Computers & Geosciences 178 (2023) 105414

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description

Sergey Nesteruk[a,*], Julia Agafonova[a], Igor Pavlov[a,b], Maxim Gerasimov[a], Nikolay Latyshev[c], Denis Dimitrov[a,b], Andrey Kuznetsov[a,b], Artur Kadurin[b], Pavel Plechov[c]

[a] *Sber AI, Russia*
[b] *AIRI (Artificial Intelligence Research Institute), Russia*
[c] *Fersman Mineralogical Museum, Russian Academy of Sciences, Leninskiy pr. 18-2, 119071 Moscow, Russia*

ARTICLE INFO

Keywords:
Mineral recognition
Computer vision
Zero-shot
Few-shot
Benchmark
Image classification
Image segmentation
Size estimation

ABSTRACT

Mineral image recognition is a challenging computer vision problem. Without external tools, even a human expert cannot distinguish some mineral species accurately. Previous research was mainly focused on processed mineral recognition. This is considered to be a simplified statement of a problem because processed minerals are more visually expressive. On the contrary, in a raw sample, the target mineral can appear in the form of thinly represented inclusions. In real life, the raw samples usually require automatic mineral species identification.

Another difficulty in raw mineral recognition is the shortage of publicly available training and validation data. It is impossible to compare different deep learning approaches when the results are evaluated on dissimilar data.

The main contribution of this paper is providing an open benchmark for zero-shot raw mineral visual recognition. Besides the evaluation-only zero-shot classification dataset, we publish subsets for segmentation, mineral size estimation, and few-shot classification. For all of the provided computer vision problems, we publish baseline solutions we offer for the community to beat.

1. Introduction

About 6000 minerals and their varieties are known in nature. Only a couple of hundred of these are rock-forming, and a few hundred more are of interest to industry. The remaining minerals are so rare that they have so far been found only in single grains. Most mineralogists have never seen these rare minerals or have only seen such minerals in photographs.

Mineral diagnostics is an essential part of geological work. With it, geological mapping and the search for deposits are possible. However, the exact identification of a mineral is a complex and time-consuming task that requires high skill. Geologists spend from 30 min to several days per item on the sample preparations and various types of analysis. With a huge volume of geological samples, any optimization of this process seems attractive. In geology, visual diagnostics of minerals and rocks with further selective instrumental verification is very common. This approach provides a significant saving in time. However, simple visual diagnostics contain 10%–20% errors even if performed by experienced mineralogists. The involvement of machine intelligence in visual diagnostics can help solve two problems: (1) free up the time of professional mineralogists in solving routine tasks, and (2) help identify obvious errors in visual diagnostics. For these purposes, the most relevant is the diagnosis of minerals in raw samples since it does not require additional labor costs for their preparation. If it is possible to solve such a problem and create an algorithm that allows visual diagnostics of minerals at the level of an ordinary geologist, then this would make it possible in the future to create search robots that could automatically explore hard-to-reach places on Earth, and possibly on other planets. The methods of IR spectroscopy, Raman spectroscopy, and remote chemical analysis of rocks and minerals are actively used to study other planets. The addition of the possibility of visual diagnostics of minerals will help in the selection of objects for analysis and will significantly expand the possibilities of research programs.

The contributions of this paper are as follows: we provide an open benchmark for zero-shot raw mineral visual recognition; we publish subsets for segmentation, mineral size estimation, and few-shot classification; for all of the provided computer vision problems, we publish baseline solutions.

[*] Corresponding author.
E-mail address: SVNesteruk@sberbank.ru (S. Nesteruk).

https://doi.org/10.1016/j.cageo.2023.105414
Received 27 February 2023; Received in revised form 4 June 2023; Accepted 13 July 2023
Available online 20 July 2023

The novelty is in the large-scale multi-task dataset accompanied with baseline solutions. To the best of our knowledge, we are the first to apply zero-shot methods for mineral recognition. In the paper we show that for the correct comparison of mineral recognition methods, it is necessary to have unified data for the experiments. Knowledge transfers between mineral datasets poorly because of the domain shift.

Wherein, the most of the papers use different subsets with different sizes. Some of the reasons for that is in limited computational resources of the researchers. Our dataset allows choosing a subset of different size and complexity. Therefore, for the cost of mineral recognition experiments, researchers can find a subset of our dataset that suits their needs.

The rest of the paper is organized as follows. Section 2 provides a comprehensive review of existing mineral image datasets and visual mineral recognition methods. Section 3 describes the presented dataset including data collection, data filtering, image pre-processing, and available annotations. Section 4 provides the results of few-shot and zero-shot mineral classifications. Section 5 shows a baseline pipeline for zero-shot mineral size estimation. Section 6 contains the results of zero-shot mineral segmentation.

2. Literature review

Deep learning methods prove their efficiency for automatic object recognition. In general domain, neural networks are usually easier to apply that in domain-specific cases (Illarionova et al., 2022b). The core limitation for their successful appliance is the shortage of task-specific training data (Nesteruk et al., 2022). The availability of domain-specific datasets is very important (Illarionova et al., 2023). For many computer vision problems, transfer learning from a related dataset is more effective than from the generic ImageNet (Lemikhova et al., 2022). But even with model pre-training and proper data augmentation, we still need sufficient number of training samples from the target domain (Illarionova et al., 2022a).

In this section, we overview the existing mineral recognition datasets and papers on automated mineral recognition. We show that today there is no clear benchmarking for mineral image analysis and the importance of the benchmarking system.

2.1. Mineral image datasets

Deep learning models need much training data to exploit their potential. However, besides the quantity of the data, its quality is also crucial. The key assumption in practical DL is that during the model inference, the data distribution will be similar to the one in the test set (Nesteruk et al., 2023). Therefore, not any data is equally useful. In the mineral recognition domain, the primary application of automated species identification is for the raw samples in the wild. We review existing datasets and assess them by their size and complexity in Table 1. We compare the number of classes, the number of images, and whether the samples are mostly raw or processed. Additionally, we specify if a dataset has auxiliary information for tasks other than classification. The formed open-source datasets are marked as ready for download. Other data sources that require site parsing and extensive data preprocessing are marked as not ready datasets.

In the Gemstones dataset (Chemkaeva, 2019) the samples are faceted. There are over 3200 images and 87 gem classes.

The Mindat dataset (Ralph, 1993) consists of both raw and processed minerals. It is the largest dataset, with over one million images of 5848 mineral species. Most of the samples in this dataset are very visually expressive. This collection has many extraordinary specimens, and usually, they are clean and polished. The main mineral is usually dominant in the image, and the recognition is not complicated by satellites or inclusions.

The Minerals Identification Dataset (Brempong, 2019) has 7 classes, and over 700 images in total. All the specimens in this dataset are processed.

Natural Diamonds dataset (Lakhani, 2020) has over eight thousand images of diamonds. It has only one gemstone class, but the images are split according to nine cut types.

Rock Classification dataset (Hossain et al., 2021) contains over two thousand images of seven raw mineral classes.

Mineral, Rock, and Fossil dataset (NIMRF, 2003) is a collection of 245 mineral classes. It has over 64 thousand images, but they are protected with a watermark. This makes them less suitable for training and testing.

Our dataset has over three thousand mineral species. We share more than a hundred thousand images. The samples in our dataset are unprocessed. They are the closest to real-world applications. Beyond images and class labels, we publish auxiliary annotations for some samples. This includes mineral masks, mineral sizes, and text descriptions.

In Fig. 1 one can find examples of images from three datasets: The gemstones dataset, the Mindat dataset, and our dataset. These examples showcase the difference between them. In addition to the dataset shift, one can see high inter-class variability even within a single dataset.

In some papers on visual mineral recognition, authors use private and small datasets they do not publish. One of the core problems in the mineral recognition domain is the lack of clear benchmarking. It is incorrect to compare the performance of two methods if it is measured on different test sets. Therefore, most of the existing papers cannot be compared with each other.

