EARTHSCAPE: A MULTIMODAL DATASET FOR SURFICIAL GEOLOGIC MAPPING AND EARTH SURFACE ANALYSIS

Anonymous authors
Paper under double-blind review

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. Current workflows are labor-intensive, subjective, and difficult to scale. We introduce EarthScape, an AI-ready multimodal dataset for advancing SG mapping and surface-aware geospatial learning. EarthScape integrates digital elevation models, aerial imagery, multi-scale terrain derivatives, and vector data for hydrologic and infrastructure features. We provide an end-to-end processing pipeline for reproducibility and report baseline benchmarks across single-modality, multi-scale, and multimodal configurations. Results show that terrain-derived features are highly predictive and that generalization across geologically diverse regions remains a key open challenge, positioning EarthScape as a benchmark for multimodal fusion and domain adaptation.

1 INTRODUCTION

Surficial geologic (SG) maps depict the spatial distribution of mostly unconsolidated materials on the Earth's surface (Compton, 1985). 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 & 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 cover less than 14% of the United States (U.S. Geological Survey, 2025), and coverage is even more limited globally.

The modern SG mapping workflow relies on manual fieldwork coupled with visual interpretation of remote sensing (RS) imagery (Compton, 1985; Lisle et al., 2011). Because SG maps depend on expert interpretation and annotation, they often reflect subjective judgment rather than reproducible criteria. Moreover, financial costs are prohibitive, with one standard 1:24k-scale map estimated at $123k (Berg, 2025). These limitations highlight the need for scalable, automated approaches.

Advancements in deep learning and the proliferation of RS imagery present an opportunity to transform SG mapping and overcome current limitations. Recent studies have demonstrated the potential of deep learning to identify or segment geologic hazards such as landslides (Prakash et al., 2021; Wang et al., 2021; Liu et al., 2023) and sinkholes (Rafique et al., 2022), and a few have extended these ideas to mapping multiple classes of geologic materials (Behrens et al., 2018; Latifovic et al., 2018; Wang et al., 2021; Liu et al., 2024b). While these works highlight the promise of computer vision (CV), they remain constrained by narrow scope, limited modality integration, and the absence of standardized benchmarks. Addressing these limitations requires methods that connect directly to ongoing advances in multimodal and multi-scale CV.

The challenges of SG mapping align closely with current directions in CV. Multimodal fusion of heterogeneous inputs is required to capture features invisible to any single modality (Baltrušaitis et al., 2018; Steyaert et al., 2023; Li & Wu, 2024). Strong spatial dependencies make it a natural testbed for attention mechanisms and multi-scale architectures (Dosovitskiy, 2020; Niu et al., 2021;

Fan et al., 2021; Hassanin et al., 2024; Liu et al., 2024a), while extreme class imbalance and geographic variability mirror open challenges in long-tail learning and domain adaptation (Lin, 2017; Ghosh et al., 2024). Beyond SG mapping, surface morphology is an underutilized signal across domains such as medical imaging where shape descriptors from CT or MRI improve disease prediction (Van Timmeren et al., 2020), autonomous navigation where terrain interpretation guides safe decision-making (Meng et al., 2023), and RS where benchmarks often underemphasize topography (Wang et al., 2025). SG mapping is not just a niche application but a challenging benchmark for multimodal, surface-aware learning.

The rapid progress in CV has been driven by 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 these are primarily focused on anthropogenic features and land use. Several geologic datasets have been introduced for landslide classification and segmentation (Ji et al., 2020), earthquake detection (Rege Cambrin & Garza, 2024), and flood mapping (Montello et al., 2022). While valuable, these resources focus on discrete events, leaving a critical gap in datasets tailored to continuous materials and landforms.

EarthScape is a multimodal dataset developed for SG mapping, with applicability to other surface-aware domains. It integrates publicly available RGB and near-infrared (NIR) imagery, digital elevation models (DEMs), DEM-derived shape-centric features computed at multiple scales, and transportation and hydrological networks from vector geographic information system (GIS) sources. This design captures the complexity of geologic processes and provides a robust benchmark for advancing multimodal learning, geospatial vision, and geological analysis. Our contributions are as follows:

- We present EarthScape, the first AI-ready, multimodal, multi-scale benchmark dataset specifically designed for SG mapping, developed as a living resource with broader applicability to surface analysis.
- EarthScape integrates vector GIS, imagery, elevation, and DEM-derived shape features across scales, explicitly capturing challenges of class imbalance, geographic heterogeneity, and surface morphology that make it an unusually challenging benchmark.
- We establish reproducible baselines across unimodal, multi-scale, and multimodal configurations, enabling systematic evaluation of fusion strategies, cross-region generalization, and future extensible research on Earth surface analysis.

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). 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 et al., 2005; Brigham & Crider, 2022). Several studies have leveraged this terrain-geologic material relationship using traditional methods such as logistic regression, random forests, and support vector machines for classification or segmentation of binary hazards (e.g., landslides, sinkholes) (Kirkwood et al., 2016; Zhu & Pierskalla Jr, 2016; Crawford et al., 2021) or SG maps (Cracknell & Reading, 2014; Johnson & Haneberg, 2025). However, these approaches depend on hand-crafted features, are restricted to small geographic extents, and fail to generalize beyond the training region. More recently, deep learning methods using convolutional neural networks (CNNs) and CNN-Transformer hybrids have been applied to related 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 site-specific, lack standardized datasets, and rely on limited input modalities.

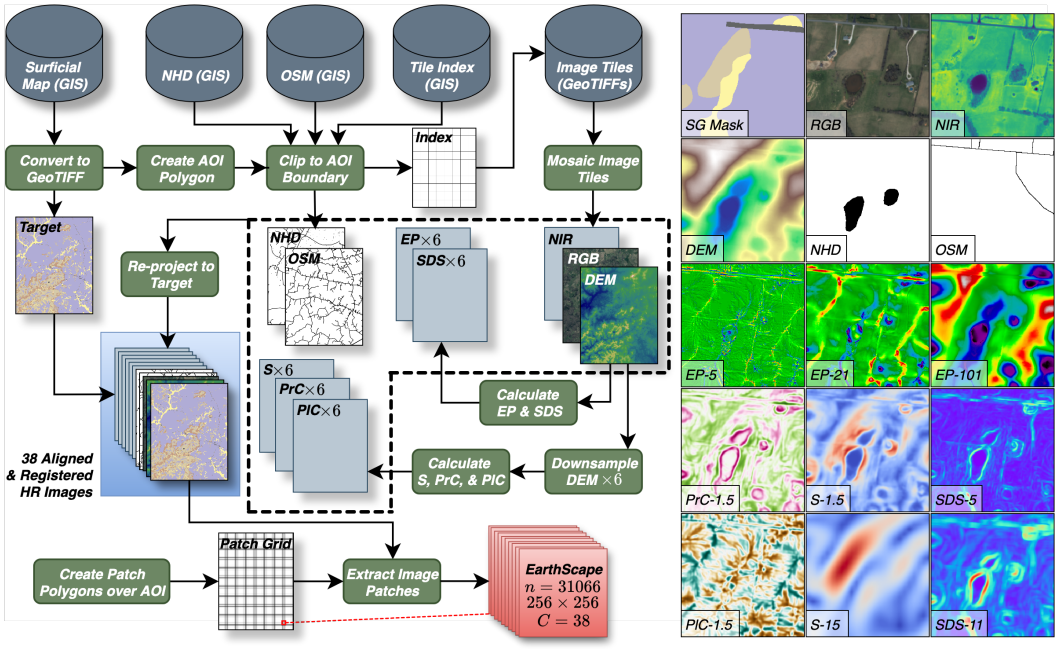


Figure 1: EarthScape data processing pipeline (left) and selected modalities from a single 256×256 patch (right). The SG map is rasterized and used to define the area of interest (AOI), from which all predictive features (DEM, RGB+NIR imagery, NHD hydrology, and OSM infrastructure) are clipped and aligned. 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. See Supp. A.2.4 and Figures 5 and 6 for additional examples.

Multimodal Learning for Geologic Tasks: Multimodal learning has become a central paradigm in RS and geospatial CV, where combining diverse data sources such as 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 has often meant 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). Although effective for detecting discrete hazards such as landslides or sinkholes, these approaches tend to overfit to absolute elevation or local spectral cues and fail to generalize to new regions. Other modalities have also been tested, including 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 useful in specific studies, these resources are site-specific and lack standardized availability for ML workflows. Together, this underscores the need for standardized, multimodal benchmarks that capture continuous surface materials and landforms across scales, rather than narrowly focusing on event-specific hazards.

RS and Geologic Datasets: RS benchmarks such as SpaceNet (Van Etten et al., 2018), xView (Lam et al., 2018), and the 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 such as vegetation or developed areas, rather than physical topographic characteristics, and therefore lack representations of Earth's surface that are essential for interpreting geological processes.

Several geoscience-specific datasets have been introduced in recent years, including MMFlood for flood delineation (Montello et al., 2022), QuakeSet for earthquake event detection (Rege Cambrin & Garza, 2024), and landslide detection datasets based on overhead imagery and DEMs (Ji et al., 2020;

Liu et al., 2023; Zhou et al., 2023). While valuable for their respective domains, these resources are narrowly scoped to discrete hazards or events, often limited to small geographic areas, and rely on shallow modality combinations. As such, they do not provide standardized, multimodal benchmarks for continuous SG mapping, where the goal is to delineate overlapping geologic units formed by surface processes, rather than to detect singular hazard events.

3 EARTHSCAPE DATASET

3.1 COMPOSITION AND FEATURES

Surficial Geologic Maps: The EarthScape dataset currently includes high-resolution (1:24.000-scale[1]) surficial geologic maps from Warren and Hardin Counties, Kentucky, compiled by the Kentucky Geological Survey (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). These maps provide the multilabel targets and segmentation masks (Fig. 1; also see Figs. 4, 5, and 6). 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 geodatabase format. Detailed unit descriptions are provided in the Supplemental A.2.3.

Aerial imagery and DEM: High-resolution aerial RGB+NIR imagery and LiDAR-derived DEMs form the primary RS inputs for EarthScape (Commonwealth of Kentucky, 2024). The aerial imagery consists of RGB and NIR channels with a ground sampling distance (GSD) of 0.15 m ($\approx$ 6 in). It is particularly useful for identifying anthropogenic features (e.g., af1) that are easily distinguished from natural landscapes (Fig. 1; also see Figs. 5, 6). The NIR band further enhances the detection of hydrological features, including alluvial deposits (Qal, Qaf, and Qat) and stream channels, by highlighting vegetation patterns that can indicate water presence or recent sediment deposition. 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 GSD of 1.52 m ($\approx$ 5 ft) GSD spatial resolution, is a critical feature for SG mapping (Fig. 1; also see Figs. 5 and 6). Both datasets are distributed as publicly accessible GeoTIFF tiles.

Geomorphometric Terrain Features: The DEM provides a foundation for deriving five key terrain features widely used in geomorphometric analysis and essential for delineating SG units (Fig. 1; also see Figs. 5 and 6) (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 the Supplemental A.2.4 for additional information.

Hydrography and Infrastructure: To support downstream tasks involving fluvial and anthropogenic processes, EarthScape includes vector data for hydrographic and infrastructure features (Fig. 1; also see Figs. 5 and 6). Stream centerlines and waterbody polygons from the U.S. Geological Survey's National Hydrography Dataset (NHD) (U.S. Geological Survey, 2024) 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 af1. These features also help characterize geologic disturbance near infrastructure, including slope undercutting and landslide susceptibility. Both datasets are included as binary raster images.

3.2 DATA PROCESSING

Targets: Each SG map was obtained as a vector GIS geodatabase, from which the relevant feature class was extracted (Fig. 1). The vector polygons were inspected for topological correctness, en-

[1]Map scale refers to cartographic accuracy, rather than raster resolution. At 1:24,000, one map unit represents 24,000 real-world units, and is considered the gold-standard geologic mapping scale.

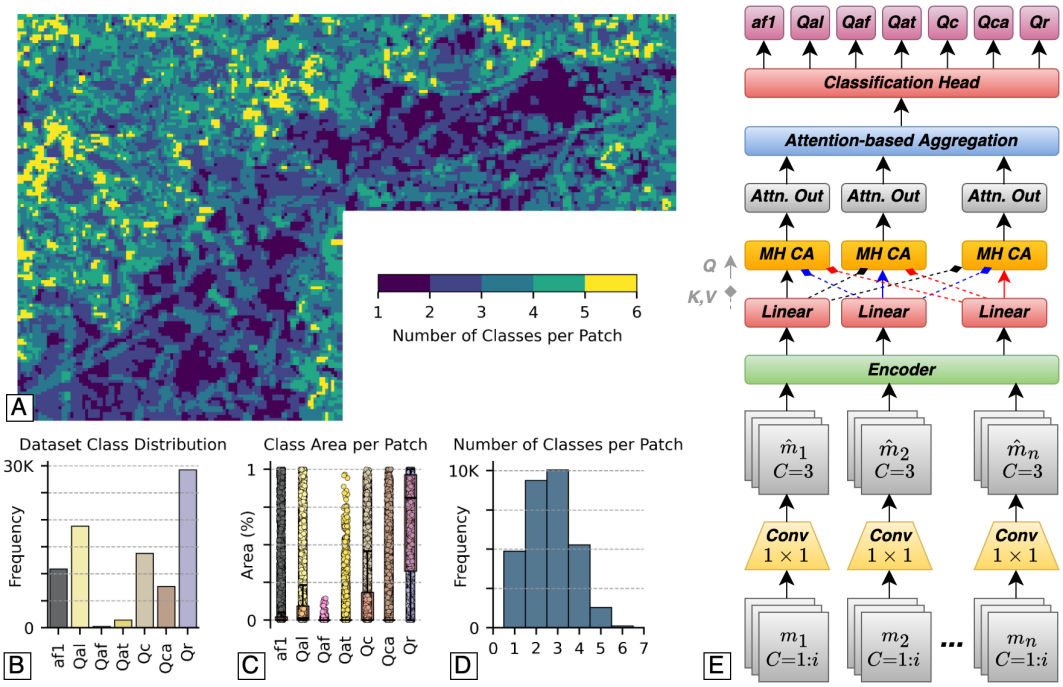


Figure 2: EarthScape dataset characteristics (A–D) and SGMap-Net architecture (E). A. Choropleth map of Warren County showing the number of classes per patch, illustrating spatial heterogeneity and task complexity. B. Dataset-wide class distribution, highlighting significant class imbalance. C. Proportional area of each class per patch, showing that many patches include low-exposure classes, increasing task 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 mid-level cross-attention.

suring no overlaps, no gaps, and valid geometries. The validated data were saved as a standalone GeoJSON file and used to generate a boundary polygon defining the area of interest (AOI) for clipping and extracting corresponding portions of other datasets. SG target classes were encoded as ordinal values in the GeoJSON. Finally, the vector data were rasterized to a GeoTIFF with a GSD of 1.52 m, matching the native resolution of the DEM.

Features: Vector datasets, including the NHD, OSM, and RS image tile index, were obtained, clipped to the target AOI, and saved as standalone GeoJSON files (Fig. 1). The tile index defines the locations of aerial RGB+NIR and DEM tiles. Using the AOI, the relevant tile footprints were selected and the corresponding DEM and imagery tiles downloaded (Fig. 1). DEM tiles were merged into a single GeoTIFF mosaic at their native GSD of 1.52 m. RGB and NIR imagery underwent the same process, with additional downsampling to GSD 1.52 m to match the DEM resolution. NHD stream centerlines and waterbody polygons, along with OSM road and railway centerlines, were rasterized into two binary GeoTIFFs with a GSD of 1.52 m.

