Available Online: 26 August 2026 DOI: 10.54254/2977-3903/2026.36417

How many orientations do you need? A dual-polarisation few-shot benchmark for thin-section mineral classification on MUMDMC2025

Zhebin Yu[1,], Haichao Du[1], Guang Miao[1], Ang Li[1], Liming Sun[2]*

[1]China National Gold Group Geology Co., Ltd., Beijing, China
[2]School of Computer Science and Engineering, Northeastern University, Shenyang, China

*Corresponding Author. Email: 157231750@qq.com

Abstract. Automated mineral identification in thin section promises to relieve a slow and subjective bottleneck in petrography, yet most deep-learning classifiers use a single polarisation mode recorded at a single stage orientation, discarding information that human petrographers actively exploit. The recently released MUMDMC2025 dataset Pairs Plane-Polarised (PPL) and Crossed-Nicols (XPL) photomicrographs of five rock-forming minerals across a full 360-degree rotation, but no classification benchmark has been reported on it. This study establishes that benchmark. Five ImageNet-pre-trained backbones are compared under a rotation-aware, orientation-disjoint evaluation protocol. We find that when many orientations are labelled the five-mineral task saturates at perfect accuracy, so the informative setting is few-shot orientation classification, in which only one or a few labelled rotation views per class are available. In that regime Swin-Tiny is the strongest single-polarisation backbone (96.3 macro-F1), and a minimal two-stream PPL-XPL late-fusion model raises one-shot macro-F1 from 92.2 to 100.0, with the gain concentrated on the orientation-sensitive feldspars, where plagioclase F1 improves by 17.6 points. Feature-level fusion outperforms early channel stacking by 13.8 points. We further verify that the reported accuracy is not an artefact of the splitting protocol, apply Grad-CAM to confirm that predictions attend to diagnostic grain interiors instead of background, and show that the pipeline transfers to a seven-class hand-sample dataset. Code and configurations are released for reproducibility.

Keywords: mineral classification, polarised light microscopy, dual-polarisation fusion, few-shot learning, transfer learning

1. Introduction

Identifying rock-forming minerals in thin section is a routine but labour-intensive step in petrology and mineral exploration. A trained petrographer inspects each grain under a polarising microscope, rotating the stage and switching between Plane-Polarised Light (PPL) and Crossed Nicols (XPL) to read colour, relief, cleavage, twinning and interference colours. A complete thin section can take hours to analyse, and the outcome depends on operator experience, which makes the process difficult to scale and hard to reproduce

across laboratories. These constraints motivate automated classifiers that are not only accurate but also interpretable enough for practitioners to trust.

Deep learning has been applied to thin-section mineral and rock classification with considerable success, typically by fine-tuning ImageNet-pre-trained backbones. However, most existing classifiers use a single polarisation mode at a single stage orientation, discarding the complementary physics that a petrographer actively exploits. PPL and XPL are not redundant views of the same signal: PPL reports absorption colour, pleochroism, relief and grain morphology, whereas XPL reports birefringence and interference colours that vary with crystallographic orientation as the stage turns. Discarding either mode removes evidence that is diagnostic for the mineral pairs that are hardest to separate.

The recently released MUMDMC2025 dataset [1] is the first to provide, for five common rock-forming minerals, systematically paired PPL and XPL images across a full 360-degree rotation. To our knowledge no classification benchmark has yet been established on it, leaving open both the choice of backbone and the practical value of using both polarisation modes. Establishing such a benchmark requires care, because the dense angular sampling that makes the dataset valuable also makes it easy to report optimistic numbers: adjacent rotation views of the same grain are near-duplicates, so a careless split can place almost identical images in both training and test sets.

We provide that benchmark. Using each grain's full rotation sweep, a pre-trained backbone separates the five minerals almost perfectly, and perfect test accuracy is reached from as few as two labelled views per class. The scientifically interesting question is therefore not whether the task can be solved when data are abundant, but how few labelled orientations are needed and what helps when they are scarce. This mirrors practice: a petrographer photographing a slide typically captures a handful of orientations rather than a complete rotation sweep, so the low-label regime is the operationally relevant one. We accordingly study a few-shot orientation regime and show that a minimal PPL-XPL feature-level fusion improves recognition when only one or a few orientations per class are labelled, with the gain concentrated on the visually similar feldspars. Throughout, we use a rotation-aware evaluation protocol and verify model behaviour with Grad-CAM. Our contributions are:

(1) The first classification benchmark on MUMDMC2025, systematically comparing five modern CNN and transformer backbones under a reproducible, orientation-disjoint protocol, and characterising how accuracy scales with the number of labelled rotation views per class.

