Automatic Laboratory Martian Rock and Mineral Classification Using Highly-Discriminative Representation Derived from Spectral Signatures

Juntao Yang 1,2,3, Zhizhong Kang 2,3,4,* Ze Yang 2,3,4, Juan Xie 2,3,4, Bin Xue 5, Jianfeng Yang 5 and Jinyou Tao 5

1 College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
2 School of Land Science and Technology, China University of Geosciences, No. 29 Xueyuan Road, Haidian District, Beijing 100083, China
3 Subcenter of International Cooperation and Research on Lunar and Planetary Exploration, Center of Space Exploration, Ministry of Education of The People’s Republic of China, No. 29 Xueyuan Road, Haidian District, Beijing 100083, China
4 Lunar and Planetary Remote Sensing Exploration Research Center, China University of Geosciences, No. 29 Xueyuan Road, Haidian District, Beijing 100083, China
5 Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an 710119, China

* Correspondence: zzkang@cugb.edu.cn

Abstract: The optical properties of rocks and minerals provide a reliable way to measure their chemical and mineralogical composition due to the specific reflection behaviors, which is also the key insight behind most automatic identification and classification approaches. However, the inter-category spectral similarity poses a great challenge to the automatic identification and classification tasks because of the diversity of rocks and minerals. Therefore, this paper develops a recognition and classification approach of rocks and minerals using the highly discriminative representation derived from their raw spectral signatures. More specifically, a transformer-based classification approach integrated with category-aware contrastive learning is constructed and trained in an end-to-end manner, which would force instances of the same category to remain close-by while pushing instances of a dissimilar category far apart in the high-dimensional feature space, in order to produce the highly discriminative feature representation of the rocks and minerals. From both qualitative and quantitative views, experiments are conducted on the laboratory sample dataset with 30 types of rocks and minerals shared from the National Mineral Rock and Fossil Specimens Resource Center, and the spectral information of the laboratory rocks and minerals is captured using a multi-spectral sensor, with a duplicated payload of the counterpart onboard the Zhurong rover. Quantitative results demonstrate that the developed approach can effectively distinguish 30 types of rocks and minerals, with a high overall accuracy of 96.92%. Furthermore, the developed approach is remarkably superior to other existing methods, with average differences of 4.75% in the overall accuracy. Furthermore, we also visualized the derived highly discriminative features of different types of rocks and minerals by projecting them onto a two-dimensional map, where the same categories tend to be modeled by nearby locations and the dissimilar categories by distant locations with high probability. It can be observed that, compared with those in the raw spectral feature space, the clusters are formed better in the derived highly discriminative feature space, which further confirms the promising representation capability.

Keywords: identification and classification; transformer; highly discriminative representation; contrastive learning; multi-spectral sensor; Mars exploration

1. Introduction

Rocks and minerals are one of the major planetary surface features. Since minerals are stable under the known ranges of temperature and pressure, a rock, made of specific
minerals, can be used to identify and deduce the climate or weathering when the rocks were being formed and deposited, based on individual mineral stability ranges and the presence or absence of equilibrium between them [1]. Obviously, the characterization of the chemical and mineralogical composition of rocks and minerals, as well as their alteration over time, are highly complementary to the investigation of the climatic and geologic history of the planetary bodies [2–5]. With the development of deep space exploration, spectroscopic systems (e.g., Raman, near-infrared spectroscopy), as one of most important scientific payloads onboard the orbiters, landers, and rovers, provide a powerful tool for the investigation and characterization of the chemical and mineralogical nature of planetary surfaces [6–9], since the reflection behavior of rocks and minerals is primarily determined by their compositions and physical properties [4].

Accurate determination of the presence of rocks and minerals from spectroscopic data is a fundamental task of planetary exploration and science [10]. Viviano et al. [11] corrected spectral reflectance at key wavelengths in a compact reconnaissance imaging spectrometer for Mars (CRISM) to produce the revised spectral parameters, which revealed an increasingly diverse suite of minerals present on the Martian surface. Jain and Chauhan [12] adopted a spectral reflectance-based mineral detection method to reveal the presence of minerals from CRISM data, which hinted of the aqueous history of the Noachian period on Mars. Fox et al. [13] detected Fe$^{3+}$-Mg$^{2+}$ smectite deposits in Marathon Valley using five CRISM observations. Núñez et al. [14] obtained the spectral information of gully sediments from CRISM data and used them to identify a variety of mineralogies for constraining the gully-forming processes. Xue et al. [15] identified the category of minerals and mapped their distributions from CRISM near-infrared spectral data in the Martian Gale and Nili Fossae regions. Amador et al. [16] used factor analysis and target transformation methods to produce the global distribution of minerals associated with serpentinization from CRISM spectral data. Lin et al. [17] jointly identified the hydrated mineral deposits and determined their abundances with a large endmember library using CRISM near-infrared data. These aforementioned identification methods generally use simple mathematical functions to judge whether specific spectral features are absent or not for achieving the mineralogical analysis. While these existing classification methods are effective for the manual analysis, the spectral artifacts and noises might affect their performance. Thus, automatic identification and classification of rocks and minerals based on their spectral information has been an active issue to date.

Automatic rock and mineral classification based on their spectral signatures is greatly challenging due to the inter-category spectral similarity [18] and lots of machine learning-based methods, generally consisting of feature representation and recognition systems, have been proposed using spectral information, such as Tsallis entropy [19], fractal dimension [20], Sato’s maximum Lyapunov exponent [21], color histogram-based features [22], etc. To further enhance the descriptive capability of the feature representation, several feature selection and optimization strategies, including information gain ranking [23], principal component analysis [24], genetic optimization [24], and K-means [25] have also been developed to autonomously learn the high-level feature representation. Following this, some recognition and classification systems of rocks and samples [18,22,25–29] have then focused on fast and reliable identification without human intervention.

Recently, deep learning technology has achieved great successes and has attempted to deal with the classification and recognition issues because of its inspirational performance in the field of computer vision. As a state-of-the-art deep learning model, the transformer provides a strong representation capability to model and encode heterogeneous interactions [30], and its tremendous success in the language domain has led researchers to investigate its adaptation to computer vision [31]. Hence, this work concentrates on using it to generate the highly discriminative representation of rocks and minerals from their raw spectral signatures. Pascual [32] explored the superior capability of convolutional neural network for classifying different types of rock images, especially in the natural scenes. Saranathan and Parente [33] developed a generative adversarial network-based feature
representation learning method to map mineral signatures of interest across the CRISM image database. To deal with the problem of limited samples, Li et al. [18] adopted the idea of transfer learning to fine-tune the super-parameters of the pre-trained convolutional neural network.

To date, some institutes and organizations have been also devoted to establishing the laboratory spectral database of rocks and minerals (e.g., the Berlin Emissivity Database, United States Geological Survey (USGS) spectral library version 7) to study their spectral behaviors [5,34–38]. In this way, the reflectance spectra of terrestrial rocks and minerals, covering at the range of visible to near-infrared wavelengths, are explored and analyzed in laboratory conditions, which is beneficial to further supporting the research of planetary origin and geological evolution [4,27,39,40]. Similarly, this work captures the spectral information of the laboratory rocks and minerals using the multi-spectral sensor, which duplicates the payload of its counterpart onboard the Zhurong rover [41], and focuses on investigating the highly discriminative feature representations from the raw spectral information for the classification task.

Although there have been numerous presented studies related to detection and classification [2,25,29,42–45], automatic rock and mineral classification remains greatly challenging due to the following aspects. First, according to the literature statistics, the majority of existing classification approaches tended to only use a few types of rocks and minerals, where the number of the category in these work is generally less than 10, to verify their own robustness and effectiveness [25,29,43]. As a matter of fact, the increasing number of categories might result in the severe inter-category similarity, which will pose a challenge in the task of accurate classification. In this case, a dataset containing as many types of rocks and minerals as possible is very necessary for verifying the generalization of the classification approaches. Second, although lots of feature extraction and selection strategies have been developed for describing the spectral and texture information of rocks and minerals [25,29,43], the inter-category similarity might still affect the descriptive capability and reduce the classification quality [46] as the number of categories increases. Obviously, a highly discriminative feature representation plays a significant role in the classification task to enhance the intra-category similarity and enlarge the inter-category variants.

