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.

Published online: 24 April 2025

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.

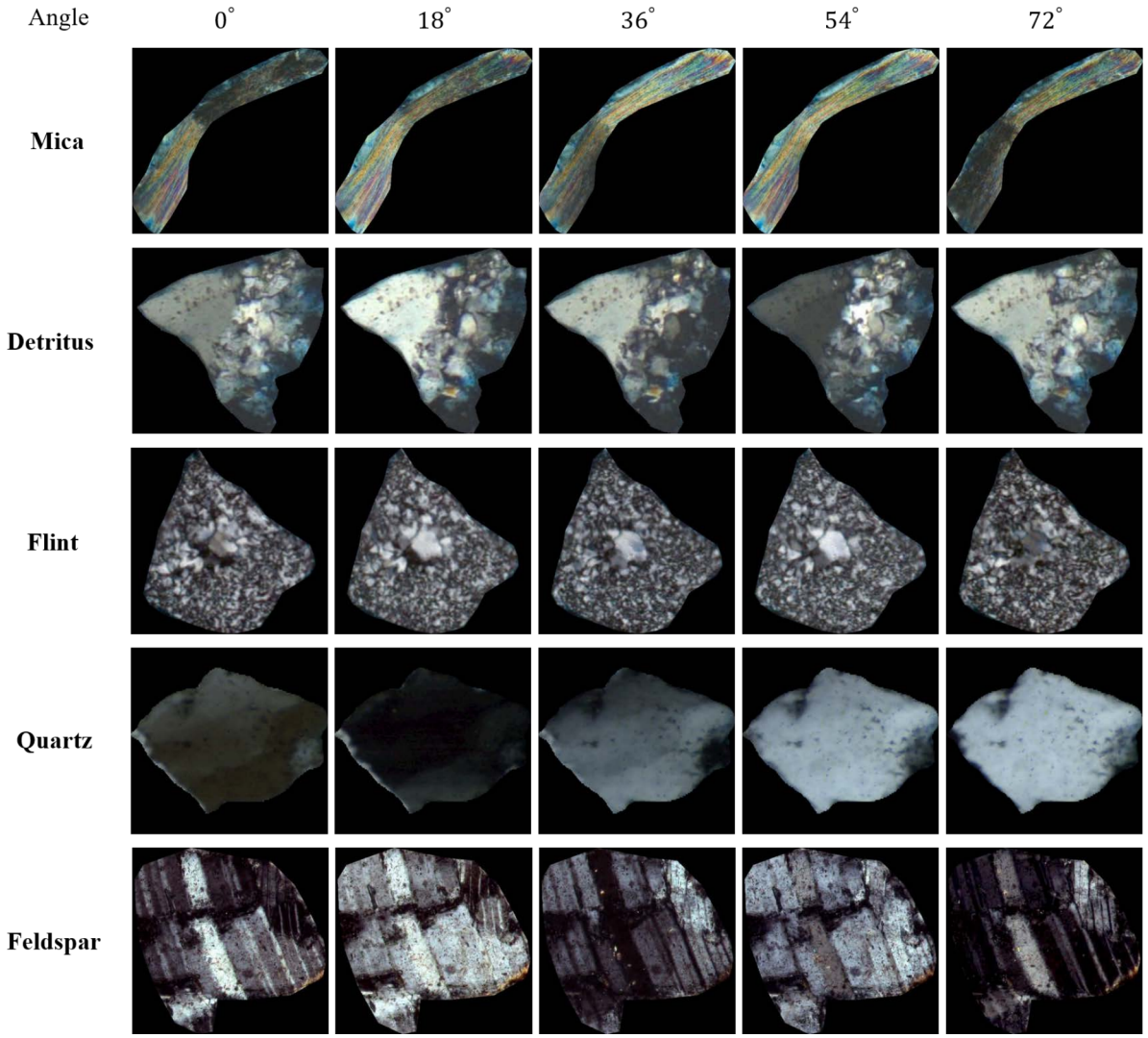

Fig. 1 Rock thin section images of five mineral particle types

