remote sensing

MDPI

Article

State-Space-Based Mamba and Convolutional Neural Network Integration for Anthropogenic Geomorphic Feature Extraction from Land Surface Parameters

Sarah Farhadpour * and Aaron E. Maxwell

Department of Geology and Geography, West Virginia University, Morgantown, WV 26505, USA; aaron.maxwell@mail.wvu.edu

* Correspondence: sf00039@mix.wvu.edu

Highlights

What are the main findings?

- Across agricultural terraces, mine benches, and valley fill faces mapped from LiDAR-derived land surface parameters, no single architecture was best on every task; the hybrid Mamba–UNetC provided the most consistent performance overall, while a randomly initialized UNet remained highly competitive and achieved the strongest results on some tasks.
- Mamba-based models improved spatial continuity and interior recovery of elongated, fragmented features, whereas the convolutional UNet more precisely delineated feature boundaries; the pure Mamba–UNet attained competitive accuracy at substantially lower computational cost.

What are the implications of the main findings?

- State-space Mamba architectures are a competitive alternative to CNN-based models for high-resolution terrain segmentation tasks requiring broad spatial context, though CNNs remain competitive, particularly where precise boundary delineation matters.
- Benchmarking existing CNN, Mamba, and hybrid architectures provides practical guidance for architecture selection in scalable LiDAR-based geomorphic mapping across diverse landform types.

Abstract

Accurate and scalable extraction of geomorphic features from LiDAR-derived terrain data is important for environmental monitoring, land use planning, and geohazard assessment. Recent deep learning advances have introduced state-space models, such as Mamba, which can capture broad spatial dependencies with linear complexity and offer an alternative to conventional convolutional neural network (CNN)- or Transformer-based approaches. This study evaluates Mamba-based architectures for geomorphic segmentation across three landform types: agricultural terraces, mine benches, and valley fill faces. Using high spatial resolution terrain derivatives, or land surface parameters (LSPs), as an input feature space, we compare a pure Mamba–UNet, a hybrid architecture with a Mamba-based encoder and CNN decoder, and a baseline CNN-based UNet with a ResNet-34 encoder, each trained with frozen and unfrozen encoders. Results indicate that no single architecture was best across all tasks; however, the hybrid Mamba-CNN architecture provided the most consistent performance overall, improving spatial continuity and interior feature recovery for tasks requiring broad spatial context, although its advantage over the baseline UNet was modest and task-dependent, and boundary-level analysis indicated that CNN-based

Academic Editor: Gareth Rees

Received: 8 June 2026
Revised: 9 July 2026
Accepted: 15 July 2026
Published: 21 July 2026

configurations delineated feature edges most precisely, while the randomly initialized UNet remained highly competitive and achieved the strongest results for some tasks. Further, Mamba-based encoders converge quickly and maintain strong performance when frozen, whereas CNN-based UNets benefit more from encoder fine-tuning.

Keywords: LiDAR; land surface parameters; semantic segmentation; CNNs; UNet; state-space models; Mamba; geomorphic feature extraction

1. Introduction

1.1. Geomorphic Mapping

Accurate and scalable mapping of geomorphic features is critical for understanding Earth surface processes, monitoring environmental change, and supporting hazard mitigation and land management strategies. As remotely sensed datasets continue to expand in spatial and temporal resolution, the development of efficient and transferable deep learning models for terrain feature extraction has become increasingly essential for large-scale environmental monitoring and sustainable land use planning.

Convolutional neural networks (CNNs) have become the dominant architecture for semantic segmentation tasks in remote sensing due to their efficiency, strong local feature modeling, and success across a wide range of geospatial applications, including land cover classification [1,2], building change detection [3], landslide mapping [4,5], and extraction of geomorphic features [6,7]. Architectures such as UNet [8] and its variants have demonstrated strong performance when segmenting terrain-related patterns from elevation data and derived land surface parameters (LSPs) as opposed to imagery or spectral data [9,10]. However, CNNs inherently struggle to model long-range spatial dependencies due to their localized receptive fields [11,12]. This limitation can hinder performance on extended or irregular landforms that span multiple spatial scales or exhibit gradual morphological transitions, such as valley fill margins or sinuous mine benches.

Transformer-based models have recently gained traction in remote sensing by enabling global attention across spatial dimensions [13,14]. Vision Transformers (ViTs) and hybrid CNN–Transformer designs have shown success in capturing context-rich representations that improve segmentation accuracy in comparison to traditional machine learning-based methods, particularly in complex or heterogeneous landscapes [15,16]. Nevertheless, their quadratic computational complexity with respect to input size or number of tokens poses challenges for large-scale or high spatial resolution imagery, where inputs often exceed 512 $\times$ 512 pixels [11]. These models demand significant computational resources and large annotated datasets, limiting their applicability in many remote sensing workflows with constrained hardware or labeled data. Consistent with this, in a companion study, we evaluated a Transformer-based segmentation model (SegFormer) against UNet on the same three datasets used here and did not observe a clear advantage for the Transformer-based approach for this task, which we attribute in part to the relatively limited size of the available training data (1000 chips per dataset) and to the strong inductive biases of convolutional models under limited-data conditions [17]. Accordingly, the present study focuses on benchmarking Mamba-based and CNN-based architectures.

Emerging state-space models (SSMs) offer a promising alternative by modeling global spatial context with linear complexity relative to the number of tokens processed [18,19]. Among these, the Mamba architecture has recently drawn attention for its ability to efficiently encode long-range dependencies while maintaining high segmentation performance and scalability [20,21]. By replacing self-attention mechanisms with structured state-space

transitions, Mamba-based models reduce inference cost while preserving global contextual awareness [20], making them well-suited for terrain segmentation tasks that require both boundary precision and contextual understanding. However, their applications in geomorphic mapping contexts remain largely unexplored, and their effectiveness on LiDAR-derived LSP feature spaces, as opposed to spectral data, is still unknown. In this study, we address this gap by systematically evaluating both Mamba-based and hybrid CNN–Mamba architectures for the segmentation of complex geomorphic features and benchmarking their performance against conventional UNet architectures.

1.2. Semantic Segmentation with UNet

CNN-based architectures, particularly UNet [8], have become foundational tools for semantic segmentation, or pixel-level labeling. UNet is conceptualized in Figure 1. In a typical UNet architecture, the encoder applies successive 3×3 convolutions followed by batch normalization and nonlinearities (often the rectified linear unit (ReLU)), interleaved with 2×2 max-pooling to downsample spatial dimensions and enlarge the receptive field. The decoder reverses this process with transposed convolutions to upsample features, while skip connections concatenate encoder features at each scale to the corresponding decoder layers. This strategy bridges the semantic gap between shallow and deep layers, preserving fine spatial detail that would otherwise be lost during pooling. The final classification head projects the decoder output to per-pixel logits using a 1×1 convolution, supporting both binary and multiclass segmentation. As further discussed below in the context of Mamba-based architectures, the UNet architecture can be conceptualized as a general framework that can be modified to incorporate other modules or non-CNN-based components.

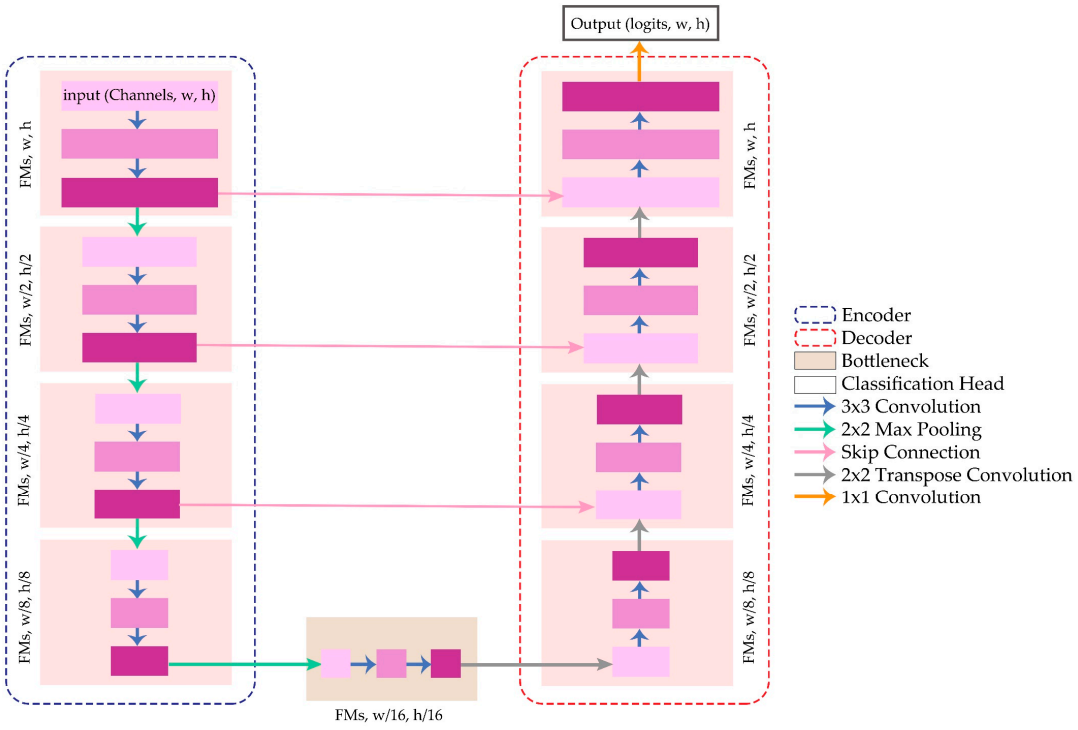

Figure 1. Conceptual illustration of the UNet architecture [8].

In geomorphic mapping and feature extraction specifically, researchers have documented that UNet trained using LSP-based feature spaces performs well in semi-automated mapping, though manual refinement is sometimes needed [22,23]. Our previous work [9] demonstrated that even standard UNet architectures can be competitive with more advanced alternatives when properly tuned and trained on sufficient data.

1.3. Mamba Architecture and Mamba–UNet

Recent advancements in semantic segmentation have sought to address the limitations of CNNs (i.e., limited receptive fields) and Transformer-based models (i.e., computational cost). A promising direction in this regard is the emergence of SSMs, particularly the Mamba architecture, which introduces a paradigm shift in sequential data modeling. Unlike self-attention mechanisms in Transformers, which scale quadratically with sequence length, SSMs process inputs using linear transformations of hidden states, resulting in linear computational complexity relative to the number of tokens processed. These properties enable Mamba-based models to capture both fine-scale boundaries and global spatial dependencies across large-scale spatial data more efficiently, making them particularly attractive for remote sensing applications involving high spatial resolution representations [24–27].

The core design of Mamba involves transforming inputs into hidden state representations and updating these states through structured linear transitions that selectively propagate information along specific sequence dimensions. By focusing on efficient context modeling across long distances, Mamba avoids the high computational overhead typical of Transformer-based approaches while also addressing the locality constraints of standard CNNs. Such efficiency and flexibility make Mamba especially compelling for dense prediction tasks, such as semantic segmentation, which demand both detailed boundary accuracy and broad contextual awareness [27–29].

Building on these principles, recent studies have adapted Mamba to vision tasks with encouraging results. We specifically make use of Mamba–UNet in this study, which was introduced by Wang et al. [30]. It is a fully visual Mamba-based symmetric encoder–decoder network originally designed for medical image segmentation. As shown in Figure 2, it begins by partitioning the input image into patches and embedding them into a higher-dimensional space using a linear embedding layer. The encoder consists of hierarchical stages where the spatial dimensions are progressively reduced through patch merging, and each stage incorporates stacked Visual State-Space (VSS) blocks to model long-range dependencies efficiently. At the network's core, a bottleneck module with additional VSS blocks extracts deep semantic features. The decoder mirrors the encoder structure, employing patch-expanding operations and VSS blocks to gradually restore the spatial resolution. Skip connections link corresponding encoder and decoder stages to preserve spatial detail and enable effective feature fusion. The architecture concludes with a linear projection layer that maps the final feature representation to the desired number of output classes. The authors reported superior results relative to traditional CNN- and Transformer-based UNet models in their case study.

