GeologyCLIP: A Hierarchical CLIP trained on geological information for Airborne LiDAR data

Takayuki Shinohara
shinohara.takayuki@aist.go.jp

National Institute of Advanced Industrial Science and Technology, AIST
Tsukuba, Japan

Abstract

Deep-learning models tailored to individual airborne LiDAR applications are gaining traction in Earth-science research, yet they face two persistent hurdles: scarce task-specific labels and poor generalisation across geographic regions. We address both issues with GeologyCLIP, a contrastive pre-training framework that jointly learns from airborne LiDAR point clouds and accompanying textual descriptions. GeologyCLIP employs a transformer encoder to capture rich, geometry-aware representations; the encoder is first trained on a large, heterogeneous corpus and then fine-tuned on limited-label downstream datasets. Across multiple regional benchmarks for geohazard detection, GeologyCLIP consistently surpasses task-specific baselines, demonstrating superior transferability and label efficiency. These results position GeologyCLIP as a promising foundation model for geological applications and open new avenues for data-efficient Earth-science analytics.

1 Introduction

The field of Earth science is entering the big data era and artificial intelligence (AI) offers substantial potential not only for solving traditional Earth science problems but also for enhancing our understanding of the Earth's complex, interactive, and multiscale processes [2, 34]. The availability of massive volumes of Earth system data, which already exceed dozens of petabytes in scale and have hundreds of terabytes transmitted daily, has led to the widespread adoption of AI, including machine learning and deep learning methods, in data-driven Earth science [13, 58]. For example, deep learning has been effectively applied to identify extreme weather patterns [78], develop competitive weather prediction models ranging from precipitation nowcasting [57, 90] to medium-range weather forecasting [5, 12, 32, 33], and predict climate phenomena such as El Niño-southern oscillation [21] or monsoon onsets [44]. Additionally, in recent years, machine learning and deep learning methods have proven effective in almost every subfield of seismology [3, 46]. These methods have consistently outperformed classical approaches on a wide range of tasks, including denoising [52, 71, 77, 81, 94], earthquake detection [60, 82, 83], phase picking [7, 17, 37, 47, 50, 53, 59, 74], phase association [41, 42, 61, 84], localization

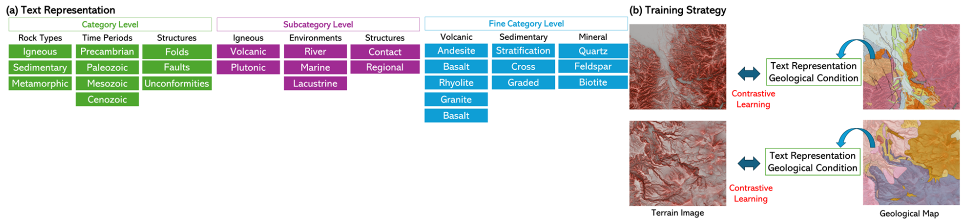

Figure 1: (a) Hierarchical text representations for geological condition. (b) Hierarchical representations of geological labels are fed into the standard contrastive pre-training objective and are matched with image representations of airborne LiDAR data.

[15, 38, 40, 45, 72, 88, 89], event classification [10, 26, 27, 29, 35, 36], focal mechanism determination [22, 30, 59, 69, 70, 91], and earthquake prediction [9, 25, 62, 66, 75, 76]. Many existing methods focus on training specific models for individual tasks.

To leverage the relationships between related tasks effectively, some researchers have proposed methods to address multiple interrelated tasks simultaneously, such as earthquake detection and phase picking [48, 92, 93], earthquake monitoring [54, 67, 95], as well as localization and magnitude estimation [49]. Although deep learning has been actively utilized in meteorology and seismology, its application in the field of geology has been relatively limited. Existing research primarily focuses on extracting geological hazards such as landslides from satellite imagery [39] and classifying rock types based on image data [1].

Despite these advances, most current workflows still rely on *task-specific* networks retrained from scratch, an approach that suffers from three fundamental limitations: (i) label scarcity—many geological targets (e.g., rare lithologies, incipient slope failures) have only dozens to hundreds of annotated examples; (ii) poor cross-regional generalisation—models tuned for one tectonic or climatic setting often fail when applied elsewhere; and (iii) computational inefficiency—repeated training for every new sensor or task wastes both energy and research time. A **pre-trained foundation model** that learns generic, geometry-aware representations from vast, heterogeneous airborne LiDAR archives offers a direct remedy. Such a model can be fine-tuned with minimal supervision, ported seamlessly across regions, and serve as a unifying backbone for diverse downstream tasks ranging from landslide detection to rock-type classification. A **pre-trained foundation model** that learns generic, geometry-aware representations from vast, heterogeneous airborne LiDAR archives offers a direct remedy. Crucially, recent progress in vision–language pre-training has shown that contrastive models such as CLIP [56] can align images with natural language descriptions, and nascent extensions have begun to port CLIP to 3-D point clouds. Yet existing CLIP for point cloud, such as PointCLIP [85], are confined to indoor or small object-centric scans; to our knowledge, no method scales the paradigm to *outdoor, kilometre-scale* airborne LiDAR scenes. We bridge this gap by rasterising the LiDAR point cloud into multi-view, terrain-aware image projections, enabling direct use of mature vision–language architectures while retaining the geometric richness of the original data.

Such a model can be fine-tuned with minimal supervision, ported seamlessly across regions, and serve as a unifying backbone for diverse downstream tasks ranging from landslide detection to rock-type classification. In this paper we therefore introduce *GeologyCLIP*, the first contrastively pre-trained LiDAR–text model designed to provide a transferable representation for geological applications.