To address these challenges mentioned above, we develop a rock and mineral classification approach using the highly discriminative representation derived from their original spectral signatures. Moreover, experiments are carried out on the multi-spectral data captured from 30 types of rocks and minerals for both qualitatively and quantitatively evaluating the robustness and reliability of the developed approach. Our main contributions in this work are as follows.

1. To efficiently achieve the classification task, we design a transformer-based classification approach for generating the highly discriminative feature representation of both rocks and minerals, where the inter-category representation variant is enlarged and the intra-category representation similarity is aggregated;

2. A category-aware contrastive learning is integrated within the developed transformer-based classification approach. In this case, the super-parameters of the whole network are learned and trained in an end-to-end multi-task manner. Consequently, the remarkable distinctions among different types of rock and minerals occur in their high-dimensional feature space;

3. We demonstrate the reliability and robustness of the developed approach on a dataset containing rocks and minerals with complicated categories. It is of significance for the investigation of the developed approach’s generalization ability.

The rest of this paper is organized as follows. Section 2 describes the experimental data and the developed classification approach in detail. Section 3 presents the experimental results and analysis for evaluating the developed classification approach both quantitatively and qualitatively. Section 4 discusses the sensitivity of parameters and analyzes the
strengths and limitations of the developed classification approach. This paper concludes with a summary of future research considerations in Section 5.

2. Materials and Methods

In this section, we will introduce the proposed method in detail. This section is organized as follows. First, the details of data acquisition are described. Second, we introduce the overview of the proposed method. Subsequently, we present other detailed implementation of the proposed method.

2.1. Data Acquisition

When the number of categories is large, the probability of classification error becomes relatively large. This is because each category is surrounded by a large number of neighboring categories [47], which results in severe inter-category spectral similarity. Consequently, a sample dataset of rocks and minerals with as many categories as possible is acquired from the National Mineral Rock and Fossil Specimens Resource Center. The National Mineral Rock and Fossil Specimens Resource Center is devoted to the digitization and sharing of China’s rock mineral specimens, offering more than 170,000 state-owned specimens with scientific value, including minerals, rocks, and fossils. As a data provider, its goal is to promote the understanding of geo-resources and to provide resources for academia, education, and scientific popularization in the field of geoscience through collecting, organizing, and sharing rock mineral specimen resources.

As for the implemented experiments in this work, we reviewed the relevant literature about the distribution of rocks and minerals on the Martian surface and selected 30 types of rocks and minerals, e.g., hydrous minerals. All the used rocks and minerals are from the National Mineral Rock and Fossil Specimens Resource Center, including muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smaltite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr). Moreover, we also enrich the number of categories of rocks and minerals as much as possible in the sample dataset, although a few rocks and minerals on the list are very uncommon and even unlikely to be found on Mars. By establishing a complicated sample dataset, the generalization of the proposed method can be better evaluated.

Generally, most rocks and minerals can be characterized and classified by their unique physical properties (e.g., hardness, luster, color, cleavage, fracture). For instance, barite can be reliably recognized based on its three directions of right-angle cleavage and the sugary appearance. Gypsum is characterized by its softness and its three directions of unequal cleavage. Biotite vermiculite is almost always much darker in color than muscovite. Talcum has a greasy feel. Psilomelane occurs as botryoidal and stalactitic masses with a smooth shining surface and submetallic luster. Quartz is characterized by its glassy luster, conchooidal fracture, and crystal form. The most obvious physical properties of serpentine are its green color, patterned appearance, and slippery feel. Stibnite typically forms coarse, irregular masses or radiating sprays of needlelike crystals, and its distinguishing characterization is easy fusibility, a bladed habit, perfect cleavage in one direction, lead-gray color, and soft black streaks. Figure 1 illustrates examples of different types of both rocks and minerals. It is obvious that similar mineralogy or appearance characteristics among some types of rocks and minerals make them very challenging to differentiate via direct visual interpretation. For example, both aragonite and calcite with the same formula are tabular, prismatic, or needlelike, often with steep pyramidal or chisel-shaped ends, and can form columnar or spreading aggregates. Basalt consists mainly of plagioclase and pyroxene minerals similar to gabbro, but the former is fine-grained, and the latter is coarse-grained. The yellow color and metallic luster are the most obvious physical properties of chalcopyrite, which shows a similar appearance to pyrite. Chlorite is usually green in color, has a
foliated appearance, inelastic cleavage, and an oily to soapy feel, but its variable chemical composition makes it a difficult specimen to. Galena can be easily recognized by its metallic cleavage planes, lead gray and silvery color, and black streaks, which are often associated with sphalerite, calcite, and fluorite. Hematite has an extremely variable appearance, from earthy to submetallic to metallic luster, but it always produces a reddish streak. Smaltite crystallizes in the cubic system with the same hemihedral symmetry as pyrite.

Figure 1. Instances of different types of rocks and samples in the laboratory. For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smallite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr).

The reflection behavior of rocks and minerals primarily depends on their own mineralogical compositions and physical properties [4]. To capture the reflectance spectra of rocks and minerals in this work, the eight-band multi-spectral camera, which is the alternative payload of its counterpart onboard the Zhurong rover, is used as the multi-spectral sensor [9], and it is able to provide spectral data at the following wavelengths: 480 nm, 525 nm, 650 nm, 700 nm, 800 nm, 900 nm, 950 nm, and 1000 nm. The specifications of the used multispectral camera are listed in Table 1. In our implementation, we only use the multi-spectral camera to capture the eight-band multispectral images of all rock/mineral samples at approximately 12:12 pm, during a period of bright sunshine. Solar elevation angle and the shooting angle of camera is approximately 60° and 37°, respectively. All rock/mineral samples are placed on the ground, and the vertical height of camera from the ground is 1.8 m. As a result, the size of the captured image corresponding to each rock or mineral is \( H \times W \times 8 \), where \( H \) denotes the image height, \( W \) denotes the image width.

Table 1. Characteristics of the multispectral camera onboard the Zhurong rover.

| Parameters                  | Values                      |
|-----------------------------|-----------------------------|
| Number of channels          | 9 (panchromatic and multispectral) |
| Weight                      | 1.2 kg                      |
| Geometric resolution        | \( 2048 \times 2048 \)       |
| Radiometric resolution      | 10 bits                     |
| Pixel size                  | 5.5 \( \mu m \)             |
| Focal length                | 50 mm                       |
| Imaging distance            | \( (1.5 \text{ m, } \infty) \) |

Because Spectralon’s optical properties make it ideal as a reference surface in remote sensing and spectroscopy, we used Spectralon as a near-Lambertian reference standard. The images of both the standard reference Spectralon and rocks/minerals are captured.
where $G_{\text{mn}}$ denotes the average gray values of the query sample, $G_{\text{ref}}$ denotes the average gray values of the standard reference Spectralon, $R_{\text{ref}}$ denotes the reflectance coefficient of the standard reference Spectralon that is generally known and measured in the laboratory, and $R$ denotes the computed reflectance of the query sample.

Figure 2 demonstrates some typical examples of the spectral curves corresponding to rocks and minerals, and shows the changes in the band position and shape of compositions because of the unique spectral signatures due to their own physical characteristics. In the real world, small variations in the composition of rocks and minerals might occur, which often causes shifts in the position and shape of absorption bands in the spectrum. We put the spectral curves of the used (partial) samples together into one subfigure due to the similar spectral characteristics. It can be observed that these shifts in the position and shape of absorption bands might result in the similarity and even the overlap of spectral curves, i.e., severe inter-category similarity, such as in the case of chalcopyrite and fluorite, or with pyrite and saponite.

![Figure 2](image_url)

**Figure 2.** Illustration of spectral curves corresponding to rocks and minerals. For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smallite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr).

### 2.2. The Developed Classification Approach of Rocks and Minerals

It is common knowledge that the reflection behaviors of rocks and minerals depend on their own physical characteristics, which is the key insight behind most existing rock and mineral classification approaches. Despite their unique spectral signatures, the inter-category spectral similarity among different types of rocks and minerals might still reduce the classification quality as the number of categories increases. Therefore, using a transformer encoder as the backbone, we integrate it with contrastive learning and develop a classification method of rocks and minerals based on the highly discriminative representation derived from the original spectral signatures. Figure 3 indicates the pipeline of the developed classification method, which consists of a transformer-based feature encoder
module and multi-task loss function for optimization. The former is able to remarkably enhance the descriptive capability of the derived feature representation [31] while the latter is to cause the feature representations of similar categories to aggregate with each other while the feature representations of dissimilar categories separate from each other. In our developed approach, an image patch $x \in \mathbb{R}^{h \times w \times d}$ (where $h$ is the height, $w$ is the weight, and $d$ is the channels; in this work, $d = 8$) is first operated via the simple linear flatten method, serving as the input of the transformer-based feature encoder module.

