This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Generalized Unsupervised Clustering of Hyperspectral Images of Geological Targets in the Near Infrared

Angela F. Gao, Brandon Rasmussen, Peter Kulits
Eva L. Scheller, Rebecca Greenberger, Bethany L. Ehlmann
Caltech
{afgao, brasmuss, kulits, eschelle, rgreenbe, behlmann}@caltech.edu

Abstract

The application of infrared hyperspectral imagery to geological problems is becoming more popular as data become more accessible and cost-effective. Clustering and classifying spectrally similar materials is often a first step in applications ranging from economic mineral exploration on Earth to planetary exploration on Mars. Semi-manual classification guided by expertly developed spectral parameters can be time consuming and biased, while supervised methods require abundant labeled data and can be difficult to generalize. Here we develop a fully unsupervised workflow for feature extraction and clustering informed by both expert spectral geologist input and quantitative metrics. Our pipeline uses a lightweight autoencoder followed by Gaussian mixture modeling to map the spectral diversity within any image. We validate the performance of our pipeline at submillimeter-scale with expert-labelled data from the Oman ophiolite drill core and evaluate performance at meters-scale with partially classified orbital data of Jezero Crater on Mars (the landing site for the Perseverance rover). We additionally examine the effects of various preprocessing techniques used in traditional analysis of hyperspectral imagery. This pipeline provides a fast and accurate clustering map of similar geological materials and consistently identifies and separates major mineral classes in both laboratory imagery and remote sensing imagery. We refer to our pipeline as "Generalized Pipeline for Spectroscopic Unsupervised clustering of Minerals (GyPSUM)."

1. Introduction

Unlike a three-channel RGB image, imaging spectroscopy, or hyperspectral imaging (HSI), typically measures radiance values for hundreds of narrow wavelength bands. Material composition and physical properties determine light absorption and scattering behavior [23], allowing HSI to be used for identification of materials through their spectra. Well-designed instruments can resolve individual absorption features, and images can be used to quantify material abundance and physical properties – tasks that are generally difficult with multispectral imaging, which measures radiance for a few, typically broad wavelength bands [9, 32, 28]. Geological HSI is used for natural-hazard risk assessment and mitigation, mineral and oil exploration and production, and Earth system modeling, among other applications [2]. With increased governmental, industrial, and academic interest in and access to this technology, advanced analysis techniques for the rapidly growing wealth of hyperspectral images are becoming more valuable [2].

Typical tasks for HSI analysis of geological targets include classification, segmentation, anomaly detection, and unmixing [2] and use instruments that target the visible to mid-infrared wavelengths (VIS-MIR, ∼400-20000 nm) for mineral, ice, and atmospheric gas identification [9]. Semi-manual investigation is still common in these tasks [4, 6, 22]. One common approach for classification by expert spectral geologists is to apply knowledge of likely geologic processes occurring in the study target to isolate important known absorptions and use simple algebraic operations ("spectral parameters") to map relative abundances of materials. Then, a system of thresholds and rules is used to classify pixels at a granularity determined by the goals of the study. Spectral parameters are also commonly used to guide basic template matching approaches of classification, such as spectral angle mapping or spectral feature fitting, which rely on extracted type spectra from the image or library spectra [11]. While many attempts to partially automate this interactive and often labor-intensive workflow have been made, the same three issues are typically left unresolved: (1) analysis is time-consuming; (2) human bias can lead to false positives or negatives; and (3) expert knowledge is needed to understand the interactions between non-unique absorptions across a large space of possible materials [11, 6]. Additionally, many geologists do not have the programming or machine learning experience required to develop more sophisticated approaches.

Supervised learning partially resolves the issues with semi-manual analysis by automating feature extraction [19]. However, creation of ground-truth classification maps is ex-

pensive and requires expert input, which depends on interpretation and can be biased. Due to these challenges, only a limited collection of high-quality, fully-labeled training images for geological targets are publicly available. Large spectral libraries of minerals have been developed [28, 32], but the full range of natural spectra is large [43] and libraries do not contain the spatial context that exists in HSI. There are infinitely many naturally-occurring infrared spectra because they are combinations of pure mineral spectra which are modified further by varying abundances and physical properties. Thus, large spectral variability and limited training data availability make generalizable supervised classification a difficult task in geological HSI analysis.

In this work, we develop a novel autoencoder-based (AE) feature extraction technique coupled with Gaussian mixture modeling (GMM) for fully unsupervised classification of single HSI images in the near-infrared (NIR, $\sim$ 1000-2500 nm) . In addition, we quantify the effects of traditional preprocessing methods on the pipeline. We employ both quantitative metrics and qualitative expert interpretation to determine algorithmic performance on two geological HSI datasets, which include a labelled laboratory image from the Oman Drilling Project [27] and satellite images from the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) [34], including Jezero Crater, where the Perseverance rover recently landed (Section 3).

2. Previous Works

Unsupervised clustering of HSI is challenging for three reasons: (1) Noise profiles vary widely depending on target, imaging conditions, instrumentation, and calibration, and can have distinct spatial or spectral structure [7, 29, 34]; (2) Performing dimensionality reduction of HSI data without losing important information is difficult [43]; (3) Many clustering approaches are computationally intensive and rely on distance metrics which become less meaningful as dimensionality increases [35, 47].

A commonly used unsupervised clustering algorithm is principal component analysis (PCA) and k-means clustering [39]. However, k-means favors equal sized clusters (typically a poor assumption for geological targets) and PCA works well in the case of well-separated, convex clusters in the original space, which is uncommon in hyperspectral data [3]. Deep Embedded Clustering (DEC) [46] and Sparse Manifold Clustering and Embedding (SMCE) [17] fail to consistently outperform PCA + k-means in HSI datasets [47]. Spectral-Spatial Diffusion Learning (DLSS) [35] is a state-of-the-art unsupervised learning algorithm for HSI [47], but does not scale well with respect to memory and is only applicable to small datasets [35, 47]. Mou et al. proposed a Conv–Deconv (convolutional/encoder - deconvolutional/decoder) network for unsupervised spectral-spatial feature learning, but it requires labels for fine-tuning [33]. Nalepa et al. [36] used a KL divergence based objective to learn parameters for clustering. However, this method does not outperform GMM-based methods on the Salinas Valley dataset, which is most similar to the Oman and CRISM datasets, and it is significantly slower than both k-means and GMM [36]. Infinite mixture of infinite GMM, I^{2}GMM, is an unsupervised method that has been shown to be successful on CRISM data [13, 30]. This method results in component features that are more difficult to interpret in comparison to AE extracted features [30].