2 Related Work

AI in GeoScience. Across the geosciences, artificial intelligence (AI) is increasingly applied in remote sensing [64, 65] and in seismology [24], notably for seismic waveform analysis. With the growing application of large foundational models to general-purpose tasks, the exploration of foundational models tailored to remote-sensing-based geoscience tasks has garnered significant attention from the research community. Here, we review recent advancements in geoscience foundational models (GFMs), covering key techniques for constructing GFMs and summarizing existing foundational models from the perspectives of large language models [14], large vision models [20], and large language-vision models [31, 87]. These vision and language foundational models are primarily trained on satellite data and there is a growing need to develop foundational models specifically tailored to geological data.|

Vision and Language. Multimodal Foundational Model like CLIP [56] has achieved state-of-the-art performance on vision tasks by training on noisy, web-scale datasets containing over 100 million image-text pairs using a contrastive objective optimized for image retrieval. Subsequent models such as ALIGN [23] and BASIC [55] expanded the number of training examples to 400 million and 6.6 billion, respectively, further enhancing the quality of vision representations. However, recent studies [16, 18, 51, 79, 80] have demonstrated that dataset diversity and improved alignment between image and caption semantics are more critical than dataset size, leading to superior performance on downstream tasks.

Hierarchical Structure. The concept of hierarchies has been well explored in computer vision, primarily because ImageNet [63] classes are derived from the hierarchical structure of WordNet [43]. For example, Bilal et al. [6] analyzed model predictions on ImageNet and discovered that model confusion patterns often corresponded to hierarchical class structures. By incorporating this hierarchical information into AlexNet's architecture [28], they achieved an absolute improvement of 8% in the top-1 error rate on ImageNet. Similarly, Bertinetto et al. [4] examined the severity of errors made by image classifiers and proposed alternative training objectives that integrate hierarchical information. Although this approach only slightly increased the top-1 error rate, it successfully reduced the severity of mistakes. In another study, Zhang et al. [86] introduced a contrastive objective that aligns the hierarchical distances between labels with the corresponding distances in the embedding space. This method outperformed traditional cross-entropy loss on both ImageNet and iNat17 [73].

3 Proposed Method

3.1 Architecture of GeologyCLIP

Our objective is to develop a foundational model for geology by leveraging the success of contrastive learning as demonstrated by CLIP (Figure 2). We propose adapting the vision encoder of CLIP to process terrain images derived from digital terrain models (DTMs). By representing 3D terrain data as 2D images, we can harness the computational efficiency and scalability of existing computer vision techniques.

To improve computational efficiency, we divide terrain images into smaller patches, similar to the approach used in the original CLIP model. These patches are then fed into a

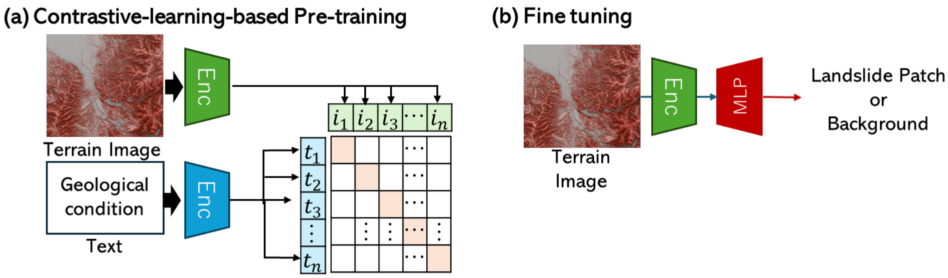

Figure 2: Summary of GeologyClip. (a) GeologyClip consists of two encoders that are pre-trained jointly using contrastive learning on multi-modal data comprising terrain image and corresponding geological and geomophogical information. (b) In a downstream task, the pre-trained image encoder is used to generate features, which are then fed into MLP for geohazard (landslide) classification.

transformer encoder, where they are processed to extract meaningful representations. By pre-training this model on a large and diverse dataset of terrain images, we aim to learn generalizable visual representations that can be adapted to various downstream geological tasks.

Our proposed GeologyCLIP model employs a dual-branch encoder architecture (Figure 3(a)). One branch processes terrain images using a pre-trained ViT-small network as its backbone, leveraging knowledge from the ImageNet dataset. The other branch processes geological text data, including terrain and geological condition information, and encodes it into a compact 1D vector using an MLP. These two encoders are jointly trained using a contrastive learning objective, aligning image and information features.

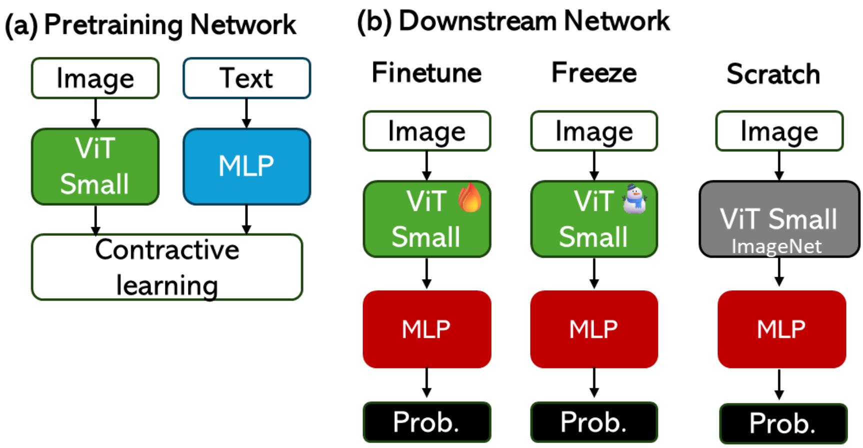

Figure 3: Transformer-based Encoder and Decoder. (a) Detailed network architectures of the two encoders during the pre-training phase. (b) Three different training strategies and the corresponding network architectures for the downstream task of landslide classification.