2.2. Mineral visual recognition

The approaches to automating visual mineral recognition can be split into two groups by the type of input data. The approaches of the first group require special equipment and domain-specific knowledge (Plechov et al., 2019). The second group uses only RGB images from regular cameras.

It is natural to expect higher results from the methods that utilize special equipment. In Baykan and Yılmaz (2011) authors apply multi-layer perceptron to microscopic images with polarized light and achieve 94% accuracy on 5 mineral classes. In Izadi et al. (2017) authors explore a similar type of input data and achieve 93% accuracy on 23 mineral classes with a cascaded model approach. In Maitre et al. (2019) they combine optical microscopy with automated Scanning Electron Microscopy (SEM) (Gottlieb et al., 2000) and propose a pipeline for individual mineral grain segmentation. Their pipeline includes simple linear iterative clustering (SLIC) (Achanta et al., 2012) for unsupervised image segmentation, feature extraction based on color space change, and multiple machine learning algorithms for superpixel classification, and it reaches 90% accuracy on 27 mineral classes. In Yousefi et al. (2020) authors successfully applied machine learning to the Infrared longwave spectra for automated mineral identification. In Zhang et al. (2019) it is shown that deep learning methods can be efficiently combined with machine learning for mineral recognition problems. The authors use Inception-v3 (Szegedy et al., 2016) deep convolutional model for feature extraction with a stack of machine learning algorithms for final classification, and state 90% accuracy on 4 classes. In Hao et al. (2022), authors apply a metric learning paradigm for heavy mineral grains classification.

Another popular data type in mineral recognition is Raman spectroscopy. Raman spectra are usually analyzed either by peaks matching (Li et al., 2020) or by full spectrum matching (Carey et al., 2015).

Both microscopy and spectrometry approaches show significant accuracy, but their practical application is limited. While they can produce micron-scale resolution (Kim et al., 2021), these methods are difficult to scale. Massive equipment and time-consuming data collection process (Bukharev et al., 2018b) lead to low-class coverage for model training and complexity for model inference (Bukharev et al., 2018a).

On the contrary, regular macro-scale images are easy and cheap to obtain (Baraboshkin et al., 2020). This is crucial in many practical

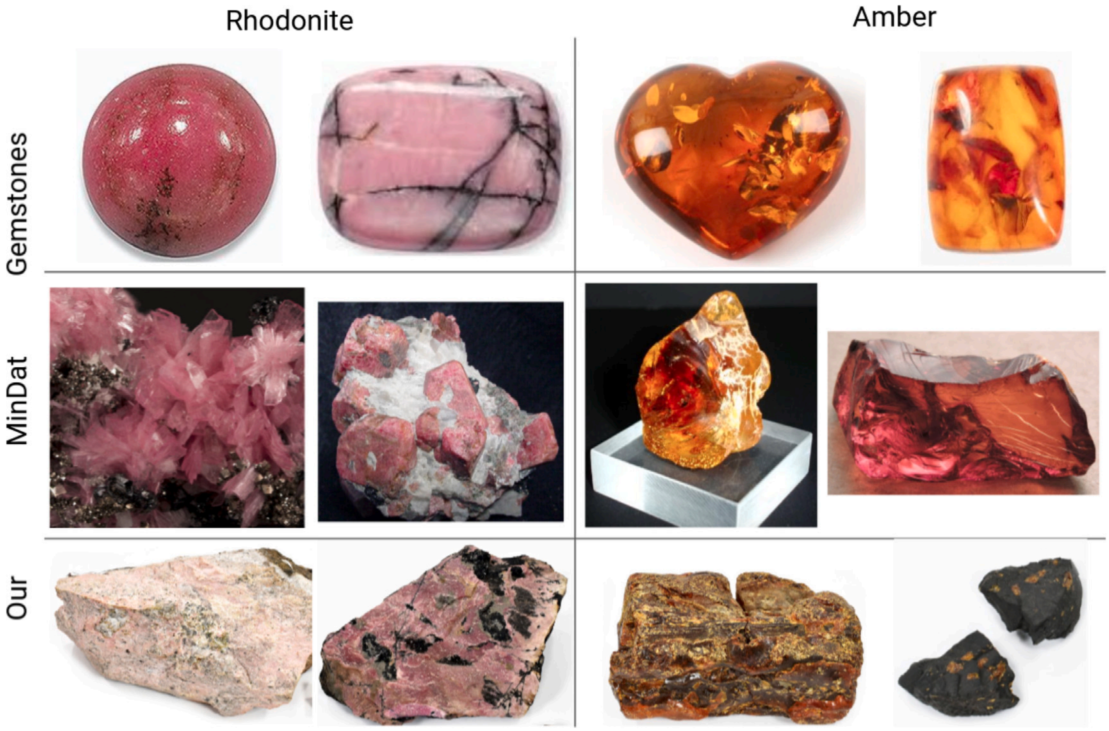

Fig. 1. Image examples from different datasets.

Table 1
Mineral recognition datasets.

Dataset	#Classes	#Images	Samples condition	Samples type	Auxiliary labels	Ready for download
Gemstones images (Chemkaeva, 2019)	87	>3200	Processed	Gem	No	Yes
Mindat (Ralph, 1993)	>5K	>1.2M	Processed&Raw	Mineral&Gem	No	No
Minerals identification (Brempong, 2019)	7	>700	Processed	Mineral	No	Yes
Natural diamonds (Lakhani, 2020)	1[a]	>8K	Processed	Gem	No	Yes
Mineral, Rock and Fossil (NIMRF, 2003)	245	>64K	Raw	Mineral	No	No
Rock classification (Hossain et al., 2021)	7	>2K	Raw	Mineral	No	Yes
Ours	>5K	>44K	Raw	Mineral&Gem	Yes	Yes

[a] 9 classes of cuts within 1 specie of a gemstone.

use cases (Jin et al., 2022). Experiments show that for some cases image-based methods can successfully solve mineral recognition problem without expensive equipment. Azarafza et al. (2021). In addition, image analysis enables the determination of both the geometry and structure of mineral materials, not just their composition. Azarafza et al. (2019). It makes image-based mineral recognition a perspective area of research (Patel et al., 2017). It remains very challenging due to a large number of mineral classes and high inter-class variability (Ivchenko et al., 2018), and many recent pieces of research started to address this issue. Chanou et al. (2014) showed the importance of texture features for individual rock classification. In Peng et al. (2019) authors use Inception-v3 to raw images of 16 mineral classes and reach 86% accuracy. In Zeng et al. (2021) authors propose to use multiple data types to obtain higher accuracy. They combine images with Mohs hardness (Wenk and Bulakh, 2016) and increase accuracy on 36 Mindat classes from 71.2% to 89%. This is a considerable improvement; however, in practice, its application is limited because determining Mohs hardness requires special tools and scratching the examined sample. In Liu et al. (2019) they decided to generate a new input type from an image itself. They extract the Histogram of Oriented Gradient features (HOG) (Dalal and Triggs, 2005) from the original image and combine it with the raw image using a series of machine learning and deep learning algorithms. The authors claim that this approach increases the accuracy from 64.1% to 74.1% on 12 mineral species classification problems. In Jia et al. (2021) authors utilize multi-level image features for 22 mineral species classification and reach 88% accuracy. In Shu et al. (2017) they utilize unlabeled data to simplify the application of a rock classification pipeline.

In summary, advanced data collection methods can utilize simple data processing algorithms, while simple data collection methods require more complex computations. Image collection is the easiest approach for collecting mineral data, and classical or deep learning methods can be used. Classical methods are lightweight and do not require a GPU (Nesteruk et al., 2020), but deep vision models are more robust and accurate, especially for a high number of classes (Nesteruk et al., 2021). Inception-v3 is the most popular choice for mineral recognition among deep vision models due to its ability to analyze images on different scales and higher input resolution, which is crucial for detecting variations in textures, geometry, satellite materials, and colors in mineral images.