Automated determination of the number of endmembers is crucial to successful HSI analysis. The unsupervised Hyperspectral Signal identification by minimum error (HySime) algorithm for endmember optimization infers the signal subspace in hyperspectral imagery [1], which yields comparable or superior results compared with Harsanyi-Farrand-Chang (HFC) [24] and Noise-Whitened HFC [8] eigen-based Neyman–Pearson detectors. HySURE [38] outperforms HySime for low SNR synthetic data settings but is less consistent on real datasets. Thus, we use HySime to determine the optimal number of clusters to produce.

To best address the three issues outlined above, we develop an AE-GMM methodology with optional post-processing (GMM+) with HySime optimization of endmembers. We directly compare the performance of this new methodology, which we call GyPSUM, to semi-manual workflows and the industry standard PCA and k-means methodology.

3. Background

3.1. Autoencoder

Autoencoders (AE) [44] are a feature-extraction technique that consists of an encoder network E_ϕ and decoder network D_θ parameterized by ϕ and θ respectively. The encoder E_ϕ maps inputs $\mathbf{x} \in \mathbb{R}^N$ to a latent representation $\mathbf{z} \in \mathbb{R}^D$. For a single hidden layer encoder, $\mathbf{z}$ is given by

$$\mathbf{z} = \mathbf{W_2}\sigma(\mathbf{W_1}\mathbf{x} + \boldsymbol{\beta}_1) + \boldsymbol{\beta}_2 \tag{1}$$

where $\mathbf{W}_1$ and $\boldsymbol{\beta}_1$ are the weights and the bias vector of the hidden layer, $\mathbf{W}_2$ and $\boldsymbol{\beta}_2$ are the weights and the bias vector of the output layer, and σ represents the nonlinear activation function. The encoded representation $\mathbf{z}$ is used to produce a reconstruction $\hat{\mathbf{x}}$ through the decoder D_θ. For a single hidden layer decoder, $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = \mathbf{V_2}\sigma(\mathbf{V_1}\mathbf{z} + \boldsymbol{\gamma}_1) + \boldsymbol{\gamma}_2 \tag{2}$$

where $\mathbf{V}_1$ and $\boldsymbol{\gamma}_1$ are the weights and bias vector of the hidden layer, $\mathbf{V}_2$ and $\boldsymbol{\gamma}_2$ are the weights and bias vector of the output layer, and σ represents the nonlinear activation function. Typically, the loss function is a reconstruction loss $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$.

3.2. Expectation Maximization Clustering Models

Expectation maximization (EM) [3] methods calculate the maximum likelihood estimate (MLE) parameters $\boldsymbol{\theta}$ (independent of the decoder D_θ parameters) of a statistical model. This model depends on unobserved latent variables $\mathbf{Z}$ for a dataset $\mathbf{X}$ and likelihood function $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$. In particular, EM clustering algorithms solve for the optimal clustering given a number of clusters and cluster probability distribution.

Maximizing the marginal likelihood function $L(\boldsymbol{\theta}|\mathbf{X}) = \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$ is often intractable. Thus, EM algorithms iteratively solve the marginal likelihood. These steps t are repeated until convergence:
(1) Defining

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]. \tag{3}$$

(2) Improved estimates of $\boldsymbol{\theta}$ are computed through

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}^{(t)}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{4}$$

GMM [3] is an example of an EM clustering algorithm with a Gaussian distribution for the cluster probability model. Unlike GMM, k-means does not optimize a probabilistic model. This biases k-means towards equal sized clusters.

3.3. Quantitative Metrics

We use multiple unsupervised cluster-separation metrics for evaluation. The Calinski-Harabasz (CH) index [5] for data E with n_E pixels and k clusters is

$$s = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n_E - k}{k - 1} \tag{5}$$

where

$$W_k = \sum_{q=1}^{k} \sum_{\mathbf{x} \in C_q} (\mathbf{x} - \mathbf{c}_q)(\mathbf{x} - \mathbf{c}_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (\mathbf{c}_q - \mathbf{c}_E)(\mathbf{c}_q - \mathbf{c}_E)^T$$

with C_q as the set of points in cluster q, $\mathbf{c}_q$ the center of cluster q, $\mathbf{c}_E$ the center of data E, and n_q the number of points in cluster q. CH scores are higher when clusters are dense and well-separated but penalize non-convex clusters.

The Davies-Bouldin (DB) index [12] is defined by

$$b = \frac{1}{k} \sum_{i=1}^{k} \max_j R_{ij} \tag{6}$$

where

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

is a cluster similarity measure between clusters i and j. s_i is the cluster diameter and d_{ij} is the distance between the centroids of clusters i and j. This score is lower when clusters are dense and well-separated, unlike the CH index.

3.4. Oman Dataset

Ocean crust is typically technologically challenging to drill and investigate in situ. However, the Oman ophiolite was tectonically thrust on top of stable continental crust [40, 20] and it is the subject of the International Continental Scientific Drilling Program's Oman Drilling Project [27]. Our first dataset consists of a few images selected from $>$ 31 TB of drill core images ($>$ 3.2 km of imaged slices of $\sim$ 5 cm diameter extracted rock core). Images were collected in-lab with Caltech's custom imaging spectrometer with a spectral resolution of $\sim$6 nm and a spatial resolution of $\sim$0.25 mm/pixel at wavelengths from 900 to 2600 nm. These high SNR images have high spectral variance compared with typical imagery, but contain rare, low spatial occurrence mixtures that are hard to isolate.