1.4. Objectives and Contributions

Despite their promising performance in other domains, the application of Mamba-based architectures for geomorphic feature extraction remains largely unexplored. Our recent work has demonstrated the potential of deep learning-based semantic segmentation workflows for extracting geomorphic features from digital terrain model (DTM)-derived LSPs and has shown that Mamba-based encoders can be integrated into LSP-based terrain segmentation architectures [31]. However, systematic evaluation of pure Mamba-based and hybrid Mamba–CNN architectures across multiple anthropogenic geomorphic feature types remains limited. Given the distinct strengths of CNNs in capturing fine-scale, localized spatial context and the efficient global-context modeling capabilities of Mamba-based architectures, combining these approaches may yield improvements in segmentation performance. Importantly, our goal is not to propose a novel algorithmic architecture but to benchmark existing CNN-based, Mamba-based, and hybrid Mamba–CNN architectures for this specific and previously underexplored use case: the segmentation of anthropogenic

geomorphic features from LiDAR-derived LSPs. Prior studies have demonstrated the value of hybrid CNN–Mamba designs in other remote sensing domains [32,33], yet these hybrid strategies have not been systematically assessed for terrain-based geomorphic mapping using land surface parameters. This combination is motivated by the complementary strengths of the two paradigms: CNNs are well-suited to modeling local texture and fine-scale boundary detail, whereas Mamba-based encoders efficiently capture broader spatial context, and the relative benefit of pairing them has not been established for this task. Thus, we systematically evaluate and compare a baseline CNN-based UNet, Mamba–UNet, and a hybrid architecture with a Mamba-based encoder and a CNN-based decoder for the specific task of geomorphic feature extraction using LSP predictor variables as opposed to spectral image bands.

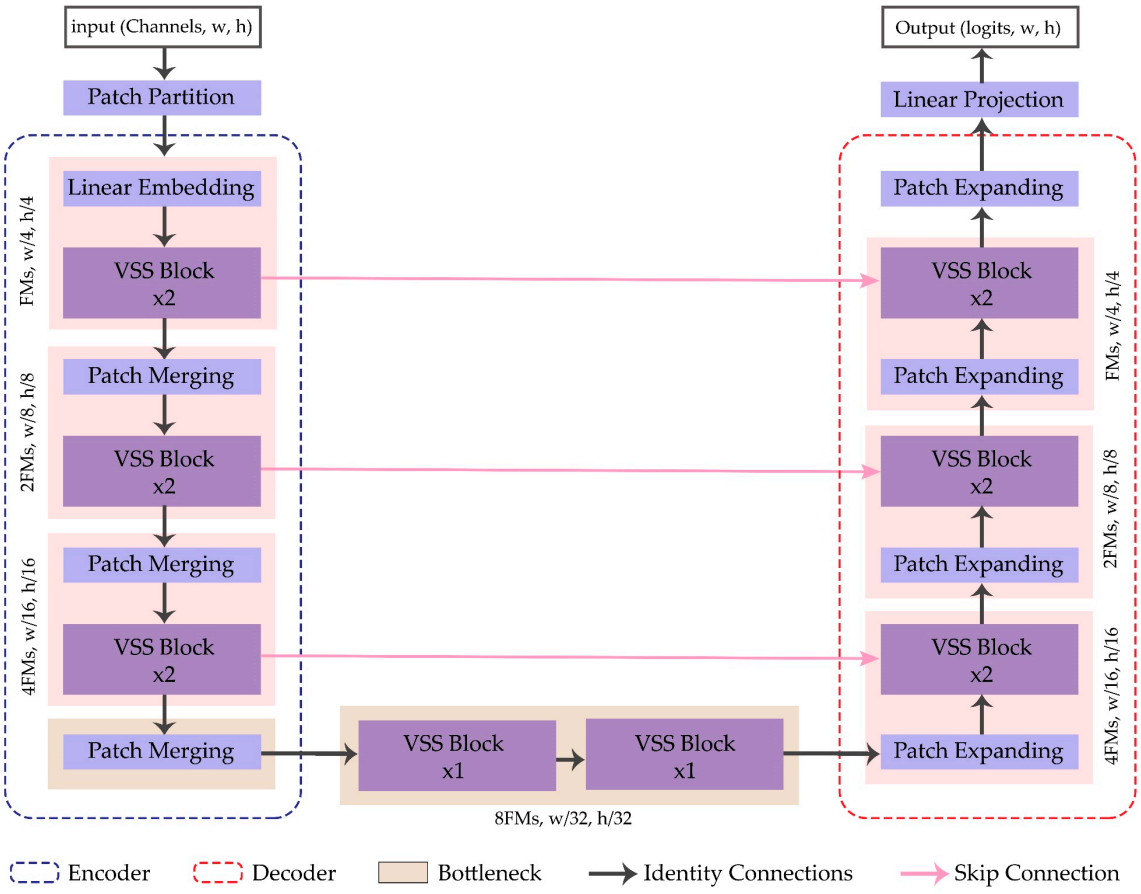

Figure 2. Conceptualization of the Mamba–UNet architecture [30], which consists of an encoder, a bottleneck, a decoder, and skip connections, with all core components built entirely using Visual Mamba blocks.

Transfer learning has proven effective in various remote sensing applications [34–37] and, in our prior work, ImageNet initialization consistently outperformed random and terrain-specific pre-training [38]. However, performance remains task-dependent; a mismatch between the prior and current task may limit the utility of pre-training [39,40], making systematic comparisons essential for robust evaluation. As a result, we also compare models with randomly initialized and ImageNet-based encoder parameter values. For those initialized with ImageNet-based encoder parameters, we also compare results with the encoder frozen and unfrozen during training. For all architectures, the pretrained encoder weights were obtained from ImageNet so that the CNN-based (ResNet-34) and Mamba-based encoders are initialized from a common pretraining source and remain comparable. We acknowledge that ImageNet pretraining on natural RGB imagery differs substantially from the three-

band LiDAR-derived LSP composites used here; however, our prior work [38] found that ImageNet initialization consistently outperformed both random initialization and terrain-specific pretraining for this class of task, which motivates its use. Because the pretraining source is held constant across architectures in this study, we do not attempt to disentangle the separate contributions of architecture and pretraining data, and we interpret claims regarding the transferability of pretrained encoders accordingly.

2. Data and Study Area

This study leverages a collection of high spatial resolution terrain datasets representing diverse geomorphic features across the Appalachian and Midwestern United States to evaluate the effectiveness of CNN-based, Mamba-based and hybrid deep learning models for landform segmentation. Each dataset corresponds to a distinct type of anthropogenic geomorphic landform: terraces, mine benches, and valley fill faces, distributed across varying topographic and environmental settings. These datasets were selected to represent diverse terrain conditions, spatial configurations, and landscape modification histories, allowing for a rigorous evaluation of model robustness and generalization. All datasets are constructed from LiDAR-derived DTMs and include pixel-level annotations of the target geomorphic features. Each annotated feature is represented as a binary mask and paired with a multiband raster stack of terrain derivatives, or LSPs, to form the input–label image pairs used in training and evaluation.

2.1. Study Areas and Geomorphic Context

The datasets used in this study originate from multiple physiographically distinct regions that pose unique challenges for semantic segmentation due to their complex topography, anthropogenic disturbance, and heterogeneous feature morphology. TerraceDL [41] (Figure 3A,D,G) represents agricultural terraces designed to control soil erosion in rolling farmland landscapes. This dataset covers portions of Iowa, USA, where extensive terracing systems have been constructed to stabilize hillslopes. These terraces are typically narrow, linear, and aligned with contour lines, producing sharp topographic transitions. Their edge-dominated geometries and strong directional patterns create class imbalance and introduce boundary delineation challenges for pixel-based classifiers.

MineBenchDL [42] (Figure 3B,E,H) captures remnant mining benches from legacy contour surface mining operations throughout north–central West Virginia, USA. These landforms were created by extracting coal seams along the slope contour, leaving behind narrow, linear ledges that are often obscured by vegetation and land cover change but remain detectable in terrain data. The benches occur on steep terrain and vary widely in length and curvature. Their subtle appearance in orthophotos, combined with high variability in topographic setting, makes them an ideal test case for evaluating model sensitivity to fine-scale topographic variation.

VfillDL [43] (Figure 3C,F,I) focuses on valley fill faces, man-made landforms created during mountaintop removal surface coal mining, where excess overburden material is placed in adjacent hollows. The dataset includes sites from West Virginia, Kentucky, and Virginia, USA. Valley fill faces tend to have smoother elevation profiles than benches or terraces but show significant variability in shape and roughness based on fill age, compaction, and vegetative cover. Their similarity to natural landforms further complicates the task of precise boundary segmentation. All three datasets were manually digitized, and the study areas were then partitioned geographically into disjoint regions for training, validation, and testing to ensure spatial independence and evaluate model generalization across unseen terrain contexts.

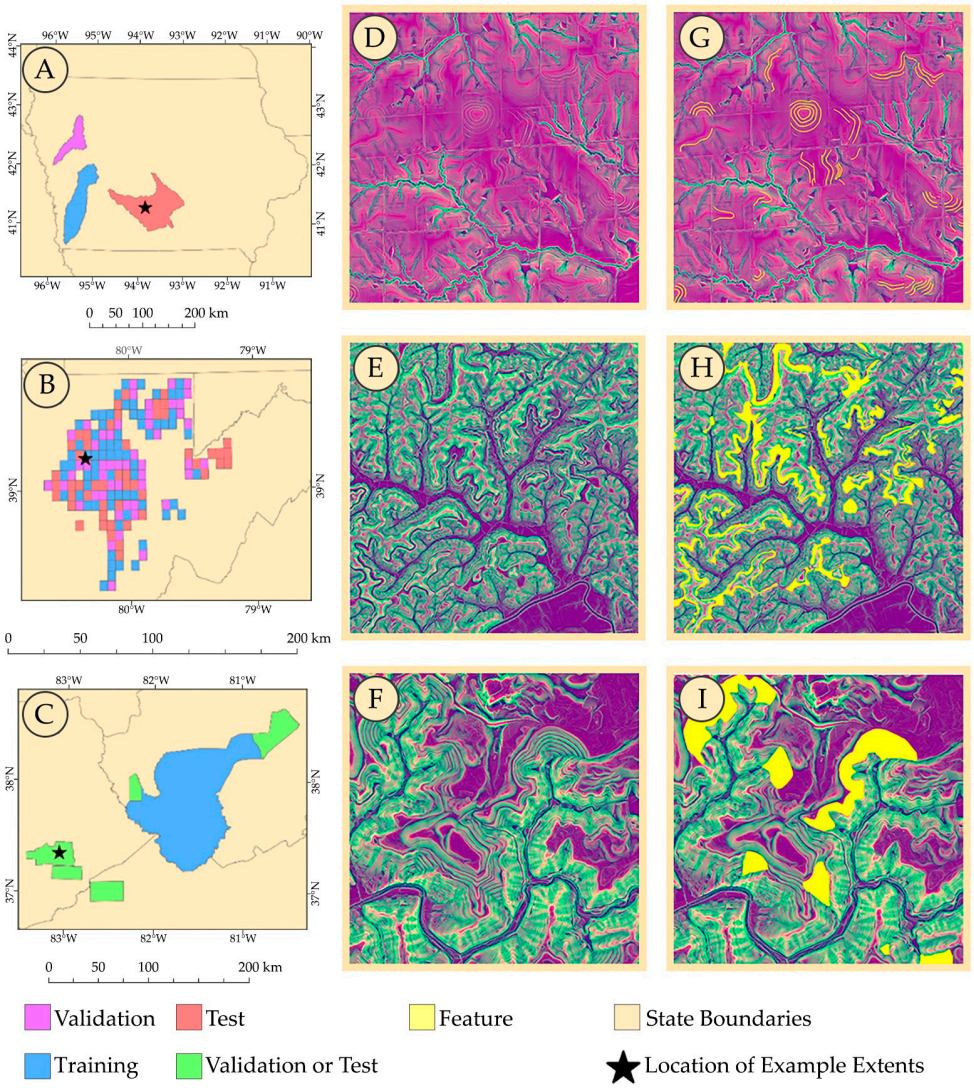

Figure 3. Training, validation, and test regions for (**A**) the agricultural terraces dataset (terraceDL) [41] located in Iowa, USA; (**B**) the historic mine benches dataset (minebenchDL) [42] from north–central West Virginia, USA; and (**C**) the valley fills dataset (vfillDL) [43] derived from surface coal mining areas in southern West Virginia, eastern Kentucky, and southwestern Virginia, USA; (**D**–**I**) illustrate example terrain surfaces and their corresponding geomorphic features.

2.2. Terrain Features and Preprocessing

To train deep learning–based semantic segmentation models, larger spatial extents must be divided into smaller, standardized input sections known as "chips". In this study, multi-band raster stacks of LSPs, derived from 1 m spatial resolution LiDAR-based DTMs, were overlaid with manually labeled geomorphic features to extract non-overlapping chips of 512 × 512 pixels. This chip size balances computational efficiency and local spatial context while aligning with the input requirements of the architectures evaluated in this study. For the Mamba-based architectures, this chip size allows global context to be modeled within 512 × 512 m (~0.25 km^2) areas.

