EarthScape: A Multimodal Dataset for Surficial Geologic Mapping and Earth Surface Analysis

Matthew A. Massey
Kentucky Geological Survey
University of Kentucky
Lexington, KY 40506-0053
matthew.massey@uky.edu

Nusrat Munia
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0633
nusrat.munia@uky.edu

Abdullah-Al-Zubaer Imran
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0633
aimran@uky.edu

Abstract

Surficial geologic (SG) maps are critical for understanding Earth surface processes, supporting infrastructure planning, and addressing challenges related to climate change and natural hazards. Advancements in artificial intelligence (AI) and the proliferation of remote sensing imagery present an opportunity to transform SG mapping and overcome many of the limitations (e.g., labor-intensive, not scalable, etc.) of current workflows. We introduce **EarthScape**, *a new AI-ready multimodal dataset designed to advance SG mapping*. EarthScape integrates digital elevation models, aerial imagery, multi-scale terrain derivatives, and vector data for hydrologic and infrastructure features. We present a complete data processing pipeline to support reproducibility and benchmarking and report baseline results across single-modality, multi-scale, and multimodal configurations. Our experiments highlight the predictive value of terrain-derived features and the challenge of generalizing across geologically diverse regions.
Code: https://github.com/masseygeo/earthscape
Dataset: https://uknowledge.uky.edu/kgs_data/16/

1 Introduction

Surficial geologic (SG) maps depict the spatial distribution of mostly unconsolidated materials on the Earth's surface [Compton, 1985, Lisle et al., 2011, Pavlis and Mason, 2017]. These maps are essential to address a range of contemporary challenges, such as supporting economic and national security interests in critical mineral resources [Brimhall et al., 2005, Schulz, 2017], informing mitigation and response planning for geologic hazards [Alcántara-Ayala, 2002, Van Westen et al., 2003], and providing a foundation on which to understand climate change [Anderson and Ferree, 2010]. SG maps are also relevant to more practical applications like urban land use planning [Dai et al., 2001, Hokanson et al., 2019] and engineering projects [Keaton, 2013]. Despite the demonstrable social benefit and scientific merit [Bernknopf, 1993], detailed SG maps ($\geq$ 1:100,000-scale) cover less than 14% of the United States [U.S. Geological Survey, 2025].

The modern SG mapping workflow relies on manual fieldwork and visual interpretation of remote sensing (RS) imagery [Compton, 1985, Lisle et al., 2011]. Because these maps rely on visual interpretation and field annotation, they often reflect expert judgment rather than reproducible criteria,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

complicating efforts to scale mapping to national or global extents [Jones et al., 2004]. Finally, financial resources prevent large-scale initiatives to collect and compile SG map data, where one standard 1:24k-scale map may cost up to $123k [Berg, 2025]. In Kentucky alone, despite prioritizing SG mapping since 2004, fully mapping the remainder of the state at the current pace and workforce capacity would require over 175 years and an estimated $31 million [U.S. Geological Survey, 2024b].

Advancements in deep learning and the proliferation of RS imagery present an opportunity to transform SG mapping, overcoming current limitations. Recent studies have showcased the power of this type of approach to identify or segment landslides [Prakash et al., 2021, Wang et al., 2021, Liu et al., 2023] and sinkholes [Rafique et al., 2022]. Several studies have extended these ideas to segment maps of multiple classes of geologic materials [Behrens et al., 2018, Latifovic et al., 2018, Wang et al., 2021, Liu et al., 2024b]. These studies have demonstrated the utility of computer vision (CV) for geological investigations, but this area of research is still in its infancy.

The challenges presented by SG mapping align closely with current trends in CV research. Multimodal fusion of diverse geological datasets is necessary to accurately capture geologic map features [Baltrušaitis et al., 2018, Steyaert et al., 2023, Li and Wu, 2024]. The spatial dependencies of geological features resonate with recent advances in attention mechanisms [Dosovitskiy, 2020, Niu et al., 2021, Hassanin et al., 2024], multi-scale architectures [Chen et al., 2017, Fan et al., 2021, Liu et al., 2024a], and contrastive learning frameworks [Chen et al., 2020, Le-Khac et al., 2020, Song et al., 2024] that capture context and structural relationships. Moreover, the scale-dependent and highly localized nature of geological processes demands robust methods for handling extreme class imbalance and ensuring geographic generalizability [Ghosh et al., 2024, Lin, 2017].

The rapid progress in CV has been driven primarily due to the availability of large-scale, standardized datasets. General-purpose benchmarks, such as ImageNet [Deng et al., 2009] and COCO [Lin et al., 2014], have catalyzed advances in classification, detection, and segmentation by offering vast repositories of labeled imagery and clear evaluation protocols. However, performance on real-world, domain-specific tasks often plateaus without datasets that reflect their unique characteristics, sensing modalities, and physical constraints. In the geospatial domain, several specialized datasets have emerged for land cover classification and urban scene analysis [Schmitt et al., 2019, Cordts et al., 2016, Demir et al., 2018, Van Etten et al., 2018, Sumbul et al., 2019]. But they are primarily focused on detecting anthropogenic features and land use. Only a single publicly available geologic dataset exists, and it is limited to landslide detection from a narrow set of features [Ji et al., 2020]. This underscores a critical gap in datasets tailored for Earth surface processes.

EarthScape is a multimodal dataset developed for SG mapping, with broad applicability to planetary surface analysis. It integrates publicly available overhead RGB and near-infrared (NIR) imagery, digital elevation models (DEMs), geomorphometric terrain features derived at multiple spatial scales, and transportation and hydrological networks from vector geographic information system (GIS) sources. *This multimodal, multi-scale design captures the complexity of Earth surface (ES) processes and provides a robust benchmark for advancing multimodal learning, geospatial vision, and geological analysis.* Our specific contributions are summarized as follows:

- EarthScape, the first AI-ready dataset specifically designed for SG mapping and ES analysis.
- Design and release of a rich set of input features that span multiple spatial scales and modalities, enabling models to learn representations of surface shape that generalize better across local and regional terrain variations.
- Establishing baseline benchmarks for multilabel classification using both unimodal and multimodal configurations. These include individual modality tests, multi-scale fusion within a single modality, and cross-modality fusion strategies.