As one can see, many research papers on mineral identification published over the last decades. However, most of them use their private and non-reproducible datasets (Peng et al., 2019), and therefore cannot compare the results with the previous research. Also, datasets are usually very small (Singh et al., 2010) to declare confident results, and need to prevent overfitting (Xu et al., 2021). In this paper, we make a step to eliminate this problem. More precisely, we shape a new large mineral identification dataset and show how it can be used along with other existing datasets to assess mineral recognition models' quality.

3. Dataset collection

We used the dataset of the Fersman mineralogical museum (http://www.fmm.ru). The museum funds contain more than 170 thousand samples (about 5000 mineral species). This is one of the largest mineralogical collections in the world. The first electronic database containing sample descriptions was created in 1996. For several years, records from old file cabinets were transferred to it. In 2017, the Museum created a new information system based on the semantic web (Plechov et al., 2019). The new system is not limited by the amount of information on a single exhibit. All data from the old system is automatically translated into the new one. In 2018–2019, the Museum began systematically revising the collection, and old photos were added to the information system for about 3000 samples. In 2021, all the necessary equipment for photography was purchased and mass photography of samples began by museum curators and about ten volunteers. At the moment about 4000 samples are photographed every month (40 thousand per year). Photographs of the first 50 thousand samples were used for MineralImage5k and we hope to check and refine the model over the next few years.

Our sample contains fewer images than Mindat, but it has several advantages for solving the problem. First, it is more homogeneous since the photographs of the samples were taken under similar conditions by a small team of photographers in a short period of time. The original images were used and not enhanced with Photoshop or other graphic editors. No watermarks or other graphic elements have been inserted into the images. Secondly, our dataset gives a broader view of mineral diversity, as it contains not only the best samples of any mineral species but also many hundreds of representative samples. Thousands of collectors add Mindat images and try to select only the best crystals. In nature, such outstanding samples are extremely rare, and minerals are most often observed in nondescript clusters. Thirdly, our dataset is linked to the museum's collection, which makes it possible to further study the sample and correct its description in case of questions. During this work, dozens of inaccuracies in the description of the samples have already been identified, and the connection of the dataset with specific samples available for study made it possible to correct these descriptions. Only a dataset of photographs exists in Mindat, and the samples themselves are not available for study. It remains to take the word of those lovers of minerals who posted a photo that it is exactly the mineral that they indicated. Mineral lovers are most often limited only to visual diagnostics, and it can lead to 10%–20% errors, as noted above. Also, our dataset contains raw samples, not polished or faceted minerals, unlike the Gemstone Images dataset. On the one hand, this makes the recognition task more difficult, but on the other hand, such samples are much closer to those found in mountains or riverbeds.

Fig. 2 reflects the overall image preprocessing scheme. Its purpose is to make a cleaner version of the dataset. We describe the data preprocessing pipeline in detail because it can be useful for other similar datasets.

On the high level, the pipeline is intended to:

- remove uninformative images from the dataset;
- removes uninformative areas from the remaining images.

First, we remove corrupted images. Then, we remove images with high aspect ratios because most of the computer vision models work with square inputs. If the difference between the image sides is too high, one should either add padding or make a square crop. If this is a rare case in a dataset, it can be more suitable to remove such images.

The next important step is to remove duplicate images. Here we consider similar images, but not only exact copies to be duplicated. It is often the case to have multiple images of the same object with slightly different viewpoints. For most computer vision problems, such near-duplicated images must be treated carefully. If there are duplicates in the training set, they do not contribute new information. It is especially evident in a few-shot learning scenario. However, it is more dangerous if two similar images land in different subsets: a training set and a testing set. Deep learning models tend to overfit the data in the training set, most significantly when the training set is small. In this case, testing on the same image will produce overstated results. A researcher must protect experiments from unreasonably optimistic conclusions consequent from data leakages. In our pipeline, we apply perceptual hashing algorithm (pHash) (Niu and Jiao, 2008) to find visually close images in the dataset. This is an efficient algorithm that can be used on the large scale. For smaller datasets, an alternative might be the MS-SSIM (Rouse and Hemami, 2008) algorithm.

The next step in our pipeline is to resize the remaining images. We set the biggest side of an image to be 1024 pixels.

Ideally, we want a dataset to be as representative as possible. An image in a mineral classification dataset should contain a mineral of a known class. Unrelated objects waste image space. As we show later in this paper, minerals often have essential small details. Therefore, it is better if the mineral occupies the whole image space. Unfortunately, in real-world datasets, images are usually imperfect.

To eliminate this issue, we propose the following solution. We apply pre-trained object detectors to find key objects on an image and process them accordingly. Three main types of objects in our dataset are:

- a mineral;
- a reference cube;
- a text plate.

Text on the image can be harmful because some neural networks are able to read it. For the training and evaluation processes, it means that a model can learn and predict the text instead of recognizing a mineral itself. For text detection, we use the CRAFT model (Baek et al., 2019). It produces a heatmap that has high values around letters. We transform this heatmap into a box that covers all the detected text. Then we simply overlay this area with a uniform box.

Two other types of objects that interest us are a mineral itself and a reference cube. The reference cube is a metal cube with a known size. Later we use it for mineral size estimation. At this stage, we cannot distinguish a mineral from a cube because we did not find a pre-trained model that can distinguish these classes. But we still can detect both of them without specifying a class. For this purpose, we apply the OWL-ViT model (Minderer et al., 2022). It allows using an open-vocabulary request for object detection. For each image, we request the following phrases: "a stone", "a rock", "a mineral", "a gem", "a crystal", and "a mineral ore". We need any of the above requests to trigger. If there are multiple detections, we choose the dominant ones via non-maximum suppression (Neubeck and Van Gool, 2006). The resulting image (Fig. 2) is a crop of the original image that contains only detected objects.

Having the filtered and cropped images, we use them to form multiple datasets (Table 2). The major one is a zero-shot classification dataset. It includes all of the images. The primary goal of this subset is to evaluate pre-trained models.

We also share several subsets for few-shot classification. They have 360, 98, and 10 classes accordingly. They are already split into training and evaluation subsets.

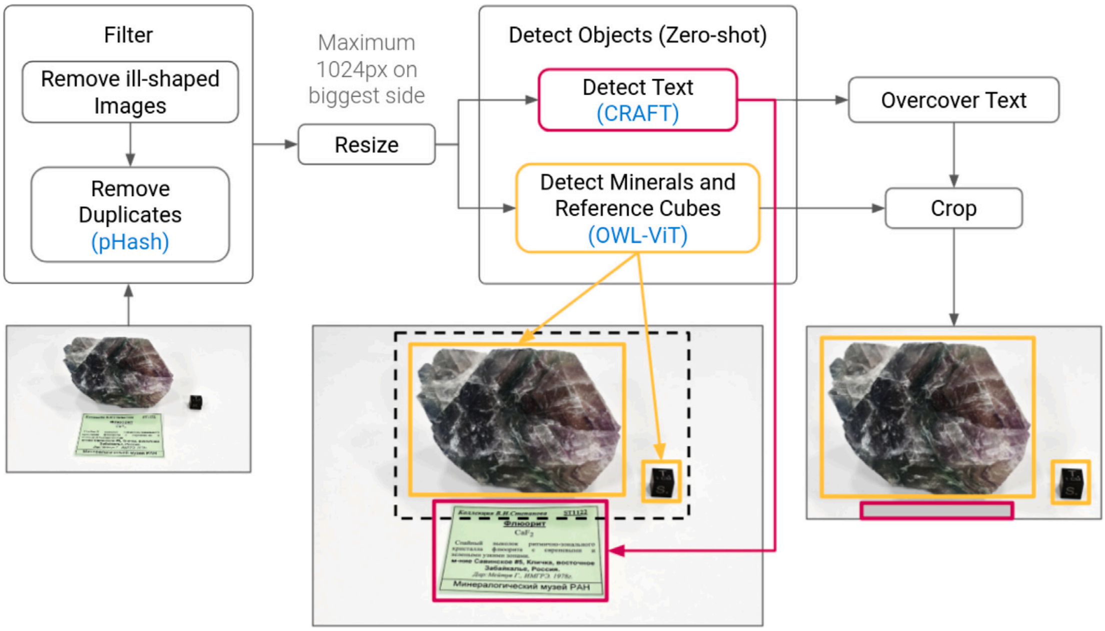
Filter
Remove ill-shaped Images
Remove Duplicates (pHash)
Maximum 1024px on biggest side
Resize
Detect Objects (Zero-shot)
Detect Text (CRAFT)
Detect Minerals and Reference Cubes (OWL-ViT)
Overcover Text
Crop
Флюорит
CaF₂
Минералогический музей РАН

Fig. 2. Data Preprocessing Scheme.

