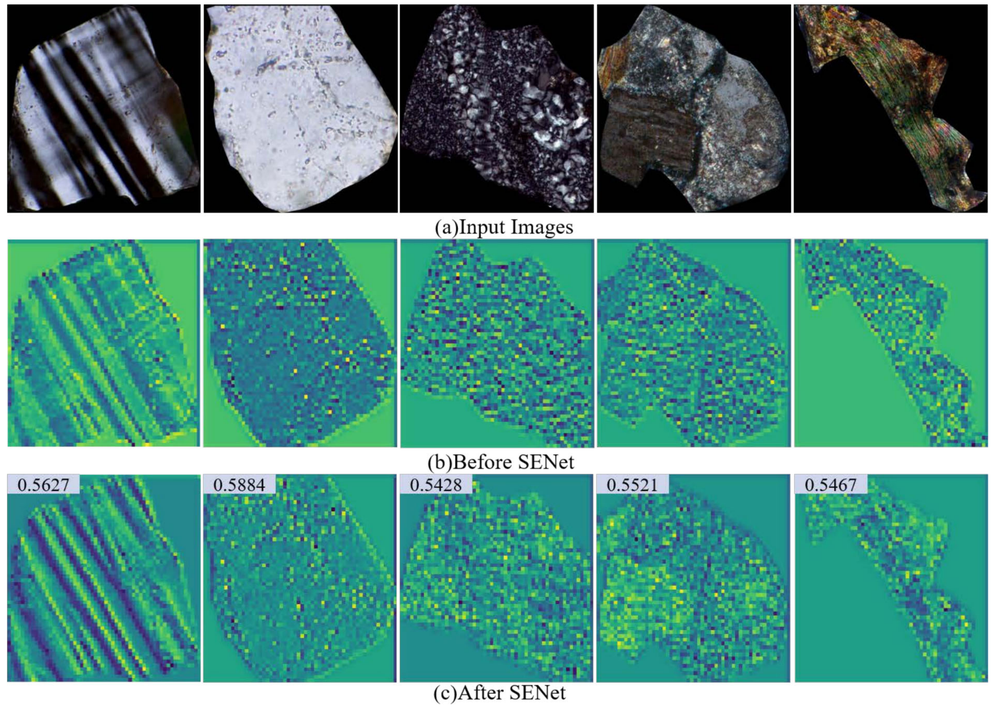


Fig. 6 Visualization of feature map before and after SENet application

3.8 Analysis of data imbalance

The dataset exhibits a significant imbalance in the number of categories, as illustrated in Table 1. To reduce the risk of the model's classification performance being biased toward categories with larger sample sizes, we employed data augmentation and reweighting techniques.

Data Augmentation: As shown in Table 1, the sample sizes of the Flint and Mica categories are significantly smaller than those of other categories. To improve this imbalance, we expanded the training set for these two categories using four data augmentation techniques: adjusting image brightness, contrast, rotation, and saturation [20]. The distribution of the training set and validation set after expansion is shown in Table 9.

Reweighting: Considering the imbalance in the number of categories in the dataset, we try to improve this situation through reweighting technology. Specifically, we employed the Focal Loss function, assigning weights based on the sample size of each category. This approach allows the model to focus more on categories with fewer samples, thereby effectively mitigating the impact of data imbalance.

Table 9 Category distribution after data augmentation

Mineral type	Training data	Testing data
Detritus	19715	4925
Quartz	15313	3826
Feldspar	10067	2513
Flint	4195	209
Mica	3330	165

表10 性能指标采用数据平衡技术的MSCET模型

优化方法	准确率(%) 云母	碎屑	燧石	石英	长石	总体
–	97.58	92.57	78.95	95.64	85.2	91.81
数据增强	96.97	92.87	78.95	96.52	80.78	91.27
权重调整	100.00	79.51	90.91	95.66	83.13	86.10

