minerals

MDPI

Communication

Mineral Photos Recognition Based on Feature Fusion and Online Hard Sample Mining

Liqin Jia [1], Mei Yang [2], Fang Meng [3], Mingyue He [3] and Hongmin Liu [1,4,*]

[1] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China; lqjia@hpu.edu.cn
[2] Sciences Institute, China University of Geosciences, Beijing 100083, China; yangmei@cugb.edu.cn
[3] Gemological Institute, China University of Geosciences, Beijing 100083, China; mengfang@cugb.edu.cn (F.M.); hemy@cugb.edu.cn (M.H.)
[4] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: hmliu_82@163.com

Abstract: Mineral recognition is of importance in geological research. Traditional mineral recognition methods need professional knowledge or special equipment, are susceptible to human experience, and are inconvenient to carry in some conditions such as in the wild. The development of computer vision provides a possibility for convenient, fast, and intelligent mineral recognition. Recently, several mineral recognition methods based on images using a neural network have been proposed for this aim. However, these methods do not exploit features extracted from the backbone network or available information of the samples in the mineral dataset sufficiently, resulting in low recognition accuracy. In this paper, a method based on feature fusion and online hard sample mining is proposed to improve recognition accuracy by using only mineral photo images. This method first fuses multi-resolution features extracted from ResNet-50 to obtain comprehensive information of mineral photos, and then proposes the weighted top-k loss to emphasize the learning of hard samples. Based on a dataset consisting of 14,986 images of 22 common minerals, the proposed method with 10-fold cross-validation achieves a Top1 accuracy of 88.01% on the validation image set, surpassing those of Inception-v3 and EfficientNet-B0 by a margin of 1.88% and 1.29%, respectively, which demonstrates the good prospect of the proposed method for convenient and reliable mineral recognition using mineral photos only.

Keywords: mineral recognition; feature fusion; online hard sample mining; deep learning; image recognition

check for updates

Citation: Jia, L.; Yang, M.; Meng, F.; He, M.; Liu, H. Mineral Photos Recognition Based on Feature Fusion and Online Hard Sample Mining. *Minerals* **2021**, *11*, 1354. https://doi.org/10.3390/min11121354

Academic Editor: Amin Beiranvand Pour

Received: 19 October 2021
Accepted: 25 November 2021
Published: 30 November 2021

1. Introduction

Mineral recognition is a basic yet important aspect in geological surveys. It can not only enrich the map of mineral resources on the earth, but also be used to estimate the hidden mining volume and potential economics of minerals, providing geological information for subsequent mineral exploration. Traditionally, mineral recognition is professional work, which distinguishes minerals according to shape, optical properties, and mechanical properties, requiring rich knowledge and experience or special equipment. However, this process is susceptible to human experience, inefficient, and costly. Recently, with the rapid development of artificial intelligence, a considerable number of methods have been proposed to solve geological problems in a smarter and more convenient way by using an artificial neural network (ANN) [1–8].

According to the difference in input images, the current research for mineral recognition methods with ANN can be organized into three groups: Microscopic Image-Based methods, Raman Spectra Image-Based methods, and Photo Image-Based methods.

(1) Microscopic Images-Based Methods: Baykan and Yılmaz [9] employed the multilayer perceptron neural network (MLPNN) with one hidden layer for mineral classification, which is based on the RGB data of plane-polarized and cross-polarized microscope images and achieved 94.07% average accuracy for five minerals. An idea to use cluster algorithms and morphological analysis to determinate colors and shapes for computing the composition of rocks from micrographs was proposed without providing the number of mineral types and accuracy [10]. Izadi et al. [11] presented a two-level cascade neural network classification approach, which first recognized the minerals based on color parameters and then identified those minerals rejected from the first level based on texture features of plane and cross-polarized light. Overall accuracy of 93.81% for the recognition of 23 test minerals was obtained. Maitre et al. [12] proposed an approach to automate mineral grain recognition using an optical microscope image, which relied on data processing such as superpixel generation, feature extraction and data cleaning, and machine-learning algorithms that classify vectors of mineral features, identifying eight kinds of mineral particles with an accuracy of approximately 90%. A complex ensemble model was proposed in [13], which used Inception-v3 [14] to extract features of the microscopic images, selected logistic regression (LR), support vector machine (SVM), and multilayer perceptron (MLP) as the basic models, and chose the LR model as the final prediction meta classifier. The composed model recognized four minerals with an accuracy of 90.9%. The work of [15] used five machine-learning algorithms to classify the scanning electron microscope images of 12 minerals, and reached accuracies of 86–92%. Among all the above methods, the acquisition of microscopic images is equipment dependent and inconvenient. Additionally, the types of recognized minerals are few due to the limited samples [9,11–13,15].

(2) Raman Spectra Image-Based methods: Raman spectroscopy has been widely used as a mature auxiliary tool for mineral recognition. An artificial neural network was trained using Raman spectra of minerals to distinguish six minerals in igneous rocks, achieving 83% accuracy [16]. The work of [17] proposed full-spectrum matching algorithms realizing 96.5% average accuracy of six minerals without model training. Due to the lack of a large-scale Raman spectrum image set, it is difficult for learning-based methods to train the network, and it is also tough for testing to obtain the Raman spectrum of the sample in the wild. So, mineral recognition based on Raman spectra has difficulty in extensive applications.

(3) Photo Image-Based methods: Compared to the above two groups, mineral photo images can be obtained conveniently due to the popularity of digital cameras and smartphones. Therefore, mineral recognition based on mineral photo images has attracted increasing attention. Recently, Zeng et al. [18] employed mineral photos combined with Mohs hardness to achieve a Top1 accuracy rate of 90.6% for 36 common minerals using a deep neural network, which input the Mohs hardness of the corresponding mineral into the model manually to assist the image recognition. Without the Mohs information, the Top1 accuracy of model dropped drastically to 78.3%. Although useful, the use of Mohs hardness reduces the adaptability and universality of the algorithm. Liu et al. [19] extracted the texture features of images using the Inception-v3 model [14] and established a color model by the K-means algorithm, and then combined the two models to obtain a comprehensive recognition model, which achieves a Top1 accuracy of 74.2% of 12 minerals. Peng et al. [20] also used the Inception-v3 model but combined the softmax loss with the center loss to identify 19 minerals. This method obtained a Top1 accuracy of 86%, 5 percentage points higher than that of the softmax loss alone. Although the center loss improves the recognition accuracy by reducing the intra-class distance, it slows the convergence of the model greatly and the model training becomes more difficult.

To solve the above issues for mineral photo image recognition, such as the use of additional geological information [18], incomplete feature extraction [19], and loss function improvement [20], a deep learning model based on feature fusion and online hard sample mining using mineral photos only is proposed in this paper. Here, ResNet-50 is used to extract features of the mineral images, and then, the low-level features are merged with the high-level features to improve the model performance due to the fact that the low-level features, such as color and texture, are important for mineral recognition. Meanwhile, a weighted top-*k* loss is also proposed to exploit the available information of hard and easy samples, improving the recognition accuracy further.

The remainder of this paper is constructed as follows. Section 2 introduces a detailed presentation of the proposed method. Section 3 provides the mineral dataset and experimental results, followed by the experimental analysis in Section 4. Finally, we conclude in Section 5.

