DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

Revision 1

The application of "transfer learning" in optical microscopy: the petrographic classification of metallic minerals

Word Count: 5753

Yi-Wei Cai[1], Kun-Feng Qiu[1*], Maurizio Petrelli[2], Zhao-Liang Hou[3], M. Santosh[1,4], Hao-Cheng Yu[1], Ryan T. Armstrong[5], Jun Deng[1, 6]

[1] *Frontiers Science Center for Deep-time Digital Earth, State Key Laboratory of Geological Processes and Mineral Resources, School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China*

[2] *Department of Physics and Geology, University of Perugia, Perugia 06100, Italy*

[3] *Department of Geology, University of Vienna, Vienna 1090, Austria*

[4] *Department of Earth Science, University of Adelaide, Adelaide SA 5005, Australia*

[5] *School of Minerals and Energy Resources Engineering, University of New South Wales, Sydney, NSW 2052, Australia*

[6] *Geological Research Institute of Shandong Gold Group Co., Ltd., Jinan 250013, China*

*Corresponding author

Kun-Feng Qiu kunfengqiu@qq.com

Professor, China University of Geosciences, Beijing

No. 29 Xueyuan Road, Haidian District, Beijing, 100083, P.R. China

Abstract

Analysis of optical microscopic image data is crucial for the identification and characterization of mineral phases, and thus directly relevant to the subsequent methodology selections of further detailed petrological exploration. Here we present a novel application of Swin Transformer, a deep learning algorithm to classify metal mineral phases such as arsenopyrite, chalcopyrite, gold, pyrite, and stibnite, in images captured by optical microscopy. To speed up the training process and improve the generalization capabilities of the investigated model, we adopt the "transfer learning" paradigm by pretraining the algorithm using a large, general-purpose, image dataset named ImageNet-1k. Further, we compare the performances of the Swin Transformer with those of two well-established Convolutional Neural Networks (CNNs) named MobileNetv2 and ResNet50, respectively. Our results highlight a maximum accuracy of 0.92 for the Swin Transformer, outperforming the CNNs. To provide an interpretation of the trained models, we apply the so-called Class Activation Map (CAM), which points to a strong global feature extraction ability of the Swin Transformer metal mineral classifier that focuses on distinctive (e.g., colors) and microstructural (e.g., edge shapes) features. The results demonstrate that the deep learning approach can accurately extract all available attributes, which reveals the potential to assist in data exploration and provides an opportunity to carry out spatial quantization at a large scale (cm-mm). Simultaneously, boosting the learning processes with pre-trained weights can accurately capture relevant attributes in mineral classification, revealing the potential for application in mineralogy and petrology, as well as enabling its use in resource explorations.

Keywords: Swin Transformer metal mineral classifier; Microscopy images; Transfer learning; Deep learning; Class Activation Map

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

Introduction

Petrographic studies at the microscopic scale and mineral identification constitute the fundamental step in many geological studies (e.g., igneous, metamorphic and sedimentary petrology or mineral exploration) and industrial productions (Schrader and Zega 2019; Deng et al. 2020a; dos Anjos et al. 2021; Sheldrake and Higgins 2021; Azeuda Ndonfack et al. 2022). In petrographic investigations at microscopic scale, the first step mostly relies on optical microscopy involving the identification of mineral phases and textures (Su et al. 2020; Leichter et al. 2022; Faria et al. 2022; Qiu et al. 2023c). In recent years, optical microscopic observations are further supported by more advanced techniques, like electron-based imaging and X-ray techniques (Fu and Aldrich 2019). Meanwhile, many software like ImageJ (Schneider et al. 2012) or scripting languages, such as Python (Petrelli 2021) or MATLAB (Trauth et al. 2007) now support quantitative petrographic investigations like the segmentation processes or crystal size distribution analyses (Santosh et al. 2009; Tarquini and Favalli 2010; Jungmann et al. 2014; Y. Wang et al. 2021; Zhang et al. 2021). Despite the recent analytical advancements and the possibilities of automation in quantitative petrographic studies, the initial investigation of new samples still relies, mostly, on the manual identification of mineral phases by expert petrologists, by optical microscopy (Deng et al. 2020b). This procedure is time-consuming, often subjective, and sometimes biased since many minerals share similar textural and optical properties (Santosh 2010; Młynarczuk et al. 2013; Xu et al. 2021; Zhong et al. 2021; Qiu et al. 2023b).

In the framework detailed above, the development of automatic identification techniques can significantly support the handling and processing of large raw microscopy images (Alférez et al. 2021; Faria et al. 2022). To achieve this goal, the use of Machine Learning (ML) techniques deserves attention, since these have been successfully applied in many fields of visual data investigations (Petrelli and Perugini 2016; Endert et al. 2017; Acosta et al. 2019; Y. D. Wang et al. 2021; Zhou et al. 2022; Qiu et al. 2023a).

Among ML techniques, the developments of deep learning algorithms have drastically boosted the application of the Artificial Intelligence (AI) in many scientific fields, including image analysis and processing (Xing et al. 2018; Zhichao Liu et al. 2021). Examples are image classification (Obaid et al. 2020), object detection (Zhao et al. 2019; Wu et al. 2020) and image segmentation (Ghosh et al. 2019; Leichter et al. 2022; Tang et al. 2022). In particular, the recent development of new network algorithms in natural language processing (NLP) favored the growth of an architecture named Transformer (Vaswani et al. 2017). Transformers are at the base of the so-called foundation models (Bommasani et al. 2021), implementing the concept of "transfer learning" (Thrun and Mitchell 1995; Polat et al. 2021). The idea behind "transfer learning" is to use the "knowledge" that is learned from one task, and apply it to solve a different problem. In deep learning, the transfer learning is often achieved by the so-called "pretraining" (Bommasani et al. 2021) on large data set. More specifically, a deep learning model is typically trained to solve a non-specific task, and then adapted to the problem of interest through fine-tuning, drawn by a specific and more focused data set (Bommasani et al. 2021).

In this study, we investigate the application of the "transfer learning" paradigm using a Transformer known as "Swin Transformer" (Ze Liu et al. 2021). The investigated Transformer has been previously pre-trained using a large, general-purpose, public computer vision dataset. Our main aim is to tap the benefit from the "transfer learning" paradigm by fine-tuning the "Swin Transformers" in the classification of five metallic minerals (i.e., arsenopyrite, chalcopyrite, gold, pyrite and stibnite) in optical microscopy images. To achieve this goal, we set up and trained the Swin Transformer plus two widely-used Convolutional Neural Networks (CNNs) models. Then, we evaluated and compared the performances of each model. Finally, we used a feature visualization technique named Class Activation Map (Zhou et al. 2016) to attempt at understanding the internal behavior of the investigated models, which is often perceived as a “black box” (Castelvecchi 2016).

Materials

Raw images were captured using optical microscopes (AXIOSCOPE-A1, Leica DM4P, and Olympus BX51) through employing cellSens Entry and Stream Essentials software under reflected light conditions. The raw data resolutions varied from 1608*1608, 1936*1216, and to 4800*3600 pixels. The process of collecting images involved several manual steps. Firstly, the thin-sections were observed through optical microscopy, the target minerals were located, and the mineral images were captured, particularly focusing on grains larger than 10 μm in width to ensure accurate mineral identification under the microscope. Notably, the collected images may show different colors for the same mineral, whereas the same/similar colors are seen for different minerals. This discrepancy arises because the original data are from different deposits and the thin-sections vary in white balance and brightness.

Data composition

Microscopy images of arsenopyrite, chalcopyrite, gold, pyrite, and stibnite were selected as the research material for our target aimed at classification. The dataset consists of 481 optical microscopy images of five unprocessed metal minerals from two gold deposits in the Jiaodong Peninsula of North China (Linglong, and Longkou gold deposits) and six gold deposits in West Qinling in Central China (Jiagantan, Liba, Xiakanmucang, Yidinan, and Zaorendao gold deposits; and the Zaozigou gold stibnite deposit). A total of 159 arsenopyrite images, 128 chalcopyrite

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

images, 159 gold images, 145 pyrite images and 131 stibnite images of different sizes were used as the raw data. The characteristics of these five types of metallic minerals under microscope – that is, the information that manual classification uses–are given in Table 1.

Dataset characteristics

As stated above, metal minerals were imaged using optical microscopy. The produced images contain basic information that characterizes each phase, i.e., the reflected color, microstructural characteristics, and mineral paragenesis. These characteristics constitute the building blocks for the classification of the metal mineral phase with direct observation.

In detail, gold (Au; Figures 1a-d) is bright yellow under the microscope, with relatively smooth edges. Gold mainly coexists with arsenopyrite, pyrite, chalcopyrite and stibnite. Pyrite (FeS_2; Figures 1e-h) is a homogeneous mineral appearing light yellow under the microscope with the edges of the grains being relatively smooth, as for gold. It is widely distributed with a number of minerals especially gold, arsenopyrite and chalcopyrite. In addition, the images show that chalcopyrite ($CuFeS_2$; Figures 1i-l) is characterized by copper-yellow, weakly homogeneous, and often heteromorphic granular aggregates. The reflective color always shows yellow-green. Under the microscope, chalcopyrite has broken edges. Arsenopyrite (FeAsS; Figures 1m-p) is bright white with cream, yellow, or red color (i.e., is weakly polychromatic) and radial aggregates can be observed. Arsenopyrite has smooth edges and is often arsenic-bearing pyrite and arsenopyrite. Stibnite (Sb_2S_3; Figures 1q-t) of the white or light off-white variants can be easily confused with arsenopyrite. Similar to arsenopyrite, the strongly homogenous stibnite coexisted with pyrite with smooth edges under the microscope.

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

Gold, pyrite, and chalcopyrite are all yellow under the microscope and tend to coexist in gold deposits. The gold and pyrite grains in the dataset have relatively smooth edges, whereas chalcopyrite has broken boundaries. The arsenopyrite and stibnite all have a reflective color of gray with relatively smooth edges and similar mineral morphology.

From the above-mentioned features, it is obvious that large-scale manual classification by observing these metal minerals with the naked eye (or directly under the microscope) poses a significant challenge. The similarity in reflected colors and morphology complicates their classification, and would consume long of time with manual studies, especially when dealing with large volumes of images for examination, such as in batch studies.

Methods

Swin Transformer and Convolutional Neural Networks

Swin Transformer. In image analysis, "Transformers" (Dosovitskiy et al. 2020) rely on a self-attention mechanism to model the correlation between various regions within an image. The self-attention mechanism, often referred to as scaled dot-product attention, stands as a fundamental concept in the field of deep learning, allowing the model to gauge the significance of distinct elements in an input sequence, dynamically regulating their impact on the output. Compared with the local receptive field mechanism of convolution used in CNNs, transformers can learn the correlation among relatively distant areas, and capture the long-distance dependence of the whole feature map, and therefore they are characterized by a high global modeling ability (Vaswani et al. 2017). The multi-head self-attention of the transformer that gives the model the ability to focus on different locations allows the model to learn relevant information in different subspaces and extract richer feature information (Devlin et al. 2019), thus alleviating the complexity of the neural network. The model does not need to input all the information into the neural network for calculation, but selectively enters some task-related information into the network. However, "plain" transformers require a large amount of computation for the training (Vyas et al. 2020). The Swin Transformer algorithm aims to improve upon this by using a hierarchical approach (shifted window) to decrease the cost of computing the self-attention which is exponential in the image size. Specifically, Swin Transformer computes self-attention on a local window which is then moved across the image, and also a multi-scale feature computation method. Swin Transformer has been used as a feature extraction network in various image tasks with good results (Jiang et al. 2022; Yang and Yang 2023).