2 Related work

SG Mapping with Machine Learning: SG mapping focuses on unconsolidated materials formed by active surface processes such as weathering, erosion, sediment transport, and deposition [Compton, 1985, Lisle et al., 2011, Pavlis and Mason, 2017]. These materials are closely tied to landform structure and surface morphology, as terrain shape governs the energy available to drive these processes and influences the way sediments are generated, transported, and deposited [Odeh et al., 1991, Schomberg

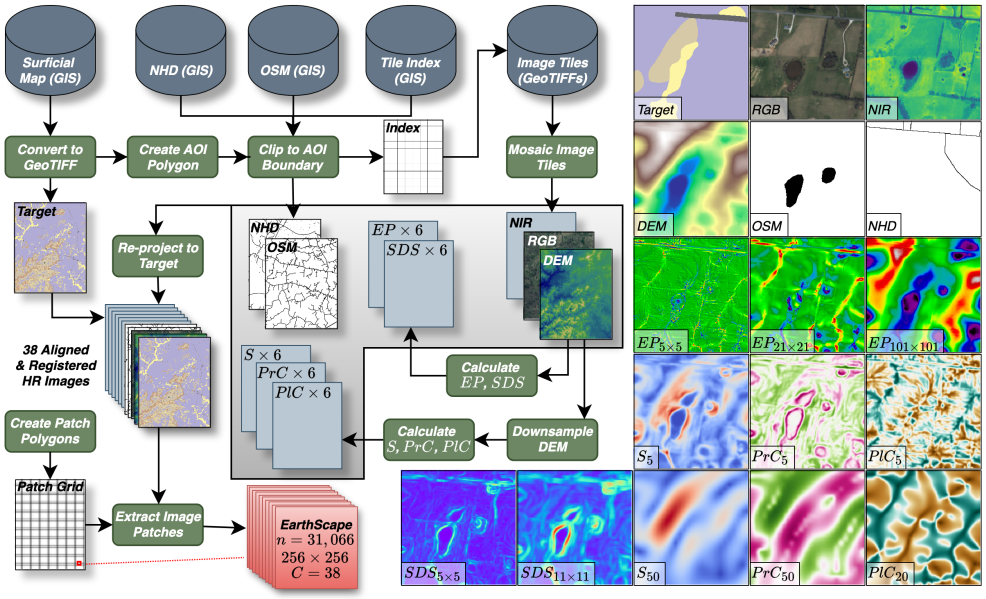

Figure 1: EarthScape data processing pipeline (left) and selected modalities from a single 256×256 patch (right). The SG map target is rasterized and used to define the area of interest (AOI), from which all predictive features (DEM, RGB+NIR imagery, NHD hydrology, and OSM infrastructure) are clipped and aligned. Terrain derivatives are then computed from the DEM at multiple spatial scales. A regular grid is applied to extract 38 co-registered channels per patch.

et al., 2005, Brigham and Crider, 2022]. Several studies have leveraged this terrain-material relationship using traditional machine learning methods, including logistic regression, random forests, and support vector machines, for classification or pixel-wise segmentation of single-class features (e.g., landslides, sinkholes) [Kirkwood et al., 2016, Zhu and Pierskalla Jr, 2016, Crawford et al., 2021] or multiclass geologic maps [Cracknell and Reading, 2014, Johnson and Haneberg, 2025]. However, these models rely on hand-crafted features, are limited to small geographic extents, and fail to generalize beyond the training region.

More recently, deep learning approaches using convolutional neural networks (CNNs) and CNN-transformer hybrids have been applied to these tasks [Prakash et al., 2021, Ji et al., 2020, Liu et al., 2023, Latifovic et al., 2018, Zhou et al., 2023, Rafique et al., 2022]. While these models better capture spatial dependencies critical to geologic interpretation [Bishop et al., 1998, Behrens et al., 2018], they remain constrained to narrow geographic domains, lack publicly available datasets or reproducible pipelines, and often rely on limited input modalities.

Remote Sensing Datasets: RS benchmarks such as SpaceNet [Van Etten et al., 2018], xView [Lam et al., 2018], and Functional Map of the World [Christie et al., 2018] provide high-resolution satellite imagery annotated for object detection and scene classification in urban environments. These datasets are optimized for anthropogenic features such as roads, buildings, and vehicles, and are widely used for infrastructure monitoring and disaster response. Other RS datasets like BigEarthNet [Sumbul et al., 2019], DeepGlobe [Demir et al., 2018], and SEN12MS [Schmitt et al., 2019], extend the domain to land cover classification and segmentation using multispectral or synthetic aperture radar (SAR) imagery. However, these datasets target coarse semantic categories like vegetation or developed areas, rather than physical topographic characteristics. These datasets lack representations of Earth's surface, which are essential for interpreting geological processes.

Multimodal Learning for Geologic Tasks: Multimodal learning has become a central paradigm in RS and geospatial CV, where combining diverse data sources like optical imagery, SAR, and DEMs can enhance model robustness through complementary information [Astruc et al., 2024, Bi et al., 2022, Jain et al., 2022, Han et al., 2024]. In geological applications, this often involves pairing overhead RGB imagery with DEMs, fused using early- or mid-level strategies [Prakash et al., 2021, Ji et al., 2020, Liu et al., 2023, Latifovic et al., 2018, Zhou et al., 2023, Rafique et al., 2022]. These

modalities have proven effective for detecting recent geomorphic events such as landslides, where strong topographic and visual signals are present. However, model performance often deteriorates when features are older, vegetated, or eroded, limiting their interpretability and transferability [Ji et al., 2020, Liu et al., 2023, Zhou et al., 2023].

Several studies have explored additional modalities such as elevation contours [Zhou et al., 2023], geochemical field data [Latifovic et al., 2018, Wang et al., 2021], and aeromagnetic imagery [Liu et al., 2024b]. While successful, these studies were site-specific, and the datasets are not commonly available or standardized for machine learning workflows. Rafique et al. [2022] evaluated several elevation-based parameters, including DEMs, slope, and shaded relief, finding that raw DEMs performed best. However, their geographically-limited validation suggests that models may have relied on absolute elevation rather than generalizable terrain patterns.