Each training sample consists of a three-channel terrain image and associated geological information. The image is processed by the ViT-small encoder, while the geological information is encoded by the MLP. The resulting image and information features are then used as inputs for contrastive learning, which encourages the model to learn representations that are similar for samples from the same class and dissimilar for samples from different classes.

3.2 Data Preparation for Pre-training and Validation

Terrain Image Data. In this study, we first convert the raw airborne LiDAR point clouds to a **Digital Terrain Model** (DTM) by ground–surface filtering and grid-based interpolation. The resulting DTM is a single–channel raster in which each pixel records bare-earth elevation at 1 m resolution. We use the nationwide DTM tiles released by the Geospatial Information Authority of Japan, a standard reference for high-precision terrain analysis[1].

To enrich the geomorphometric information, we derive auxiliary terrain attributes—slope, hill-shade, relief degree of land surface, and various curvatures—directly from the DTM. Stacking these derivatives with the elevation band yields a multi-channel *terrain image* (Fig. 4) that preserves the metric fidelity of the LiDAR data while remaining compatible with convolution-based encoders.

From the full Japanese archive we extract 2,935 spatial extents that cover both common and rare geological features. The data are randomly partitioned into 2,364 training extents, 571 validation extents, and 483 held-out extents for cross-regional testing. Each extent is further tiled into 512×512 patches to provide sufficient spatial context; patches from different splits do not overlap, preventing information leakage during training and evaluation.

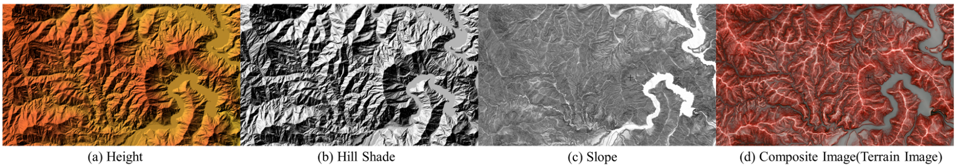

Figure 4: Input terrain image of our ViT-based encoder. Terrain images are a combined image of height, hill shade, and slope.

Text Data. A key advantage of CLIP is its ability to accept free-form text descriptions. In the context of geology, this allows us to incorporate a wide range of textual information, including stratigraphic names, scientific classifications, and common geological terms.

Geological Condition. In geology, unlike other classification tasks, category names are diversely formatted. We consider the following categories[2]:

- Stratigraphic Names: To represent the hierarchical nature of stratigraphic units, we concatenate all labels from the highest (e.g., eon) to the lowest level (e.g., member) into a single string. For example, the stratigraphic name "Paleozoic Era, Carboniferous Period, Mississippian Epoch" would be used as a single text input.

[1]Terrain data was obtained from the AIST geological database.

[2]The Geological data was obtained from the AIST geological database.

- Scientific Classifications: We include scientific classifications based on the composition, texture, and formation process of rocks and minerals. These classifications provide detailed information regarding the geological materials present.

- Common Classifications: In addition to scientific classifications, we also incorporate common geological terms that are more widely understood. Terms such as "sandstone" or "limestone" may not always have a direct correspondence with specific stratigraphic units but can provide valuable context for image-text matching.

Geomorphological condition. Additionally, we consider geomorphological text as follows[3]:

- Geomorphological hierarchy: A standard hierarchy in geomorphology from higher to lower levels may include continental, regional, sub-regional, and local landforms. For each landform, we "flatten" this hierarchy by concatenating all labels from the broadest to the most specific into a single string, which we call the *geomorphological name*.

- Scientific classification: Scientific classifications of landforms are based on their origin, structure, and process of formation (e.g., *fluvial terrace, aeolian dune*). These classifications are used in the same manner as taxonomic names in biology.

- Common classification: Geomorphological classifications are often highly technical and specific, which may not be reflected in generalist image-text pre-training datasets. Common classifications such as "mountain," "valley," or "plain" are more widespread. Note that common classifications may not have a one-to-one mapping to specific landforms because a single landform may have multiple common names or the same common name may refer to different types of landforms.

4 Results and Discussion

4.1 Geohazard Classification

We evaluated the effectiveness of CLIP for geohazard (landslide) classification by first pre-training our model on a large dataset and then fine-tuning it on a specific task. Specifically, we evaluated the effectiveness of using a CLIP model fine-tuned on landslide patch/background patch classification task. To assess the impact of fine-tuning the CLIP encoder, we conducted experiments comparing the classification performance of three different approaches: fine-tuning the pre-trained image encoder (Finetune in Figure 2), training the same encoder from scratch (Scratch in Figure 2), and applying transfer learning from a pre-trained CLIP model without further fine-tuning (Frozen in Figure 2).

The results indicate that fine-tuning the CLIP encoder on domain-specific geological data significantly enhances its performance for classifying geological formations such as rock types and stratigraphic layers (Table 1). The fine-tuned model (Finetune in Table 1) outperformed both the model trained from scratch (Scratch in Table 1) and the transfer learning model without fine-tuning (Freeze in Table 1). Specifically, the fine-tuned CLIP model demonstrated superior accuracy and robustness in terms of handling the diverse and complex nature of geological data, which often involve subtle distinctions between classes.