In the present study, we optimize computational expenses by employing the Swin Transformer-base version (Ze Liu et al. 2021). The network architecture adopts a hierarchical design encompassing four stages (Figure 2). Initially, the RGB image undergoes the Patch Partition module, segmenting it into non-overlapping patches. Every pixel adjacent to 4*4 is a Patch. The image is then flattened across the color channel dimension, reshaping the color channel values into an elongated one-dimensional vector. This sequential vector captures the color information corresponding to each pixel. Subsequently, the channel data are transformed by a Linear Embedding layer. This layer is a technique for representing images as dense embedding vectors and these vectors capture visual features of the image. Following this, feature maps of varying sizes are constructed through four stages. Except that the image first passes through a Linear Embedding layer in stage 1, for the remaining three stages, the image is subsampled through a Patch Merging layer and then through the Swin Transformer Block. The Swin Transformer-base version is described in detail by Ze Liu et al. (2021).

Convolutional Neural Networks. Convolutional neural networks (CNNs) constitute a neural network architecture that is often used to extract features from image data. The Convolutional layer is the fundamental aspect of CNNs which involves the application of convolution operation to the input data and plays a crucial role in feature extraction from images with spatial relationships. The CNNs were first proposed by LeCun et al. (1989, 2015) for handwritten digital image recognition. In the present study, two general-purpose and widely-used CNN models have been investigated, namely ResNet50 and MobileNetv2 (Sandler et al. 2018).

ResNet50, a 50-layer variant of ResNet (He et al. 2015) operates through five processing stages (Figure 3). Stage 1 can be considered as a preprocessing step for the input images. In detail, for a three-channel RGB input image, it performs a preliminary feature extraction via 64 convolutional layers. Feature normalization is then carried out by a Batch Normalization (BN) layer that can convert interlayer outputs of a neural network into a standard format by subtracting the batch mean and then dividing by the batch's standard deviation, and effectively 'resets' the distribution of the output of the previous layer to be more efficiently processed by the subsequent layer (Chang and Chen 2015). Thus, the training convergence speed of the model is made faster and training process becomes more stable. Feature normalization is followed by a nonlinear feature mapping via the ReLU activation function layer and, next, a Maximum Pooling layer. Subsequently, the feature map size is further reduced to a quarter of the input image to reduce spatial information and parameters, increase computational speed, and reduce the risk of overfitting. In ResNet50, the remaining four stages have a similar structure, all of which are made up of different numbers of residual modules (Feng 2017). Finally, the extracted features pass through a fully connected (FC) layer to integrate feathers together for easy submission to the final classifier.

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

The MobileNetv2 is a common lightweight CNN characterized by the structure reported in Figure 4. The activation function ReLU is used within the MobileNetv2 because of its simplicity and efficiency of calculation. In addition, MobileNetv2 has bottleneck layers in the residual connections that obtain a representation of the input with reduced dimensionality (Sze et al. 2017). The lightweight depth-wise convolutions are used by the intermediate layer (Alain and Bengio 2016) to filter features as the source of nonlinearity. MobileNetv2 has 32 filters and an initial fully connected convolution layer followed by 19 residual bottleneck layers. The fully connected convolution layer, also known as convolutional kernels, can find the most effective filters based on the task and then combine these filters into more complex patterns. In addition, the output of some neurons will be 0, which reduces the interdependence of parameters and alleviates the overfitting problem.

Compared with the traditional CNNs, Swin Transformer has the unique shifted window which enables the model to gain strong global modeling ability and fewer parameters (Vaswani et al. 2017; Devlin et al. 2019). Due to the hierarchical approach, Swin Transformer has strong scalability in processing images of different scales. Meanwhile, the shifted window brings high computational complexity (Vyas et al. 2020). The CNNs usually have lower computational complexity and memory consumption, which can effectively extract local features in images (Zhichao Liu et al. 2021). Also, the CNNs can effectively extract local features in images through weight sharing (Abdel-Hamid et al. 2012; Miao et al. 2016).

The same epoch, batch size and optimizer as Swin Transformer are used in our experiments of ResNet50 and MobileNetv2.

Classification workflow

As mentioned in previous sections, we aimed at exploring the effectiveness of deep learning techniques as a possible substitute to the "by-hand" classification of metal mineral phases in images acquired by optical microscopes. To achieve our goal, we proposed a four-step workflow (Figure 5). From the raw images, we cropped the areas and selected those areas that contain only one mineral phase for further analysis (step 1); from these images, we constructed the training and test datasets, also using data augmentation techniques (step 2); then, we used the obtained datasets to train and test the investigated models (step 3); the trained models are finally used to infer the five metal mineral classes that are the object of the present study (step 4). In the following, we are going to detail the 4 steps of the proposed workflow (Figure 5).

Step 1: Image Stacks. We started with 481 raw light microscopy images of the five different metal mineral phases from different outcrops (see Materials Section). We next used the sliding window technique from OpenCV (Rosebrock 2015) to capture the mineral phases that are present in the image which we next cropped to equal-sized, non-overlapping, sub-images (256×256 pixels each, RGB; Figure 5). Subsequently, we removed the images that contain more than one mineral. The images spatially dominated by a single mineral were manually selected and labeled. Accordingly, 4524 images were obtained (Table 2), each being labeled with the name of the (dominant) mineral phase.

Step 2: Preprocessing and Data Augmentation. We resized the size of images to 224×224 because of the requirement of model input. And we then divided the above-mentioned image dataset into a training set, a validation set, and a test set with a ratio of 3:1:1 (see Table 2). The scope of the training set is to "educate" the model by determining the weight and bias learning parameters. The validation set is used to tune hyperparameters and check whether the effect of model training goes in a "good" or "bad" direction. In addition, the validation set is used and data is unseen during the training process. Finally, the test set is used to evaluate the generalization ability of the final model. In addition, the samples in the test set were selected from different ore deposits than those in the training set, which alleviates the potential issue of data leakage.

However, due to the limited number of images in the dataset, training can be challenging. To improve the model's generalization ability and reduce overfitting, we increased the size and the robustness of the training set and introduced variations that could be found in "real world data" using six typical offline data augmentation methods (Supplemental Materials). The first is named "random erasing" (Zhong et al. 2020). It consists of randomly selecting a rectangular region in an image and replacing its pixels with random values. This procedure generates new training images with various levels of occlusion, which, when used for training, reduces the risk of over-fitting and also makes the model less sensitive to occlusions (i.e., missing portions). The second approach is the "flipping". It consists of mirroring the images both horizontally and vertically, along the vertical and horizontal axes, respectively. The third augmentation method is named "brightness adjust". To note, the coloring of a picture can be set using three parameters: hue (H), saturation (S) and value (V), with the latter mainly governing the brightness. By using the Auto Gamma Correction method, a non-linear operation $S = T(R) = R^{\gamma}$ (where S and R are the values of brightness in output and original image, respectively, that are mapped to [0 1]) is to automatically lighten or darken the image (Babakhani and Zarei 2015). The fourth approach is "random zoom", which zooms into an image at a random location within the image. The fifth is "random contrast", which adjusts the contrast of the images by a random factor. And the last is "random saturation", which can adjust the saturation of images by a random factor. These methods can also improve the model's ability to classify based on the color (Supplemental Materials). At the end of the augmentation process, we increased the number of the training set to 18991 images.

Step 3: Model Training. For the present study, we trained two standard CNNs (i.e., ResNet50 and MobileNetv2) and a "Swin Transformer" algorithm (Figure 5). These architectures are followed by the max pooling and a fully connected (FC) layer. From the latter, a softmax function performs the final prediction, selecting the category with the largest softmax value. We trained the three models using adam gradient descent algorithm (Kingma and Ba 2017), using the first-order momentum parameter of 0.9, the second-order momentum parameter of 0.999, a learning rate of 1e-6 and a batch size of 16, respectively. The task was set for 20 epochs, each of which contains 136 iterations. At each training iteration, the image fed into the model is forward-propagated to the output layer, after which the difference between the ground-truth label and predictive label is measured by a Cross Entropy (CE) loss function. The loss value is then reduced by back-propagating and updating the model's parameters. To accelerate the training convergence and possibly increase the generalization capabilities of the models, we used a "transfer learning" approach by initializing the weights of the models to those from ImageNet-1k (Deng et al. 2009). In the process of model evaluation, the accuracy for the validation set is calculated after every epoch and the model's final accuracy is the highest among 20 epochs (Wu and Chen 2015; Ruby and Yendapalli 2020; Zhong et al. 2020).

Step 4: Getting the Output. We finally obtained a metal mineral classifier and we evaluated its performance.

Model evaluation

For the evaluation of the investigated models, we utilized 5 metrics, i.e., accuracy, precision, recall, F1-score, and training loss, defined as follows.

The model accuracy is defined as

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

TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative occurrences, respectively. The accuracy is the most common evaluation index in deep learning for classifiers, i.e., deep learning models used in classification tasks, due to its simple intuition.

The model precision is defined as

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

The precision is the proportion of positive samples predicted correctly by the model from all samples predicted as positive.

The model recall is defined as

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

The recall is a coverage measure, representing the classification accuracy of positive samples.

The F1-score is given by

$$F1 = 2 \times \left[\frac{Precision \times Recall}{Precision + Recall}\right] \ (4)$$

The F1-score is the harmonic mean between precision and recall. A macro F1-score computes the metric for each category independently and then takes the average (all categories are treated equally).

Finally, the Cross Entropy function is used to calculate the training loss of the model, which is defined as

$$CE(p,q) = -\sum_{i=1}^{C} p_i \log(q_i) \quad (5)$$

where C denotes the total number of classes, p_i denotes the i-th prediction class probability, and q_i denotes the i-th true class of training samples. The smaller the Cross Entropy (CE) loss is, the distributions of the two probabilities are approximately close, indicating that the model has a good performance.

Transfer learning

To get benefit from the "transfer learning" paradigm (Torrey and Shavlik 2010; Zhichao Liu et al. 2021), we set the initial weights of the investigated models to the pre-trained values that have been obtained after a full training on the natural ImageNet-1k dataset. The ImageNet-1k is a large public computer vision dataset that is often used as a benchmark to evaluate the performances of different deep learning models. It consists of 10 million images, characterized by thousand categories (Deng et al. 2009). The weights trained by ImageNet-1k, when used as the pre-training weight initialization, can quickly extract the shallow general image features (such as shape, brightness, and size of underlying image structures), thus future improving the initial accuracy of the model and accelerating the convergence of training models (Deng et al. 2009; Torrey and Shavlik 2010). Pre-trained weights for the ResNet50, MobileNetv2 and Swin Transformer are publicly available at the following repository: https://download.pytorch.org/models.