Five terrain features were derived from the DEM at multiple spatial scales (Fig. 1), enabling models to capture both local and regional landform structure. We adopted a roughly logarithmic progression of resolutions, which is common in geomorphometric analysis when no "correct" scale exists. The DEM mosaic (native 1.52 m GSD) was downsampled using cubic convolution to five coarser resolutions of 3.05, 6.10, 15.24, 30.48, and 60.96 m GSD. S, PrC, and PlC were computed with 5×5 kernels on each of the six DEMs, upsampled back to 1.52 m GSD using cubic convolution, and smoothed with a Gaussian filter to reduce resampling artifacts. SDS and EP were calculated as neighborhood statistics on the original 1.52 m DEM using kernels of 5×5, 11×11, 21×21, 51×51, 101×101, and 201×201 pixels. These kernel sizes correspond to receptive fields comparable to the effective resolutions used for S, PrC, and PlC, allowing direct comparison across modalities.

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 patch was constructed so that it covers an area of exactly 256×256 pixels (390.14×390.14 m), overlaps adjacent cells 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. EarthScape supports classification (labels), segmentation (masks), regression (class proportions), object detection via derived bounding boxes, and multi-task extensions. This flexibility positions it as a general-purpose benchmark for surface-aware geospatial learning.

3.3 Dataset 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 currently 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 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, where a class is marked positive if any part of its mask intersects the patch (i.e., even a single pixel), without applying a proportional threshold. The model learns a mapping $f : X \to [0, 1]^K$ 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 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 introduce SGMap-Net, a lightweight fusion-based model designed as a transparent baseline. Its simplicity allows us to isolate the contributions of modality, scale, and fusion strategy without architectural confounds, while ensuring that results are reproducible and easily extendable. Figure 2E 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.

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

Each standardized modality $\hat{m_i}$ is passed through a shared encoder to extract feature maps $f_{m_i} = \mathrm{Encoder}(\hat{m_i})$. 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 representations $z_i = f_{m_i} + e_i$. 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).

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

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 $w = \mathrm{Softmax}(v^T A)$. 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 (see Supp. A.3, Fig. 7, Table 3). 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 it 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 a 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 A.4 and A.5 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 ResNeXt-50 and ViT-B/16 backbones (Table 1; also see Supp. A.6.1, Fig. 10, Tables 5, 6, and

Table 1: Macro-averaged F1, 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 single-modality, multi-scale, and multimodal models. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	F1			Precision			AUC			Accuracy		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-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	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	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-ms+S-ms+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-5+S-3+SDS-5	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-201+S-61+SDS-201	0.626	0.582	0.044	0.588	0.529	0.059	**0.885**	**0.812**	0.073	0.858	0.852	0.006

7). EP, S, and SDS consistently outperformed DEM and RGB+NIR, underscoring the value of domain-specific terrain derivatives. Model performance declined under cross-domain testing, with ViT exhibiting smaller generalization gaps (ΔF1 = 0.018) than ResNeXt (ΔF1 = 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 classes, motivating the evaluation of multi-scale fusion. Models trained with the ResNeXt-50 backbone (Table 1; also see Supp. A.6.2, Fig. 10, Tables 8, 9, and 10) demonstrate that early fusion via channel stacking is consistently more effective than attention-based fusion for cross-domain generalization. However, with ViT-B/16, multi-scale fusion achieves results comparable to single-modality models. EP, S, and SDS were the best overall performers in both backbones. With ResNeXt-50, EP continued to show the weakest generalization, consistent with single-modality results and indicating that EP retains local elevation signatures that transfer poorly across regions compared to S and SDS. Interestingly, PrC performed better with ViT-B/16 than with ResNeXt-50. PrC provides a relatively weak local signal, but ViT's patch-based tokenization and global attention appear to leverage it more effectively than convolutional filters. Overall, these results suggest that multi-scale fusion of S and SDS mitigates sensitivity to region-specific relief variation, and that early channel stacking remains the most reliable and stable strategy across backbones.

Multimodal Fusion: We evaluated multimodal fusion using ResNeXt-50 and ViT-B/16 backbones with three representative modality configurations: RGB+DEM, three variants of EP+S+SDS, and RGB+DEM+EP+S+SDS (Table 1; also see Supp. A.6.3, Fig. 10, Tables 11, 12, and 13). Multi-scale EP+S+SDS inputs consistently outperformed RGB+DEM, improving in-domain macro-F1 to 0.657 and yielding the best cross-generalization (0.059). Incorporating multi-scale versions of these features yielded the best cross-domain performance, improving macro-F1 from 0.380 to 0.598 and reducing the generalization gap from 0.271 to 0.059. Even reduced single-scale variants of this modality set ranked highly and slightly outperformed single-modality and multi-scale versions of the same modalities. Early channel stacking produced the highest in-domain scores, while mid-level concatenation and cross-attention yielded the smallest overall domain shifts (0.028 and 0.029). Overall, multimodal fusion with terrain-derived features substantially improves generalization to unseen regions compared to RGB or DEM inputs alone.

Class-Level Trends: Per-class analysis shows strong variation in discriminability across the seven units (see Supp. A.6.4, Fig. 11, Tables 14 and 15). Qc, Qca, and Qr are consistently well-identified, whereas minority classes Qat and Qaf remain the most challenging. Multimodal models typically yield the highest in-domain scores across backbones and reduce the generalization gap.

Comparisons with Existing Models: We conducted exploratory baselines with recent multimodal foundation models, including SatMAE (Cong et al., 2022), SatMAE++ (Noman et al., 2024), DOFA (Xiong et al., 2024), and Panopticon (Waldmann et al., 2025) (see Supp. A.6.5, Table 16). While SatMAE and SatMAE++ achieved competitive in-domain macro F1 scores (0.614, 0.656), their cross-domain performance degraded sharply (0.427, 0.454). DOFA and Panopticon performed even

worse. In contrast, our best SGMap-Net variants consistently outperformed these models in both in-domain and cross-domain tests, highlighting the importance of shape-centric terrain features.

5 Challenges and Limitations

Geographic Scope and Extensibility: EarthScape is currently limited to two regions in Kentucky, USA, reflecting the availability of 1:24,000-scale SG maps in standardized GIS formats. While this geographic scope is narrow, the dataset is designed as a living resource. The patch-based curation workflow supports continuous expansion by our team and by external contributors, provided quality-control and assurance checks are met. Planned updates will triple the number of patches by the end of 2025 and extend coverage to additional regions in 2026, enabling broader cross-domain evaluation.

Breadth vs. Depth: EarthScape is modest in area, but each patch contains 38 co-registered channels spanning imagery, elevation, terrain derivatives, and vector data. This balance of limited spatial breadth and high modality depth presents a unique challenge where models must learn to integrate rich, heterogeneous inputs while generalizing across sparse geographic coverage.

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 geologic 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 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, class boundaries are often approximate. The 1:24,000-scale mapping reflects geologic certainty, which propagates into patch-level labels. In our benchmarks, we employ a one-hot labeling scheme, where a class is marked as present even if it occupies only a single pixel. We provide class-area proportions per patch, which allows future work to explore thresholding and probabilistic label assignment.

Temporal Inconsistency: Input features were acquired between 2019 and 2024, introducing potential temporal mismatches across modalities. While the main source of temporal variability is anthropogenic (af1), but the underlying geology and SG classes are inherently stable on these timescales. This stability provides a consistent foundation for benchmarking, while still enabling evaluation of model robustness to asynchronous inputs.

6 Conclusions

We introduced EarthScape, an AI-ready, multimodal benchmark dataset for SG mapping. EarthScape integrates aerial imagery, DEMs, multi-scale terrain derivatives, and GIS vector data, providing a unique resource for multimodal geospatial learning. The dataset reflects real-world challenges such as class imbalance, spatial heterogeneity, and geographic variability, making it a robust testbed for AI models. Through baseline experiments, we established 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. Designed as a living dataset, EarthScape is extensible in both geographic and modality space, and while geographically compact (31k patches), it is unusually deep, with 38 co-registered channels per patch that present a distinctive multimodal learning challenge. Ongoing work includes expanding coverage, incorporating globally available features, and experimenting with segmentation. By releasing data, code, and benchmarks, we aim to foster reproducible research and cross-disciplinary collaboration, positioning EarthScape as a benchmark for multimodal fusion and domain adaptation in geospatial AI.

Ethics Statement

This work adheres to the ICLR Code of Ethics. EarthScape is built exclusively from publicly available, government or community datasets under open licenses; no human subjects, personal data, or sensitive information are involved. All source attributions and licensing terms are respected, and no conflicts of interest are present. We caution that models trained on EarthScape should be applied with geological domain expertise, particularly outside regions with similar surficial processes, to avoid misinterpretation in decision-making contexts. We report implementation details in the supplemental to promote awareness of environmental impact and enable informed replication.

Reproducibility Statement

We support reproducibility through precise documentation of data sources and preprocessing, patch generation and spatially independent splits, model and training configurations, and comprehensive results. Upon acceptance, the full EarthScape dataset and code will be publicly released with a data dictionary and README. These materials are intended to allow end-to-end reproduction of all reported experiments.

References

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

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

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

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

Richard 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/`.

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

Meiqiao Bi, Minghua Wang, Zhi Li, and Danfeng 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.

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

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

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

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

Wes Buchanan, Meredith Swallom, Antonia Bottoms, Matthew Massey, Bailee Nicole Hodelka, and Emily 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.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 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.

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

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Matthew J Cracknell and Anya 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.

Matthew M Crawford, Jason M Dortch, Hudson J Koch, Ashton A Killen, Junfeng Zhu, Yichuan Zhu, Lindsey S Bryson, and William 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.

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

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh 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*, pp. 172–181, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

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

Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113 (7):4845–4901, 2024.

Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27852–27862, 2024.

Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad Shahbaz Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108:102417, 2024.

Bailee Hodelka, Matthew Massey, Meredith Swallom, Steve Martin, Charles Wells, and Emily Morris. Surficial geologic map of the bristow 7.5-minute quadrangle, kentucky. Accepted for publication, 2024.

Kelly J Hokanson, CA Mendoza, and KJ 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.

Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*, 2022.

Shunping Ji, Dawen Yu, Chaoyong Shen, Weile Li, and Qiang Xu. Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides*, 17:1337–1352, 2020.

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

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

Charlie Kirkwood, Mark Cave, David Beamish, Stephen Grebby, and Antonio Ferreira. A machine learning approach to geochemical mapping. *Journal of Geochemical Exploration*, 167:49–61, 2016.

Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.

Rasim Latifovic, Darren Pouliot, and Janet 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.

Hui Li and Xiao-Jun 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence 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*, pp. 740–755. Springer, 2014.

Richard J. Lisle, Peter Brabham, and John 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.

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong 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*, pp. 26658–26668, 2024a.

Xinran Liu, Yuexing Peng, Zili Lu, Wei Li, Junchuan Yu, Daqing Ge, and Wei 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.

Yao Liu, Jianyuan Cheng, Qingtian Lü, Zaibin Liu, Jingjin Lu, Zhenyu Fan, and Lianzhi Zhang. Deep learning for geological mapping in the overburden area. *Frontiers in Earth Science*, 12: 1407173, 2024b.

Matthew Massey, Antonia Bottoms, Max Hammond, Emily Morris, and Michelle McHugh. Surficial geologic map of the sonora 7.5-minute quadrangle, central kentucky. *Kentucky Geological Survey Contract Report*, 13(44), 2021.

Matthew Massey, Meredith Swallom, Antonia Bottoms, Wes Buchanan, Bailee Nicole Hodelka, and Emily Morris. Surficial geologic map of the hadley 7.5-minute quadrangle, warren county, kentucky. *Kentucky Geological Survey Contract Report*, 13(56), 2023.

Matthew Massey, Meredith Swallom, Bailee Hodelka, Hannah Hayes, Charles Wells, Steve Martin, and Emily Morris. Surficial geologic map of the bowling green south 7.5-minute quadrangle, kentucky. Accepted for publication, 2024.

Xiangyun Meng, Nathan Hatch, Alexander Lambert, Anqi Li, Nolan Wagener, Matthew Schmittle, JoonHo Lee, Wentao Yuan, Zoey Chen, Samuel Deng, Greg Okopal, Dieter Fox, Byron Boots, and Amirreza Shaban. Terrainnet: Visual modeling of complex terrain for high-speed, off-road navigation, 2023. URL https://arxiv.org/abs/2303.15771.

Fabio Montello, Edoardo Arnaudo, and Claudio Rossi. Mmflood: A multimodal dataset for flood delineation from satellite imagery. *IEEE Access*, 10:96774–96787, 2022.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27811–27819, 2024.

IOA Odeh, DJ Chittleborough, and AB 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.

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

Muhammad Usman Rafique, Junfeng Zhu, and Nathan Jacobs. Automatic segmentation of sinkholes using a convolutional neural network. *Earth and Space Science*, 9(2):e2021EA002195, 2022.

Daniele Rege Cambrin and Paolo Garza. Quakeset: A dataset and low-resource models to monitor earthquakes through sentinel-1. *Proceedings of the International ISCRAM Conference*, May 2024. ISSN 2411-3387. doi: 10.59297/n89yc374. URL http://dx.doi.org/10.59297/n89yc374.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang 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.

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

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

Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4):351–362, 2023.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE, 2019.

Meredith Swallom, Matthew Massey, Wes Buchanan, Bailee Nicole Hodelka, Hannah Hayes, Charles Wells III, and Emily Morris. Surficial geologic map of the bowling green north 7.5-minute quadrangle, warren county, kentucky. *Kentucky Geological Survey Contract Report*, 13 (55), 2023.

Meredith Swallom, Bailee Hodelka, Matthew Massey, Hannah Hayes, Charles Wells, and Emily 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`, 2024. Stream centerlines and waterbody polygons. Accessed: 2024-08-01.

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

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

Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):91, 2020.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2204–2214, 2025.

Jiayu Wang, Ruizhi Wang, Jie Song, Haofei Zhang, Mingli Song, Zunlei Feng, and Li Sun. Rs3dbench: A comprehensive benchmark for 3d spatial perception in remote sensing, 2025. URL `https://arxiv.org/abs/2509.18897`.

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

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Yiming Zhou, Yuexing Peng, Wei Li, Junchuan Yu, Daqing Ge, and Wei 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.

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

A Supplemental Material

A.1 Code Availability and Reproducibility

All code used for data preprocessing, patch extraction, model training, and evaluation will be publicly available for the camera-ready version pending acceptance. The repository includes clear documentation and instructions for reproducing all experiments presented in the main paper and supplemental material. The codebase provides tools for downloading and aligning multimodal data (including GeoTIFF imagery and vector layers), generating spatially independent patch splits, and computing terrain derivatives. It also includes baseline model implementations of SGMap-Net using both ResNeXt-50 and ViT-B/16 backbones, along with scripts for training, evaluation, and visualization. Additional utilities support focal loss configuration, per-class performance metrics, and spatial overlays of predictions.

The full EarthScape dataset will also be publicly available upon acceptance. The dataset archive includes geospatially registered input images, multilabel target masks, class proportion tables, a README, and a detailed data dictionary describing all included modalities.

A.2 Exploring the EarthScape Dataset

A.2.1 Current Status and Roadmap

Figure 3 illustrates the current and planned geographic extent of the EarthScape dataset. Version 1.0 includes two spatially independent regions in central Kentucky: Warren County, which contains the largest number of image patches, and Hardin County, which serves as an independent test area with similar geologic and geomorphic conditions. This separation enables evaluation of cross-region generalizability. Version 1.1 (expected Q4 2025) will nearly triple the number of patches (Fig. 3), while Version 1.2 (expected Q2 2026) will extend coverage beyond Kentucky into adjacent regions. EarthScape is intended as a "living" resource. We anticipate and encourage external users to collaborate with us in contributing additional high-quality data, thereby broadening the dataset's geographic coverage and strengthening its value for the research community.