3 EarthScape Dataset

3.1 Composition and Features

Surficial Geologic Maps: The Kentucky Geological Survey has conducted high-resolution SG mapping since 2004, targeting rapidly developing regions and transportation corridors across the state. Mapping is performed at a scale of 1:24,000 or finer, widely considered the gold standard for detailed geological surveys. The EarthScape dataset currently includes SG map data from Warren and Hardin Counties [Buchanan et al., 2023, Massey et al., 2023, Swallom et al., 2023, Massey et al., 2024, Hodelka et al., 2024, Swallom et al., 2024, Bottoms et al., 2021, Massey et al., 2021], which provide the multilabel targets and segmentation masks (Fig. 1). Seven SG map units are represented, capturing three dominant surface processes: fluvial deposition, gravitational transport, and in-situ weathering. These include *alluvium (Qal)* and *terrace deposits (Qat)* from river activity; *alluvial fans (Qaf)* associated with debris flow hazards; *colluvium (Qc)* and *colluvial aprons (Qca)* from hillslope processes; *residuum (Qr)* from bedrock weathering; and *artificial fill (af1)* from anthropogenic modification. All maps are publicly available as vector polygons in ESRI file geodatabase format. See the supplement for detailed unit descriptions.

Aerial imagery and DEM: The KyFromAbove program has been acquiring high-resolution aerial imagery and DEMs for the state of Kentucky, USA since 2010 [Commonwealth of Kentucky, 2024]. Aerial imagery consists of RGB and NIR channels with a 6-inch spatial resolution. Its utility is in identifying anthropogenic features (such as af1) that are easily distinguished from natural landscapes (Fig. 1). The NIR band further enhances the detection of hydrological features, such as alluvial deposits and stream channels, by highlighting vegetation patterns that can indicate water presence or recent sediment deposition (Fig. 1). However, the utility of aerial RGB and NIR in delineating detailed SG map units is limited. In contrast, the DEM, generated from airborne LiDAR with a 5ft/pixel spatial resolution, is a critical feature for SG mapping and ES analysis (Fig. 1). Both the DEM and the aerial imagery are available as publicly accessible GeoTIFF tiles.

Geomorphometric Terrain Features: The DEMs provide a foundation for deriving five key terrain features widely used in geomorphometric analysis and essential for delineating SG units (Fig. 1) [Florinsky, 2016]. These include: *slope (S)* measures terrain steepness; *profile curvature (PrC)* and *planform curvature (PlC)* are directional second derivatives capturing flow acceleration and divergence; *elevation percentile (EP)* is a relative topographic position metric; *standard deviation of slope (SDS)* is a measure of terrain roughness quantifying local variability of slope angles. Each feature was calculated at multiple spatial scales to capture both localized and regional landform structure (see supplement for detailed definitions and scale parameters).

Hydrography and Infrastructure: To support downstream tasks involving fluvial and anthropogenic processes, EarthScape includes vector data for hydrographic and infrastructure features (Fig. 1). Stream centerlines and waterbody polygons from the U.S. Geological Survey's National Hydrography Dataset (NHD) [U.S. Geological Survey, 2024a] provide context for identifying alluvial units within stream valleys. Road and railway centerlines from OpenStreetMap (OSM) [OpenStreetMap contributors, 2024] delineate areas modified by human activity, such as artificial fill. These features also help characterize geologic disturbance near infrastructure, including slope undercutting and landslide susceptibility. Both datasets are included as binary raster channels aligned to the patch grid.

3.2 Data Processing

Targets: Each SG map was downloaded as a vector GIS geodatabase, the relevant feature class extracted, and the vector polygons inspected for topological correctness, ensuring no overlaps, no gaps, and valid polygon geometries (Fig. 1). The validated data was saved as a standalone GeoJSON file, which was then used to generate a boundary polygon defining the area of interest (AOI) for clipping and extracting relevant portions of other datasets. SG target classes were encoded with ordinal values in the GeoJSON. Finally, the vector GeoJSON was rasterized to a GeoTIFF image with a 5ft/pixel spatial resolution, matching the native resolution of the DEM.

Features: Vector datasets, including NHD, OSM, and the KyFromAbove tile index, were downloaded, clipped to the target AOI, and saved as standalone GeoJSON files (Fig. 1). NHD stream centerlines and waterbody polygons, and OSM road and railway centerlines were rasterized into two binary GeoTIFFs representing hydrography and infrastructure, respectively. The KyFromAbove tile index defines the locations of aerial RGB, NIR, and DEM data tiles across Kentucky. Using the AOI, relevant locations were selected and the corresponding image tiles downloaded (Fig. 1). DEM tiles were merged into a single GeoTIFF mosaic at 5ft/pixel resolution. RGB and NIR imagery underwent similar processing, with additional downsampling from 6in/pixel to 5ft/pixel resolution.

Five terrain features were calculated at six different spatial scales directly from the DEM mosaic (Fig. 1). S, PrC, and PlC were created using 5×5 kernels applied to the original 5ft/pixel DEM and five additional DEMs downsampled with cubic convolution to resolutions of 10, 20, 50, 100, and 200ft/pixel. A Gaussian filter was applied to each downsampled DEM to smooth potential artifacts, the relevant terrain feature was calculated, then upsampled back to the original resolution of 5ft/pixel using cubic convolution, and another Gaussian filter was applied to minimize resampling artifacts. SDS and EP were calculated using six kernel sizes of 5×5, 11×11, 21×21, 51×51, 101×101, and 201×201 pixels, applied only to the original 5ft/pixel DEM. These kernel sizes capture receptive fields similar to those represented by the coarser-resolution DEMs used for S, PrC, and PlC, but are better suited for SDS and EP due to their reliance on the number of neighbors.

Spatial Alignment and Registration: The target SG map GeoTIFF images served as the spatial reference for aligning all other features in the dataset (Fig. 1). Once each feature was collected and compiled into its respective GeoTIFF image file, they were reprojected to align with the reference image coordinates using cubic convolution interpolation. All images were checked to ensure that their bounding coordinates and spatial resolutions were identical across all other images.

Image Patches: Vector polygon patches were systematically constructed in a grid pattern to cover the target AOI using the same coordinate reference system as the target GeoTIFF (Fig. 1). Each grid cell polygon was assigned a unique patch ID, and then all patches were saved as a GeoJSON file. Each grid cell polygon patch was constructed so that it covers an area of exactly 1280×1280 feet (256×256 pixels), overlaps adjacent patches by 50%, and is fully contained within the target AOI. Each cell was assigned a unique patch ID and used to extract 38 corresponding channels, including target mask, aerial RGB and NIR, DEM, the five terrain features calculated at six scales, NHD, and OSM. Target masks were then used to extract one-hot encoded class labels and the proportional areas occupied by each class within each patch.