Class Activation Map

Although deep learning models are often characterized by good performances, they are subject to criticism because of their "black box" nature. To unblur the "black box" nature of the proposed models, we adopted the Class Activation Map (CAM; Zhou et al. 2016), also known as Class Thermal Maps. The CAM is a feature visualization technique that aims at highlighting the contribution of the different image regions to a given classification outcome. In detail, a CAM for a particular class of objects highlights the image regions used by the model to identify the specific class and shows which feature maps the model is based on for classification. Using a network architecture comprising convolutional layers, the feature map is extracted and the feather map up-sampled, which could be used as mask information to obtain the model's response value to the image in the target class. By linearly weighting the feature map with the obtained response values, the visual CAM mappings are obtained (Wang et al. 2020). In our specific case, the areas characterized by low- and high-discriminative powers have been highlighted by progressively shifting the thermal map from the blue to the red color. CAM is fully described in Zhou (2016).

Codes and Libraries

All the tasks are implemented using Python (https://www.python.org; version 3.8) and Pytorch (https://pytorch.org; version 1.8.1), and finished by Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and NVIDIA GeForce RTX 3090 GPU. All the codes and the dataset are available at the https://doi.org/10.5281/zenodo.10441351 repository.

The following libraries were used to complete the code: Pytorch (https://pytorch.org/) for computing tensors on graphics processing units; NumPy (https://numpy.org/) for data analysis; TorchAudio (https://pytorch.org/audio/stable/index.html) and SciPy (https://scipy.org/) for data processing functions; TensorBoard (https://tensorflow.google.cn/tensorboard) for data visualization; OpenCV (https://opencv.org/) and Pillow (https://pypi.org/project/Pillow/) for image processing; Torchvision (https://pytorch.org/vision/stable/index.html) for image classification; timm (https://timm.fast.ai/) for loading image model; Safetensors (https://pypi.org/project/safetensors/) for parameters and weights saving; tqdm (https://pypi.org/project/tqdm/) for progress prompt; Matplotlib (https://matplotlib.org/) for plotting the diagrams.

Results

Figure 6 displays the evolution of the training loss and validation accuracy for the Swin Transformer, ResNet50, and MobileNetv2, respectively.

Swin Transformer

For the Swin Transformer, the accuracy of the validation set gradually increases from 0.74 at the beginning of the training to 0.92 after 16 epochs (Figure 6a). After seven epochs, the accuracy reached 0.90. On the training set, the Cross Entropy (CE) loss which is a function that minimizes the model's loss is initially 1.42 (Figure 6b). The CE loss on the training set, initially at 1.60, decreases considerably during the seven epochs and slowly decreases over the course of the following epochs, reaching a minimum value of 0.01 in twentieth epoch (Figure 6b).

The average F1-score of each category for the Swin Transformer model is 0.92 (Table 3). Further analysis of the performance of each category shows that gold has the best classification performance with the F1-score of 0.98, the recall of 1.00 and the precision of 0.96 while pyrite has the lowest classification performance with the recall of 0.81. Figure 7a further details these insights by showing the confusion matrix for this model.

Convolutional Neural Networks

ResNet50. The accuracy of ResNet50 on the validation set gradually increases from 0.40 to 0.91, after 20 epochs (Figure 6a). After about 13 epochs, the accuracy of the validation set has reached 0.90, and then it begins to stabilize gradually. Initially, the CE loss on the training set is 1.58. During seven epochs, the training loss decreases greatly and drops below 0.1. The training loss decreases slowly in later epochs and reaches the minimum value of 0.01 after twenty epochs (Figure 6b).

The average F1-score of each category for the ResNet50 is 0.90 (Table 3). Further analysis of the performance parameters for each class shows that stibnite has the best classification performance whose F1-score is 0.98, the recall is 0.98 and the precision is 0.97, whereas chalcopyrite and pyrite have the worst performance with an F1-scores of 0.84 and 0.80, respectively. As can be found from the confusion matrix of ResNet50 (Figure 7b), chalcopyrite and pyrite are not predicted well, while gold is the best.

MobileNetv2. For the MobileNetv2, the accuracy of the validation set gradually increases from 0.34 to 0.84, achieved after 20 epochs (Figure 6a). After six epochs, the accuracy of the validation set has reached 0.81, after which it begins to stabilize gradually. The CE loss on the training set, initially at 1.60, continues to reduce in later epochs until it reaches a minimum value of 0.01. (Figure 6b).

The average F1-score for each class in the MobileNetv2 is 0.81 (Table 3). Further analysis of the performance of each class shows that gold is the best classified mineral, while pyrite and stibnite are the worst, with recall of 0.61, and 0.62, respectively. The confusion matrix (Figure 7c) shows that the MobileNetv2 can easily predict arsenopyrite and gold.

Discussion

Model classification performance

The results of the present study demonstrate that the Swin Transformer is characterized by an excellent prediction performance and a higher accuracy than the other tested models (Table 3). Comparing the specific scores of the three models, the Swin Transformer greatly improves the F1-scores of chalcopyrite, gold and pyrite, which are difficult to classify by the investigated CNNs (Table 3). This occurrence results in final average class accuracies of 0.92, 0.91, and 0.81 for the Swin Transformer, ResNet50, and MobileNetv2, respectively. The Swin Transformer also provides the lowest final training loss (Figure 6).

Moreover, misclassification occurrences of chalcopyrite and pyrite, often recognized as gold by ResNet50 (F1-score equal to 0.84 and 0.80, respectively; Table 3), MobileNetv2 (F1-score equal to 0.83 and 0.75, respectively; Table 3) as highlighted in Figure 7, were greatly reduced by the use of the Swin Transformer (F1-score of chalcopyrite and pyrite exceeding 0.85). It is also seen that stibnite is much less likely to be misclassified as arsenopyrite by the Swin Transformer than the MobileNetv2 (Figure 7). As a drawback, the confusion matrix demonstrates that stibnite is more likely misclassified as chalcopyrite by the Swin Transformer than the Resnet50 (Figure 7). This occurrence results in a slightly lower precision value of chalcopyrite of Swin Transformer (i.e., 0.85) than those characterizing the ResNet50 (i.e., 0.97). By analyzing the cause of the classification error, it can be inferred that the model easily classifies pyrites as chalcopyrite, probably due to the limited quality of the input images, thus the two minerals in the images have a similar yellow color (Figures 8a, b). This feature can confuse the models. Furthermore, arsenopyrite and stibnite are often misclassified by all models. The visual analysis of these situations (Figures 8c, d) shows that there are two main reasons for misclassification: (1) As stibnite images are from different samples with different image collection parameters, and they show a variety of colors some of which are similar to the grey reflection color of the arsenopyrite. (2) For some samples, both minerals have similar crystal forms from euhedral to subhedral, which can confuse the feature identification of the network. Simultaneously, pyrite and arsenopyrite also have similar reflection colors which are not easy to distinguish (Figures 8e, f). Also, other minerals present in the image cause interference, leading to model classification error (Figures 8e, f). However, these similarities do not affect the overall classification capabilities of the Swin Transformer, which clearly outperforms the investigated CNNs in most of the scores for the single classes and all the average performance metrics (Table 3 and Figure 7). In conclusion, our study supports the Swin Transformer as a metal mineral classifier (abbreviated ST-MMC).

Transfer learning in optical microscopy for the study of metal minerals

A large number of studies have shown that using the transfer learning paradigm to set the initial weights of a model before starting the training can effectively help in achieving the convergence. Transfer learning can fine-tune the parameters of the entire model to get initial high accuracy and a low loss value. (Figure 9; Supplemental Materials). Also, it allows for improving the generalization capability of a model (Kora Venu 2022). As an example, a number of studies demonstrated that using the "knowledge" acquired on natural images effectively improves the capabilities on a model in solving specific problems like the processing of medical or remote sensing images, even when using limited training sets (Xie et al. 2016; Raghu et al. 2019; Kora et al. 2021). Collecting metal mineral images with optical microscopes is a time-consuming task, thus resulting in a limited dataset size. As reported in the method section, we adopted the "pre-trained" weights for the investigated models deriving from a training on the ImageNet-1k dataset. As highlighted in Figure 6, the accuracy of the first validation, i.e., deriving from the pretraining only, was 0.34, 0.40, and 0.74 for the MobileNetv2, ResNet50 and the Swin Transformer, respectively. The Swin Transformer has a greater response to transfer learning and a higher initial accuracy.

To further outline the added value of transfer learning in achieving a solution for the problem investigated in the present manuscript, we trained the Swin Transformer without pre-trained weights. The training epoch, batch size, and other parameters of both models were the same. As a result, with the same number of training iterations, we obtained an initial and maximum accuracy of 0.56 and 0.88, respectively (Figure 9a), less than the accuracy achieved with the support of the "transfer learning" paradigm, i.e., 0.92. And the use of "transfer learning" paradigm supports lower initial and minimum loss values, which are 1.42 and 0.01, respectively (Figure 9b).

Model interpretation

Figure 10 shows five images that have been fed as unknowns, classified by the three models investigated in the present manuscript and output the probabilities. For each image, a blue-to-red heatmap points to the contribution of the different regions to the classification output. In detail, Figure 10 highlights that, for minerals with broken edges such as chalcopyrite (Figure 10a), thermal maps specifically focus on mineral edges. For the other cases (Figures 10b, c, d), the thermal maps highlight different regions, often focusing on the edges.

Based on the evidence reported above, it can be inferred that the shape of edges effectively influences the classification and the networks pay attention to their smoothness or sharpness. Moreover, the occurrences of misclassifications can now be better explained: Arsenopyrite and stibnite are both grey in color, and they also share smooth edges (Figures 8c, d). Also, both stibnite and arsenopyrite have void development on their surfaces. Despite the insights provided by CAMs do not directly lead us to improved models, they point to the causes that generate misclassifications and, therefore, suggest a direction for the possible improvement.

Notably, the different investigated networks show significant differences in their CAMs for the same inputs (Figure 10). The thermal maps of ResNet50 and MobileNetv2 almost focus on the edge of the minerals, and not inside them. As a consequence, the reflected color and texture may not be the most important distinctive features of these two models. However, CAMs for the Swin Transformer also cover the interior of the mineral, rather than just the edge, which suggests that ST-MMC effectively uses the reflected color and texture of the minerals for its inference, and it has better global performance. In the thermal map, the mineral area of middle and edge of ST-MMC are redder than the other two CNNs, indicating that these domains have stronger model response, which reveals that mineral reflection color and texture contribute more to the classification output of the model. The Swin Transformer also achieved a classification predict probability over 0.95 for unknown minerals, significantly outperforming the other two CNNs (Figure 10). This effectiveness in handling unknown samples also demonstrates its capability for efficient batch image processing.

Implications

Large-dimensional image analyses are dominantly based on digital image datasets, the automatic identification of the optical microscopic data is still poorly examined, and the mineral image data is also difficult to collect. Deep learning-based approach (Swin Transformer) with the transfer learning paradigm fully explores the information of different metal mineral phases to produce a well-behaved mineral classifier with high accuracy and strong global ability. To circumvent the ‘black box’ problem commonly associated with deep learning models, CAM (Class Activation Map) tool was introduced to explain individual predictions. With the increasing amount

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

of high-throughput mineral image data produced by modern analytical techniques, our ST-MMC offers the potential to make more data driven decisions such as transparent minerals classification. Moreover, the “transfer learning” paradigm on large images captured by optical microscopy, will possibly liberate researchers from tiresome labor, sharpen the accuracy, and increase the productivity. More widely, the use of “transfer learning” may disclose new perspectives in petrology and mineralogy, possibly providing a paradigm shift over the current applications of deep learning in petrology and mineralogy.