(2) A minimal and effective dual-polarisation fusion: a two-stream PPL-XPL late-fusion model that, with no architectural tricks, improves few-shot orientation classification over either polarisation alone, together with an ablation of late, early and score fusion.

(3) Honest and interpretable evaluation: we adopt an orientation-disjoint splitting protocol, verify that the reported few-shot accuracy is robust to the splitting choice rather than an artefact of it, and use Grad-CAM to expose where the model looks and why particular minerals are confused.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the model and the evaluation protocol. Section 4 reports the benchmark, the fusion results and the ablations. Section 5 discusses implications and limitations, and Section 6 concludes.

2. Related work

2.1. Deep learning for thin-section mineral classification

Convolutional networks and, more recently, vision transformers have become standard tools for petrographic image analysis. Transfer learning from ImageNet-pre-trained backbones such as ResNet [2], EfficientNet [3], ConvNeXt [4] and Swin [5] is widely used because annotated thin-section data are scarce and expensive to

produce, requiring expert labelling at the grain level. Representative studies apply convolutional networks, sometimes coupled with classical classifiers such as support vector machines [6], to discriminate rock-forming minerals, and semantic-segmentation models have been used to delineate grains in whole thin sections [7]. A recurring difficulty is the visual similarity of co-occurring minerals. Lou and Zhang [8] study the quartz, biotite and K-feldspar confusion in granite, which is also central to our dataset, and report that multi-angle dynamic imaging improves both accuracy and reproducibility relative to single-shot imaging.

2.2. Dual-polarisation and multi-angle imaging

Polarised-light microscopy yields two complementary modalities, and a growing line of work exploits the additional physics instead of collapsing it to a single image. Extinction-angle and polarisation-sequence features have been used to identify plagioclase from rotated polarised images [9], demonstrating that the way appearance evolves under rotation is itself diagnostic. Dual-encoder petrographic models that jointly process multiple modalities have recently appeared [10]. These methods are often architecturally heavy, which makes it difficult to attribute their gains to the extra input rather than to the extra capacity. In contrast, we deliberately use the simplest possible fusion, two backbones of identical architecture followed by a concatenation, so that the measured improvement is attributable to the dual-polarisation input itself rather than to a more expressive model.

2.3. Datasets, evaluation protocols and interpretability

The MUMDMC2025 dataset [1] provides photomicrographs of five rock-forming minerals under both PPL and XPL across a full rotation, but to our knowledge no classification benchmark has been reported on it. Two methodological points motivate our study. First, because the same grain is imaged at many close orientations, evaluation must avoid leaking near-duplicate rotation views across splits, an issue analogous to patient-level or specimen-level splitting in medical imaging, where per-image random splits are known to inflate reported performance. Second, interpretability matters for adoption in the geosciences, where a practitioner must be able to audit why a label was assigned. Grad-CAM [11] lets us verify that predictions attend to the grain and not to background or mounting artefacts. We bring these threads together into the first classification benchmark on MUMDMC2025, a minimal yet effective PPL-XPL fusion, an explicit analysis of the splitting protocol, and a Grad-CAM-based case study of the residual confusions.

3. Method

3.1. Overview

Figure 1 summarises our pipeline. Each thin-section grain is imaged under two polarisation modes, PPL and XPL, at a sequence of stage-rotation angles. A sample is a single rotation view, or a PPL-XPL pair captured at one angle. Every view is resized to 224 x 224 and passed through an ImageNet-pre-trained backbone used as a feature extractor; a linear head maps the pooled feature to the five mineral classes. For the dual-polarisation model the PPL and XPL streams are encoded by two backbones of identical architecture and combined by feature concatenation before the head. We deliberately keep the architecture minimal so that any performance change can be attributed cleanly to the input, that is single versus dual polarisation, rather than to added capacity or training tricks.