3.3 Dataset and Statistics

EarthScape currently comprises 31,018 image patch locations, each measuring 256×256 pixels with 50% spatial overlap with adjacent locations (Fig. 1). Each patch contains 38 channels, stored as individual 32-bit float GeoTIFF files with embedded geospatial metadata. Patch geometries are defined in an accompanying GeoJSON file to support spatial querying and GIS-based evaluation. The dataset spans two regions in Kentucky: a large contiguous subset of 23,566 locations in Warren County (Fig. 2A) and 7,452 locations in Hardin County. This geographic partitioning enables cross-region generalization studies and domain adaptation experiments, with additional regions planned as new SG maps become available. The dataset exhibits significant spatial and statistical heterogeneity. Most patches contain multiple SG units, with up to six unique classes per patch, and pronounced spatial variability across the AOIs in class co-occurrence (Fig. 2A, 2D). The dataset is highly imbalanced, with common units like Qr dominating the distribution and minority classes Qaf and Qat appearing infrequently (Fig. 2B). Intra-patch complexity is further reflected in the proportional area each class occupies per patch (Fig. 2C), with many units contributing small but

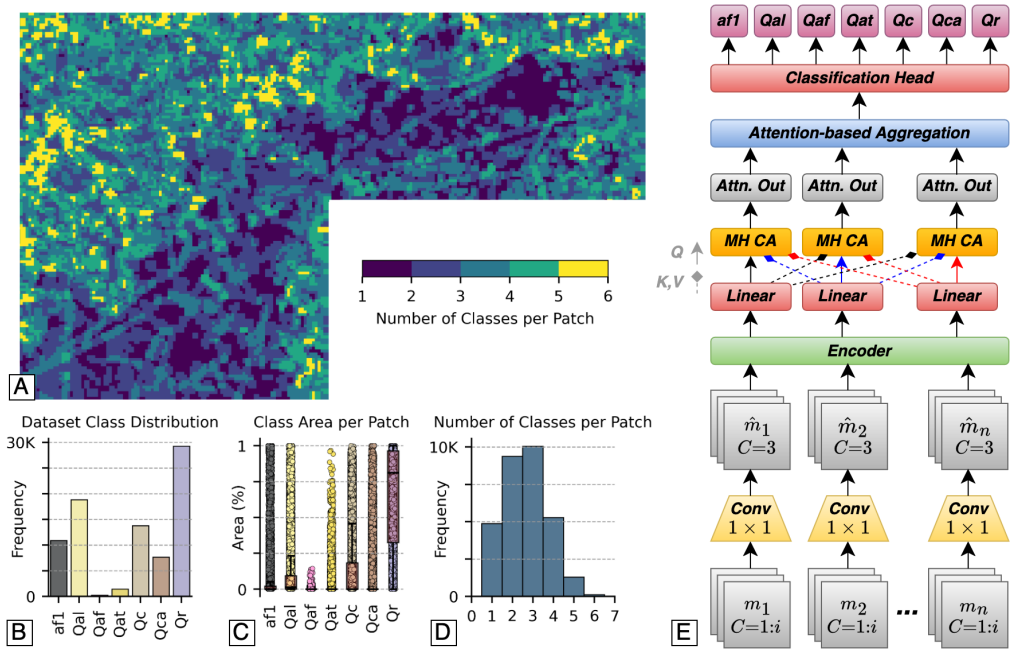

Figure 2: EarthScape dataset characteristics (A–D) and SGMap-Net architecture (E). A. Choropleth map of Warren County showing the number of classes per patch, illustrating spatial heterogeneity and multilabel complexity. B. Dataset-wide class distribution, highlighting significant class imbalance. C. Proportional area of each class per patch, showing that many patches include low-exposure classes, increasing classification difficulty. D. Histogram of class counts per patch, further illustrating multilabel and intra-patch complexity. E. SGMap-Net architecture comprising a standardization module, shared encoder, and multilabel classification head. Fusion is implemented via early channel stacking and intermediate attention-based strategies.

meaningful fractions to the total label. These properties make EarthScape well-suited for evaluating multilabel models under realistic geological class imbalance and spatial heterogeneity.

4 Experiments

4.1 Methods

Task Definition: We formulate SG mapping as a multilabel classification task over multimodal geospatial inputs. Each input sample corresponds to a 256×256 image patch with co-registered modalities and a label vector indicating the presence or absence of each of the SG units. Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ denote the dataset, where each $x_i = m_1, m_2, \ldots, m_n$ is a collection of n modality-specific input tensors (e.g., DEM, EP, PlC, etc.) and each modality m_i can have multiple scaled images that we consider as channels C_i. The $y_i \in 0, 1^K$ is a binary label vector over $K = 7$ classes. The model learns a mapping $f : X \rightarrow [0, 1]^K$ to predict per-class probabilities, enabling multi-class label assignment for each patch. This formulation allows us to systematically evaluate how different modality combinations contribute to geologic feature recognition and serves as a tractable benchmark for future tasks such as semantic segmentation.

Surficial Geologic Mapping Network (SGMap-Net): Our dataset comprises multiple geospatial image modalities with varying channel dimensionalities (e.g., RGB, DEM, terrain derivatives), which we aim to classify into seven geologic classes. To effectively integrate the complementary information across modalities, we propose a fusion-based model, SGMap-Net, which incorporates both early and intermediate fusion mechanisms to capture fine-grained spatial cues and high-level semantic relationships. Figure 2 (E) illustrates the overall architecture of SGMap-Net, which consists of three key components: a standardization module, a feature extractor, and a classification head. As part of our early fusion strategy, we first stack all channels of each modality m_i and then apply a

1×1 convolution followed by batch normalization and ReLU activation to standardize the input to a common channel dimension $C = 3$. This ensures compatibility with a shared encoder while preserving modality-specific spatial patterns through independent convolutions.

$$\hat{m_i} = ReLU(BN(Conv1 \times 1(m_i))). \tag{1}$$

Each standardized modality $\hat{m_i}$ is passed through a shared encoder to extract feature maps $f_{m_i} = Encoder(\hat{m_i})$. The shared encoder is initialized with ImageNet-pretrained weights, and we experiment with ResNeXt-50 [Xie et al., 2017] and Vision Transformer (ViT-B/16) [Dosovitskiy, 2020] architectures. Next, each feature vector f_{m_i} is projected into a common latent space of dimension d using a fully connected layer and augmented with a learnable modality embedding e_i to get the final representaions $z_i = f_{m_i} + e_i$. Then we apply modality-specific multi-head attention (MHA) [Vaswani et al., 2017] mechanisms to enable intermediate fusion across modalities. For each modality m_i, attention is computed using z_i as the query (Q), and the embeddings from all other modalities as keys (K) and values (V).