Acknowledgments

We gratefully acknowledge editors for handling our manuscript, and the constructive comments of the anonymous reviewers. This research was financially supported by the National Key Research and Development Program (SQ2023YFE0103286), the National Natural Science Foundation (42261134535 and 42072087), the Frontiers Science Center for Deep-time Digital Earth (2652023001), the 111 Project of the Ministry of Science and Technology (BP0719021) and the Shandong Provincial Engineering Laboratory of Application and Development of Big Data for Deep Gold Exploration (SDK202211 and SDK202214). MP kindly acknowledge the MUR PRIN2020 project “Dynamics and timescales of volcanic plumbing systems: a multidisciplinary approach to a multifaceted problem” (202037YPCZ_001).

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., and Penn, G. (2012) Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 4277–

4280. Presented at the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://doi.org/10.1109/ICASSP.2012.6288864.

Acosta, I.C.C., Khodadadzadeh, M., Tusa, L., Ghamisi, P., and Gloaguen, R. (2019) A Machine Learning Framework for Drill-Core Mineral Mapping Using Hyperspectral and High-Resolution Mineralogical Data Fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12, 4829–4842, https://doi.org/10.1007/s00521-021-05849-3.

Alain, G., and Bengio, Y. (2016) Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, https://doi.org/10.48550/arXiv.1610.01644.

Alférez, G.H., Vázquez, E.L., Martínez Ardila, A.M., and Clausen, B.L. (2021) Automatic classification of plutonic rocks with deep learning. Applied Computing and Geosciences, 10, 100061, https://doi.org/10.1016/j.acags.2021.100061.

Azeuda Ndonfack, K.I., Xie, Y., and Goldfarb, R. (2022) Gold occurrences of the Woumbou–Colomine–Kette district, eastern Cameroon: ore-forming constraints from petrography, SEM–CL imagery, fluid inclusions, and C–O–H–S isotopes. Mineralium Deposita, 57, 83–105, https://doi.org/10.1007/s00126-021-01050-7.

Babakhani, P., and Zarei, P. (2015) Automatic gamma correction based on average of brightness, 4, 156-159.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., and others (2021) On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258, https://doi.org/10.48550/arXiv.2108.07258.

Cameron, E.N. (1961) Ore microscopy.

Castelvecchi, D. (2016) Can we open the black box of AI? Nature News, 538, 20.

Chang, J.-R., and Chen, Y.-S. (2015) Batch-normalized maxout network in network. arXiv preprint arXiv:1511.02583, https://doi.org/10.48550/arXiv.1511.02583.

Chen, Z., Chen, D., and Li, C. (1979) Color index of the reflection color of ore minerals. Acta Petrologica Sinica, 219–233, https://doi.org/10.19762/j.cnki.dizhixuebao.1979.03.005.

Craig, J.R., Vaughan, D.J., and Hagni, R.D. (1981) Ore microscopy and ore petrography Vol. 406. Wiley New York.

Criddle, A.J., and Stanley, C.J. (2012) Quantitative data file for ore minerals. Springer Science & Business Media.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009) ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition pp. 248–255. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPR.2009.5206848.

Deng, J., Yang, L.-Q., Groves, D.I., Zhang, L., Qiu, K.-F., and Wang, Q.-F. (2020a) An integrated mineral system model for the gold deposits of the giant Jiaodong province, eastern China. Earth-Science Reviews, 208, 103274, https://doi.org/10.1016/j.earscirev.2020.103274.

Deng, J., Qiu, K.-F., Wang, Q.-F., Goldfarb, R., Yang, L.-Q., Zi, J.-W., Geng, J.-Z., and Ma, Y. (2020b) In situ dating of hydrothermal monazite and implications for the geodynamic controls on ore formation in the Jiaodong gold province, eastern China. Economic Geology, 115, 671–685, https://doi.org/10.5382/econgeo.4711.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, https://doi.org/10.48550/arXiv.1810.04805.

dos Anjos, C.E., Avila, M.R., Vasconcelos, A.G., Pereira Neta, A.M., Medeiros, L.C., Evsukoff, A.G., Surmas, R., and Landau, L. (2021) Deep learning for lithological classification of carbonate rock micro-CT images. Computational Geosciences, 25, 971–983, https://doi.org/10.1007/s10596-021-10033-6.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., and others (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, https://doi.org/10.48550/arXiv.2010.11929.

Endert, A., Ribarsky, W., Turkay, C., Wong, B.L.W., Nabney, I., Blanco, I.D., and Rossi, F. (2017) The State of the Art in Integrating Machine Learning into Visual Analytics. Computer Graphics Forum, 36, 458–486, https://doi.org/10.1111/cgf.13092.

Faria, E.L., Coelho, Juliana.M., Matos, T.F., Santos, B.C.C., Trevizan, W.A., Gonzalez, J.L., Bom, C.R., de Albuquerque, Márcio P., and de Albuquerque, Marcelo P. (2022) Lithology identification in carbonate thin section images of the Brazilian pre-salt reservoirs by the computational vision and deep learning. Computational Geosciences, 26, 1537–1547, https://doi.org/10.1007/s10596-022-10168-0.

Feng, V. (2017) An overview of resnet and its variants. Towards data science, 2.

Fu, Y., and Aldrich, C. (2019) Quantitative Ore Texture Analysis with Convolutional Neural Networks. IFAC-PapersOnLine, 52, 99–104, https://doi.org/10.1016/j.ifacol.2019.09.171.

Ghosh, S., Das, N., Das, I., and Maulik, U. (2019) Understanding Deep Learning Techniques for Image Segmentation. ACM Computing Surveys, 52, 73:1-73:35, https://doi.org/10.1145/3329784.

He, K., Zhang, X., Ren, S., and Sun, J. (2015) Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385, https://doi.org/10.48550/arXiv.1512.03385.

Jiang, Y., Zhang, Y., Lin, X., Dong, J., Cheng, T., and Liang, J. (2022) SwinBTS: A Method for 3D Multimodal Brain Tumor Segmentation Using Swin Transformer. Brain Sciences, 12, 797, https://doi.org/10.3390/brainsci12060797.

Jungmann, M., Pape, H., Wißkirchen, P., Clauser, C., and Berlage, T. (2014) Segmentation of thin section images for grain size analysis using region competition and edge-weighted region merging. Computers & Geosciences, 72, 33–48, https://doi.org/10.1016/j.cageo.2014.07.002.

Kingma, D.P., and Ba, J. (2017, January 29) Adam: A Method for Stochastic Optimization. arXiv.

Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P., and Rajendra Acharya, U. (2021) Transfer learning techniques for medical image analysis: A review. Biocybernetics and Biomedical Engineering, 42, 79–107, https://doi.org/10.1016/j.bbe.2021.11.004.

Kora Venu, S. (2022) Improving the Generalization of Deep Learning Classification Models in Medical Imaging Using Transfer Learning and Generative Adversarial Networks. In A.P. Rocha, L. Steels, and J. van den Herik, Eds., Agents and Artificial Intelligence pp. 218–235. Springer International Publishing, Cham, https://doi.org/10.1007/978-3-031-10161-8_12.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1989) Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1, 541–551. Presented at the Neural Computation, https://doi.org/10.1162/neco.1989.1.4.541.

LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning. Nature, 521, 436–444, https://doi.org/10.1038/nature14539.

Leichter, A., Almeev, R.R., Wittich, D., Beckmann, P., Rottensteiner, F., Holtz, F., and Sester, M. (2022) Automated Segmentation of Olivine Phenocrysts in a Volcanic Rock Thin Section Using a Fully Convolutional Neural Network. Frontiers in Earth Science, 10, 740638, https://doi.org/10.3389/feart.2022.740638.

Liu, Zhichao, Jin, L., Chen, J., Fang, Q., Ablameyko, S., Yin, Z., and Xu, Y. (2021) A survey on applications of deep learning in microscopy image analysis. Computers in Biology and Medicine, 134, 104523, https://doi.org/10.1016/j.compbiomed.2021.104523.

Liu, Ze, Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows pp. 10012–10022. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, https://doi.org/10.48550/arXiv.2103.14030.

Lu, J., and Peng, X. (2010) Manual of microscopic identification of metal minerals. Geology Press, Beijing.

Miao, S., Wang, Z.J., and Liao, R. (2016) A CNN Regression Approach for Real-Time 2D/3D Registration. IEEE Transactions on Medical Imaging, 35, 1352–1363. Presented at the IEEE Transactions on Medical Imaging, https://doi.org/10.1109/TMI.2016.2521800.

Młynarczuk, M., Górszczyk, A., and Ślipek, B. (2013) The application of pattern recognition in the automatic classification of microscopic rock images. Computers & Geosciences, 60, 126–133, https://doi.org/10.1016/j.cageo.2013.07.015.

Obaid, K.B., Zeebaree, S.R.M., and Ahmed, O.M. (2020) Deep Learning Models Based on Image Classification: A Review. International Journal of Science and Business, 4, 75–81, https://doi.org/10.5281/zenodo.4108433.

Petrelli, M. (2021) Introduction to Python in Earth Science Data Analysis: From Descriptive Statistics to Machine Learning. Springer Nature.

Petrelli, M., and Perugini, D. (2016) Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contributions to Mineralogy and Petrology, 171, 81, https://doi.org/10.1007/s00410-016-1292-2.

Picot, P., and Johan, Z. (1977) Atlas des Mineraux Metalliques.

Piller, H. (1966) Colour measurements in ore-microscopy. Mineralium Deposita, 1, 175–192, https://doi.org/10.1007/BF00204546.

——— (2012) Microscope photometry. Springer Science & Business Media.

Polat, Ö., Polat, A., and Ekici, T. (2021) Automatic classification of volcanic rocks from thin section images using transfer learning networks. Neural Computing and Applications, 33, 11531–11540, https://doi.org/10.1007/s00521-021-05849-3.

Qiu, K., Zhou, T., Chew, D., Hou, Z., Müller, A., Yu, H., Lee, R.G., Chen, H., and Deng, J. (2023) Apatite trace element composition as an indicator of ore deposit types: a machine learning approach. American Mineralogist, http://doi.org/10.2138/am-2022-8805.

Qiu, K.-F., Deng, J., Laflamme, C., Long, Z.-Y., Wan, R.-Q., Moynier, F., Yu, H.-C., Zhang, J.-Y., Ding, Z.-J., and Goldfarb, R. (2023a) Giant Mesozoic gold ores derived from subducted oceanic slab and overlying sediments. Geochimica et Cosmochimica Acta, 343, 133–141, https://doi.org/10.1016/j.gca.2023.01.002.