2. Method

In this section, a mineral photo-recognition model based on the deep residual network ResNet-50, combining multi-resolution feature fusion and online hard sample mining weighted top-*k* loss, is designed.

2.1. Backbone Network ResNet-50

In the past years, many excellent backbone networks have been proposed, for example LeNet [21], ALexNet [22], VGGNet [23], Inception [14,24], ResNet [25], EfficientNet [26], and so on. By introducing a residual structure, ResNet can improve the network performance by increasing the network layers while avoiding gradient explosion/disappearance. It has become one of the most widely used convolutional neural network (CNN) backbones for feature extraction. The structure of ResNet-50 is shown in Table 1. The input image is resized to 224 × 224, then it goes through a convolution layer (conv1) and a max pooling process with a stride size of 2. Features are subsequently extracted through four residual layers (Layer1, Layer2, Layer3, and Layer4). Next, a global average pooling (GAP) operation is conducted to obtain a 1 × 1 × 2048 feature, which is then flattened and input into a full convolutional (FC) layer and the probabilities of mineral types are the output.

Table 1. Network structure of ResNet-50.

Layer	Process	Output Size
conv1	$7 \times 7,\ 64,\ stride = 2$	$112 \times 112 \times 64$
-	max pooling, 3×3, $stride = 2$	$56 \times 56 \times 64$
Layer1	$\begin{bmatrix} 1 \times 1,\ 64 \\ 3 \times 3,\ 64 \\ 1 \times 1,\ 256 \end{bmatrix} \times 3$	$56 \times 56 \times 256$
Layer2	$\begin{bmatrix} 1 \times 1,\ 128 \\ 3 \times 3,\ 128 \\ 1 \times 1,\ 512 \end{bmatrix} \times 4$	$28 \times 28 \times 512$
Layer3	$\begin{bmatrix} 1 \times 1,\ 256 \\ 3 \times 3,\ 256 \\ 1 \times 1,\ 1024 \end{bmatrix} \times 6$	$14 \times 14 \times 1024$
Layer4	$\begin{bmatrix} 1 \times 1,\ 512 \\ 3 \times 3,\ 512 \\ 1 \times 1,\ 2048 \end{bmatrix} \times 3$	$7 \times 7 \times 2048$
-	global average pooling	$1 \times 1 \times 2048$
FC	FC + softmax	Num_classes

2.2. Feature Fusion

The convolution operation and design of convolution neural networks mean the extracted features in the network are of a hierarchical nature. That is to say, the low layers respond to basic features, such as the color and edges. With the increase in the number of layers, the complexity of features increases, and more class-specific features are extracted. Generally, high-level features are used for the classification of different kinds of objects, such as objects given in the ImageNet dataset. However, due to the differences in chemical composition, crystallization, and chemical properties, minerals present a variety of colors, crystal forms, hardness, and luster, which are shown intuitively in different colors, shapes, transparencies, and textures of mineral photo images. So, for mineral recognition, low-level features such as colors, shapes, and textures are still important for mineral recognition, which are ignored by existing methods.

In this paper, a mineral recognition method fusing low-level features and high-level features is proposed to improve the performance of high-level features and produce an increase in recognition accuracy. In detail, mineral photo images are input into the ResNet-50, which is pretrained by the ImageNet dataset to obtain the features of four layers.Layer3 and Layer4 output the high-level features, and the features of Layer4 are usually used as the most discriminative features to classify the objects since they include the largest receptive field and the richest semantic information. The features from Layer1 and Layer2 are often considered low-level features. Since features from different layers have different feature sizes and numbers of channels, before feature fusion, GAP is performed to resize the features to a uniform height and width (1 × 1). Then, the low-level features, denoted as F_L, and the high-level features, denoted as F_H, are merged via concatenation to obtain fused features $F_{fused} = [F_H, F_L]$. The fused features thus contain not only rich high-level semantic information, but also much low-level details information. Finally, the model output is obtained by passing the fusion features through the full connection (FC) layer. Figure 1 displays the example model when the features from Layer2 and Layer4 are fused.

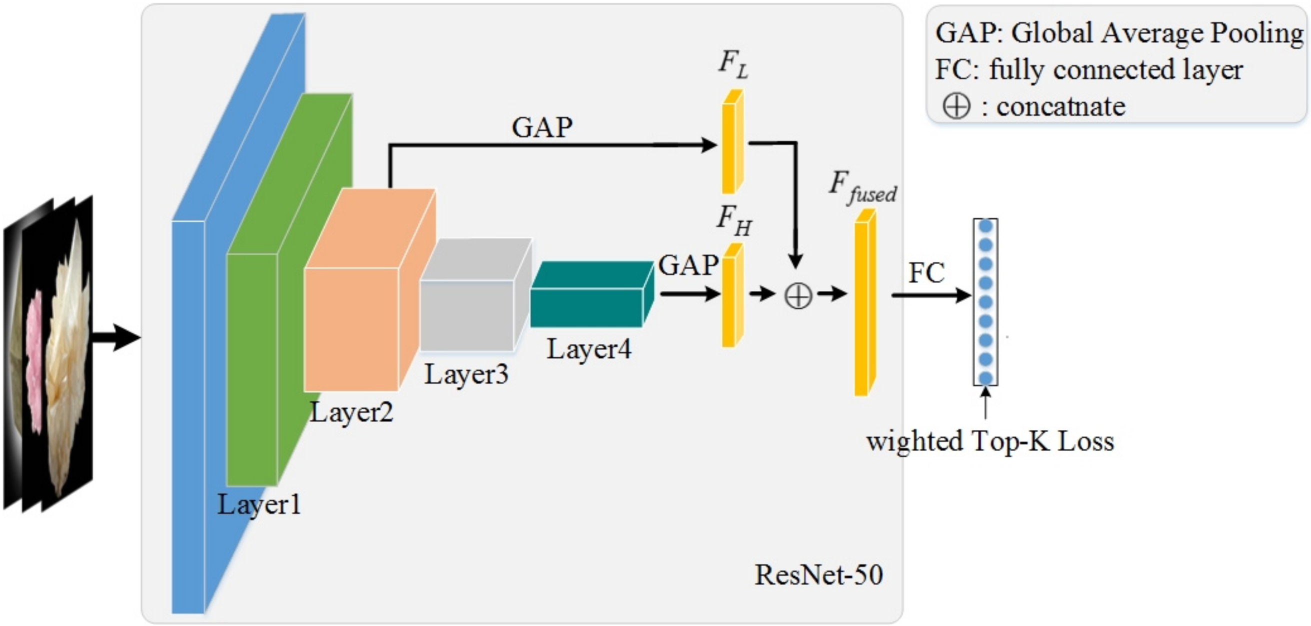


Figure 1. Mineral photos recognition based on feature fusion and weighted Top-*k* flow chart.

2.3. Loss Function

The loss function assigns the goal of network learning. Usually, the proportion of simple samples, which display the clear features of the minerals, is much larger than that of hard ones, the features of which are shown in a confusing way due to inappropriate imaging or unapparent characteristics of the mineral itself. So, in the training process, the network can easily learn the obvious features from the simple samples, while further mining of the features from hard samples is ignored because the small number of hard samples brings low weights in the total loss. So, when network learning reaches a certain level, the existing loss functions cannot impose the network to learn the implicit information contained in the hard samples further, restricting the improvement of the network performance.