$$a_i = MHA(Q = z_i, K = [z_j]_{j \neq i}, V = [z_j]_{j \neq i}). \tag{2}$$

Next, we perform attention-weighted aggregation over the set of modality-specific attention outputs a. We begin by concatenating all outputs $A = [a_i]$. To determine the relative importance of each modality, we apply a learnable linear projection v_i followed by a softmax operation to obtain attention weights $w = Softmax(v^T A)$. The final fused representation is then computed using these weights, $z_{fused} = \sum_{i=1}^{N} w_i a_i$. This attention-weighted aggregation adaptively emphasizes the most informative modalities for each sample. The fused embedding z_{fused} is then passed through a classification head consisting of two fully connected layers to predict the geologic class logits $\hat{y}$. In addition to our proposed attention-based fusion strategy, we evaluate two alternative approaches, cross-modality channel stacking and concatenation. We stack selected channels from different modalities, extract a joint representation using the encoder, and feed it into the classification head. In another approach, we concatenate the modality embeddings from the encoder and pass them directly to the classification head. These variants serve as comparative baselines to assess the impact of modality-aware attention in our fusion framework.

Data Splits and Selection: We define training, validation, and in-domain test splits using the Warren County subset. A total of 1,536 patch locations were randomly selected for the in-domain test set. Next, 768 non-intersecting locations were randomly sampled for validation. All remaining patches that did not intersect the in-domain test patches or validation patches were used for training (8,416). To evaluate geographic generalization to a geologically similar, but previously unseen region, we sampled an additional cross-domain test set of 1,536 patches from the Hardin County subset. While this split uses less than half of the available EarthScape patches, it was chosen to balance typical dataset proportions and maintain spatial independence between training and evaluation regions.

Training Procedure: All patches were normalized using modality-specific means and standard deviations computed over the in-domain dataset to ensure consistent input scaling. Data augmentation included random horizontal and vertical flips and 90◦ rotations, reflecting that geologic features are not orientation-dependent. Restricting rotations to right angles preserves label accuracy by preventing small classes along edges from being cropped due to padding. To address class imbalance, we adopted focal loss [Lin, 2017] with $\alpha = 0.25$ and $\gamma = 2.0$ for all experiments. Oversampling was tested but degraded performance, so training used the original distribution. Models were trained for 15 epochs using the Adam optimizer, a fixed learning rate of 0.001, and batch size of 16. The model with the lowest validation loss was used for testing. After training, label-wise thresholds were optimized for F1 on the validation set and applied to both in-domain (Warren) and cross-domain (Hardin) test sets. Performance was evaluated using per-class and macro-averaged accuracy, precision, recall, F1 score, average precision (AP), and area under the ROC curve (AUC). See the supplemental material for focal loss tuning, training time, and compute details.

4.2 Results and Discussion

Single Modality Benchmarks: We first evaluated single-modality models using SGMap-Net with both ResNeXt-50 and ViT-B/16 backbones (Table 1; also see supplemental material). Among the ResNeXt-50 models, the best in-domain performance was achieved using EP 51×51, EP 5×5, and EP 21×21, all of which outperformed the top ViT models. Most classes benefited from the relative elevation signal captured by EP, except for Qc, which performed best with slope (S),

Table 1: Macro-averaged F1 scores, precision, AUC, and accuracy on in-domain (Warren County, WC) and cross-domain (Hardin County, HC) test sets, along with differences between WC and HC (Δ) for each metric. Results are reported for the top three models in each experimental setting: single-modality, multi-scale fusion, and multimodal. Parentheses indicate the spatial scales used for fusion—ms: all spatial scales; s: smallest only; l: largest only. All models use a ResNeXt backbone and fusion with early channel stacking. The best and second-best scores in each column are indicated in **bold** and underlined, respectively. Additional results are provided in the supplemental material.

Model	F1			Precision			AUC			Accuracy		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP 51×51	0.651	0.380	0.271	0.612	0.382	0.230	0.876	0.663	0.213	0.862	0.818	0.044
EP 5×5	0.648	0.357	0.291	0.617	0.450	0.167	0.872	0.582	0.290	0.858	0.831	0.027
EP 21×21	0.645	0.384	0.261	**0.629**	0.455	0.174	0.877	0.695	0.182	0.860	0.828	0.032
EP (ms)	0.640	0.425	0.215	0.606	**0.556**	**0.050**	0.862	0.717	0.145	0.865	0.828	0.037
S (ms)	0.637	0.594	**0.043**	0.607	0.535	0.072	0.864	0.804	0.060	0.856	0.860	**-0.004**
SDS (ms)	0.636	0.588	0.048	0.588	0.509	0.079	0.878	0.792	0.086	0.846	0.839	0.007
EP+S+SDS (ms)	**0.657**	**0.598**	0.059	0.626	0.546	0.080	0.882	0.806	0.076	**0.875**	**0.867**	0.008
EP+S+SDS (s)	0.641	0.568	0.073	0.606	0.531	0.075	0.848	**0.812**	**0.036**	0.865	0.856	0.009
EP+S+SDS (l)	0.626	0.582	0.044	0.588	0.529	0.059	**0.885**	**0.812**	0.073	0.858	0.852	0.006

aligning with its gravity-driven depositional process. Rare classes such as Qaf and Qat remained poorly identified across all experiments, suggesting the need for targeted loss strategies, additional training data, or synthetic augmentation Across all modalities, DEM-derived features EP, S, and SDS consistently outperformed raw DEM inputs, reinforcing the value of domain-specific terrain derivatives over implicit feature learning. Additionally, no single spatial kernel was optimal across all classes (e.g., af1 performed best with EP 11×11, while Qal favored EP 51×51), highlighting the importance of multi-scale inputs. Model performance declined under cross-domain testing in Hardin County. However, the ViT backbone showed better generalization ($\Delta F1_{ViT} = 0.018$ vs. $\Delta F1_{ResNeXt} = 0.043$). S and SDS exhibited the best cross-region transfer, while raw DEM inputs underperformed, likely due to overfitting region-specific topography.