Qiu, K.-F., Deng, J., Sai, S.-X., Yu, H.-C., Tamer, M.T., Ding, Z.-J., Yu, X.-F., and Goldfarb, R. (2023b) Low-Temperature Thermochronology for Defining the Tectonic Controls on Heterogeneous Gold Endowment Across the Jiaodong Peninsula, Eastern China. Tectonics, 42, e2022TC007669, https://doi.org/10.1029/2022TC007669.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019) Transfusion: Understanding Transfer Learning for Medical Imaging. In Advances in Neural Information Processing Systems Vol. 32. Curran Associates, Inc, https://doi.org/10.48550/arXiv.1902.07208.

Ramdohr, P. (2013) The ore minerals and their intergrowths. Elsevier.

Rosebrock, A. (2015) Sliding windows for object detection with python and opencv. PylmageSearch, Navigation.

Ruby, U., and Yendapalli, V. (2020) Binary cross entropy with deep learning technique for Image classification. International Journal of Advanced Trends in Computer Science and Engineering, 9.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks pp. 4510–4520. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.48550/arXiv.1801.04381.

Santosh, M. (2010) Assembling North China Craton within the Columbia supercontinent: The role of double-sided subduction. Precambrian Research, 178, 149–167, https://doi.org/10.1016/j.precamres.2010.02.003.

Santosh, M., Wilde, S.A., and Li, J.H. (2007) Timing of Paleoproterozoic ultrahigh-temperature metamorphism in the North China Craton: Evidence from SHRIMP U–Pb zircon

DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

geochronology. Precambrian Research, 159, 178–196, https://doi.org/10.1016/j.precamres.2007.06.006.

Santosh, M., Maruyama, S., and Sato, K. (2009) Anatomy of a Cambrian suture in Gondwana: Pacific-type orogeny in southern India? Gondwana Research, 16, 321–341, https://doi.org/10.1016/j.gr.2008.12.012.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. Nature methods, 9, 671–675, https://doi.org/10.1038/nmeth.2089.

Schrader, D.L., and Zega, T.J. (2019) Petrographic and compositional indicators of formation and alteration conditions from LL chondrite sulfides. Geochimica et Cosmochimica Acta, 264, 165–179, https://doi.org/10.1016/j.gca.2019.08.015.

Shang, J., and Lin, J. (1990) Gold minerals and their occurance. Journal of Changchun University of Earth Science, 20, 273–278.

Sheldrake, T., and Higgins, O. (2021) Classification, segmentation and correlation of zoned minerals. Computers & Geosciences, 156, 104876, https://doi.org/10.1016/j.cageo.2021.104876.

Su, C., Xu, S., Zhu, K., and Zhang, X. (2020) Rock classification in petrographic thin section images based on concatenated convolutional neural networks. Earth Science Informatics, 13, 1477–1484, https://doi.org/10.1007/s12145-020-00505-1.

Sze, V., Chen, Y.-H., Yang, T.-J., and Emer, J.S. (2017) Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE, 105, 2295–2329. Presented at the Proceedings of the IEEE, https://doi.org/10.1109/jproc.2017.2761740.

Tang, K., Wang, Y.D., Mostaghimi, P., Knackstedt, M., Hargrave, C., and Armstrong, R.T. (2022) Deep convolutional neural network for 3D mineral identification and liberation analysis. Minerals Engineering, 183, 107592, https://doi.org/10.1016/j.mineng.2022.107592.

Tarquini, S., and Favalli, M. (2010) A microscopic information system (MIS) for petrographic analysis. Computers & Geosciences, 36, 665–674, https://doi.org/10.1016/j.cageo.2009.09.017.

Thrun, S., and Mitchell, T.M. (1995) Lifelong robot learning. Robotics and autonomous systems, 15, 25–46, https://doi.org/10.1016/0921-8890(95)00004-y.

Torrey, L., and Shavlik, J. (2010) Transfer Learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques pp. 242–264. IGI Global.

Trauth, M.H., Gebbers, R., Marwan, N., and Sillmann, E. (2007) MATLAB recipes for earth sciences Vol. 34. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017) Attention is All you Need. In Advances in Neural Information Processing Systems Vol. 30. Curran Associates, Inc, https://doi.org/10.48550/arXiv.1706.03762.

Vyas, A., Katharopoulos, A., and Fleuret, F. (2020) Fast Transformers with Clustered Attention. In Advances in Neural Information Processing Systems Vol. 33, pp. 21665–21674. Curran Associates, Inc, https://doi.org/10.48550/arXiv.2007.04825.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020, April 13) Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. arXiv, https://doi.org/10.48550/arXiv.1910.01279.

Wang, Y., Qiu, K.-F., Müller, A., Hou, Z.-L., Zhu, Z.-H., and Yu, H.-C. (2021) Machine learning prediction of quartz forming-environments. Journal of Geophysical Research: Solid Earth, 126, e2021JB021925, https://doi.org/10.1029/2021JB021925.

Wang, Y.D., Shabaninejad, M., Armstrong, R.T., and Mostaghimi, P. (2021) Deep neural networks for improving physical accuracy of 2D and 3D multi-mineral segmentation of rock micro-CT images. Applied Soft Computing, 104, 107185, https://doi.org/10.1016/j.asoc.2021.107185.

Wu, M., and Chen, L. (2015) Image recognition based on deep learning. In 2015 Chinese Automation Congress (CAC) pp. 542–546. Presented at the 2015 Chinese Automation Congress (CAC).

Wu, X., Sahoo, D., and Hoi, S.C.H. (2020) Recent advances in deep learning for object detection. Neurocomputing, 396, 39–64, https://doi.org/10.1016/j.neucom.2020.01.085.

Xie, M., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2016) Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. Proceedings of the AAAI Conference on Artificial Intelligence, 30, https://doi.org/10.1609/aaai.v30i1.9906.

Xing, F., Xie, Y., Su, H., Liu, F., and Yang, L. (2018) Deep Learning in Microscopy Image Analysis: A Survey. IEEE Transactions on Neural Networks and Learning Systems, 29, 4550–4568. Presented at the IEEE Transactions on Neural Networks and Learning Systems, https://doi.org/10.1109/tnnls.2017.2766168.

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

Yang, H., and Yang, D. (2023) CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. Expert Systems with Applications, 213, 119024, https://doi.org/10.1016/j.eswa.2022.119024.

Zhang, L., Qiu, K., Hou, Z., Pirajno, F., Shivute, E., and Cai, Y. (2021) Fluid-rock reactions of the Triassic Taiyangshan porphyry Cu-Mo deposit (West Qinling, China) constrained by QEMSCAN and iron isotope. Ore Geology Reviews, 132, 104068, https://doi.org/10.1016/j.oregeorev.2021.104068.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019) Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems, 30, 3212–3232. Presented at the IEEE Transactions on Neural Networks and Learning Systems, https://doi.org/10.1109/tnnls.2018.2876865.

Zhong, R., Deng, Y., Li, W., Danyushevsky, L.V., Cracknell, M.J., Belousov, I., Chen, Y., and Li, L. (2021) Revealing the multi-stage ore-forming history of a mineral deposit using pyrite geochemistry and machine learning-based data interpretation. Ore Geology Reviews, 133, 104079, https://doi.org/10.1016/j.oregeorev.2021.104079.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020) Random Erasing Data Augmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 34, 13001–13008, https://doi.org/10.1609/aaai.v34i07.7000.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016) Learning Deep Features for Discriminative Localization pp. 2921–2929. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/cvpr.2016.319.

Zhou, T., Qiu, K., Wang, Y., Yu, H., and Hou, Z. (2022) Apatite Eu/Y-Ce discrimination diagram: A big data based approach for provenance classification. Acta Petrol. Sin, 38, 291–299, https://doi.org/10.18654/1000-0569/2022.01.19.

Zussman, J. and others (1967) Physical methods in determinative mineralogy. Physical methods in determinative mineralogy, https://doi.org/10.1111/j.1365-2818.1979.tb00182.x.

Figure Captions

Figure 1. The representative images for five minerals which are collected from gold deposits. (a-d) Gold; (e-h) Pyrite; (i-l) Chalcopyrite; (m-p) Arsenopyrite; (q-t) Stibnite. Apy: arsenopyrite; Au: gold; Cal: calcite; Ccp: chalcopyrite; Py: pyrite; Qz: quartz; Ser: sericite; Stb: stibnite; Tur: tourmaline.

Figure 2. Architecture of Swin Transformer. The blocks with different colors represent different functions. The network has four stages and the last three stages have same structure. The olive block is patch partition module and patch merging layer. The sage is linear embedding layer. The light salmon is fully connected layer. And the orange is the classifier. H: height; W: width; C: color.

Figure 3. Architecture of ResNet50. The blocks with different colors represent different functions. The network has five stages and the last four stages have same structure. Stage 1: the dark salmon block is the convolutional layer. the salmon is the normalization. the light one is the activation function. and the sage one is the max pooling layer; Stage 2 to stage 5: olive one is the convolutional block with one convolutional layer; the salmon block is the identity block with two,

three, five and two convolutional layers, respectively. And the last light salmon is the pooling layer, and the orange is the classifier。

Figure 4. Architecture of MobileNetv2. It contains two units: stride=1 and stride=2. Conv 1x1 is the 1x1 convolutional kernel. ReLU is nonlinear activation function. Dwise 3x3 is depth-wise convolution with 3x3 convolutional kernel.

Figure 5. Workflow of the proposed automatic classification. Step 1: dataset compiling. Crop raw images using OpenCV. Select the processed images that contain only one mineral phase; Step 2: data splitting and augmentation. The dataset was divided into training set, validation set and test set (3:1:1). The data augmentation methods include random erasing, flipping, brightness adjust, random zoom, random contrast and random saturation; Step 3: model training and evaluating. Swin Transformer, ResNet50 and MobileNetv2 algorithms were used to train the classification models. The model evaluation metrics include accuracy, precision, recall, and F1-score; Step 4: model predicting. Put the images to the trained model to predict the five metal mineral classes.

Figure 6. Changes of (a) validation accuracy and (b) training loss of three algorithms using the method of transfer learning. The lines reflect the changes of different algorithms' performance within 20 epochs (green: Swin Transformer algorithm; red: ResNet50 algorithm; blue: MobileNetv2 algorithm).

Figure 7. Confusion matrix of the test set used to evaluate the three algorithms. (a) Swin Transformer; (b) ResNet50; (c) MobileNetv2. Indicated values are the number of images. The

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9092. http://www.minsocam.org/

horizontal axis represents the predicted label, while the vertical axis denotes the true label. The horizontal axis is the predicted label, while the vertical axis is the true label. Apy: arsenopyrite; Ccp: chalcopyrite; Au: gold; Py: pyrite; Stb: stibnite.

Figure 8. Presentation of erroneous classification results from Swin Transformer metal mineral classifier. The model misclassified (a) pyrite and (b) chalcopyrite, (c) arsenopyrite and (d) stibnite, as well as (e) pyrite and (f) arsenopyrite. Apy: arsenopyrite; Ccp: chalcopyrite; Py: pyrite; Stb: stibnite.

Figure 9. Changes in (a) validation accuracy and (b) training loss of Swin Transformer with transfer learning and without transfer learning respectively. The lines reflect the changes of different algorithms' performance within 20 epochs (dark green: Swin Transformer with transfer learning; light green: Swin Transformer without transfer learning).