![Figure 3. A pipeline showing the developed rock and mineral classification approach.](image-url)

### 2.2.1. Transformer-Based Feature Encoder Module

The vision transformer as a deep learning model shows a remarkable capability to capture rich dependency information between variables [49]. We explore and investigate the adaptation of the vision transformer for basic visual feature extraction in order to generate the highly discriminative feature representation. It is found that, unlike the convolutional neural networks (CNNs) that gradually expand the field of view by repeatedly “convoluting” the information around the kernel layer by layer, the transformer-based method uses the stacked multi-head attention module that allows its strong ability to model the long-range dependencies. Recently, the vision transformer has adopted the attention mechanism and achieved promising results for image classification. Generally, an input image is first converted into a sequence of tokens by dividing it with a certain patch size and then linearly projecting each patch into tokens. Then, when a sequence of tokens is passed into a vision transformer model, attention weights are calculated between every token simultaneously. That is to say, the attention weight $a_{ij}$ of token $z_i$ with respect to token $z_j$ is learned, which suggests the relevant information to each token. After calculating the $a_{ij}$ value for all $i$ and $j$ pairs, we update each token $z_j$ to $z'_j$ using a weighted sum of all tokens followed by a nonlinear ReLU layer. This is defined in the following Equations (2)–(4):

$$a_{ij} = \text{softmax} \left( \frac{(W^d z_i)(W^h z_j)}{\sqrt{d}} \right) \quad (2)$$

$$\mathbf{z}_i = \sum_{j=1}^{h \times w} a_{ij} W^o z_j \quad (3)$$

$$z'_j = \text{ReLU} (\mathbf{z}_i W^e + b_1) W^o + b_2 \quad (4)$$
where \(d\) denotes the dimension of the key vector, \(W^k\) denotes the key weight matrix, \(W^q\) denotes the query weight matrix, \(W^v\) denotes the value weight matrix, \(W^r\) and \(W^o\) are the transformation matrices, and \(b_1\) and \(b_2\) are the bias terms.

One set of \((W^k, W^q, W^v)\) matrices is called an attention head, and each layer in a vision transformer model has multiple attention heads, a module for attention mechanisms which runs through an attention mechanism several times in parallel. While each attention head attends to the tokens that are relevant to each token, the model can do this for different definitions of “relevance” with multiple attention heads.

As shown in the subfigure about the highly discriminative representation learning in Figure 3, a transformer-based feature encoder module is composed of a sequence of blocks where each block contains the multi-head attention module. Following this, three successive feedforward networks are used to produce the highly discriminative feature representation for the final label predictions. Empirically, the size of three successive fully connected layers is set to \(h \times w \times d, 360\), and the number of classes, respectively. As a result, the output of the whole network, i.e., the last fully-connected layer, is considered as the feature representation of rock and mineral derived from the developed method.

### 2.2.2. Multi-Task Loss Function for Optimization

The super-parameters within the whole network are learned for mapping a set of the feature representations of rocks and minerals to a set of categories from massive high-quality labeled training data. General speaking, the problem of learning is cast as a search or optimization problem, which navigates the space of possible sets of super-parameters within the whole network in order to make good or good enough predictions. In the context of an optimization process, the function used to evaluate a candidate solution (i.e., a set of super-parameters) is referred to as the loss function, and the value calculated by the loss function is referred to simply as “loss”. Typically, a deep learning model is learned using the stochastic gradient descent optimization algorithm, and super-parameters are updated using the backpropagation of error algorithm so that the next evaluation reduces the error, which means that we are searching for a candidate solution that has the lowest score. Here, we primarily discuss how to design an effective loss function, which jointly optimizes both cross entropy loss and category-aware contrastive loss. It enables us to guide the developed model to move towards convergence via measuring the difference between the predicted output and the ground truth. As a consequence, a highly discriminative feature representation is generated from the learned model.

Due to its fast convergence, the cross-entropy loss \(L_{\text{cls}}\) is used to guarantee accurate classification of the categories, as defined in the following Equation (5):

\[
L_{\text{cls}} = \sum_{i=0}^{C} \left\{ y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right\}
\]

(5)

where \(C\) denotes the total number of categories, \(\hat{y}_i\) denotes the \(i\)th true category of the training samples, and \(y_i\) denotes the \(i\)th predicted category from our developed model.

In addition to the cross-entropy loss, the class separation in the latent feature space would also be an ideal characteristic to discriminate among different types of rocks and minerals. Therefore, we define the category-aware contrastive loss, as defined in Equation (6). The right subfigure in Figure 3 gives an explicit description about the role of category-aware contrastive loss. Intuitively, the goal of the category-aware contrastive loss is to force instances of the same category to remain close-by while pushing ones in the dissimilar category far apart in their latent feature space.

More specifically, we consider the set of categories as \(K = \{1, 2, \ldots, C\} \subset N^+\), where \(N^+\) represents the set of positive integers. For each category \(i \in K\), a mean feature representation \(p_i\) is represented and maintained as the specific formed cluster in the latent feature space, which makes up a set of category-specific mean feature representations, namely \(P = \{p_1, \ldots, p_C\}\), for computing the category-aware contrastive loss. During the
training procedure, the mean of feature representation for each category makes up a set of category-specific mean feature representations, namely \( P = \{ p_1, \ldots, p_C \} \), for computing the category-aware contrastive loss. Since the super-parameters of the whole network is learned in an end-to-end manner, the mean feature representation corresponding to each category would gradually evolve at the training time. Inspired by the contrastive clustering method [50], a queue \( q_i \) with a fixed length is maintained for the \( i \) th category to store its associated feature representation. As a result, a feature representation store \( F_{\text{store}} = \{ q_1, \ldots, q_C \} \) stores the category-specific feature representations in the corresponding queues. After every \( I_p \) iterations, a set of new category-specific mean feature representation \( P_{\text{new}} \) is calculated. Following this, the existing set of category-specific mean feature representations, namely \( P \), is updated by weighting \( P \) and \( P_{\text{new}} \) with a momentum constant \( \eta \), as defined in Equation (6). In our implementation, \( \eta \) is set to 0.99, while \( I_p \) is set to 2000.

\[
P = \eta P + (1 - \eta) P_{\text{new}} \tag{6}
\]

After obtaining the set of category-specific mean feature representations, let \( f_c \) denote a feature representation produced by an intermediate layer of the used transformer-based feature encoder module, for an instance of category \( c \). To force instances of the same category to remain close-by while pushing instances of the dissimilar category far apart in the high-dimensional feature space, the category-aware contrastive loss is defined as follows:

\[
L_{\text{cont}}(f_c) = \sum_{i=0}^{C} l(f_c, p_i) \tag{7}
\]

\[
l(f_c, p_i) = \begin{cases} 
D(f_c, p_i), i = c \\
\max\{0, \Delta - D(f_c, p_i)\}, \text{otherwise}
\end{cases} \tag{8}
\]

where \( D(\cdot) \) denotes any distance function (e.g., Euclidean, cosine), and \( \Delta \) denotes how close a similar and dissimilar instance can be. In our implementation, the value of \( \Delta \) is empirically set to 2.0.

Finally, we simultaneously learn the shared super-parameters through the backpropagation algorithm [51] for optimization to provide better generalization of the developed classification approach [52]. The total loss used in this work is defined as a weighted sum of both the category-aware contrastive loss and the cross-entropy loss as the following Equation (9):

\[
\text{Loss}_{\text{total}} = \lambda \times L_{\text{cont}} + L_{\text{cls}} \tag{9}
\]

where \( \lambda \) is experimentally set to 0.01, \( L_{\text{cont}} \) denotes the category-aware contrastive loss, and \( L_{\text{cls}} \) denotes the cross-entropy loss by measuring the cross entropy between the ground truth and the predicted output of our developed model. By minimizing the total loss, our developed classification approach is progressively trained for optimization until convergence.