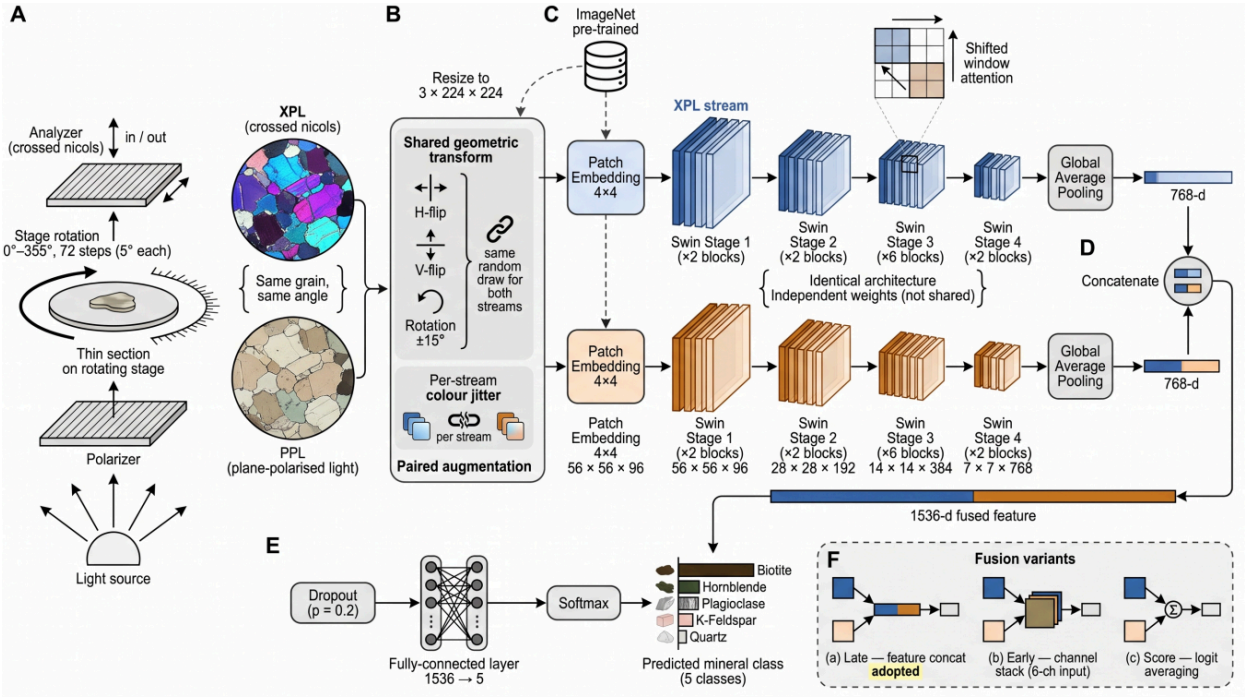

Figure 1. Overview of the dual-stream PPL-XPL late-fusion model. Each polarisation mode is encoded by an ImageNet-pre-trained backbone of the same architecture; the pooled features are concatenated and classified by a single linear head. Single-polarisation baselines use one stream only

3.2. Backbones and transfer learning

We compare five widely used ImageNet-pre-trained backbones spanning convolutional and transformer families and a range of capacities: ResNet-50 [2], EfficientNet-B0 [3], ConvNeXt-Tiny [4], Swin-Tiny [5] and MobileNetV3-Large [12]. Each backbone is loaded headless, that is with its classification layer removed so that it emits a globally pooled feature vector, and is fine-tuned end to end together with a freshly initialised linear classifier. All backbones share the same input size, optimiser, schedule and augmentation, so that the comparison in Table 1 isolates the effect of architecture. The feature dimension is determined empirically by a probe forward pass instead of being read from the model metadata, because for some networks the reported width differs from the true pooled output width. The strongest backbone is then carried forward as the shared encoder for the fusion experiments.

3.3. PPL-XPL late fusion

The two polarisation modes are physically complementary: PPL reveals colour, pleochroism, relief and morphology, whereas XPL reveals birefringence and interference colours that depend on crystallographic orientation. Our core model encodes the two modes with separate backbones and concatenates their pooled features, followed by a single linear classifier. This is late, or feature-level, fusion. We contrast it with two alternatives. Early fusion stacks the two modes on the channel axis, giving a six-channel input to a single backbone whose stem is adapted accordingly. Score fusion classifies the two streams independently and averages their logits. Geometric augmentation is shared across the two streams so that a paired XPL and PPL view stays spatially aligned, since the two images depict the same grain at the same angle; photometric jitter is applied independently per stream, because the modes legitimately differ in colour.

