

ZERO-SHOT MINERAL SEGMENTATION IN REFLECTED-LIGHT MICROSCOPY USING THE SEGMENT ANYTHING MODEL

ADRIAAN CAMPO^{✉1,2} AND ERIC PIRARD¹

¹GeMMe, Georesources and Geoimaging, Université de Liège, Sart Tilman B52, Allée de la Découverte 9, 4000 Liège, Belgium; ²Faculty of Arts and Philosophy, Institute for Psychoacoustics and Electronic Music (IPEM), Ghent University, Miriam Makebaplein 1, 9000 Ghent, Belgium
e-mail: adriaan.campo@ugent.be, eric.pirard@uliege.be

(Received July 1, 2026; revised September 14, 2026; accepted September 16, 2026)

ABSTRACT

Automated mineral segmentation underpins quantitative mineralogy and stereological characterization, but supervised deep-learning methods require annotated data and may generalize poorly across mineralogical domains. This study evaluated the Segment Anything Model (SAM), a general-purpose foundation model, for reflected-light microscopy (RLM) mineral-image analysis in a purely zero-shot setting, without task-specific training, fine-tuning, or domain adaptation. Two complementary case studies were conducted. First, three SAM encoder variants (ViT-B, ViT-L, and ViT-H) were evaluated on publicly available iron-ore (FeM) and copper-ore (Cu) RLM datasets using SEM-derived reference masks. SAM achieved mean F1-scores of 0.77–0.88, with ViT-L and ViT-H reaching 93–94% of the previously reported supervised DeepLabv3+ F1-score benchmark for FeM and approximately 93% for Cu. Second, SAM was applied to polarized RLM image stacks of dolomite to assess its influence on downstream stereological characterization. No significant differences from the gradient-based method were detected for any RGB configuration or PCA ViT-B in major or minor intercept length or bireflectance. The largest deviations occurred for PCA-preprocessed ViT-H, with median major and minor intercept lengths approximately 21% and 17% higher, respectively, and bireflectance approximately 16% lower. No significant method-dependent difference was detected in principal anisotropy orientation. These results show that zero-shot SAM can support mineral-image segmentation and downstream stereological characterization without annotated training data, while highlighting the influence of preprocessing and encoder choice on the derived grain representation.

Keywords: Automated mineralogy; image analysis; ore microscopy; reflected-light microscopy; Segment Anything Model; zero-shot segmentation.

INTRODUCTION

Quantitative mineralogical characterization relies on image segmentation to delineate mineral phases, individual grains, grain boundaries, pores, and larger textural domains. These representations provide the basis for describing mineral textures and microstructures, including grain-size distributions, liberation and intergrowth characteristics, lineal intercept statistics, and grain orientation and anisotropy (Cruz-Orive, 2017). Such descriptors help in understanding and predicting material behavior during mineral processing (Pérez-Barnuevo *et al.*, 2013). In these workflows, segmentation is often an intermediate representation from which quantitative descriptors are derived. The delineation of grains and their boundaries determines the geometrical information available for subsequent measurements, linking segmentation directly to quantitative characterization (Pirard *et al.*, 2007; Jaimes Contreras *et al.*, 2010).

Quantitative mineral characterization commonly relies on automated scanning electron microscopy (SEM)-based mineralogy systems or reflected-light

microscopy (RLM) combined with image-analysis approaches (Gu, 2003; Fandrich *et al.*, 2007). More advanced optical methods, including multispectral and multiradial reflected-light imaging, have also been developed to exploit wavelength-dependent reflectance and polarization-dependent optical anisotropy for improved mineral discrimination, although these techniques remain relatively specialized (Pirard, 2004).

Although SEM-based automated mineralogy provides highly detailed quantitative mineralogical information, its acquisition cost, instrument complexity, and acquisition time can limit routine deployment (Fandrich *et al.*, 2007; Schulz *et al.*, 2020). RLM, on the other hand, remains attractive because of its accessibility, rapid image acquisition, and ability to capture textural information over comparatively large sample areas (Pirard *et al.*, 2007; Jaimes Contreras *et al.*, 2010). However, reliable segmentation of RLM images remains challenging because of polishing artifacts, variable illumination, weak grain-boundary contrast, and the complexity of mineral assemblages (Álvarez Iglesias *et al.*, 2011; Filippo *et al.*, 2021).

To address these challenges, a wide range of classical image-segmentation approaches has been developed, including thresholding, watershed transforms, edge detection, and graph-based methods (e.g., Otsu, 1979; Canny, 1986; Beucher, 1992; Felzenszwalb and Huttenlocher, 2004). More recently, deep-learning architectures such as U-Net, SegNet, and DeepLab have become widely used for semantic segmentation of complex images (e.g., Ronneberger *et al.*, 2015; Badrinarayan *et al.*, 2017; Chen *et al.*, 2018).

While supervised deep-learning methods have substantially improved image segmentation, their application to mineralogical imaging remains constrained by the need for representative annotated datasets (Álvarez Iglesias *et al.*, 2019; Koh *et al.*, 2021; De Castro *et al.*, 2024). Generating representative mineralogical annotations can require substantial expert interpretation or additional analytical information. For example, objective pixel-wise reference masks for RLM images can be generated through correlative microscopy using spatially coregistered RLM and SEM acquisitions (Filippo *et al.*, 2021). Moreover, models trained on specific ore types, mineral assemblages, or imaging conditions may generalize poorly to other domains, potentially requiring additional annotation, retraining, or domain adaptation (Filippo *et al.*, 2021; Quiñero-Candela *et al.*, 2009).

Foundation models offer a potential alternative to task-specific supervised training. They are trained on broad datasets at scale and can be adapted to a range of downstream tasks (Bommasani *et al.*, 2021). In computer vision, large-scale pretraining has demonstrated strong transfer to previously unseen tasks and domains (Kolesnikov *et al.*, 2020; Azizi *et al.*, 2021). General-purpose vision foundation models have extended this approach to image segmentation, allowing previously unseen image domains to be segmented without task-specific training (Kirillov *et al.*, 2023).

Among vision foundation models, the Segment Anything Model (SAM) has attracted particular attention (Kirillov *et al.*, 2023). Its potential for scientific microscopy has subsequently been demonstrated across a range of imaging applications. Archit *et al.* (2025) introduced Segment Anything for Microscopy (μ SAM), extending SAM through models fine-tuned for light and electron microscopy, while Li *et al.* (2025) developed a training-free SAM-based approach for segmentation of material microstructures. Within mineralogical and petrographic microscopy, SAM-based approaches have begun to emerge for transmitted-light thin-section imaging and automated grain segmentation (Vellappally *et al.*, 2024; Frayssinet *et al.*, 2026). These developments highlight the growing potential of foundation-model segmentation for quantitative microscopy. However, the application of SAM to RLM remains comparatively

unexplored, particularly in a strictly zero-shot setting in which the pretrained model is applied directly without retraining, fine-tuning, or domain adaptation.

Segmentation methods can be evaluated at different levels depending on the intended application. Where an independent reference mask is available, segmentation performance can be assessed directly using measures of agreement between predicted and reference labels. In quantitative mineralogical and stereological applications, however, segmentation also provides the representation from which downstream descriptors are derived. Consequently, differences between segmentation methods may propagate into measurements of the underlying microstructure even when direct grain-level ground truth is unavailable. Despite the increasing application of foundation models to microscopy, comparatively little attention has been given to how foundation-model-derived segmentations influence stereologically relevant descriptors, including directional intercept lengths and derived measures of grain size and anisotropy.

This study investigated the applicability of SAM to RLM mineral images through two complementary case studies. In the first, three SAM encoder variants (ViT-B, ViT-L, and ViT-H) were evaluated in a purely zero-shot setting on publicly available iron-ore and copper-ore microscopy datasets using conventional pixel-wise segmentation metrics (Gomes *et al.*, 2021a,b). In the second, SAM was applied to polarized RLM image stacks of dolomite to compare stereological descriptors derived from SAM-generated grain masks with those obtained using a gradient-based method. Each stack comprised six polarization orientations, requiring preprocessing to produce input compatible with SAM. Two preprocessing strategies were evaluated, each in combination with the three SAM encoder variants.