我们分别采用数据增强和重加权技术训练MSCET模型。实验结果如表10所示，其中"-"表示未使用任何平衡策略训练的MSCET模型。

实验结果表明，数据增强虽然提升了石英类别的分类性能，却导致云母和长石类别的准确率下降，从而降低了整体分类效果。这种下降可能源于数据增强改变了原始数据分布，未能有效契合模型的训练需求。相比之下，重加权技术显著改善了燧石、云母等小样本类别的分类性能，但由于过度关注小样本类别，导致碎屑类别的分类准确率下降，最终造成整体分类性能降低。

尽管数据增强和重加权技术能在一定程度上缓解数据不平衡问题，但二者均存在局限性。数据增强可能通过改变类别分布引入偏差，从而对某些类别的分类性能产生负面影响；而重加权策略可能过度强调样本量较小的类别，导致大样本类别分类性能下降，最终削弱整体表现。因此基于实验结果，本研究未采用数据增强或重加权技术作为优化方法。

4 结论

本文提出了一种结合多尺度通道增强变换器(MSCET)与序列一致性优化(SCO)策略的岩石薄片分类新方法。MSCET模型综合CNN、SENet和Transformer架构的优势，能有效提取并处理多尺度特征，提升不同粒度下的特征识别能力；同时SCO策略通过高置信度数据调整低置信度结果，有效解决了不同角度偏光图像带来的分类挑战。该方法显著提高了岩石薄片图像的分类可靠性与准确性，为地质勘探提供了更高效精准的技术支持。

本文聚焦于岩石薄片图像分类方法。在数据不平衡分析中，我们仅尝试了基础数据增强技术。未来研究可探索更先进的数据增强方法以缓解数据不平衡影响，进一步提升分类性能。在SCO策略中，后续工作可考虑为不同矿物颗粒类别设置差异化阈值 δ，以适应其独特纹理和复杂特性，从而增强分类效果与模型鲁棒性。

致谢 作者感谢四川省智能油气勘探开发工程研究中心及西南石油大学高性能计算平台的支持。

作者贡献 郭晓尧：提出原创思路，编写代码与论文；陈岩：获取输入数据，验证结果；何世鹏：论文修订；张星鹏：论文评议与指导；周静：数据分析；鲍旭成：数据整理与可视化。

Table 10 Performance of MSCET Model with Data Balancing Techniques

Optimization method	Accuracy (%) Mica	Detritus	Flint	Quartz	Feldspar	Overall
–	97.58	92.57	78.95	95.64	85.2	91.81
Data augmentation	96.97	92.87	78.95	96.52	80.78	91.27
Reweighting	100.00	79.51	90.91	95.66	83.13	86.10

We trained the MSCET model using data augmentation and reweighting techniques respectively. The experimental results are presented in Table 10, where "–" indicates the MSCET model trained without any balancing strategies.

Experimental results indicate that while data augmentation enhanced the classification performance of the Quartz category, it resulted in a decrease in accuracy for the Mica and Feldspar categories, thereby diminishing the overall classification performance. This decline may be attributed to the fact that data augmentation alters the original data distribution, which fails to align effectively with the training requirements of the model. In contrast, the reweighting technique significantly improves the classification performance of categories with fewer samples, such as Flint and Mica. However, due to its excessive focus on smaller sample categories, the classification accuracy of the Detritus category decreased, ultimately leading to a reduction in overall classification performance.

Although data augmentation and reweighting techniques can alleviate the problem of data imbalance to a certain extent, both have limitations. Data augmentation may introduce biases by altering the class distribution, thereby negatively affecting the classification performance of some categories. On the other hand, reweighting strategies may overly emphasize categories with smaller sample sizes, leading to a decline in the classification performance of categories with larger sample sizes, ultimately weakening overall performance. Therefore, based on the experimental results, this study did not adopt data augmentation or reweighting techniques as optimization methods.