3.4. Training and evaluation protocol

Orientation-disjoint splitting. The dataset images each grain at many closely spaced rotation angles. Under dense sampling a naive per-image random split would place near-duplicate orientations of the same grain in both training and test, leaking information and inflating the reported accuracy. We therefore split by rotation sector: for every class the 0 to 355 degree circle is partitioned into contiguous arcs assigned to training, validation and test, so that test orientations are angularly separated from training orientations. Section 4.4 compares this protocol against the naive alternative.

Few-shot orientation regime. Because each class in the released subset corresponds to a single physical grain, a backbone given many training orientations separates the five minerals almost perfectly. We therefore centre the benchmark on the realistic few-shot orientation regime, in which only K labelled rotation views per class are available and the model must generalise to all unseen orientations. This is the situation a petrographer faces when only a few orientations of a grain are captured. We report results as a function of K and use $K = 1$ as the headline setting.

Augmentation and optimisation. Training uses random horizontal and vertical flips, small rotations and colour jitter. Flips and rotations are natural for thin-section grains, which have no canonical up direction. Models are optimised with AdamW using a cosine-annealed learning rate and mixed precision, with early stopping on the validation loss and a minimum-improvement threshold so that runs whose validation loss has plateaued terminate instead of drifting to the epoch cap. The checkpoint with the best validation loss is evaluated on the test set.

Metrics. We report Overall Accuracy (OA) and macro-averaged F1, together with per-class F1 and confusion matrices. For a test set of N views with predictions and ground-truth labels, overall accuracy is defined in Equation (1), where the indicator function counts correctly classified views:

$$OA = \frac{1}{N} \sum_{i=1}^{N} I(y_i = t_i) \tag{1}$$

Macro-averaged F1 gives every mineral class equal weight regardless of how many test views it contributes, which matters because per-class support is not identical after the angular split. It is defined in Equation (2) over the C classes, with per-class precision and recall:

$$macro - F1 = \frac{1}{C} \sum_{c=1}^{C} \frac{2P_c R_c}{P_c + R_c} \tag{2}$$

Because the few-shot regime is sensitive to which orientation happens to be labelled, every configuration is run over multiple random seeds, with each seed selecting a different set of training orientations, and we report the mean and standard deviation.

4. Experiments

4.1. Dataset and experimental setup

MUMDMC2025 [1] contains photomicrographs of five rock-forming minerals, namely biotite, hornblende, plagioclase, potassium feldspar (K-feldspar) and quartz, each imaged under PPL and XPL at 72 stage-rotation angles in 5-degree steps. The publicly released subset used here provides one physical grain per class, imaged at full frame resolution. We parse the rotation angle from the file naming convention of each modality and pair PPL with XPL at matching angles, which yields 357 paired views and 360 XPL views in total after discarding a small number of unpaired frames and the duplicate 360-degree frame that repeats the 0-degree view.

We resize all views to 224 x 224 and split each class rotation circle into contiguous training, validation and test arcs in a 70/15/15 ratio, so that test orientations are angularly separated from training ones as described in

Section 3.4. For the few-shot study we retain only K evenly spaced training orientations per class, rotating the selection offset with the seed so that different seeds label different orientations. All models are fine-tuned with AdamW using a base learning rate of 3×10^{-4}, weight decay of 0.05, a cosine schedule, batch size 32, label smoothing 0.1, mixed precision and early stopping on validation loss, on a single NVIDIA RTX 4090 GPU. Unless noted otherwise we report mean and standard deviation over three seeds, and the headline regime is $K = 1$. Under this configuration a single run completes in a few minutes, so the full benchmark of 69 runs is inexpensive to reproduce.

4.2. Backbone comparison

Table 1 compares the five ImageNet-pre-trained backbones in the $K = 1$ regime. With abundant orientations the task saturates, since the best backbone already reaches 100% test accuracy from only two labelled views per class, as shown at the right of Figure 2, so the single-view numbers separate the architectures far better. Swin-Tiny is the strongest with 96.3 macro-F1, ahead of EfficientNet-B0 at 89.4 and well ahead of ConvNeXt-Tiny at 63.7. The ordering does not follow parameter count: MobileNetV3-Large reaches 82.4 macro-F1 with 4.2 million parameters and 0.22 GFLOPs, outperforming ResNet-50 which uses roughly five times the parameters and nearly twenty times the computation. The large standard deviations reflect the genuine sensitivity of one-shot learning to which orientation is labelled, not noise in optimisation. We carry Swin-Tiny forward as the shared encoder for the fusion experiments.