3.5. CRISM Dataset

CRISM is a push-broom visible/infrared imaging spectrometer aboard the Mars Reconnaissance Orbiter that has been orbiting Mars since 2006. It has collected hundreds of thousands of images of the planet's surface in varying image modes ranging from coarse multispectral imagery to high spatial resolution targeted imagery at 18-40 m/pixel at spectral sampling of $\sim$6.5 nm across the wavelength range of 362-3920 nm [34]. HRL000040FF ($\sim$40 m/pixel) is an image of the Jezero Crater rim and an ancient crater lake delta that is the main target of exploration for the recently landed Perseverance rover [18, 21]. We also include FRT0000634B[1] ($\sim$18 m/pixel), an image from the Claritas Rise that shows evidence of hydrothermal alteration [15]. Challenges to clustering include systematic cross-track dependent noise, pixels that always represent only mixtures due to coarse spatial resolution, subdued absorptions indicative of minerals, and atmospheric residual absorptions which overlap with mineralogically important absorption features.

4. GyPSUM Method

Our pipeline consists of preprocessing detailed in Algorithm 1, feature extraction and clustering detailed in Algorithm 2, and an optional post-processing step.

4.1. Preprocessing

We implement two similar preprocessing workflows in the case of laboratory imaging versus orbital CRISM imagery detailed in Algorithm 1. We begin with reflectance data and clip values between 0 and 1 to remove nonphysical outlier spikes before trimming the data in the spectral dimension to between 1050 and 2550 nm; this spectral subset contains the absoprtion features of greatest interest for our studies. We normalize each spectrum with the per-pixel

[1] included in the supplementary materials

ℓ_2 norm and reshape the data to a flat vector of pixels. We then apply a mask created from previous work to remove background material. While not necessary for generalized implementation, we find that this increases the total variance captured by a fixed dimension of latent components. Finally, we optionally divide out a linear convex hull computed for each spectrum, a technique (continuum removal, CR) commonly used to visually interpret subtle absorption features [10].

Algorithm 1 Preprocessing

Data: $\mathbf{X}$ HSI cube n × m × w, optional mask $\mathbf{M}$
Result: $\hat{\mathbf{X}}$ preprocessed vectorized image p × w
$\mathbf{X} \leftarrow$ ClipReflectance($\mathbf{X}$) *% clip values from 0 to 1*
if *CRISM* **then**
  $\mathbf{X} \leftarrow$ RatioImage($\mathbf{X}$) *% divide by ratio spectrum*
end
$\mathbf{X} \leftarrow$ ClipWavelengths($\mathbf{X}$) *% clip from 1050 to 2550 nm*
$\mathbf{X} \leftarrow \mathbf{X}./\|\mathbf{X}\|_2$ *% per pixel normalization*
if *mask* **then**
  $\mathbf{X} \leftarrow$ mask($\mathbf{X}$, $\mathbf{M}$) *% mask out unwanted pixels*
end
if *not CRISM* **then**
  $\mathbf{X} \leftarrow$ RemoveContinuum($\mathbf{X}$) *% get convex hull*
end
$\hat{\mathbf{X}}$ = Flatten($\mathbf{X}$)

In the case of orbital CRISM imagery, we start with MTR3 products from the Planetary Data System [41], considered the highest-fidelity publicly available CRISM imagery. A sophisticated set of empirical and statistical corrections have been pre-applied to this data to remove spikes, correct for imaging geometry and gimbal motion, and remove atmospheric contamination to retrieve approximate surface reflectance. It is noteworthy that the method used for atmospheric correction is imperfect and leaves considerable CO_2 absorption residuals [31]. One common way that atmospheric residuals and systematic cross-track dependent noise is removed in CRISM imagery is through spectral ratioing with bland material within the image, which also emphasizes minor mineral components [15, 4]. We optionally manually develop a ratio spectrum from a mean of many bland pixels for each image and divide it out of every pixel of the image.

4.2. Feature Extraction

In order to develop a general approach that performs well across different noise settings, we use an autoencoder framework, which has been shown to perform well at denoising [44]. We train on a per-image basis to leverage the inherent mineralogical similarity within an image target. For each image, we train a lightweight autoencoder to learn a per-pixel embedding that is of a lower dimension d than the input image space w. The training data for the image-specific autoencoder consists of the entire hyperspectral image of interest with p pixels with w channels each. The input to the autoencoder is a single pixel $\mathbf{x} \in \mathbb{R}^w$, which is a single w channel spectrum. The encoder network E_ϕ, parameterized by ϕ, generates learned features $\mathbf{z} \in \mathbb{R}^d$. Then, the decoder network D_θ, parameterized by θ, generates a reconstructed spectrum $\hat{\mathbf{x}} \in \mathbb{R}^w$. The dimensionality d of the learned feature space $\mathbf{Z} \in \mathbb{R}^d$ is determined by HySime [1]. The learned features $\mathbf{z}$ are then used for clustering.

Algorithm 2 Feature Extraction and Clustering

Data: $\hat{\mathbf{X}}$ vectorized image p × w, optional spectral angle threshold λ, optional embedding size d, optional number of clusters k
Result: $\mathcal{C}$ clusters p × k
if d *is not given* **then**
  $d \leftarrow$ HySime($\hat{\mathbf{X}}$)
end
$\mathbf{Z} \leftarrow$ Autoencoder($\hat{\mathbf{X}}$, d) *% train with preprocessed data*
if k *is not given* **then**
  $k \leftarrow 2d$
end
$\mathcal{C} \leftarrow$ GaussianMixtureModel($\mathbf{Z}$, k)
if *PostProcess* **then**
  $\mathcal{C} \leftarrow$ PostProcess($\mathcal{C}$, λ)
end

We use a lightweight two-hidden-layer encoder and decoder architecture for training efficiency as well as implicit regularization. We use rectified linear unit (ReLU) as the activation function. Using ReLU instead of output-constrained activation functions, like sigmoid and hyperbolic tangent, allows for faster convergence [25]. We use Adam [49] as our optimizer with a learning rate of 10^{-3} and train until convergence. Instead of the typical mean squared error reconstruction loss, we use the spectral angle (SA) between the input $\mathbf{x}$ and reconstruction $\hat{\mathbf{x}}$, which is defined by

$$\mathrm{SA}(\mathbf{x}, \hat{\mathbf{x}}) = \arccos\left(\frac{\langle \mathbf{x}, \hat{\mathbf{x}} \rangle}{\|\mathbf{x}\|_2 \, \|\hat{\mathbf{x}}\|_2}\right). \tag{7}$$

Using spectral angle allows for invariance to relative magnitude, which results in decreased sensitivity to network initialization, and increases the cost in small scale variation, which helps to capture small features in the spectra. Learning a per-pixel embedding for each image allows for robustness to various mineral abundance distributions and varying systematic instrumental and environmental noise.