3. Experimentation and Analysis

To evaluate the reliability and robustness of the developed classification method, in this section, we performed both qualitative and quantitative analysis on the multi-spectral images of 30 types of rocks and minerals from the National Mineral Rock and Fossil Specimens Resource Center. First, a brief description of evaluation criteria is given. Then, we offer a detailed experimental setting. Finally, we qualitatively and quantitatively analyze the performance of category classification results, where the experiments are conducted using the optimal parameters obtained and discussed in Section 5.

3.1. Evaluation Criteria

To measure the classification quality of rocks, we conduct the evaluation solution using precision (Pr), recall (Re), \( F_1 \)-score and overall accuracy (OA), as defined in Equations (10)–(13). The recall represents a measure of completeness, the precision denotes a measure of correct-
ness, and the $F_1$-score, also called balanced $F$-score, is a weighted average of the precision and recall; overall accuracy denotes the sum of the true positives plus true negatives divided by the total number of queried individuals. The equations are as follows:

$$
\text{Precision} = \frac{TP}{TP + FP} \quad (10)
$$

$$
\text{Recall} = \frac{TP}{TP + FN} \quad (11)
$$

$$
F_1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (12)
$$

$$
\text{Overall accuracy} = \frac{TP + FN}{TP + TN + FP + FN} \quad (13)
$$

where $TP$ denotes the number of positive objects that are correctly determined as positive ones, $TN$ denotes the number of positive objects that are correctly determined as negative ones, $FN$ denotes the number of negative objects that are incorrectly classified as negative ones, and $FP$ denotes the number of negative objects that are incorrectly determined as positive ones.

### 3.2. Implementation Details

Computer configuration in the experiment comprised two NVIDIA GeForce RTX 3080Ti GPUs, 128 Gb of memory, Ubuntu 20.04 operating system, Cuda10.0, and Cudnn7.5. Our developed classification approach to generating the highly discriminative features is constructed using Pytorch library [53] and trained in an end-to-end manner. We use Adam for the optimizer with betas $= (0.9, 0.999)$ and weight decay $= 1 \times 10^{-6}$. We train the developed approach with a batch size of 64, an initial learning rate of $10^{-4}$, and maintain a more optimal learning rate throughout the training procedure using a cosine annealing scheduler. We use a dropout with $p = 0.1$ for regularization. The number of heads in the multi-head attention modules is 8. We use a $L = 3$ layer transformer with a residual layer around each embedding update and layer normalization.

With regards to our laboratory sample dataset, we imaged the 30 rocks and minerals in it. To compare the spectra of rocks and minerals as well as to reduce the spectral noises, the image patches with fixed-size coverage are obtained. For example, in our implementation, the image patch is a group of pixels $9 \times 9 \times 8$ if its size is set to $9 \times 9$, which will serve as the input for the developed transformer-based classification approach in Section 3.3. The effect of the size of image patch on the classification performance will be discussed in Section 4.1. In our implementation, we randomly selected some of the image patches as training, and then used the rest as validation and test to classify the rocks and minerals. The ratio of training, validation, and test split is 60%, 10%, and 30%, respectively. We count out the number of image patches in the training set, validation set, and test set; Table 2 lists the detailed statistics.
Table 2. The statistics of image patches in training set, validation set and test set. For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galenite (Gal), talcumin (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smallite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorspar (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr).

| Category | Mu  | He  | Ca  | Gal | Ka  | Ta  | Pyr | Cha | GT  | St  | Gab |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Training set | 3011 | 5508 | 3149 | 4019 | 3667 | 3486 | 3360 | 3894 | 3824 | 4272 | 6044 |
| Validation set | 502  | 918  | 525  | 670  | 612  | 581  | 560  | 649  | 639  | 712  | 1008 |
| Test set    | 1506 | 2754 | 2575 | 2009 | 1834 | 1743 | 1680 | 1947 | 1917 | 2136 | 3022 |

| Category | Qu  | Chl | Se  | Sm  | Te  | Gy  | Gr  | Cr  | Sua | Ar  | Bas |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Training set | 4350 | 4882 | 3131 | 3686 | 5186 | 3378 | 3311 | 4043 | 3272 | 3900 | 4100 |
| Validation set | 725  | 814  | 522  | 614  | 864  | 563  | 552  | 674  | 545  | 650  | 683  |
| Test set    | 2175 | 2441 | 1566 | 1843 | 2593 | 1689 | 1655 | 2022 | 1637 | 1950 | 2050 |

| Category | Fl  | Ps  | Sa  | Go  | Bar | Sul | BV  | Pr  | Sum | In Total |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|
| Training set | 3840 | 3378 | 5280 | 4572 | 4050 | 4356 | 6038 | 4708 | 128,045 |
| Validation set | 640  | 563  | 880  | 762  | 675  | 726  | 1006 | 785  | 21,344 |
| Test set    | 1920 | 1689 | 2640 | 2286 | 2025 | 2178 | 3020 | 2355 | 65,032 |

3.3. Rock Classification Results

The classification performance relies on the descriptive ability of the generated feature representation of rocks and minerals to some extent. In our implementation, we conduct the classification task to distinguish 30 types of rock and mineral samples. Table 3 lists rock classification results generated by the developed method in precision, recall, and $F_1$-score per class, as well as overall accuracy. From the quantitative perspective, we can observe that the developed method can effectively distinguish 30 types of rock and minerals samples, with a high overall accuracy of 96.92% and an average $F_1$-score of 97.11%. It is worth noting that some types of rocks and minerals, such as muscovite, talcumin, pyrite and so on, even achieve 100% in their $F_1$-score. Experimental results suggest that the feature representation derived from the developed method in this paper is highly discriminative for carrying out the classification task, as summarized in Table 3. That is to say, the generated feature representation shows the capability to reliably and robustly describe the differences among different types of rocks and minerals from the quantitative perspective.

Inevitably, there are some cases of incorrect classification. From the classification results derived from the trained optimal model, it is observed that the sulfur is easily misclassified into crystal, which results in the low recall. Furthermore, the low precision and low recall which occur between chlorite and saponite is primarily due to the mixed classification. For the visualization comparison and analysis (as shown in Figure 4), there exists largely-overlapping regions between sulfur and crystal, as well as between chlorite and saponite in raw spectral signature signatures, which illustrates the high similarity between them. This is one of most important factors responsible for low precision or recall in the classification results.
Table 3. Classification results generated by the developed method in precision, recall and \( F_1 \)-score per class as well as overall accuracy (the unit: %). For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smaltite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), saponite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr).

| Category | Mu | He | Ca | Gal | Ka | Ta | Pyr | Cha | GT | St | Gab |
|----------|----|----|----|-----|----|----|-----|-----|----|----|-----|
| Precision | 100.0 | 96.62 | 97.07 | 93.84 | 99.89 | 100.0 | 100.0 | 100.0 | 100.0 | 92.54 | 99.93 |
| Recall | 100.0 | 99.93 | 98.98 | 97.95% | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 99.70 |
| F1-score | 100.0 | 98.25 | 98.01 | 95.86 | 99.94 | 100.0 | 100.0 | 100.0 | 100.0 | 96.13 | 99.81 |

| Category | Qu | Chl | Se | Sm | Te | Gy | Gr | Cr | Sua | Ar | Bas |
|----------|----|----|----|----|----|-----|-----|----|-----|----|-----|
| Precision | 99.95 | 89.58 | 99.23 | 100.0 | 95.22 | 100.0 | 100.0 | 100.0 | 100.0 | 99.51 | 100.0 |
| Recall | 100.0 | 82.09 | 99.93 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| F1-score | 99.97 | 85.67 | 99.58 | 100.0 | 97.55 | 100.0 | 100.0 | 100.0 | 99.76 | 100.0 | 100.0 |

| Category | Fl | Ps | Sa | Go | Bar | Sul | BV | Pr | Overall accuracy |
|----------|----|----|----|-----|-----|-----|----|----|------------------|
| Precision | 100.0 | 99.12 | 96.35 | 100.0 | 100.0 | 99.29 | 93.34 | 88.28 | 96.92% |
| Recall | 100.0 | 100.0 | 100.0 | 99.65 | 100.0 | 58.40 | 85.03 | 98.59 | 100.0 |
| F1-score | 100.0 | 99.55 | 98.14 | 99.82 | 100.0 | 73.54 | 88.99 | 93.16 | 100.0 |

**Figure 4.** Spectral analysis of some wrong classification examples. The bars are rendered in different colors according to different categories. (a) Crystal–sulfur. (b) Chlorite–saponite.