Table 1. Backbone comparison on MUMDMC2025 (XPL single-polarisation, orientation-disjoint split, $K = 1$). Mean ± standard deviation over three seeds. OA and macro-F1 in percent

Backbone	OA (%)	Macro-F1 (%)	Params (M)	GFLOPs
ResNet-50	80.0 ± 12.9	77.2 ± 14.9	23.5	4.11
EfficientNet-B0	90.3 ± 8.7	89.4 ± 9.8	4.0	0.40
ConvNeXt-Tiny	67.3 ± 11.8	63.7 ± 16.0	27.8	4.47
Swin-Tiny	96.4 ± 3.0	96.3 ± 3.1	27.5	4.51
MobileNetV3-L	84.8 ± 4.8	82.4 ± 8.2	4.2	0.22

4.3. Dual-polarisation fusion

Table 2 reports the core comparison on the paired XPL and PPL set. In the one-shot regime XPL alone reaches 92.2 macro-F1 and PPL alone reaches 96.8, itself a notable observation given that XPL is the mode most commonly used in prior single-modality work. Combining both via late fusion reaches 100.0 with zero variance across seeds, a gain of 7.8 points over XPL. The improvement is concentrated on the feldspars: plagioclase F1 rises from 82.4 under XPL to 100, an increase of 17.6 points, and K-feldspar rises from 96.7 to 100. This is consistent with the PPL stream supplying complementary colour and relief cues that disambiguate the orientation-sensitive XPL appearance of feldspar, whose interference colours change markedly as the stage rotates.

Table 2. Dual-polarisation fusion versus single polarisation on the paired XPL and PPL set (backbone: Swin-Tiny, $K = 1$). Mean ± standard deviation over three seeds

Input	OA (%)	Macro-F1 (%)
XPL only	92.7 ± 3.0	92.2 ± 3.4
PPL only	97.0 ± 4.3	96.8 ± 4.5
Dual (late fusion)	100.0 ± 0.0	100.0 ± 0.0

Table 3 ablates the fusion mechanism. Feature-level (late) fusion and score-level fusion both reach 100%, whereas early channel stacking drops to 86.2, a deficit of 13.8 points. Early fusion is also the only variant that must modify the pretrained stem in order to accept six input channels, which discards part of the ImageNet initialisation at the layer where it is most transferable. The results indicate that keeping the two polarisation streams separate until after feature extraction is what matters, while the specific choice between combining features and combining scores is not critical on this dataset. Late fusion is preferable to score fusion in principle because a joint head can learn interactions between the two modalities rather than assuming their decisions are independent, and we therefore adopt it as the default.

Table 3. Fusion-strategy ablation (dual XPL and PPL, Swin-Tiny, $K = 1$). Mean ± standard deviation over three seeds

Fusion strategy	OA (%)	Macro-F1 (%)
Late (feature concatenation)	100.0 ± 0.0	100.0 ± 0.0
Early (channel stacking)	86.1 ± 5.2	86.2 ± 5.0
Score (logit averaging)	100.0 ± 0.0	100.0 ± 0.0

4.4. Orientation budget and splitting protocol

Figure 2 traces accuracy as a function of the number of labelled training orientations K. A single labelled view per class already yields strong accuracy. The dual-polarisation model dominates the single-polarisation baseline at $K = 1$, reaching 100 against 92.7 overall accuracy, and the two converge from $K = 2$ onward where both reach 100%. The benefit of fusion is thus concentrated where labels are scarcest, which is also where an automated tool is most useful in practice. From an annotation-budget perspective the result is actionable: capturing a second polarisation image at one orientation is cheaper than capturing and labelling a second orientation, because it requires only inserting the analyser rather than rotating the stage and re-focusing.

Table 4 examines the splitting protocol. Our angle-sector split holds out a contiguous block of unseen orientations, so test views are never near-duplicates of training views. A naive per-view random split yields a slightly lower accuracy, 89.1 against 96.4 overall accuracy, rather than an inflated one. With only one labelled orientation per class there are no near-duplicate views available for the random split to leak, so the two protocols agree that the few-shot accuracy reflects genuine across-orientation generalisation and is not a splitting artefact. We nevertheless adopt the orientation-disjoint protocol on principle, because in denser-sampling or multi-grain settings a per-view split would leak adjacent orientations and the resulting numbers would not be comparable across studies.