Table 2
The subsets of our mineral recognition dataset.

#Classes	#Images	Minimum images per class	Task	Subsets
5139	44 784	1	Classification	Test
360	31 982	17	Classification	Train&Test
98	23 496	78	Classification	Train&Test
36	16 001	187	Classification	Train&Test
10	8549	462	Classification	Train&Test
–	>100	–	Segmentation	Test
–	18 076	–	Regression (size estimation)	Test

Additionally, we have a 36-classes dataset that has the matching subsets in Mindat and Gemstones Images datasets. Their purpose is to provide a cross-dataset comparison.

The auxiliary segmentation dataset has manually labeled mineral and satellite masks. The auxiliary regression dataset has manually measured mineral sizes.

4. Mineral classification

4.1. Few-shot mineral classification

In Table 3 we report the classification results for 10, 98, and 360 classes datasets. We compare different types of models and various input resolutions. The reported metric is top-N accuracy with $N \in \{1, 3, 5, 10\}$. Conventional classification models sort the possible answers according to their confidence. Top-N metrics for a given sample consider being 1 if the correct answer is within the model's first N predictions, and 0 otherwise.

We compare the following classification models: ResNet50 (He et al., 2016), HRNet-W18 (Wang et al., 2020), and Google ViT-L/16 (Dosovitskiy et al., 2020). The resolutions vary from 336 to 768 pixels. The results show that higher resolution consistently provides higher accuracy. This may indicate that mineral samples have important small details.

The training pipeline for classification models is as follows. First, we split the dataset into 5 folds to avoid overfitting on the validation set. The folds are stratified (Diamantidis et al., 2000). All of the models are trained with AdamW optimizer (Kuen et al., 2019) for 20 epochs. We use one cycle learning rate scheduler (Smith and Topin, 2019) starting from ImageNet pre-trained checkpoint, an initial learning rate of $2e^{-5}$, and the scheduler division factor of 25. The model is trained with cross-entropy loss function (Good, 1952). With the A100-SXM4-80 GB card, we have a batch size of 80.

Table 3
Mineral few-shot classification metrics.

Model	Image resolution	#Classes	Acc@1	Acc@3	Acc@5	Acc@10
ResNet50	336	10	58 ± 1	87 ± 1	95 ± 1	100
ResNet50	576	10	62 ± 2	89 ± 1	96 ± 1	100
HRNet-W18	576	10	62 ± 2	87 ± 1	95 ± 1	100
ResNet50	768	10	66 ± 2	90 ± 1	96 ± 1	100
ResNet50	336	98	37 ± 1	57 ± 1	65 ± 1	76 ± 1
Google ViT-L/16	384	98	41 ± 2	58 ± 1	65 ± 1	74 ± 1
ResNet50	576	98	44 ± 2	62 ± 1	70 ± 1	80 ± 1
HRNet-W18	576	98	39 ± 1	58 ± 1	66 ± 1	76 ± 1
ResNet50	768	98	43 ± 1	63 ± 1	71 ± 1	81
ResNet50	336	360	29 ± 1	44	51 ± 1	61 ± 1
Google ViT-L/16	384	360	33 ± 2	47 ± 1	54 ± 1	62 ± 1
ResNet50	576	360	33 ± 1	49 ± 1	56 ± 1	66 ± 1
HRNet-W18	576	360	30 ± 1	46 ± 1	53 ± 1	62 ± 1
ResNet50	768	360	34 ± 1	50 ± 1	57 ± 1	66 ± 1

4.2. Zero-shot mineral classification

Most of the classes in our dataset do not have enough samples to train a classifier. However, we can use them for model evaluation. For this purpose, we use the CLIP model (Radford et al., 2021). It is pre-trained on a large dataset of text-image pairs. In Fig. 3 we provide the results of the model without fine-tuning any mineral domain-specific dataset.

We calculate top-N accuracy for the 36-classes dataset and the whole 5193-classes dataset. Also, for the comparison, we plot the accuracy of a random guess for 5193 classes. As one can see, it is possible to have a zero-shot model which performs much stronger than a random guess. The top-1 accuracy of the model is 6.5% which is 3250 times better. The CLIP model works by comparing an image with a text string in a natural language. It allows using real mineral names instead

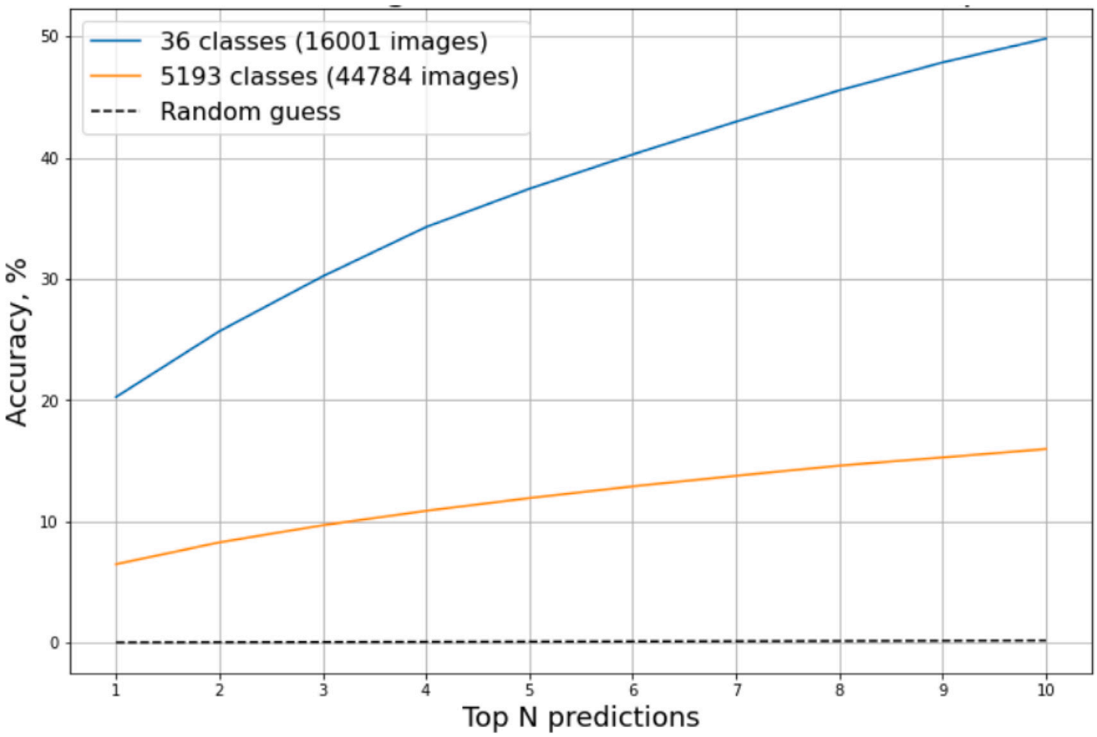

Fig. 3. Zero-shot mineral classification results.

Table 4
Cross-Dataset classification metrics.

Model	Training dataset	Test dataset		
		Gemstones	Mindat	Ours
ResNet50	Gemstones	**91.8**	22.4	20.5
ResNet50	Mindat	35.3	78.7	54.1
ResNet50	Ours	24.1	40.8	**64.1**
CLIP	–	39.4	20.2	20.7
CLIP	Mindat	36.5	**87**	38.4

of one-hot representation. This approach is the most suitable for the real-world scenario.

This technique involves identifying the most suitable text description for an image among the given options. Therefore, the main limitation is that the correct mineral name must be in the list of the provided options. Nonetheless, this is not a significant issue as there exists a comprehensive list of all known mineral varieties.