Each chip contains a three-band composite of terrain predictors carefully selected to capture both fine-scale surface morphology and broader hillslope position. We adopted a three-layer configuration developed by researchers at the U.S. Geological Survey (USGS), which has been shown to outperform conventional terrain visualizations in machine learning applications [7]. The three layers include: (1) a topographic position index (TPI) computed using a 50 m circular moving window to capture mid-scale hillslope context;

(2) the square root of slope, a nonlinear transformation that enhances gradients in steep areas while smoothing overly sharp transitions; and (3) an annular-window TPI, calculated using an inner radius of 2 m and an outer radius of 5 m, to capture localized relief and roughness variations. We note that the selection of this input feature space was itself the subject of a dedicated prior study [7], in which alternative terrain visualizations and predictor combinations were systematically compared for CNN-based geomorphic feature extraction; we therefore adopt this composite rather than re-evaluating feature-space combinations here and treat the input representation as fixed in order to isolate the effect of model architecture.

All terrain derivatives were generated in R [44] using standardized preprocessing workflows, including the terra [45] and MultiscaleDTM packages [46]. The resulting multiband raster grids form a compact but information-rich representation of terrain structure suitable for deep learning–based segmentation. To ensure consistency across datasets, all input layers were normalized to a [0, 1] range.

To address class imbalance while preserving representative background context, each training set was constructed to include 1000 image chips per dataset, comprising 800 chips that contained at least one pixel labeled as the positive class and 200 chips that consisted entirely of background (negative class). This strategy balanced foreground representation while preserving background context for effective learning.

Even with this sampling strategy, the target classes remain highly underrepresented at the pixel level. Within the 1000 training chips, positive (feature) pixels accounted for 3.46% of pixels for agricultural terraces, 5.77% for mine benches, and 4.27% for valley fill faces, corresponding to background-to-foreground ratios of approximately 28:1, 16:1, and 22:1, respectively. These values quantify the severe class imbalance inherent to these tasks and motivate both the sampling strategy and the use of a focal Tversky loss. To further characterize the scale of the target features, we measured the area of connected positive-pixel components within the 1000 training chips of each dataset (1 m pixel resolution). Because features that cross chip boundaries are divided during tiling, these values represent within-chip feature fragments and provide a lower-bound characterization of landform size rather than complete per-feature areas. The resulting feature fragments were smallest for agricultural terraces (median 0.04 ha, mean 0.05 ha, range 0.0001–0.63 ha; n = 17,832), consistent with their narrow, contour-parallel geometry, and larger for mine benches (median 0.36 ha, mean 0.73 ha, range 0.0001–19.77 ha; n = 2059) and valley fill faces (median 0.52 ha, mean 0.92 ha, range 0.0001–13.11 ha; n = 1223), which form broader, more compact features. The right-skewed distributions, in which mean areas exceed medians, reflect the presence of a small number of comparatively large features alongside many small ones.

In contrast, for validation and testing, all chips were included regardless of class composition to evaluate model performance under realistic and spatially diverse conditions. The chip extraction process was implemented using a custom R function tailored to efficiently partition the terrain predictor grids and corresponding pixel-wise labels, preserving spatial alignment and data integrity. A summary of the number of training, validation, and test chips used in each dataset is presented in Table 1. All datasets were partitioned to maintain spatial independence between training and testing areas, and the same sampling logic was applied across agricultural terraces, mine benches, and valley fill faces. This preprocessing pipeline provides a consistent and replicable foundation for evaluating model performance across distinct landform types, spatial scales, and topographic complexity.

Table 1. Overview of the number of image chips allocated to training, validation, and testing for each dataset.

Dataset	Training	Validation	Testing
Agricultural Terraces	1000	2268	2770
Mine Benches	1000	2306	3078
Valley Fill Faces	1000	1081	1644

2.3. Landscape Context

The study areas span regions with highly contrasting geomorphic and land-use characteristics. The Appalachian sites feature steep, dissected terrain with a legacy of surface mining and vegetation regrowth, whereas the Iowa region exhibits engineered agricultural landscapes characterized by mechanically modified slopes and uniform drainage patterns. These contrasting environments were deliberately selected to test the ability of models trained in one geomorphic context to generalize across others. Additionally, differences in feature morphology, such as the long and narrow geometries of terraces and mine benches versus the larger, more compact shapes of valley fills, introduce variation in edge density, object compactness, and class imbalance. These properties make the datasets especially well-suited for benchmarking the relative performance of CNNs, Mamba-based models, and hybrid architectures in modeling structured terrain features under real-world variability.

3. Methods

3.1. Baseline and Proposed Models

To evaluate the effectiveness of Mamba-based architectures for geomorphic feature segmentation, we implemented and compared three categories of models: (1) a standard CNN-based UNet with a ResNet-34 encoder, (2) a pure Mamba–UNet architecture [30], and (3) a hybrid model (Mamba–UNetC) that combines a Mamba–UNet encoder with a CNN-based decoder.

The baseline UNet follows a standard encoder–decoder structure. The encoder consists of a ResNet-34 backbone that extracts hierarchical features through stacked residual blocks, while the decoder progressively upsamples the feature maps using transposed convolutions. Lateral skip connections link encoder and decoder layers of matching resolution to preserve spatial detail and support accurate boundary reconstruction. This architecture is widely used in geospatial segmentation and provides a strong reference point for performance comparison. Residual connections, either identity or projection shortcuts, help maintain gradient flow and improve optimization in deeper networks [9,47], leading to consistent performance gains across remote sensing tasks including land use classification [48], urban segmentation [49], fault detection [50], and tree species mapping [51].

The Mamba–UNet represents a fully state-space model adapted for semantic segmentation. The encoder replaces traditional convolutional or self-attention mechanisms with stacked Visual Mamba blocks, which model long-range spatial dependencies through a linear state-space formulation. Input images are partitioned into non-overlapping patches, linearly embedded, and passed through a multi-stage encoder where patch merging is applied to reduce spatial resolution. The decoder mirrors this structure using patch-expanding operations to reconstruct the original spatial resolution.

The Mamba–UNetC architecture is a hybrid design that retains the Mamba-based encoder used in Mamba–UNet but replaces the decoder with a conventional CNN-based UNet decoder. This model aims to leverage the global context modeling and efficiency of the Mamba–UNet encoder while utilizing the proven localization and boundary refinement capabilities of convolution-based decoding and skip connection structures. Encoder–

decoder skip connections are retained to bridge semantic and spatial representations. Figure 4 presents the conceptual design of the Mamba–UNetC architecture.

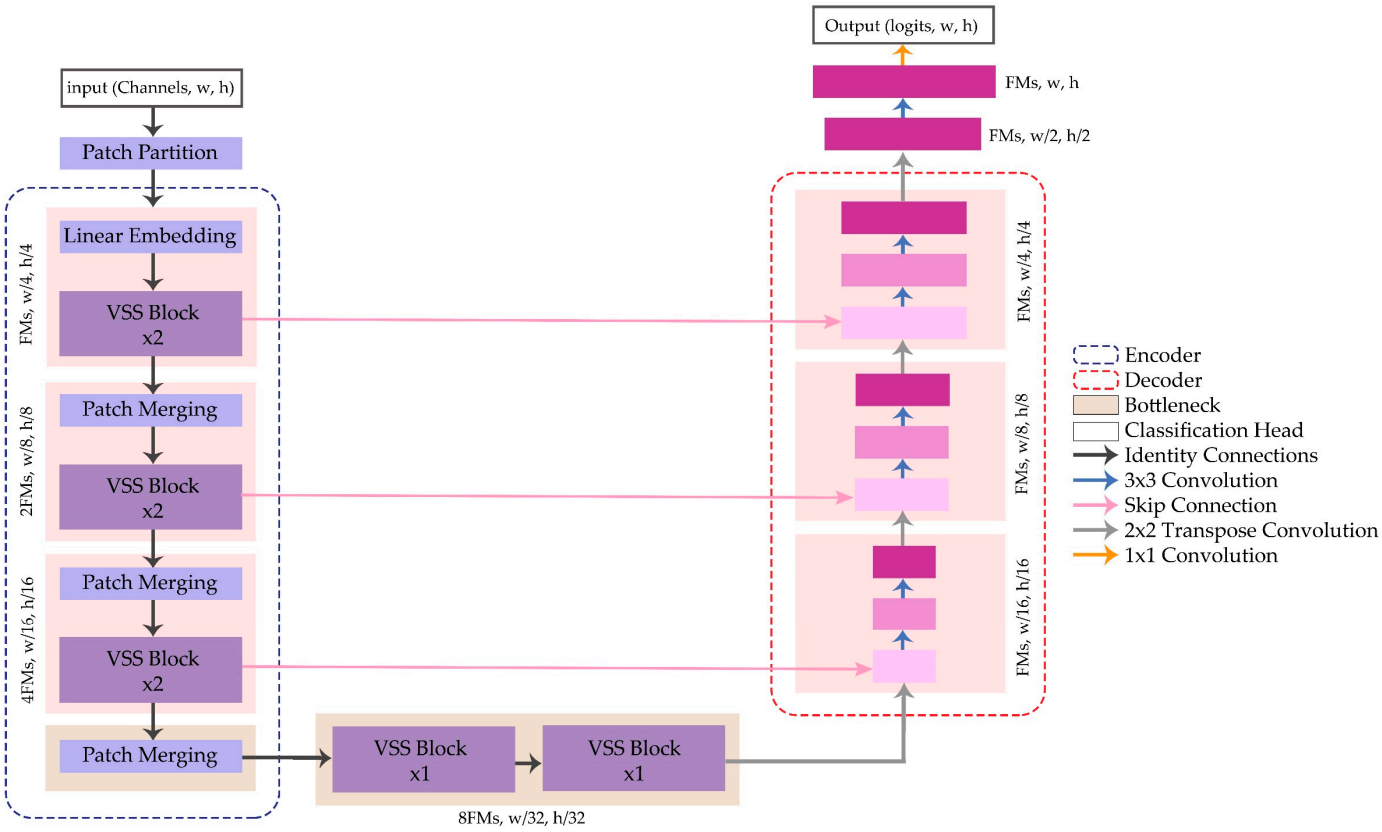

Figure 4. Conceptual design of a UNet model with a Mamba-based encoder, illustrating the hierarchical Mamba encoder connected via skip connections to a UNet-style, CNN-based decoder.

Each model was evaluated under multiple initialization and training strategies. Specifically, we considered both ImageNet-pretrained and randomly initialized encoders, as well as frozen and unfrozen training schemes. In the frozen setting, encoder weights were kept fixed during training, allowing only the decoder to be updated. In the unfrozen setting, the entire model was fine-tuned. The frozen random-initialization configuration was excluded because freezing randomly initialized encoder weights offers no practical learning benefit.

3.2. Training Strategy

All model training and evaluation were conducted using the PyTorch (2.7.1 + CUDA 12.8) deep learning [52] framework within a Python (3.11.9) environment [53]. Model development was performed on a Linux-based workstation equipped with an AMD Ryzen Threadripper Pro 3955WX 16-core processor, 128 GB of RAM, and three NVIDIA RTX A5000 GPUs, providing a combined 72 GB of GPU memory. CUDA-enabled GPU acceleration was used to facilitate efficient model training. The UNet architecture was implemented using the Segmentation Models library [54], while the Mamba–UNet model was implemented using the original codebase made publicly available by the developers [30]. The Mamba–UNetC model was developed by the authors based on the Mamba–UNet architecture. Input data consisted of 512 × 512-pixel image chips with three-band LSP composites as input and binary masks as ground reference labels. All models were trained for 50 epochs. Data augmentation was applied probabilistically, with horizontal and vertical flipping each having a 30% chance of occurring per image. Augmentation was restricted to geometric transformations (flipping), consistent with prior terrain-mapping studies

using LSPs. Radiometric or image-enhancement augmentations commonly used for optical imagery, such as contrast stretching, brightness or color jittering, and blurring, are not physically meaningful for LSP predictor variables and were therefore not applied, as they would distort the terrain-derived values rather than represent plausible variation.