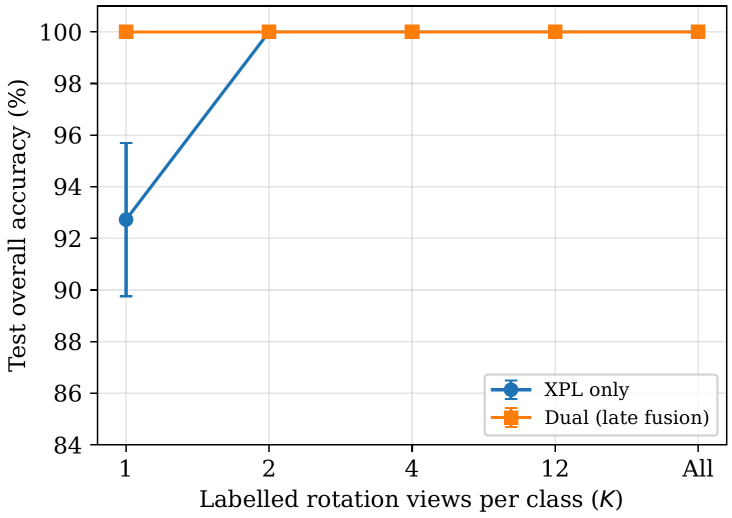

Figure 2. Test accuracy versus the number of labelled rotation views per class (K) for the best backbone, single-polarisation (XPL) against dual-polarisation late fusion. Fusion helps most when orientations are scarce; both saturate as K grows

Table 4. Robustness to the data-splitting protocol (Swin-Tiny, XPL, $K = 1$). Mean ± standard deviation over three seeds

Split protocol	OA (%)	Macro-F1 (%)
Angle-sector (orientation-disjoint)	96.4 ± 3.0	96.3 ± 3.1
Per-view random	89.1 ± 2.6	89.1 ± 2.7

4.5. Case study: confusion and Grad-CAM

Figure 3 shows the confusion matrix of the single-polarisation XPL model in the one-shot regime. Its errors concentrate on plagioclase, with a recall of 0.82 and confusions towards biotite and hornblende, and on biotite, with a recall of 0.88. These are the minerals whose one-shot XPL appearance is most orientation-dependent, while quartz and hornblende are recognised perfectly. The dual-polarisation model removes these errors entirely, as reported in Table 2, with plagioclase F1 rising from 82.4 to 100.

Figure 4 overlays Grad-CAM for a representative correctly classified view of each class. Because gradient-based class activation maps are designed for convolutional feature maps, we visualise the best convolutional backbone, EfficientNet-B0; applying the same procedure to a hierarchical transformer produces maps dominated by the patch grid rather than by image content, and we therefore do not use them. The maps show that the model attends to the grain interior, specifically the strong absorption colour of biotite and hornblende and the cleavage and twinning traces of the feldspars, rather than to background, grain boundaries or mounting medium. This is the behaviour a petrographer would expect and supports the interpretation that the classifier has learned mineralogically meaningful evidence rather than incidental image cues.

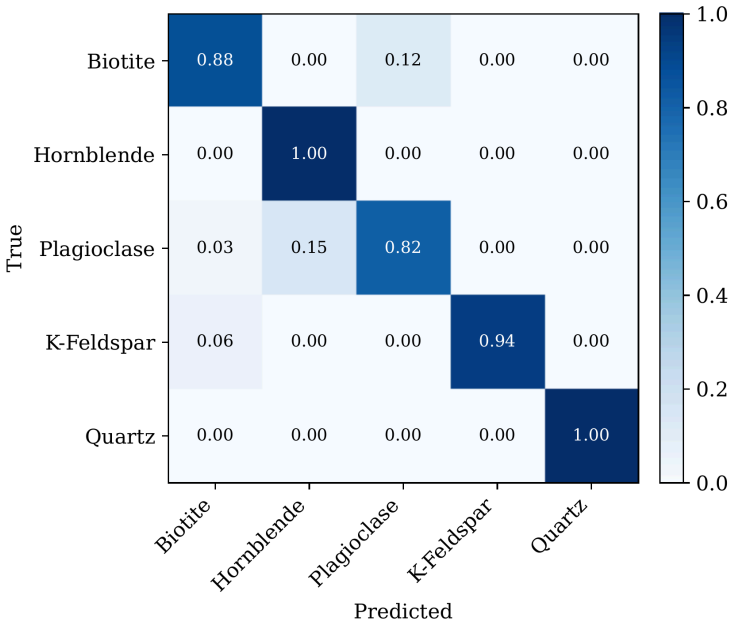

Figure 3. Confusion matrix of the single-polarisation (XPL) model in the one-shot regime, row-normalised and summed over seeds. Errors concentrate on plagioclase, which is confused with biotite and hornblende, and on biotite; dual-polarisation fusion removes them