4 Conclusion

In this paper, we propose a novel classification method for rock thin sections by combining the Multi-Scale Channel Enhanced Transformer (MSCET) with the Sequence Consistency Optimization (SCO) strategy. Leveraging the integrated capabilities of CNN, SENet, and Transformer architectures, the MSCET model adeptly extracts and processes multi-scale features, enhancing feature recognition across varying granularities. Concurrently, the SCO strategy addresses the challenges posed by cross-polarized images at different angles, refining prediction accuracy by adjusting low-confidence results based on high-confidence data. This approach significantly improves the classification reliability and accuracy of rock thin section images, facilitating more effective and precise geological explorations.

The focus of this paper is on the classification method for rock thin section images. In the analysis of data imbalance, we only experimented with basic data augmentation techniques. Future research may explore more advanced data augmentation methods to mitigate the effects of data imbalance and further improve the classification performance. In the SCO strategy, future work could explore setting different thresholds δ for different categories to adapt to the unique textures and complexities of mineral particles, thereby enhancing classification performance and model robustness.

Acknowledgements The authors acknowledge the support from the Engineering Research Center for Intelligent Oil & Gas Exploration and Development of Sichuan Province and the High-Performance Computing platform of Southwest Petroleum University.

Author Contributions Xiaoyao Guo: Original idea, Wrote the code and the manuscript. Yan Chen: obtained the input data, Verified the results. Shipeng He: Edited the manuscript. Xingpeng Zhang: Commented on the manuscript, Supervision. Jing Zhou: Data analysis. Xucheng Bao: Data curation, Visualization.

基金资助 本研究得到国家自然科学基金（编号62441610）、四川省科技厅重点研发项目（编号2023YFG0129）、西南石油大学自然科学启动项目（编号2022QHZ023）的资助。

数据可用性 本研究所用数据由合作方提供。根据保密协议，原始数据集暂不公开，但论文中详细描述了数据构成与预处理步骤以确保透明度。

代码可用性 源代码下载地址：https://github.com/swpugxy/MSCET.git

声明

利益冲突 作者声明无任何相关财务或非财务利益冲突。

伦理批准与参与同意 不适用

不适用发表同意书

参考文献

1. Pereira Borges, H., Aguiar, M.S.: Mineral classification using machine learning and images of microscopic rock thin section. In: Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence, MICAI 2019, Xalapa, Mexico, October 27–November 2, 2019, Proceedings 18, pp. 63–76 (2019)
2. Lepistö, L., Kunttu, I., Visa, A.: Rock image classification based on k-nearest neighbour voting. IEE Proc.-Vis. Image Signal Process. **153**(4), 475–482 (2006)
3. Zhang, J., Li, J., Hu, Y., Zhou, J.Y.: The identification method of igneous rock lithology based on data mining technology. Adv. Mater. Res. **466**, 65–69 (2012)
4. Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., Ding, X.: Application of deep learning algorithms in geotechnical engineering: A short critical review. Artif. Intell. Rev. 1–41 (2021)
5. Cheng, G., Guo, W.: Rock images classification by using deep convolution neural network. In: Journal of Physics: Conference Series, vol. 887, p. 012089 (2017)
6. Zhang, Y., Li, M., Han, S.: Automatic identification and classification in lithology based on deep learning in rock images. Yanshi Xuebao/Acta Petrol. Sin. **34**(2), 333–342 (2018)
7. Xu, Y., Dai, Z., Luo, Y.: Research on application of image enhancement technology in automatic recognition of rock thin section. In: IOP Conference Series: Earth and Environmental Science, vol. 605, p. 012024 (2020)
8. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv:2103.11886 (2021)
9. Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv:2201.04676 (2022)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
13. Huang, H., Zhou, X., Cao, J., He, R., Tan, T.: Vision transformer with super token sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22690–22699 (2023)
14. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

19 Page 18 of 19 https://doi.org/10.1007/s10596-025-10356-8

Funding This work was supported by the National Natural Science Foundation of China (No. 62441610), Key Research and Development Project of Sichuan Provincial Department of Science and Technology (No.2023YFG0129), Natural Science Starting Project of SWPU (No.2022QHZ023).

Data Availability The data used in this study was provided by our collaborators. Due to confidentiality agreements, the dataset itself cannot be made publicly applicable. However, detailed descriptions of the dataset's composition and preprocessing steps are provided in the manuscript to ensure transparency.