### 3.4. T-SNE Visualization in the Discriminative Feature Space

In this work, the key insight behind the developed method is to make a remarkable distinction among different types of rocks and minerals in their high-dimensional feature space. The T-distributed stochastic neighbor embedding (T-SNE) technique [54] is a dimensionality reduction technique, whose goal is to project high-dimensional data into a two-dimensional map space. In the two-dimensional map space, the same categories tend to be modeled by nearby locations and the dissimilar categories by distant locations with high probability. That is to say, the feature representation is highly discriminative if the clusters belonging to the same categories are formed well. For a more detailed description of the T-SNE technique, please refer to Appendix A. To qualitatively analyze the performance of the highly discriminative feature representation derived from the transformer-based classification method, in this subsection, we visualize the quality of formed clusters in their
high-dimensional feature space using T-SNE visualization. Theoretically, the number of formed clusters within the two-dimensional visualization maps should be associated with the number of categories in the rock and mineral dataset. In our implementation, the output of the developed method, i.e., the last fully-connected layer, is considered as the derived feature representation describing both rocks and minerals and serves as the input of the scikit-learn T-SNE package [55] for visualization. For comparison, we project the original spectral signatures and the feature representation derived from the developed method into the two-dimensional map space for visualization. Figure 5 demonstrates a visualization comparison between the original spectral signatures and the feature representation derived from the developed method using the T-SNE technique. As shown in Figure 5a, there occur severe overlapping regions among different types of rocks in the two-dimensional visualization maps. It can be concluded that inter-category spectral similarity among different types of rocks and minerals might appear, which would result in the degradation of the classification performance. Compared with those from the original spectral signatures, as shown in Figure 5b, the clusters are formed well from the derived feature representation for different type of both rocks and minerals, which suggests the highly discriminative representation capability.

Figure 5. Visualization comparison using the T-SNE technique. (a) Original spectral signatures; (b) highly discriminative features derived from the developed method.

4. Discussion

In this section, the parameter sensitivity of the developed classification approach is provided, primarily including the size of image patches, the number of transformer layers, the number of transformer heads, and the integration of category-aware contrastive loss. Additionally, we compare the proposed method with other methods. Afterwards, both the strengths and limitations of the developed rock classification method are briefly analyzed and discussed.

4.1. Effect of the Size of Image Patches on Classification Results

To compare the spectra of rocks and minerals for automatic classification, in our implementation, spectra were extracted from the image patches with fixed-size coverage. In this subsection, we set different sizes of image patches, including $5 \times 5$, $7 \times 7$, $9 \times 9$, and $11 \times 11$, to discuss the effect of the size of image patch on the classification results. Table 4 lists the effect of the size of image patches on classification results, which demonstrate the fluctuation of classification results from 95.96% to 96.92%. As mentioned in Section 2.1, small changes in the compositions of rocks and minerals cause shifts in the position and shape of absorption bands in the spectrum. With the increasing size, the image patches
with fixed-size coverage alleviate the effect of composition variants, which enhance the classification performance to some extent, although it brings computing burdens. Taking the accuracy and computational cost into consideration, we set the image patch size to $9 \times 9$ as the optimal parameters in our implementation.

Table 4. Effect of the size of image patches on classification results in $F_1$-score per class and overall accuracy (the unit: %). For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smaltite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr). The best performance is highlighted in **BOLD**.

| Patch Size | Mu  | He  | Ca  | Gal | Ka  | Ta  | Pyr | Cha | GT  | St  | Gab |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Size = 5   | 95.18 | 95.86 | 92.49 | 94.19 | 98.66 | 99.97 | 100.0 | 99.80 | 99.44 | 97.45 | 98.33 |
| Size = 7   | 99.94 | 97.57 | 92.23 | 95.21 | 99.65 | 100.0 | 100.0 | 100.0 | 99.93 | 97.04 | 99.31 |
| Size = 9   | 100.0 | **98.25** | 98.01 | 95.86 | **99.94** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** |
| Size = 11  | 100.0 | 97.97 | 87.88 | **97.47** | 99.88 | 100.0 | 100.0 | 99.95 | 99.73 | **98.15** | 97.77 |

| Patch size | Qu  | Chl | Se  | Sm  | Te  | Gy  | Gr  | Cr  | Sua | Ar  | Bas |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Size = 5   | 99.58 | 81.66 | 98.54 | 99.66 | 91.09 | 99.91 | 99.34 | 92.01 | 99.54 | 99.83 | 99.77 |
| Size = 7   | 99.56 | 84.49 | 99.56 | 100.0 | 96.81 | 99.97 | 99.27 | 89.97 | 99.70 | 99.97 | 99.97 |
| Size = 9   | **99.97** | 85.67 | **99.58** | **100.0** | **97.55** | **100.0** | **100.0** | **89.82** | **99.76** | **100.0** | **100.0** |
| Size = 11  | 99.33 | **86.88** | 96.55 | 99.93 | 96.99 | 100.0 | 100.0 | **99.21** | 100.0 | 99.96 | 99.94 |

| Patch size | Fl  | Ps  | Sa  | Go  | Bar | Sul | BV  | Pr  | Overall accuracy |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|------------------|
| Size = 5   | 100.0 | 98.88 | 97.17 | 99.32 | 99.98 | 67.59 | 85.33 | 93.13 | 95.96 |
| Size = 7   | 100.0 | 99.18 | 96.87 | 99.74 | 100.0 | 67.27 | **89.58** | 92.51 | 96.48 |
| Size = 9   | 100.0 | 99.55 | **98.14** | 99.82 | **100.0** | **73.54** | 88.99 | **93.16** | **96.92** |
| Size = 11  | 100.0 | **99.93** | 96.36 | **100.0** | 100.0 | 68.77 | 89.33 | 87.46 | 96.83 |

4.2. Effect of the Number of Transformer Layers on Classification Results

In our implementation, we extract the highly discriminative feature representation from image patches encoded by a series of transformer layers. To evaluate the effect of the number of transformer layers, we set it with steps of 1 from 1 to 7. As with deeper convolutional neural networks, the developed method is capable of perfectly fitting training data, and also performed well on test data as the number of transformer layers increased. Figure 6 demonstrates the experimental results. When the number of Transformer layers rises from 1 to 3, the overall accuracy fluctuates by approximately 0.77%. However, the increase in the number of transformer layers also makes the model convergence very difficult due to the gradient vanishing, which might degrade the classification performance when the number of transformer layers increases from 3 to 7.
4.3. Effect of the Number of Transformer Heads on Classification Results

In addition to the number of transformer layers, the developed transformer-based classification method aggregates the inter-channel and intra-channel information through the multi-head self-attention mechanism. Thus, the accuracy might depend on the number of transformer heads for the developed transformer-based classification method. To demonstrate the effect of the number of transformer heads on classification results, we set it to 1, 2, 4, 8, respectively, since the number of multi-spectral bands must be guaranteed to be divisible. In this way, the multi-spectral bands are split across the multiple attention heads so that each can process them independently, while the correlations are extracted and aggregated for each attention head. From the experimental results shown in Figure 7, we can conclude that the accuracy of classification is easily affected by the number of transformer heads, with difference in the overall accuracy of more than 1.01%. This enables the transformer-based feature encoder module to capture richer interpretations of the multi-spectral bands.

Figure 6. Effect of the number of transformer layers on classification results.

Figure 7. Effect of the number of transformer heads on classification results.
4.4. Effect of Category-Aware Contrastive Loss on Classification Results

To generate the highly discriminative feature representation, we combine the cross-entropy loss with the category-aware contrastive loss to optimize a multi-task loss. In this way, we attempt to force instances of the same category to remain close-by while pushing instances of dissimilar category far apart by introducing the category-aware contrastive loss, as shown in Figure 3. Table 5 compares the results between no contrastive loss and the developed method in $F_1$-score per class, and the overall accuracy. From the comparative results, an improvement of approximately 0.72% can be observed, which suggests that the generated feature representation becomes more discriminative than that without the category-aware contrastive loss. Hence, we can draw a conclusion that the category-aware contrastive loss is beneficial to enhancing the descriptive capability of the generated feature representation.