For mineral recognition, due to some subtle differences, certain minerals visually display large intra-class differences and small inter-class differences resulting in misrecognition. So, paying more attention to these hard samples can help the model achieve higher accuracy. Top-*k* loss [27] is proposed by Zhang to solve online hard sample mining (OHSM) for face detection. The core of the OHEM algorithm is to select hard samples with large loss values as training samples to learn the network parameters. Here, we try to consider top-k loss, denoted as $Loss_{\text{top-}k}$, in our study for mineral photo recognition. L_i denotes the softmax loss value of the ith sample in a batch, which can be described as (1), where a_i represents the ith sample's output from the model, which is a $num\_calsses \times 1$ vector. The softmax loss of a batch is the average of all losses in the batch, shown as (2), where N is the batch size. Denote L'_i as L_i in descending order like (3). The $Loss_{\text{top-}k}$ is defined as the average of the top $k \times N$ loss values, as shown in Equation (4), where k is a percentage.

$$L_i = -\log\left(\frac{e^{a_i}}{\sum_{j=1}^{num\_classes} e^{a_j}}\right) \tag{1}$$

$$Loss_{\text{softmax}} = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{2}$$

$$sort(L_i) = \{L'_1, \cdots, L'_i, L'_{i+1}, \cdots, L'_N\}, L'_i \geq L'_{i+1}, i \in (1, \cdots, N-1) \tag{3}$$

$$Loss_{\text{top-}k} = \frac{1}{k \times N}\sum_{i=1}^{k \times N} L'_i \tag{4}$$

In the top-k loss, only the loss values of the top $k \times N$ samples, considered as hard samples, are selected for learning parameters in each training batch. Then the gradients are computed from these hard samples in backward propagation, which can ensure the network pays more attention to hard samples and effectively excavates more implicit information, improving the performance of the network. Although top-k loss shows good effectiveness in face detection, a task paying more attention to semantic information, mineral recognition is a task that pays close attention to low-level features, clearly presented in the simple samples. In order to effectively utilize all samples and balance the roles of hard and simple samples, a weighted top-k loss is proposed in this paper, which ensures the network pays attention to the hard samples while taking simple samples into account, improving the performance of the network. The weighted top-k loss function is shown in Equation (5), where the top $k \times N$ loss values are hard samples and the latter $N - k \times N$ are easy samples. In (5), $\alpha(\alpha \in [0,1])$ is the weight coefficient. When $\alpha = 0$, the weighted top-k loss is the top-k loss, and when $\alpha = 1$, the weighted top-k loss is the softmax loss.

$$Loss_{\text{weighted top-}k} = \frac{1}{k \times N}\sum_{i=1}^{k \times N} L'_i + \alpha \times \frac{1}{(1-k) \times N}\sum_{i=k \times N+1}^{N} L'_i \tag{5}$$

3. Experiments

3.1. Data

In total, 14,986 images of 22 common minerals collected from the National Mineral Rock and Fossil Specimens Resource Center of China [28] and Mindat.org [29] were used in our experiments. The names and numbers of the minerals are shown in Table 2. The biotite and phlogopite is a mixture of biotite and phlogopite. To recognize minerals correctly, most of the minerals are single minerals, and some minerals contain surrounding rocks or symbiotic minerals. Some examples of images are shown in Figure 2, where (c), (e)–(i) are single minerals, (a) and (b) contain surrounding rocks, (d) contains symbiotic minerals, and (j) includes a base. It can be seen from the images that the typical mineral photo has a large object and clear features such as color, shape, and texture.

Table 2. Names and quantities of minerals in the dataset.

	Name	Number		Name	Number
1	pyrite	959	12	muscovite	227
2	realgar	319	13	biotite and phlogopite	150
3	orpiment	217	14	feldspar	701
4	stibnite	669	15	calcite	1045
5	galena	860	16	barite	1210
6	quartz	1315	17	turquoise	346
7	spinel	622	18	tourmaline	874
8	corundum	726	19	malachite	918
9	garnet	279	20	azurite	783
10	olivine	215	21	rhodochrosite	859
11	beryl	791	22	fluorite	901
Total			14,986		

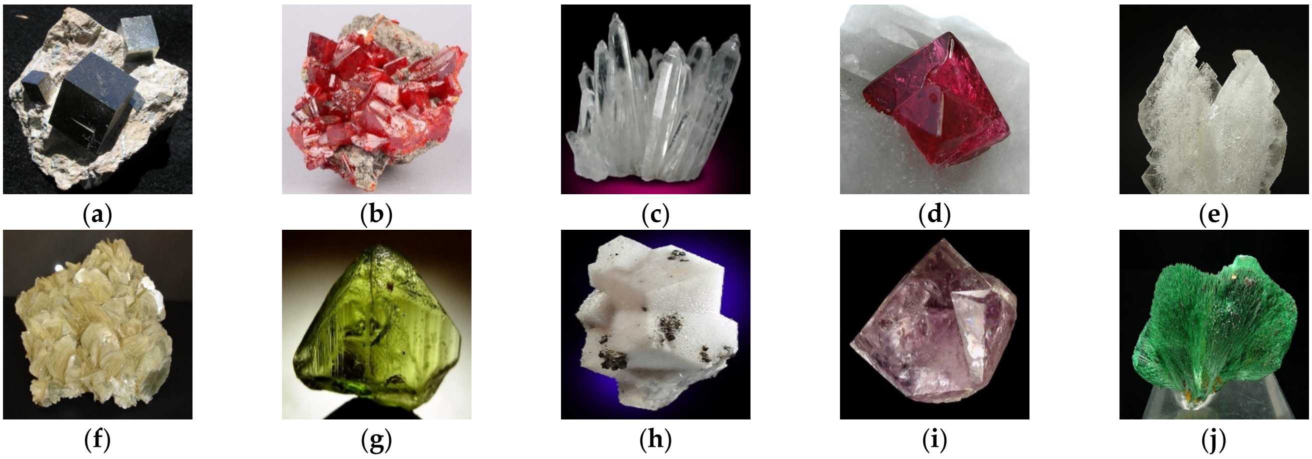


Figure 2. Mineral image samples. (**a**) Pyrite, (**b**) Realgar, (**c**) Quartz, (**d**) Spinel, (**e**) Barite, (**f**) Muscovite, (**g**) Olivine, (**h**) Calcite, (**i**) Fluorite, (**j**) Malachite. (**c**,**e**–**i**) are single minerals, (**a**,**b**) contain surrounding rocks, (**d**) contains symbiotic minerals, and (**j**) includes a base.

3.2. Evaluation

The Top1 accuracy and mean average precision (mAP) are employed to evaluate the performance of the proposed method in the experiments. The ratio of the number of times the maximum value of the output probability vector matches the correct label to the number of the validation set is called the Top1 accuracy of the dataset, and it is the same for a certain mineral category. Top1 accuracy indicates the recognition accuracy of the algorithm for the overall dataset and a certain mineral category. The mAP is the mean of all mineral categories' Top1 accuracies, which pays attention to the accuracy of the categories with small sample size, showing the performance of the algorithm in each category.