[3] The Geological data was obtained from the Geographical Survey Institute.

In contrast, the model trained from scratch exhibited lower classification performance, likely due to the limited size of the geological dataset compared with the large-scale data typically used in pre-training CLIP models. This result underscores the importance of leveraging pre-trained models, especially when handling specialized datasets that may not have extensive labeled data available.

The frozen approach, while better than training from scratch, did not match the performance of the fine-tuned model. This suggests that while the pre-trained CLIP model contains valuable general visual and textual representations, fine-tuning is crucial for adapting these representations to the specific nuances of geological data.

Overall, these findings highlight the effectiveness of fine-tuning pre-trained models such as CLIP models for domain-specific tasks in the geosciences, offering a promising approach for improving classification accuracy in applications where data diversity and complexity are prevalent.

Table 1: The performance of geohazard classification

Models	F1 Score
Scratch (GeologyClip)	88.1
Freeze (GeologyClip)	92.2
Finetune (GeologyClip)	94.7

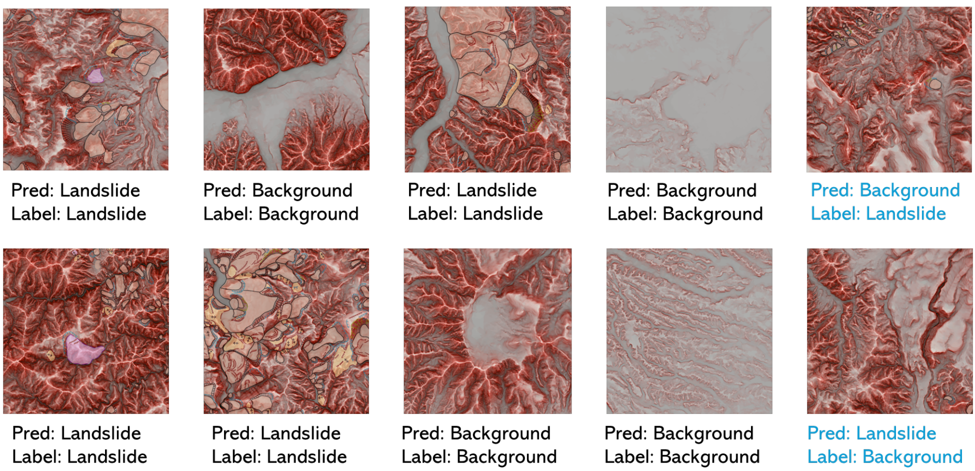

Figure 5: Geohazard (landslide) Classification Results. Our fine-tuned GeologyCLIP classifed landslide patch or background (not landslide) patch.

4.2 Zero-Shot Classification of Geological Categories

Here, we use our pre-trained GeologyCLIP model to classify geological categories. We consider four geological categories: sedimentary rock, igneous rock, accretionary rock, and metamorphic rock.

Chao et al. [11] introduced the concept of *generalized zero-shot learning (GZSL)*, which we used to classify unseen and seen terrain images. We selected a set of 400 *seen* terrain image samples from our dataset and performed the classification of geological categories on these terrain images. The zero-shot accuracy is 26.4 in $A_{\mathcal{U}\to\mathcal{T}}$ and 63.2 in $A_{\mathcal{S}\to\mathcal{T}}$. We then augmented our testing set by gathering an additional 4000 *unseen* terrain images from various areas, without removing other images of these areas from our dataset. The classification results for the unseen dataset are presented in Figure 6. Following the methodology of [11], we evaluated the conventional zero-shot accuracy of GeologyCLIP.