Mini-batch sizes were adjusted according to GPU memory requirements. Because the Mamba-based models required more memory, all models were trained with a mini-batch size of 8 to ensure stable training and consistent throughput across architectures. For models initialized with pretrained ImageNet weights [55], input data were normalized using standard ImageNet statistics (mean: [0.485, 0.456, 0.406]; standard deviation: [0.229, 0.224, 0.225]) in the UNet-based experiments. In contrast, the Mamba-based experiments used the preprocessing configuration employed by the corresponding architecture and therefore did not include ImageNet-based normalization. The AdamW optimizer was used for all models with different learning rates based on the models and encoder initialization strategy. For all models trained from random initialization, a fixed learning rate of 1×10^{-4} was applied uniformly to all parameters. To stabilize optimization in pretrained models and mitigate erratic learning dynamics commonly observed when using a uniform learning rate across heterogeneous components, a tiered learning rate strategy was adopted for UNet variants. The early encoder blocks were trained with a lower learning rate (1×10^{-6}), middle layers with a moderate rate (1×10^{-5}), and the decoder and classification head with a higher rate (1×10^{-4}). This differential scheme was selected based on empirical results and past studies noting instability when pretrained backbones are updated too aggressively [56]. For the Mamba–UNet and Mamba–UNetC models, which were initialized using pretrained weights, we used a consistent learning rate of 1×10^{-4} across all trainable layers.

To handle severe class imbalance, we employed a focal Tversky loss tailored to each model family. We note that this was not simply a tuning preference: in initial experiments, we attempted to train all architectures using a single, shared loss parameterization, but the Mamba-based models failed to converge under the loss configuration that was effective for the UNet-based models. We therefore found it necessary to parameterize the loss function differently when including Mamba-based components, and we treat this requirement itself as a finding regarding the differing optimization behavior of these architectures. As a consequence, the precision–recall balance is not directly comparable across model families, and we interpret the corresponding comparisons with this caveat in mind. For UNet-based models, we used $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.75$, balancing false positives and false negatives with a moderate emphasis on underrepresented classes. For the Mamba–UNet and Mamba–UNetC architectures, which exhibited sharper prediction behavior and benefited from greater recall emphasis in initial experiments, we used $\alpha = 0.95$, $\beta = 0.05$, and $\gamma = 1.5$. These settings more heavily penalized false negatives and down-weighted easier background predictions, thereby improving segmentation performance for sparse, edge-dominated geomorphic features. Finally, we retained the model checkpoint corresponding to the epoch that yielded the highest validation macro-averaged F1-score during training, ensuring that the model's performance on both the feature and background classes contributed equally [57]. Because the loss functions were parameterized differently across architectures, loss magnitudes are not directly comparable between model families, and we therefore assessed convergence and fit using the validation macro-averaged F1 trajectories rather than raw loss curves. These trajectories did not exhibit the sustained late-epoch decline in validation performance that would indicate overfitting.

3.3. Accuracy and Computational Assessment

Model performance was assessed using a set of standard pixel-level classification metrics derived from the binary confusion matrix, which is a common framework for evaluating semantic segmentation accuracy. In the context of binary geomorphic segmentation, each pixel prediction falls into one of four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These values form the basis for calculating the evaluation metrics used in this study that reflect not only overall classification success but also the model's sensitivity to underrepresented geomorphic features and its tendency toward false detection. Overall accuracy (OA) measures the proportion of correctly classified pixels (both feature and background) over the total number of pixels (Equation (1)). While widely reported, OA can be misleading in class-imbalanced settings, as it is heavily influenced by the dominant background class [57,58].

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{1}$$

Precision, or 1—commission error (Equation (2)), quantifies the proportion of positive predictions that are actually correct. It reflects the model's ability to avoid false positives. High precision indicates that most predicted feature pixels correspond to true features, which is especially important for reducing over-segmentation in sparse classes. Recall, or 1—omission error (Equation (3)), measures the fraction of actual feature pixels that are successfully identified. It reflects the model's sensitivity to true positive detection and is critical in applications where missing small or subtle features is undesirable.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

F1-score (Equation (4)) is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. It provides a more informative summary metric when class imbalance is present and is commonly used in segmentation tasks with sparse positive classes.

$$\mathrm{F1} - \mathrm{score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{4}$$

All metrics were calculated using the TorchMetrics library and accumulated on a per-image basis. Precision, recall, and F1-score were computed and reported for the positive (foreground) class only. Each model was evaluated on an independently held-out test set corresponding to its respective dataset. No test samples were used during model training or validation, ensuring unbiased performance estimates. Together, these metrics offer a comprehensive assessment of both segmentation accuracy and error structure, allowing us to interpret how different architectural choices affect the ability to detect geomorphic features from LiDAR-derived terrain data.

In addition to segmentation accuracy, we also assessed each model's computational efficiency by recording the number of trainable parameters, the number of multiply accumulate operations (MACs), and the estimated floating-point operations per forward pass (FLOPs). These metrics serve as indicators of model size, memory demand, and runtime complexity, respectively. Parameter count reflects the architecture's capacity and optimization burden, MACs offer a proxy for inference efficiency, and FLOPs quantify the total computational workload. While these values are not direct indicators of segmentation

quality, they help contextualize the trade-offs between model complexity and performance, especially in remote sensing applications where computational resources may be limited.

4. Results

4.1. Training Dynamics and Validation Trends

Figure 5 presents the validation F1-score trajectories over 50 epochs for all model configurations across the three geomorphic segmentation tasks. For the agricultural terraces dataset, most models converged to relatively similar validation F1-scores, generally between 0.77 and 0.82. The randomly initialized UNet achieved the highest and most stable performance, while the ImageNet-initialized UNet and the pretrained Mamba-based models were slightly lower but still competitive. The randomly initialized Mamba–UNet and Mamba–UNetC models improved steadily but showed greater fluctuation, particularly during the early and middle training epochs. For the mine benches dataset, validation performance was more variable across epochs and model types. The Mamba–UNetC configuration initialized with pretrained weights generally produced the highest validation F1-scores, particularly in the unfrozen setting, while the pretrained Mamba–UNet and randomly initialized UNet also remained competitive. In contrast, the ImageNet-initialized UNet showed lower and flatter validation F1-score trajectories. The randomly initialized Mamba-based models exhibited the greatest instability on this dataset, suggesting that mine bench segmentation was more sensitive to initialization and training strategy. For the valley fill faces dataset, several models achieved high validation F1-scores, with the randomly initialized UNet and the pretrained Mamba-based configurations generally reaching the highest values.

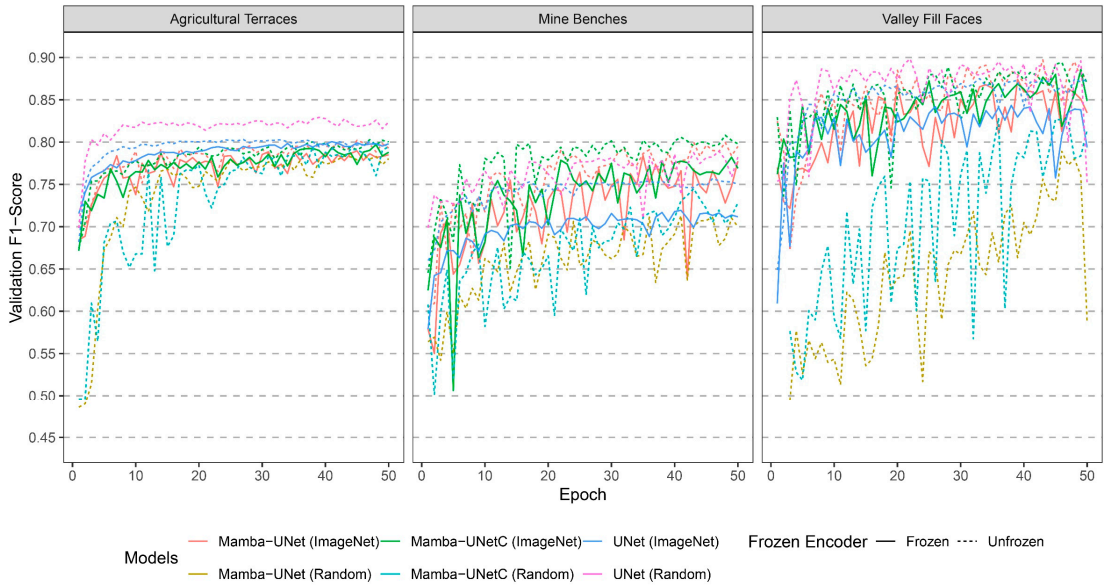

Figure 5. Validation F1-score curves over 50 epochs for all models on the terraceDL, mineBenchDL, and vfillDL datasets. Solid lines indicate frozen encoders; dashed lines indicate unfrozen encoders.

In general, validation F1-scores increased rapidly during the early epochs and then stabilized, although the degree of stability varied by dataset and model configuration. Across the three tasks, the results indicate that validation performance was more dataset-dependent than training loss alone suggested, with no single model dominating under all conditions. Nonetheless, the pretrained Mamba-based models, particularly Mamba–UNetC, showed strong and competitive generalization, while the randomly initialized UNet also performed remarkably well, especially for agricultural terraces.

4.2. Test Set Performance and Statistical Robustness

Figure 6 and Table 2 summarize test set performance across all model configurations for the agricultural terraces, mine benches, and valley fill faces segmentation tasks. Each configuration was evaluated on a spatially independent test split using the checkpoint that achieved the highest validation F1-score during training. Because these segmentation tasks were highly imbalanced, OA is reported for completeness, but F1-score, recall, and precision were calculated and reported for the positive class only, making them more informative for evaluating feature detection performance.

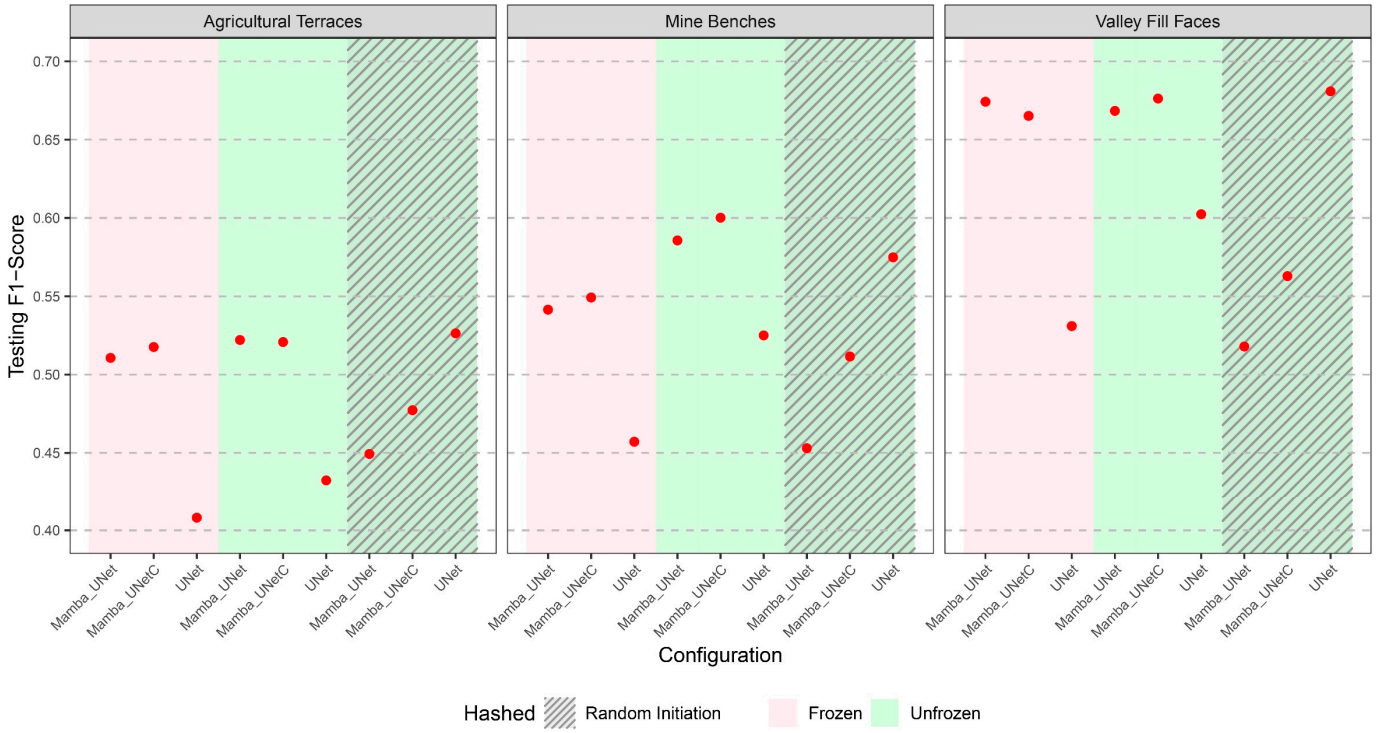

Figure 6. Testing F1-score for all model configurations across the three geomorphic feature segmentation tasks: agricultural terraces, mine benches, and valley fill faces. Each red dot represents the final model's F1-score after 50 training epochs. Configurations are grouped by encoder freezing strategy: frozen (pink) and unfrozen (green).