A.2.2 Geologic Generalization and Transferability

Although EarthScape 1.0 is geographically limited, the geologic processes and terrain surface types it represents are not unique. The dataset is directly applicable to the surficial geology exposed in the Interior Low Plateaus and Appalachian Plateaus (Fig. 3). Comparable landscapes characterized by carbonate bedrock, dissected plains, and mixed fluvial–colluvial systems occur globally, including the Ozark Plateau (USA), parts of the Carpathians (Eastern Europe), the Dinaric Alps (Balkans), and areas of central China and southeastern Australia. However, differences in geologic processes do constrain transferability. For instance, the Central Lowlands (Fig. 3) contain fundamentally different surficial materials and geomorphic processes as a result of widespread glaciation (rather than non-glaciated weathering and erosion), limiting the direct applicability of EarthScape 1.0. Accordingly, we recommend that applications of EarthScape 1.0 to new regions be guided by domain expertise to ensure geological validity and meaningful interpretation.

A.2.3 Surficial Geology and Surface Morphology

Figure 4 presents two examples of SG maps from the EarthScape dataset, shown as semi-transparent overlays atop multi-directional hillshade images. This visualization emphasizes the relationship between SG and topography. Distinct landforms, such as river valleys, plains, and steep hillslopes, are spatially correlated with specific surficial geologic units. EarthScape leverages this relationship to frame surficial geologic mapping as a vision task, where computer vision models can learn to associate surface patterns with underlying geological processes. The EarthScape dataset currently includes seven surficial geologic map units, each representing distinct surface processes. Although the maps are from Kentucky, the units reflect fluvial deposition, gravitational transport, and in-situ weathering processes that are active in many landscapes worldwide.

1. *Artificial fill (af1):* Manmade deposits consisting of transported or excavated material placed or removed for engineering, mining, or other anthropogenic structures. Includes

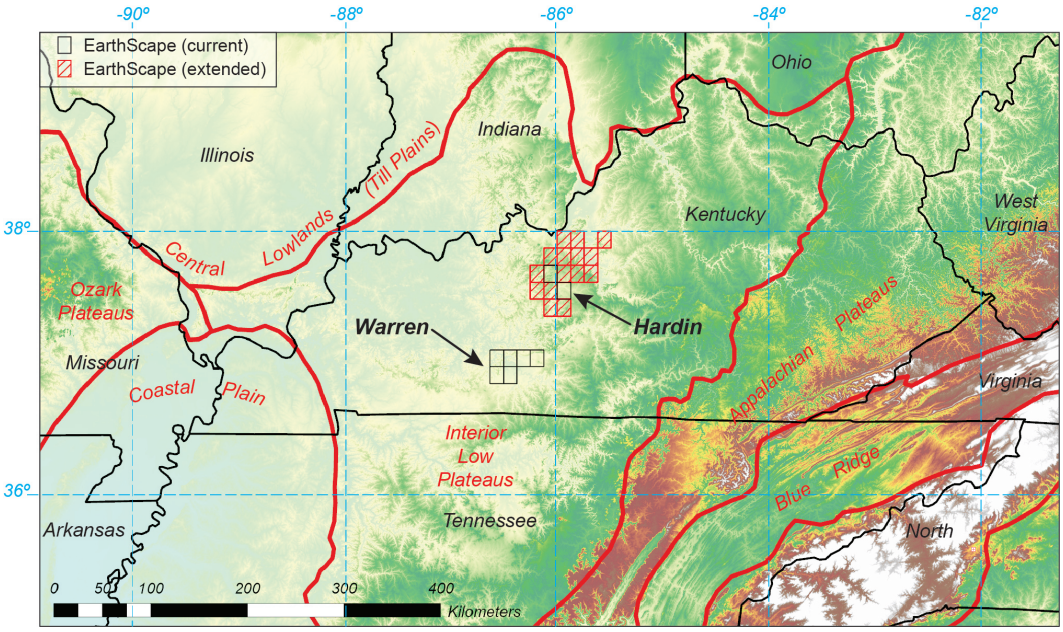
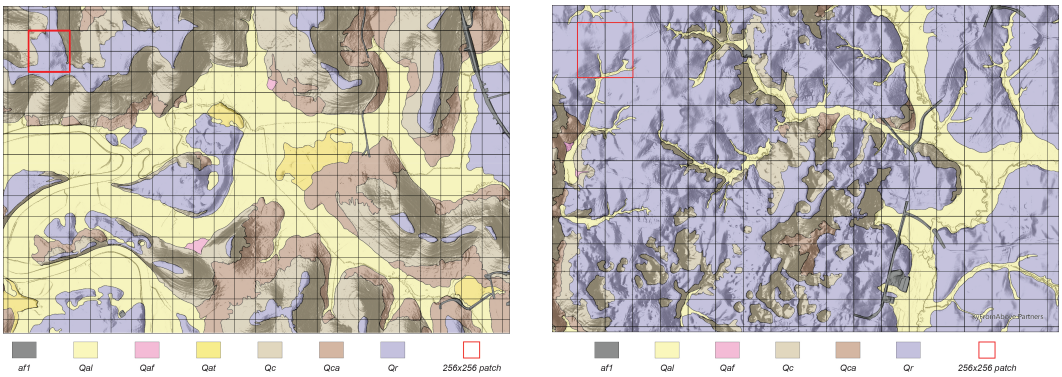


Figure 3: Map of the central United States showing the publicly available 1:24,000-scale surficial geologic maps. Red lines show boundaries of major geologic provinces, which provide geological constraints for generalizability. EarthScape-trained models are expected to generalize effectively throughout the Interior Low Plateaus and adjacent Appalachian Plateaus, based on shared terrain, bedrock, and geomorphic processes. In contrast, the glaciated Central Lowlands and Coastal Plain are characterized by fundamentally different surficial processes and materials.

road embankments, building pads, quarries, and areas of significant topographic modification. Often exhibits sharp, angular boundaries. The spatial extent of af1 can be below the mapping resolution and inconsistently captured on expert-curated surficial geologic maps.

2. *Alluvium (Qal):* Unconsolidated sediments, typically consisting of clay-, silt-, sand-, and gravel-sized particles, deposited by modern rivers and streams. Qal is commonly found in active floodplains and valley bottoms and reflects recent sedimentation from overbank flooding and channel migration. These areas are generally flat, vegetated, and hydrologically dynamic.
3. *Alluvial fans (Qaf):* Fan-shaped deposits formed at the base of tributaries or drainages, where sediment-laden water rapidly spreads and loses energy. These deposits are typically coarse-grained, poorly sorted, and associated with debris flows or flash floods. Although geologically significant, Qaf are often small, making them inconsistently represented on typical 1:24,000-scale maps.
4. *Terrace deposits (Qat):* Relict alluvial sediments preserved on elevated flat surfaces above modern stream channels. These deposits reflect former floodplain levels and subsequent stream incision. Compositionally similar to Qal, but usually expressed as distinct landforms above modern flood plains.
5. *Colluvium (Qc):* Hillslope-derived sediments that accumulate at the base of slopes due to gravity-driven processes such as soil creep, slopewash, and shallow landslides. Qc deposits are unsorted and variable in thickness, typically found on slopes $> 12°$. Qc is considered an active geomorphic unit.
6. *Colluvial aprons (Qca):* Slope-derived material deposited across lower hillslopes. Qca typically occurs downslope from Qc and is more stable, having accumulated over longer time periods. These deposits may be partially weathered, with poorly defined lower boundaries that grade into Qr due to extended weathering and lower erosion rates.
7. *Residuum (Qr):* Weathered material formed in place from the physical, chemical, and biological breakdown of underlying bedrock or older unconsolidated deposits. Qr lacks signif-



(a) Surficial geologic map of part of Warren County. (b) Surficial geologic map of part of Hardin County.

Figure 4: Example surficial geologic maps showing the distribution of unconsolidated materials overlaid on hillshade images to emphasize topographic context. The spatial correspondence between SG map units and landscape features, such as valleys and slopes, is visually apparent. The black grid indicates the layout of EarthScape patches, each measuring 256×256 pixels (390.14×390.14 m) with 50% overlap. Red squares in the upper left corners highlight a single patch

Table 2: Descriptions of surficial geologic units represented in EarthScape.

Class	Name	Dominant Process	Visual Cues
af1	Artificial fill	Anthropogenic	Sharp, angular edges; linear or rectilinear shapes; DEM anomalies inconsistent with natural terrain.
Qal	Alluvium	Water-dominated	Relatively wide, flat-bottomed valleys; active stream channels; low relative elevations.
Qaf	Alluvial fans	Water-dominated (acute)	Small, isolated, lobate landforms; located at slope-base transitions.
Qat	Terrace deposits	Water-dominated (relict)	Flat benches above floodplains; stepped margins; often dissected.
Qc	Colluvium	Gravity-dominated (active)	Steep slopes ($> 12°$); may include landslides or erosional hazards.
Qca	Colluvial aprons	Gravity-dominated (stable)	Wedge-shaped landforms along slope bases with concave profiles; transitional between slope and plain.
Qr	Residuum	In-situ weathering	Broad, low-relief uplands; little drainage or erosion; variable surface texture.

icant sediment transportation and is commonly found in upland areas with minimal active erosion. Qr is commonly gradational and poorly defined where it grades into Qc or Qca, leading to interpretive ambiguity during mapping.

A.2.4 EarthScape Modalities

Figs. 5 and 6 showcase the diverse, multimodal data available for each of the 31,018 EarthScape patches. Each patch includes 38 co-registered channels, comprising expert-labeled geologic masks, high-resolution aerial RGB and NIR imagery, a DEM, terrain features derived from the DEM at multiple spatial scales, and rasterized vector data representing hydrologic and infrastructure features. Among these modalities, the DEM and its derived terrain features provide critical context for understanding surface processes and interpreting surficial geologic units. Five terrain variables were computed at six spatial scales to capture localized and regional landform variability.

1. *Slope (S)* is the first derivative of elevation, measuring the rate of change of elevation over a horizontal distance. It quantifies the steepness of the terrain, providing insight into processes like erosion and material movement.

$$S = \tan^{-1}\left(\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\right) \tag{3}$$

Where $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ are the partial derivatives of elevation in the x and y directions, respectively.

2. *Profile curvature (PrC)* is a directional second derivative of elevation, measured along the direction of the steepest slope. It quantifies how slope changes in that direction, reflecting the acceleration or deceleration of flow, and influencing erosion and deposition patterns.

$$PrC = \frac{p^2 r + 2pqs + q^2 t}{(p^2 + q^2)^{3/2}} \tag{4}$$

Where $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$ are the first-order partial derivatives of elevation in the x and y directions, and $r = \frac{\partial^2 z}{\partial x^2}$, $s = \frac{\partial^2 z}{\partial x \partial y}$, and $t = \frac{\partial^2 z}{\partial y^2}$ are the corresponding second-order partial derivatives.

3. *Planform curvature (PlC)* is another directional second derivative of elevation, measured perpendicular to the direction of the steepest slope. It describes the curvature of contour lines (lines of equal elevation) and reflects how flow paths converge or diverge across the landscape.

$$PlC = \frac{q^2 r - 2pqs + p^2 t}{(p^2 + q^2)^{3/2}} \tag{5}$$

Where $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$ are the first-order partial derivatives of elevation in the x and y directions, and $r = \frac{\partial^2 z}{\partial x^2}$, $s = \frac{\partial^2 z}{\partial x \partial y}$, and $t = \frac{\partial^2 z}{\partial y^2}$ are the corresponding second-order partial derivatives.

4. *Elevation percentile (EP)* measures the relative elevation of a point within a defined neighborhood, expressed as a percentile rank (0–100%) of the elevation among neighboring values. EP helps distinguish between landforms defined by relative topography, such as ridges, valleys, or sinkholes.

$$EP = 100 \cdot \frac{|\{z_i \in Z \mid z_i < z\}|}{N} \tag{6}$$

Where z is the elevation at the center cell, Z is the set of elevations in the neighborhood, z_i are the individual neighboring elevations, and N is the total number of neighbors. The numerator counts the number of neighbors with elevation less than z.

5. *Standard deviation of slope (SDS)* is a measure of roughness and quantifies the variability in slope angle within a local window. SDS represents how rugged or uneven the surface is, highlighting areas with complex topography that may correlate with diverse geologic materials or processes.

$$SDS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_i - \bar{S}\right)^2} \tag{7}$$

Where S_i is the slope angle (in degrees or radians) of the i^{th} cell in the neighborhood, $\bar{S}$ is the mean slope within that neighborhood, and N is the total number of cells used in the calculation window.

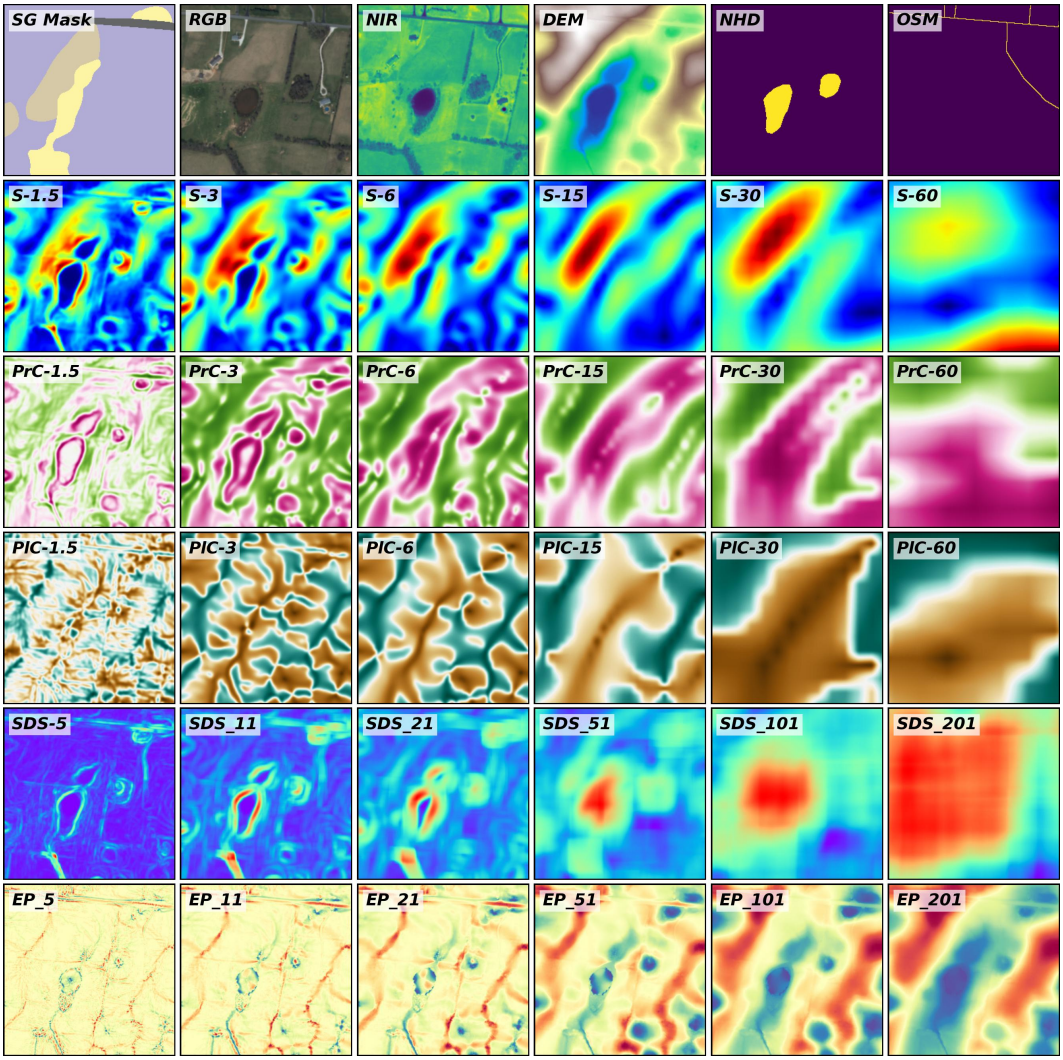


Figure 5: Example patch from the Warren County area showcasing the 38 channels available in EarthScape. Channels are displayed from top left to bottom right: target mask, RGB aerial imagery, NIR aerial imagery, DEM, NHD hydrologic features, OSM infrastructure, six spatial scales of S, PrC, and PlC derived from downsampled DEMs, and multiple scales of SDS and EP calculated using six kernel sizes with the original DEM.