Cross-validation is an important method of evaluating models and parameters, and k-fold cross-validation is usually used. k-fold cross-validation means that all data are randomly divided into k groups, where $k - 1$ groups are used as the training set, and the one remaining group is used as the validation set. The accuracy results are obtained by averaging the results of k validation sets. The experimental results in this paper are the results of 10-fold cross-validation, that is, the mineral dataset is divided into the training set and validation set according to 9:1, and the result is the average of ten validation sets.

3.3. Experiments Results

3.3.1. Network Setting

In experiments, the 64-bit Ubuntu 18.04 operating system and 11G GeForce RTX 2080ti are used. We use version 1.4.0 of the torch deep-learning library and ResNet-50 as the backbone network. We calculate the mean and variance of the RGB channels of all photo images in the training and validation sets. Data augmentation is employed with random resizing and cropping, random rotation, and random horizontal flipping. Mineral photo images are preprocessed by data augmentation, cutting the center of the images to 224 × 224 and normalizing them with the computed mean and variance.

The initial learning rate is set as 1.0×10^{-3} and reduced to 0.5 times the current learning rate when the highest validation set accuracy remains unchanged for five epochs. The training batch size is set to 50, the optimizer is SGD, and the training epoch is 100. The fully connected layer in the network consists of one layer of the linear network.

3.3.2. Backbone Selection

Vgg16 employs the fixed convolution kernel 3 × 3 and increases the channels of features gradually, which has proved that the network depth is critical in image recognition. ResNet adds a residual structure on the network, such as Vgg, which ensures the training and performance of deeper networks are excellent. Inception broadens the network structure. EfficientNet considers the depth, width, and resolution of the network simultaneously. These networks play important roles in the development of image recognition and are considered classic backbones. The mineral recognition results based on these classic networks are listed in Table 3. As we can see, ResNet-50 is more suitable for our task. Both the Top1 (87.15%) and mAP (86.54%) of ResNet-50 surpass those of Vgg16 and Inception-v3 used in [19,20] by a large margin, and are also higher than that of Efficient-Net-B0, which is an excellent deep-learning network model proposed recently. So, the following work is conducted on the ResNet-50 network.

Table 3. Mineral recognition results based on classical deep-learning backbone networks (%).

Networks	Top1	mAP
Vgg16 [23]	83.17	82.23
Inception-v3 [14]	86.07	85.35
EfficientNet-B0 [26]	86.72	86.27
ResNet-50 [25]	87.15	86.54

3.3.3. Feature Fusion

For convenience, F_1, F_2, F_3, F_4 denotes features obtained by Layer1, Layer2, Layer3, and Layer4 passing through the GAP. As F_4 has the largest receptive field and the best semantic information of the objects in the mineral photos, it is naturally used to distinguish mineral categories as the most discriminative feature. We take the low-level features to merge with F_4 and the Top1 and mAP results are shown in Table 4. It can be observed that, when the feature from the lowest layer, F_1, is fused with that from the highest layer, F_4, such as the combinations of $F_1 + F_4$ and $F_1 + F_2 + F_3 + F_4$, the performance of the model decreases. This may be due to the redundant and more detailed feature of F_1 damaging the semantic feature structure of high-level. As a comparison, fusion with the features from the lower level, F_2, make obvious improvements. The combination $F_2 + F_4$ reaches the highest Top1 accuracy of 87.6%, which is 0.45% higher than the result of not using the fused feature (87.15%), and the mAP improved by a margin of 0.38%. Moreover, since F_3 is located in a higher layer, it includes fewer details compared to F_2 as well as incomplete semantics compared to F_4, so its limited role brings a limited performance improvement. Consequently, the following work is based on $F_2 + F_4$ fusion.

Table 4. Mineral recognition results using different features (%).

Features	Top1	mAP
F_4	87.15	86.54
$F_1 + F_4$	87.13	86.36
$F_2 + F_4$	87.6	86.92
$F_3 + F_4$	87.29	86.45
$F_2 + F_3 + F_4$	87.47	86.63
$F_1 + F_2 + F_3 + F_4$	87.05	86.36

To illustratee the Grad-CAM (Gradient-weighted Class Activation Mapping) [30] and confusion matrix in the following section conveniently, and with the accuracies of dataset-1 in the 10-fold cross-validation dataset being closest to the accuracies of the 10-fold cross-validation dataset, we take dataset-1 instead of the 10-fold cross-validation dataset in some experiments. The recognition accuracies of 22 minerals employing F_4 feature and $F_2 + F_4$ fused features of dataset-1 are listed in the left two columns of Table 5, both of which use the softmax loss as the goal of learning. Benefitting from the feature fusion, the accuracies of 9 minerals from 22 minerals are increased, and the accuracies of the other 9 minerals are retained. In addition, four minerals were sensibly declined. Muscovite achieves the largest increase (8.70%) for the fused features complementing the insufficient features extracted from the small sample size. Nevertheless, the accuracy of biotite and phlogopite is significantly reduced (6.66%) and it may be that the number of samples is too small to extract informative features with both low-level and high-level data, resulting in poor fused features.

Table 5. The Top1 accuracies of minerals obtained by three methods (%).

Name	F_4	F_2 + F_4	F_2 + F_4 + Weighted Top-k
Pyrite	89.58	87.50	94.79
Realgar	96.88	100	100
Orpiment	72.73	77.27	86.36
Stibnite	98.51	100	100
Galena	94.19	94.19	91.86
Quartz	80.3	81.06	84.85
Spinel	80.95	84.13	82.54
Corundum	84.93	91.78	91.78
Garnet	71.43	71.43	78.57
Olivine	90.91	90.91	90.91
Beryl	87.50	82.50	85.00
Muscovite	78.26	86.96	69.57
Biotite & Phlogopite	73.33	66.67	86.67
Feldspar	87.32	87.32	90.14
Calcite	75.24	76.19	76.19
Barite	83.47	83.47	85.95
Turquoise	97.14	97.14	97.14
Tourmaline	87.50	88.64	89.77
Malachite	98.91	98.91	98.91
Azurite	98.73	98.73	98.73
Rhodochrosite	93.02	91.86	94.19
Fluorite	80.22	80.22	79.12
Top1	87.13	87.59	88.98
mAP	86.41	87.13	88.77

Grad-CAM is an algorithm that provides a visual interpretation of the areas that CNN focuses on when making predictions. In back propagation of the CNN networks, the gradients are computed to obtain the weights, which capture the "importance" of a feature map for a target class. The Grad-CAM is now widely used to visualize network performance. We illustrate some activation maps of samples with the gradients of F_4 and fused features

$F_2 + F_4$ based on dataset-1 using the Grad-CAM algorithm in Figure 3. The images in rows represent the original mineral photos, the Grad-CAM results of F_4 (wrongly recognized), and the Grad-CAM results of fused feature $F_2 + F_4$ (correctly recognized), respectively. Images (a)–(f) are realgar, orpiment, spinel, corundum, muscovite, and muscovite, respectively. In the second row, by using F_4, (a)–(f) are misrecognized as rhodochrosite, realgar, tourmaline, tourmaline, spinel, and calcite, respectively. As we can see, by exploiting F_4 only, the model ignores many details, which leads to misrecognition. Fused features $F_2 + F_4$ allow the model to utilize more information from the images, especially details that accurately reflect the essential characteristics of minerals. For example, in Figure 3a, F_4 mainly focuses on the surrounding rocks, but $F_2 + F_4$ pays closer attention to the mineral itself. In addition, the corundum in Figure 3d has fewer features extracted from the model with F_4, which is similar to tourmaline. However, fused features ($F_2 + F_4$) extract more details from the color, shape, and texture, and obtain the correct decision.