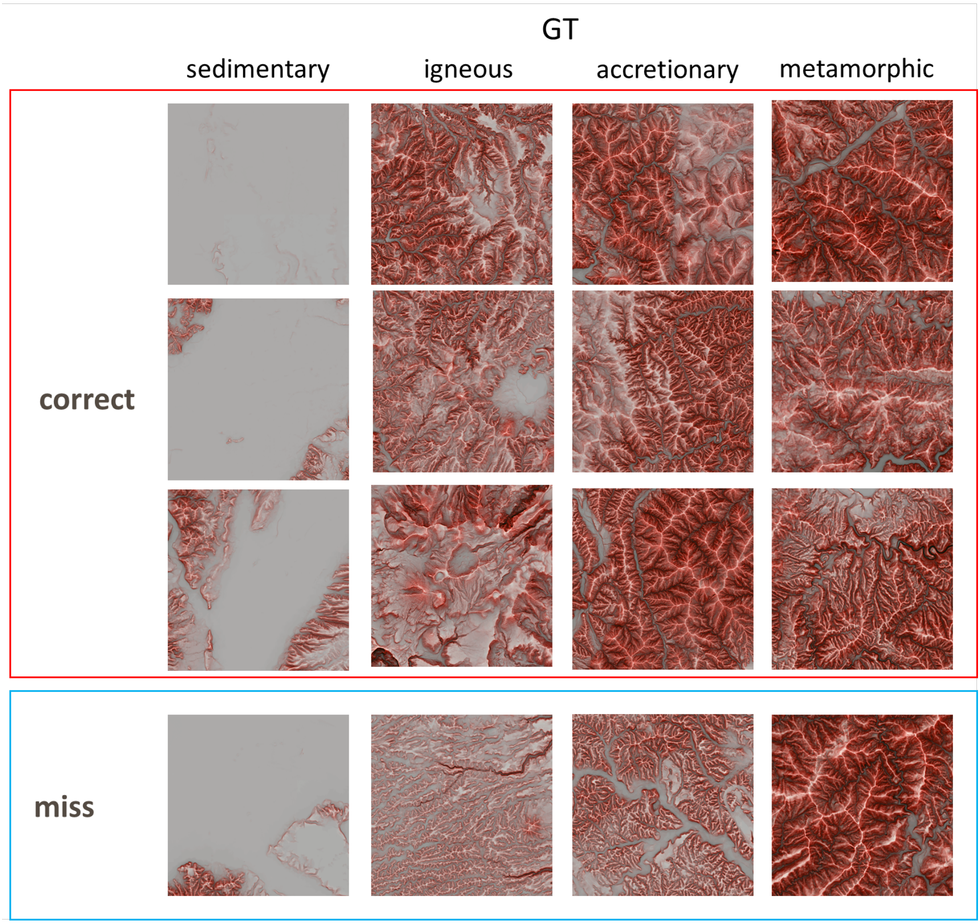

Figure 6: Zeroshot Classification Results of geological category for terrain image.

4.3 Is the CLIP Objective Necessary?

Using the CLIP objective function to pre-train on a labeled image dataset is an unintuitive decision. (Goyal et al. [19] performed fine-tuning using the CLIP objective but did not perform pre-training). We justify this decision by training two ViT-small models on our terrain image and text dataset using cross-entropy classification loss and a multitask hierarchical variant, and then evaluate these models against the CLIP objective in a few-shot setting. The multitask hierarchical training objective is to predict the labels for rock type, time periods, etc. down to fine categories using cross-entropy for each level of the taxonomy, and then sum those losses [8].

We evaluated each model on the 4,000 unseen terrain images used in Zero-Shot Classification section. One-shot and five-shot settings were evaluated because non-CLIP models cannot perform zero-shot classification. We report mean accuracy values in Table 2. The hierarchical classification model outperformed simple classification and is comparable to the CLIP baseline (see Table 2). However, the CLIP objective massively outperforms both baselines, strongly justifying our repurposing of the CLIP objective.

Objective	Mean 1-Shot	Mean 5-shot
Cross-entropy	16.5	26.2
Hier. cross-entropy	19.3	30.5
CLIP	**44.7**	**63.8**

Table 2: One- and five-shot classification top-1 accuracy for different pre-training objectives on our dataset. Results are macro-averaged over all the test sets.

5 Conclusion

We introduce a novel foundation model for Earth-science workflows that learns directly from airborne LiDAR point clouds rasterised into multi-channel terrain images. Each point cloud is projected into a set of georeferenced 2-D rasters—elevation, slope, curvature, hillshade, and relief—so that the rich 3-D geometry remains intact yet becomes compatible with convolutional backbones. Using these LiDAR-derived images paired with expert text descriptions, we train a CLIP-style contrastive encoder that aligns terrain appearance with geological semantics. Consequently, the pre-trained model acquires a robust, language-grounded understanding of geomorphological patterns and can be adapted to diverse downstream tasks—such as landslide mapping, lithology classification, or fault-scarp recognition—with only a few fine-tuning iterations instead of training from scratch for each task.

However, this study has some limitations. First, the training and testing data were constructed from a limited dataset collected in Japan. Therefore, our model was not trained to handle a variety of classes, unlike the original CLIP model, and it does not map global-scale data to geological conditions. Therefore, the training data will need to be expanded in the future. If a country has a database of geological information, the proposed method can be used to prepare teacher signals for pre-training at a low cost. Additionally, considering the amount of data and computational resources available, we were unable to use a rich transformer-based encoder for image feature extraction. When the dataset is expanded, it will be necessary to perform training with a larger model.

Our model consistently outperformed baseline methods on multiple datasets from diverse regions, demonstrating its generalizability and adaptability. This work establishes a new benchmark for Earth science deep learning and paves the way for future research on developing more comprehensive and universal geological AI systems.

Acknowledgements

The geological data utilized in this study were provided by the Geological Survey of Japan (GSJ), and the geomorphological data were made available by the Geospatial Information Authority of Japan (GSI). We used ABCI 3.0 provided by AIST and AIST Solutions[68]. This work was supported by a grant from The Maeda Engineering Foundation.

References

[1] Korhan Ayranci, Isa E. Yildirim, Umair bin Waheed, and James A. MacEachern. Deep learning applications in geosciences: Insights into ichnological analysis. *Applied Sciences*, 11(16), 2021. ISSN 2076-3417. doi: 10.3390/app11167736. URL https://www.mdpi.com/2076-3417/11/16/7736.

[2] Karianne J Bergen, Paul A Johnson, Maarten V de Hoop, and Gregory C Beroza. Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433): eaau0323, 2019.

[3] Gregory C Beroza, Margarita Segou, and S Mostafa Mousavi. Machine learning and earthquake forecasting—next steps. *Nature communications*, 12(1):4761, 2021.

[4] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12506–12515, 2020.

[5] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pages 1–6, 2023.

[6] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):152–162, 2018. doi: 10.1109/TVCG.2017.2744683.

[7] Muhammad Atif Bilal, Yanju Ji, Yongzhi Wang, Muhammad Pervez Akhter, and Muhammad Yaqub. Early earthquake detection using batch normalization graph convolutional neural network (bngcnn). *Applied Sciences*, 12(15):7548, 2022.

[8] Kim Bjerge, Quentin Geissmann, Jamie Alison, Hjalte MR Mann, Toke T Høye, Mads Dyrmann, and Henrik Karstoft. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77:102278, 2023.