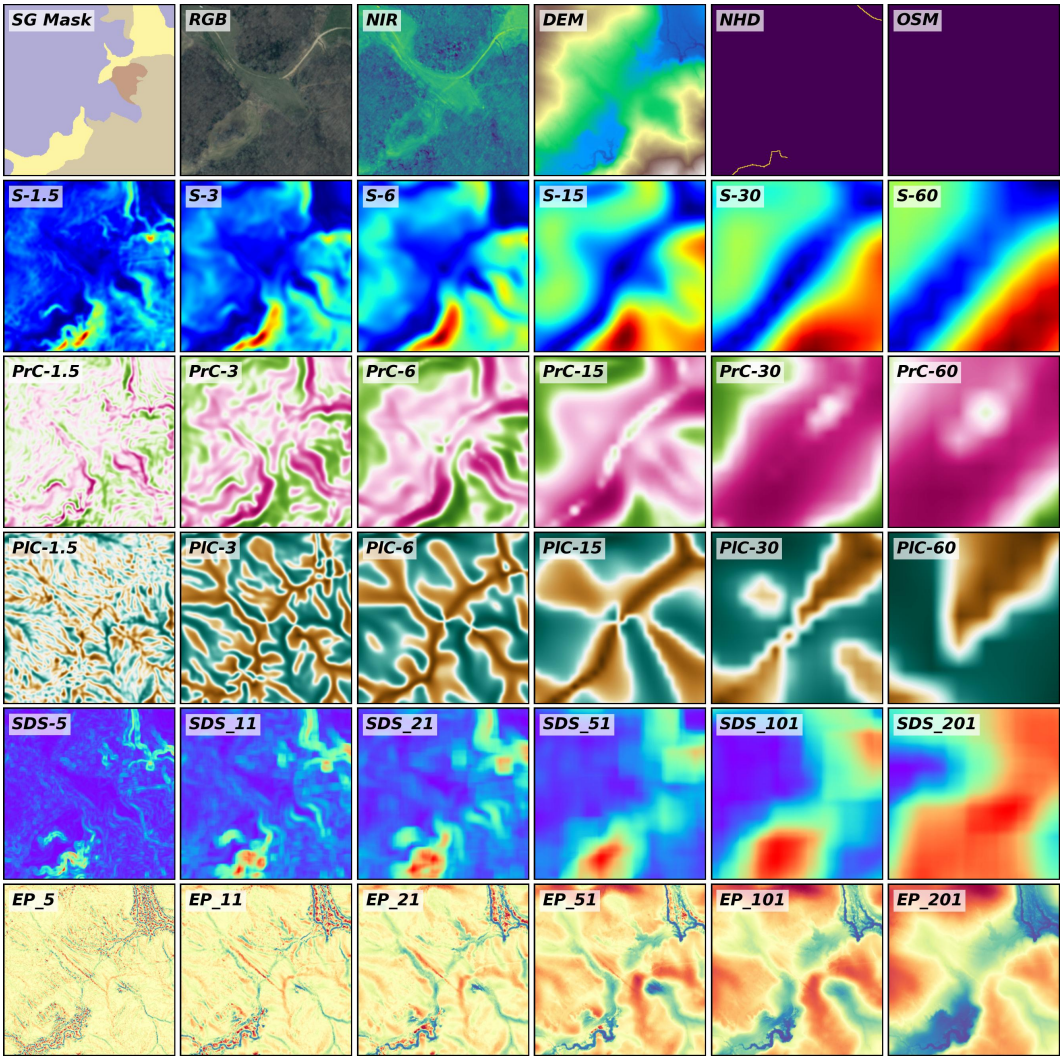


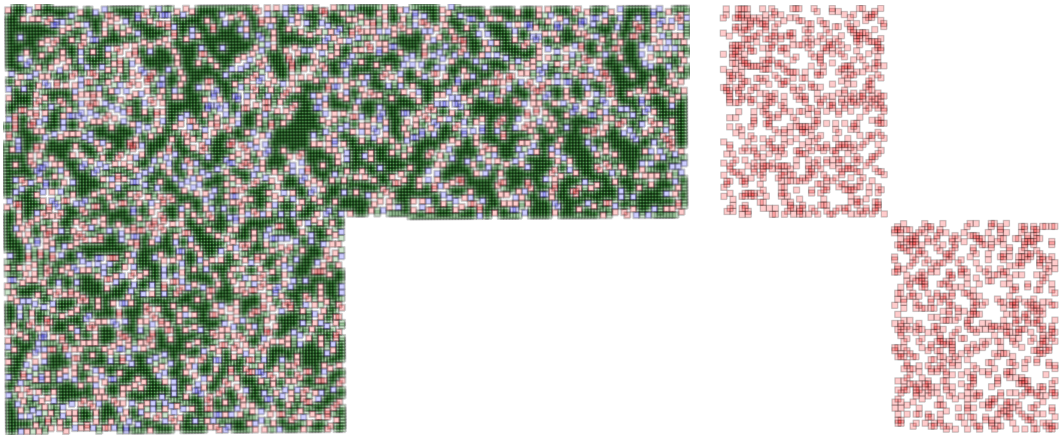
Figure 6: Example patch from the Hardin County area showcasing the 38 channels available in EarthScape. Channels are displayed from top left to bottom right: target mask, RGB aerial imagery, NIR aerial imagery, DEM, NHD hydrologic features, OSM infrastructure, six spatial scales of S, PrC, and PlC derived from downsampled DEMs, and multiple scales of SDS and EP calculated using six kernel sizes with the original DEM.

A.3 Geospatial Patch Selection and Experimental Design

To ensure robust and geographically fair model evaluation, EarthScape patches were split into spatially independent training, validation, and test sets. The Warren County region was used for in-domain training and evaluation due to its broader spatial coverage and diversity of surficial geologic units. We first randomly selected 1,536 test patches, followed by 768 validation patches that did not spatially intersect with the test set, and then assigned the remaining 8,416 non-overlapping patches to the training set (Fig. 7). These split sizes were chosen through iterative selection to satisfy several practical constraints: (1) all splits had to be spatially non-overlapping; (2) patch counts needed to be divisible by common batch sizes (e.g., 16 or 32) to support efficient model training; (3) the resulting proportions had to be reasonably balanced and typical for supervised learning workflows (Table 3).

To assess geographic generalization, we created a cross-domain test set consisting of 1,536 randomly selected patches from the Hardin County region (Fig. 7). Although geologically similar, Hardin County is located approximately 85 km from Warren County and is spatially independent. This separate region enables testing model performance under domain shift, simulating real-world conditions in which models are applied beyond the area used for training.

Figure 8 shows the class distributions for each data split. All subsets reflect the inherent class imbalance typical of surficial geologic mapping, driven by the localized nature of surface processes. Importantly, the class distributions are consistent across the training, validation, and both test sets, ensuring that evaluation performance is not biased by differences in class representation.



(a) Training, validation, and in-domain test patches from the Warren County region.

(b) Cross-domain test patches from the Hardin County region.

Figure 7: Spatial distribution of selected patches for EarthScape experiments. All splits are spatially independent: no patch overlaps between splits, though patches within the same split may partially overlap due to the 50% patch stride. See Figure 3 for geographic locations.

Table 3: Patch counts and split proportions for training, validation, and testing based on the total number of patches used for in-domain training and evaluation. An additional test set from the spatially independent Hardin County region was used to assess cross-domain generalization.

Split	Region	Patch Count (n)	In-domain Proportion (%)
Training	Warren	8,416	78.5
Validation	Warren	768	7.2
In-domain Testing	Warren	1,536	14.3
Cross-domain Testing	Hardin	1,536	-

A.4 Hardware and Training Configuration

All experiments were implemented in Python using the PyTorch framework. Models were trained and evaluated on a machine equipped with an Intel Xeon processor, 128 GB of RAM, and two

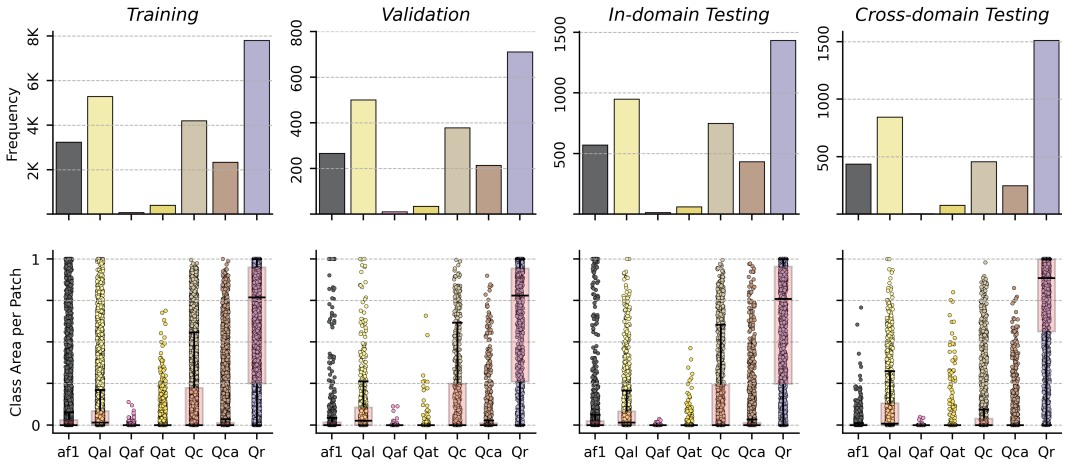
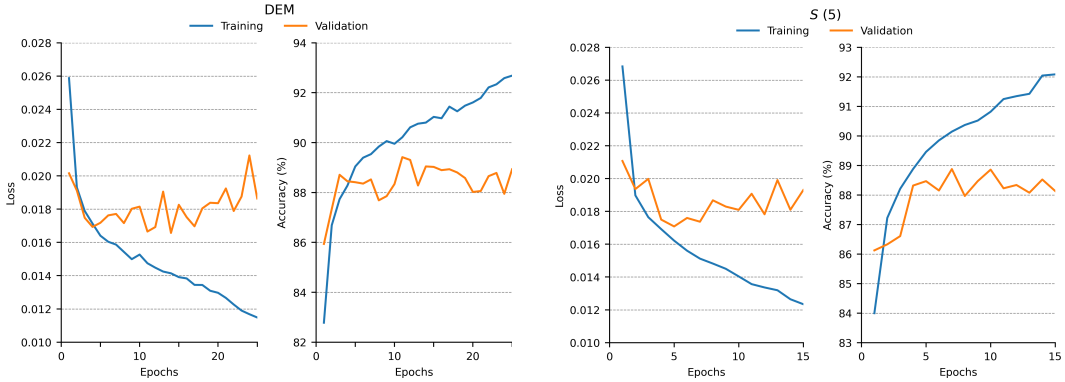


Figure 8: Class distribution and intra-patch composition across EarthScape data splits. Top row: Bar plots showing the frequency of each surficial geologic unit in the training, validation, in-domain test, and cross-domain test sets. Bottom row: Swarm plots overlaid with box plots showing the proportion of each patch occupied by each class. All splits display consistent patterns in both overall frequency and within-patch composition, supporting fair evaluation across subsets.

NVIDIA RTX A4000 GPUs. Initial training experiments were run for 25 epochs to observe convergence behavior (Fig. 9). Across all configurations, we found that model performance generally stabilized within the first 10 epochs (Fig. 9). Based on these observations, we standardized all subsequent experiments to 15 epochs, which provided a balance between sufficient training and computational efficiency.



(a) DEM model trained for 25 epochs. Early convergence is evident by epoch 10, with decreased performance thereafter.

(b) S (5) model trained for 15 epochs, demonstrating stable convergence and alignment between training and validation performance.

Figure 9: Training and validation loss and accuracy curves across epochs. Each subplot shows model loss (left panel) and accuracy (right panel) behavior for a different input modality, with training curves shown in blue and validation curves in orange.

A.5 Focal Loss

To address the significant class imbalance in EarthScape, we adopted focal loss. Initial tuning was conducted using the validation set and DEM modality only, a ResNeXt-50 backbone, the Adam optimizer, and a fixed learning rate of 0.001 to explore the effects of focal loss parameters. We evaluated values of $\gamma \in 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0$ and tested several strategies for the class-balancing factor (α), including a fixed scalar ($\alpha = 0.25$), inverse class frequency (ICF), square root of ICF ($\sqrt{\text{ICF}}$), and class-balanced focal loss with $\beta = 0.999$ (CBFL) (Table 4). The combination of

$\alpha = \sqrt{\text{ICF}}$ and $\gamma = 2.0$ yielded the best performance for the DEM-only configuration. However, when this setting was applied to other modalities, training became unstable, and convergence was inconsistent. To ensure comparability across all experiments and isolate the effects of modality and fusion design, we adopted the original focal loss settings ($\alpha = 0.25$, $\gamma = 2.0$) for all remaining runs.

Table 4: Per-class and macro-averaged validation set F1 scores for different focal loss configurations using the DEM modality and a ResNeXt-50 backbone. These results were used to guide focal loss tuning, although the best-performing configuration did not generalize well across modalities. As a result, we adopted $\alpha = 0.25$, $\gamma = 2.0$ for all subsequent experiments.