4.3. Cross-dataset evaluation

When we have a model for mineral classification, we need it to work in a real-world scenario. If we assume that the data distribution during the inference is similar to the data distribution in the test set, we can say that the result in the test set is the approximation of the real-world performance. Unfortunately, in practice, data can differ significantly. To prove it, we show Table 4. In this table, we selected 36 overlapping classes in three datasets. As one can see, all of the tested models poorly generalize to new datasets with the same classes.

5. Zero-shot mineral size estimation

Beyond classification, we study zero-shot mineral size estimation. Automatic specimen size estimation is important for museum specimen storage procedures. We can get the approximate length and width of the sample from the photo, which corresponds to the minimum size of the box in which this sample can be stored. Having these data for all samples, we can think over the optimal storage system, as well as purchase or manufacture boxes of the right size in the correct quantity. Now, measuring the dimensions of the sample is carried out using a ruler manually, it takes a lot of time, and the work is far from complete. The results of the work described in this section will be immediately used in the work of the museum. In addition, there is a possibility of detecting errors when entering dimensions that have already been measured manually. Usually, zero-shot means that a model was not trained on the current dataset. For the baseline, we propose a model-free pipeline (Fig. 4). The key idea is to estimate the relative mineral size and compare it with an object of a known size.

The pipeline uses bounding boxes obtained during dataset preprocessing (Section 3). The size estimation pipeline starts with assigning classes to the bounding boxes. We need to distinguish two classes: a mineral and a reference cube. Here we assume that text plates are already filtered out in the previous step.

We assign a bounding box as a reference cube if at least one of two tests shows that it is a reference cube. The tests are based on the fact that raw minerals do not have a perfect cubic shape, but the reference cube does.

The first test is based on perimeter analysis. For this method, we need to find a contour of an object. We use the border following algorithm (Suzuki et al., 1985) to find the contour. Image preprocessing for this algorithm includes conversion from an RGB image to a grayscale image, image binarization, and application of morphological operations (Maragos, 1987).

Having the contour of the object, we compare how different it is from the contour of a perfect square. We assume that the contour is likely to be a square if it satisfies Eq. (1).

$$\left|\frac{L^2}{S} - 16\right| \leq m, \tag{1}$$

where L is the length of the contour, S is the area, and m is a margin. The margin can be greater than zero for a cube for two main reasons. First, a cube can be rotated, but we treat it as a two-dimensional

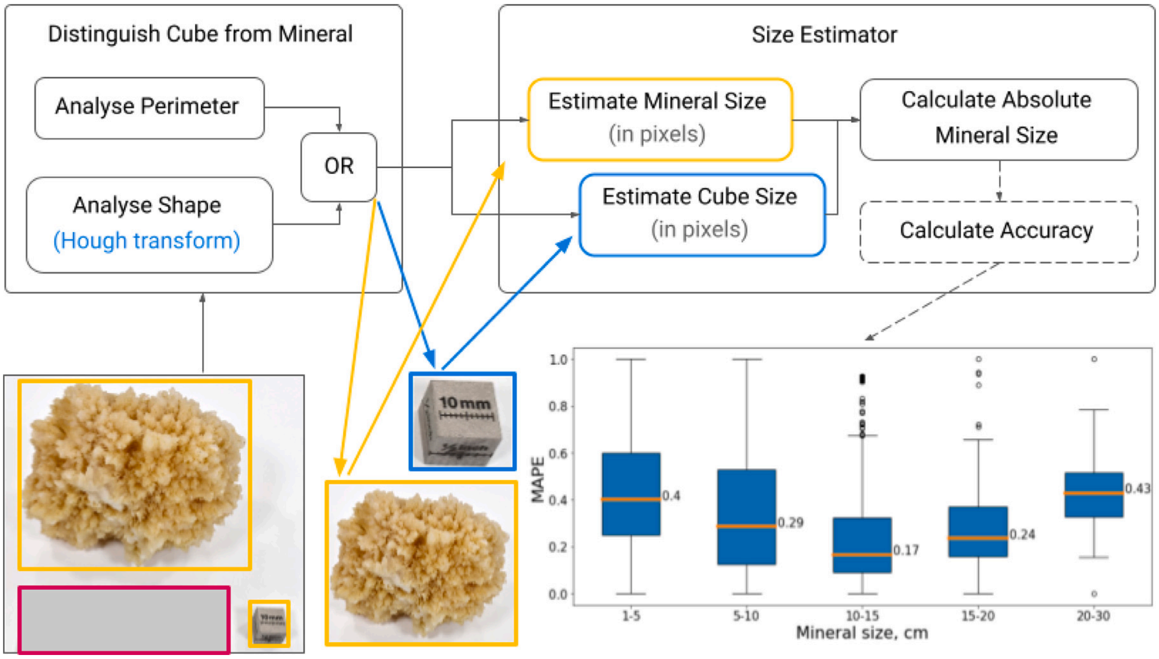

Fig. 4. Zero-shot mineral size estimation pipeline.

object for simplification. Second, the found contour can be not perfectly accurate. We set $m = 1$.

The second test is based on shape analysis. It tries to find an even number of pairs of parallel lines on an image. For this purpose, we use the Hough transform (Illingworth and Kittler, 1988). More precisely, we apply Hough line detection algorithm (Duda and Hart, 1972) and set additional constraints to select pairs of lines that have a small difference in orientation.

If an object is a reference cube, the other object is a target mineral. The images with more than two bounding boxes are eliminated for this stage. The last step is to compare the sizes of two bounding boxes. The size of the reference cube is 1 inch, and it is fixed. According to it, we can estimate the size of one pixel in centimeters for the current image. Having the size of a pixel and the number of pixels in the detected bounding box, we calculate the absolute size of the mineral.

The main limitation of this pipeline is that it requires a reference object on the image. Among the advantages are the simplicity and the ability to filter some of the bad cases. The computational simplicity allows using only the CPU. The self-filtering means that an algorithm can distinguish some of the images where it should not be applied. We define two types of such cases:

- no reference cube was found on the image;
- the size of the reference cube differs too much from the predicted size of a mineral.

To skip a mineral according to its predicted size, it must be either 2 times smaller that the reference cube or 25 times bigger. The lower limit is set as just 2 times the difference because small samples are usually stored in capsules, and the zero-shot bounding estimator will trigger on a capsule rather than a mineral inside it. We report the results excluding automatically filtered images. The metric is mean absolute percentage error (MAPE). As Eq. (2) shows, it provides an error in percent and allows comparison errors for the samples with different scales.

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\frac{|T_n - P_n|}{T_n}, \tag{2}$$

where N is the number of samples in the test set; T_n is the true size value of the nth sample; P_n is the predicted size value of the nth sample.

In Fig. 4 one can see MAPE for different groups of mineral samples split according to their true size. In our dataset, we have ground truth size labels for 18 076 images. Among them, 7637 were not filtered out by the algorithm described above. The MAPE on the remaining samples is 0.378.

6. Zero-shot mineral segmentation

In Liu et al. (2021) and Jia et al. (2021) they already proposed to apply GradCAM-based methods for mineral classification model visualization. However, to the best of our knowledge, we are the first to propose zero-shot mineral segmentation via GradCAM.

GradCAM is a method that attempts to explain the prediction of a classifier model for a particular image. The idea behind it is to identify the pixels that strongly influence the classifier decision. We assume that important pixels are more likely to correspond to the target object. On the contrary, less important pixels are more likely to belong to the background. To determine important pixels we calculate the gradient of the class score function with respect to the input image. The described approach produces a coarse localization map that highlights the important regions.

One can see the overall mineral segmentation pipeline in Fig. 5. We propose two major options. The main zero-shot pipeline uses only mineral classification labels without any segmentation annotations. The auxiliary few-shot variant can use available segmentation data to tune the original pipeline.