For the agricultural terraces task, the highest test F1-score was achieved by the randomly initialized unfrozen UNet (F1 = 0.5264), followed closely by the ImageNet-initialized unfrozen Mamba–UNet (F1 = 0.5221) and ImageNet-initialized unfrozen Mamba–UNetC (F1 = 0.5208). The frozen ImageNet-initialized Mamba-based models also remained competitive (F1 = 0.5107 for Mamba–UNet and 0.5176 for Mamba–UNetC), whereas the ImageNet-initialized UNet configurations performed notably worse, especially with a frozen encoder (F1 = 0.4080). The terrace results also revealed a precision–recall tradeoff across architectures: the Mamba-based models generally produced higher recall values, with the highest recall observed for the frozen ImageNet-initialized Mamba–UNet (Recall = 0.6586), whereas the randomly initialized UNet achieved the highest precision (Precision = 0.5567) and the highest overall F1-score.

For the mine benches task, the best performance was obtained by the ImageNet-initialized unfrozen Mamba–UNetC (F1 = 0.6002), which also achieved the highest OA (0.9842) and the highest precision (0.5797) for this dataset. The ImageNet-initialized unfrozen Mamba–UNet ranked second (F1 = 0.5857), followed by the randomly initialized unfrozen UNet (F1 = 0.5749). In contrast, the randomly initialized Mamba-based models performed less favorably, particularly the randomly initialized Mamba–UNet

(F1 = 0.4530). Among all mine bench configurations, the highest recall was obtained by the frozen ImageNet-initialized Mamba–UNetC (Recall = 0.7210), although its lower precision reduced its overall F1-score relative to the unfrozen counterpart. As in the terraces dataset, the frozen ImageNet-initialized UNet was among the weakest performers (F1 = 0.4571), while unfreezing improved its results (F1 = 0.5251).

Table 2. Test set performance metrics for all model configurations across the three geomorphic segmentation tasks. Each configuration is defined by model type, encoder initialization, and whether the encoder was frozen or unfrozen during training.

Problem	Model	Initiation	Frozen/ Unfrozen	OA	F1-Score	Recall	Precision
Agricultural Terraces	Mamba–UNet	ImageNet	Frozen	0.9914	0.5107	0.6586	0.4170
			Unfrozen	0.9924	0.5221	0.6071	0.4580
		Random	Unfrozen	0.9908	0.4494	0.5530	0.3785
	Mamba–UNetC	ImageNet	Frozen	0.9920	0.5176	0.6297	0.4394
			Unfrozen	0.9926	0.5208	0.5895	0.4664
		Random	Unfrozen	0.9915	0.4773	0.5725	0.4093
	UNet	ImageNet	Frozen	0.9926	0.4080	0.3750	0.4474
			Unfrozen	0.9935	0.4325	0.3664	0.5277
		Random	Unfrozen	0.9939	0.5264	0.4991	0.5567
Mine Benches	Mamba–UNet	ImageNet	Frozen	0.9781	0.5415	0.6787	0.4504
			Unfrozen	0.9816	0.5857	0.6798	0.5145
		Random	Unfrozen	0.9723	0.4530	0.6015	0.3634
	Mamba–UNetC	ImageNet	Frozen	0.9774	0.5492	0.7210	0.4436
			Unfrozen	0.9842	0.6002	0.6222	0.5797
		Random	Unfrozen	0.9804	0.5116	0.5388	0.4869
	UNet	ImageNet	Frozen	0.9793	0.4571	0.4560	0.4583
			Unfrozen	0.9825	0.5251	0.5069	0.5445
		Random	Unfrozen	0.9837	0.5749	0.5757	0.5741
Valley Fill Faces	Mamba–UNet	ImageNet	Frozen	0.9881	0.6742	0.7294	0.6267
			Unfrozen	0.9903	0.6684	0.5811	0.7865
		Random	Unfrozen	0.9795	0.5180	0.6520	0.4297
	Mamba–UNetC	ImageNet	Frozen	0.9895	0.6652	0.6196	0.7179
			Unfrozen	0.9900	0.6762	0.6208	0.7426
		Random	Unfrozen	0.9850	0.5629	0.5716	0.5544
	UNet	ImageNet	Frozen	0.9878	0.5310	0.4088	0.7572
			Unfrozen	0.9893	0.6024	0.4806	0.8069
		Random	Unfrozen	0.9906	0.6808	0.5931	0.7991

For the valley fill faces task, the strongest performance was achieved by the randomly initialized unfrozen UNet (F1 = 0.6808), which also yielded the highest OA (0.9906). However, several pretrained Mamba-based configurations performed very similarly, including the ImageNet-initialized unfrozen Mamba–UNetC (F1 = 0.6762), the frozen ImageNet-initialized Mamba–UNet (F1 = 0.6742), and the unfrozen ImageNet-initialized Mamba–UNet (F1 = 0.6684). The highest recall for this dataset was again produced by the frozen ImageNet-initialized Mamba–UNet (Recall = 0.7294), while the highest precision was obtained by the ImageNet-initialized unfrozen UNet (Precision = 0.8069), although its comparatively low recall limited its overall F1-score (0.6024). The randomly initialized Mamba-based models performed substantially worse than their pretrained counterparts on this dataset, with F1-scores of 0.5180 for Mamba–UNet and 0.5629 for Mamba–UNetC.

Overall, the results indicate that no single model family consistently dominated all three tasks. The ImageNet-initialized Mamba-based models, particularly Mamba–UNetC, showed the most consistent performance across datasets and generally remained competitive even when the encoder was frozen. In contrast, the randomly initialized UNet performed remarkably well, achieving the best F1-scores for both agricultural terraces and valley fill faces and remaining competitive for mine benches. Another important pattern is that random initialization affected the Mamba-based models more negatively than the UNet, whereas freezing the encoder was generally more detrimental for UNet,

especially for the ImageNet-initialized configurations. Although OA values were uniformly high across all models, reflecting the dominance of background pixels, the positive-class F1-score, recall, and precision revealed clearer differences in the ability of each architecture to detect geomorphic features. Table 2 provides the full set of test performance metrics for all configurations.

To further assess robustness to sampling variation in the held-out test data, we recalculated the performance metrics using 10 random subsamples of the test set for each task, with each subsample containing 70% of the available test chips. For every model configuration, we recomputed OA, F1-score, precision, and recall on each resample. OA is reported for all pixels, whereas F1-score, precision, and recall were calculated for the positive class only.

Figure 7 summarizes these results, where each red point represents the mean across the 10 resamples and each error bar shows the full min-max range. Taken together, the resampling results strengthen the conclusions drawn from Table 2 and Figure 6. Although OA was consistently high across all configurations, the positive-class F1-score, precision, and recall provided a more meaningful assessment of model behavior. The ImageNet-initialized Mamba-based models, particularly Mamba–UNetC, showed the most consistent and robust performance across tasks, while the randomly initialized UNet remained highly competitive. The generally narrow to moderate min-max ranges indicate that these performance patterns were stable across repeated test-set subsamples, supporting the statistical robustness of the main findings.

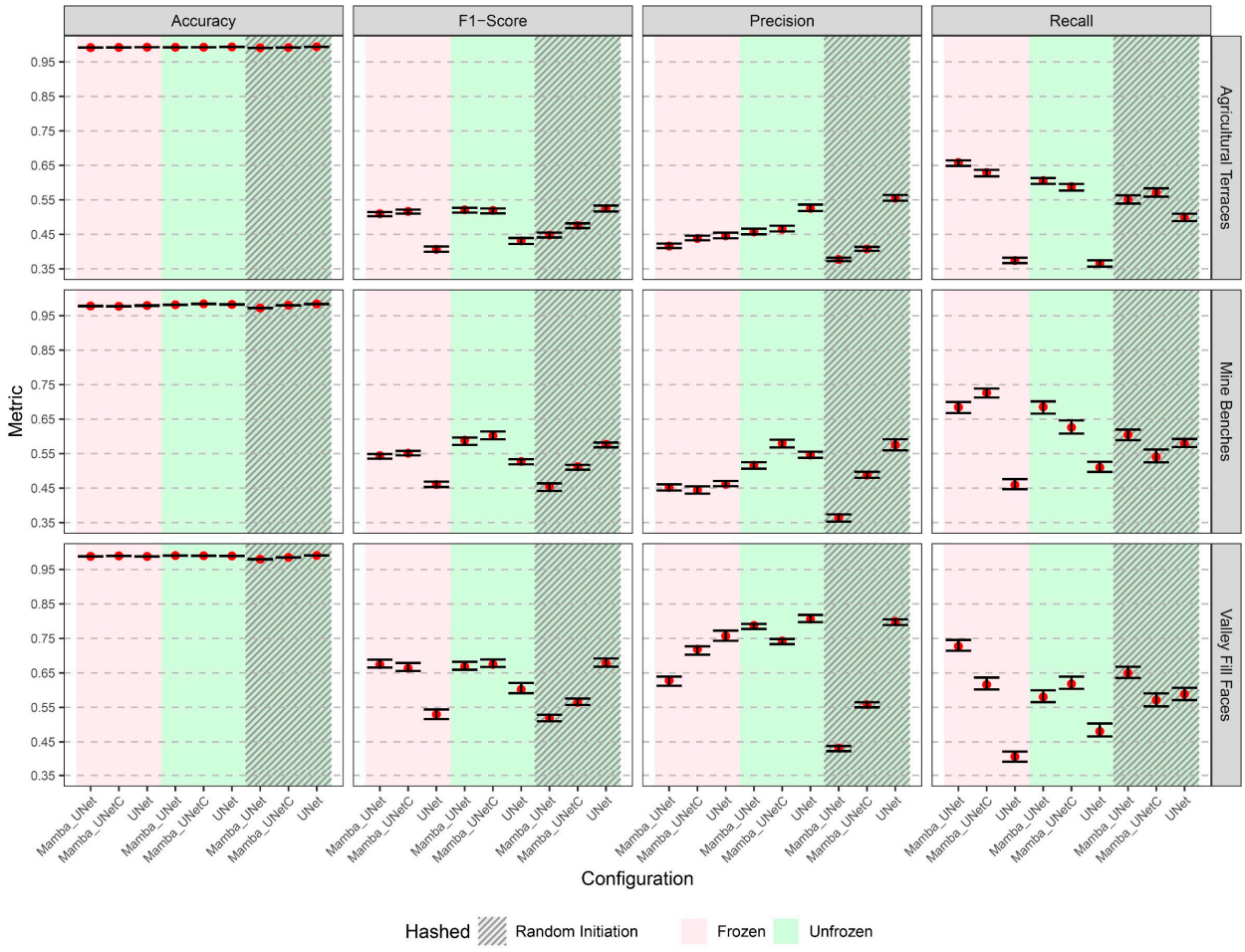

Figure 7. Analysis of test metrics using subsamples of the test set. Red points denote the mean across 10 resamples; error bars show the full min–max range. Panels display OA, F1-score, precision, and recall for each configuration and grouped by encoder freezing.

4.3. Visual Evaluation of Spatial Prediction Quality

To complement the quantitative results, we qualitatively assessed spatial predictions on representative subsets from each task. These comparisons, shown in Figure 8 (agricultural terraces), Figure 9 (mine benches), and Figure 10 (valley fill faces), provide visual insights into the boundary quality, spatial coverage, and noise artifacts across configurations.