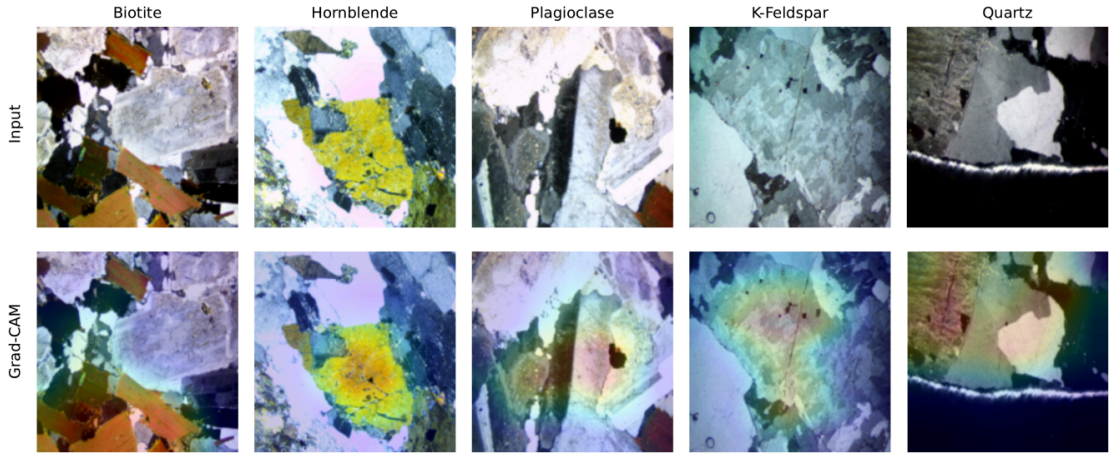

Figure 4. Grad-CAM for one representative view per class (XPL stream). Top row: input; bottom row: Grad-CAM overlay. Attention falls on diagnostic grain regions

4.6. Cross-dataset generalisation

Finally, Table 5 applies the same pipeline to the Minerals Identification hand-sample dataset, which contains seven mineral classes photographed as macroscopic hand specimens rather than as thin sections under polarised light. This is a substantially different imaging modality, with no rotation series and no polarisation information. The pipeline reaches 87.3 overall accuracy and 85.5 macro-F1 with a small standard deviation, indicating that the transfer-learning recipe is not specific to photomicrographs. Since this dataset provides only one imaging mode, it serves as a check on the general pipeline rather than on the fusion contribution.

Table 5. Cross-dataset generalisation: the same pipeline applied to the Minerals Identification hand-sample dataset (seven classes). Mean ± standard deviation over three seeds

Dataset	OA (%)	Macro-F1 (%)
Minerals Identification (7 classes)	87.3 ± 1.4	85.5 ± 1.5

5. Discussion

Three findings have practical consequences for building petrographic classifiers. First, benchmark design matters more than model capacity on this task. Once several orientations per class are labelled, every backbone we tested reaches perfect accuracy, so a paper reporting only that number would convey little information about the relative merits of architectures. Reporting accuracy as a function of the labelled-orientation budget, as in Figure 2, is far more informative and exposes differences that a single saturated number conceals.

Second, the value of the second polarisation mode is largest where data are scarcest. When a single orientation is labelled, the classifier cannot observe how interference colours evolve under rotation, and the complementary PPL evidence substitutes for that missing temporal signal. As the orientation budget grows the model can infer the same information from multiple XPL views, and the two curves converge. This suggests a simple acquisition guideline: when annotation effort is limited, capture both polarisation modes at one orientation rather than one mode at two orientations.

Third, the architectural comparison favours efficiency. Swin-Tiny performs best overall, but MobileNetV3-Large attains 82.4 macro-F1 at roughly one twentieth of the computational cost, which is attractive for deployment on microscope-attached hardware or in field laboratories where a workstation-class GPU is unavailable. ConvNeXt-Tiny performs worst in the one-shot regime despite its capacity, which is consistent with larger modern convolutional networks being comparatively data-hungry when fine-tuned on very few labelled examples.