Multi-scale Fusion: Unimodal experiments showed that no single spatial scale consistently performed best across all classes, with each SG map unit exhibiting distinct preferences for both modality and resolution. To explore whether combining spatial scales could improve performance, we evaluated the effects of fusing all six spatial resolutions for terrain features. Models were trained using both ResNeXt-50 and ViT-B/16 backbones (Table 1; also see supplemental material). Early fusion with channel stacking and ResNeXt yielded the most reliable results. For SDS, fusion slightly improved in-domain F1 (from 0.633 to 0.636) and enhanced cross-domain generalization ($\Delta F1$ decreased from 0.060 to 0.048). EP experienced a modest drop in in-domain F1 (0.651 to 0.640), but showed a substantial improvement in generalization ($\Delta F1$ decreased from 0.271 to 0.215), indicating that multi-scale fusion can mitigate its sensitivity to regional relief variation. Mid-level attention-based fusion underperformed in all cases, suggesting that early fusion is both more effective and more stable for combining spatial scales.

Multimodal Fusion: We evaluated multimodal fusion using both ResNeXt-50 and ViT-B/16 backbones, testing three fusion strategies: early fusion via channel stacking, mid-level attention-based fusion, and mid-level fusion via feature concatenation. We tested three modality configurations: (1) RGB + DEM, a common baseline in geospatial literature; (2) $EP + S + SDS$, selected based on unimodal performance; and (3) a full configuration combining DEM, RGB, EP, S, and SDS. For the $EP + S + SDS$ configuration, we tested three variants: one using all six spatial scales for each modality, one with three representative scales, and one with a single scale per modality (Table 1; also see supplemental material).

The best-performing model, according to both in-domain and cross-domain F1, was the $EP + S + SDS$ configuration using three selected spatial scales. While the in-domain macro-F1 improved only modestly (from 0.651 to 0.657), the cross-domain F1 increased dramatically from 0.380 to 0.598, reducing the generalization gap ($\Delta F1$) from 0.271 to just 0.059. This result underscores the strength of terrain-based, multi-scale inputs for learning region-invariant surface structure. Two reduced variants of the same modality set, using single scales per modality, ranked second and third, confirming the robustness of shape-centric features and the benefit of multi-scale representations. The next best performers were the $EP + S + SDS$ models using mid-level concatenation, followed by the full model (DEM+RGB+$S + SDS + EP$) with the same fusion strategy. In contrast, the

RGB+DEM configuration performed worst across all fusion methods and backbone combinations, reinforcing the limited generalizability of location-sensitive visual and elevation inputs. Despite its architectural sophistication, the attention-based fusion strategy consistently underperformed, suggesting that early fusion, and even simpler mid-level concatenation, can be more effective than complex attention mechanisms for integrating geospatial modalities in this domain.

5 Challenges and Limitations

Geographic Scope and Extensibility: EarthScape is currently limited to two regions in Kentucky, USA, reflecting both the availability of high-resolution SG maps and the natural variability of geological processes. While this constraint is typical of geospatial datasets, EarthScape was designed with a modular, patch-based architecture to support expansion. Kentucky is the only state in the region with SG maps of this terrain type available in standardized GIS formats, but the dataset curation workflow is broadly applicable. Ongoing efforts aim to incorporate additional regions and globally available features to improve geographic coverage and enable cross-domain model development.

Class Imbalance: The dataset includes seven SG units with highly imbalanced distributions that reflect real-world conditions. At the patch level, the number of co-occurring classes ranges from one to six, and many units occupy only a small fraction of a given patch. This results in both inter-class imbalance and intra-patch heterogeneity, offering a challenging testbed for multilabel and segmentation models that must handle sparse and noisy labels.

Geographic Generalization: SG varies significantly across regions due to localized geomorphic and depositional processes. Unlike many AI benchmarks that assume spatial homogeneity, EarthScape explicitly supports the evaluation of cross-region generalization. The inclusion of two distinct geographic subsets allows for benchmarking spatial transfer and domain adaptation performance under realistic conditions.

Multi-scale Complexity: SG features are scale-dependent, with different processes operating at distinct spatial resolutions. EarthScape includes terrain derivatives computed at six spatial scales, enabling models to learn both local and regional landform patterns. This supports research in multi-scale fusion, resolution-aware architectures, and feature relevance across spatial hierarchies.

Interpretation Variability: Although EarthScape relies on expert-labeled SG maps, geological interpretation is inherently uncertain, particularly in regions with limited field validation or ambiguous unit boundaries. This introduces structured label noise, which poses a challenge for supervised learning but also provides an opportunity to develop models that are robust to real-world uncertainty.

Temporal Inconsistency: The DEM, imagery, and vector layers were acquired between 2019 and 2024, introducing potential temporal mismatches across modalities. While this may reduce fine-grained alignment in some patches, it offers an opportunity to evaluate model resilience to asynchronous data and supports future work in temporal generalization.

6 Conclusions

We introduced EarthScape, a new AI-ready, multimodal benchmark dataset for SG mapping and ES analysis. EarthScape integrates aerial imagery, DEMs, multi-scale terrain derivatives, and GIS vector data, offering a unique resource for multimodal geospatial learning. The dataset presents real-world challenges like class imbalance, spatial heterogeneity, and geographic variability, making it a robust testbed for developing and evaluating AI models. Through baseline experiments, we established performance benchmarks across individual modalities, multi-scale fusion, and multimodal inputs, highlighting both the predictive value of terrain-based features and the difficulty of cross-region generalization in geologic settings. Designed as a living dataset, EarthScape is extensible in both geographic and modality space. Ongoing work includes expanding regional coverage, incorporating globally available features, and improving fusion strategies. Future directions include high-resolution segmentation tasks, pretraining pipelines, and region-specific fine-tuning to support applied geological workflows. By releasing data, code, and benchmarks, we aim to foster reproducible research, cross-disciplinary collaboration, and the development of generalizable models for geospatial AI.

References

I. Alcántara-Ayala. Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. *Geomorphology*, 47(2-4):107–124, 2002.

M. G. Anderson and C. E. Ferree. Conserving the stage: climate change and the geophysical underpinnings of species diversity. *PloS one*, 5(7):e11554, 2010.

G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

T. Behrens, K. Schmidt, R. A. MacMillan, and R. A. Viscarra Rossel. Multi-scale digital soil mapping with deep learning. *Scientific reports*, 8(1):15244, 2018.

R. C. Berg. *Economic Analysis of the Costs and Benefits of Geological Mapping in the United States of America from 1994 to 2019*. American Geosciences Institute, Alexandria, VA, 2025. URL `https://profession.americangeosciences.org/reports/geological-mapping-economics/`.