Table 5. Effect of category-aware contrastive loss on classification results in $F_1$-score per class (%). For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcum (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smaltite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluoride (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr). The best performance is highlighted in BOLD.

| Methods          | Mu  | He  | Ca  | Gal | Ka  | Ta  | Pyr | Cha | GT  | St  | Gab |
|------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| No contrastive loss | 100.0 | 97.57 | 91.21 | 95.87 | 98.84 | 100.0 | 100.0 | 100.0 | 100.0 | 98.84 | 100.0 |
| Proposed method   | 100.0 | 98.25 | 98.01 | 95.86 | 99.94 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 99.84 |

| Methods          | Qu  | Chl | Se  | Sm  | Te  | Gy  | Gr  | Cr  | Sua | Ar  | Bas |
|------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| No contrastive loss | 99.97 | 80.76 | 99.45 | 100.0 | 95.85 | 100.0 | 100.0 | 93.03 | 99.87 | 99.97 | 99.97 |
| Proposed method   | 99.97 | 85.67 | 99.58 | 100.0 | 97.55 | 100.0 | 100.0 | 89.82 | 99.76 | 100.0 | 100.0 |

| Methods          | Fl  | Ps  | Sa  | Go  | Bar | Sul | BV  | Pr  | Average | Overall accuracy |
|------------------|-----|-----|-----|-----|-----|-----|-----|-----|----------|------------------|
| No contrastive loss | 100.0 | 99.97 | 94.70 | 99.91 | 100.0 | 68.77 | 89.14 | 91.49 | 96.40 | 96.25 |
| Proposed method   | 100.0 | 99.55 | 98.14 | 99.82 | 100.0 | 73.54 | 89.99 | 93.16 | 97.12 | 96.92 |

4.5. Comparisons with Other Methods

To further evaluate the descriptive performance of the developed method, we also compare it with other methods which implement the frequently-used classifiers, such as the decision tree [56], random forest [57], and support vector machine (SVM) [58], based on the raw spectral signatures of rock samples. Furthermore, we compare our results with other neural networks [59], namely ConvNet, in this work. In our implementation, only one simple convolutional neural network is selected and used because the small image patches in this work are the input of the developed method which cannot be conducted on more complex architectures with many pooling operations. Table 6 lists the comparisons of the classification results between different methods. For the experimental results, we can observe that the overall accuracy and average $F_1$-score derived from the developed method exceeds those of other methods, with average differences of 4.75% and 4.97%, respectively. Moreover, although the developed method achieves a lower $F_1$-score for crystal than other methods, the developed method remarkably improves the classification performance, especially for identifying hematite, chalcopyrite, chlorite, serpentine, smaltite, saponite, etc., which further confirms the superior discriminative ability of the produced feature representation.
Table 6. Comparisons with other methods in $F_1$-score per class and overall accuracy (%). For abbreviation, muscovite (Mu), hematite (He), calcite (Ca), galena (Gal), kaolinite (Ka), talcumin (Ta), pyrite (Pyr), chalcopyrite (Cha), gray tiemannite (GT), stibnite (St), gabbro (Gab), quartz (Qu), chlorite (Chl), serpentine (Se), smaltite (Sm), tennantite (Te), gypsum (Gy), graphite (Gr), crystal (Cr), suanite (Sua), aragonite (Ar), basalt (Bas), fluorite (Fl), psilomelane (Ps), saponite (Sa), goethite (Go), barite (Bar), sulfur (Sul), biotite vermiculite (BV), and pyroxenite (Pr). The best performance is highlighted in BOLD fonts.

| Methods              | Mu  | He  | Ca  | Gal | Ka  | Ta  | Pyr | Cha | GT  | St  | Gab |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Decision tree [56]   | 95.98 | 85.82 | 90.06 | 90.86 | 99.20 | 95.20 | 99.97 | 87.30 | 97.62 | 95.06 | 86.29 |
| Random forest [57]   | 97.98 | 88.83 | 95.60 | 91.39 | 99.91 | 98.44 | 100.0 | 88.82 | 99.47 | 94.72 | 90.52 |
| SVM [58]             | 99.70 | 88.44 | 97.99 | 95.68 | 98.86 | 98.78 | 99.97 | 88.62 | 99.63 | 96.09 | 96.17 |
| ConvNet [59]         | 99.33 | 91.66 | 93.26 | 94.96 | 97.39 | 99.45 | 100.0 | 87.03 | 99.76 | 97.62 | 98.10 |
| Developed method     | 100.0 | 98.25 | 98.01 | 95.86 | 99.94 | 100.0 | 100.0 | 100.0 | 100.0 | 96.13 | 99.81 |

| Methods              | Qu  | Chl | Se  | Sm  | Te  | Gy  | Gr  | Cr  | Su  | Ar  | Bas |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Decision tree [56]   | 99.15 | 57.40 | 62.62 | 85.76 | 90.48 | 99.20 | 97.58 | 95.95 | 87.53 | 96.26 | 92.66 |
| Random forest [57]   | 99.38 | 65.13 | 70.33 | 82.94 | 91.69 | 100.0 | 99.16 | 98.92 | 93.77 | 96.27 | 96.28 |
| SVM [58]             | 99.70 | 59.31 | 77.72 | 72.96 | 83.12 | 100.0 | 97.66 | 97.84 | 93.84 | 99.82 | 98.77 |
| ConvNet [59]         | 100.0 | 74.56 | 76.19 | 99.53 | 96.26 | 100.0 | 99.84 | 98.38 | 99.31 | 99.58 | 99.72 |
| Developed method     | 99.97 | 85.67 | 99.58 | 100.0 | 97.55 | 100.0 | 100.0 | 98.82 | 100.0 | 99.76 | 100.0 |

| Methods              | Fl  | Ps  | Sa  | Go  | Bar | Sul | BV  | Pr  | Average | Overall accuracy |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|---------|------------------|
| Decision tree [56]   | 98.23 | 96.93 | 86.79 | 97.42 | 98.99 | 69.38 | 80.17 | 91.37 | 90.24 | 90.22 |
| Random forest [57]   | 99.89 | 98.39 | 91.67 | 97.62 | 99.38 | 71.56 | 83.98 | 94.00 | 92.53 | 92.41 |
| SVM                  | 99.94 | 95.74 | 87.25 | 98.55 | 100.0 | 30.72 | 77.18 | 90.77 | 90.79 | 90.79 |
| ConvNet [59]         | 99.03 | 99.59 | 95.29 | 98.55 | 98.15 | 91.88 | 79.34 | 98.32 | 95.04 | 95.24 |
| Developed method     | 100.0 | 99.55 | 98.14 | 99.82 | 100.0 | 73.54 | 88.99 | 93.16 | 97.12 | 96.92 |

5. Summary and Outlook

To address the challenges due to the inter-category spectral similarity, we present a category recognition approach of both rocks and minerals using the discriminative representation derived from its spectral signatures. The developed method combines a transformer-based recognition approach with category-aware contrastive learning, which is trained in an end-to-end multi-task manner. The advantages of our proposed approach are as follows: (1) Different from convolutional neural network, the transformer shows its strong ability to model long-range dependencies. (2) We defined a category-aware contrastive loss, which would force instances of the same category to remain close-by while pushing instances of the dissimilar category far apart. Therefore, the derived highly discriminative feature representation from the developed approach is beneficial to enhancing the descriptive capability and alleviating the inter-category spectral similarity for the classification task. Furthermore, we establish a rock sample database with 30 types of rocks and then carry out the experimental analysis both quantitatively and qualitatively. Experimental results confirm the robustness and reliability of the developed method. It can be concluded that the developed method can effectively distinguish 30 types of rock samples, with a high overall accuracy of 96.92%. Additionally, the overall accuracy and average $F_1$-score derived from the developed method exceed those of other common methods, with average differences of 5.78% and 5.93%, respectively. It is well-known that the more rock categories, the more severe the inter-category spectral similarity. The establishment of rock databases with more types of rock samples will enable us to test the generalization of the developed method in our future work. Additionally, for the developed method in this work, the category-aware contrastive loss and cross-entropy loss is aggregated, and the weight is set empirically. In our future research, an adaptive weighting solution is required for the performance stability of the method.
Author Contributions: Methodology, J.Y. (Juntao Yang) and Z.K.; validation, Z.Y., B.X., J.Y. (Jianfeng Yang) and J.T.; visualization, J.X.; formal analysis, J.Y. (Juntao Yang) and Z.K.; supervision, J.Y. (Juntao Yang) and Z.K.; writing—original draft, J.Y. (Juntao Yang) and Z.Y.; writing—review and editing, J.Y. (Juntao Yang) and Z.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly supported by the National Key Research and Development Program of China (No. 2019YFE0123300), National Natural Science Foundation of China (No. 41872207), Civil Aerospace Technology Advance Research Project of National Defense Science and Engineering (No. D020103), Beijing Science and Technology Project (No. Z19110004319001), the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 871149, and Introduction plan of high-end foreign experts (No. G2021025006L).