[9] Prabhav Borate, Jacques Rivière, Chris Marone, Ankur Mali, Daniel Kifer, and Parisa Shokouhi. Using a physics-informed neural network and fault zone acoustic monitoring to predict lab earthquakes. *Nature Communications*, 14(1):3693, 2023.

[10] Y Bregman, O Lindenbaum, and N Rabin. Array based earthquakes-explosion discrimination using diffusion maps. *Pure and Applied Geophysics*, 178:2403–2418, 2021.

[11] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 52–68. Springer, 2016.

[12] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

[13] Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170, 2024.

[14] Huseyin Denli, Hassan A Chughtai, Brian Hughes, Robert Gistri, and Peng Xu. Geoscience language processing for exploration. In *Abu Dhabi International Petroleum Exhibition and Conference*, page D031S102R003. SPE, 2021.

[15] Phoebe MR DeVries, Fernanda Viégas, Martin Wattenberg, and Brendan J Meade. Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720): 632–634, 2018.

[16] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *International Conference on Machine Learning*, pages 6216–6234, 2022.

[17] Tian Feng, Saeed Mohanna, and Lingsen Meng. Edgephase: A deep learning model for multi-station seismic phase picking. *Geochemistry, Geophysics, Geosystems*, 23(11): e2022GC010453, 2022.

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[19] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

[20] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv preprint arXiv:2312.10115*, 2023.

[21] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019.

[22] Shota Hara, Yukitoshi Fukahata, and Yoshihisa Iio. P-wave first-motion polarity determination of waveform data in western japan using deep learning. *Earth, Planets and Space*, 71(1):1–11, 2019.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[24] Pengcheng Jiao and Amir H Alavi. Artificial intelligence in seismology: advent, performance and future trends. *Geoscience Frontiers*, 11(3):739–744, 2020.

[25] Paul A Johnson, Bertrand Rouet-Leduc, Laura J Pyrak-Nolte, Gregory C Beroza, Chris J Marone, Claudia Hulbert, Addison Howard, Philipp Singer, Dmitry Gordeev, Dimosthenis Karaflos, et al. Laboratory earthquake forecasting: A machine learning competition. *Proceedings of the national academy of sciences*, 118(5):e2011362118, 2021.

[26] Gwantae Kim, Bonhwa Ku, Jae-Kwang Ahn, and Hanseok Ko. Graph convolution networks for seismic events classification using raw waveform data from multiple stations. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[27] Qingkai Kong, Ruijia Wang, William R Walter, Moira Pyle, Keith Koper, and Brandon Schmandt. Combining deep learning with physics based features in explosion-earthquake discrimination. *Geophysical Research Letters*, 49(13):e2022GL098645, 2022.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, 2012.

[29] Bonhwa Ku, Gwantae Kim, Jae-Kwang Ahn, Jimin Lee, and Hanseok Ko. Attention-based convolutional neural network for earthquake event classification. *IEEE Geoscience and Remote Sensing Letters*, 18(12):2057–2061, 2020.

[30] Wenhuan Kuang, Congcong Yuan, and Jie Zhang. Real-time determination of earthquake focal mechanism via deep learning. *Nature communications*, 12(1):1–8, 2021.

[31] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2023.

[32] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–11, 2023.

[33] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[34] Yunyue Elita Li, Daniel O'malley, Greg Beroza, Andrew Curtis, and Paul Johnson. Machine learning developments and applications in solid-earth geosciences: Fad or future?, 2023.

[35] Zefeng Li, Men-Andrin Meier, Egill Hauksson, Zhongwen Zhan, and Jennifer Andrews. Machine learning seismic wave discrimination: Application to earthquake early warning. *Geophysical Research Letters*, 45(10):4773–4779, 2018.

[36] Lisa Linville, Kristine Pankow, and Timothy Draelos. Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, 46(7):3643–3651, 2019.

[37] Min Liu, Miao Zhang, Weiqiang Zhu, William L Ellsworth, and Hongyi Li. Rapid characterization of the july 2019 ridgecrest, california, earthquake sequence from raw seismic data using machine-learning phase picker. *Geophysical Research Letters*, 47 (4):e2019GL086189, 2020.

[38] Anthony Lomax, Alberto Michelini, and Dario Jozinović. An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network. *Seismological Research Letters*, 90(2A):517–529, 2019.

[39] Zhengjing Ma and Gang Mei. Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Science Reviews*, 223: 103858, 2021. ISSN 0012-8252. doi: https://doi.org/10.1016/j.earscirev.2021.103858. URL https://www.sciencedirect.com/science/article/pii/S0012825221003597.

[40] Ian W McBrearty and Gregory C Beroza. Earthquake phase association with graph neural networks. *Bulletin of the Seismological Society of America*, 113(2):524–547, 2023.

[41] Ian W McBrearty, Andrew A Delorey, and Paul A Johnson. Pairwise association of seismic arrivals with convolutional neural networks. *Seismological Research Letters*, 90(2A):503–509, 2019.

[42] Ian W McBrearty, Joan Gomberg, Andrew A Delorey, and Paul A Johnson. Earthquake arrival association with backprojection and graph theoryearthquake arrival association with backprojection and graph theory. *Bulletin of the Seismological Society of America*, 109(6):2510–2531, 2019.

[43] George A. Miller. WordNet: a lexical database for english. *Commun. ACM*, 38(11): 39–41, nov 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

[44] Takahito Mitsui and Niklas Boers. Seasonal prediction of indian summer monsoon onset with echo state networks. *Environmental Research Letters*, 16(7):074024, 2021.