6. Conclusion

We presented the first classification benchmark on the MUMDMC2025 dual-polarisation thin-section dataset. Comparing five ImageNet-pre-trained backbones under a rotation-aware, orientation-disjoint protocol, we found that with full angular coverage the five-mineral task is almost trivially separable, so the informative regime is few-shot orientation classification. In that regime a minimal PPL-XPL late-fusion model improves over either polarisation alone, raising one-shot macro-F1 from 92.2 to 100.0, with the gain concentrated on the visually similar feldspars, and feature-level fusion outperforms early channel stacking by 13.8 points. We evaluated under an orientation-disjoint splitting protocol, verified that the few-shot accuracy is not an artefact of the split, and used Grad-CAM to confirm that predictions attend to diagnostic grain regions rather than to background.

Limitations. The publicly released subset used here contains a single physical grain per class, so our results measure generalisation across orientations of a fixed grain rather than across distinct grains of the same mineral. Absolute accuracies are correspondingly optimistic for cross-grain deployment, and the five-class, single-lithology scope is narrow. The few-shot numbers are also sensitive to which orientation is labelled, which we mitigate by averaging over seeds but cannot eliminate with one grain per class. Reported perfect scores should therefore be read as saturation of this particular benchmark, not as evidence that mineral identification is solved.

Future work. The natural next steps are to evaluate on the full multi-grain MUMDMC2025 release and on broader lithologies, to model the rotation sequence explicitly using temporal or rotation-equivariant encoders rather than treating orientations independently, and to explore self-supervised pre-training on unlabelled photomicrographs in order to reduce the labelled-orientation budget further.

Declarations

Data availability. The MUMDMC2025 dataset is publicly available from the original publication [1] under its stated licence. The Minerals Identification dataset is publicly available from its original source. No new data were collected for this study.

Code availability. The source code is not publicly released. Configuration files and code supporting the reported experiments are available from the corresponding author upon reasonable request.

Authorship

Zhebin Yu: conceptualisation, study design, supervision, and manuscript review and editing. Haichao Du: data curation, petrographic validation and investigation. Ang Li: formal analysis, visualisation and investigation. Liming Sun: software implementation, experiments and writing of the original draft. All authors read and approved the final manuscript.

References

[1] Amer, B. G., Mousa, H. M., Dawoud, M., & Youssef, A. (2025). A photomicrographic dataset of rocks for the accurate classification of minerals. *Scientific Data, 12*(1), Article 1775. https://doi.org/10.1038/s41597-025-05879-9

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

[3] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (Volume 97)* (pp. 6105–6114). PMLR.

[4] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11966–11976). IEEE. https://doi.org/10.1109/CVPR52688.2022.01167

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012–10022). IEEE. https://doi.org/10.1109/ICCV48922.2021.00986

[6] Aydin, I., Sener, T. K., Kilic, A. D., & Dervis, H. (2025). Classification of thin-section rock images using a combined CNN and SVM approach. *Minerals, 15*(9), 976. https://doi.org/10.3390/min15090976

[7] Morales, J., Saldivia, C., Lobo, R., del Pino, M., Muñoz, M., Caniggia, G., Catalán, J., Hayde, R., & Poblete, V. (2025). Visual analysis of deep learning semantic segmentation applied to petrographic thin sections. *Scientific Reports, 15*(1), Article 14612. https://doi.org/10.1038/s41598-025-99767-2

[8] Lou, W., & Zhang, D. (2025). Applications of deep learning in mineral discrimination: A case study of quartz, biotite and K-feldspar from granite. *Journal of Earth Science, 36*(1), 29-45. https://doi.org/10.1007/s12583-022-1672-7

[9] Shu, J., He, X., Su, G., He, H., Yue, F., & Teng, Q. (2024). Identification of plagioclase extinction-angle features from polarized images using deep neural network. *Physical Review E, 110*, 045303.

https://doi.org/10.1103/PhysRevE.110.045303

[10] Chacón, I. D., Puentes, P. R., Pearse, J., & Arbeláez, P. (2025). *Towards automated petrography*. arXiv. https://doi.org/10.48550/arXiv.2511.00328

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

[12] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324). IEEE. https: //doi.org/10.1109/ICCV.2019.00140