Together, these case studies evaluate direct segmentation performance where independent reference masks are available and downstream quantitative texture characterization where grain-level ground truth is unavailable. The aim was to determine how well a general-purpose foundation model supports quantitative mineral-image analysis in a zero-shot setting and to identify limitations and method-dependent differences in its application to RLM.

MATERIALS AND METHODS

Segment Anything Model

SAM is a foundation-model-based image segmentation framework developed by Meta AI (Kirillov *et al.*, 2023). The model consists of an image encoder, a prompt encoder, and a mask decoder that together generate segmentation masks from visual input data. Three variants of the SAM image encoder were investigated:

ViT-B (Base, 93 million parameters), ViT-L (Large, 308 million parameters), and ViT-H (Huge, 636 million parameters). These Vision Transformer encoders differ primarily in network depth, embedding dimensionality, and the number of trainable parameters, resulting in increasing model capacity and computational cost (Kirillov *et al.*, 2023). SAM was used in automated segmentation mode in this study.

Case study I

The first case study evaluated SAM on two publicly available RLM mineral-image datasets introduced by Filippo *et al.* (2021). Both contain paired RGB RLM images and binary masks distinguishing ore grains from embedding resin. The reference masks were generated by spatially coregistering RLM images with SEM

acquisitions and thresholding the SEM-derived information to produce objective binary labels.

Cu dataset

The Cu dataset contains 121 spatially coregistered RLM images and SEM-derived reference masks of a copper ore from Yauri (Cusco, Peru) characterized by complex mineralogy including sulfides, oxides, silicates, and native copper (Gomes *et al.*, 2021a). Samples were size-classified (+74 –100 μm), cold-mounted in epoxy, and prepared through grinding and polishing. Images were acquired using a reflected-light microscope equipped with a 20 $\times$ objective (NA 0.40). The resulting registered image pairs have dimensions of 1017 $\times$ 753 pixels and a spatial resolution of approximately 0.53 $\mu\text{m}/\text{pixel}$ (see Fig. 1).



Fig. 1. Representative images from the Cu dataset: (a) RLM image and (b) corresponding binary reference mask. Ore grains are shown in black and embedding resin in white. Data reproduced from Gomes *et al.* (2021a).

FeM dataset

The FeM dataset contains 81 correlated image pairs acquired from an itabiritic iron-ore concentrate from the Quadrilátero Ferrífero region (Brazil), consisting predominantly of hematite and quartz, with minor magnetite and goethite phases (Gomes *et al.*, 2021b). Samples were size-classified and density-concentrated

+105 –149 μm , density >3.2), cold-mounted in epoxy, and subsequently ground and polished. Imaging was performed using a reflected-light microscope equipped with a 10 $\times$ objective (NA 0.20). The spatially coregistered RLM images and SEM-derived reference masks have dimensions of 999 $\times$ 756 pixels and a spatial resolution of approximately 1.05 $\mu\text{m}/\text{pixel}$ (see Fig. 2).

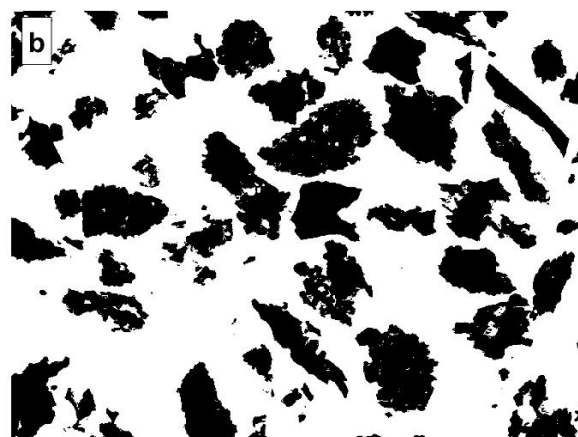
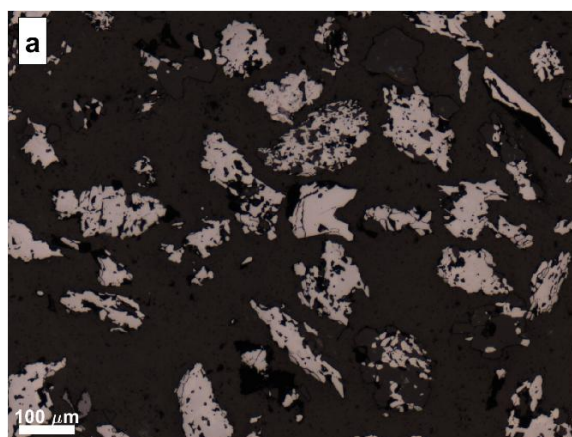


Fig. 2. Representative images from the FeM dataset: (a) RLM image and (b) corresponding binary reference mask. Ore grains are shown in black and embedding resin in white. Data reproduced from Gomes *et al.* (2021b).

Evaluation metrics

Segmentation performance was evaluated using overall accuracy (OA), precision, recall, and F1-score. These metrics were calculated on a pixel-wise basis using the SEM-derived reference masks provided with each dataset.

OA quantifies the proportion of correctly classified pixels:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Precision and recall were calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The F1-score was used as the primary segmentation metric and is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

In addition to segmentation accuracy, wall-clock inference time per image was recorded as an indicator of computational performance under identical hardware conditions.

For each dataset and encoder variant, performance statistics were calculated on a per-image basis and summarized using the mean and standard deviation (SD).

Segmentation results were also assessed qualitatively. For each image, a 2×2 overview panel showed agreement and disagreement with the reference mask, highlighting false-positive (FP) and false-negative (FN) regions for each method.

Case study II

The dataset consists of RLM image stacks acquired at six polarizer rotation angles (0° , 30° , 60° , 90° , 120° , and 150°). Each sample therefore contains six monochrome 8-bit images representing the same field of view under different polarization conditions. In addition, each stack is accompanied by an 8-bit grain-boundary map generated using a gradient-based method (Jaimes Contreras *et al.*, 2010).

The complete dataset contains seven image stacks and corresponding grain-boundary maps. In the grain-boundary maps, pixels are labeled as pore = 100, resin/matrix = 50, and grain = 0 (see Fig. 4). All images have a resolution of 1280×1024 pixels. Because this resolution exceeded the practical memory limitations for SAM inference on the available workstation, each image stack was subdivided into 12 tiles of 300×300 pixels, yielding a total working dataset of 84 tile stacks (see Fig. 3).

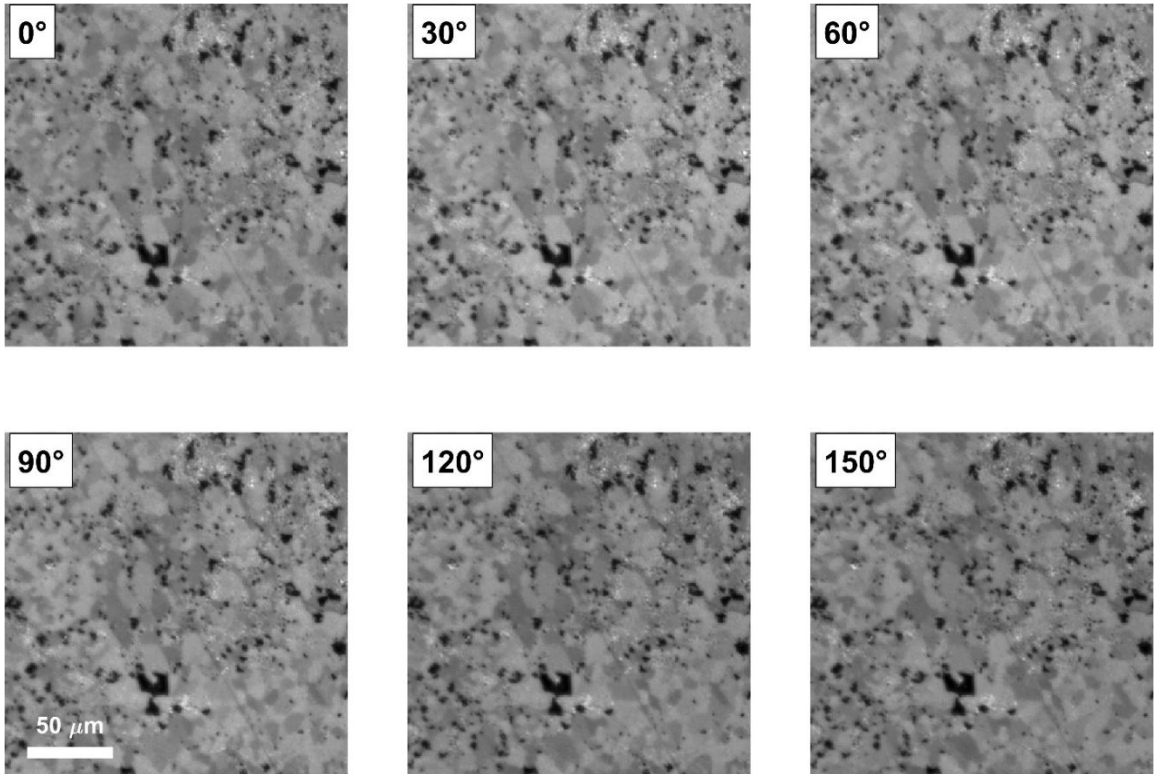


Fig. 3. Representative polarized RLM 300×300 -pixel tile stack of dolomite acquired at polarizer rotation angles of 0° , 30° , 60° , 90° , 120° , and 150° .