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

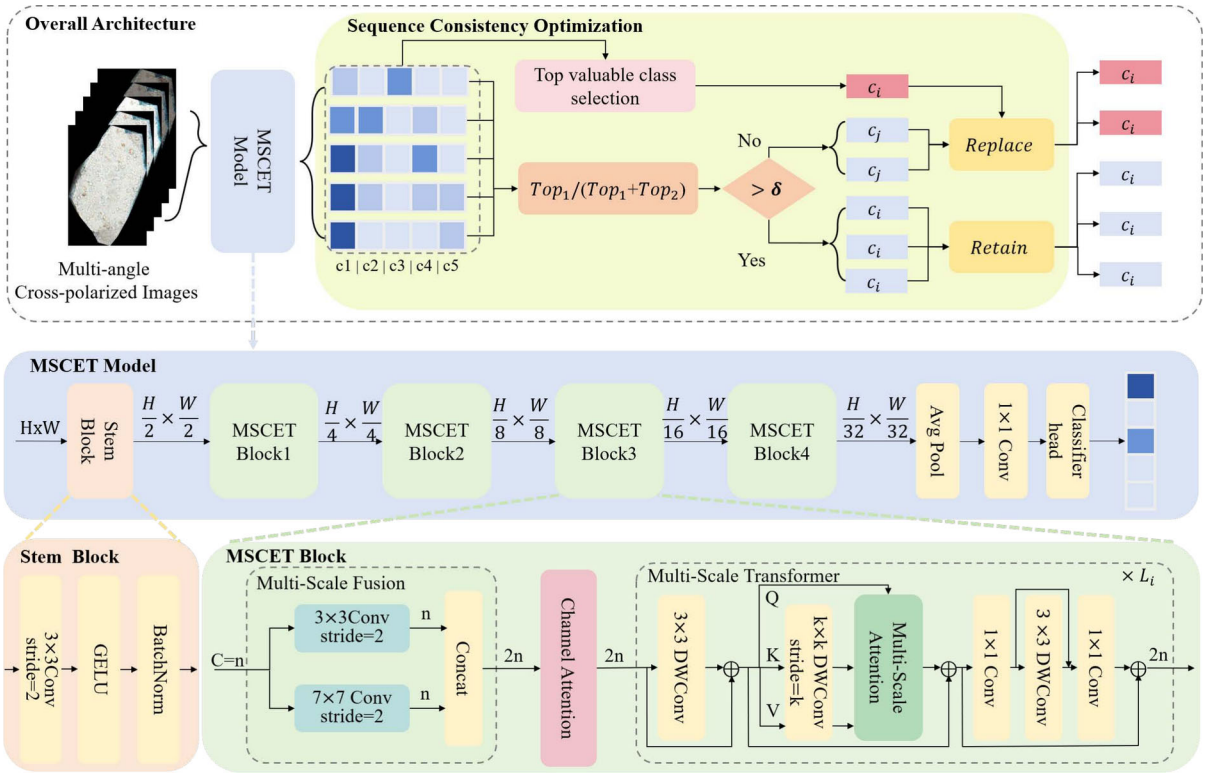

Fig. 2 The proposed rock thin section image identification method

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 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.