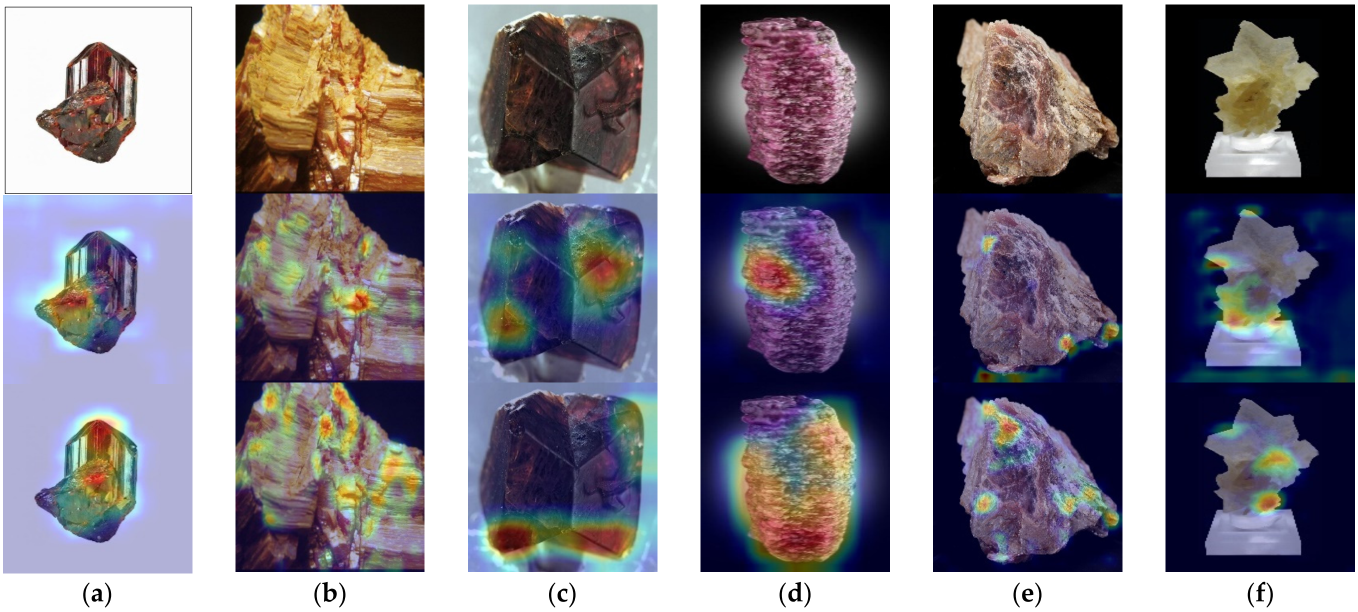


Figure 3. Grad-cam results using F_4 and fused features $F_2 + F_4$ based on ResNet-50. (**a**) Realgar; (**b**) Orpiment; (**c**) Spinel; (**d**) Corundum; (**e**) Muscovite; (**f**) Muscovite; the first row are original images, the second and third rows are Grad-cam results of F_4 and $F_2 + F_4$ recognition, respectively. In the second row, using F_4, (**a**–**f**) are misrecognized as rhodochrosite, realgar, tourmaline, tourmaline, spinel, and calcite, respectively.

3.3.4. Weighted Top-*k* Loss

Weighted Top-*k* Loss

We conducted experiments based on fused features and top-*k* loss firstly, and the results of different *k* are shown in Table 6. The number of loss values for backpropagation changes with the *k* values of the top-*k* loss, that is the smaller the *k* value, the fewer losses are included, where more attention is paid to samples with large loss values; the larger the *k* value, the more losses included, where the attention paid to difficult samples is reduced, resulting in important information being ignored. It can be observed from the cross-validation result that Top1 is 87.69% when $k = 0.5$, which exceeds the result of the softmax loss (Top1 = 87.60% when $k = 1$), indicating the top-*k* loss can obtain a higher accuracy rate when the appropriate ratio of hard samples is selected. However, mAP of the top-*k* loss is reduced by 0.36%, showing that although top-*k* loss mines the information of hard samples, these hard samples may come from the same category and do not help to improve mAP.

Table 6. Mineral recognition results based on fused features and top-*k* loss (%).

k	Top1	mAP
0.1	86.93	85.63
0.2	87.48	86.21
0.3	87.53	86.77
0.4	87.57	86.75
0.5	87.69	86.56
0.6	87.30	86.43
0.7	87.45	86.80
0.8	87.23	86.53
0.9	87.20	86.36
1.0	87.60	86.92

Further, we conducted experiments based on fused features and the weighted top-*k* loss defined by equation (3) when k is fixed to 0.5, and the comparison results of different α are shown in Table 7. As can be seen from Table 7, when $\alpha = 0.2$, the weighted top-*k* loss ensures the proposed model utilizes the information of both the hard and easy samples simultaneously, reaching the highest Top1 accuracy 88.01% and mAP 87.15%.

Table 7. Mineral recognition results based on fused features and weighted top-*k* loss when $k = 0.5$ (%).

α	Top1	mAP
0.1	87.41	86.49
0.2	88.01	87.15
0.3	87.53	86.76
0.4	87.74	86.91
0.5	87.32	86.36
0.6	87.17	86.46
0.7	87.45	86.28
0.8	87.53	86.84
0.9	87.34	86.61
1.0	87.60	86.92

Comparison with Other Loss Functions

To illustrate the superiority of the weighted top-*k* loss, we compared it with other loss functions used in mineral recognition. The experiments are all based on fused features, and the comparison results are listed in Table 8. It can be observed that the weighted top-*k* loss proposed in this paper achieves optimal Top1 and mAP accuracies, surpassing the results of the softmax loss and top-*k* loss by a considerable margin. As a comparison, training with the combination of the softmax loss and center loss used in [20] and testing in validation obtained a Top1 accuracy of 87.00% and mAP of 85.75%, which are much lower than the result of the weighted top-*k* loss, with a large gap of 1.01% in Top1 and 1.40% in mAP accuracy, indicating that the weighted top-*k* loss had an excellent performance in mining information and the center loss does not work in our model based on feature fusion.

Table 8. Mineral recognition results using different loss functions based on fused features.

Loss Function	Top1 (%)	mAP(%)	Parameter
softmax loss	87.60	86.92	-
top-*k* loss	87.69	86.56	$k = 0.5$
softmax loss + center Loss	87.00	85.75	same as [19]
weighted top-*k* loss(we proposed)	88.01	87.15	$k = 0.5, \alpha = 0.2$

The recognition results of 22 minerals by fused features $F_2 + F_4$ with the softmax loss and weighted top-*k* loss are shown in the last two columns of Table 5. As we can observe, paying more attention to hard samples and taking easy ones into account simultaneously causes both Top1 and mAP to obviously improve. The accuracies of 10 minerals from 22 minerals are greatly increased and those of 8 minerals are maintained. Surprisingly, the Top1 accuracy of biotite and phlogopite is improved by the largest margin, at 20%, the reason for which is the large loss values resulting from the poorly fused features meaning the biotite and phlogopite samples are more emphasized as hard samples. It is worth noting that the accuracies of 4 minerals from 22 minerals decreased. The muscovite had the largest drop, and there are two reasons for this. The first is the number of muscovite is less than those of most other minerals, producing a smaller effect on the loss values in each batch. Additionally, the loss values of muscovite weighted by α contribute little to the total loss because the loss values are not large enough to be considered as hard samples, and this is the same reason for the Top1 decrease of galena, spinel, and fluorite. Both of these mean the model cannot be trained enough for muscovite, and a large sample size with samples possessing discriminative features can ensure the muscovite recognition accuracy improves.