Image preprocessing

Two image preprocessing strategies were applied prior to SAM inference.

RGB projection

The six polarization images were linearly combined into three RGB channels for input to SAM, which cannot directly process six-channel polarization stacks. This preprocessing strategy is hereafter referred to as RGB projection.

Principal component analysis

Principal component analysis (PCA) was applied to the complete polarization stack, and the first three principal components were assigned to the red, green, and blue channels of an RGB image. This concentrated variance associated with polarization-dependent reflectance changes into three orthogonal components suitable for visualization and segmentation. This strategy is hereafter referred to as PCA preprocessing.

Stereological analysis

SAM generated closed object masks representing grain domains. Only the external boundaries of the

segmented grain domains were extracted and converted into grain-boundary maps for subsequent intercept-based stereological analysis (Underwood, 1970). The following descriptors were extracted, as described in Jaimes Contreras *et al.* (2010):

- Porosity (%)
- Major intercept length (L_a)
- Minor intercept length (L_b)
- Principal anisotropy orientation (Φ)
- Bireflectance
- Intercept count

Porosity was estimated exclusively from the gradient-based grain-boundary maps by counting pixels labeled as pores according to the Delesse principle (Delesse, 1848). The bireflectance value was calculated as the maximum pixel intensity difference over the six polarizer orientations and normalized using the anisotropy descriptors L_a and L_b (see Fig. 4) (Jaimes Contreras *et al.*, 2010). The intercept count corresponds to the number of individual intercept elements accumulated across all evaluated directions within each 300×300 -pixel tile (directions range from 0° to 180° with 2° increments).

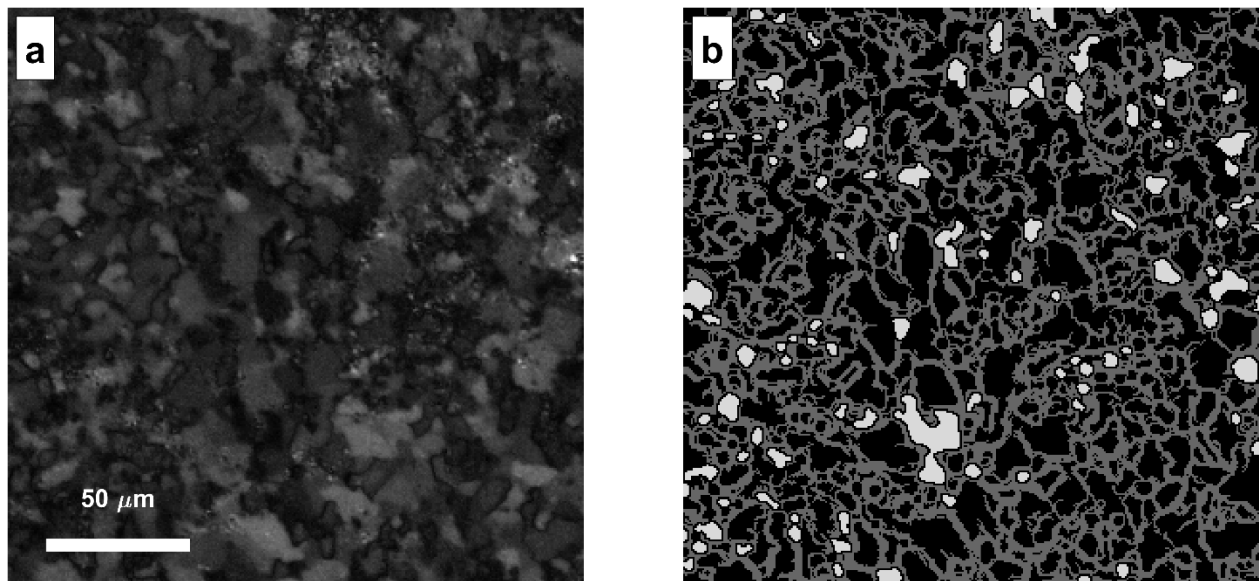


Fig. 4. Representative images used in case study II: (a) bireflectance image and (b) corresponding gradient-based grain-boundary map. Grain interiors are shown in black, grain boundaries in gray, and pores in light gray.

Statistical analysis

For the polarized dolomite case study, stereological descriptors were first computed for each tile. The median across the 12 tiles in each original image stack was then calculated. These image-level medians were used for statistical comparisons between the gradient-based method and the six SAM encoder–preprocessing configurations.

Differences among methods were assessed using the Friedman test, and effect sizes were quantified using Kendall's coefficient of concordance (W) (Friedman, 1937; Kendall and Smith, 1939). When significant differences were detected, pairwise post hoc comparisons were performed using Dunn–Šidák-adjusted multiple comparisons (Dunn, 1961). Statistical significance was defined as $p < 0.05$.

Experimental workflow

The workflow for each case study is summarized below.

Case study I workflow

RLM images were processed by each of the three SAM encoders to generate binary masks, which were then evaluated pixel-wise for the two datasets.

Case study II workflow

Polarized image stacks underwent RGB projection or PCA preprocessing before inference with each SAM encoder; the resulting grain masks were used to calculate L_a , L_b , Φ , bireflectance, and intercept count.

RESULTS

Case study I

Table 1 summarizes the segmentation performance of the three SAM encoders on the Cu and FeM datasets.

For the Cu dataset, mean OA values ranged between 0.90 and 0.91, while mean F1-scores ranged between 0.86 and 0.87 (Table 1). The ViT-B encoder achieved the highest mean precision (0.91 ± 0.06) and F1-score (0.87 ± 0.05), whereas the ViT-L and ViT-H encoders achieved slightly higher recall values (0.89 ± 0.02).

For the FeM dataset, larger differences between encoder variants were observed (Table 1). The ViT-B encoder achieved a mean OA of 0.86 ± 0.03 and a mean

F1-score of 0.77 ± 0.07 . In contrast, ViT-L and ViT-H achieved mean F1-scores of 0.87 ± 0.02 and 0.88 ± 0.02 , respectively, and mean OA values of approximately 0.92. These improvements were accompanied by substantially higher recall values (0.83 ± 0.03 and 0.84 ± 0.03 for ViT-L and ViT-H, respectively) compared with ViT-B (0.65 ± 0.09), while mean precision remained at or above 0.92 for all models.

Mean inference time increased with encoder size (Table 1). Across both datasets, inference times ranged from 6.52 ± 0.18 s per image for ViT-B to 11.00 ± 0.73 s per image for ViT-H.

Representative qualitative segmentation results for the Cu dataset are shown in Fig. 5. All encoder variants produced grain masks that closely matched the reference segmentation, with only limited false-positive and false-negative regions visible along grain boundaries. These qualitative observations are consistent with the results in Table 1. Representative qualitative segmentation results for the FeM dataset are shown in Fig. 6. Compared with the Cu dataset, larger disagreement regions are visible, particularly for the ViT-B encoder. The ViT-L and ViT-H encoders produced more continuous grain masks and visibly reduced false-negative regions relative to ViT-B, consistent with the higher recall and F1-scores reported in Table 1.

Table 2 compares the zero-shot SAM results with the supervised DeepLabv3+ benchmark reported by Filippio *et al.* (2021).