Code Availability The source codes are available for downloading at the link: https://github.com/swpugxy/MSCET.git.

Declarations

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval and Consent to Participate Not applicable

Consent for Publication Not applicable

References

1. Pereira Borges, H., Aguiar, M.S.: Mineral classification using machine learning and images of microscopic rock thin section. In: Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence, MICAI 2019, Xalapa, Mexico, October 27–November 2, 2019, Proceedings 18, pp. 63–76 (2019)
2. Lepistö, L., Kunttu, I., Visa, A.: Rock image classification based on k-nearest neighbour voting. IEE Proc.-Vis. Image Signal Process. **153**(4), 475–482 (2006)
3. Zhang, J., Li, J., Hu, Y., Zhou, J.Y.: The identification method of igneous rock lithology based on data mining technology. Adv. Mater. Res. **466**, 65–69 (2012)
4. Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., Ding, X.: Application of deep learning algorithms in geotechnical engineering: A short critical review. Artif. Intell. Rev. 1–41 (2021)
5. Cheng, G., Guo, W.: Rock images classification by using deep convolution neural network. In: Journal of Physics: Conference Series, vol. 887, p. 012089 (2017)
6. Zhang, Y., Li, M., Han, S.: Automatic identification and classification in lithology based on deep learning in rock images. Yanshi Xuebao/Acta Petrol. Sin. **34**(2), 333–342 (2018)
7. Xu, Y., Dai, Z., Luo, Y.: Research on application of image enhancement technology in automatic recognition of rock thin section. In: IOP Conference Series: Earth and Environmental Science, vol. 605, p. 012024 (2020)
8. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv:2103.11886 (2021)
9. Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv:2201.04676 (2022)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
13. Huang, H., Zhou, X., Cao, J., He, R., Tan, T.: Vision transformer with super token sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22690–22699 (2023)
14. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

Springer Comput Geosci (2025) 29:19

18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
19. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
20. Ma, H., Han, G., Peng, L., Zhu, L., Shu, J.: Rock thin sections identification based on improved squeeze-and-excitation networks model. Comput. Geosci. **152**, 104780 (2021)

作者及从属机构

郭晓尧 [1,2]. 陈岩 [1,2]. 何世鹏 [3]. 张兴鹏 [1,2]. 周静 [1]. 鲍旭成 [1,2]

C 陈岩
carly_chen@126.com
郭晓尧
xiaoyao1206@foxmail.com
何世鹏
heshipeng@petrochina.com.cn
张星鹏
xpzhang@swpu.edu.cn
周静
shenwangxiaobai@163.com
鲍旭成
xucheng.bao@foxmail.com

西南石油大学计算机科学与软件工程学院，成都 610500，中国
四川省智能油气勘探开发工程研究中心，成都610500，中国

中国石油西南油气田公司川西北气矿，绵阳 621700

https://doi.org/10.1007/s10596-025-10356-8 Page 19 of 19 19

18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
19. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
20. Ma, H., Han, G., Peng, L., Zhu, L., Shu, J.: Rock thin sections identification based on improved squeeze-and-excitation networks model. Comput. Geosci. **152**, 104780 (2021)

Authors and Affiliations

Xiaoyao Guo[1,2] · Yan Chen[1,2] · Shipeng He[3] · Xingpeng Zhang[1,2] · Jing Zhou[1] · Xucheng Bao[1,2]

✉ Yan Chen
carly_chen@126.com

Xiaoyao Guo
xiaoyao1206@foxmail.com

Shipeng He
heshipeng@petrochina.com.cn

Xingpeng Zhang
xpzhang@swpu.edu.cn

Jing Zhou
shenwangxiaobai@163.com

Xucheng Bao
xucheng.bao@foxmail.com

1 School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China

2 Engineering Research Center for Intelligent Oil & Gas Exploration and Development of Sichuan Province, Chengdu 610500, China

3 Northwest Sichuan Gas District of Southwest Oil and Gasfield Company, Mianyang 621700, China