Figure 10. CAMs of three models with five-classes metal minerals image classification. The redder the mapping, the higher the response of the corresponding area of the original image to the model's classification output. The numbers on the mappings represent the output probability of the model for unknown minerals. Ccp: chalcopyrite; Py: pyrite; Au: gold; Apy: arsenopyrite; Stb: stibnite.

Table

Table 1 Characteristics of Metal Minerals Under the Microscope

	gold	**pyrite**	**chalcopyrite**	**arsenopyrite**	**stibnite**
Chemical composition	Au	FeS_2	$CuFeS_2$	FeAsS	Sb_2S_3
Reflectivity	Gold 480: 33.97; 546: 70.67; 589: 80.09; 656: 85.88	White: 54.5; 470: 46; 546: 53; 589: 54; 650:5 4	White: 44~46.1; 470: 34; 546: 47; 589: 48; 650: 49	White: 51.7~55.7; 470: 51~55; 546: 52~54; 589: 53~54; 650: 53	White: 30.2~40; 470: 31~53; 546: 31~48; 589: 30~45; 650: 30~42
Reflection color	Golden yellow, bright yellow	Light yellow	Copper yellow	Bright white with cream or red color, weak polychromatic	White to light off-white
Homogeneity and heterogeneity	Homogenous	Homogenous	Weak heterogeneity	Strong heterogeneity	Strong heterogeneity
Morphological characteristics	Polymeric crystals between octahedral, hexahedral, tetrahexahedral, triangular trioctahedral, and rhomboid dodecahedron; Irregularly granular	Euhedral crystals in the form of cubes, pentagonal dodecahedral and octahedron	Irregular granular crystals	Diamond-shaped, elongated columnar, spear-headed and other euhedral crystals	Columnar long and granular crystals
Mineral combination	Arsenopyrite, pyrite, chalcopyrite, pyrrhotite, galena, sphalerite, stibnite, calcite, tellurite and quartz	Iron, copper, lead, zinc, silver sulfide, gold, rutile, graphite, etc.	Associated with sulfides	Pyrite, loellingite, tetrahedrite, magnetite, galena, sphalerite, stibnite, etc.	Berthierite, pyrite, arsenopyrite, sphalerite, tetrahedrite, scheelite, gold, realgar, orpiment, etc.
References	(Shang and Lin 1990; Piller 2012; Ramdohr 2013)	(Cameron 1961; Chen et al. 1979; Ramdohr 2013)	(Piller 1966; Zussman and others 1967; Santosh et al. 2007)	(Picot and Johan 1977; Lu and Peng 2010)	(Craig et al. 1981; Criddle and Stanley 2012)

Note: Reflectivity is the reflective quality or power of the mineral surface; Reflection color is the color of the mineral polished surface under vertical illumination; Homogeneity is the property of crystalline mineral of isometric system and amorphous mineral. The mineral is dark in the field of view under the polarizer, and the darkness level does not change when the platform is rotated; Heterogeneity is the property of crystalline mineral of non-isometric system. When the orientation of the mineral is changed by rotating the platform, the brightness and the color through the upper polarizer will change with the variation of orientation.

Table 2 Summary of the Training, Validation and Test Sets of Image Dataset

	Arsenopyrite	Chalcopyrite	Gold	Pyrite	Stibnite	Total
Training	548	505	490	574	596	2713
Validation	182	170	164	191	198	905
Test	183	169	163	192	199	906
Total	913	844	817	957	993	4524

Table 3 Mineral Classification Performance on the Test Set

Method	Metric	Arsenopyrite	Chalcopyrite	Gold	Pyrite	Stibnite	Metric Value
Swin-Transformer	Acc						**0.92**
	Pre	0.85	0.83	0.96	1.00	0.99	**0.93**
	Rec	0.99	0.97	1.00	0.81	0.84	**0.92**
	F1	0.91	0.89	0.98	0.90	0.91	**0.92**
ResNet50	Acc						**0.91**
	Pre	0.97	0.78	0.88	0.94	0.97	**0.91**
	Rec	0.98	0.89	0.99	0.70	0.98	**0.91**
	F1	0.97	0.84	0.93	0.80	0.98	**0.90**
MobileNetv2	Acc						**0.81**
	Pre	0.64	0.77	0.91	0.98	1.00	**0.86**
	Rec	1.00	0.90	0.99	0.61	0.62	**0.82**
	F1	0.78	0.83	0.95	0.75	0.76	**0.81**

Note: Acc: abbreviation for model evaluation indicator accuracy; Pre: abbreviation for model evaluation indicator precision; Rec: abbreviation for model evaluation indicator recall; F1: abbreviation for model evaluation indicator F1-score.

Figure 1

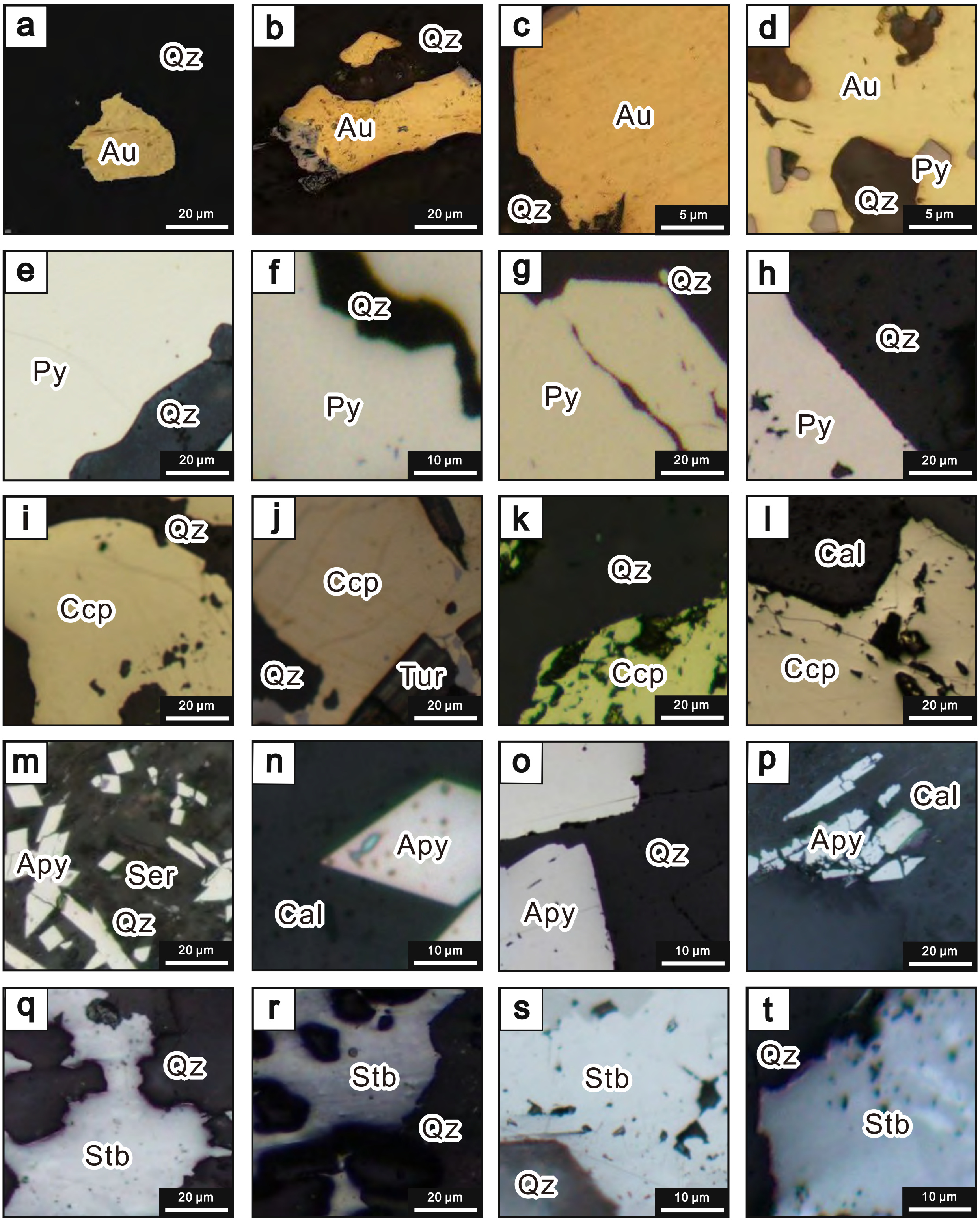
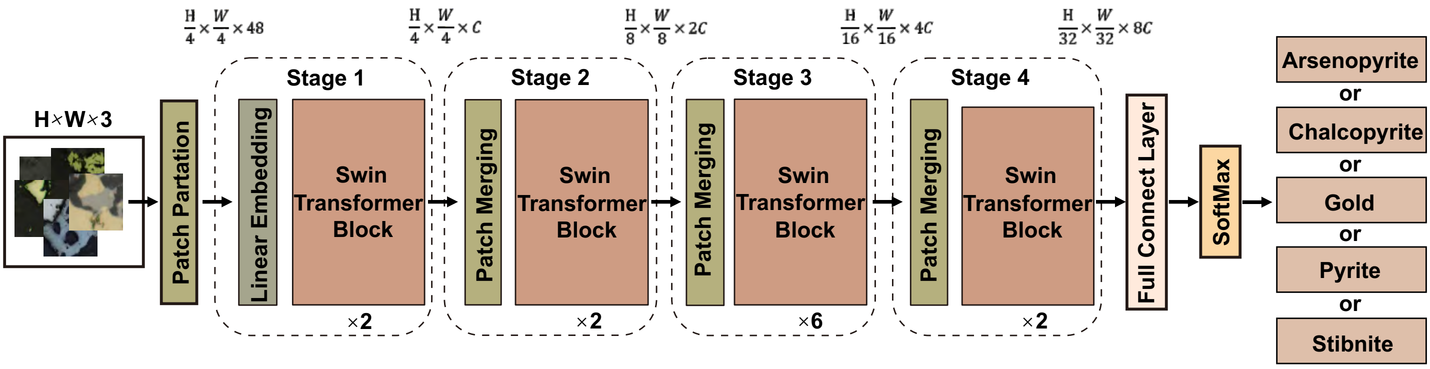

a
Qz
Au
20 µm
b
Qz
Au
20 µm
c
Au
Qz
5 µm
d
Au
Py
Qz
5 µm
e
Py
Qz
20 µm
f
Qz
Py
10 µm
g
Qz
Py
20 µm
h
Qz
Py
20 µm
i
Qz
Ccp
20 µm
j
Ccp
Qz
Tur
20 µm
k
Qz
Ccp
20 µm
l
Cal
Ccp
20 µm
m
Apy
Ser
Qz
20 µm
n
Apy
Cal
10 µm
o
Qz
Apy
10 µm
p
Cal
Apy
20 µm
q
Qz
Stb
20 µm
r
Stb
Qz
20 µm
s
Stb
Qz
10 µm
t
Qz
Stb
10 µm