4.3. Clustering

We cluster the learned embeddings from the autoencoder on a per-pixel basis using a Gaussian mixture model with full covariance matrices. Unlike k-means, GMMs are robust

	F1		NMI		ARI	
	NCR	CR	NCR	CR	NCR	CR
11 Cl.	**0.260**	0.258	0.254	0.289	0.122	0.185
20 Cl.	0.243	0.236	0.282	0.313	0.109	0.149
GMM+	0.157	0.149	0.332	**0.402**	0.170	**0.221**
PCA+	0.132	0.129	0.223	0.238	0.148	0.181

Table 1: **Supervised Metrics: Oman.** F1, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) using expert labels (11 classes). Clusters generated using non-CR (NCR) and CR embeddings from GMM (11 and 20 clusters), GMM with twice the HySime output (52) and post-processed to 11 (GMM+, full GyPSUM pipeline), and PCA (20 components) + k-means (52 clusters) + post-processing to 11 (PCA+). Best score for each metric is bolded.

to small clusters and can separate clusters that are not well-separated in space. Since the features can be interpreted as corresponding to different minerals, enforcing feature independence would artificially prevent mixtures of minerals to be identified. To determine the number of clusters, we use twice the number estimated by HySime [1] on the spectral data since HySime tends to underestimate the number of distinct endmembers [38]. We can then use post-processing to reduce the number of clusters to a manageable number.

4.4. Post-Processing

Because we fix the number of classes before clustering to twice the estimate from HySime [1], we optionally post-process output clusters by merging redundant classes based on their mean spectra. We use spectral angle as a similarity metric between mean spectra of each cluster. We iteratively combine pairs of clusters with the smallest spectral angle between cluster means until the minimum spectral angle between cluster means exceeds a user-defined threshold λ.

4.5. Evaluation

Classification of geological materials is interpretive by nature, with necessary class specificity dependent on the application. For example, initial expert classification maps for the Oman core data did not differentiate two spectrally distinct zeolite minerals because they did not inform the scientific goal of determining trends in hydration, formation temperatures, and water chemistry with depth. To assess performance of varying preprocessing techniques, autoencoder architectures, embedding sizes, and number of clusters, we apply visual qualitative analytical methods and compare against PCA + k-means, which is used widely in software for unsupervised HSI classification. We hand-select regions of interest (ROIs) in each image covering the range of important mineralogical diversity determined by two spectral geologists in the case of the laboratory data, and through both a partially expert-labelled image and literature results across many publications for the CRISM data [45, 26, 21, 15, 42]. We visually assess each class present within the ROIs to determine if the class is consistently mapping similar material across a large subset of the image. Additionally, we assess whether cluster means of the classes comprising the pixels from the ROIs contain the absorption features representative of the mineralogy, and if mixing with other distinct mineralogy is muted.

In addition to the qualitative evaluation, we employ several quantitative metrics to evaluate our methods. We compute metrics that evaluate the separation and density of clusters for each variation of our pipeline. In particular, we use unsupervised metrics (CH index (Eq. 5) and DB index (Eq. 6)) and supervised metrics (F1, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI)) [37] for the Oman dataset, which has spatially complete expert labels available.

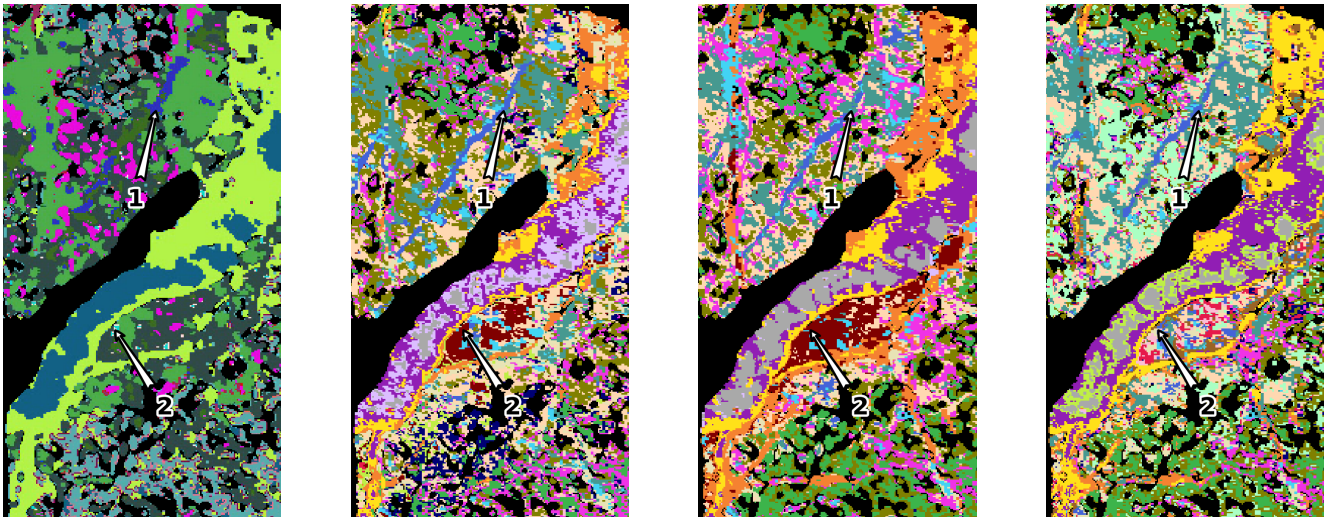
5. Experiments and Results

We apply our methods to two datasets and show results for one image from each with additional images in the supplement. For consistency in presentation of classification images, we show the results for 20 latent embeddings and 20 clusters in Figures 2 and 4 despite better quantitative metrics for different combinations for both images (Tables 1, 2, and 3). Colors are matched between images to maximize the pixel-wise color similarity of the largest clusters. We perform no post-processing to selectively remove redundant endmembers for classification images. The 11-class expertly labeled image for the Oman dataset (Fig. 1a) and the 6-class partially classified Jezero crater image 3a are displayed without color matching.