α	γ	F1								AUC							
		af1	Qal	Qaf	Qat	Qc	Qca	Qr	AVG.	af1	Qal	Qaf	Qat	Qc	Qca	Qr	AVG.
0.25	1	0.743	0.848	0.267	0.436	0.899	0.778	0.968	0.706	0.861	0.862	0.907	0.923	0.967	0.923	0.937	0.911
0.25	1.5	0.726	0.855	0.250	0.354	0.914	0.751	0.968	0.688	0.866	0.874	0.915	0.884	0.964	0.909	0.932	0.906
0.25	2	0.749	0.841	0.229	0.400	0.914	0.778	0.965	0.697	0.868	0.859	0.929	0.919	0.970	0.929	0.912	0.912
0.25	2.5	0.690	0.866	0.275	0.387	0.895	0.767	0.971	0.693	0.844	0.887	0.944	0.895	0.965	0.920	0.945	0.914
0.25	3	0.709	0.851	0.267	0.323	0.890	0.772	0.970	0.683	0.853	0.863	0.895	0.890	0.962	0.925	0.924	0.902
ICF	1	0.524	0.804	0.204	0.390	0.831	0.640	0.961	0.622	0.639	0.730	0.921	0.851	0.912	0.828	0.851	0.819
ICF	2	0.596	0.805	0.286	0.314	0.839	0.687	0.961	0.641	0.731	0.737	0.934	0.828	0.916	0.854	0.869	0.838
ICF	2.5	0.589	0.799	0.267	0.326	0.843	0.671	0.962	0.637	0.711	0.716	0.923	0.838	0.919	0.842	0.848	0.828
$\sqrt{\text{ICF}}$	1	0.696	0.845	0.286	0.348	0.879	0.763	0.965	0.683	0.843	0.867	0.912	0.905	0.955	0.925	0.922	0.904
$\sqrt{\text{ICF}}$	1.5	0.688	0.838	0.333	0.409	0.877	0.766	0.974	0.698	0.834	0.844	0.961	0.909	0.951	0.914	0.924	0.905
$\sqrt{\text{ICF}}$	2	0.726	0.841	0.444	0.460	0.905	0.749	0.962	0.727	0.850	0.853	0.945	0.931	0.961	0.921	0.913	0.911
$\sqrt{\text{ICF}}$	2.5	0.709	0.835	0.293	0.487	0.901	0.760	0.963	0.707	0.849	0.844	0.956	0.940	0.962	0.926	0.893	0.910
CBFL	1	0.720	0.831	0.412	0.427	0.893	0.733	0.973	0.713	0.864	0.839	0.965	0.903	0.962	0.902	0.924	0.908
CBFL	1.5	0.715	0.841	0.286	0.412	0.908	0.764	0.971	0.700	0.844	0.854	0.940	0.906	0.971	0.920	0.947	0.912
CBFL	2	0.727	0.866	0.357	0.455	0.914	0.792	0.965	0.725	0.867	0.890	0.918	0.923	0.971	0.921	0.914	0.915
CBFL	2.5	0.711	0.844	0.455	0.372	0.911	0.753	0.968	0.716	0.846	0.857	0.970	0.908	0.967	0.928	0.930	0.915

A.6 Comprehensive Results

A.6.1 Single Modality

Tables 5, 6, and 7 present single-modality results across F1, AUC, precision, recall, mAP, and accuracy for both in-domain (Warren County) and cross-domain (Hardin County) test sets, using ResNeXt-50 and ViT-B/16 backbones (see also Fig. 10). Results highlight substantial performance differences between modalities and backbones, particularly under domain shift.

Imagery-based models (RGB and NIR) degrade sharply when transferred across regions. For example, RGB drops from 0.599 to 0.394 in macro-averaged F1 ($\Delta_{F1} = -0.205$), with a corresponding AUC decline of 0.258. NIR shows a smaller but still notable loss. In contrast, models trained on DEM inputs retain more performance across domains, with only a 0.105 decline in F1 and a 0.153 decline in AUC.

Terrain features derived from the DEM consistently outperform both raw elevation and spectral imagery in terms of accuracy and generalization. Among these, slope (S) and elevation percentile (EP) stand out as the most informative and stable. For instance, S-5 achieves an in-domain F1 of 0.645 and a cross-domain F1 of 0.575 with ResNeXt-50, indicating strong transferability. EP-51 provides high in-domain scores but suffers larger drops under domain shift, consistent with its reliance on local elevation signatures. Standard deviation of slope (SDS) and, to a lesser extent, curvature measures (PlC, PrC) also contribute, though curvature features remain weaker overall.

Backbone choice further influences performance. ResNeXt-50 generally achieves higher in-domain scores, capturing localized patterns more effectively, whereas ViT-B/16 narrows the generalization gap (e.g., $\Delta F1_{ViT} = 0.018$ vs. $\Delta F1_{ResNeXt} = 0.043$). This suggests that convolutional backbones may better exploit site-specific features, while transformer backbones offer slightly greater robustness in unfamiliar geologic settings.

A.6.2 Multi-scale Fusion

Tables 8, 9, and 10 present results for multi-scale fusion experiments across F1, AUC, precision, recall, mAP, and accuracy, using both ResNeXt-50 and ViT-B/16 backbones. Two fusion strategies were evaluated: early channel stacking and attention-based fusion.

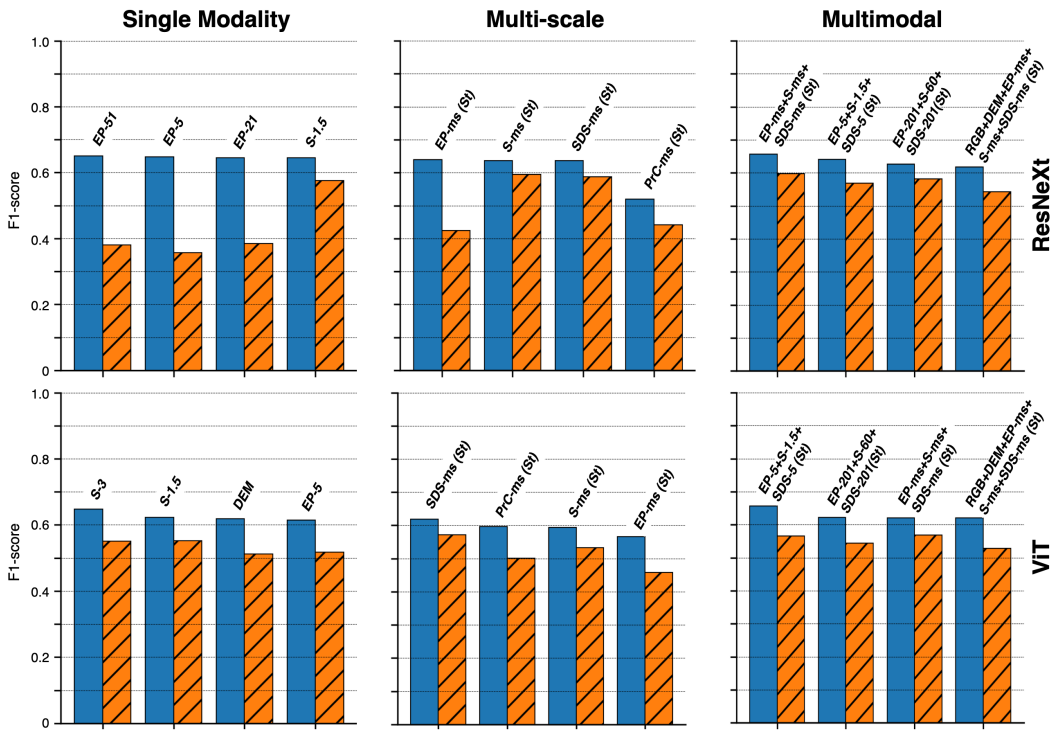


Figure 10: In-domain (blue) and cross-domain (orange, hatched) F1 scores for the top four models for single-modality, multi-scale fusion, and multimodal fusion experiments. Rows show comparisons of ResNeXt-50 (top) vs. ViT-B/16 (bottom) backbones. Each subplot shows the four best-performing models based on in-domain F1 scores. Cross-domain bars illustrate domain shift using the same models selected based on in-domain performance. Model configurations are shown above each group and indicate the input modality, or modality combination and fusion strategy.

Across nearly all configurations, early channel stacking consistently outperforms attention-based fusion in in-domain performance and often improves generalization. For example, with ResNeXt-50 on EP, stacking increases in-domain F1 from 0.494 (attention-based) to 0.640, while cross-domain results remain stable (0.426 vs. 0.425). Terrain features S and SDS achieve strong in-domain scores (0.636–0.637) with comparatively small performance drops across domains ($\Delta_{F1} = 0.043$–0.048), underscoring their robustness. In contrast, curvature measures (PlC and PrC) remain weak performers, even with multi-scale inputs, reinforcing earlier findings that they are less discriminative in isolation.

Results using the ViT-B/16 backbone largely mirror those of ResNeXt-50. Early stacking again yields modest gains, but PrC emerges as a relative outlier: it performs best under ViT despite being poor with ResNeXt. This contrast suggests that sensitivity to certain terrain features may depend more on backbone architecture than on fusion strategy.

Overall, multi-scale fusion helps mitigate the limitations of single-resolution inputs, with early channel stacking proving to be the most reliable and effective strategy. Attention-based approaches fail to match its simplicity and stability, highlighting the value of straightforward fusion mechanisms for integrating information across spatial scales.

A.6.3 Multimodal Fusion

Tables 11, 12, and 13 present results for multimodal fusion experiments across F1, AUC, precision, recall, mAP, and accuracy, using ResNeXt-50 and ViT-B/16 backbones. Four fusion strategies were tested: early channel stacking, mid-level concatenation, and mid-level attention with either a shared or separate encoder.

Fusion strategy plays a critical role in shaping both peak performance and generalization. Early channel stacking delivers the highest in-domain results (e.g., F1=0.657 with ResNeXt-50) but incurs a moderate domain gap ($\Delta F1 = 0.059$). Mid-level attention with a shared encoder narrows this gap ($\Delta F1 = 0.029$) but at the cost of lower in-domain performance (F1=0.561). Mid-level concatenation provides the most balanced compromise, achieving moderate in-domain scores (F1=0.596) with the smallest domain shift observed ($\Delta F1 = 0.028$). Overall, attention-based fusion underperforms compared to simpler strategies, suggesting that architectural sophistication does not necessarily translate into better integration of geospatial modalities.

Backbone choice also influences outcomes. ResNeXt-50 generally provides a slight in-domain advantage over ViT-B/16 (e.g., 0.657 vs. 0.621), while ViT offers similar generalization. More importantly, modality selection has a larger effect than backbone choice. Pairing RGB and DEM produces poor generalization, with domain gaps as large as $\Delta F1 = 0.211$. In contrast, engineered terrain features (EP, S, and SDS) consistently yield the best overall results. Their multi-scale combination achieves an in-domain F1 of 0.657 and reduces the cross-domain drop to just $\Delta F1 = 0.059$ (Fig. 10). Strong performance persists even with single-scale variants, underscoring the robustness of shape-centric terrain features relative to raw elevation or spectral imagery.

While strong single-modality models exist, multimodal fusion offers slight gains in peak accuracy and, more importantly, substantially improves generalization to unseen regions. These results emphasize the importance of shape-based terrain features and demonstrate that simple, well-chosen fusion strategies can outperform more complex attention mechanisms for multimodal geospatial learning.

A.6.4 Class-Level Trends

Class-wise AUC analysis across backbones and fusion strategies reveals broadly consistent patterns for both ResNeXt-50 and ViT-B/16 (Fig. 11; Tables 14, 15). Units such as Qc, Qca, and Qr consistently achieve the highest discriminability, whereas Qal, Qat, and Qaf are more challenging. Single-modality models typically yield the best in-domain scores but are more sensitive to domain shift. Multi-scale and multimodal fusions generally reduce this gap, though sometimes at the cost of peak in-domain performance. Interestingly, several units perform better in cross-domain testing than in-domain, suggesting that the training region (Warren County) may be geomorphologically more complex than the test region (Hardin County), or that models learn transferable representations despite this complexity.

Preferred modalities vary by class and backbone. For ResNeXt, most classes favor slope (S), but at different scales. For example, af1 and Qal perform best with S at 1.52 m GSD, Qat favors EP at the smallest kernel size (consistent with its position above floodplains), and Qc/Qca prefer S, though Qca benefits from a coarser 30.48 m GSD, reflecting its broader morphology. Qr also prefers S, aligning with its expression as a low-relief deposit. Qaf is an exception, performing best with PlC at 60.96 m GSD, possibly because large-scale curvature helps identify depositional fans in their broader geomorphic context. With ViT-B/16, af1, Qal, and Qat favor EP, Qaf performs best with S at 15.24 m GSD, Qc and Qca again prefer S at varying scales, and Qr shows a unique preference for PrC, consistent with its occurrence in relatively flat terrain.

Multimodal models reinforce the importance of terrain-based features. With ResNeXt, nearly all classes perform best with EP+S+SDS, with the exception of Qca, which benefits from including RGB and DEM. Although Qca is typically slope-derived and not visually distinct, the added value of RGB may reflect incidental correlations with infrastructure or vegetation patterns. ViT-based multimodal models were tested on fewer configurations but reveal similar trends: af1 and Qat perform best with RGB+DEM, reflecting their distinctive visual patterns and human modification, while the other five classes favor EP+S+SDS.

Overall, these results highlight the complexity of class-specific modality preferences. Optimal configurations vary by unit, backbone, and fusion strategy, reflecting differences in geomorphic expression and internal variability. While no single input or fusion method works best for all classes, shape-derived terrain features (EP, S, and SDS) emerge as consistently strong predictors of unit separability.

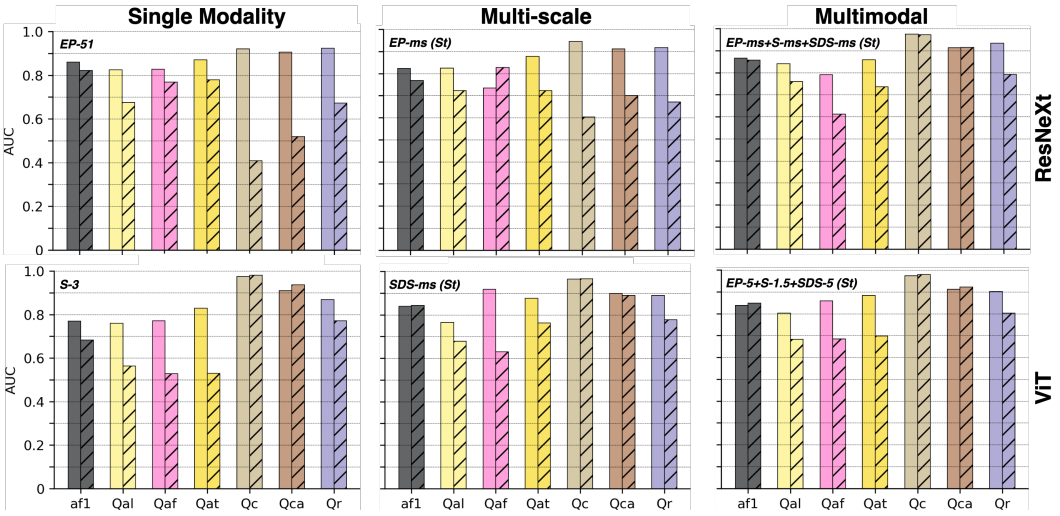


Figure 11: In-domain (solid) and cross-domain (hatched) class-wise AUC scores for the single best-performing models across different experiment types and backbone architectures. Rows show comparisons of ResNeXt-50 (top) vs. ViT-B/16 (bottom) backbones. Each subplot shows the best-performing model based on in-domain F1 scores. Cross-domain bars illustrate domain shift using the same model selected based on in-domain performance. Model configurations are shown above each group and indicate the input modality, or modality combination and fusion strategy.

A.6.5 Comparisons with Existing Models

We conducted exploratory experiments with recent multimodal foundation models, including SatMAE (Cong et al., 2022), SatMAE++ (Noman et al., 2024), DOFA (Xiong et al., 2024), and Panopticon (Waldmann et al., 2025). These models were developed for grouped multispectral or multisensor satellite imagery and are not natively configured to handle LiDAR-derived terrain features at multiple spatial scales. Our goal was not exhaustive hyperparameter optimization, but rather to provide indicative baselines for how existing large-scale models perform on EarthScape. We present these comparisons as exploratory and encourage future work on adapting foundation models to the challenges highlighted by EarthScape.