3.3.5. Comparison with the Previous Methods

Zeng et al. [18] employed mineral Mohs hardness for recognition, which is inconvenient for users without the corresponding knowledge. Liu et al. [19] employed manual extraction features and the recognition categories were few (12 categories) while the Top1 accuracy was also low (74.2%). Peng et al. [20] employed Inception-v3 and softmax loss combined with center loss, obtaining 86% Top1 accuracy on 19 minerals. We compare our method with [20] in this section. Table 9 shows the results using the same settings as that in [20] and those of our method. We can observe that both Top1 and mAP of our method are higher than those of [20] by a large margin. Our method uses a backbone network more suitable for classification, then makes full use of the features extracted from the network, and modifies the loss function by considering the weight of hard and easy samples, so that it can outperform other mineral photo-recognition methods and achieve a promising mineral recognition performance.

Table 9. Recognition result of our method and the method proposed by [20] (%).

Method	Top1	mAP
Inception-v3 + softmax loss + center loss [20]	85.99	84.52
ResNet-50 + fused features + weighted top-*k* loss (our method)	88.01	87.15

4. Experimental Analysis

4.1. Confusion Matrix Analysis

Table 10 shows the confusion matrix of the validation set of dataset-1 based on the method we propose. Here, the abbreviations of the minerals are provided except for turquoise and tourmaline, whose abbreviations are the same. In Table 10, the values in rows represent the probability of one mineral being judged as other minerals and the diagonal values denote the Top1 accuracies of the minerals. It can be seen that almost all minerals with obvious characteristics can be accurately recognized, such as minerals with specific colors (Top1 accuracies reach 100%). For example, realgar is orange-red, turquoise is green-blue, malachite looks green like malachite, and azurite presents as dark blue. The recognition accuracies of minerals with special shapes are also satisfactory. For example, pyrite (95%) is cubic and octahedron, and antimonite (100%) usually shows emission. Minerals with obvious texture have also been accurately identified, for example rhodochrosite (94%) has a special ring.

Table 10. The confusion matrix of 22 minerals in validation set of the dataset-1 (%).

Names	Py	Rel	Opm	Stb	Gn	Qtz	Spl	Crn	Grt	Ol	Brl	Ms	Bt&Phl	Fsp	Ct	Brt	Turquoise	Tourmaline	Mi	Azr	Rds	Fl
Py	94.79	0	0	0	1.04	0	0	0	1.04	0	0	0	0	0	0	1.04	0	0	0	0	0	2.08
Rel	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Opm	0	9.09	86.36	0	0	0	0	0	0	4.55	0	0	0	0	0	0	0	0	0	0	0	0
Stb	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gn	3.49	0	0	0	91.86	0	2.33	0	0	0	0	1.16	0	0	0	0	0	0	0	1.16	0	0
Qtz	0	0	0	0	0.76	84.85	0	1.52	0.76	0	2.27	0.76	0	1.52	3.79	1.52	0	0	0	0	0	2.27
Spl	1.59	0	0	0	0	1.59	82.54	4.76	3.17	0	0	0	1.59	1.59	0	0	0	0	0	0	3.17	0
Crn	0	0	0	0	1.37	0	0	91.78	0	0	1.37	0	1.37	0	1.37	1.37	0	0	0	0	1.37	0
Grt	0	0	0	0	3.57	3.57	3.57	3.57	78.57	3.57	0	0	0	0	3.57	0	0	0	0	0	0	0
Ol	4.55	0	0	0	0	0	0	0	0	90.91	0	0	0	0	0	0	0	0	4.55	0	0	0
Brl	0	0	0	1.25	0	0	0	1.25	0	2.5	85	0	0	1.25	1.25	3.75	0	2.5	1.25	0	0	0
Ms	0	0	0	0	0	4.35	4.35	0	0	0	0	69.57	8.7	0	4.35	8.7	0	0	0	0	0	0
Bt&Phl	6.67	0	0	0	0	0	0	0	6.67	0	0	0	86.67	0	0	0	0	0	0	0	0	0
Fsp	1.41	0	0	0	1.41	1.41	4.23	0	0	0	0	0	0	90.14	1.41	0	0	0	0	0	0	0
Ct	0.95	0	0	0	0	8.57	0	0	0	0	0	0	0	0	76.19	7.62	0	0	0	0	3.81	2.86
Brt	0	0	0	0	0	5.79	0	0	0	0	0.83	0.83	0	0.83	4.96	85.95	0	0	0.83	0	0	0
Turquoise	0	0	2.86	0	0	0	0	0	0	0	0	0	0	0	0	0	97.14	0	0	0	0	0
Tourmaline	0	0	0	1.14	2.27	0	1.14	1.14	1.14	1.14	1.14	0	0	0	0	0	0	89.77	0	0	1.14	0
Mi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98.91	0	0	1.09
Azr	0	0	0	0	1.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98.73	0	0
Rds	1.16	0	1.16	0	0	0	1.16	0	1.16	0	0	0	0	0	0	1.16	0	0	0	0	94.19	0
Fl	0	0	0	0	0	6.59	0	2.2	0	1.1	0	0	0	0	5.49	3.3	0	0	2.2	0	0	79.12

Obviously, minerals with wide color coverage, large shape differences, and no clear characteristics are easily confused with others. The model cannot determine a clear classification boundary between these minerals, resulting in a low accuracy. For example, quartz, calcite, and barite can be transparent, white, yellow, pink, and other colors, and have many crystal shapes, which are prone to being confused with each other (8.57% of calcite is misjudged as quartz and 7.62% is misjudged as barite).

4.2. Feature Extraction Performance Analysis

t-Distributed Stochastic Neighbor Embedding (t-SNE) [31] is a technique for dimensionality reduction. It has been widely used to visualize the effectiveness of the algorithm by mapping the high-dimensional features to 2-d or 3-d vectors, and those can be displayed conveniently. For each type of the 22 minerals, we selected five images with typical characteristics (some samples are shown in Figure 4). Next, the images were input into the trained network to extract their feature vectors $v \in \mathbb{R}^C$ and v that uniquely represents the image, where C is 2560 due to feature fusion. Then, the feature vectors of the 110 images were reduced to two dimensions with t-SNE and displayed subsequently in Figure 5a, where each image is marked with the serial number of the mineral, as given in Table 2. In order to illustrate the misrecognition, we choose two images with indistinct characteristics, one of which is quartz and the other is calcite. The misrecognized images are marked with red cycles in Figure 5a.