5.1. Oman Core Evaluation

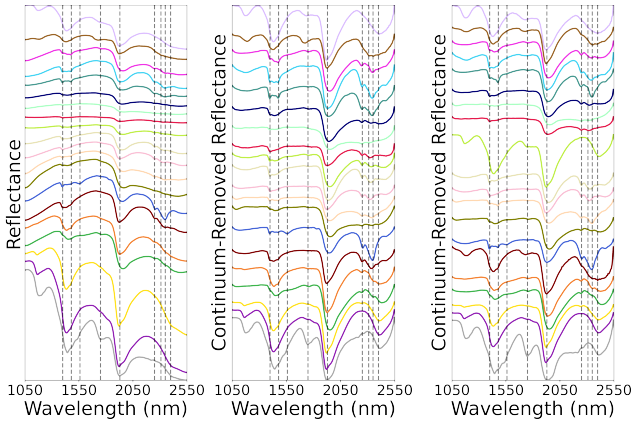
In this image, we expect to map different assemblages of the mineral groups chlorite, pyroxene, zeolite, epidote, prehnite, amphibole, and gypsum, with other minerals present but spectrally inactive in this wavelength range. Several distinct veins not clustered using PCA and k-means or mapped in the expert classification are clearly identified with our pipeline (i.e. multiple orange ■ zeolite veins towards the bottom of the image in Fig. 1c), and speckle noise abundant in the PCA case (Fig. 1b) is clearly reduced. The autoencoder uniquely maps subtle cross-cutting prehnite mixing in an epidote-chlorite vein (labeled 1 in Fig. 1a). The autoencoder also cleanly separates gypsum from zeolite in a large vein in the center of the image where PCA struggles (neon green ■, slate blue ■ in Fig. 1a, grey ■, purple ■ in Fig. 1c).

The main challenge for the current implementation is



(a) Expert Labels (b) PCA + K-Means (c) AE + GMM on Spectral Data (d) AE + GMM on CR Data

Figure 1: **Oman core ROI with different methods** (c) and (d) clearly capture the epidote vein (dark green ■ below 1 in (a), cobalt blue ■ in (c) (d)), while separating distinct prehnite/epidote mixtures (dark blue ■ in (a), blue-green ■ in (c) (d), labeled 1). PCA + k-means (b) fails to identify this distinct mixture. All methods struggle with distinctly mapping amphibole (labeled 2, turquoise ■ in (b)). Full size core images are available in the supplement.



(a) (b)

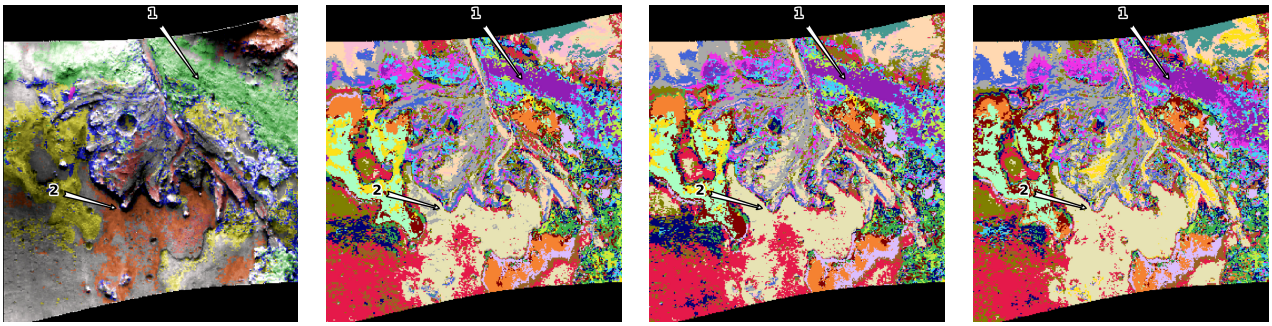
Figure 2: **Comparison of mean spectra for continuum removed (CR) and non-CR clustering on Oman core image (offset for clarity)** (a) Cluster means on spectral data from Fig. 1c (left), with CR duplicate (right) (b) Cluster means of CR data from clusters in Fig. 1d. Key unique absorptions are at 1560 nm (epidote), 1750 nm (gypsum), and 1480 nm (prehnite). Important features are correlated with the vertical dotted lines.

separating out rare, low spatial area mineral classes. With all combinations of hyperparameters, there are no distinctive amphibole (turquoise ■ in expertly labeled image from Fig. 1a, combination of absorptions at 1390 nm, 2320 nm, and 2390 nm) or pyroxene (pink ■ in expertly labeled image from Fig. 1a, strong 1050 nm with no sharp absorptions 2200-2400 nm) clusters mapped. These classes each represent $< 1\%$ of pixels in the image, and the amphibole is only present in subtle mixtures with other minerals, resulting in weak diagnostic absorptions.

Although CR exaggerates spectral features of interest, the CR-learned feature space does not seem to have stronger clustering properties. The non-CR clusters outperform the CR clusterings for the unsupervised metrics (Table 2). Methods using non-CR embeddings also perform better in with the F1 supervised metrics (Table 1), though NMI and ARI are both slightly improved by CR. NMI and ARI metrics for the full implementation of GyPSUM for both spectral and CR data are substantially better than PCA + k-means results (Table 1). Supervised metrics are comparable to results presented for a suite of unsupervised methodologies on a small subset of a Cuprite, Nevada AVIRIS image which contains fewer, arguably more distinct classes [48].