Figure 2



$\frac{H}{4}\times\frac{W}{4}\times 48$
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
H×W×3
Patch Partation
Stage 1
Linear Embedding
Swin Transformer Block
×2
Stage 2
Patch Merging
Swin Transformer Block
×2
Stage 3
Patch Merging
Swin Transformer Block
×6
Stage 4
Patch Merging
Swin Transformer Block
×2
Full Connect Layer
SoftMax
Arsenopyrite
or
Chalcopyrite
or
Gold
or
Pyrite
or
Stibnite

Figure 3

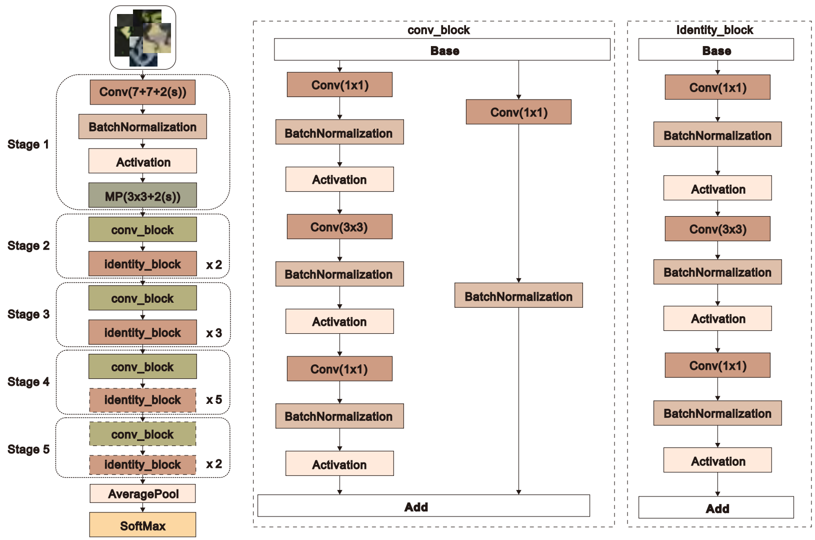

Stage 1
Conv(7+7+2(s))
BatchNormalization
Activation
MP(3x3+2(s))
Stage 2
conv_block
identity_block
x 2
Stage 3
conv_block
identity_block
x 3
Stage 4
conv_block
identity_block
x 5
Stage 5
conv_block
identity_block
x 2
AveragePool
SoftMax
conv_block
Base
Conv(1x1)
BatchNormalization
Activation
Conv(3x3)
BatchNormalization
Activation
Conv(1x1)
BatchNormalization
Activation
Conv(1x1)
BatchNormalization
Add
Identity_block
Base
Conv(1x1)
BatchNormalization
Activation
Conv(3x3)
BatchNormalization
Activation
Conv(1x1)
BatchNormalization
Activation
Add

Figure 4

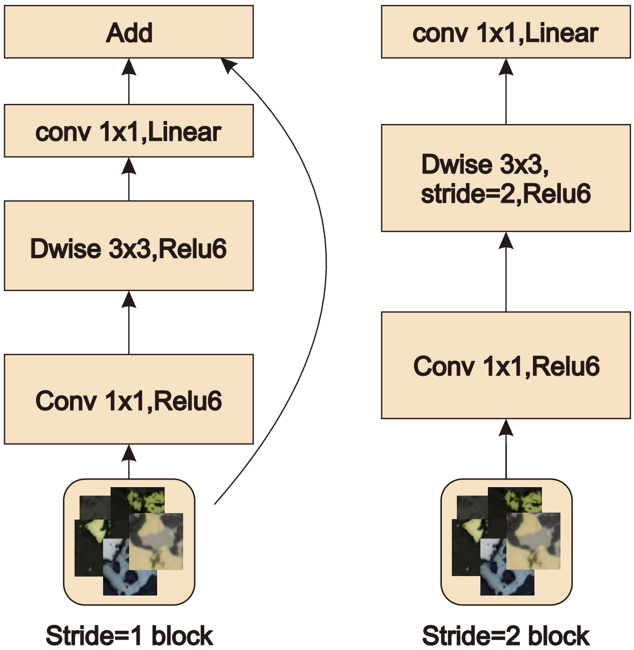

Add
conv 1x1,Linear
Dwise 3x3,Relu6
Conv 1x1,Relu6
Stride=1 block
conv 1x1,Linear
Dwise 3x3,
stride=2,Relu6
Conv 1x1,Relu6
Stride=2 block

Figure 5

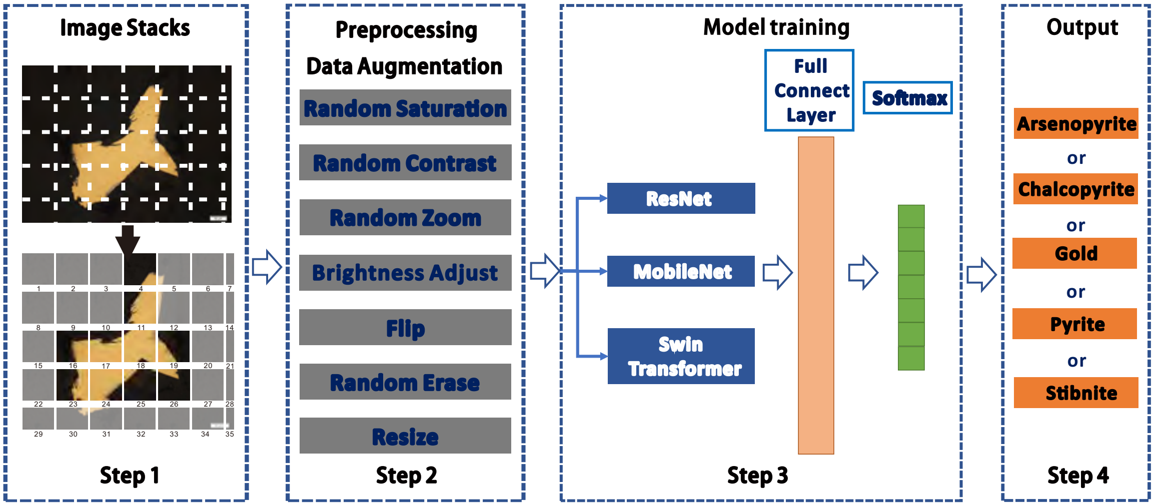
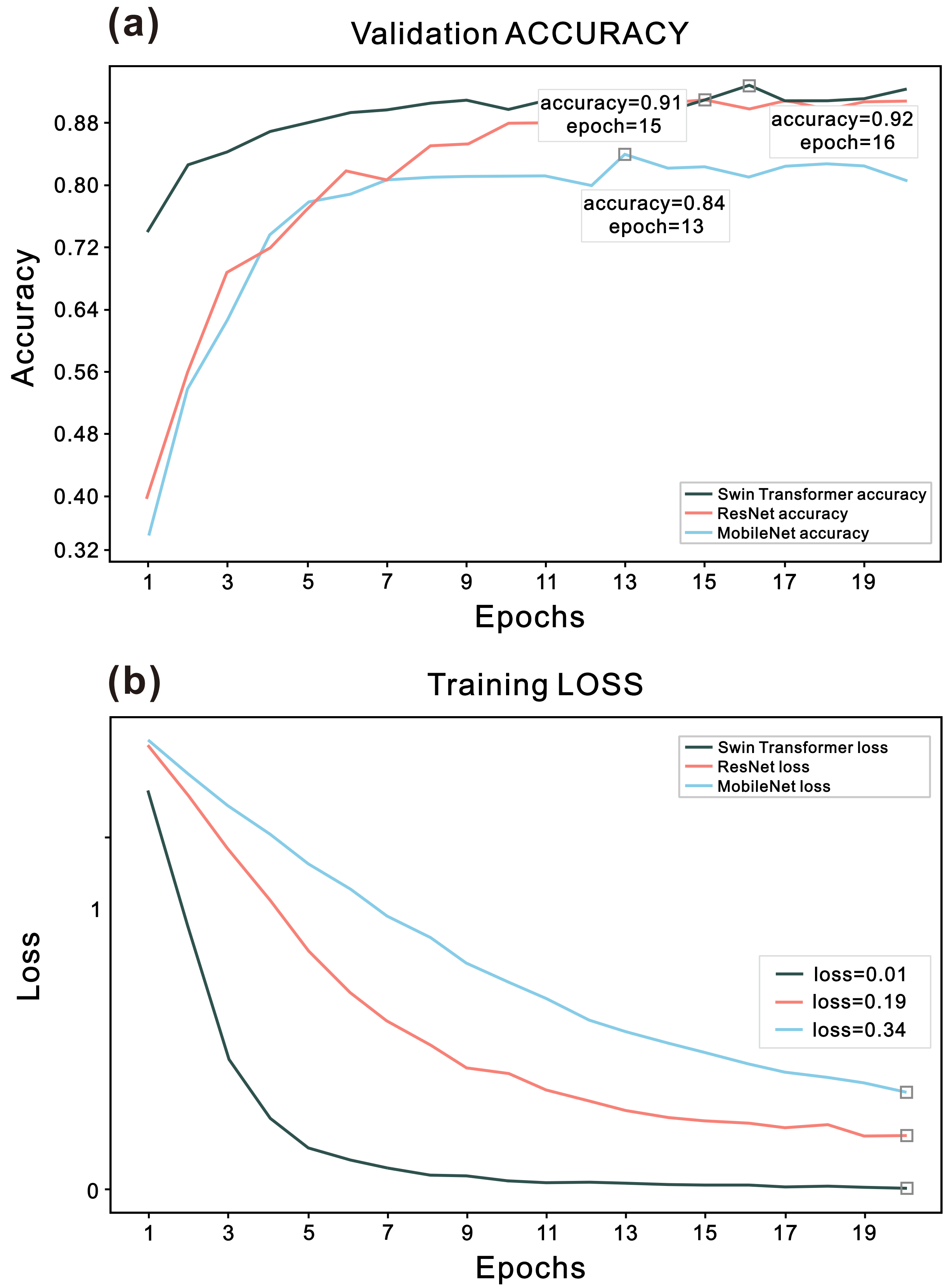


Image Stacks
Step 1
Preprocessing
Data Augmentation
Random Saturation
Random Contrast
Random Zoom
Brightness Adjust
Flip
Random Erase
Resize
Step 2
Model training
ResNet
MobileNet
Swin Transformer
Full Connect Layer
Softmax
Step 3
Output
Arsenopyrite
or
Chalcopyrite
or
Gold
or
Pyrite
or
Stibnite
Step 4

Figure 6



(a)
Validation ACCURACY
accuracy=0.91
epoch=15
accuracy=0.92
epoch=16
accuracy=0.84
epoch=13
Swin Transformer accuracy
ResNet accuracy
MobileNet accuracy
Accuracy
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32
1
3
5
7
9
11
13
15
17
19
Epochs
(b)
Training LOSS
Swin Transformer loss
ResNet loss
MobileNet loss
loss=0.01
loss=0.19
loss=0.34
Loss
1
0
Epochs