R. L. Bernknopf. *Societal value of geologic maps*, volume 1111. DIANE Publishing, 1993.

M. Bi, M. Wang, Z. Li, and D. Hong. Vision transformer with contrastive learning for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:738–749, 2022.

M. P. Bishop, J. F. Shroder Jr, B. L. Hickman, and L. Copland. Scale-dependent analysis of satellite imagery for characterization of glacier surfaces in the karakoram himalaya. *Geomorphology*, 21 (3-4):217–232, 1998.

A. Bottoms, M. Hammond, M. Massey, E. Morris, and M. McHugh. Surficial geologic map of the howe valley 7.5-minute quadrangle, central kentucky. *Kentucky Geological Survey Contract Report*, 13(43), 2021.

C. A. Brigham and J. G. Crider. A new metric for morphologic variability using landform shape classification via supervised machine learning. *Geomorphology*, 399:108065, 2022.

G. H. Brimhall, J. H. Dilles, and J. M. Proffett. The role of geologic mapping in mineral exploration. 2005.

W. Buchanan, M. Swallom, A. Bottoms, M. Massey, B. N. Hodelka, and E. Morris. Surficial geologic map of the rockfield 7.5-minute quadrangle, warren, logan, and simpson counties, kentucky. *Kentucky Geological Survey Contract Report*, 13(57), 2023.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

Commonwealth of Kentucky. Kyfromabove: Kentucky's elevation data aerial photography program, 2024. URL `https://kyfromabove.ky.gov`. Aerial RGB+NIR imagery and DEM. Accessed: 2024-08-01.

R. R. Compton. *Geology in the Field*. John Wiley & Sons, New York, 1985. Classic field geology manual covering mapping techniques8203;:contentReference[oaicite:41]index=41.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

M. J. Cracknell and A. M. Reading. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22–33, 2014.

M. M. Crawford, J. M. Dortch, H. J. Koch, A. A. Killen, J. Zhu, Y. Zhu, L. S. Bryson, and W. C. Haneberg. Using landslide-inventory mapping for a combined bagged-trees and logistic-regression approach to determining landslide susceptibility in eastern kentucky, usa. *Quarterly Journal of Engineering Geology and Hydrogeology*, 54(4):qjegh2020–177, 2021.

F. Dai, C. Lee, and X. Zhang. Gis-based geo-environmental evaluation for urban land-use planning: a case study. *Engineering geology*, 61(4):257–271, 2001.

I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

I. Florinsky. *Digital terrain analysis in soil science and geology*. Academic Press, 2016.

K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, 2024.

B. Han, S. Zhang, X. Shi, and M. Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27852–27862, 2024.

M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108:102417, 2024.

B. Hodelka, M. Massey, M. Swallom, S. Martin, C. Wells, and E. Morris. Surficial geologic map of the bristow 7.5-minute quadrangle, kentucky. Accepted for publication, 2024.

K. J. Hokanson, C. Mendoza, and K. Devito. Interactions between regional climate, surficial geology, and topography: characterizing shallow groundwater systems in subhumid, low-relief landscapes. *Water Resources Research*, 55(1):284–297, 2019.

U. Jain, A. Wilson, and V. Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*, 2022.

S. Ji, D. Yu, C. Shen, W. Li, and Q. Xu. Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides*, 17:1337–1352, 2020.

S. E. Johnson and W. C. Haneberg. Machine learning for surficial geologic mapping. *Earth Surface Processes and Landforms*, 50(1):e6032, 2025.

R. R. Jones, K. J. W. McCaffrey, R. W. Wilson, and R. E. Holdsworth. Digital field data acquisition: towards increased quantification of uncertainty during geological mapping. In A. Curtis and R. Wood, editors, *Geological Prior Information: Informing Science and Engineering*, volume 239, pages 43–56. Geological Society of London, 2004. doi: 10.1144/GSL.SP.2004.239.01.04. URL `https://doi.org/10.1144/GSL.SP.2004.239.01.04`.

J. R. Keaton. Engineering geology: fundamental input or random variable? In *Foundation Engineering in the Face of Uncertainty: Honoring Fred H. Kulhawy*, pages 232–253. 2013.

C. Kirkwood, M. Cave, D. Beamish, S. Grebby, and A. Ferreira. A machine learning approach to geochemical mapping. *Journal of Geochemical Exploration*, 167:49–61, 2016.

D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.

R. Latifovic, D. Pouliot, and J. Campbell. Assessment of convolution neural networks for surficial geology mapping in the south rae geological region, northwest territories, canada. *Remote sensing*, 10(2):307, 2018.

P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020.

H. Li and X.-J. Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

T. Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

R. J. Lisle, P. Brabham, and J. W. Barnes. *Basic Geological Mapping*. John Wiley & Sons, Chichester, UK, 5th edition, 2011. ISBN 9780470686348. Field guide to mapping geology, updated with modern techniques8203;:contentReference[oaicite:42]index=428203;:contentReference[oaicite:43]index=43.

S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26658–26668, 2024a.

X. Liu, Y. Peng, Z. Lu, W. Li, J. Yu, D. Ge, and W. Xiang. Feature-fusion segmentation network for landslide detection using high-resolution remote sensing images and digital elevation model data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

Y. Liu, J. Cheng, Q. Lü, Z. Liu, J. Lu, Z. Fan, and L. Zhang. Deep learning for geological mapping in the overburden area. *Frontiers in Earth Science*, 12:1407173, 2024b.

M. Massey, A. Bottoms, M. Hammond, E. Morris, and M. McHugh. Surficial geologic map of the sonora 7.5-minute quadrangle, central kentucky. *Kentucky Geological Survey Contract Report*, 13 (44), 2021.

M. Massey, M. Swallom, A. Bottoms, W. Buchanan, B. N. Hodelka, and E. Morris. Surficial geologic map of the hadley 7.5-minute quadrangle, warren county, kentucky. *Kentucky Geological Survey Contract Report*, 13(56), 2023.

M. Massey, M. Swallom, B. Hodelka, H. Hayes, C. Wells, S. Martin, and E. Morris. Surficial geologic map of the bowling green south 7.5-minute quadrangle, kentucky. Accepted for publication, 2024.

Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

I. Odeh, D. Chittleborough, and A. McBratney. Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma*, 49(1-2):1–32, 1991.

OpenStreetMap contributors. Openstreetmap road and railway centerlines. `https://www.openstreetmap.org`, 2024. Road and railway centerlines. Accessed: 2024-08-01.