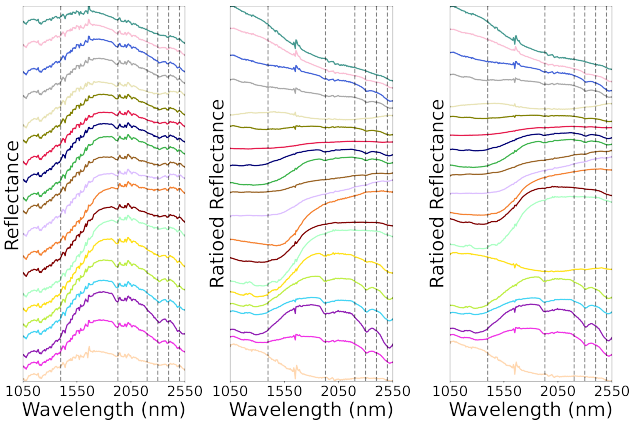
5.2. CRISM Evaluation

For the Jezero Crater image, we expect to map different assemblages of olivine, pyroxene, carbonate, Fe/Mg smectite, hydrated silica, and Al-rich phyllosilicates. We effectively differentiate distinct units of varying olivine and pyroxene, which are primary minerals that have not been altered by interaction with water. In this scene, distinct units with variable chemistry or grain size of these spectrally active components have been identified [16, 21, 26]. Additionally, we map varying characteristics of carbonate/olivine mixtures (purple ■, cyan ■, magenta ■ in Fig.



(a) Partial Expert Classification (b) PCA + K-Means (c) AE + GMM on Unratioed Spectral Data (d) AE + GMM on Ratioed Data

Figure 3: **Map projected Jezero Crater clustering with different methods** (a) Partial expert classified mineral map overlaying greyscale image (olivine, yellow ■; pyroxene, orange ■; carbonate, green ■; Fe/Mg smectite, blue ■; silica, magenta ■; unclassified, gray ■. 6 total classes). Note the delta feature in the top center of the image, with distinct pyroxene-bearing unit bounding its edge below (tan ■, labeled 2). A spatially coherent carbonate unit branches off to the right at its top (purple ■, labeled 1).



(a) (b)

Figure 4: **Comparison of mean spectra for ratioed and unratioed clustering on Jezero Crater image (offset for clarity)** (a) Cluster means on spectral data from Fig. 3c (left) with ratioed duplicate (right) (b) Cluster means of ratioed data from clusters in Fig. 3d. Key unique absorptions are at 1900 nm (water in minerals), 2300 nm and 2500 nm (carbonate), 2300 nm (and no 2500 nm; Fe/Mg smectite), a broad absorption from 1050 nm to 1800 nm (olivine), and a broad absorption from 1300 nm to 2300 nm (pyroxene). Important features are correlated with the vertical dotted lines.

3d), and various Fe/Mg smectite mixtures with both pyroxene and olivine (Fig. 4 grey ■, blue ■, dark blue ■). The ratioing process better defines continuous, unaltered or very weakly altered units (i.e. tan ■ in Fig. 3d).

Ratioing performs better in the unsupervised clustering metrics for both embeddings and spectral data (Table 3). These metrics show that ratioing produces dense clusters in the learned feature space. Ratioed embeddings performed substantially better on the spectral data, and comparison

Unsupervised Clustering Metrics: Oman

CH Scores Using Embeddings ($\times 10^5$)			
	PCA	Non-CR Data	CR Data
15 Clusters	1.858	**1.016**	0.958
20 Clusters	1.627	0.818	**0.876**
25 Clusters	1.464	0.744	**0.801**
DB Scores Using Embeddings			
	PCA	Non-CR Data	CR Data
15 Clusters	1.239	**1.977**	1.986
20 Clusters	1.254	**2.165**	2.280
25 Clusters	1.260	**2.256**	2.266
CH Scores Using Spectral Data ($\times 10^5$)			
	PCA	Non-CR Data	CR Data
15 Clusters	0.392	**0.840**	0.501
20 Clusters	0.330	**0.656**	0.355
25 Clusters	0.279	**0.597**	0.333
DB Scores Using Spectral Data			
	PCA	Non-CR Data	CR Data
15 Clusters	5.887	**5.843**	6.944
20 Clusters	6.320	**6.618**	7.876
25 Clusters	8.284	**5.691**	9.488

Table 2: CH (Eq. 5) and DB (Eq. 6) scores on embeddings $\mathbf{Z} \subseteq \mathbb{R}^{20}$ and spectral data $\mathbf{X}$ varying by number of clusters. Embeddings generated from the following methods: PCA + k-means on spectral data, AE + GMM on spectral data and CR data. The best score between non-CR data and CR data embeddings is in bold. PCA scores are included as reference. Note that PCA, Non-CR, and CR embedding spaces are all different, so scores on embeddings are not directly comparable.

with PCA and k-means shows that our methodology is less sensitive to push-broom sensor striping noise[2]. When comparing the PCA + k-means and AE + GMM on spectral data, the clustering is nearly identical while the spectral data embedding scores are consistently better (Table 3). This seems to indicate that the autoencoder is learning an embedding space that produces better distinct clusters, but further investigation is necessary.

With all combinations of hyperparameters we are unable to uniquely identify hydrated silica or Al-rich phyllosilicates, (which have distinct absorptions near 2200 nm) which instead appear as subtle mixtures with carbonate-dominated clusters. These minerals, and even rarer detections of jarosite and akageneite, are not abundant in the scene and have only recently been mapped exhaustively with new expertly-guided methodologies [42, 14]. These classes are also not mapped in the partial expert classification provided here (Fig. 3a).

6. Conclusions

In this work we find that the GyPSUM pipeline effectively clusters most of the important spectral diversity in both drill-core imagery and remote-sensing imagery. Our pipeline performs comparably to other modern unsupervised classification algorithms and is relatively fast and memory efficient. Spectral ratioing of CRISM imagery unambiguously increases both clustering performance and spectral interpretability, while continuum removal results show similar clustering performance and slightly better NMI and ARI metrics. Overall, this lightweight architecture provides a relatively fast ($\sim$8.5 minutes for 4003x275x249 Oman image, $\sim$3.5 minutes for 455x751x228 Jezero Crater image), effective initial clustering for guiding in-depth work, and provides flexibility for semi-supervised learning by separating the feature extraction and clustering processes. GyPSUM enables rapid determination of distinct mineral classes across multiple imaging systems and noise profiles, demonstrating that the technique is highly generalizable. Its main shortcoming is non-identification of spectrally distinct but spatially rare ($< 1\%$) mineral classes that can be geologically significant. The current implementation also requires user input of a spectral angle stopping condition for optional post-processing to determine a final number of clusters. Future work will include weighted sampling for clustering to improve computation time and windowing the data or hierarchical clustering to better identify spatially rare classes.