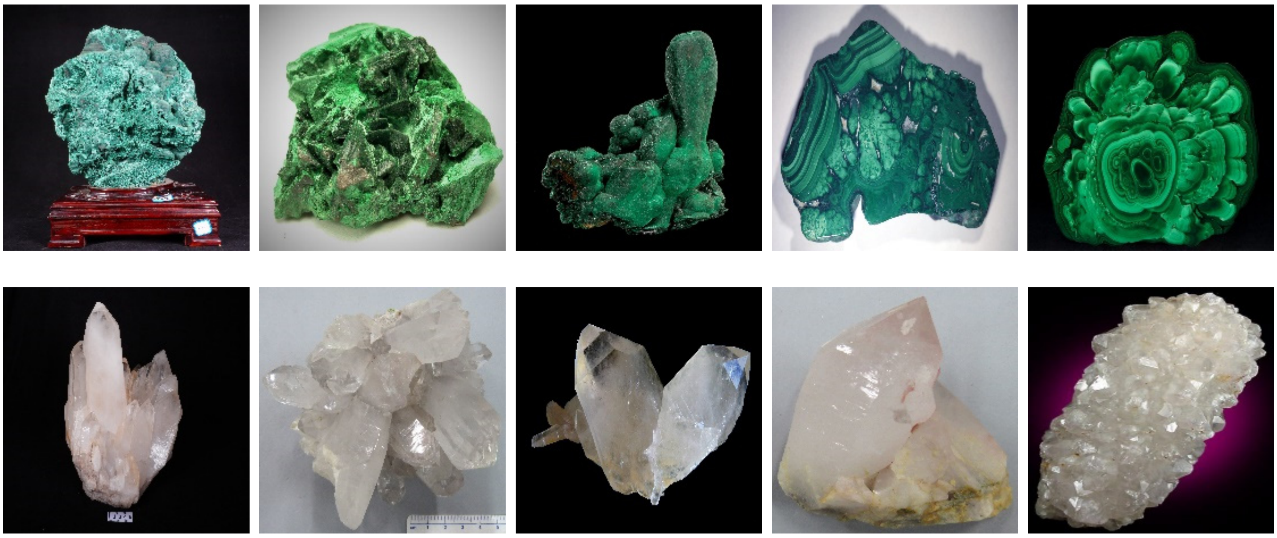
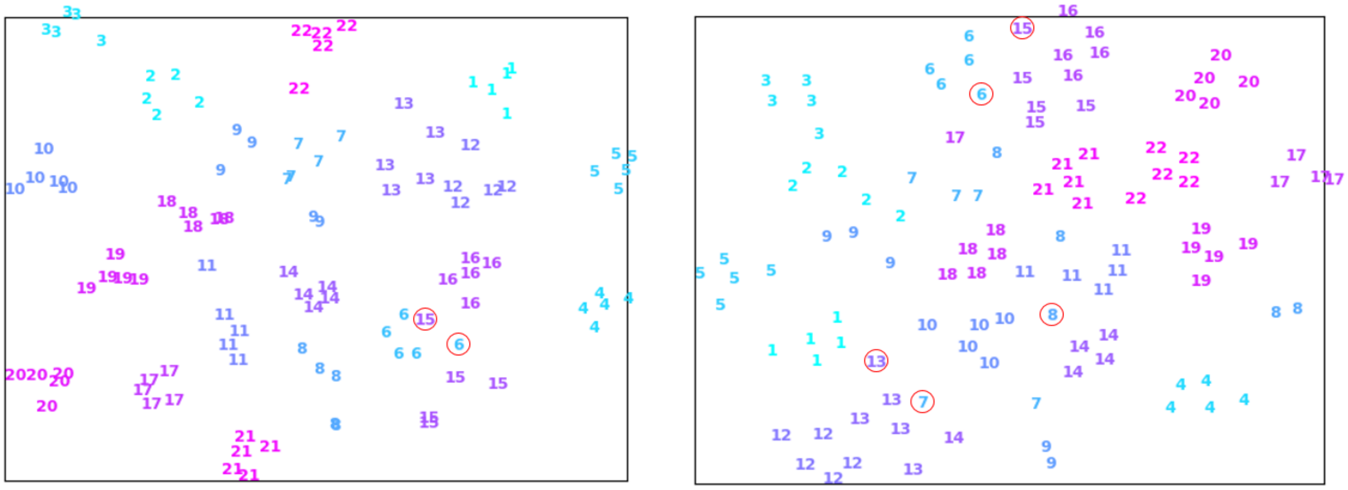


Figure 4. Samples of minerals with typical features. The first and the second row are malachite and quartz, respectively.

On the whole, the features from the same kind are well aggregated and those from different kinds are distinguished obviously. More specifically, for minerals with discriminative characteristics, such as stibnite (No. 4), turquoise (No. 17), malachite (No. 19), and azurite (No. 20), the intra-class distances are small and the inter-class distances are obviously large. However, for minerals with similar appearances, such as quartz (No. 6) and calcite (No. 15), the inter-class margin is small. In Figure 5a, a calcite image is misrecognized as quartz and a quartz image is misrecognized as calcite. Similarly, for muscovite (No. 12) and biotite and phlogopite (No. 13), although they can be divided into two clusters, the inter-class distance is small and the intra-class distance is large.

As a comparison, we use ResNet-50 combined with the softmax loss to obtain features and map them to two-dimensional vectors with t-SNE, as shown in Figure 5b. It can be seen that the features extracted by our model have better aggregation and accuracy than ResNet-50.

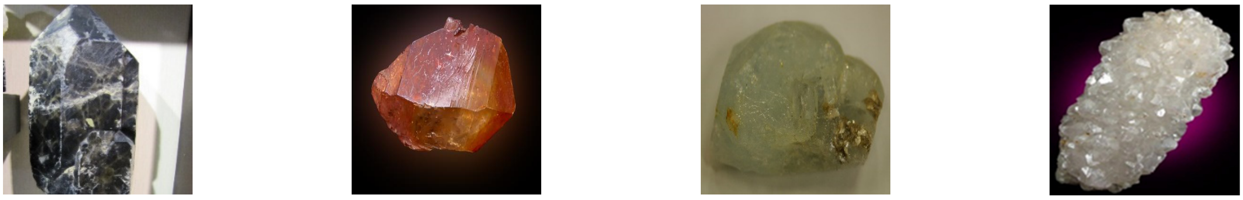


(**a**) our model (**b**) ResNet-50

Figure 5. Visualization of different latent representations of mineral images with t-SNE. Five images for each kind of mineral are given. Misrecognized images are marked by red circles. (**a**) and (**b**) are t-SNE embedding based on our model and ResNet-50 with softmax loss, respectively.

4.3. Misrecognized Images Analysis

Quartz is often in the shape of a long prism with a sharp tapered tip and there are stripes vertical to the direction of crystal extension. Figure 6 shows misrecognized quartz samples of the dataset-1 validation set, in which (a)–(d) was misidentified as muscovite, garnet, beryl, and calcite, respectively. The misrecognition of images can be attributed to the extracted features from the quartz images being more similar to other categories. For example, Figure 6a is similar to the characteristics of mica, the color of (b) and (c) are more close to that of garnet and beryl, and the shape of (d) is more similar to calcite. In the misrecognized photos, most of them do not show the characteristics of the mineral due to the inappropriate shooting angle, clarity, and lighting conditions, etc., so it is also difficult for geologists to distinguish them from the photos alone. More mineral photos showing the correct characteristics can clearly guide the neural network to learn the discriminative features of minerals well and improve recognition accuracy.