The core of the pipeline is a classifier model. We use the best model from the mineral classification experiment described in Section 4. It is ResNet50 trained on 360 mineral classes with an image resolution of 768 × 768 pixels. A very important part of the unsupervised algorithm is the ability to filter the cases, where it works poorly. The first filtering line in our algorithm is to exclude images, where the classifier predicts wrong answers. Next, we obtain class activation maps and score them with RoadCombined (Rong et al., 2022) metric. We use only samples with $RoadCombined > 0.2$ to filter out poor visualizations. The last step is to binarize a class activation map. To find an adaptive binarization rule without segmentation annotations, we apply Otsu thresholding algorithm (Otsu, 1979). It is unsupervised and non-parametric. If one has training segmentation masks, they can be used to tune the threshold by multiplying it with some constant.

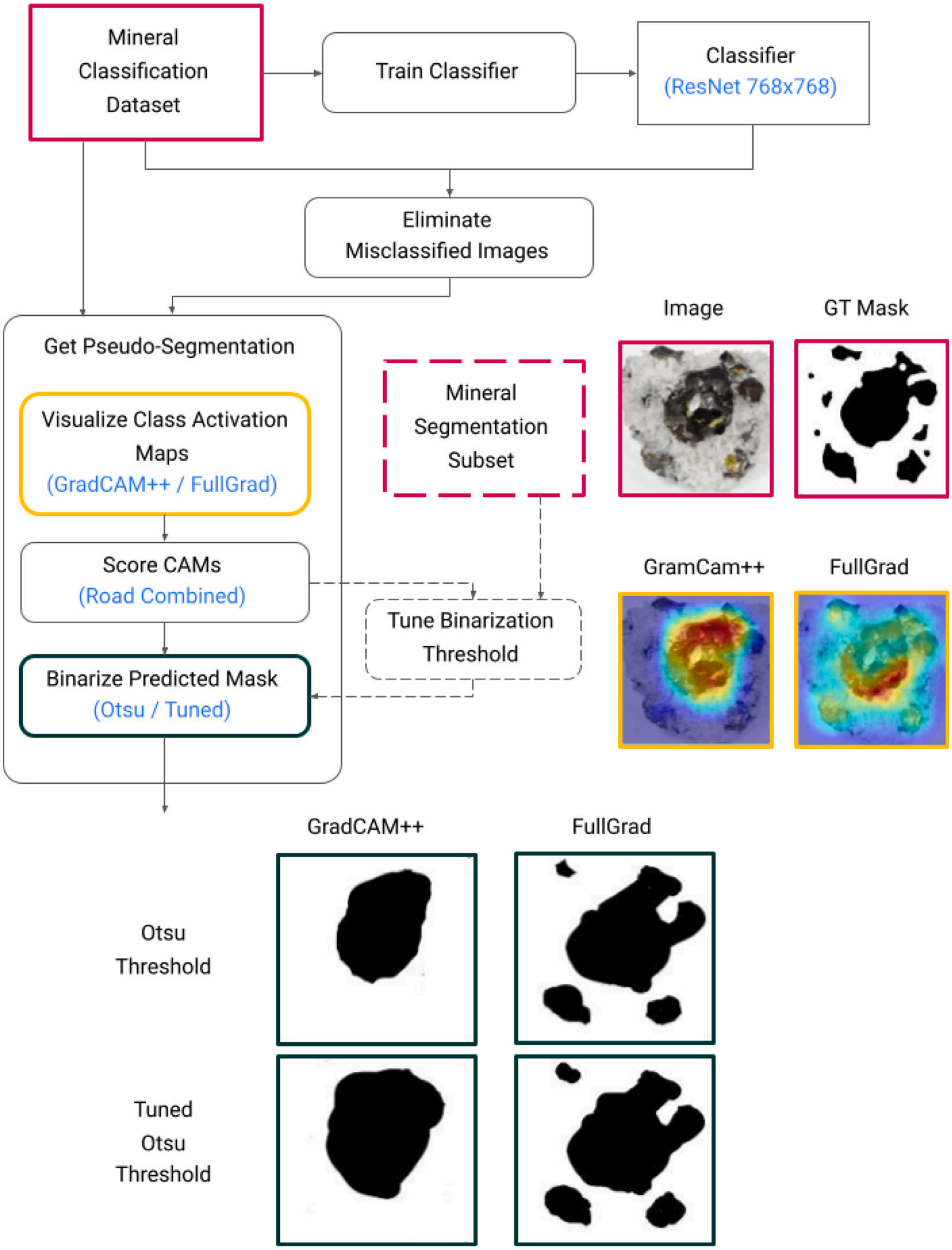

Fig. 5. Mineral Segmentation Scheme.

In Table 5 we report the results of the zero-shot segmentation pipeline. We tested several GradCAM-based methods, and show the results for the top two ones, namely FullGrad (Srinivas and Fleuret, 2019) and GradCAM++ (Chattopadhay et al., 2018). FullGrad computes gradients from all layers of the model and sums them up to produce more detailed results. GradCAM++ uses second-order derivatives. For both methods, we additionally apply eigensmoothing (Smilkov et al., 2017) to remove noise.

The metrics are: precision (Eq. (3)), recall (Eq. (4)), F1-score (Eq. (5)), IoU (Eq. (6)), and RoadCombined. Note that RoadCombined is not a segmentation quality metric. It is an intermediate metric that shows how important are the selected pixels for the classifier.

$$P = \frac{TP}{TP + FP}, \quad (3)$$

where TP is the number of correctly segmented (true positive) pixels; FP is the number of false positive pixels.

$$R = \frac{TP}{TP + FN}, \quad (4)$$

where FN is the number of false negative pixels.

$$F1 = 2\frac{P \cdot R}{P + R}, \quad (5)$$

$$IoU = \frac{TP}{TP + FP + FN}, \quad (6)$$

One can see that FullGrad has better overall performance, but GradCAM++ has higher precision. Also, FullGrad is very time-consuming. Therefore, the specific method's choice may differ depending on the application.

Table 5
Zero-shot mineral segmentation metrics.

Method	F1-score	Precision	Recall	IoU	RoadCombined
FullGrad	**0.78** ± **0.16**	0.67 ± 0.19	**0.98** ± **0.04**	**0.66** ± **0.18**	0.34 ± 0.13
GradCAM++	0.63 ± 0.13	**0.79** ± **0.16**	0.56 ± 0.17	0.47 ± 0.14	0.41 ± 0.03

The suggested method is zero-shot as it does not necessitate the use of segmentation data for training. However, it does rely on classification annotations since the classification model training entails learning class activation maps. This approach can be further improved with weak supervision and zero-shot segmentation algorithms (Mukhamadiev et al., 2023).

7. Conclusion

Mineral recognition in the real-world scenario remains a challenging problem. To help a geologist or a non-professional with its automation, we need a robust and non-invasive method. The most affordable approach is to apply computer vision. A regular camera available in any smartphone is enough to examine a sample without destruction.

However, many complications do not allow us to obtain a perfect score. They include high interclass variability in conjunction with the low difference between some classes. A raw mineral can in addition be covered with dust or it can be presented as tiny inclusion in a sample. The single most significant difficulty is the lack of training and testing data. Moreover, we show that existing datasets are very different from each other, and knowledge transfers poorly between them. The inference data distribution must match the training one to avoid overstated results. Another issue is that most of the available datasets are small. Unfortunately, good results on several most popular mineral classes do not guarantee scalability.

To overcome the stated limitations, we share the MineralImage5k dataset. It provides more than 5 thousand classes and a total of 44 thousand images. The classes of minerals are verified by geologists. Some of the images have information about mineral variety, and size and are accompanied by a segmentation mask. Some images have a detailed natural language description. The datasets mostly consist of unprocessed samples that are close to the minerals in the wild.

We cannot split classes with rare mineral species into the training set and the testing set. Instead, we introduce multiple subsets of our dataset for different purposes. The whole 5K+ classes dataset is designed for the evaluation-only task. One can use other datasets for training and measuring the performance of our dataset. Three other subsets are both for training and testing. The difference between them is in the number of mineral classes. The subset with 360 classes is guaranteed to have at least 17 images per class, the subset with 98 classes is guaranteed to have at least 78 images per class, and the subset with 10 classes is guaranteed to have at least 462 images per class.