Table 1. Segmentation performance and inference time of SAM on the Cu and FeM RLM datasets in a zero-shot setting. Values are reported as image-wise mean $\pm$ SD for OA, precision, recall, F1-score, and wall-clock inference time per image.

Dataset	Model	OA	Precision	Recall	F1-score	Inference time (s)
Cu	ViT-B	0.91 ± 0.04	0.91 ± 0.06	0.84 ± 0.06	0.87 ± 0.05	6.52 ± 0.18
Cu	ViT-L	0.90 ± 0.04	0.85 ± 0.08	0.89 ± 0.02	0.86 ± 0.04	8.74 ± 2.21
Cu	ViT-H	0.90 ± 0.04	0.84 ± 0.08	0.89 ± 0.02	0.86 ± 0.05	10.78 ± 0.26
FeM	ViT-B	0.86 ± 0.03	0.95 ± 0.02	0.65 ± 0.09	0.77 ± 0.07	6.61 ± 0.18
FeM	ViT-L	0.92 ± 0.01	0.92 ± 0.03	0.83 ± 0.03	0.87 ± 0.02	8.33 ± 0.33
FeM	ViT-H	0.92 ± 0.02	0.92 ± 0.04	0.84 ± 0.03	0.88 ± 0.02	11.00 ± 0.73

Table 2. Comparison of zero-shot SAM performance with previously reported supervised DeepLabv3+ segmentation results for RLM mineral images (Filippo et al., 2021; DL_8_2 model variant). SAM F1-scores are reported as mean $\pm$ SD, while the DeepLabv3+ values correspond to the supervised benchmark reported by Filippo et al. (2021). Relative F1 values are expressed as percentages of the corresponding DeepLabv3+ benchmark.

Method	Training paradigm	FeM F1-score	Cu F1-score	FeM relative F1 (%)	Cu relative F1 (%)
DeepLabv3+	Fully supervised	0.93	0.92	100	100
SAM ViT-B	Zero-shot	0.77 $\pm$ 0.07	0.87 $\pm$ 0.05	83	94
SAM ViT-L	Zero-shot	0.87 $\pm$ 0.02	0.86 $\pm$ 0.04	93	93
SAM ViT-H	Zero-shot	0.88 $\pm$ 0.02	0.86 $\pm$ 0.05	94	93

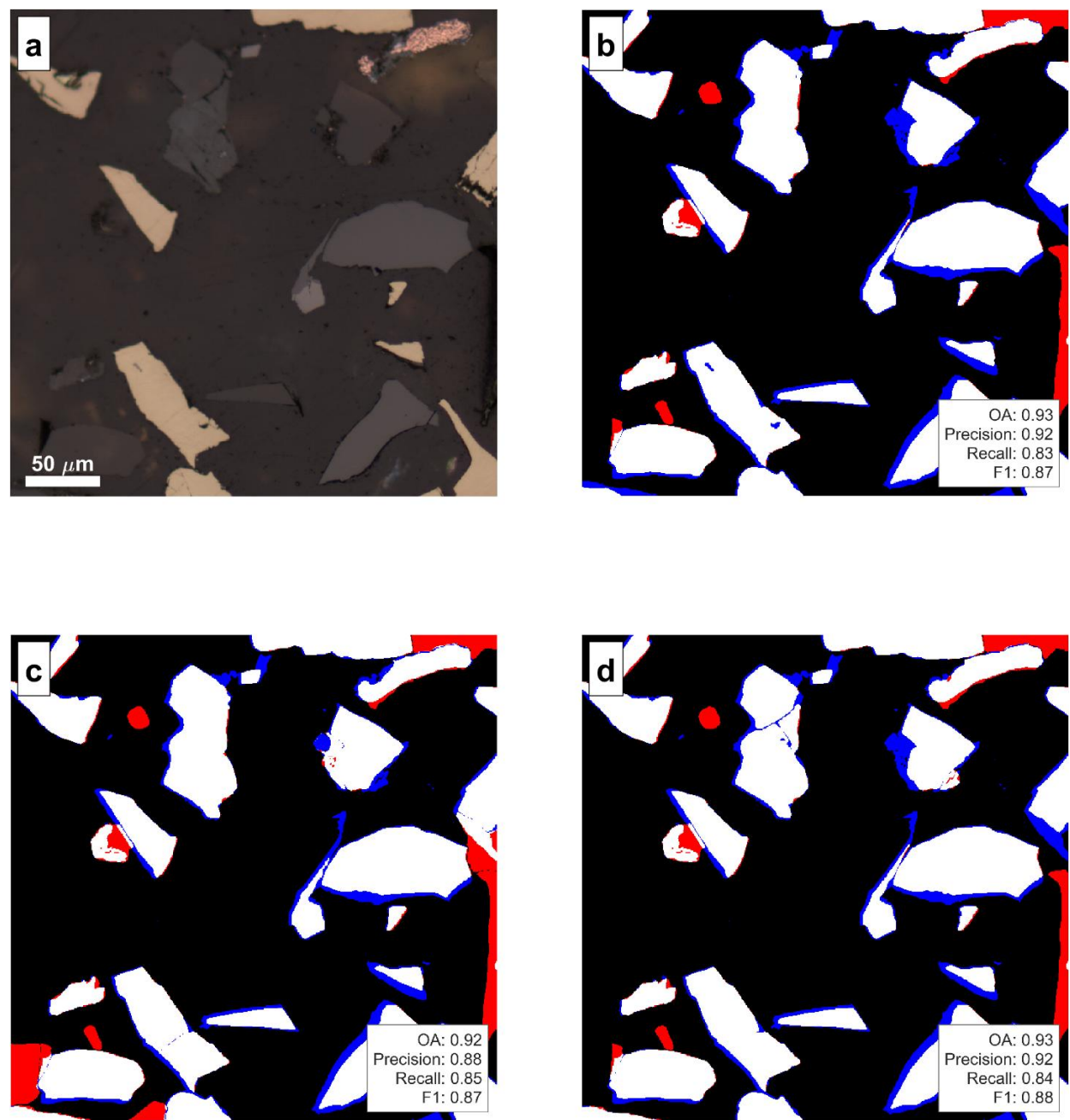


Fig. 5. Qualitative comparison of SAM segmentation results for a representative Cu dataset image: (a) RLM image, (b) ViT-B result, (c) ViT-L result, and (d) ViT-H result. White pixels indicate true positives, black pixels true negatives, blue pixels false negatives, and red pixels false positives relative to the reference mask.