Acknowledgements

The authors would like to thank the Oman Drilling Project and the CRISM team for access to the data used in this work. We would also like to thank Richard Murray and Sara Beery for additional comments which improved this paper.

[2]See FRT0000634B in Supplementary Materials

Unsupervised Clustering Metrics: Jezero Crater

CH Scores Using Embeddings ($\times 10^5$)			
	PCA	Spectral Data	Ratioed Data
10 Clusters	1.242	1.391	**1.772**
15 Clusters	1.047	1.247	**1.719**
20 Clusters	0.917	1.164	**1.563**
25 Clusters	0.821	1.153	**1.449**
DB Scores Using Embeddings			
	PCA	Spectral Data	Ratioed Data
10 Clusters	1.098	1.078	**1.010**
15 Clusters	1.168	1.116	**1.011**
20 Clusters	1.178	1.106	**1.071**
25 Clusters	1.248	**1.095**	1.208
CH Scores Using Spectral Data ($\times 10^5$)			
	PCA	Spectral Data	Ratioed Data
10 Clusters	0.119	0.012	**0.179**
15 Clusters	0.085	0.008	**0.130**
20 Clusters	0.070	0.007	**0.102**
25 Clusters	0.057	0.006	**0.084**
DB Scores Using Spectral Data			
	PCA	Spectral Data	Ratioed Data
10 Clusters	4.545	15.179	**4.636**
15 Clusters	5.657	15.874	**6.399**
20 Clusters	6.727	16.344	**6.543**
25 Clusters	7.126	15.935	**5.956**

Table 3: CH (Eq. 5) and DB (Eq. 6) scores on embeddings $\mathbf{Z} \subseteq \mathbb{R}^{20}$ and spectral data $\mathbf{X}$ varying by number of clusters. Embeddings generated from the following methods: PCA + k-means on spectral data, AE + GMM on spectral data and ratioed data. The best score between spectral data embeddings and ratioed data embeddings is in bold. PCA scores are included as reference. Note that PCA, unratioed, and ratioed embedding spaces are all different, so scores on embeddings are not directly comparable.

References

[1] José M. Bioucas-Dias and José M. P Nascimento. Hyperspectral Subspace Identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445, Aug. 2008.

[2] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, June 2013. Conference Name: IEEE Geoscience and Remote Sensing Magazine.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[4] Janice L. Bishop, Alberto G. Fairén, Joseph R. Michalski, Luis Gago-Duport, Leslie L. Baker, Michael A. Velbel, Christoph Gross, and Elizabeth B. Rampe. Surface clay formation during short-term warmer and wetter conditions on a largely cold ancient Mars. *Nature Astronomy*, 2(3):206–213, Mar. 2018.

[5] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[6] W. M. Calvin. Band Parameterization for Imaging Spectrometer Systems: Lessons Learned from CRISM at Mars. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 8356–8358, July 2018. ISSN: 2153-7003.

[7] Xavier Ceamanos and Sylvain Douté. Spectral smile correction of CRISM/MRO hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):3951–3959, 2010.

[8] Chein-I Chang and Qian Du. Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 42(3):608–619, Mar. 2004.

[9] Roger N. Clark, Trude V. V. King, Matthew Klejwa, Gregg A. Swayze, and Norma Vergo. High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research: Solid Earth*, 95(B8):12653–12680, 1990.

[10] Roger N. Clark and Ted L. Roush. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research: Solid Earth*, 89(B7):6329–6340, 1984.

[11] Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Raymond F. Kokaly, Steve J. Sutley, J. Brad Dalton, Robert R. McDougal, and Carol A. Gent. Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. *Journal of Geophysical Research: Planets*, 108(E12), 2003.

[12] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[13] M. Dundar and B. L. Ehlmann. Rare jarosite detection in CRISM imagery by non-parametric bayesian clustering. In *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5, 2016.

[14] Murat Dundar, Bethany L. Ehlmann, and Ellen K. Leask. Machine-Learning-Driven New Geologic Discoveries at Mars Rover Landing Sites: Jezero and NE Syrtis. *arXiv:1909.02387 [astro-ph, stat]*, Sept. 2019. arXiv: 1909.02387.

[15] B. L. Ehlmann, J. F. Mustard, and S. L. Murchie. Geologic setting of serpentine deposits on Mars. *Geophysical Research Letters*, 37(6), 2010.

[16] Bethany L. Ehlmann, John F. Mustard, Gregg A. Swayze, Roger N. Clark, Janice L. Bishop, Francois Poulet, David J. Des Marais, Leah H. Roach, Ralph E. Milliken, James J. Wray, Olivier Barnouin-Jha, and Scott L. Murchie. Identification of hydrated silicate minerals on Mars using MRO-CRISM: Geologic context near Nili Fossae and implications for aqueous alteration. *Journal of Geophysical Research: Planets*, 114(E2), 2009.

[17] Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[18] Kenneth A. et al. Farley. Mars 2020 Mission Overview. *Space Science Reviews*, 216(8):142, Dec. 2020.

[19] Utsav B. Gewali, Sildomar T. Monteiro, and Eli Saber. Machine learning based hyperspectral image analysis: A survey. *arXiv:1802.08701 [cs, eess]*, Feb. 2019. arXiv: 1802.08701.

[20] K. W. Glennie, M. G. A. Boeuf, M. W. Hughes Clarke, M. Moody-Stuart, W. F. H. Pilaar, and B. M. Reinhardt. Late Cretaceous Nappes in Oman Mountains and Their Geologic Evolution1. *AAPG Bulletin*, 57(1):5–27, 01 1973.