To prove that it is possible to have a zero-shot baseline which is much stronger than a random guess, we evaluate a vision-language model pre-trained on the data from the general domain. We also report that fine-tuning this model on a domain-specific dataset considerably increases the accuracy. Our classification experiments show that cross-dataset evaluation is a promising way to assess a mineral recognition model.

Beyond the classification task, we show zero-shot pipelines for mineral segmentation and size estimation. An important feature of these pipelines is that they can filter the cases for which they perform poorly. The ability to produce trustworthy predictions is useful in practice. It helps to specify the limits and to separate samples that can be processed automatically from the samples that require manual control.

The proposed baseline approaches have their limitations. Zero-shot classification implies that correct mineral name must be in the list of the provided options. Zero-shot size estimation requires a reference object on the image. Zero-shot segmentation rely on classification annotations. Our dataset aims to support further research in the field of mineral recognition, and we encourage the scientific community to develop novel approaches that can enhance the accuracy of our results.

This dataset is challenging because it is class-imbalanced, and in many cases it is not trivial to distinguish mineral varieties. In future work, we are going to increase the dataset. We suggest that future studies should focus on developing more diverse datasets that include a wider range of minerals and rocks, as well as incorporating other types of data to improve accuracy. We also recommend exploring new approaches that can address the limitations of manual annotations, such as using semi-supervised or unsupervised learning techniques. Finally, we emphasize the importance of collaboration between researchers in mineralogy, computer vision, and machine learning to advance the field of mineral recognition.

CRediT authorship contribution statement

Sergey Nesteruk: Project administration, Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Visualization. **Julia Agafonova:** Software, Formal analysis, Investigation, Visualization. **Igor Pavlov:** Data curation, Software, Formal analysis, Investigation. **Maxim Gerasimov:** Software, Formal analysis, Investigation, Visualization. **Nikolay Latyshev:** Resources, Validation. **Denis Dimitrov:** Supervision, Validation. **Andrey Kuznetsov:** Supervision, Validation. **Artur Kadurin:** Conceptualization, Writing – review & editing. **Pavel Plechov:** Supervision, Validation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to the code. The dataset will be shared in the same repository.

Code availability section

Name of the code: Zero-Shot Raw Mineral Visual Recognition and Description.

Contact: SVNesteruk@sberbank.ru

Program language: Python3

The source codes are available for download at the link: https://github.com/ai-forever/mineral-recognition

The dataset is available at the link: https://disk.yandex.ru/d/KapicF_MEysifg

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Azarafza, M., Ghazifard, A., Akgün, H., Asghari-Kaljahi, E., 2019. Development of a 2D and 3D computational algorithm for discontinuity structural geometry identification by artificial intelligence based on image processing techniques. Bull. Eng. Geol. Environ. 78, 3371–3383.

Azarafza, M., Nanehkaran, Y.A., Akgun, H., Mao, Y., 2021. Application of an image processing-based algorithm for river-side granular sediment gradation distribution analysis. Adv. Mater. Res. 10 (3), 229–244.

Baek, Y., Lee, B., Han, D., Yun, S., Lee, H., 2019. Character region awareness for text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9365–9374.

Baraboshkin, E.E., Ismailova, L.S., Orlov, D.M., Zhukovskaya, E.A., Kalmykov, G.A., Khotylev, O.V., Baraboshkin, E.Y., Koroteev, D.A., 2020. Deep convolutions for in-depth automated rock typing. Comput. Geosci. 135, 104330. http://dx.doi.org/10.1016/j.cageo.2019.104330, URL: https://www.sciencedirect.com/science/article/pii/S0098300419304686.

Baykan, N.A., Yılmaz, N., 2011. A mineral classification system with multiple artificial neural network using k-fold cross validation. Math. Comput. Appl. 16 (1), 22–30.

Brempong, K.A., 2019. Minerals identification dataset. URL: https://www.kaggle.com/datasets/asiedubrempong/minerals-identification-dataset.

Bukharev, A., Budennyy, S., Lokhanova, O., Belozerov, B., Zhukovskaya, E., 2018a. The task of instance segmentation of mineral grains in digital images of rock samples (thin sections). In: 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). IEEE, pp. 18–23.

Bukharev, A., Budennyy, S., Pachezhertsev, A., Belozerov, B., Zhuk, E., 2018b. Automatic analysis of petrographic thin section images of sandstone. In: ECMOR XVI-16th European Conference on the Mathematics of Oil Recovery, Vol. 2018. EAGE Publications BV, pp. 1–10.

Carey, C., Boucher, T., Mahadevan, S., Bartholomew, P., Dyar, M., 2015. Machine learning tools formineral recognition and classification from Raman spectroscopy. J. Raman Spectrosc. 46 (10), 894–903.

Chanou, A., Osinski, G., Grieve, R., 2014. A methodology for the semi-automatic digital image analysis of fragmental impactites. Meteorit. Planet. Sci. 49 (4), 621–635.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 839–847.

Chemkaeva, D., 2019. Gemstones neural network - multiclass classification. https://github.com/LSIND/Gemstones-Convolutional-Neural-Network.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1. IEEE, pp. 886–893.

Diamantidis, N., Karlis, D., Giakoumakis, E.A., 2000. Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence 116 (1–2), 1–16.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duda, R.O., Hart, P.E., 1972. Use of the hough transformation to detect lines and curves in pictures. Commun. ACM 15 (1), 11–15.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc. Ser. B Stat. Methodol. 14 (1), 107–114, URL: http://www.jstor.org/stable/2984087.

Gottlieb, P., Wilkie, G., Sutherland, D., Ho-Tun, E., Suthers, S., Perera, K., Jenkins, B., Spencer, S., Butcher, A., Rayner, J., 2000. Using quantitative electron microscopy for process mineralogy applications. Jom 52 (4), 24–25.

Hao, H., Jiang, Z., Ge, S., Wang, C., Gu, Q., 2022. Siamese adversarial network for image classification of heavy mineral grains. Comput. Geosci. 159, 105016. http://dx.doi.org/10.1016/j.cageo.2021.105016, URL: https://www.sciencedirect.com/science/article/pii/S0098300421002983.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hossain, S., Uddin, J., Nahin, R.A., 2021. Rock classification dataset. http://dx.doi.org/10.34740/KAGGLE/DS/1293628, URL: https://www.kaggle.com/ds/1293628.

Illarionova, S., Shadrin, D., Ignatiev, V., Shayakhmetov, S., Trekin, A., Oseledets, I., 2022a. Augmentation-based methodology for enhancement of trees map detalization on a large scale. Remote Sens. 14 (9), 2281.

Illarionova, S., Shadrin, D., Shukhratov, I., Evteeva, K., Popandopulo, G., Sotiriadi, N., Oseledets, I., Burnaev, E., 2023. Benchmark for building segmentation on up-scaled sentinel-2 imagery. Remote Sens. 15 (9), 2347.

Illarionova, S., Shadrin, D., Tregubova, P., Ignatiev, V., Efimov, A., Oseledets, I., Burnaev, E., 2022b. A survey of computer vision techniques for forest characterization and carbon monitoring tasks. Remote Sens. 14 (22), 5861.

Illingworth, J., Kittler, J., 1988. A survey of the hough transform. Comput. Vis. Graph. Image Process. 44 (1), 87–116.

Ivchenko, A.V., Baraboshkin, E.E., Ismailova, L.S., Orlov, D.M., Koroteev, D.A., Baraboshkin, E.Y., 2018. Core photo lithological interpretation based on computer analyses. In: Proceedings of the IEEE Northwest Russia Conference on Mathematical Methods in Engineering and Technology, Russia, Saint-Petersburg. pp. 10–14.

Izadi, H., Sadri, J., Bayati, M., 2017. An intelligent system for mineral identification in thin sections based on a cascade approach. Comput. Geosci. 99, 37–49.