Acknowledgments: We also would like to thank the National Mineral Rock and Fossil Specimens Resource Center (http://www.nimrf.net.cn) (accessed on 25 August 2022) for the share and free use of their rock samples. The data that support the findings of this work are available from the corresponding author upon the reasonable request.

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

Appendix A. T-Distributed Stochastic Neighbor Embedding Visualization Technique

The T-distributed stochastic neighbor embedding (T-SNE) technique [54] is a non-linear dimensionality reduction technique for projecting high-dimensional data onto a two-dimensional map space for visualization, where the same categories tend to be modeled by nearby locations and the dissimilar categories by distant locations with high probability. The T-SNE technique consists of two stages, namely the construction of a probability distribution over pairs of high-dimensional data, and the definition of a similar probability distribution over the data in the two-dimensional map space.

Given a set of \( N \) high-dimensional data \( x_1, \ldots, x_N \), the probabilities \( p_{ij} \) that are proportional to the similarity of data \( x_i \) and \( x_j \) are calculated as defined in the following Equation (A1):

\[
p_{ij} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2 / 2\sigma_i^2)}
\]  

(A1)

where \( \sigma_i \) denotes the bandwidth of the Gaussian kernels set using a predefined perplexity. In our implementation, \( p_{ii} \) is set to 0, and \( \sum_{i} p_{ij} = 1 \) for all \( i \). Define \( p_{ij} = \frac{(p_{ij} + p_{ji})}{2N} \), then \( p_{ij} = p_{ji} \), \( p_{ii} = 0 \), and \( \sum_{ij} p_{ij} = 1 \).

Following this, a two-dimensional map \( y_1, \ldots, y_N \), is learned to represent the similarities as well as possible. The similarities \( q_{ij} \) between \( y_i \) and \( y_j \) is defined as follows:

\[
q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_k \sum_{l \neq k} (1+||y_k - y_l||^2)^{-1}}
\]  

(A2)

The locations of the point \( y_i \) in the two-dimensional map is determined by minimizing the Kullback–Leibler (KL) divergence between the distribution \( P \) and \( Q \) for the visualization.

References

1. Blake, D.; Bristow, T.; Sarrazin, P.; Zacny, K. In-Situ Mineralogical Analysis of the Venus Surface using X-ray Diffraction. Bull. Am. Astron. Soc. 2021, 53, 018. [CrossRef]
2. Yang, J.; Kang, Z. A Gradient-Region Constrained Level Set Method for Autonomous Rock Detection from Mars Rover Image. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Enschede, The Netherlands, 10–14 June 2019.
3. Christensen, P.R.; Bandfield, J.; Clark, R.N.; Edgett, K.; Hamilton, V.E.; Hoefen, T.; Kieffer, H.H.; Kuzmin, R.O.; Lane, M.D.; Malin, M.C.; et al. Detection of crystalline hematite mineralization on Mars by the Thermal Emission Spectrometer: Evidence for near-surface water. J. Geophys. Res. Earth Surf. 2000, 105, 9623–9642. [CrossRef]
4. Kodikara, G.R.; McHenry, L.J. Machine learning approaches for classifying lunar soils. Icarus 2020, 345, 113719. [CrossRef]
5. Alemanno, G.; Maturilli, A.; D’Amore, M.; Helbert, J. A new laboratory emissivity and reflectance spectral library for the interpretation of Mars thermal infrared spectral data. Icarus 2021, 368, 114622. [CrossRef]
6. Anderson, R.C.; Jandura, L.; Okon, A.B.; Sunshine, D.; Roumeliotis, C.; Beegle, L.; A. Hurowitz, J.; Kennedy, B.P.; Limonadi, D.; McCloskey, S.; et al. Collecting Samples in Gale Crater, Mars; an Overview of the Mars Science Laboratory Sample Acquisition, Sample Processing and Handling System. *Space Sci. Rev.* 2012, 170, 57–75. [CrossRef]

7. Caudill, C.M.; Pontefract, A.J.; Osinski, G.R.; Tornabene, L.L.; Pilles, E.A.; Battler, M.; Francis, R.; Godin, E.; Gallofre, A.G.; Haltigin, T.; et al. CanMars mission Science Team operational results: Implications for operations and the sample selection process for Mars Sample Return (MSR). *Planet. Space Sci.* 2019, 172, 43–56. [CrossRef]

8. Osinski, G.R.; Battler, M.; Caudill, C.M.; Francis, R.; Haltigin, T.; Hipkin, V.J.; Kerrigan, M.; Pilles, E.A.; Pontefract, A.; Tornabene, L.L.; et al. The CanMars Mars Sample Return analogue mission. *Planet. Space Sci.* 2019, 166, 110–130. [CrossRef]

9. Bishop, J.L.; Bell, J.F.; Io Moersch, J.E. Remote Compositional Analysis: Techniques for Understanding Spectroscopy, Mineralogy, and Geochemistry of Planetary Surfaces; Cambridge U. Press: Cambridge, UK, 2020.

10. Viviano, C.E.; Seelos, F.P.; Murchie, S.L.; Kahn, E.G.; Seelos, K.D.; Taylor, H.W.; Taylor, K.; Ehlmann, B.L.; Wiseman, S.M.; Mustard, J.F.; et al. Revised CRISM spectral parameters and summary products based on the currently detected mineral diversity on Mars. *J. Geophys. Res.-Planet* 2014, 119, 1403–1431. [CrossRef]

11. Jain, N.; Chauhan, P. Study of phyllosilicates and carbonates from the Capri Chasma region of Valles Marineris on Mars based on Mars Reconnaissance Orbiter-Compact Reconnaissance Imaging Spectrometer for Mars (MRO-CRISM) observations. *Icarus* 2015, 250, 7–17. [CrossRef]

12. Fox, V.K.; Arvidson, R.E.; Guinness, E.A.; McLennan, S.M.; Catalano, J.G.; Murchie, S.L.; Powell, K.E. Smectite deposits in Marathon Valley, Endeavour Crater, Mars, identified using CRISM hyperspectral reflectance data. *Geophys. Res. Lett.* 2016, 43, 4885–4892. [CrossRef]

13. Núñez, J.L.; Barnounin, O.S.; Murchie, S.L.; Seelos, F.P.; McGovern, J.A.; Seelos, K.D.; Buczkowski, D.L. New insights into gully formation on Mars: Constraints from composition as seen by MRO/CRISM. *Geophys. Res. Lett.* 2016, 43, 8893–8902. [CrossRef]

14. Xue, Y.; Yang, Y.; Yu, L. Mineral composition of the Martian Gale and Nili Fossae regions from Mars Reconnaissance Orbiter CRISM images. *Planet. Space Sci.* 2018, 163, 97–105. [CrossRef]

15. Fox, V.K.; Arvidson, R.E.; Guinness, E.A.; McLennan, S.M.; Catalano, J.G.; Murchie, S.L.; Powell, K.E. Smectite deposits in Marathon Valley, Endeavour Crater, Mars, identified using CRISM hyperspectral reflectance data. *Geophys. Res. Lett.* 2016, 43, 4885–4892. [CrossRef]

16. Xue, Y.; Yang, Y.; Yu, L. Mineral composition of the Martian Gale and Nili Fossae regions from Mars Reconnaissance Orbiter CRISM images. *Planet. Space Sci.* 2018, 163, 97–105. [CrossRef]