Following the grouping strategy of SatMAE and SatMAE++, we organized EarthScape modalities into three groups: (1) RGB + DEM, (2) elevation percentile (EP) at four scales (5, 51, 101, 201), and (3) slope (S-5) and standard deviation of slope (SDS-5). This configuration included ten modalities drawn from the strongest single-modality performers. Both SatMAE and SatMAE++ were fine-tuned on the same training, validation, and testing splits used in our main experiments. Despite achieving competitive in-domain macro F1 scores of 0.614 and 0.656, respectively, cross-domain performance dropped sharply to 0.427 and 0.454 (Table 16), underscoring the difficulty of transferring pretrained representations designed for spectral imagery to geologically diverse terrain settings.

We also evaluated DOFA and Panopticon, two transformer-based foundation models for multimodal Earth observation. Both underperformed SGMap-Net across in-domain and cross-domain tests, reflecting the limitations of imagery-centric architectures when applied to tasks dominated by terrain derivatives.

For comparison, our best SGMap-Net variants consistently outperformed these foundation models in both in-domain and cross-domain settings, illustrating the strong generalization of shape-centric features. While masked autoencoder and foundation model architectures remain promising for Earth observation, their robustness does not readily transfer to surface-process analyses without geologically informed modality design and fusion strategies.

Table 5: Macro-averaged F1 and AUC for *single modality* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones. WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	F1 (ResNeXt)			F1 (ViT)			AUC (ResNeXt)			AUC (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
DEM	0.632	0.527	0.105	0.618	0.512	0.237	**0.883**	0.730	0.153	**0.857**	0.620	0.237
RGB	0.599	0.394	0.205	0.579	0.332	0.267	0.815	0.557	0.258	0.793	0.526	0.267
NIR	0.613	0.468	0.145	0.579	0.275	0.274	0.815	0.650	0.166	0.784	0.509	0.274
NHD	0.515	0.434	0.081	0.492	0.428	0.064	0.659	0.576	0.083	0.496	0.509	-0.013
OSM	0.530	0.463	0.067	0.500	0.428	0.072	0.653	0.587	0.066	0.545	0.513	0.032
EP-5	0.648	0.357	0.291	0.614	0.518	0.117	0.872	0.582	0.290	0.854	0.738	0.117
EP-11	0.639	0.425	0.214	0.603	0.519	0.082	0.879	0.675	0.203	0.850	**0.768**	0.082
EP-21	0.645	0.384	0.261	0.608	0.503	0.079	0.877	0.695	0.183	0.838	0.759	0.079
EP-51	**0.651**	0.380	0.271	0.604	0.489	0.078	0.876	0.663	0.213	0.835	0.757	0.078
EP-101	0.619	0.476	0.143	0.589	0.477	0.075	0.857	0.739	0.118	0.819	0.744	0.075
EP-201	0.610	0.391	0.219	0.584	0.472	0.062	0.869	0.724	0.145	0.799	0.737	0.062
PlC-1.5	0.491	0.425	0.066	0.517	0.452	0.013	0.514	0.513	**0.001**	0.603	0.590	0.013
PlC-3	0.494	0.426	0.068	0.524	0.457	0.007	0.501	0.500	**0.001**	0.621	0.614	0.007
PlC-6	0.495	0.425	0.070	0.513	0.453	**0.005**	0.488	0.485	0.002	0.632	0.627	**0.005**
PlC-15	0.488	0.425	0.063	0.495	0.426	0.016	0.472	0.459	0.013	0.560	0.544	0.016
PlC-30	0.488	0.420	0.068	0.484	0.422	-0.008	0.511	0.470	0.041	0.532	0.540	-0.008
PlC-60	0.488	0.433	0.055	0.495	0.427	-0.039	0.474	0.528	-0.054	0.500	0.539	-0.039
PrC-1.5	0.493	0.433	0.060	0.494	0.426	-0.039	0.554	0.516	0.038	0.407	0.446	-0.039
PrC-3	0.492	0.421	0.071	0.497	0.425	0.023	0.486	0.520	-0.034	0.517	0.493	0.023
PrC-6	0.496	0.415	0.081	0.495	0.426	-0.055	0.508	0.463	0.046	0.389	0.444	-0.055
PrC-15	0.492	0.417	0.074	0.494	0.426	-0.022	0.440	0.398	0.042	0.466	0.487	-0.022
PrC-30	0.510	0.418	0.092	0.540	0.431	0.035	0.553	0.491	0.062	0.613	0.578	0.035
PrC-60	0.495	0.425	0.071	0.549	0.431	0.028	0.417	0.428	-0.011	0.626	0.599	0.028
S-1.5	0.645	**0.575**	0.070	0.623	0.552	0.093	0.876	**0.808**	0.068	0.855	0.762	0.093
S-3	0.619	0.570	0.049	**0.647**	0.551	0.127	0.875	0.779	0.096	0.841	0.713	0.127
S-6	0.617	0.555	0.061	0.614	**0.555**	0.102	0.861	0.804	0.057	0.833	0.731	0.102
S-15	0.612	0.537	0.075	0.600	0.554	0.081	0.841	0.744	0.096	0.812	0.731	0.081
S-30	0.594	0.536	0.058	0.578	0.528	0.061	0.811	0.710	0.102	0.765	0.705	0.061
S-60	0.543	0.485	0.058	0.578	0.514	0.093	0.601	0.578	0.023	0.770	0.676	0.093
SDS-5	0.613	0.567	0.045	0.569	0.513	0.072	0.850	0.804	0.046	0.786	0.713	0.072
SDS-11	0.631	**0.575**	0.056	0.599	0.543	0.080	0.846	0.786	0.061	0.803	0.723	0.080
SDS-21	0.633	0.573	0.060	0.591	0.552	0.074	0.854	0.786	0.067	0.809	0.735	0.074
SDS-51	0.603	0.533	0.069	0.554	0.536	0.038	0.841	0.746	0.095	0.727	0.689	0.038
SDS-101	0.611	0.571	**0.040**	0.535	0.502	0.037	0.848	0.756	0.092	0.718	0.681	0.037
SDS-201	0.613	0.527	0.086	0.548	0.508	0.064	0.837	0.713	0.124	0.735	0.671	0.064

Table 6: Macro-averaged precision and recall for *single modality* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones. WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	Precision (ResNeXt)			Precision (ViT)			Recall (ResNeXt)			Recall (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
DEM	0.621	0.460	0.161	0.551	0.432	0.125	0.661	0.653	0.008	0.800	0.674	0.125
RGB	0.553	0.405	0.148	0.522	0.296	0.235	0.672	0.418	0.254	0.664	0.429	0.235
NIR	0.564	0.486	0.078	0.521	0.273	0.384	0.698	0.514	0.184	0.668	0.284	0.384
NHD	0.419	0.353	0.066	0.390	0.334	0.056	0.725	0.691	0.034	0.857	0.881	-0.024
OSM	0.442	0.373	0.069	0.395	0.334	0.061	0.846	0.853	-0.007	0.971	0.949	0.022
EP-5	0.617	0.450	0.167	0.556	0.452	0.112	0.706	0.333	0.373	0.733	0.621	0.112
EP-11	0.602	0.474	0.128	0.552	0.449	0.060	0.748	0.428	0.320	0.690	0.631	0.060
EP-21	**0.629**	0.455	0.173	0.548	0.435	0.089	0.737	0.416	0.321	0.706	0.617	0.089
EP-51	0.612	0.382	0.230	0.565	0.440	0.087	0.705	0.389	0.316	0.664	0.577	0.087
EP-101	0.570	0.480	0.090	0.539	0.421	0.102	0.727	0.551	0.176	0.674	0.572	0.102
EP-201	0.593	0.465	0.127	0.520	0.425	0.092	0.634	0.364	0.270	0.707	0.615	0.092
PlC-1.5	0.390	0.333	0.057	0.419	0.359	0.078	0.837	0.829	0.007	0.806	0.728	0.078
PlC-3	0.391	0.333	0.059	0.432	0.370	0.119	**1.000**	**1.000**	**0.000**	0.871	0.752	0.119
PlC-6	0.393	0.333	0.060	0.429	0.365	0.052	0.892	0.889	0.003	0.853	0.801	0.052
PlC-30	0.390	0.332	0.058	0.392	0.334	-0.045	0.856	0.809	0.047	0.795	0.840	-0.045
PlC-15	0.390	0.334	0.057	0.403	0.338	-0.029	0.823	0.834	-0.010	0.765	0.794	-0.029
PlC-60	0.389	0.337	**0.052**	0.393	0.335	-0.022	0.842	0.921	-0.079	0.973	0.995	-0.022
PrC-1.5	0.392	0.341	**0.052**	0.391	0.333	**0.000**	0.967	0.946	0.021	**1.000**	**1.000**	**0.000**
PrC-3	0.394	0.335	0.059	0.406	0.336	**0.000**	0.819	0.853	-0.034	0.919	0.919	**0.000**
PrC-6	0.396	0.328	0.068	0.392	0.333	-0.001	0.739	0.719	0.020	0.997	0.998	-0.001
PrC-15	0.392	0.331	0.061	0.391	0.333	**0.000**	0.759	0.718	0.041	**1.000**	**1.000**	**0.000**
PrC-30	0.430	0.337	0.092	0.456	0.348	0.074	0.679	0.639	0.040	0.731	0.657	0.074
PrC-60	0.392	0.332	0.060	0.464	0.350	0.100	0.896	0.854	0.042	0.748	0.648	0.100
S-1.5	0.616	0.506	0.110	0.578	0.489	0.051	0.681	0.687	-0.006	0.726	0.674	0.051
S-3	0.590	**0.507**	0.084	**0.614**	0.490	0.041	0.654	0.662	-0.009	0.693	0.653	0.041
S-6	0.592	0.497	0.095	0.553	**0.491**	0.072	0.670	0.671	0.001	0.791	0.720	0.072
S-15	0.550	0.478	0.072	0.537	0.484	-0.027	0.749	0.664	0.085	0.774	0.801	-0.027
S-30	0.523	0.464	0.059	0.508	0.464	0.054	0.744	0.679	0.065	0.717	0.663	0.054
S-60	0.469	0.409	0.060	0.500	0.436	0.064	0.697	0.651	0.047	0.736	0.672	0.064
SDS-5	0.580	0.487	0.093	0.518	0.435	-0.025	0.661	0.707	-0.047	0.641	0.666	-0.025
SDS-11	0.596	0.499	0.097	0.545	0.460	0.084	0.689	0.698	-0.008	0.769	0.685	0.084
SDS-21	0.578	0.486	0.092	0.529	0.469	-0.006	0.768	0.740	0.027	0.690	0.696	-0.006
SDS-51	0.578	0.471	0.108	0.482	0.443	0.022	0.638	0.646	-0.008	0.740	0.718	0.022
SDS-101	0.566	0.490	0.075	0.459	0.409	-0.009	0.775	0.716	0.058	0.710	0.719	-0.009
SDS-201	0.558	0.452	0.107	0.459	0.411	0.044	0.709	0.660	0.048	0.796	0.752	0.044

Table 7: Mean average precision (mAP) and macro-averaged accuracy for *single modality* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones. WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	mAP (ResNeXt)			mAP (ViT)			Accuracy (ResNeXt)			Accuracy (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
DEM	0.554	0.442	0.111	0.516	0.431	0.022	**0.873**	0.827	0.046	0.808	0.785	0.022
RGB	0.509	0.367	0.143	0.489	0.336	0.109	0.832	0.781	0.051	0.815	0.706	0.109
NIR	0.513	0.387	0.125	0.485	0.337	0.020	0.833	0.809	0.025	0.812	0.792	0.020
NHD	0.403	0.339	0.064	0.391	0.333	0.058	0.682	0.634	0.048	0.523	0.468	0.055
OSM	0.435	0.367	0.068	0.395	0.334	0.061	0.647	0.548	0.099	0.545	0.406	0.139
EP-5	0.549	0.385	0.164	0.516	0.417	0.019	0.858	0.831	0.026	0.829	0.810	0.019
EP-11	0.551	0.397	0.154	0.510	0.409	0.024	0.854	0.832	0.022	0.829	0.805	0.024
EP-21	**0.565**	0.386	0.179	0.504	0.398	0.029	0.860	0.828	0.031	0.827	0.798	0.029
EP-51	0.546	0.377	0.169	0.507	0.395	0.034	0.862	0.818	0.044	0.837	0.803	0.034
EP-101	0.528	0.401	0.128	0.500	0.385	0.034	0.835	0.812	0.024	0.818	0.784	0.034
EP-201	0.535	0.381	0.154	0.476	0.367	0.041	0.858	0.838	0.019	0.791	0.750	0.041
PlC-1.5	0.391	0.333	0.058	0.411	0.354	0.015	0.551	0.502	0.049	0.643	0.628	0.015
PlC-3	0.391	0.333	0.059	0.418	0.353	0.005	0.392	0.333	0.059	0.631	0.626	0.005
PlC-6	0.393	0.333	0.060	0.416	0.353	**-0.001**	0.494	0.452	0.043	0.617	0.619	**-0.001**
PlC-15	0.391	0.334	0.057	0.397	0.335	0.053	0.533	0.482	0.051	0.644	0.591	0.053
PlC-30	0.392	0.333	0.059	0.392	0.334	0.064	0.524	0.467	0.057	0.586	0.521	0.064
PlC-60	0.390	0.335	0.055	0.393	0.335	0.062	0.525	0.471	0.054	0.456	0.395	0.062
PrC-1.5	0.392	0.340	0.052	0.391	0.333	0.059	0.411	0.402	**0.009**	0.392	0.333	0.059
PrC-3	0.393	0.332	0.060	0.400	0.334	0.051	0.527	0.466	0.061	0.452	0.401	0.051
PrC-6	0.392	0.333	0.059	0.392	0.333	0.062	0.645	0.581	0.064	0.395	0.334	0.062
PrC-15	0.393	0.334	0.059	0.391	0.333	0.059	0.644	0.591	0.054	0.392	0.333	0.059
PrC-30	0.406	0.339	0.067	0.431	0.345	0.055	0.714	0.674	0.040	0.726	0.671	0.055
PrC-60	0.392	0.333	0.059	0.433	0.345	0.045	0.510	0.463	0.047	0.723	0.677	0.045
S-1.5	0.552	0.468	0.084	0.525	0.456	0.021	0.871	0.848	0.023	0.840	0.819	0.021
S-3	0.543	**0.472**	0.071	**0.542**	0.465	0.025	0.867	**0.852**	0.015	**0.850**	**0.825**	0.025
S-6	0.539	0.463	0.077	0.523	**0.466**	0.019	0.857	0.844	0.013	0.812	0.793	0.019
S-15	0.517	0.455	0.062	0.506	0.463	0.012	0.807	0.799	0.008	0.794	0.781	0.012
S-30	0.501	0.447	0.053	0.485	0.452	**-0.001**	0.793	0.784	0.009	0.792	0.793	**-0.001**
S-60	0.450	0.398	0.052	0.481	0.435	0.003	0.742	0.752	-0.010	0.784	0.780	0.003
SDS-5	0.527	0.459	0.068	0.484	0.420	0.011	0.853	0.833	0.020	0.820	0.809	0.011
SDS-11	0.533	0.466	0.068	0.504	0.434	0.011	0.850	0.839	0.011	0.806	0.795	0.011
SDS-21	0.531	0.454	0.078	0.491	0.435	0.007	0.836	0.819	0.017	0.816	0.809	0.007
SDS-51	0.529	0.436	0.093	0.459	0.418	0.002	0.855	0.824	0.031	0.754	0.752	0.002
SDS-101	0.525	0.461	0.064	0.448	0.400	-0.017	0.820	0.808	0.012	0.734	0.751	-0.017
SDS-201	0.520	0.427	0.093	0.446	0.402	-0.019	0.834	0.805	0.030	0.710	0.729	-0.019