Figure 7

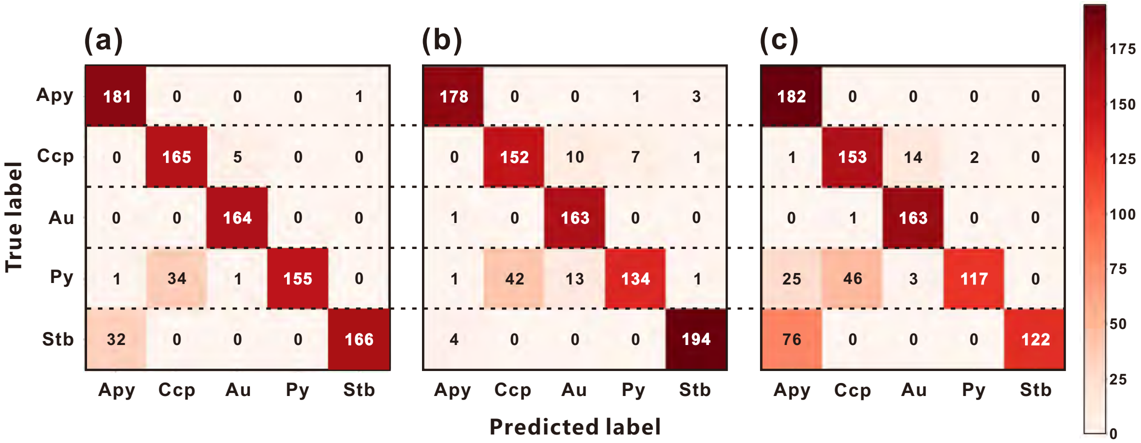
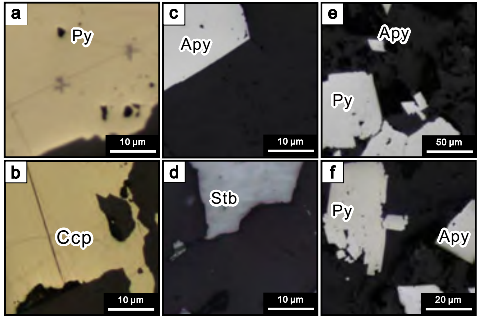

(a)
(b)
(c)
True label
Predicted label
Apy
Ccp
Au
Py
Stb
181 0 0 0 1
0 165 5 0 0
0 0 164 0 0
1 34 1 155 0
32 0 0 0 166
178 0 0 1 3
0 152 10 7 1
1 0 163 0 0
1 42 13 134 1
4 0 0 0 194
182 0 0 0 0
1 153 14 2 0
0 1 163 0 0
25 46 3 117 0
76 0 0 0 122
175
150
125
100
75
50
25
0

Figure 8



a
Py
10 μm
b
Ccp
10 μm
c
Apy
10 μm
d
Stb
10 μm
e
Apy
Py
50 μm
f
Py
Apy
20 μm

Figure 9

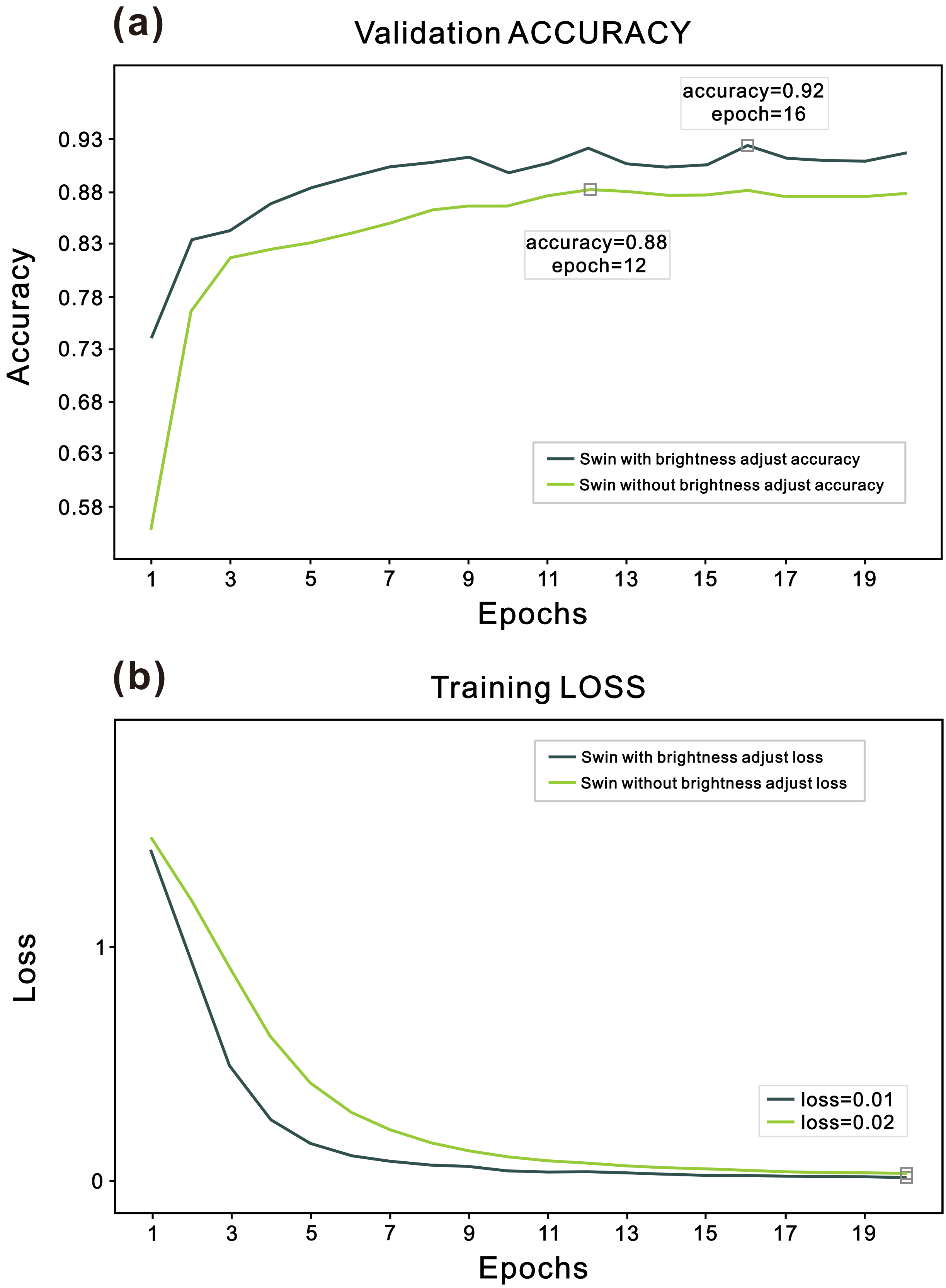
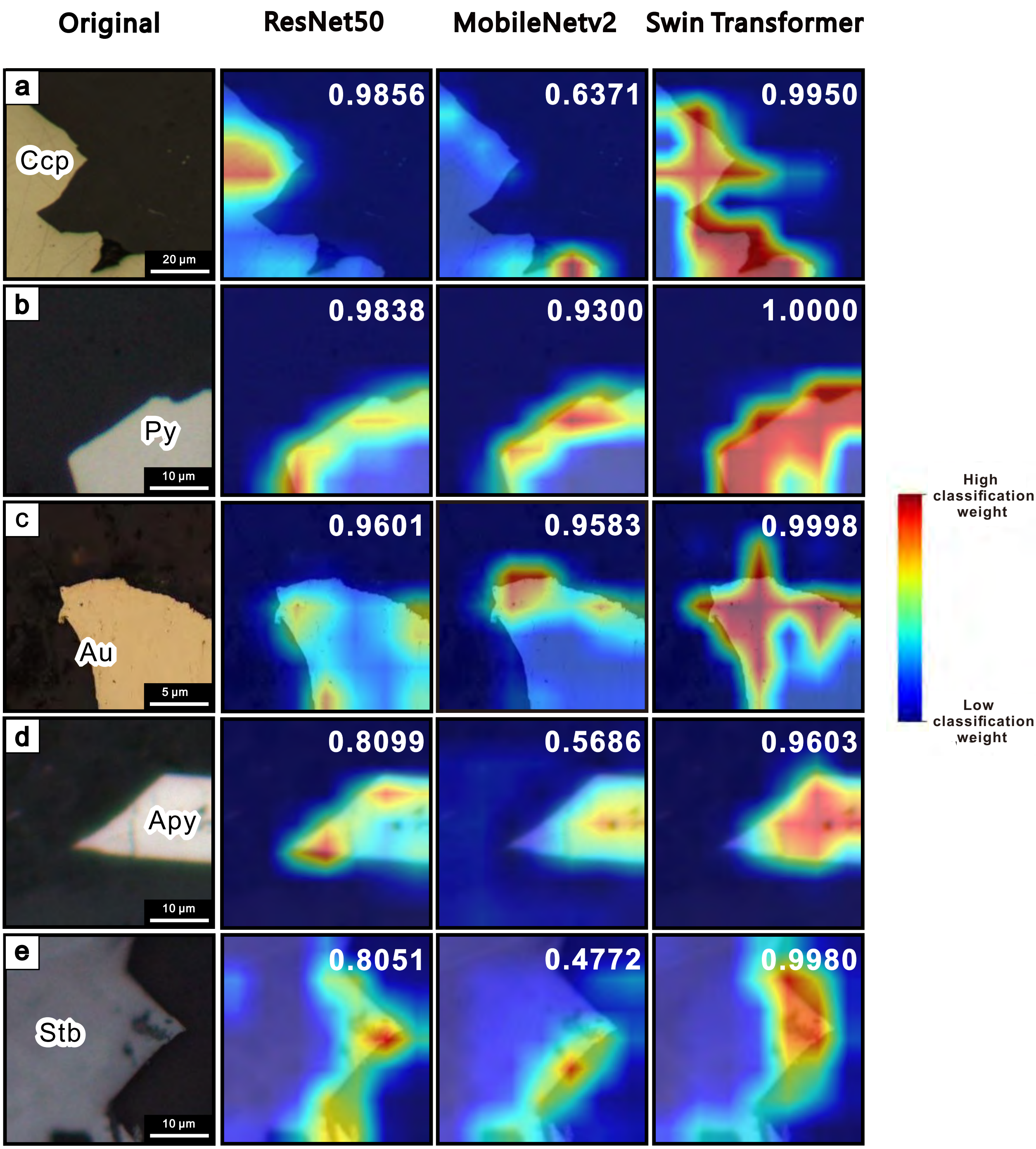

(a)
Validation ACCURACY
Accuracy
0.93
0.88
0.83
0.78
0.73
0.68
0.63
0.58
accuracy=0.92
epoch=16
accuracy=0.88
epoch=12
Swin with brightness adjust accuracy
Swin without brightness adjust accuracy
1
3
5
7
9
11
13
15
17
19
Epochs
(b)
Training LOSS
Loss
1
0
Swin with brightness adjust loss
Swin without brightness adjust loss
loss=0.01
loss=0.02
1
3
5
7
9
11
13
15
17
19
Epochs

Figure 10



Original
ResNet50
MobileNetv2
Swin Transformer
a
Ccp
20 μm
0.9856
0.6371
0.9950
b
Py
10 μm
0.9838
0.9300
1.0000
c
Au
5 μm
0.9601
0.9583
0.9998
d
Apy
10 μm
0.8099
0.5686
0.9603
e
Stb
10 μm
0.8051
0.4772
0.9980
High classification weight
Low classification weight