$$X_{c}^{i} = s_c \cdot X_{m}^{i} \tag{3}$$

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

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

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

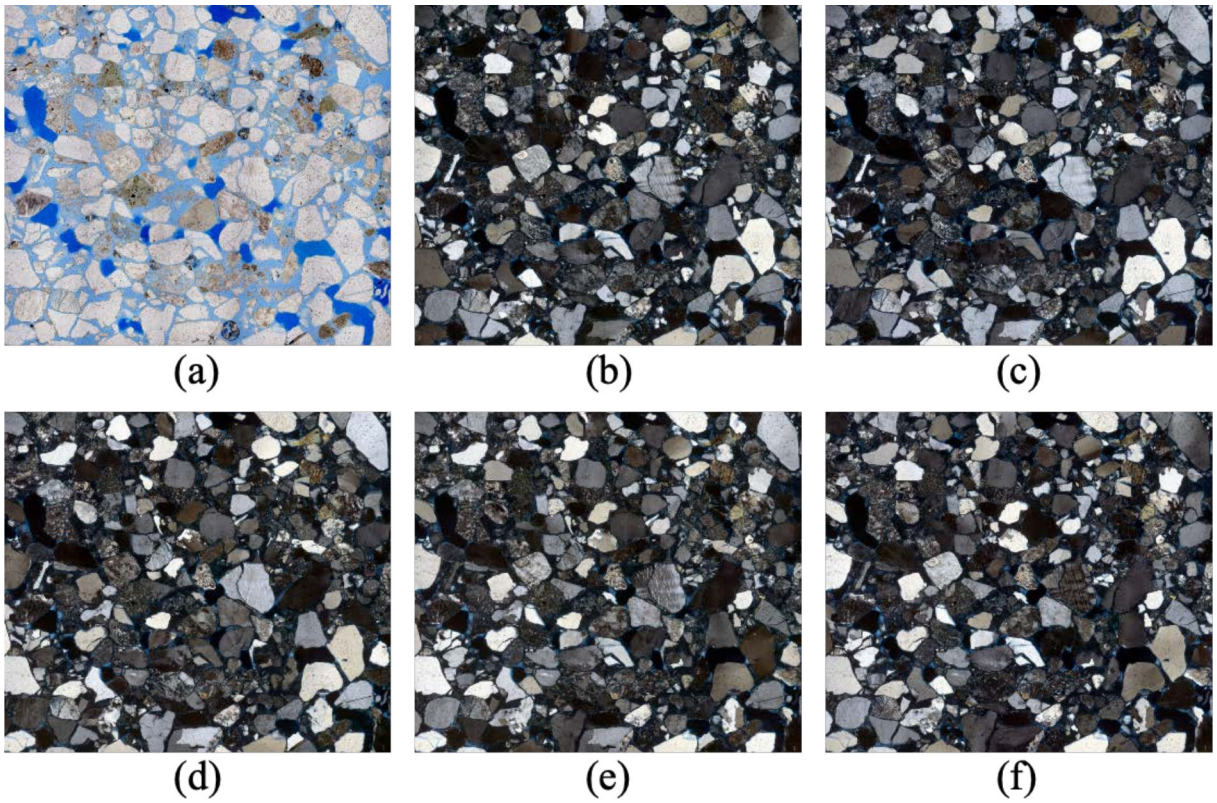

Fig. 3 Single-polarized and cross-polarized rock thin section images at different angles