[21] Timothy A. Goudge, John F. Mustard, James W. Head, Caleb I. Fassett, and Sandra M. Wiseman. Assessing the mineralogy of the watershed and fan deposits of the Jezero crater paleolake system, Mars. *Journal of Geophysical Research: Planets*, 120(4):775–808, 2015.

[22] Rebecca N. Greenberger, Bethany L. Ehlmann, Gordon R. Osinski, Livio L. Tornabene, and Robert O. Green. Compositional Heterogeneity of Impact Melt Rocks at the Haughton Impact Structure, Canada: Implications for Planetary Processes and Remote Sensing. *Journal of Geophysical Research: Planets*, 125(10):e2019JE006218, 2020.

[23] Bruce Hapke. Bidirectional reflectance spectroscopy: 1. Theory. *Journal of Geophysical Research: Solid Earth*, 86(B4):3039–3054, 1981.

[24] J. Harsanyi, W. Farrand, and C.-I Chang. Determining the number and identity of spectral endmembers; an integrated approach using Neyman-Person eigen-thresholding and iterative constrained RMS error minimization. In *9th Thematic Conference on Geologic Remote Sensing*, 1993.

[25] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training, 2019.

[26] Briony H. N. Horgan, Ryan B. Anderson, Gilles Dromart, Elena S. Amador, and Melissa S. Rice. The mineral diversity of Jezero crater: Evidence for possible lacustrine carbonates on Mars. *Icarus*, 339:113526, Mar. 2020.

[27] Peter et al. Kelemen. *Oman Drilling Project, Scientific Drilling in the Samail Ophiolite, Sultanate of Oman: Proceedings of the Oman Drilling Project*. 01 2020.

[28] Raymond F. Kokaly, Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Todd M. Hoefen, Neil C. Pearson, Richard A. Wise, William M. Benzel, Heather A. Lowers, Rhonda L. Driscoll, and Anna J. Klein. USGS Spectral Library Version 7. USGS Numbered Series 1035, U.S. Geological Survey, Reston, VA, 2017. IP-075936.

[29] EK Leask, BL Ehlmann, MM Dundar, SL Murchie, and FP Seelos. Challenges in the search for perchlorate and other

hydrated minerals with 2.1-μm absorptions on mars. *Geophysical research letters*, 45(22):12–180, 2018.

[30] E. K. Leask, B. L. Ehlmann, M. M. Dundar, S. L. Murchie, and F. P. Seelos. Challenges in the search for perchlorate and other hydrated minerals with 2.1-m absorptions on mars. *Geophysical Research Letters*, 45(22):12,180–12,189, 2018.

[31] Patrick C. et al. McGuire. An improvement to the volcano-scan algorithm for atmospheric correction of CRISM and OMEGA spectral data. *Planetary and Space Science*, 57(7):809–815, June 2009.

[32] Susan K. Meerdink, Simon J. Hook, Dar A. Roberts, and Elsa A. Abbott. The ECOSTRESS spectral library version 1.0. *Remote Sensing of Environment*, 230:111196, Sept. 2019.

[33] L. Mou, P. Ghamisi, and X. X. Zhu. Unsupervised Spectral–Spatial Feature Learning via Deep Residual Conv–Deconv Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):391–406, Jan. 2018.

[34] S. et al. Murchie. Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) on Mars Reconnaissance Orbiter (MRO). *Journal of Geophysical Research: Planets*, 112(E5), 2007.

[35] James M. Murphy and Mauro Maggioni. Unsupervised Clustering and Active Learning of Hyperspectral Images with Nonlinear Diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, PP:1–17, Oct. 2018.

[36] Jakub Nalepa, Michal Myller, Yasuteru Imai, Ken-ichi Honda, Tomomi Takeda, and Marek Antoniak. Unsupervised Segmentation of Hyperspectral Images Using 3D Convolutional Autoencoders. *CoRR*, abs/1907.08870, 2019.

[37] Reihaneh Rabbany and Osmar Zaane. A general clustering agreement index: For comparing disjoint and overlapping clusters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[38] B. Rasti, M. O. Ulfarsson, and J. R. Sveinsson. Hyperspectral subspace identification using sure. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2481–2485, 2015.

[39] Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2):115–122, 2002.

[40] Mike Searle and Jon Cox. Tectonic setting, origin, and obduction of the Oman ophiolite. *GSA Bulletin*, 111(1):104–122, 01 1999.

[41] F. P. Seelos, S. L. Murchie, D. C. Humm, O. S. Barnouin, F. Morgan, H. W. Taylor, C. Hash, and CRISM Team. CRISM Data Processing and Analysis Products Update — Calibration, Correction, and Visualization. *42nd Annual Lunar and Planetary Science Conference*, 42:1438, Mar. 2011.

[42] J. D. Tarnas, J. F. Mustard, Honglei Lin, T. A. Goudge, E. S. Amador, M. S. Bramble, C. H. Kremer, X. Zhang, Y. Itoh, and M. Parente. Orbital Identification of Hydrated Silica in Jezero Crater, Mars. *Geophysical Research Letters*, 46(22):12771–12782, 2019.

[43] David R. Thompson, Joseph W. Boardman, Michael L. Eastwood, and Robert O. Green. A large airborne survey of Earth's visible-infrared spectral dimensionality. *Optics Express*, 25(8):9186, Apr. 2017.

[44] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

[45] C. E. Viviano-Beck, S. L. Murchie, A. W. Beck, and J. M. Dohm. Compositional and structural constraints on the geologic history of eastern Tharsis Rise, Mars. *Icarus*, 284:43–58, Mar. 2017.

[46] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised Deep Embedding for Clustering Analysis. *arXiv:1511.06335 [cs]*, May 2016. arXiv: 1511.06335.

[47] Himanshi Yadav, Alberto Candela, and David Wettergreen. A study of unsupervised classification techniques for hyperspectral datasets. pages 2993–2996, 07 2019.

[48] Himanshi Yadav, Alberto Candela, and David Wettergreen. A Study of Unsupervised Classification Techniques for Hyperspectral Datasets. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 2993–2996, Yokohama, Japan, July 2019. IEEE.

[49] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. 2020. https://d2l.ai.