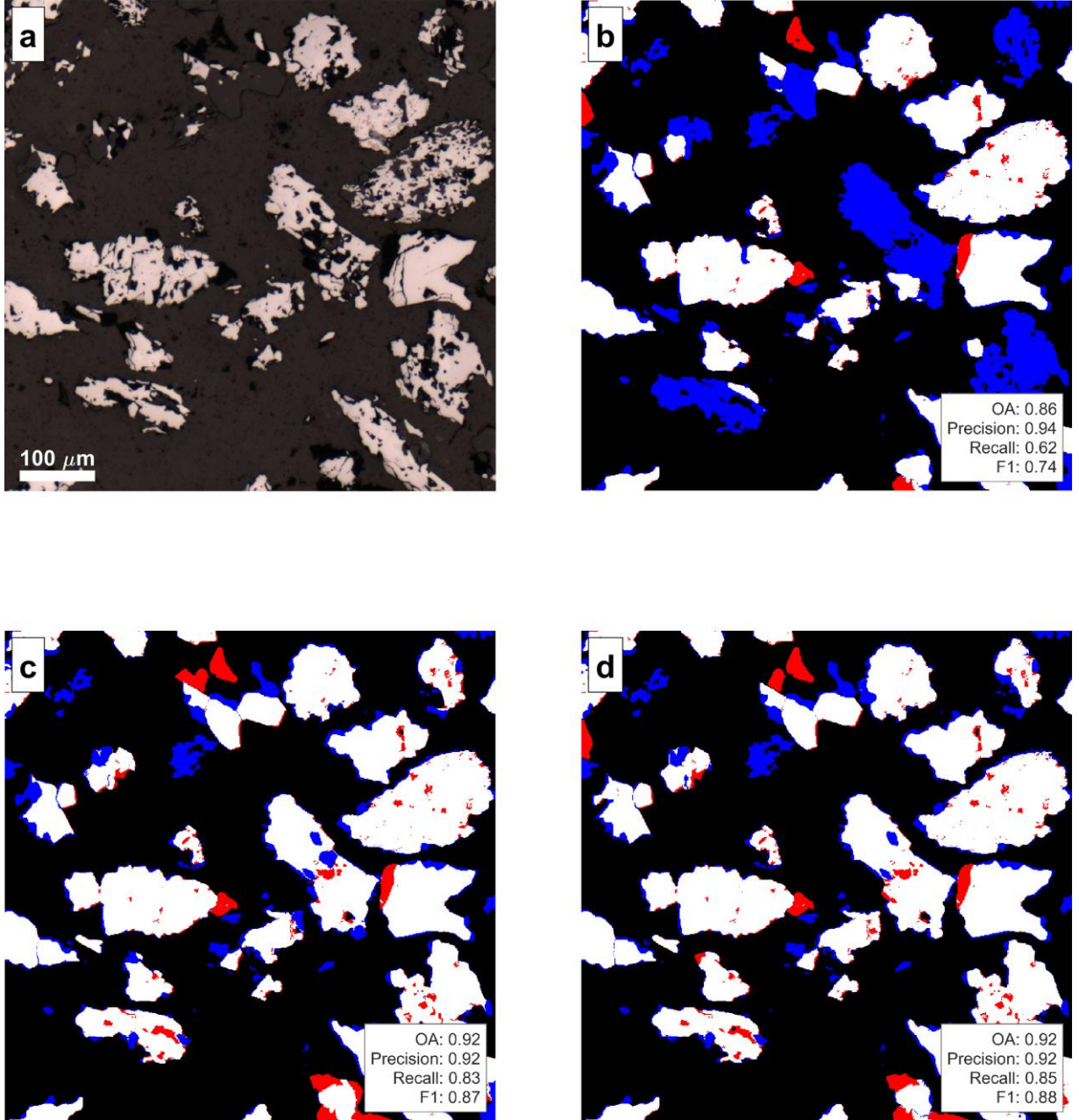


Fig. 6. Qualitative comparison of SAM segmentation results for a representative FeM dataset image: (a) RLM image, (b) ViT-B result, (c) ViT-L result, and (d) ViT-H result. White pixels indicate true positives, black pixels true negatives, blue pixels false negatives, and red pixels false positives relative to the reference mask.

Case study II

Representative segmentation results obtained from the three SAM encoders using RGB and PCA preprocessing are shown in Figs. 7 and 8, respectively. Grain masks were generated for all encoder-preprocessing combinations. Compared with the RGB results presented in Fig. 7, the PCA results provided greater apparent local contrast and resulted in a greater number of large grain regions. These differences can be observed across all three SAM encoder variants.

The stereological descriptors derived from the gradient-based grain-boundary maps and SAM-generated segmentations are summarized in Table 3. The median

intercept count was 83,734 [81,423, 85,671] for the gradient-based method. The SAM-based methods produced lower intercept counts, ranging from 36,938 [32,137, 38,951] for RGB ViT-B to 62,818 [56,268, 63,223] for PCA ViT-H.

Median L_a was 9.73 [8.83, 10.12] pixels for the gradient-based method. Among the SAM-based methods, median L_a ranged from 9.39 [9.18, 9.94] pixels for RGB ViT-B to 11.77 [11.21, 11.97] pixels for PCA ViT-H.

Median L_b was 9.50 [8.66, 9.70] pixels for the gradient-based method and ranged from 8.62 [8.48, 9.27] pixels for RGB ViT-B to 11.10 [10.48, 11.43] pixels for PCA ViT-H.

Φ showed variability among individual images but no consistent method-dependent shift across segmentation approaches.

Median bireflectance was 16.54 [16.23, 17.06] for the gradient-based method and ranged from 13.87 [13.47, 14.11] for PCA ViT-H to 16.93 [16.60, 17.56] for RGB ViT-B (Table 3).

Representative intercept-length statistics derived from the gradient-based method and SAM-generated segmentations are shown in Fig. 9 for one image (12 tiles), while elliptical anisotropy representations derived from omnidirectional intercept measurements are shown in Fig. 10 for the same image. Differences in orientation visible in this representative image reflect image-specific variation and are not indicative of a systematic shift across the complete image set. The greater contribution of short intercepts for the gradient-based method is apparent in Fig. 9a, while the corresponding shift toward longer intercept lengths for the SAM-derived segmentations is visible in Fig. 9b. This trend is consistent across the complete dataset and is apparent in the median L_a and L_b values in Table 3.

Image-level Friedman tests revealed significant overall effects of segmentation method on L_a ($p = 1.11$

$\times 10^{-6}$, Kendall's $W = 0.91$), L_b ($p = 2.21 \times 10^{-6}$, $W = 0.87$), and bireflectance ($p = 1.47 \times 10^{-6}$, $W = 0.89$). In contrast, no significant overall effect of segmentation method was observed for Φ ($p = 0.58$, $W = 0.11$).

Post hoc comparisons showed no significant differences between the gradient-based method and RGB ViT-B, RGB ViT-L, RGB ViT-H, or PCA ViT-B for L_a , L_b , or bireflectance. Significant differences relative to the gradient-based method were confined to the larger PCA variants. For PCA ViT-L, median L_a was 17% higher than for the gradient-based method (11.38 versus 9.73 pixels; $p = 0.04$), while median bireflectance was 13% lower (14.33 versus 16.54; $p = 0.04$). Its 12% higher median L_b (10.64 versus 9.50 pixels) was not statistically significant ($p = 0.13$).

The largest deviations from the gradient-based method were observed for PCA ViT-H. Median L_a was 21% higher (11.77 versus 9.73 pixels; $p = 0.0016$), and median L_b was 17% higher (11.10 versus 9.50 pixels; $p = 0.0070$). Median bireflectance was 16% lower for PCA ViT-H (13.87 versus 16.54; $p = 0.00093$). No post hoc comparisons were performed for Φ because the overall Friedman test was not significant.

Table 3. Image-level stereological descriptors from the gradient-based grain-boundary maps and SAM-generated segmentations of polarized dolomite. Values are median [interquartile range] across seven images, after first taking the median across the 12 tiles within each image. L_a and L_b denote the major and minor intercept lengths, respectively; Φ denotes the principal anisotropy orientation. Porosity, available only for the gradient-based method, was 5.32% [4.92%, 5.73%].

Method	L_a (pixels)	L_b (pixels)	Φ (radians)	Bireflectance	Intercept count
Gradient-based method	9.73 [8.83, 10.12]	9.50 [8.66, 9.70]	1.84 [1.14, 2.26]	16.54 [16.23, 17.06]	83,734 [81,423, 85,671]
RGB ViT-B	9.39 [9.18, 9.94]	8.62 [8.48, 9.27]	1.89 [1.69, 2.00]	16.93 [16.60, 17.56]	36,938 [32,137, 38,951]
RGB ViT-L	10.10 [9.67, 10.86]	8.97 [8.72, 9.61]	1.85 [1.71, 1.94]	16.02 [15.75, 16.86]	43,564 [37,155, 45,817]
RGB ViT-H	10.53 [10.10, 11.14]	9.78 [9.29, 10.34]	1.80 [1.58, 1.90]	15.45 [14.59, 15.61]	46,111 [40,168, 47,864]
PCA ViT-B	10.23 [9.97, 10.67]	9.92 [9.32, 10.16]	1.75 [1.53, 2.04]	15.37 [15.19, 15.77]	48,334 [44,486, 49,240]
PCA ViT-L	11.38 [10.74, 11.53]	10.64 [10.22, 11.09]	1.78 [1.49, 2.10]	14.33 [13.81, 14.62]	59,200 [53,377, 61,116]
PCA ViT-H	11.77 [11.21, 11.97]	11.10 [10.48, 11.43]	1.67 [1.51, 1.99]	13.87 [13.47, 14.11]	62,818 [56,268, 63,223]