[45] S Mostafa Mousavi and Gregory C Beroza. Bayesian-deep-learning estimation of earthquake location from single-station observations. *arXiv preprint arXiv:1912.01144*, 2019.

[46] S Mostafa Mousavi and Gregory C Beroza. Deep-learning seismology. *Science*, 377 (6607):eabm4470, 2022.

[47] S Mostafa Mousavi, Weiqiang Zhu, Yixiao Sheng, and Gregory C Beroza. Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Scientific reports*, 9(1):1–14, 2019.

[48] S Mostafa Mousavi, William L Ellsworth, Weiqiang Zhu, Lindsay Y Chuang, and Gregory C Beroza. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1):1–12, 2020.

[49] Jannes Münchmeyer, Dino Bindi, Ulf Leser, and Frederik Tilmann. Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. *Geophysical Journal International*, 226(2):1086–1104, 2021.

[50] Jannes Münchmeyer, Jack Woollam, Andreas Rietbrock, Frederik Tilmann, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, et al. Which picker fits my data? a quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, 127(1): e2021JB023499, 2022.

[51] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In *Advances in Neural Information Processing Systems*, volume 35, pages 21455–21469, 2022.

[52] Artemii Novoselov, Peter Balazs, and Götz Bokelmann. Sedenoss: Separating and denoising seismic signals with dual-path recurrent neural network architecture. *Journal of Geophysical Research: Solid Earth*, 127(3):e2021JB023183, 2022.

[53] Esteban Pardo, Carmen Garfias, and Norberto Malpica. Seismic phase picking using convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9): 7086–7092, 2019.

[54] Thibaut Perol, Michaël Gharbi, and Marine Denolle. Convolutional neural network for earthquake detection and location. *Science Advances*, 4(2):e1700578, 2018.

[55] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[57] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[58] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[59] Zachary E Ross, Men-Andrin Meier, and Egill Hauksson. P wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, 123(6):5120–5129, 2018.

[60] Zachary E Ross, Men-Andrin Meier, Egill Hauksson, and Thomas H Heaton. Generalized seismic phase detection with deep learningshort note. *Bulletin of the Seismological Society of America*, 108(5A):2894–2901, 2018.

[61] Zachary E Ross, Yisong Yue, Men-Andrin Meier, Egill Hauksson, and Thomas H Heaton. Phaselink: A deep learning approach to seismic phase association. *Journal of Geophysical Research: Solid Earth*, 124(1):856–869, 2019.

[62] Bertrand Rouet-Leduc, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J Humphreys, and Paul A Johnson. Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, 44(18):9276–9282, 2017.

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

[64] Takayuki Shinohara and Hidetaka Saomoto. Ground-displacement forecasting from satellite image time series via a koopman-prior autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2025.

[65] Takayuki Shinohara and Hidetaka Saomoto. Vit-koop: Vision-transformer–koopman operators for efficient time-series forecasting of earth-observation data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2025.

[66] Parisa Shokouhi, Vrushali Girkar, Jacques Rivière, Srisharan Shreedharan, Chris Marone, C Lee Giles, and Daniel Kifer. Deep learning can predict laboratory quakes from active source seismic data. *Geophysical Research Letters*, 48(12): e2021GL093187, 2021.

[67] Xu Si, Xinming Wu, Zefeng Li, Shenghou Wang, and Jun Zhu. Multi-task multi-station earthquake monitoring: An all-in-one seismic phase picking, location, and association network (plan). *arXiv preprint arXiv:2306.13918*, 2023.

[68] Ryousei Takano, Shinichiro Takizawa, Yusuke Tanimura, Hidemoto Nakada, and Hirotaka Ogawa. Abci 3.0: Evolution of the leading ai infrastructure in japan, 2024. URL https://arxiv.org/abs/2411.09134.

[69] Xiao Tian, Wei Zhang, Xiong Zhang, Jie Zhang, Qingshan Zhang, Xiangteng Wang, and Quanshi Guo. Comparison of single-trace and multiple-trace polarity determination for surface microseismic data using deep learning. *Seismological Research Letters*, 91(3):1794–1803, 2020.

[70] Takahiko Uchide. Focal mechanisms of small earthquakes beneath the japanese islands based on first-motion polarities picked using deep learning. *Geophysical Journal International*, 223(3):1658–1671, 2020.

[71] Martijn van den Ende, Itzhak Lior, Jean-Paul Ampuero, Anthony Sladen, André Ferrari, and Cédric Richard. A self-supervised deep learning approach for blind denoising and waveform coherence enhancement in distributed acoustic sensing data. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[72] Martijn PA van den Ende and J-P Ampuero. Automated seismic source characterization using deep graph neural networks. *Geophysical Research Letters*, 47(17): e2020GL088690, 2020.

[73] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[74] Jian Wang, Zhuowei Xiao, Chang Liu, Dapeng Zhao, and Zhenxing Yao. Deep learning for picking seismic arrival times. *Journal of Geophysical Research: Solid Earth*, 124 (7):6612–6624, 2019.

[75] Kun Wang, Christopher W Johnson, Kane C Bennett, and Paul A Johnson. Predicting fault slip via transfer learning. *Nature Communications*, 12(1):7319, 2021.

[76] Kun Wang, Christopher W Johnson, Kane C Bennett, and Paul A Johnson. Predicting future laboratory fault friction through deep learning transformer models. *Geophysical Research Letters*, 49(19):e2022GL098233, 2022.

[77] Tiantong Wang, Daniel Trugman, and Youzuo Lin. Seismogen: Seismic waveform synthesis using gan with application to seismic data augmentation. *Journal of Geophysical Research: Solid Earth*, 126(4):e2020JB020077, 2021.