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

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

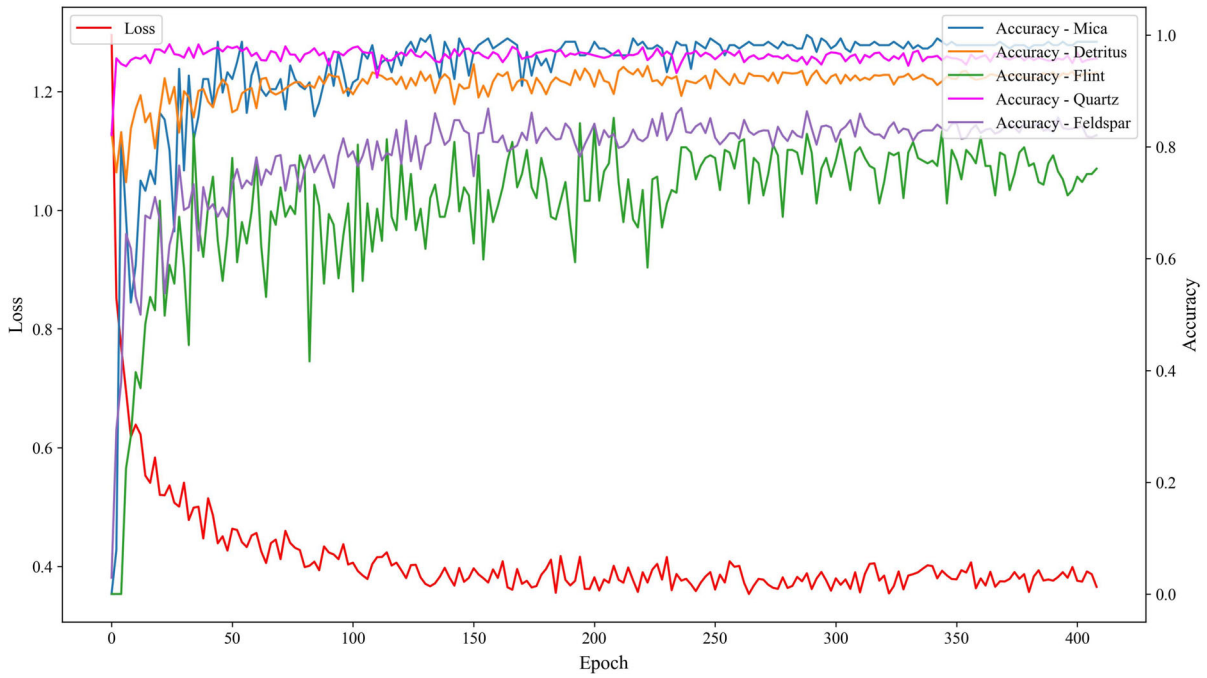

Fig. 4 The test loss and accuracy of the MSCET model.

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

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

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 (%)						Params(M)
	Mica	Detritus	Flint	Quartz	Feldspar	Overall	
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

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

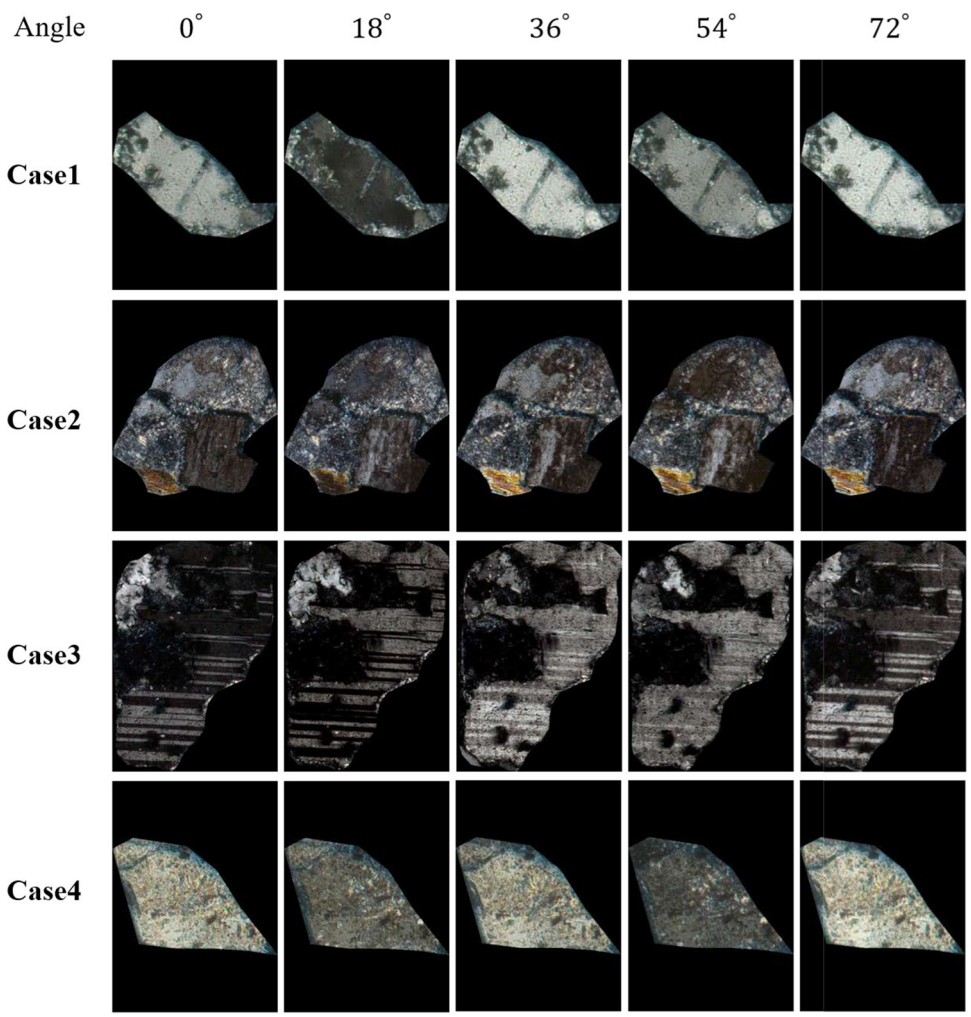

Fig. 5 Examples of four sets of cross-polarized images at different angles

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.