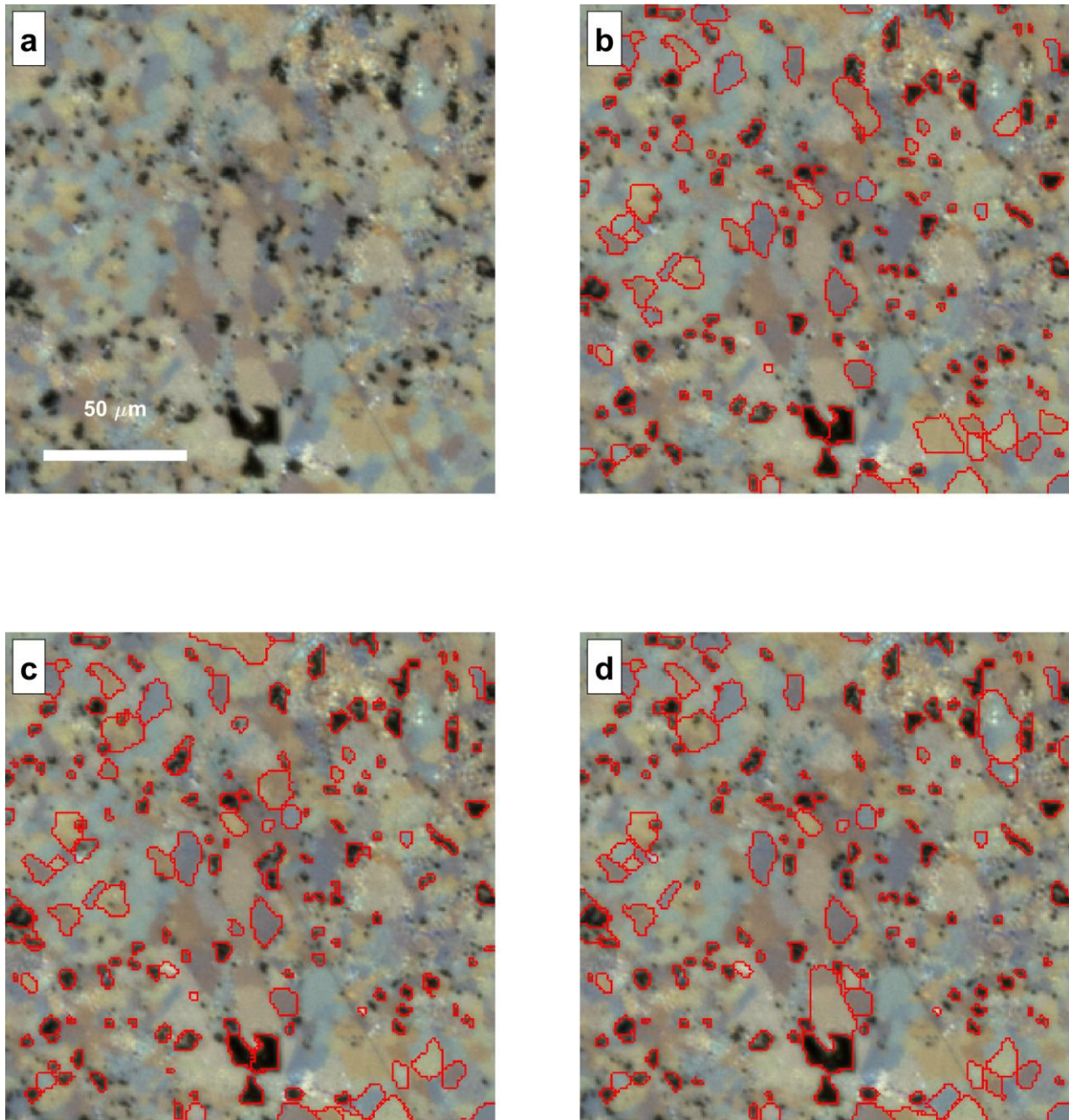


Fig. 7. SAM segmentation results obtained from RGB projection preprocessing of a representative polarized dolomite image stack: (a) RGB-projected input image, (b) ViT-B result, (c) ViT-L result, and (d) ViT-H result. Red contours indicate grain boundaries extracted from the SAM-generated masks.

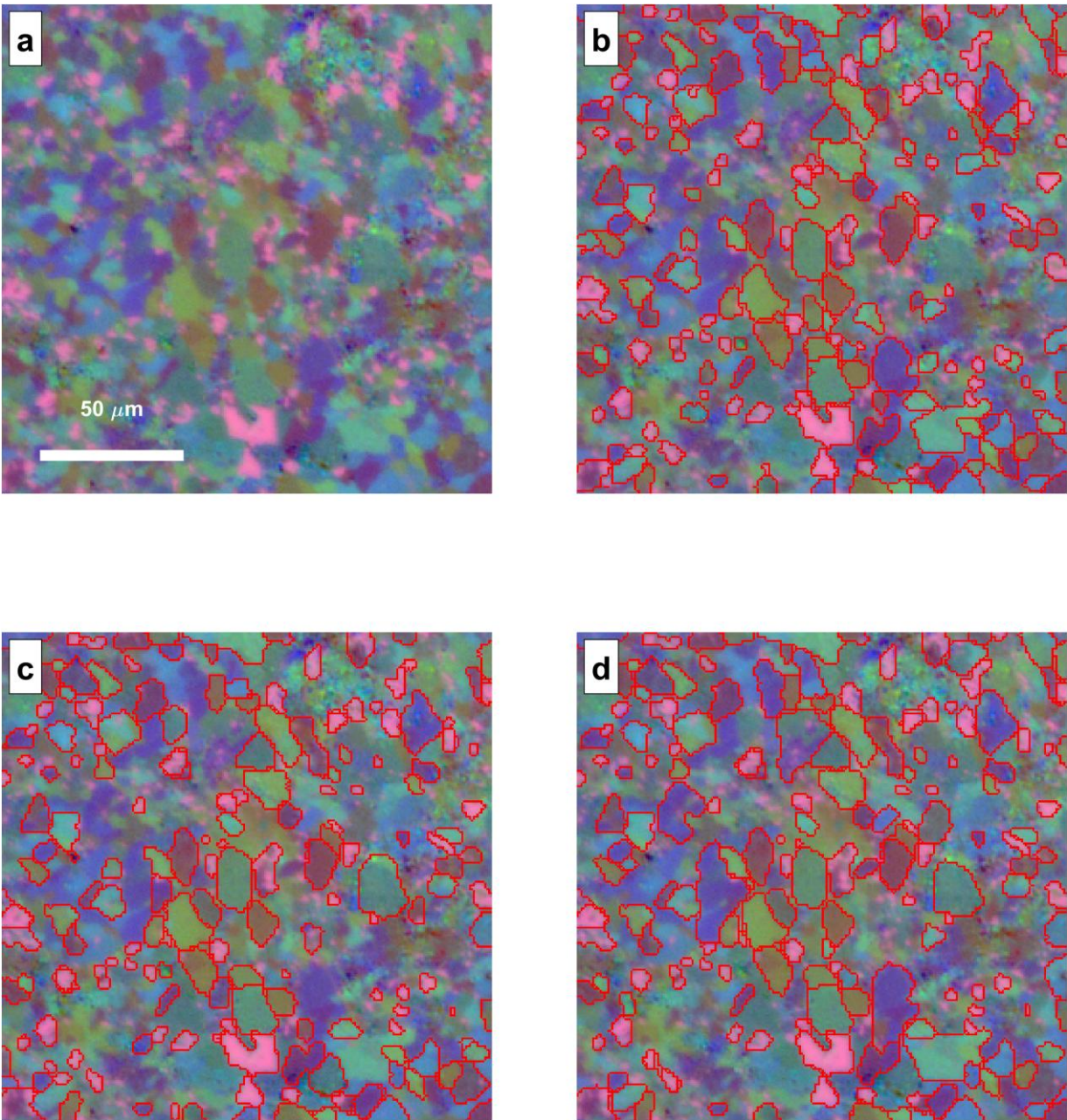


Fig. 8. SAM segmentation results obtained from PCA preprocessing of a representative polarized dolomite image stack: (a) PCA-preprocessed input image, (b) ViT-B result, (c) ViT-L result, and (d) ViT-H result. Red contours indicate grain boundaries extracted from the SAM-generated masks.