Table 8: Macro-averaged F1 and AUC for *multi-scale fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under two fusion strategies: channel stacking of input features (St) and cross-attention with a shared encoder (A1). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	F1 (ResNeXt)			F1 (ViT)			AUC (ResNeXt)			AUC (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms (St)	**0.640**	0.425	0.215	0.566	0.458	0.108	0.862	0.717	0.145	0.756	0.693	0.063
PlC-ms (St)	0.490	0.426	0.063	0.493	0.429	0.063	0.525	0.521	0.004	0.511	0.536	-0.026
PrC-ms (St)	0.519	0.441	0.078	0.596	0.501	0.095	0.579	0.497	0.082	**0.816**	**0.727**	0.089
S-ms (St)	0.637	**0.594**	0.043	0.593	0.533	0.061	0.864	**0.804**	0.061	0.798	0.705	0.093
SDS-ms (St)	0.636	0.588	0.048	**0.619**	**0.571**	0.048	**0.878**	0.792	0.086	0.672	0.644	0.028
EP-ms (A1)	0.494	0.426	0.068	0.561	0.445	0.117	0.500	0.500	**0.000**	0.759	0.664	0.095
PlC-ms (A1)	0.494	0.426	0.068	0.505	0.435	0.070	0.500	0.500	**0.000**	0.578	0.581	-0.003
PrC-ms (A1)	0.494	0.426	0.068	0.531	0.410	0.121	0.500	0.500	**0.000**	0.594	0.562	0.032
S-ms (A1)	0.494	0.426	0.068	0.557	0.519	**0.038**	0.500	0.500	**0.000**	0.615	0.594	0.021
SDS-ms (A1)	0.493	0.451	**0.042**	0.494	0.426	0.068	0.618	0.618	0.001	0.500	0.500	**0.000**

Table 9: Macro-averaged precision and recall for *multi-scale fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under two fusion strategies: channel stacking of input features (St) and cross-attention with a shared encoder (A1). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	Precision (ResNeXt)			Precision (ViT)			Recall (ResNeXt)			Recall (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms (St)	0.606	**0.556**	**0.051**	0.493	0.380	0.112	0.703	0.426	0.277	0.712	0.636	0.076
PlC-ms (St)	0.391	0.335	0.056	0.391	0.335	0.056	0.738	0.738	**0.000**	0.872	0.940	-0.067
PrC-ms (St)	0.429	0.353	0.076	0.530	0.435	0.095	0.697	0.694	0.003	0.743	0.642	0.101
S-ms (St)	**0.607**	0.535	0.072	0.525	0.455	0.070	0.730	0.682	0.047	0.714	0.681	0.033
SDS-ms (St)	0.588	0.509	0.079	**0.575**	**0.472**	0.103	0.742	0.729	0.013	0.675	0.674	0.001
EP-ms (A1)	0.391	0.333	0.059	0.483	0.375	0.108	**1.000**	**1.000**	**0.000**	0.700	0.612	0.088
PlC-ms (A1)	0.391	0.333	0.059	0.405	0.341	0.064	**1.000**	**1.000**	**0.000**	0.874	0.868	0.006
PrC-ms (A1)	0.391	0.333	0.059	0.431	0.325	0.106	**1.000**	**1.000**	**0.000**	0.738	0.678	0.060
S-ms (A1)	0.391	0.333	0.058	0.489	0.440	**0.049**	**1.000**	**1.000**	**0.000**	0.745	0.688	0.057
SDS-ms (A1)	0.432	0.380	0.052	0.391	0.332	0.057	0.801	0.748	0.053	**1.000**	**1.000**	**0.000**

Table 10: Mean average precision (mAP) and macro-averaged accuracy for *multi-scale fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under two fusion strategies: channel stacking of input features (St) and cross-attention with a shared encoder (A1). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	mAP (ResNeXt)			mAP (ViT)			Accuracy (ResNeXt)			Accuracy (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms (St)	0.555	0.403	0.152	0.460	0.360	0.099	**0.865**	0.828	0.037	0.774	0.724	0.050
PlC-ms (St)	0.392	0.335	**0.057**	0.392	0.335	0.057	0.634	0.588	0.046	0.534	0.465	0.069
PrC-ms (St)	0.416	0.348	0.069	0.504	0.423	0.081	0.717	0.666	0.051	0.794	0.768	0.027
S-ms (St)	**0.557**	**0.491**	0.066	0.498	**0.453**	0.045	0.856	**0.860**	**-0.004**	0.810	0.803	0.006
SDS-ms (St)	0.540	0.470	0.070	**0.522**	0.447	0.075	0.846	0.839	0.007	**0.851**	**0.826**	0.025
EP-ms (A1)	0.391	0.333	0.059	0.450	0.362	0.088	0.391	0.333	0.059	0.766	0.727	0.039
PlC-ms (A1)	0.391	0.333	0.059	0.401	0.338	0.062	0.391	0.333	0.059	0.598	0.541	0.057
PrC-ms (A1)	0.391	0.333	0.059	0.407	0.333	0.074	0.391	0.333	0.059	0.691	0.625	0.065
S-ms (A1)	0.391	0.333	0.058	0.472	0.434	**0.038**	0.391	0.333	0.059	0.742	0.747	**-0.005**
SDS-ms (A1)	0.416	0.357	0.059	0.391	0.333	0.058	0.630	0.666	-0.036	0.391	0.333	0.058

Table 11: Macro-averaged F1 and AUC for *multimodal fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under four fusion strategies: channel stacking of input features (St), concatenation of modality embeddings (C), cross-attention with a shared encoder (A1), and cross-attention with separate encoders (A2). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	F1 (ResNeXt)			F1 (ViT)			AUC (ResNeXt)			AUC (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms+S-ms+SDS-ms (St)	**0.657**	**0.598**	0.059	0.621	**0.569**	0.053	0.882	0.806	0.076	0.860	**0.774**	0.086
EP-5+S-1.5+SDS-5 (St)	0.641	0.568	0.073	**0.657**	0.566	0.092	0.848	**0.812**	0.036	0.712	0.664	0.048
EP-201+S-60+SDS-201 (St)	0.626	0.582	0.045	0.622	0.544	0.078	**0.885**	**0.812**	0.073	0.695	0.631	0.064
EP-ms+S-ms+SDS-ms (C)	0.596	0.569	**0.028**	0.613	0.532	0.081	0.829	0.750	0.079	0.686	0.622	0.064
RGB+DEM (C)	0.600	0.389	0.211	0.614	0.503	0.111	0.808	0.535	0.273	**0.870**	0.721	0.149
RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.618	0.543	0.074	0.621	0.528	0.093	0.858	0.739	0.118	0.735	0.615	0.120
EP-ms+S-ms+SDS-ms (A1)	0.561	0.532	0.029	0.567	0.538	**0.029**	0.677	0.707	-0.030	0.776	0.678	0.098
RGB+DEM (A1)	0.551	0.457	0.094	0.575	0.404	0.171	0.714	0.552	0.163	0.787	0.622	0.165
EP-ms+S-ms+SDS-ms (A2)	0.561	0.532	0.029	0.496	0.425	0.071	0.677	0.707	-0.030	0.523	0.480	**0.043**
RGB+DEM (A2)	0.559	0.474	0.085	0.581	0.464	0.118	0.763	0.641	0.122	0.810	0.724	0.085
RGB+DEM+EP-ms+S-ms+SDS-ms (A2)	0.494	0.426	0.068	0.520	0.457	0.063	0.500	0.500	**0.000**	0.572	0.511	0.061

Table 12: Macro-averaged precision and recall for *multimodal fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under four fusion strategies: channel stacking of input features (St), concatenation of modality embeddings (C), cross-attention with a shared encoder (A1), and cross-attention with separate encoders (A2). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	Precision (ResNeXt)			Precision (ViT)			Recall (ResNeXt)			Recall (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms+S-ms+SDS-ms (St)	**0.626**	**0.546**	0.080	0.568	0.491	0.077	0.735	0.666	0.068	0.761	0.711	0.050
EP-5+S-1.5+SDS-5 (St)	0.606	0.531	0.074	**0.604**	0.482	0.122	0.697	0.623	0.074	0.731	0.708	0.023
EP-201+S-60+SDS-201 (St)	0.588	0.529	0.059	0.579	**0.499**	0.080	0.721	0.674	0.048	0.686	0.610	0.076
EP-ms+S-ms+SDS-ms (C)	0.542	0.529	**0.013**	0.541	0.456	0.085	0.694	0.640	0.054	0.752	0.671	0.081
RGB+DEM (C)	0.537	0.373	0.163	0.558	0.420	0.137	0.715	0.437	0.278	0.706	0.661	0.045
RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.563	0.496	0.067	0.574	0.485	0.090	0.740	0.644	0.096	0.621	0.622	**-0.001**
EP-ms+S-ms+SDS-ms (A1)	0.487	0.451	0.036	0.507	0.466	0.041	0.734	0.723	0.011	0.752	0.693	0.059
RGB+DEM (A1)	0.495	0.445	0.050	0.515	0.387	0.129	0.647	0.555	0.092	0.686	0.582	0.105
EP-ms+S-ms+SDS-ms (A2)	0.487	0.451	0.036	0.392	0.332	0.060	0.734	0.723	0.011	**0.984**	**0.889**	0.095
RGB+DEM (A2)	0.498	0.411	0.087	0.513	0.434	0.079	0.656	0.595	0.061	0.720	0.607	0.113
RGB+DEM+EP-ms+S-ms+SDS-ms (A2)	0.391	0.333	0.059	0.448	0.420	**0.028**	**1.000**	**1.000**	**0.000**	0.873	0.689	0.184

Table 13: Mean average precision (mAP) and macro-averaged accuracy for *multimodal fusion* models on in-domain (WC) and cross-domain (HC) test sets. Results are reported for ResNeXt-50 and ViT-B/16 backbones under four fusion strategies: channel stacking of input features (St), concatenation of modality embeddings (C), cross-attention with a shared encoder (A1), and cross-attention with separate encoders (A2). WC–HC differences (Δ) are also shown. The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	mAP (ResNeXt)			mAP (ViT)			Accuracy (ResNeXt)			Accuracy (ViT)		
	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ	WC	HC	Δ
EP-ms+S-ms+SDS-ms (St)	**0.571**	**0.495**	0.076	0.534	0.463	0.070	**0.875**	**0.867**	0.008	0.834	0.823	0.011
EP-5+S-1.5+SDS-5 (St)	0.551	0.471	0.080	**0.540**	0.461	0.079	0.865	0.856	0.009	0.712	0.664	0.048
EP-201+S-60+SDS-201 (St)	0.552	0.480	0.072	0.532	**0.468**	0.064	0.858	0.852	0.006	**0.851**	**0.840**	0.011
EP-ms+S-ms+SDS-ms (C)	0.505	0.451	0.053	0.508	0.450	0.058	0.822	0.836	-0.015	0.817	0.806	0.011
RGB+DEM (C)	0.495	0.360	0.135	0.524	0.415	0.109	0.815	0.809	0.007	0.838	0.796	0.042
RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.525	0.458	0.067	0.537	0.449	0.088	0.833	0.805	0.028	0.827	0.824	0.003
EP-ms+S-ms+SDS-ms (A1)	0.474	0.442	**0.033**	0.488	0.456	**0.032**	0.747	0.758	-0.011	0.750	0.752	**-0.002**
RGB+DEM (A1)	0.459	0.389	0.070	0.478	0.360	0.118	0.784	0.776	0.008	0.799	0.745	0.054
EP-ms+S-ms+SDS-ms (A2)	0.474	0.442	**0.033**	0.392	0.333	0.059	0.747	0.758	-0.011	0.452	0.402	0.050
RGB+DEM (A2)	0.464	0.389	0.075	0.486	0.388	0.098	0.795	0.793	0.002	0.795	0.775	0.020
RGB+DEM+EP-ms+S-ms+SDS-ms (A2)	0.391	0.333	0.059	0.422	0.368	0.054	0.391	0.333	0.059	0.603	0.620	-0.017