T. L. Pavlis and K. A. Mason. The new world of 3d geologic mapping. *GSA Today*, 27(9):4–10, 2017. doi: 10.1130/GSATG313A.1. Discusses digital field mapping advances, including 3D photogrammetry and their impact on geologic mapping8203;:contentReference[oaicite:47]index=478203;:contentReference[oaicite:48]index=48.

N. Prakash, A. Manconi, and S. Loew. A new strategy to map landslides with a generalized convolutional neural network. *Scientific reports*, 11(1):9722, 2021.

M. U. Rafique, J. Zhu, and N. Jacobs. Automatic segmentation of sinkholes using a convolutional neural network. *Earth and Space Science*, 9(2):e2021EA002195, 2022.

M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.

J. D. Schomberg, G. Host, L. B. Johnson, and C. Richards. Evaluating the influence of landform, surficial geology, and land use on streams using hydrologic simulation modeling. *Aquatic Sciences*, 67:528–540, 2005.

K. J. Schulz. *Critical mineral resources of the United States: economic and environmental geology and prospects for future supply*. Geological Survey, 2017.

B. Song, Y. Xu, and Y. Wu. Vitcn: Vision transformer contrastive network for reasoning. arxiv prepr int. *arXiv preprint arXiv:2403.09962*, 2024.

S. Steyaert, M. Pizurica, D. Nagaraj, P. Khandelwal, T. Hernandez-Boussard, A. J. Gentles, and O. Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4):351–362, 2023.

G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

M. Swallom, M. Massey, W. Buchanan, B. N. Hodelka, H. Hayes, C. Wells III, and E. Morris. Surficial geologic map of the bowling green north 7.5-minute quadrangle, warren county, kentucky. *Kentucky Geological Survey Contract Report*, 13(55), 2023.

M. Swallom, B. Hodelka, M. Massey, H. Hayes, C. Wells, and E. Morris. Surficial geologic map of the smiths grove 7.5-minute quadrangle, kentucky. Accepted for publication, 2024.

U.S. Geological Survey. National hydrography dataset (nhd) – high resolution. `https://www.usgs.gov/national-hydrography`, 2024a. Stream centerlines and waterbody polygons. Accessed: 2024-08-01.

U.S. Geological Survey. Statemap award funding by fiscal year (1993–2024). `https://www.usgs.gov/media/files/statemap-award-funding-fiscal-year-1993-2024`, 2024b. Accessed May 2025.

U.S. Geological Survey. National geologic map database (ngmdb). `https://ngmdb.usgs.gov`, 2025. Accessed May 2025.

A. Van Etten, D. Lindenbaum, and T. M. Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

C. Van Westen, N. Rengers, and R. Soeters. Use of geomorphological information in indirect landslide susceptibility assessment. *Natural hazards*, 30:399–419, 2003.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Z. Wang, R. Zuo, and H. Liu. Lithological mapping based on fully convolutional network and multi-source geological data. *Remote Sensing*, 13(23):4860, 2021.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Y. Zhou, Y. Peng, W. Li, J. Yu, D. Ge, and W. Xiang. A hyper-pixel-wise contrastive learning augmented segmentation network for old landslide detection using high-resolution remote sensing images and digital elevation model data. *arXiv preprint arXiv:2308.01251*, 2023.

J. Zhu and W. P. Pierskalla Jr. Applying a weighted random forests method to extract karst sinkholes from lidar data. *Journal of Hydrology*, 533:343–352, 2016.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **The Abstract and Introduction 1 clearly state the dataset's purpose, composition, and intended use. These claims are supported by Sections 3 and 4 presented in the main text.**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **The paper includes a dedicated Challenges and Limitations section 5 discussing geographic scope, class imbalance, generalization, interpretive uncertainty, and temporal inconsistencies. These are acknowledged as challenges for model performance and dataset expansion.**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: **The paper includes standard mathematical formulations to define the task and model, but it does not present new theoretical results or proofs.**

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: **The paper provides a detailed description of the model architecture, loss functions, training setup, and data split methodology, including spatial constraints for training, validation, and cross-domain testing. The dataset is publicly available, along with all patch IDs, target labels, and 38 input channels for each patch. Baseline model implementations, training scripts, and evaluation code are provided in the accompanying GitHub repository. All experiments were run with fixed seeds and reported using standardized metrics, ensuring full reproducibility of the main results.**

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: **The paper provides public access to both the dataset and code, with links and documentation for data access, preprocessing, and benchmark experiments; we have temporarily held out one of our test sets for possible challenge competitions.**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **The paper describes all training and evaluation settings, including data splits, loss functions, hyperparameters, optimizer, and augmentation strategy.**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: **The paper reports aggregate performance metrics for all experiments, but does not include error bars or statistical significance testing. All models were trained and evaluated using fixed seeds and deterministic splits to ensure reproducibility. While we do not report variance across multiple runs, the primary goal of this work is to establish benchmark results for a new dataset, and the experimental setup is designed to support consistent replication and future extension.**

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: **The main paper does not report detailed compute environment information such as hardware specifications, memory, or runtime, but these details will be included in the Supplementary Material. While this information is not in the main text, all experiments were conducted on reproducible infrastructure, and model training scripts are available to ensure replicability.**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **This work adheres to the NeurIPS Code of Ethics. All data sources used in the EarthScape dataset are publicly available and government-released, including aerial imagery, DEMs, SG maps, and vector GIS data. No personal or private data**

were used, and no human or animal subjects were involved. The dataset is designed to support scientific understanding of Earth surface processes and does not pose foreseeable risks to individuals, communities, or the environment. Additionally, we emphasize transparency and reproducibility through open dataset access and detailed documentation.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: **The Introduction 1 highlights positive societal applications of surficial geologic mapping, while the Challenges and Limitations section 5 discusses risks related to model generalization, interpretation variability, and temporal discrepancies among the data sources.**

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: **The dataset is derived from publicly available, government-provided geospatial data and does not pose high risk for misuse.**

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **All third-party assets (e.g., NHD, OSM, KyFromAbove) are properly cited and credited.**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: **The EarthScape dataset and code are publicly released with detailed documentation on data structure, processing, and usage in both the paper and linked repositories.**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: **The paper does not involve any crowdsourcing or research involving human subjects.**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: **The paper does not involve research with human subjects and therefore does not require IRB approval.**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: **The research does not involve the use of large language models in any core methodological or experimental component.**

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.