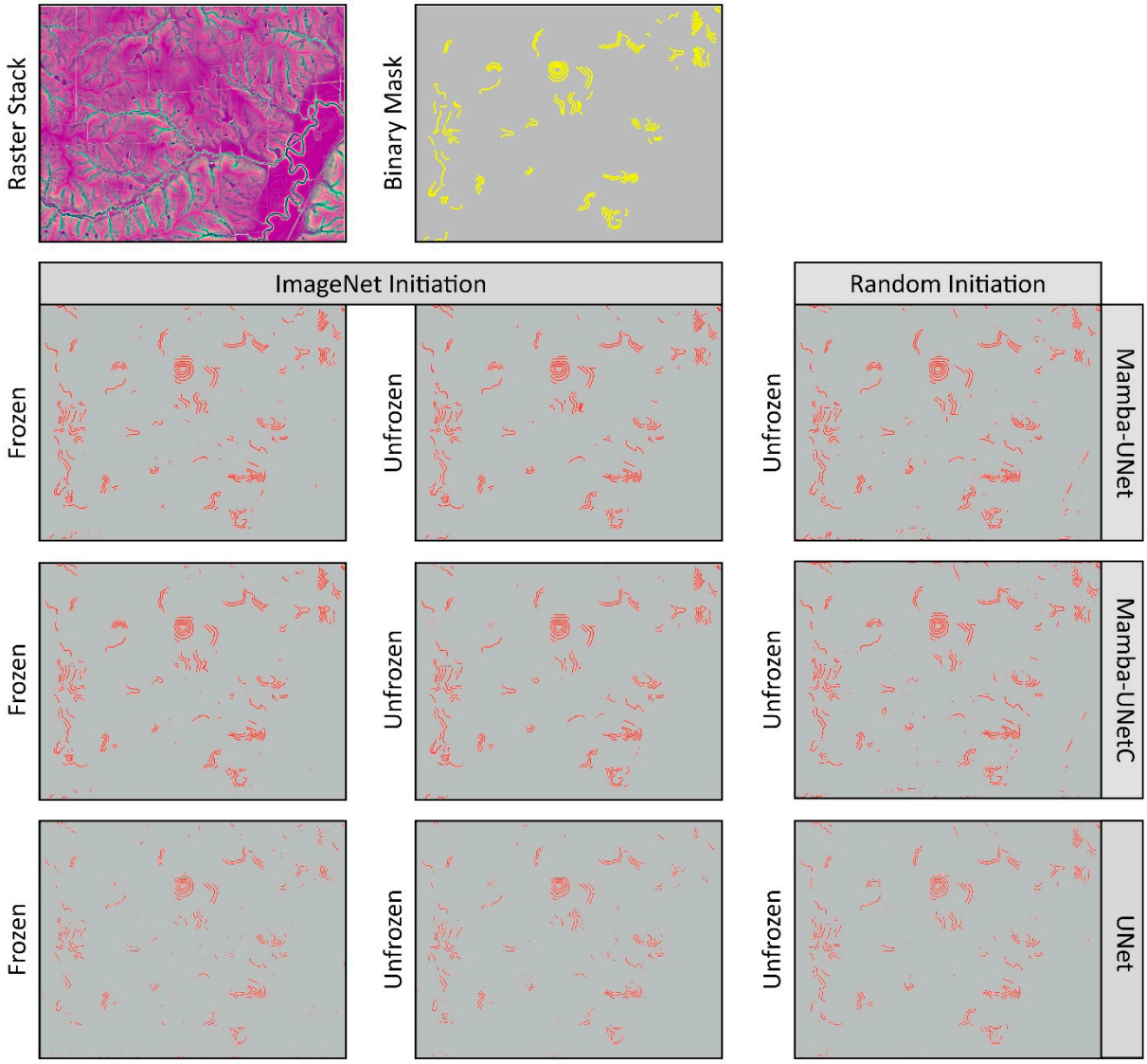

Figure 8. Spatial predictions for the agricultural terrace (terraceDL) test area. Panels compare predicted terrace segments (red overlays) from each configuration against the reference mask (bottom-right) and the three-band LSP stack (bottom-center).

For the agricultural terraces example, the target features appear as thin, contour-parallel lineaments, making continuity and boundary sharpness especially important for visual assessment. Across the predictions, the Mamba–UNet and Mamba–UNetC configurations generally recovered the terrace patterns more completely than UNet, with relatively continuous and spatially coherent lineaments that showed closer agreement with the reference mask. The ImageNet-initialized Mamba–UNet produced strong delineation in both frozen and unfrozen settings, with the unfrozen version appearing slightly more complete in weaker or partially fragmented terrace segments. Similarly, Mamba–UNetC generated clean and well-defined terrace traces in both settings, with the unfrozen model again showing somewhat improved continuity and fewer missed segments. In contrast,

the ImageNet-initialized frozen UNet was clearly more conservative, omitting many subtle terrace features and representing others as shorter, discontinuous fragments. Unfreezing improved the UNet prediction, but it still remained less complete than the Mamba-based models. The randomly initialized UNet was visually competitive and produced relatively crisp terrace traces with limited background noise, but it remained somewhat sparser overall, particularly in low-contrast areas, suggesting a more precision-oriented prediction pattern. Overall, the visual results for agricultural terraces indicate that the Mamba-based models, especially Mamba–UNetC, provided one of the closest matches to the reference patterns of the narrow and elongated terrace geometry.

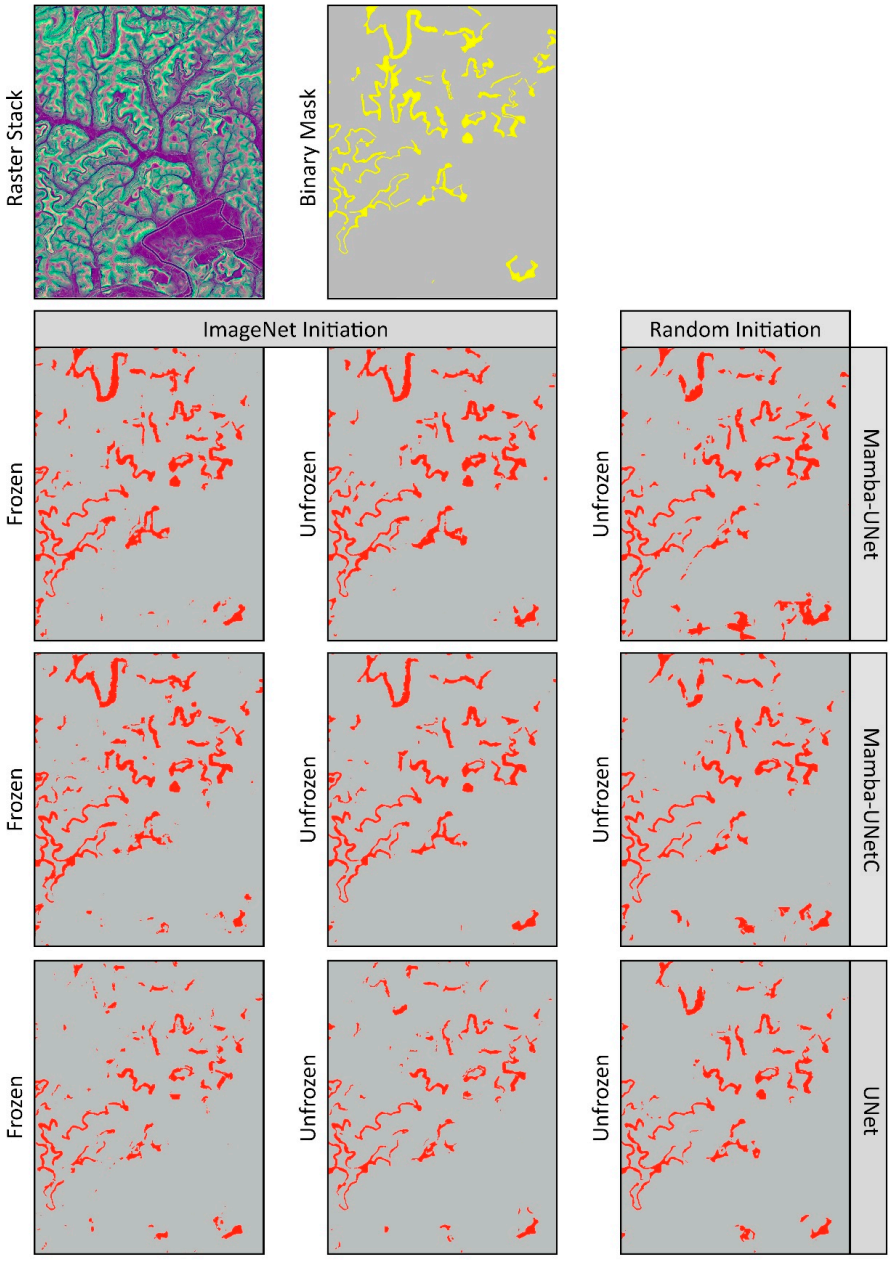

Figure 9. Spatial predictions for the mine bench (mineBenchDL) test area. Panels compare predicted mine benches (red overlays) from each configuration against the reference mask (bottom-right) and the three-band LSP stack (bottom-center).

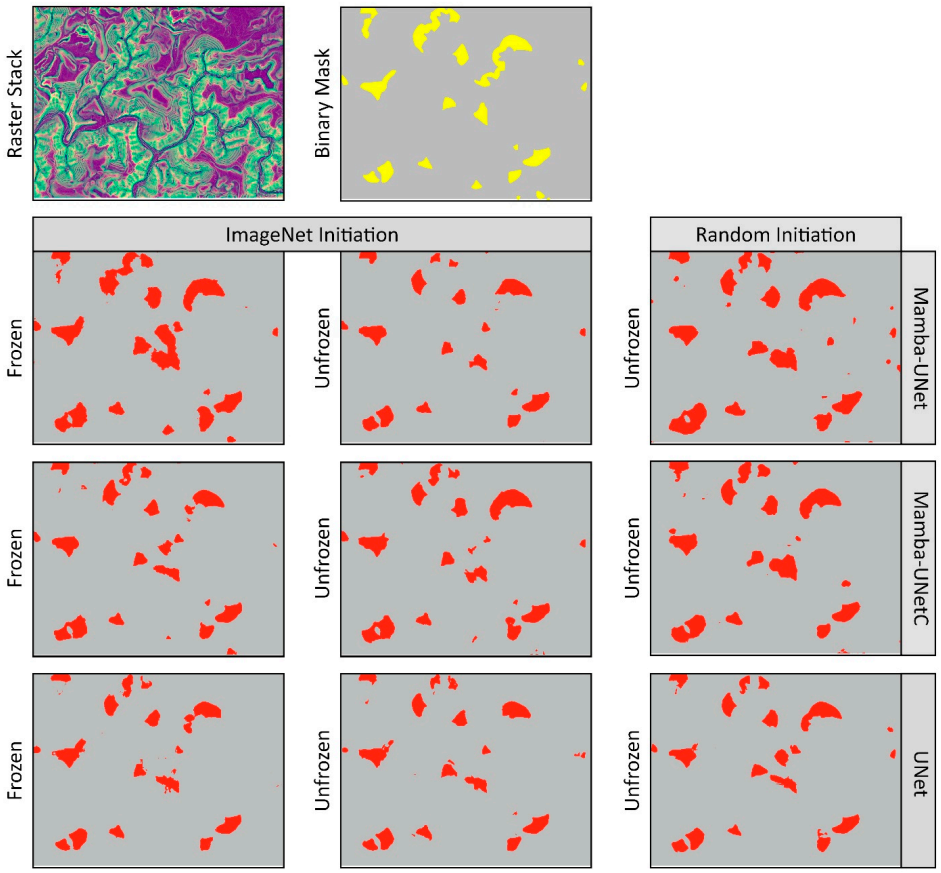

Figure 10. Spatial predictions for the valley fill face (vfillDL) test area. Panels compare predicted valley fills (red overlays) from each configuration against the reference mask (bottom-right) and the three-band LSP stack (bottom-center).

For the mine benches example, the target features form broad, sinuous, contour-parallel corridors with locally variable widths, making it important to capture both continuity and realistic feature thickness. In this case, the Mamba–UNet and Mamba–UNetC predictions generally recovered the bench network more completely than UNet, with both ImageNet-initialized and randomly initialized variants delineating most of the major bench segments across the scene. The ImageNet-initialized Mamba–UNet performed well in both frozen and unfrozen settings, preserving the main curvilinear structures with relatively good continuity, although some segments appeared slightly over-expanded and a few extra patches were introduced in lower-contrast areas. The Mamba–UNetC predictions were similarly strong and appeared somewhat cleaner in places, especially in the unfrozen setting, where the major bench traces were reproduced with good connectivity while maintaining clearer local geometry. By contrast, the ImageNet-initialized frozen UNet was noticeably more conservative, missing portions of several bench segments and fragmenting others into thinner or less continuous traces. Unfreezing improved the UNet prediction, and the randomly initialized UNet was also visually competitive, but both still appeared somewhat less complete than the Mamba-based models in broader and more sinuous bench sections. Overall, the visual results for mine benches suggest that the Mamba-based architectures, particularly Mamba–UNetC, generally reproduced the bench network more completely and with stronger continuity than the UNet configurations.