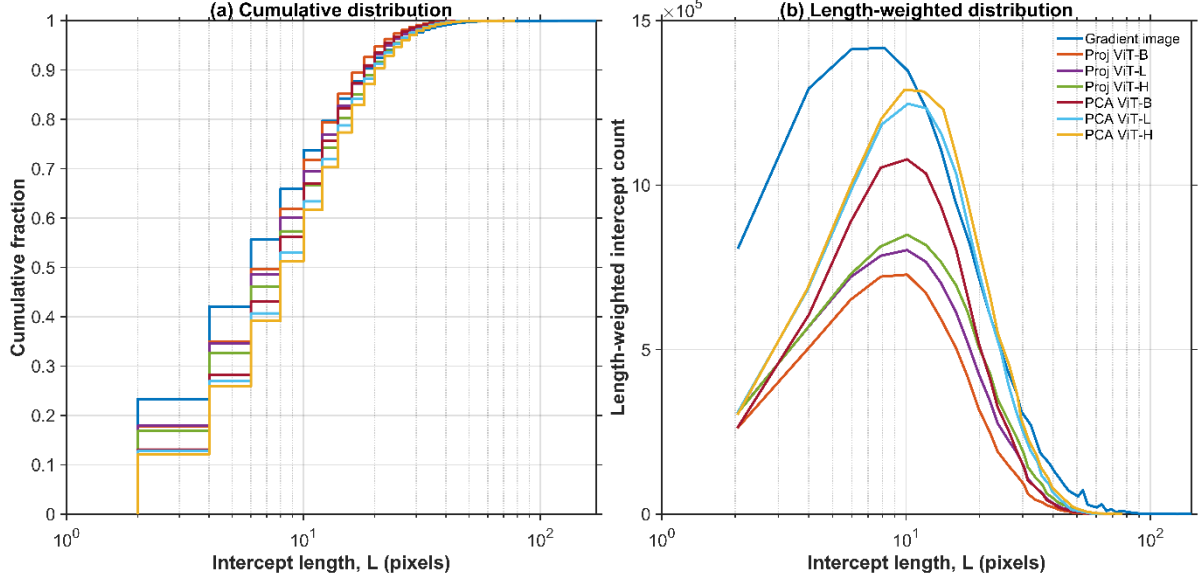


Fig. 9. Comparison of intercept-length statistics derived from the gradient-based grain-boundary map and SAM-generated segmentations of a representative image from the dataset: (a) cumulative omnidirectional intercept-length distributions and (b) length-weighted intercept-length distributions, where each intercept is weighted by its length in pixels.

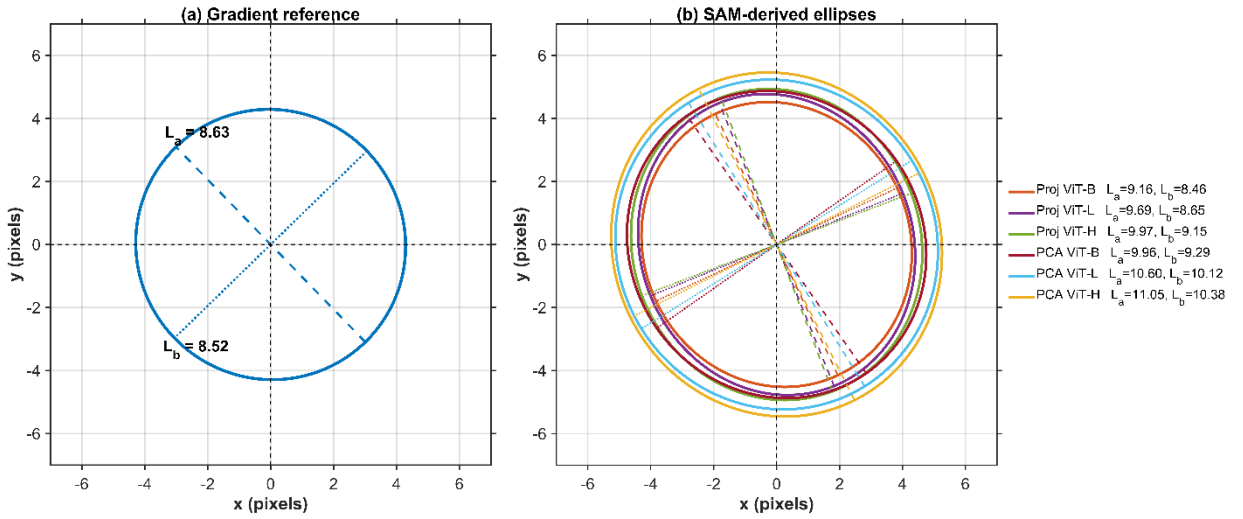


Fig. 10. Elliptical anisotropy representations derived from omnidirectional intercept measurements for the same representative image shown in Fig. 9: (a) anisotropy ellipse obtained from the gradient-based grain-boundary map and (b) corresponding anisotropy ellipses derived from SAM-generated segmentations using RGB projection and PCA preprocessing with the ViT-B, ViT-L, and ViT-H encoders.

DISCUSSION

For the Cu dataset in case study I, all SAM encoder variants produced consistently high segmentation accuracy, with mean F1-scores ranging between 0.86 and 0.87. In contrast, performance on the FeM dataset depended strongly on encoder capacity. The ViT-B encoder exhibited pronounced under-segmentation,

characterized by high precision but substantially reduced recall. Increasing encoder capacity to ViT-L and ViT-H resulted in marked improvements in recall and F1-score, indicating that, for the FeM benchmark, the larger SAM encoders were better able to recover the target regions represented in the reference masks.

Comparison with the supervised DeepLabv3+ framework reported by Filippo *et al.* (2021) further

highlights the potential of foundation-model approaches. Although supervised models achieved slightly higher peak segmentation performance when trained and evaluated within the same domain, their study demonstrated a pronounced loss of performance when applied to previously unseen mineralogical domains, with substantial reductions in F1-score during cross-domain validation experiments (Filippo *et al.*, 2021). Recovery of this lost performance would require additional domain-adaptation procedures and access to target-domain data. In contrast, SAM achieved competitive segmentation performance immediately in a zero-shot setting, without task-specific training, fine-tuning, or domain adaptation (Table 2).

In case study II, strong overall method effects were observed for the derived stereological descriptors, although post hoc comparisons demonstrated that these effects depended on the encoder-preprocessing combination. In particular, no significant differences from the gradient-based method were detected for the ViT-B configurations, whereas the largest differences in L_a , L_b , and bireflectance were observed for the PCA ViT-H configuration. Relative to the gradient-based method, PCA ViT-H produced a 21% greater median L_a , a 17% greater median L_b , and a 16% lower median bireflectance. In contrast, no significant method-dependent difference was observed in principal anisotropy orientation.

Inspection of the grain-boundary maps and SAM-generated masks suggests that these differences can be explained by the different nature of the intermediate representations. The gradient-based maps contain numerous short and fragmented grain-boundary features (Fig. 4), whereas the SAM-derived masks represent discrete grain regions whose number and size depend on preprocessing and ViT encoder (Figs. 7 and 8). These differences propagate into the intercept-length distributions, as illustrated in Fig. 9, where the gradient-based method contains a larger contribution from short intercepts, with the length-weighted distribution peaking at approximately 5-6 pixels, whereas the SAM-derived distributions peak at approximately 10 pixels. The resulting shift toward longer intercepts is consistent with the greater L_a and L_b obtained for some SAM configurations and is also qualitatively apparent when comparing the segmentation results in Figs. 7 and 8 and in Table 3. Interestingly, the principal anisotropy orientation was not significantly different between methods, suggesting that the dominant directional fabric was preserved despite differences in the scale and continuity of the segmented grain structures.

The sensitivity of SAM-derived grain representations to preprocessing raises the question of how preprocessing interacts with model adaptation. Efficient fine-tuning methods, including adapters and low-rank

adaptation (LoRA), are used to specialize foundation models for specific applications (Hu *et al.*, 2021). The present results suggest that preprocessing and model adaptation may interact in determining segmentation performance in this setting. Further work is needed to examine these effects.

The two case studies demonstrate complementary aspects of zero-shot segmentation in mineralogical microscopy. Case study I shows that SAM can achieve high pixel-wise segmentation performance without mineralogy-specific training. Case study II shows that SAM-derived segmentations can support stereological characterization in the absence of grain-level ground truth, using a gradient-based method as a reference. Because this method is not independent ground truth, differences in stereological descriptors should not necessarily be interpreted as errors in the SAM measurements; they may instead reflect alternative representations of the grain structure. Future work should compare optimized classical, supervised deep-learning, and foundation-model approaches using independent expert grain-level annotations across mineralogical domains.