Jia, L., Yang, M., Meng, F., He, M., Liu, H., 2021. Mineral photos recognition based on feature fusion and online hard sample mining. Minerals 11 (12), 1354.

Jin, C., Wang, K., Han, T., Lu, Y., Liu, A., Liu, D., 2022. Segmentation of ore and waste rocks in borehole images using the multi-module densely connected U-net. Comput. Geosci. 159, 105018. http://dx.doi.org/10.1016/j.cageo.2021.105018, URL: https://www.sciencedirect.com/science/article/pii/S0098300421003009.

Kim, J.J., Ling, F.T., Plattenberger, D.A., Clarens, A.F., Lanzirotti, A., Newville, M., Peters, C.A., 2021. SMART mineral mapping: Synchrotron-based machine learning approach for 2D characterization with coupled micro XRF-XRD. Comput. Geosci. 156, 104898. http://dx.doi.org/10.1016/j.cageo.2021.104898, URL: https://www.sciencedirect.com/science/article/pii/S0098300421001904.

Kuen, J., Perazzi, F., Lin, Z., Zhang, J., Tan, Y.-P., 2019. Scaling object detection by transferring classification weights. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV.

Lakhani, H., 2020. Natural diamonds. URL: https://www.kaggle.com/datasets/harshitlakhani/natural-diamonds-prices-images.

Lemikhova, L., Nesteruk, S., Somov, A., 2022. Transfer learning for few-shot plants recognition: Antarctic station greenhouse use-case. In: 2022 IEEE 31st International Symposium on Industrial Electronics. ISIE, pp. 715–720. http://dx.doi.org/10.1109/ISIE51582.2022.9831723.

Li, Y., McCausland, P.J., Flemming, R.L., 2020. Best fit for complex peaks (BFCP) in matlab® for quantitative analysis of in situ 2D X-Ray diffraction data and Raman spectra. Comput. Geosci. 144, 104572. http://dx.doi.org/10.1016/j.cageo.2020.104572, URL: https://www.sciencedirect.com/science/article/pii/S0098300420305604.

Liu, C., Li, M., Zhang, Y., Han, S., Zhu, Y., 2019. An enhanced rock mineral recognition method integrating a deep learning model and clustering algorithm. Minerals 9 (9), 516.

Liu, Y., Zhang, Z., Liu, X., Wang, L., Xia, X., 2021. Deep learning-based image classification for online multi-coal and multi-class sorting. Comput. Geosci. 157, 104922. http://dx.doi.org/10.1016/j.cageo.2021.104922, URL: https://www.sciencedirect.com/science/article/pii/S0098300421002120.

Maitre, J., Bouchard, K., Bédard, L.P., 2019. Mineral grains recognition using computer vision and machine learning. Comput. Geosci. 130, 84–93. http://dx.doi.org/10.1016/j.cageo.2019.05.009, URL: https://www.sciencedirect.com/science/article/pii/S0098300419301037.

Maragos, P., 1987. Tutorial on advances in morphological image processing and analysis. Opt. Eng. 26 (7), 623–632.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al., 2022. Simple open-vocabulary object detection with vision transformers. arXiv preprint arXiv:2205.06230.

Mukhamadiev, S., Nesteruk, S., Illarionova, S., Somov, A., 2023. Enabling multi-part plant segmentation with instance-level augmentation using weak annotations. Information 14 (7), 380.

Nesteruk, S., Illarionova, S., Akhtyamov, T., Shadrin, D., Somov, A., Pukalchik, M., Oseledets, I., 2022. XtremeAugment: Getting more from your data through combination of image collection and image augmentation. IEEE Access 10, 24010–24028. http://dx.doi.org/10.1109/ACCESS.2022.3154709.

Nesteruk, S., Shadrin, D., Kovalenko, V., Rodríguez-Sánchez, A., Somov, A., 2020. Plant growth prediction through intelligent embedded sensing. In: 2020 IEEE 29th International Symposium on Industrial Electronics. ISIE, pp. 411–416. http://dx.doi.org/10.1109/ISIE45063.2020.9152399.

Nesteruk, S., Shadrin, D., Pukalchik, M., Somov, A., Zeidler, C., Zabel, P., Schubert, D., 2021. Image compression and plants classification using machine learning in controlled-environment agriculture: Antarctic station use case. IEEE Sens. J. 21 (16), 17564–17572. http://dx.doi.org/10.1109/JSEN.2021.3050084.

Nesteruk, S., Zherebtsov, I., Illarionova, S., Shadrin, D., Somov, A., Bezzateev, S.V., Yelina, T., Denisenko, V., Oseledets, I., 2023. CISA: Context substitution for image semantics augmentation. Mathematics 11 (8), 1818.

Neubeck, A., Van Gool, L., 2006. Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06), Vol. 3. IEEE, pp. 850–855.

NIMRF, 2003. The national infrastructure of mineral, rock and fossil for science and technology. URL: http://www.nimrf.net.cn/en/english.

Niu, X.-m., Jiao, Y.-h., 2008. An overview of perceptual hashing. Acta Electon. Sin. 36 (7), 1405.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9 (1), 62–66. http://dx.doi.org/10.1109/TSMC.1979.4310076.

Patel, A.K., Chatterjee, S., Gorai, A.K., 2017. Development of machine vision-based ore classification model using support vector machine (SVM) algorithm. Arab. J. Geosci. 10 (5), 1–16.

Peng, W., Bai, L., Shang, S., Tang, X., Zhang, Z., 2019. Common mineral intelligent recognition based on improved InceptionV3. Geol. Bull. China 38 (12), 2059–2066.

Plechov, P.Y., Trousov, S.V., Bychkov, K.A., Konovalova, K.A., 2019. Multilayered mineralogical information in spectroscopy of minerals. In: XIX International Meeting on Crystal Chemistry, X-Ray Diffraction and Spectroscopy of Minerals. pp. 43–43.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Ralph, J., 1993. Mindat.org - mines, minerals and more. URL: https://www.mindat.org/.

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E., 2022. A consistent and efficient evaluation strategy for attribution methods. In: International Conference on Machine Learning. PMLR, pp. 18770–18795.

Rouse, D.M., Hemami, S.S., 2008. Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In: Human Vision and Electronic Imaging XIII, Vol. 6806. SPIE, pp. 410–423.

Shu, L., McIsaac, K., Osinski, G.R., Francis, R., 2017. Unsupervised feature learning for autonomous rock image classification. Comput. Geosci. 106, 10–17. http://dx.doi.org/10.1016/j.cageo.2017.05.010, URL: https://www.sciencedirect.com/science/article/pii/S0098300417305587.

Singh, N., Singh, T., Tiwary, A., Sarkar, K.M., 2010. Textural identification of basaltic rock mass using image processing and neural network. Comput. Geosci. 14 (2), 301–310.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

Smith, L.N., Topin, N., 2019. Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006. SPIE, pp. 369–386.

Srinivas, S., Fleuret, F., 2019. Full-gradient representation for neural network visualization. Adv. Neural Inf. Process. Syst. 32.

Suzuki, S., et al., 1985. Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. 30 (1), 32–46.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.

Wenk, H.-R., Bulakh, A., 2016. Minerals: Their Constitution and Origin. Cambridge University Press.

Xu, Z., Ma, W., Lin, P., Shi, H., Pan, D., Liu, T., 2021. Deep learning of rock images for intelligent lithology identification. Comput. Geosci. 154, 104799. http://dx.doi.org/10.1016/j.cageo.2021.104799, URL: https://www.sciencedirect.com/science/article/pii/S009830042100100X.

Yousefi, B., Castanedo, C.I., Maldague, X.P., Beaudoin, G., 2020. Assessing the reliability of an automated system for mineral identification using LWIR hyperspectral infrared imagery. Miner. Eng. 155, 106409.

Zeng, X., Xiao, Y., Ji, X., Wang, G., 2021. Mineral identification based on deep learning that combines image and mohs hardness. Minerals 11 (5), 506.

Zhang, Y., Li, M., Han, S., Ren, Q., Shi, J., 2019. Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms. Sensors 19 (18), 3914.