For the valley fill faces example, the target features appear as compact to lobate patches with relatively broad interiors and locally irregular margins, making both areal completeness and boundary placement important for visual evaluation. Across the predictions, the Mamba–UNet and Mamba–UNetC configurations generally recovered the

mapped fill faces more completely than the UNet models, especially for the larger patches. The ImageNet-initialized Mamba–UNet in the frozen setting produced broad and fairly complete delineations, although some predicted patches appeared slightly enlarged relative to the reference. Its unfrozen counterpart generated somewhat more compact and sharper footprints, but with a few additional omissions in smaller or weaker patches. The Mamba–UNetC predictions were also strong in both frozen and unfrozen settings, showing good agreement with the main mapped polygons while maintaining relatively coherent shapes and limited background noise; among the Mamba-based models, the unfrozen Mamba–UNetC appeared to provide a strong balance between completeness and shape coherence. In contrast, the ImageNet-initialized frozen UNet was more conservative and tended to underestimate the extent of several fill faces, while unfreezing improved coverage but still left some smaller patches and peripheral margins less completely represented. The randomly initialized UNet was visually competitive and captured many of the major fill faces with crisp, compact boundaries, but it remained somewhat more selective overall, particularly along weaker edges and smaller isolated patches. Overall, the visual results for valley fill faces indicate that the pretrained Mamba-based models, particularly Mamba–UNetC, generally produced more complete delineations with strong spatial agreement to the reference masks, while the randomly initialized UNet remained a strong but slightly more conservative alternative.

To complement the qualitative comparison with a quantitative measure of boundary quality, we computed the Boundary Intersection over Union (Boundary IoU) for each configuration on the three spatial-prediction test areas, alongside the standard region-based IoU (Table 3). Boundary IoU restricts the IoU computation to a narrow band along object edges and therefore isolates boundary delineation quality from interior region overlap. The reference labels were rasterized from the annotation polygons onto each prediction grid (2 m spatial resolution), and Boundary IoU was computed with a two-pixel (4 m) tolerance; rankings were consistent when the tolerance was widened to three pixels (6 m), indicating that the relative differences among models are not sensitive to the specific bandwidth. As expected, Boundary IoU values are lower than region IoU values because boundary agreement is more demanding than overall region overlap; the comparison below therefore emphasizes relative differences among configurations rather than absolute magnitudes. For agricultural terraces, whose narrow, contour-parallel geometry means that most feature pixels lie on or near a boundary, the CNN-based UNet configurations achieved substantially higher Boundary IoU (up to 0.379) than the Mamba-based configurations (0.176–0.233), indicating markedly better edge delineation for this elongated feature type. For the more compact mine benches and valley fill faces, several Mamba-based configurations achieved the highest region IoU (for example, Mamba–UNetC and Mamba–UNet exceeded 0.48 and 0.61, respectively), yet their Boundary IoU values were among the lowest, whereas the UNet and the unfrozen Mamba–UNetC configurations provided the best boundary agreement. This divergence between region overlap and boundary quality indicates that models can attain comparable overall accuracy through different error structures: the global receptive field of the Mamba-based encoders supports strong interior coverage and recovery of spatially extensive features, while the local inductive bias of the convolutional UNet and of the unfrozen hybrid more precisely localizes feature edges. These boundary-level statistics provide quantitative support for the qualitative patterns described above and clarify the mechanisms underlying the performance differences among architectures.

Table 3. Region-based IoU and Boundary IoU (B-IoU) for each model configuration on the three spatial-prediction test areas. Boundary IoU was computed with a two-pixel (4 m) tolerance at 2 m spatial resolution; model rankings were consistent at a three-pixel (6 m) tolerance. Higher values indicate better agreement with the reference annotations.

			Agricultural Terraces		Mine Benches		Valley Fill Face	
Model	**Initiation**	**Frozen/ Unfrozen**	**IoU**	**B-IoU**	**IoU**	**B-IoU**	**IoU**	**B-IoU**
Mamba–UNet	ImageNet	Frozen	0.451	0.199	0.484	0.054	0.583	0.022
		Unfrozen	0.460	0.210	0.472	0.049	0.616	0.049
	Random	Unfrozen	0.385	0.176	0.391	0.041	0.518	0.009
Mamba–UNetC	ImageNet	Frozen	0.469	0.221	0.459	0.049	0.613	0.027
		Unfrozen	0.462	0.233	0.488	0.067	0.599	0.021
	Random	Unfrozen	0.412	0.201	0.389	0.045	0.560	0.014
UNet	ImageNet	Frozen	0.424	0.313	0.355	0.057	0.630	0.043
		Unfrozen	0.416	0.307	0.393	0.063	0.612	0.048
	Random	Unfrozen	0.511	0.379	0.428	0.068	0.639	0.056

4.4. Model Efficiency

To complement the segmentation performance evaluation, we assessed the computational efficiency of each model in terms of trainable parameter count, FLOPs, and MACs, as summarized in Table 4. These indicators provide insight into the scalability of each architecture and its potential suitability for deployment in resource-constrained or time-sensitive remote sensing applications. The Mamba–UNet models were the most computationally efficient. The frozen Mamba–UNet configuration required 5.37 million trainable parameters, 374.19 GFLOPs, and 145.19 GMACs, whereas unfreezing the encoder increased the number of trainable parameters to 19.12 million without changing FLOPs or MACs, since only the trainable status of the encoder layers was altered. This substantially lower parameter count relative to the other architectures highlights the architectural efficiency of Mamba–UNet and its potential suitability for lower-memory deployment.

Table 4. Computational efficiency of the evaluated model configurations.

Model	**Frozen/ Unfrozen**	**Trainable Parameters**	**FLOPs**	**Macs**
Mamba–UNet	Frozen	5.37 M	374.19 GFLOPs	145.19 GMACs
	Unfrozen	19.12 M		
Mamba–UNetC	Frozen	33.74 M	1.61 TFLOPs	781.9 GMACs
	Unfrozen	47.49 M		
UNet	Frozen	3.15 M	1.57 TFLOPs	782.34 GMACs
	Unfrozen	24.44 M		

The Mamba–UNetC models incurred considerably greater computational cost. Both frozen and unfrozen Mamba–UNetC variants required 1.61 TFLOPs and 781.9 GMACs, while the number of trainable parameters increased from 33.74 million in the frozen setting to 47.49 million in the unfrozen setting. Although less efficient in terms of raw computation, these models consistently delivered strong segmentation performance across all three tasks, suggesting that the added complexity of the CNN-based decoder provides practical benefits for boundary refinement and spatial detail preservation.

The UNet models showed computational demands similar to those of Mamba–UNetC, requiring 1.57 TFLOPs and 782.34 GMACs in both frozen and unfrozen settings. However, the number of trainable parameters differed substantially between these configurations: the frozen UNet required only 3.15 million trainable parameters, whereas the unfrozen UNet

required 24.44 million due to the trainable encoder layers. Despite their computational cost being comparable to that of Mamba–UNetC, the UNet models generally produced weaker segmentation results, particularly when the encoder was frozen, indicating a less favorable performance-to-complexity ratio.

Taken together, these comparisons indicate that the Mamba-based models achieved competitive segmentation accuracy with fewer parameters and lower computational cost, while Mamba–UNetC provided a strong trade-off between performance and complexity, particularly for boundary-rich geomorphic features. In contrast, the UNet architectures were generally less efficient relative to their segmentation performance, especially when the pretrained encoder was kept frozen.

5. Discussion

This study evaluated whether Mamba-based state-space modeling can improve geomorphic feature segmentation from LiDAR-derived LSPs and whether combining a Mamba encoder with a CNN-based decoder can recover local detail while benefiting from broader spatial context. Across three contrasting tasks—agricultural terraces, mine benches, and valley fill faces—the results document that no single architecture was uniformly best in every case; however, several clear patterns emerged. Overall, the Mamba-based models, particularly Mamba–UNetC with ImageNet initialization, provided the most consistent performance across datasets, whereas the randomly initialized UNet remained unexpectedly competitive and achieved the highest positive-class F1-scores for agricultural terraces and valley fill faces. We distill six main findings and discuss their implications for LiDAR-based terrain segmentation workflows.

1. Mamba–UNetC provided the most consistent overall balance of segmentation performance and spatial prediction quality, although its advantage over the baseline UNet was limited and task-dependent.

Across the three datasets, Mamba–UNetC consistently ranked at or near the top in test-set F1-score and produced visually coherent predictions, especially for the agricultural terrace and mine bench tasks. Its predictions generally showed stronger continuity and more faithful recovery of elongated or sinuous target features than the UNet configurations, while also avoiding excessive fragmentation. These results support the idea that a Mamba encoder can effectively capture long-range spatial dependencies, while a CNN-based decoder improves local reconstruction through multiscale upsampling and skip connections. The resulting combination appears particularly well suited to geomorphic segmentation problems in which complete feature recovery and boundary quality are both important.

2. Mamba-based architectures showed faster and more stable optimization, but lower training loss did not always translate into the highest test F1-score.

The training curves showed that the Mamba–UNet and Mamba–UNetC models converged more rapidly and reached much lower loss values than the UNet models across all datasets. This pattern was evident even when the encoder was frozen, indicating that Mamba-based encoders retained useful representational capacity without full fine-tuning. However, the final test results also showed that faster convergence alone did not guarantee the best generalization. In particular, the randomly initialized UNet achieved the highest test F1-scores for agricultural terraces and valley fill faces, despite exhibiting higher training losses than the Mamba-based models. This suggests that optimization efficiency and final detection performance should be interpreted separately, especially in highly imbalanced segmentation tasks where precision–recall trade-offs strongly influence the positive-class F1-score.

3. Initialization and fine-tuning affected the model families differently.

A key result of this study was that random initialization was more detrimental for the Mamba-based models than for UNet, whereas freezing the encoder was more detrimental for UNet, especially for the ImageNet-initialized configurations. For the Mamba-based architectures, the ImageNet-initialized models clearly outperformed their randomly initialized counterparts on mine benches and valley fill faces, and the random Mamba variants also showed greater fluctuation in validation F1-score. In contrast, the randomly initialized UNet performed strongly on two of the three tasks and remained competitive on the third. At the same time, frozen ImageNet-initialized UNet was consistently among the weakest configurations, indicating that pretrained CNN features transferred less effectively to LiDAR-derived terrain variables unless adapted through task-specific training. These findings suggest that Mamba-based models benefit more from pretrained initialization, whereas CNN-based UNet models benefit more from encoder fine-tuning.

4. The relative strengths of the architectures depended on feature morphology and on the precision–recall trade-off.

The best-performing model varied by task, indicating that geomorphic feature geometry strongly influences model behavior. For the agricultural terraces and mine benches tasks, which involve narrow, elongated, and spatially connected features, the Mamba-based models, especially Mamba–UNetC, showed stronger continuity and more complete delineation. Their higher recall values suggest an advantage in recovering weak or partially fragmented terrain features that require broader contextual reasoning. In contrast, for valley fill faces, which are more compact and patch-like, the randomly initialized UNet achieved the highest test F1-score and remained visually competitive, reflecting a more precision-oriented detection style. The Mamba-based models, especially pretrained configurations, still performed strongly on valley-fill faces, but their predictions were often slightly broader or more recall-oriented. Thus, the choice of architecture should depend not only on the overall F1-score but also on whether the application prioritizes minimizing omissions or minimizing false positives.

The absolute positive-class F1-scores reported here (best values of approximately 0.53 for agricultural terraces, 0.60 for mine benches, and 0.68 for valley fill faces) are moderate, and several characteristics of the task help to explain this. First, the target features are rare relative to the background, producing severe class imbalance; under such imbalance, achieving high precision is intrinsically difficult because the large number of background pixels creates many opportunities for false positives. Second, many of the target landforms are narrow or small (for example, contour-parallel terraces and mine benches), so a large proportion of their pixels lie near object boundaries, where edge effects, annotation ambiguity, and boundary uncertainty disproportionately increase error. Third, some features, particularly valley fill faces, can resemble natural landforms, which further complicates precise delineation. Direct comparison of these scores with prior work is limited because, to our knowledge, these specific anthropogenic geomorphic feature extraction tasks have been examined almost exclusively in our own prior studies using these datasets [9,17,31,38]; relative to that prior work, the present results are consistent in magnitude, and we therefore interpret them as reflecting the inherent difficulty of the tasks rather than underperformance of the evaluated architectures.

5. Subsampling analysis confirmed that the main conclusions were robust, while OA remained relatively uninformative.

Recomputing performance metrics on 10 random subsamples containing 70% of the test chips showed that the relative ranking of the leading configurations remained stable across resamples. The ImageNet-initialized Mamba-based models, particularly

Mamba–UNetC, maintained a strong central tendency and generally modest variability, especially for mine benches, while the randomly initialized UNet remained among the strongest configurations for agricultural terraces and valley fill faces. These results indicate that the observed differences were not driven by a particular selection of test chips. At the same time, OA remained uniformly high across tasks and configurations because of the dominance of background pixels, confirming that it was far less diagnostic than the positive-class F1-score, precision, and recall for assessing segmentation quality under severe class imbalance.