Table 14: Class-wise AUC scores for in-domain (Warren County region) performance across single-modality, multi-scale fusion, and multimodal fusion models. Results are reported for ResNeXt-50 and ViT-B/16 backbones under four fusion strategies: channel stacking of input features (St), concatenation of modality embeddings (C), cross-attention with a shared encoder (A1), and cross-attention with separate encoders (A2). The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	ResNeXt							ViT						
	af1	Qal	Qaf	Qat	Qc	Qca	Qr	af1	Qal	Qaf	Qat	Qc	Qca	Qr
DEM	0.845	0.832	0.820	0.887	0.964	0.922	0.910	0.663	0.771	**0.926**	0.871	0.956	0.923	0.888
RGB	0.834	0.713	0.684	0.815	0.912	0.857	0.886	0.816	0.679	0.744	0.780	0.891	0.834	0.805
NIR	0.816	0.698	0.782	0.793	0.907	0.866	0.842	0.760	0.664	0.797	0.799	0.886	0.816	0.763
NHD	0.549	0.655	0.682	0.782	0.618	0.630	0.697	0.497	0.571	0.441	0.354	0.506	0.502	0.505
OSM	0.807	0.586	0.702	0.586	0.708	0.627	0.557	0.505	0.484	0.693	0.606	0.5	0.513	0.487
EP-5	0.837	0.805	0.845	0.845	0.947	0.905	0.920	0.791	0.783	0.838	0.865	0.914	0.885	0.903
EP-11	**0.868**	0.816	0.833	0.888	0.936	0.905	0.902	0.778	0.781	0.834	0.882	0.891	0.889	0.898
EP-21	0.856	0.807	0.842	0.883	0.945	0.908	0.900	0.783	0.776	0.799	0.858	0.888	0.885	0.880
EP-51	0.860	0.825	0.827	0.870	0.921	0.906	0.924	0.794	0.766	0.791	0.858	0.877	0.888	0.870
EP-101	0.853	0.806	0.759	0.886	0.904	0.904	0.890	0.757	0.751	0.758	0.860	0.850	0.884	0.874
EP-201	0.846	0.812	0.844	0.879	0.901	0.894	0.904	0.734	0.750	0.756	0.830	0.789	0.872	0.864
PlC-1.5	0.440	0.491	0.610	0.515	0.513	0.514	0.516	0.438	0.509	0.719	0.610	0.575	0.725	0.645
PlC-3	0.501	0.501	0.500	0.500	0.501	0.501	0.500	0.445	0.494	0.769	0.675	0.499	0.773	0.689
PlC-6	0.459	0.516	0.491	0.497	0.455	0.505	0.490	0.451	0.478	0.746	0.712	0.668	0.719	0.649
PlC-15	0.526	0.505	0.362	0.387	0.547	0.476	0.500	0.466	0.523	0.655	0.620	0.578	0.575	0.505
PlC-30	0.517	0.490	0.604	0.473	0.501	0.524	0.465	0.469	0.567	0.650	0.515	0.531	0.529	0.465
PlC-60	0.462	0.413	0.617	0.414	0.479	0.494	0.439	0.461	0.627	0.620	0.382	0.524	0.482	0.402
PrC-1.5	0.465	0.566	0.569	0.473	0.564	0.516	0.724	0.444	0.545	0.546	0.236	0.501	0.347	0.233
PrC-3	0.549	0.555	0.324	0.537	0.341	0.554	0.539	0.545	0.501	0.630	0.400	0.420	0.613	0.508
PrC-6	0.526	0.494	0.445	0.503	0.472	0.539	0.579	0.493	0.602	0.487	0.190	0.541	0.224	0.186
PrC-15	0.443	0.423	0.602	0.522	0.145	0.377	0.567	0.501	0.429	0.477	0.378	0.501	0.499	0.476
PrC-30	0.515	0.432	0.465	0.608	0.530	0.681	0.640	0.501	0.341	0.523	0.845	0.512	0.738	0.833
PrC-60	0.482	0.499	0.494	0.244	0.473	0.474	0.253	0.511	0.326	0.558	0.859	0.601	0.682	0.846
S-1.5	0.863	0.800	0.813	0.870	0.968	0.905	0.910	0.794	0.748	0.853	0.854	0.974	0.900	0.864
S-3	0.816	0.805	0.840	0.870	0.971	0.915	0.908	0.770	0.759	0.772	0.829	0.975	0.910	0.868
S-6	0.778	0.809	0.764	0.877	0.974	0.921	0.905	0.718	0.765	0.809	0.853	0.975	0.910	0.803
S-15	0.648	0.788	0.842	0.873	0.966	**0.926**	0.842	0.641	0.750	0.826	0.796	0.974	0.908	0.789
S-30	0.619	0.750	0.803	0.831	0.957	0.912	0.807	0.623	0.707	0.791	0.725	0.947	0.869	0.696
S-60	0.416	0.535	0.681	0.595	0.838	0.815	0.324	0.626	0.666	0.818	0.750	0.909	0.880	0.738
SDS-5	0.855	0.733	0.789	0.860	0.944	0.890	0.883	0.772	0.665	0.800	0.757	0.921	0.833	0.751
SDS-11	0.839	0.751	0.774	0.866	0.946	0.877	0.871	0.792	0.671	0.817	0.757	0.933	0.853	0.800
SDS-21	0.842	0.750	0.842	0.841	0.953	0.889	0.860	0.769	0.685	0.853	0.767	0.934	0.837	0.816
SDS-51	0.832	0.719	0.851	0.800	0.951	0.883	0.852	0.675	0.620	0.777	0.684	0.889	0.759	0.689
SDS-101	0.814	0.732	0.860	0.813	0.964	0.882	0.874	0.659	0.608	0.804	0.659	0.891	0.751	0.655
SDS-201	0.802	0.679	0.812	0.833	0.967	0.897	0.870	0.633	0.605	0.855	0.666	0.913	0.741	0.729
EP-ms (St)	0.823	0.824	0.734	0.878	0.945	0.911	0.917	0.823	0.824	0.734	0.878	0.945	0.911	**0.917**
PlC-ms (St)	0.504	0.500	0.641	0.501	0.514	0.500	0.514	0.504	0.500	0.641	0.501	0.514	0.500	0.514
PrC-ms (St)	0.494	0.653	0.567	0.721	0.628	0.791	0.201	0.494	0.653	0.567	0.721	0.628	0.791	0.201
S-ms (St)	0.863	0.787	0.760	0.870	0.962	0.911	0.900	**0.863**	0.787	0.760	0.870	0.962	0.911	0.900
SDS-ms (St)	0.839	0.766	**0.917**	0.876	0.964	0.898	0.889	0.839	0.766	0.917	0.876	0.964	0.898	0.889
EP-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
PlC-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
PrC-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
S-ms (A1)	0.499	0.501	0.500	0.500	0.501	0.499	0.500	0.499	0.501	0.500	0.500	0.501	0.499	0.500
SDS-ms (A1)	0.552	0.576	0.801	0.602	0.679	0.540	0.580	0.552	0.576	0.801	0.602	0.679	0.540	0.580
EP-ms+S-ms+SDS-ms (St)	0.866	**0.840**	0.790	0.858	**0.975**	0.913	**0.933**	0.780	0.772	0.864	0.847	**0.976**	0.890	0.890
EP-5+S-1.5+SDS-5 (St)	0.845	0.797	0.712	0.829	0.964	0.904	0.886	0.837	**0.803**	0.858	0.884	0.974	0.912	0.901
EP-201+S-60+SDS-201 (St)	0.846	0.802	0.840	**0.903**	0.961	0.911	0.933	0.752	0.799	0.848	0.856	0.967	**0.937**	0.905
EP-ms+S-ms+SDS-ms (C)	0.723	0.802	0.746	0.809	0.959	0.879	0.885	0.728	0.720	0.871	0.816	0.969	0.890	0.898
RGB+DEM (C)	0.821	0.708	0.804	0.803	0.871	0.845	0.803	0.800	0.756	0.874	**0.899**	0.949	0.901	0.911
RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.837	0.774	0.842	0.827	0.963	0.899	0.860	0.746	0.755	0.875	0.878	0.975	0.910	0.921
EP-ms+S-ms+SDS-ms (A1)	0.486	0.575	0.726	0.641	0.930	0.784	0.599	0.698	0.623	0.831	0.660	0.961	0.879	0.785
RGB+DEM (A1)	0.687	0.476	0.747	0.762	0.837	0.801	0.692	0.711	0.629	0.813	0.815	0.886	0.842	0.811
EP-ms+S-ms+SDS-ms (A2)	0.486	0.575	0.726	0.641	0.930	0.784	0.599	0.500	0.500	0.623	0.534	0.500	0.500	0.494
RGB+DEM (A2)	0.752	0.617	0.780	0.816	0.825	0.786	0.764	0.704	0.692	0.856	0.806	0.930	0.841	0.839
RGB+DEM+EP-ms+S-ms+SDS-ms (A2)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.501	0.477	0.768	0.557	0.499	0.753	0.626

Table 15: Class-wise AUC scores for cross-domain (Hardin County region) performance across single-modality, multi-scale fusion, and multimodal fusion models. Results are reported for ResNeXt-50 and ViT-B/16 backbones under four fusion strategies: channel stacking of input features (St), concatenation of modality embeddings (C), cross-attention with a shared encoder (A1), and cross-attention with separate encoders (A2). The best and second-best scores in each column are indicated in **bold** and underlined, respectively.

Model	ResNeXt							ViT						
	af1	Qal	Qaf	Qat	Qc	Qca	Qr	af1	Qal	Qaf	Qat	Qc	Qca	Qr
DEM	0.804	0.613	0.612	0.472	0.969	0.907	0.733	0.587	0.549	0.379	0.210	0.958	0.947	0.710
RGB	0.757	0.576	0.403	0.486	0.654	0.515	0.507	0.575	0.527	0.782	0.650	0.270	0.381	0.494
NIR	0.733	0.519	0.490	0.550	0.703	0.824	0.727	0.502	0.578	0.474	0.641	0.466	0.348	0.554
NHD	0.556	0.642	0.630	0.722	0.485	0.494	0.504	0.494	0.506	0.538	0.498	0.516	0.510	0.618
OSM	0.833	0.518	0.479	0.572	0.624	0.586	0.496	0.503	0.5	0.543	0.553	0.5	0.505	0.584
EP-5	0.769	0.635	0.782	**0.847**	0.291	0.352	0.399	0.764	0.651	0.626	0.622	0.860	0.882	0.757
EP-11	0.790	0.687	0.801	0.763	0.463	0.563	0.662	0.763	0.698	0.734	0.667	0.807	0.870	0.840
EP-21	0.818	0.700	0.846	0.746	0.392	0.668	0.694	0.778	0.696	0.725	0.662	0.817	0.842	0.796
EP-51	0.821	0.676	0.769	0.778	0.409	0.519	0.672	0.779	0.633	0.798	0.684	0.771	0.851	0.786
EP-101	0.851	0.716	0.726	0.769	0.621	0.748	0.742	0.745	0.633	0.815	0.629	0.759	0.842	0.789
EP-201	0.786	0.737	0.805	0.752	0.573	0.697	0.717	0.698	0.676	0.821	0.729	0.718	0.818	0.701
PlC-1.5	0.492	0.501	0.599	0.548	0.487	0.509	0.453	0.514	0.340	0.650	0.518	0.561	0.792	0.752
PlC-3	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.511	0.305	0.733	0.638	0.500	0.791	0.819
PlC-6	0.517	0.480	0.529	0.478	0.474	0.492	0.426	0.530	0.304	0.758	0.701	0.627	0.703	0.766
PlC-15	0.511	0.464	0.275	0.397	0.557	0.497	0.511	0.517	0.470	0.711	0.600	0.537	0.532	0.442
PlC-30	0.513	0.514	0.324	0.516	0.497	0.472	0.454	0.517	0.527	0.809	0.512	0.536	0.527	0.349
PlC-60	0.510	0.472	0.899	0.537	0.465	0.503	0.311	0.501	0.562	0.831	0.515	0.554	0.523	0.285
PrC-1.5	0.426	0.559	0.263	0.418	0.679	0.710	0.559	0.412	0.633	0.219	0.362	0.500	0.592	0.404
PrC-3	0.597	0.379	0.797	0.612	0.277	0.363	0.614	0.574	0.508	0.507	0.448	0.372	0.539	0.505
PrC-6	0.498	0.490	0.408	0.414	0.478	0.491	0.459	0.417	0.644	0.348	0.468	0.584	0.393	0.256
PrC-15	0.493	0.493	0.426	0.551	0.136	0.248	0.438	0.500	0.458	0.476	0.496	0.500	0.500	0.482
PrC-30	0.506	0.448	0.150	0.505	0.552	0.631	0.646	0.532	0.428	0.566	0.528	0.463	0.664	0.867
PrC-60	0.467	0.543	0.464	0.435	0.431	0.429	0.225	0.534	0.424	0.569	0.574	0.573	0.612	**0.905**
S-1.5	0.863	0.737	0.611	0.754	0.975	0.915	0.801	0.759	0.579	0.646	0.667	0.981	0.923	0.778
S-3	0.781	0.731	0.531	0.696	0.976	0.922	0.815	0.683	0.563	0.528	0.530	0.981	0.937	0.772
S-6	0.713	0.704	**0.889**	0.706	0.976	0.924	0.717	0.621	0.569	0.786	0.708	0.981	0.941	0.509
S-15	0.625	0.619	0.674	0.665	0.974	0.936	0.718	0.529	0.551	**0.964**	0.477	0.971	0.952	0.673
S-30	0.550	0.549	0.746	0.537	0.965	**0.945**	0.675	0.533	0.559	0.704	0.372	0.945	0.959	0.859
S-60	0.467	0.545	0.541	0.365	0.802	0.890	0.435	0.524	0.533	0.607	0.348	0.919	**0.962**	0.842
SDS-5	0.858	0.637	0.805	0.737	0.963	0.886	0.744	0.776	0.561	0.503	0.593	0.958	0.864	0.739
SDS-11	**0.861**	0.671	0.587	0.701	0.971	0.905	0.804	0.762	0.538	0.556	0.631	0.957	0.863	0.753
SDS-21	0.838	0.673	0.749	0.794	0.969	0.869	0.613	0.741	0.543	0.658	0.694	0.952	0.853	0.704
SDS-51	0.822	0.649	0.608	0.605	0.959	0.834	0.749	0.670	0.515	0.511	0.673	0.943	0.824	0.686
SDS-101	0.809	0.611	0.443	0.788	0.960	0.886	0.795	0.656	0.474	0.491	0.644	0.954	0.871	0.677
SDS-201	0.752	0.579	0.503	0.645	0.964	0.804	0.744	0.641	0.479	0.477	0.640	0.942	0.870	0.647
EP-ms (St)	0.769	0.722	0.828	0.722	0.603	0.701	0.671	0.769	**0.722**	0.828	0.722	0.603	0.701	0.671
PlC-ms (St)	0.479	0.524	0.603	0.489	0.553	0.567	0.432	0.479	0.524	0.603	0.489	0.553	0.567	0.432
PrC-ms (St)	0.496	0.567	0.301	0.440	0.687	0.788	0.202	0.496	0.567	0.301	0.440	0.687	0.788	0.202
S-ms (St)	0.881	0.711	0.643	0.741	**0.977**	0.915	0.759	**0.881**	0.711	0.643	**0.741**	0.977	0.915	0.759
SDS-ms (St)	0.843	0.679	0.629	0.762	0.966	0.889	0.777	0.843	0.679	0.629	0.762	0.966	0.889	0.777
EP-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
PlC-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
PrC-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
S-ms (A1)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
SDS (A1)	0.558	0.592	0.699	0.679	0.626	0.602	0.568	0.558	0.592	0.699	0.679	0.626	0.602	0.568
EP-ms+S-ms+SDS-ms (St)	0.857	**0.760**	0.612	0.736	0.972	0.914	0.792	0.734	0.586	0.740	0.650	**0.982**	0.922	0.805
EP-5+S-1.5+SDS-5 (St)	0.860	0.638	0.735	0.760	0.960	0.899	0.833	0.848	0.683	0.685	0.697	0.980	0.922	0.803
EP-201+S-60+SDS-201 (St)	0.859	0.717	0.699	0.685	0.962	0.911	**0.855**	0.657	0.587	0.748	0.646	0.976	**0.962**	0.879
EP-ms+S-ms+SDS-ms (C)	0.701	0.693	0.498	0.689	0.962	0.902	0.804	0.679	0.577	0.633	0.582	0.973	0.938	0.765
RGB+DEM (C)	0.788	0.460	0.173	0.406	0.661	0.621	0.635	0.752	0.554	0.545	0.611	0.930	0.923	0.732
RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.841	0.644	0.452	0.493	0.964	0.946	0.833	0.660	0.540	0.687	0.594	0.965	0.933	0.825
EP-ms+S-ms+SDS-ms (A1)	0.555	0.525	0.674	0.552	0.921	0.907	0.816	0.653	0.483	0.500	0.377	0.973	0.955	0.805
RGB+DEM (A1)	0.708	0.527	0.274	0.130	0.836	0.740	0.647	0.671	0.513	0.271	0.548	0.916	0.901	0.531
EP-ms+S-ms+SDS-ms (A2)	0.555	0.525	0.674	0.552	0.921	0.907	0.816	0.500	0.500	0.362	0.497	0.500	0.500	0.505
RGB+DEM (A2)	0.743	0.482	0.325	0.499	0.905	0.835	0.695	0.688	0.498	0.695	0.676	0.941	0.894	0.677
RGB+DEM+EP-ms+S-ms+SDS-ms (A2)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.515	0.451	0.250	0.350	0.500	0.860	0.670

Table 16: Macro-averaged F1 and AUC for DOFA (Xiong et al., 2024), Panopticon (Waldmann et al., 2025), SatMAE (Cong et al., 2022), SatMAE++ (Noman et al., 2024), and SGMap-Net architectures. Two SGMap-Net variants are shown: one with a comparable set of input modalities and one representing the best overall configuration. Fusion in SGMap-Net is implemented using either concatenation (C) or channel stacking (St). Metrics are reported for in-domain (Warren County, WC) and cross-domain (Hardin County, HC) test sets, with WC–HC differences (Δ) also shown. The best and second-best scores in each column are highlighted in **bold** and underlined, respectively.

Model	Modalities	F1			AUC		
		WC	HC	Δ	WC	HC	Δ
DOFA	RGB+NIR	0.597	0.533	0.064	0.652	0.623	**0.029**
Panopticon-FM	RGB+NIR	0.570	0.313	0.257	0.635	0.533	0.102
SatMAE	RGB+DEM+EP-5+EP-51+EP-101+EP-201+S-1.5+SDS-5	0.614	0.427	0.187	0.864	0.735	0.129
SatMAE++	RGB+DEM+EP-5+EP-51+EP-101+EP-201+S-1.5+SDS-5	0.656	0.454	0.202	**0.904**	0.762	0.142
SGMap-Net	RGB+DEM+EP-ms+S-ms+SDS-ms (C)	0.618	0.543	0.074	0.858	0.739	0.118
SGMap-Net	EP-ms+S-ms+SDS-ms (St)	**0.657**	**0.598**	**0.059**	0.882	**0.806**	0.076