(**a**) (**b**) (**c**) (**d**)

Figure 6. Misrecognized examples of quartz photo images in dataset-1. (**a**–**d**) are misrecognized as muscovite, garnet, beryl, and calcite respectively.

5. Conclusions

This paper proposes a common mineral recognition method using only mineral photos. In this method, we merge the multi-resolution features extracted from ResNet-50 and propose a weighted top-*k* loss function to balance the importance of hard and easy samples. Since the low-level features, such as color, shape, and so on, are important for mineral recognition, the fused features can supplement the high-level information that suffers from missing details, and the weighted top-*k* loss function better balances the roles of hard and easy samples for network learning. They effectively improve the accuracy of mineral photo recognition. Of the 14,986 image datasets of 22 common minerals, the experimental results show that the proposed method achieves a Top1 accuracy of 88.01% and mAP of 87.15%, which surpasses the Top1 accuracy of Inception-v3, EfficientNet-B0, and ResNet-50 with softmax loss by a margin 1.88%, 1.29%, and 0.86%, respectively, achieving a promising mineral recognition performance. Experimental analysis illustrates the excellent feature extraction performance of the method we proposed and we know that aside from improving the performance of the algorithm, collecting more diverse samples with clear and discriminative characteristics is a feasible and effective way to increase recognition accuracy.

Author Contributions: Conceptualization, L.J. and H.L.; methodology, L.J., H.L., M.Y. and M.H.; validation, M.Y. and F.M.; writing—original draft preparation, L.J.; writing—review and editing, H.L., M.Y. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Program of National Mineral Rock and Fossil Specimens Resource Center from MOST. And the APC was funded by National Natural Science Foundation of China grant number 61472119.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: The authors show a real appreciation for the rich and clear mineral pictures provided by the National Mineral Rock and Fossil Specimens Resource Center of China and mindat.org. Our deepest gratitude goes to the anonymous reviewers for their careful work and thoughtful suggestions that have helped improve this paper substantially.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singh, N.; Singh, T.N.; Tiwary, A.; Sarkar, K.M. Textural identification of basaltic rock mass using image processing and neural network. *Comput. Geosci.* **2010**, *14*, 301–310. [CrossRef]
2. Li, N.; Hao, H.Z.; Gu, Q.; Wang, D.R.; Hu, X.M. A transfer learning method for automatic identification of sandstone microscopic images. *Comput. Geosci.* **2017**, *103*, 111–121. [CrossRef]
3. Chan, S.; Elsheikh, A.H. Parametric generation of conditional geological realizations using generative neural networks. *Comput. Geosci.* **2019**, *23*, 925–952. [CrossRef]
4. Jiang, L.S.; Zhao, Y.; Golsanami, N.; Chen, L.J.; Yan, W.C. A novel type of neural networks for feature engineering of geological data: Case studies of coal and gas hydrate-bearing sediments. *Geosci. Front.* **2020**, *11*, 1511–1531. [CrossRef]
5. Laloy, E.; Herault, R.; Lee, J.; Jacques, D.; Linde, N. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour.* **2017**, *110*, 387–405. [CrossRef]
6. Li, S.; Chen, J.P.; Xiang, J. Applications of deep convolutional neural networks in prospecting prediction based on two-dimensional geological big data. *Neural Comput. Appl.* **2020**, *32*, 2037–2053. [CrossRef]
7. Palafox, L.F.; Hamilton, C.W.; Scheidt, S.P.; Alvarez, A.M. Automated detection of geological landforms on Mars using Convolutional Neural Networks. *Comput. Geosci.* **2017**, *101*, 48–56. [CrossRef]
8. Tan, Q.L.; Huang, Y.; Hu, J.; Zhou, P.G.; Hu, J.P. Application of artificial neural network model based on GIS in geological hazard zoning. *Neural Comput. Appl.* **2021**, *33*, 591–602. [CrossRef]
9. Baykan, N.A.; Yılmaz, N. A Mineral Classification System with Multiple Artificial Neural Network Using K-Fold Cross Validation. *Math. Comput. Appl.* **2011**, *16*, 22–30. [CrossRef]
10. Baklanova, O.E.; Baklanov, M.A. Methods and Algorithms of Image Recognition for Mineral Rocks in the Mining Industry. In *Advances in Swarm Intelligence*; Springer International Publishing: New York, NY, USA, 2016; pp. 253–262.
11. Izadi, H.; Sadri, J.; Bayati, M. An Intelligent System for Mineral Identification in Thin Sections Based on a Cascade Approach. *Comput. Geosci.* **2017**, *99*, 37–49. [CrossRef]
12. Maitre, J.; Bouchard, K.; Bedard, L.P. Mineral grains recognition using computer vision and machine learning. *Comput. Geosci.* **2019**, *130*, 84–93. [CrossRef]
13. Zhang, Y.; Li, M.; Han, S.; Ren, Q.; Shi, J. Intelligent Identification for Rock-Mineral Microscopic Images Using Ensemble Machine Learning Algorithms. *Sensors* **2019**, *19*, 3914. [CrossRef]
14. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
15. Li, C.X.; Wang, D.M.; Kong, L.Y. Application of Machine Learning Techniques in Mineral Classification for Scanning Electron Microscope-Energy Dispersive X-ray Spectroscopy (SEM-EDS) Images. *J. Pet. Sci. Eng.* **2021**, *200*, 1–30. [CrossRef]
16. Ishikawa, S.T.; Gulick, V.C. An automated mineral classifier using Raman spectra. *Comput. Geosci.* **2013**, *54*, 259–268. [CrossRef]

17. Carey, C.; Boucher, T.; Mahadevan, S.; Bartholomew, P.; Dyar, M.D. Machine learning tools formineral recognition and classification from Raman spectroscopy. *J. Raman Spectrosc.* **2015**, *46*, 894–903. [CrossRef]
18. Zeng, X.; Xiao, Y.; Ji, X.; Wang, G. Mineral Identification Based on Deep Learning That Combines Image and Mohs Hardness. *Minerals* **2021**, *11*, 506. [CrossRef]
19. Liu, C.Z.; Li, M.C.; Zhang, Y.; Han, S.; Zhu, Y.Q. An Enhanced Rock Mineral Recognition Method Integrating a Deep Learning Model and Clustering Algorithm. *Minerals* **2019**, *9*, 516. [CrossRef]
20. Peng, W.H.; Bai, L.; Shang, S.W.; Tang, X.j.; Zhang, Z.y. Common mineral intelligent recognition based on improved InceptionV3. *Geol. Bull. China* **2019**, *38*, 2059–2066.
21. Lecun, Y.B.L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
23. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–11.
24. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
25. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1–12.
26. Tan, M.X.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
27. Zhang, K.P.; Zhang, Z.P.; Li, Z.F.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
28. National Infrastructure of Mineral Rock and Fossil Resources for Science and Technology. Available online: http://www.nimrf.net.cn/ (accessed on 27 November 2021).
29. Mindat.Org-Mines, Minerals and More. Available online: https://www.mindat.org/ (accessed on 27 November 2021).
30. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
31. Laurens, V.D.M.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.