[78] Jonathan A Weyn, Dale R Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019.

[79] Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. CiT: Curation in training for effective vision-language data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15180–15189, 2023.

[80] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

[81] Lei Yang, Xin Liu, Weiqiang Zhu, Liang Zhao, and Gregory C Beroza. Toward improved urban earthquake monitoring through deep-learning-based noise suppression. *Science advances*, 8(15):eabl3564, 2022.

[82] Shaobo Yang, Jing Hu, Haijiang Zhang, and Guiquan Liu. Simultaneous earthquake detection on multiple stations via a convolutional neural network. *Seismological Research Letters*, 92(1):246–260, 2021.

[83] Keisuke Yano, Takahiro Shiina, Sumito Kurata, Aitaro Kato, Fumiyasu Komaki, Shin'ichi Sakai, and Naoshi Hirata. Graph-partitioning based convolutional neural network for earthquake detection using a seismic array. *Journal of Geophysical Research: Solid Earth*, 126(5):e2020JB020269, 2021.

[84] Ziye Yu and Weitao Wang. Fastlink: a machine learning and gpu-based fast phase association method and its application to yangbi m s 6.4 aftershock sequences. *Geophysical Journal International*, 230(1):673–683, 2022.

[85] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. *arXiv preprint arXiv:2112.02413*, 2021.

[86] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022.

[87] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822*, 2024.

[88] Xiong Zhang, Jie Zhang, Congcong Yuan, Sen Liu, Zhibo Chen, and Weiping Li. Locating induced earthquakes with a network of seismic stations in oklahoma via a deep learning method. *Scientific reports*, 10(1):1–12, 2020.

[89] Xitong Zhang, Will Reichard-Flynn, Miao Zhang, Matthew Hirn, and Youzuo Lin. Spatio-temporal graph convolutional networks for earthquake source characterization. *Journal of Geophysical Research: Solid Earth*, page e2022JB024401, 2022.

[90] Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, pages 1–7, 2023.

[91] Zhu, Lihua Fang, Fajun Miao, Liping Fan, Ji Zhang, and Zefeng Li. Deep learning and transfer learning of earthquake and quarry-blast discrimination: Applications to southern california and eastern kentucky. *Authorea Preprints*, 2022.

[92] Lijun Zhu, Zhigang Peng, James McClellan, Chenyu Li, Dongdong Yao, Zefeng Li, and Lihua Fang. Deep learning for seismic phase detection and picking in the aftershock zone of 2008 mw7. 9 wenchuan earthquake. *Physics of the Earth and Planetary Interiors*, 293:106261, 2019.

[93] Weiqiang Zhu and Gregory C Beroza. Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 10 2018.

[94] Weiqiang Zhu, S Mostafa Mousavi, and Gregory C Beroza. Seismic signal denoising and decomposition using deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9476–9488, 2019.

[95] Weiqiang Zhu, Kai Sheng Tai, S Mostafa Mousavi, Peter Bailis, and Gregory C Beroza. An end-to-end earthquake detection method for joint phase picking and association using deep learning. *Journal of Geophysical Research: Solid Earth*, 127(3): e2021JB023283, 2022.

6 Supplementary Material

6.1 Training strategy

6.1.1 Pre-training GeologyCLIP using Contrastive Learning

During pre-training, GeologyCLIP's dual encoders are jointly optimized through contrastive learning, utilizing paired terrain images and their corresponding geological and geomorphologic information. For each batch of N pairs, the two encoder branches independently compute embeddings for the terrain images and geological information from their respective inputs. The contrastive learning objective aims to maximize the cosine similarity between embeddings of genuine terrain image-geological information pairs while minimizing similarity for the N^2 - N incorrect pairings [56]. This process is achieved by optimizing symmetric cross-entropy loss over the computed similarity scores. The result of this pre-training phase is a fine-tuned GeologyCLIP model consisting of both a spectrum encoder and an information encoder.

6.1.2 Adapting GeologyCLIP to Downstream Tasks

Following the pre-training phase, the terrain image encoder of our foundational model takes on a multifaceted role in the downstream task of geohazard (landslide) detection (Figure 3(b)). This critical task is fundamentally a scene classification problem, requiring the model to discern and categorize various geological hazards in terrain images.

To evaluate downstream tasks, we implemented three distinct strategies: fine-tuning, freezing, and scratch. Figure 2 delineates the training strategy and network architecture specifics for the geohazard classification task. The fine-tuning approach involves further training of the pre-trained ViT-Small-based image encoder (Finetune in Figure 2), while the freezing strategy maintains this encoder in a static form during training (Freeze in Figure 2). Conversely, the scratch model eschews the geology-based pre-trained model entirely (Scratch in Figure 2). To provide a comprehensive evaluation, we not only compared pre-trained models using different strategies but also retrained several baseline models. For the classification task, we employed the geohazard classification network [27] as a benchmark. Minor modifications were made to the network architecture to accommodate the length of the spectrum data, ensuring compatibility and fair comparisons.

6.2 Experimental Setup

The pre-training process employed a learning rate of 1e-4 and batch size of 192, with the model undergoing training for 100 epochs. Given the use of cross-entropy loss, we selected the model iteration that achieved the highest classification accuracy on the validation set (occurring at the 55th epoch) as our final pre-trained model. This optimized model was primed for deployment with its trained terrain image encoder ready to be utilized across a diverse range of downstream tasks. The learning environment used in our experiment was a parallel GPU server with four NVIDIA A100s.