17. Jain, N.; Giampiccolo, D.N.; Oru, J.P.M.; Gonçalves, W.N. Fractal dimension of maximum response filters applied to texture analysis. *Pattern Recognit. Lett.* 2015, 65, 116–123. [CrossRef]

18. Lin, H.; Mustard, J.F.; Zhang, X. A methodology for quantitative analysis of hydrated minerals on Mars with large endmember library using CRISM near-infrared data. *Planet. Space Sci.* 2016, 165, 124–136. [CrossRef]

19. Li, J.; Zhang, L.; Wu, Z.; Ling, Z.; Cao, X.; Guo, K.; Yan, F. Autonomous Martian rock image classification based on transfer deep learning methods. *Earth Sci. Inform.* 2020, 13, 951–963. [CrossRef]

20. De Albuquerque, M.P.; Esquivel, I.; Mello, A.G. Image thresholding using Tsallis entropy. *Pattern Recognit. Lett.* 2004, 25, 1059–1065. [CrossRef]

21. Sato, S.; Sano, M.; Sawada, Y. Practical methods of measuring the generalized dimension and the largest Lyapunov exponent in high dimensional chaotic systems. *Prog. Theor. Phys.* 1987, 77, 1–5. [CrossRef]

22. Patel, A.K.; Chatterjee, S. Computer vision-based limestone rock-type classification using probabilistic neural network. *Geosci. Front.* 2016, 7, 53–60. [CrossRef]

23. Shang, C.; Barnes, D. Support vector machine-based classification of rock texture images aided by efficient feature selection. In Proceedings of the 2012 International Conference on Neural Networks (IJCNN), Brisbane, QLD, Australia, 10–15 June 2012.

24. Valentine, M.B.; De Bom, C.R.; de Albuquerque, M.P.; Faria, E.L.; Correia, M.; Surmas, R. On a method for Rock Classification using Textural Features and Genetic Optimization. *arXiv* 2016, arXiv:1607.01679. [CrossRef]

25. Shu, L.; McIsaac, K.; Osinski, G.R.; Franci, R. Unsupervised feature learning for autonomous rock image classification. *Comput. Geosci.* 2017, 106, 10–17. [CrossRef]

26. 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]

27. Ishikawa, S.T.; Gullick, V.C. An automated mineral classifier using Raman spectra. *Comput. Geosci.* 2013, 54, 259–268. [CrossRef]

28. Sharif, H.; Balchem, M.; Samson, C.; Ellery, A. Autonomous rock classification using Bayesian image analysis for Rover-based planetary exploration. *Comput. Geosci.* 2015, 83, 153–167. [CrossRef]

29. Díaz, G.F.; Ortiz, J.M.; Silva, J.F.; Lobos, R.A.; Egaña, F. VarioSim-Based Descriptors for Comparison and Classification of Rock Texture Images. *Math. Geosci.* 2020, 52, 451–476. [CrossRef]

30. Parmar, N.; Vasswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 4055–4064.

31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weiss, B.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* 2020, arXiv:2010.11929. [CrossRef]

32. Pascual, A.D. Autonomous and Real Time Rock Image Classification using Convolutional Neural Networks. Electronic Thesis and Dissertation Repository. Master’s Thesis, The University of Western Ontario, London, ON, Canada, 2019. Available online: [https://ir.lib.uwo.ca/etd/6059](https://ir.lib.uwo.ca/etd/6059) (accessed on 25 July 2022).
33. Saranathan, A.M.; Parente, M. Adversarial feature learning for improved mineral mapping of CRISM data. *Icarus* **2021**, *355*, 114107. [CrossRef]

34. Grove, C.; Hook, S.; Paylor, I. Compilation of Laboratory Reflectance Spectra of 160 Minerals, 0.4 to 2.5 Micrometers. In *Jet Propulsion Laboratory*, NASA: Washington, DC, USA, 1992.

35. Meerdink, S.K.; Hook, S.J.; Roberts, D.A.; Abbott, E.A. The ECOSTRESS spectral library version 1.0. *Remote Sens. Environ.* **2019**, *230*, 111-112. [CrossRef]

36. Baldridge, A.M.; Hook, S.J.; Grove, C.I.; Rivera, G. The ASTER spectral library version 2.0. *Remote Sens. Environ.* **2009**, *113*, 711–715. [CrossRef]

37. Masoumi, F.; Eslamkish, T.; Abkar, A.A.; Honarmand, M.; Harris, J.R. Integration of spectral, thermal, and textural features of ASTER data using Random Forests classification for lithological mapping. *Remote Sens. Environ.* **2009**, *113*, 9609–9621. [CrossRef]

38. Xie, B.; Wu, L.; Mao, W.; Zhou, S.; Liu, S. An Open Integrated Rock Spectral Library (RockSL) for a Global Sharing and Matching Service. *Minerals* **2022**, *12*, 118. [CrossRef]

39. Christensen, P.R.; Bandfield, J.L.; Hamilton, V.E.; Clark, R.N. Identification of a basaltic component on the Martian surface from Thermal Emission Spectrometer data. *J. Geophys. Res. Earth Surf.* **2000**, *105*, 9609–9621. [CrossRef]

40. Michalski, J.R.; Niles, P.B.; Glotch, T.D.; Cuadros, J. Infrared Spectral Evidence for K-Metasomatism of Volcanic Rocks on Mars. *J. Geophys. Res. Lett.* **2021**, *48*, e2021GL093882. [CrossRef]

41. Mallapaty, S. What China’s Mars rover will do next. *Nature* **2021**, 593, 323–324. [CrossRef]

42. Li, R.; Di, K.; Howard, A.B.; Matthies, L.; Wang, J.; Agarwal, S. Rock modeling and matching for autonomous long-range Mars rover localization. *J. Field Robot.* **2007**, *24*, 187–203. [CrossRef]

43. Baykan, N.A.; Yılmaz, N. Mineral identification using color spaces and artificial neural networks. *Comput. Geosci.* **2010**, *36*, 91–97. [CrossRef]

44. Xiao, X.; Cui, H.; Yao, M.; Tian, Y. Autonomous rock detection on mars through region contrast. *Adv. Space Res.* **2017**, *60*, 626–635. [CrossRef]

45. Xiao, X.; Cui, H.; Yao, M.; Fu, Y.; Qi, W. Auto rock detection via sparse-based background modeling for mars rover. In Proceedings of the 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 8–13 July 2018.

46. Zhang, J.; Mei, K.; Zheng, Y.; Fan, J. Learning multi-layer coarse-to-fine representations for large-scale image classification. *Pattern Recognit.* **2019**, *91*, 175–189. [CrossRef]

47. Fukunaga, K.; Flick, T.E. Classification error for a very large number of classes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *6*, 779–788. [CrossRef]

48. Reid, R.J.; Smith, P.H.; Lemmon, M.; Tanner, R.; Burke, M.; Wegryn, E.; Weinberg, J.; Marcialis, R.; Britt, D.T.; Thomas, N.; et al. Imager for Mars Pathfinder (IMP) image calibration. *J. Geophys. Res.-Planet* **1999**, *104*, 8907–8925. [CrossRef]

49. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General Multi-label Image Classification with Transformers. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.

50. Joseph, K.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, virtual, 19–25 June 2021.

51. Leonard, J.; Kramer, M.A. Improvement of the backpropagation algorithm for training neural networks. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]

52. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]

53. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.

54. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

55. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Gramfort, A.; Grimbert, V.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

56. Zhang, J.; Li, J.; Hu, Y.; Zhou, J.Y. The identification method of igneous rock lithology based on data mining technology. *Adv. Space Res.* **2012**, *466*, 65–69. [CrossRef]

57. Masoumi, F.; Esfahandish, T.; Akbar, A.A.; Honarmand, M.; Harris, J.R. Integration of spectral, thermal, and textural features of ASTER data using Random Forests classification for lithological mapping. *J. Afr. Earth Sci.* **2017**, *129*, 445–457. [CrossRef]

58. Chatterjee, S. Vision-based rock-type classification of limestone using multi-class support vector machine. *Appl. Intell.* **2013**, *39*, 14–27. [CrossRef]

59. El-Sawy, A.; El-Bakry, H.; Loey, M. CNN for handwritten arabic digits recognition based on LeNet-5. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 24–26 October 2016; Springer: Cham, Switzerland, 2016; pp. 566–575.