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

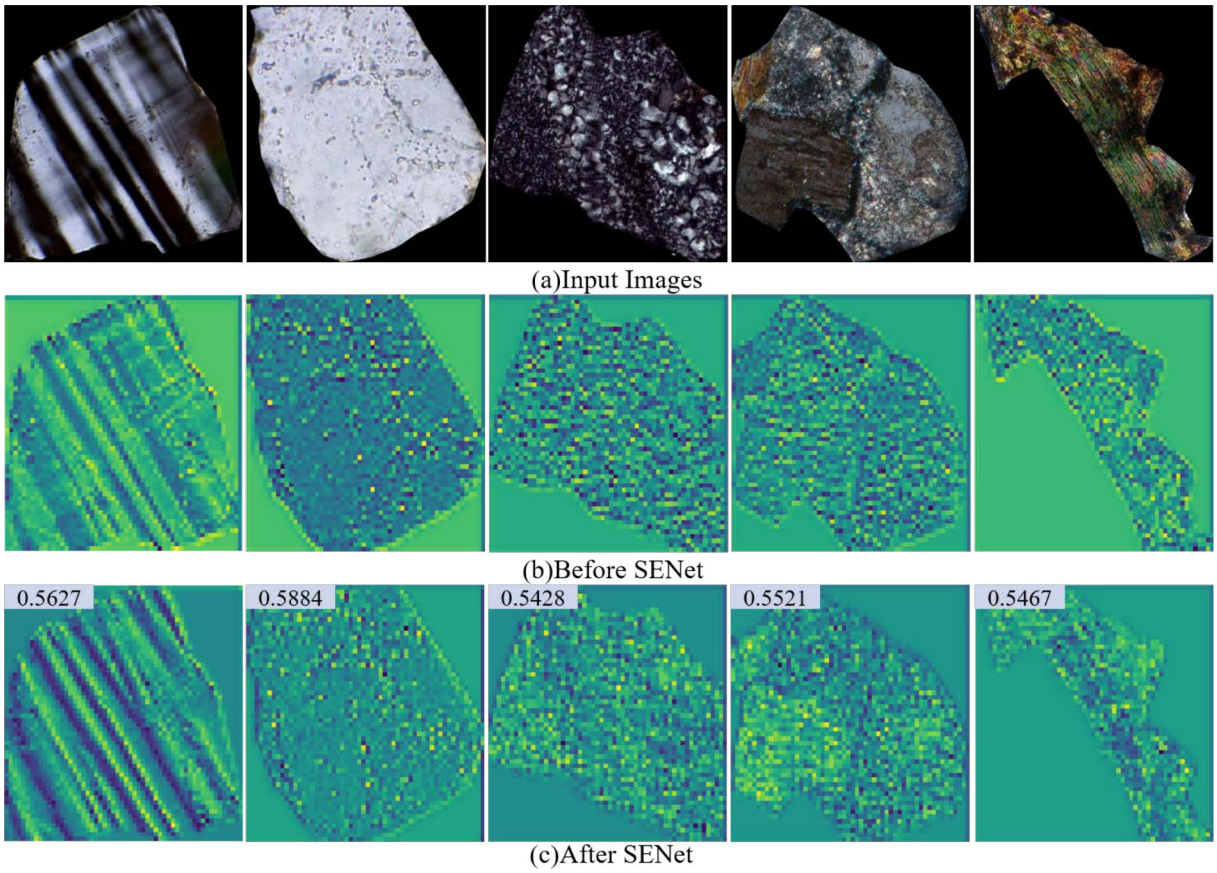

Fig. 6 Visualization of feature map before and after SENet application

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

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.

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)

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