Several limitations should be acknowledged, however. First, all experiments were conducted in a purely zero-shot setting without prompt engineering, fine-tuning, or domain-specific adaptation and therefore do not represent the maximum achievable performance of foundation-model-based segmentation.

The present study also addresses only a limited range of the structural features encountered in mineralogical imaging, specifically the segmentation and grain-boundary characteristics investigated in the two case studies. Broader evaluation across different minerals, ore types, textures, imaging conditions, and structural features is required to establish its range of applicability.

Additionally, because the six SAM encoder-preprocessing combinations were treated as individual segmentation configurations in the statistical analysis, the present analysis does not separate the individual effects of encoder architecture and preprocessing strategy or their potential interaction. A factorial analysis would be required to distinguish these effects.

Finally, the pixel-wise metrics used in case study I should be interpreted with care. Such scores can appear strong even when grain boundaries are misaligned, over-smoothed, or interrupted by small gaps or bridges, errors that can distort grain counts and phase-contact statistics (Taha and Hanbury, 2015; Reinke *et al.*, 2024).

In conclusion, SAM demonstrated strong potential for mineralogical image analysis, although it was originally developed as a general-purpose segmentation model rather than specifically for this domain (Kirillov *et al.*, 2023). Case study I showed that SAM can achieve competitive mineral segmentation in a fully zero-shot

setting, without task-specific training, fine-tuning, or domain adaptation. Case study II showed that the same zero-shot approach can be extended to grain segmentation for downstream stereological analysis when grain-level ground-truth segmentation is unavailable. The resulting stereological descriptors depended on the segmentation configuration, while principal anisotropy orientation remained comparatively stable across methods. Furthermore, the dependence of the results on preprocessing demonstrates that the SAM-derived grain representation can be influenced before segmentation, providing scope for further optimization. Given the scarcity and cost of annotated mineralogical datasets (Álvarez Iglesias *et al.*, 2019; Filippo *et al.*, 2021), the ability to obtain both useful segmentations and quantitative structural information without domain-specific training makes SAM a particularly promising tool for reflected-light mineralogical microscopy.

DATA AVAILABILITY

The data supporting the findings of this study, together with all source code and scripts required to reproduce the analyses and figures, are publicly available at <https://github.com/adriaancampo/SAMforRLM>.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- Álvarez Iglesias JC, Gomes OFM, Paciornik S (2011). Automatic recognition of hematite grains under polarized reflected light microscopy through image analysis. *Miner Eng* 24(12):1264-70.
- Álvarez Iglesias JC, Santos RBM, Paciornik S (2019). Deep learning discrimination of quartz and resin in optical microscopy images of minerals. *Miner Eng* 138:79-85.
- Archit A, Freckmann L, Nair S, Khalid N, Hilt P, Rajashekar V, Freitag M, Teuber C, Spitzner M, Tapia Contreras C, Buckley G, von Haaren S, Gupta S, Grade M, Wirth M, Schneider G, Dengel A, Ahmed S, Pape C (2025). Segment Anything for Microscopy. *Nat Methods* 22(3):579-91.
- Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T, Natarajan V, Norouzi M (2021). Big self-supervised models advance medical image classification. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3478-88.
- Badrinarayanan V, Kendall A, Cipolla R (2017). SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(12):2481-95.
- Beucher S (1992). The watershed transformation applied to image segmentation. *Scanning Microsc Suppl* 6:299-314.
- Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, et al. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*.
- Canny J (1986). A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8(6):679-698.
- Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018). DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 40(4):834-48.
- Cruz-Orive LM (2017). Stereology: a historical survey. *Image Anal Stereol* 36(3):153-77.
- De Castro B, Benzaazoua M, Roychowdhury S, Chopard A, Quintal Lauzon F, Plante B (2024). Novel technique for the preparation and analysis of powder-based polished sections by automated optical mineralogy: Part 2 - use of deep learning approach for transparent mineral detection. *Miner Eng* 206:108508.
- Delesse A (1848). Procédé mécanique pour déterminer la composition des roches. *Ann Mines* 13:379-88.
- Dunn OJ (1961). Multiple comparisons among means. *J Am Stat Assoc* 56(293):52-64.
- Fandrich R, Gu Y, Burrows D, Moeller K (2007). Modern SEM-based mineral liberation analysis. *Int J Miner Process* 84(1-4):310-20.
- Felzenszwalb PF, Huttenlocher DP (2004). Efficient graph-based image segmentation. *Int J Comput Vis* 59(2):167-81.
- Filippo MP, Gomes OFM, da Costa GAOP, Mota GLA (2021). Deep learning semantic segmentation of opaque and non-opaque minerals from epoxy resin in reflected light microscopy images. *Miner Eng* 170:107007.
- Frayssinet L, Grinberg P, Grecco HE, Comerio MA, Domené EA (2026). Automated segmentation using computer vision applied to petrographic thin sections of unconventional tight gas sandstone reservoirs. *Discov Geosci* 4:57.
- Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32(200):675-701.
- Gomes OFM, Paciornik S, Filippo MP, da Costa GAOP, Mota GLA (2021a). Cu dataset - A copper ore labeled images dataset for segmentation training and testing. Zenodo. doi:10.5281/zenodo.5020566.

- Gomes OFM, Paciornik S, Filippo MP, da Costa GAOP, Mota GLA (2021b). FeM dataset - An iron ore labeled images dataset for segmentation training and testing. Zenodo. doi:10.5281/zenodo.5014700.
- Gu Y (2003). Automated scanning electron microscope based mineral liberation analysis: an introduction to JKMRC/FEI mineral liberation analyser. *J Miner Mater Charact Eng* 2(1):33-41.
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021). LoRA: low-rank adaptation of large language models. arXiv:2106.09685.
- Jaimes Contreras R, Pilawski D, Califice A, Pirard E (2010). Quantitative microtexture analysis of carbonate rocks using bireflectance imaging. In: *Proceedings of IAMG 2010; Budapest, Hungary*.
- Kendall MG, Smith BB (1939). The problem of m rankings. *Ann Math Stat* 10(3):275-87.
- Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R (2023). Segment Anything. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4015-26.
- Koh EJY, Amini E, McLachlan GJ, Beaton N (2021). Utilising convolutional neural networks to perform fast automated modal mineralogy analysis for thin-section optical microscopy. *Miner Eng* 173:107230.
- Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, Houlsby N (2020). Big Transfer (BiT): general visual representation learning. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, 491-507.
- Li C, Han X, Yao C, Guo Y, Li Z, Jiang L, Liu W, Huang H, Fu H, Ban X (2025). A novel training-free approach to efficiently extracting material microstructures via visual large model. *Acta Mater* 290:120962.
- Otsu N (1979). A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9(1):62-6.
- Pérez-Barnuevo L, Pirard E, Castroviejo R (2013). Automated characterisation of intergrowth textures in mineral particles: a case study. *Miner Eng* 52:136-42.
- Pirard E (2004). Multispectral imaging of ore minerals in optical microscopy. *Mineral Mag* 68(2):323-33.
- Pirard E, Lebichot S, Krier W (2007). Particle texture analysis using polarized light imaging and grey level intercepts. *Int J Miner Process* 84(1-4):299-309.
- Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, eds. (2009). *Dataset shift in machine learning*. Cambridge, MA: MIT Press.
- Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE, Rädtsch T, Sudre CH, Acion L, Antonelli M, Arbel T, Bakas S, Benis A, Buettner F, Cardoso MJ, Cheplygina V, Chen J, Christodoulou E, Cimini BA, Farahani K, et al. (2024). Understanding metric-related pitfalls in image analysis validation. *Nat Methods* 21(2):182-94.
- Ronneberger O, Fischer P, Brox T (2015). U-Net: convolutional networks for biomedical image segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234-41.
- Schulz B, Sandmann D, Gilbricht S (2020). SEM-based automated mineralogy and its application in geo- and material sciences. *Minerals* 10(11):1004.
- Taha AA, Hanbury A (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 15:29.
- Underwood EE (1970). *Quantitative stereology*. Reading, MA: Addison-Wesley.
- Vellappally A, Hou S, Emmings J (2024). Automated grain segmentation and mineral classification in rock thin sections. In: *Proceedings of the EAGE Annual Conference & Exhibition*, 1-5.