6. Computational results showed a clear performance–complexity trade-off that favors Mamba-based models.

The efficiency analysis highlighted substantial differences in computational burden among architectures. The pure Mamba–UNet model was by far the most efficient, requiring far fewer trainable parameters and much lower computational cost than either Mamba–UNetC or UNet, while still delivering competitive segmentation performance. The Mamba–UNetC architecture incurred much greater computational cost because of its CNN-based decoder, but this added complexity was often justified by improved cross-dataset consistency and stronger spatial prediction quality. In contrast, UNet had computational demands similar to Mamba–UNetC yet generally delivered weaker results, particularly when the encoder was frozen, resulting in a less favorable performance-to-complexity ratio. From a practical standpoint, Mamba–UNet appears especially attractive when memory or computation is limited, whereas Mamba–UNetC offers a strong default when the goal is to maximize overall segmentation reliability.

Beyond these findings, several broader implications emerge. First, the results reinforce that global contextual reasoning and local boundary delineation are distinct and complementary requirements in LiDAR-based geomorphic mapping, and that they are not necessarily optimized by the same architecture. Incorporating Mamba-based global context improved interior feature recovery and spatial continuity, reducing fragmentation for spatially extensive or weakly expressed landforms; however, the boundary-level analysis showed that CNN-based UNet configurations achieved the most precise edge delineation, particularly for narrow, elongated features such as agricultural terraces. Architecture selection should therefore be guided by whether feature completeness or boundary accuracy is the primary objective. Second, the inconsistent behavior of ImageNet-pretrained CNN backbones suggests that pretraining mismatch remains an important issue for terrain segmentation. Domain-specific or self-supervised pretraining on large DTM or LSP corpora may improve transferability, particularly for CNN-based encoders. Third, the strong performance of frozen or partially frozen Mamba-based models points to promising resource-aware deployment options, where pretrained Mamba encoders can be paired with relatively lightweight decoders in settings with limited annotation or hardware budgets. Fourth, the observation that Mamba-based and CNN-based architectures could not be trained stably under a shared focal Tversky loss configuration and instead required architecture-specific parameterization suggests that loss configuration interacts with architecture in ways that are not yet well understood for this class of tasks. This points to a need for future work that systematically examines the disparate impact of loss parameterization across CNN, Mamba, and hybrid architectures, both to enable more fully controlled comparisons and to better understand the differing optimization behavior of these model families.

The benefit of combining CNNs and Mamba blocks has also been noted in recent hybrid segmentation studies. For example, Liu et al. [32] introduced CM-UNet, which combines CNN-based feature extraction with Mamba-based global modeling for semantic segmentation of high-resolution remotely sensed imagery. Similarly, Liu et al. [33] proposed

CWmamba for change detection, showing that Mamba-based global context modeling can complement CNN-based local feature extraction. Although those studies integrate Mamba and CNN components differently, they converge on the same broader conclusion: hybrid architectures that combine local detail modeling with long-range spatial reasoning are particularly effective for complex remote sensing tasks. Our findings extend this line of research by evaluating the inverse configuration, namely a Mamba-based encoder paired with a CNN-based decoder. The strong and consistent performance of Mamba–UNetC, especially for terraces and mine benches, suggests that the complementary strengths of Mamba and CNN modules can be leveraged effectively in either stage of the segmentation pipeline. These converging results highlight a promising direction for future hybrid architectures in geospatial analysis.

6. Conclusions

This study evaluated Mamba–UNet, UNet, and the hybrid Mamba–UNetC architectures for geomorphic feature segmentation from LiDAR-derived LSPs across three contrasting tasks: agricultural terraces, mine benches, and valley fill faces. No single architecture was best across all tasks; the randomly initialized UNet achieved the highest positive-class F1-score for agricultural terraces and valley fill faces and remained competitive for mine benches. Mamba–UNetC with ImageNet initialization provided the most consistent performance across datasets, combining effective long-range spatial context modeling with strong spatial prediction quality, though its advantage over the baseline UNet was modest and task-dependent rather than uniform. Feature morphology strongly influenced relative model performance. For the more elongated and spatially connected targets, particularly agricultural terraces and mine benches, the Mamba-based models generally provided more complete and continuous delineation, consistent with their stronger recall and better recovery of weak or fragmented features. Random initialization was more detrimental for the Mamba-based models, whereas freezing the encoder was more detrimental for UNet, especially when initialized with ImageNet weights. Mamba–UNetC is a strong default choice when the goal is to maximize segmentation reliability across diverse geomorphic targets and when computational resources permit the use of a more complex hybrid architecture, although its advantage over the baseline UNet was modest and task-dependent rather than uniform. Where precise boundary delineation is the primary objective, particularly for narrow, elongated features, CNN-based UNet configurations remained the strongest performers. In more resource-constrained settings, Mamba–UNet offers an attractive alternative because it achieved competitive performance with substantially fewer trainable parameters and lower computational cost.

Overall, this study demonstrates that Mamba-based architectures are a promising direction for LiDAR-based geomorphic mapping, and that combining a Mamba encoder with a CNN-based decoder can provide an effective balance between global spatial reasoning and local feature recovery, while boundary-level evaluation indicates that CNN-based configurations remain the most precise at delineating feature edges. We note that these results characterize within-task, spatially independent performance: each model was trained and evaluated on geographically disjoint regions within a given landform type, so the findings demonstrate generalization to new geographic areas within the same landscape and landform class rather than transfer across fundamentally different landforms or terrain contexts. We did not perform cross-region or cross-landform transfer experiments here; in related prior work, we examined transferring pretrained models across tasks and did not observe consistent improvements [38], and a fuller assessment of cross-landform generalization remains an important direction for future work. By clarifying when hybrid Mamba–CNN models are most beneficial and when conventional CNN-based models remain competitive,

this work provides practical guidance for architecture selection in high-resolution terrain segmentation workflows and supports continued development of state-space/CNN hybrid architectures for remote sensing segmentation.

Author Contributions: Conceptualization, S.F.; methodology, S.F. and A.E.M.; validation, S.F.; formal analysis, S.F.; investigation, S.F.; resources, A.E.M.; data curation, S.F.; writing—original draft preparation, S.F.; writing—review & editing, S.F. and A.E.M.; visualization, S.F. and A.E.M.; supervision, A.E.M.; project administration, A.E.M.; funding acquisition, A.E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation (Federal Award ID No. 2046059: "CAREER: Mapping Anthropocene Geomorphology with Deep Learning, Big Data Spatial Analytics, and LiDAR"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data Availability Statement: The data supporting the findings of this study are openly available in figshare. The terraceDL dataset is available at "terraceDL: A geomorphology deep learning dataset of agricultural terraces in Iowa, USA (https://figshare.com/articles/dataset/terraceDL_A_geomorpholgy_deep_learning_dataset_of_agricultural_terraces_in_Iowa_USA/22320373, accessed on 15 April 2026)". The mineBenchDL dataset is available at "mineBenchDL: A geomorphology deep learning dataset of historic surface coal mine benches in West Virginia, USA (https://figshare.com/articles/dataset/mineBenchDL_A_geomorphology_deep_learning_dataset_of_historic_surface_coal_mine_benches_in_West_Virginia_USA/26042920?file=47066854, accessed on 15 April 2026)". The vfillDL dataset is available at "vfillDL: A geomorphology deep learning dataset of valley fill faces resulting from mountaintop removal coal mining in southern West Virginia, eastern Kentucky, and southwestern Virginia, USA (https://figshare.com/articles/dataset/vfillDL_A_geomorpholgy_deep_learning_dataset_of_valley_fill_faces_resulting_from_mountaintop_removal_coal_mining_southern_West_Virginia_eastern_Kentucky_and_southwestern_Virginia_USA_/22318522, accessed on 15 April 2026)".

Acknowledgments: The authors would like to thank the Iowa Best Management Practices (BMP) Mapping Project for providing the agricultural terraces data and Muhammad Ali for assisting in the creation of the mineBenchDL dataset. Funding was provided by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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

References

1. Pashaei, M.; Kamangir, H.; Starek, M.J.; Tissot, P. Review and Evaluation of Deep Learning Architectures for Efficient Land Cover Mapping with UAS Hyper-Spatial Imagery: A Case Study over a Wetland. *Remote Sens.* **2020**, *12*, 959. [CrossRef]
2. Pan, X.; Gao, L.; Marinoni, A.; Zhang, B.; Yang, F.; Gamba, P. Semantic Labeling of High Resolution Aerial Imagery and LiDAR Data with Fine Segmentation Network. *Remote Sens.* **2018**, *10*, 743. [CrossRef]
3. Ji, S.; Shen, Y.; Lu, M.; Zhang, Y. Building Instance Change Detection from Large-Scale Aerial Images using Convolutional Neural Networks and Simulated Samples. *Remote Sens.* **2019**, *11*, 1343. [CrossRef]
4. Prakash, N.; Manconi, A.; Loew, S. Mapping Landslides on EO Data: Performance of Deep Learning Models vs. Traditional Machine Learning Models. *Remote Sens.* **2020**, *12*, 346. [CrossRef]
5. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sens.* **2019**, *11*, 196. [CrossRef]
6. Maxwell, A.E.; Bester, M.S.; Guillen, L.A.; Ramezan, C.A.; Carpinello, D.J.; Fan, Y.; Hartley, F.M.; Maynard, S.M.; Pyron, J.L. Semantic Segmentation Deep Learning for Extracting Surface Mine Extents from Historic Topographic Maps. *Remote Sens.* **2020**, *12*, 4145. [CrossRef]
7. Maxwell, A.E.; Odom, W.E.; Shobe, C.M.; Doctor, D.H.; Bester, M.S.; Ore, T. Exploring the Influence of Input Feature Space on CNN-Based Geomorphic Feature Extraction from Digital Terrain Data. *Earth Space Sci.* **2023**, *10*, e2023EA002845. [CrossRef]

8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer International Publishing: Cham, Switzerland, 2015.
9. Farhadpour, S.; Maxwell, A.E. Comparing UNet configurations for anthropogenic geomorphic feature extraction from land surface parameters. *PLoS ONE* **2025**, *20*, e0325904. [CrossRef] [PubMed]
10. Wang, Z.; Chen, L.; Shen, W.; Xiao, J.; Xu, Z.; Liu, J. RAF-Unet: A Remote Sensing Identification Method for Forest Land Information with Modified Unet. *J. Phys. Conf. Ser.* **2024**, *2868*, 012030. [CrossRef]
11. Wang, Z.; Qin, J.; Huang, C.; Zhang, Y. Multiscale Context-Aware Network for Remote Sensing Images Semantic Segmentation. *ResearchSquare* **2025**, *preprint*. [CrossRef] [PubMed]
12. Xiang, X.; Gong, W.; Li, S.; Chen, J.; Ren, T. TCNet: Multiscale Fusion of Transformer and CNN for Semantic Segmentation of Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3123–3136. [CrossRef]
13. Xu, Z.; Zhang, W.; Zhang, T.; Yang, Z.; Li, J. Efficient Transformer for Remote Sensing Image Segmentation. *Remote Sens.* **2021**, *13*, 3585. [CrossRef]
14. Tang, X.; Tu, Z.; Wang, Y.; Liu, M.; Li, D.; Fan, X. Automatic Detection of Coseismic Landslides Using a New Transformer Method. *Remote Sens.* **2022**, *14*, 2884. [CrossRef]
15. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
16. Bazi, Y.; Bashmal, L.; Al Rahhal, M.M.; Al Dayil, R.; Al Ajlan, N. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
17. Farhadpour, S.; Maxwell, A.E.; Solouki, B. SegFormer and SegFormer-UNet for anthropogenic geomorphic feature extraction from land surface parameters. *EarthArXiv* **2026**. [CrossRef]
18. Huang, Z.; Yang, Y.; Wang, Z.; Li, Y.; Chen, Z.; Ma, Y.; Zhang, S. MultiSEss: Automatic Sleep Staging Model Based on SE Attention Mechanism and State Space Model. *Biomimetics* **2025**, *10*, 288. [CrossRef] [PubMed]
19. Zhang, Y.; Jin, X.; Zhang, X.; Wu, Y.; Tu, L. EchoMamba: A new Mamba model for fast and efficient hyperspectral image classification. *PLoS ONE* **2025**, *20*, e0330678. [CrossRef] [PubMed]
20. Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; Wang, Z.; Ye, Z. A Survey on Visual Mamba. *Appl. Sci.* **2024**, *14*, 5683. [CrossRef]
21. Zhang, X.; Ou, W.; Wu, X.; Zhang, C. MHS-VIT: Mamba hybrid self-attention vision transformers for traffic image detection. *PLoS ONE* **2025**, *20*, e0325962. [CrossRef] [PubMed]
22. van der Meij, W.M.; Meijles, E.W.; Marcos, D.; Harkema, T.T.L.; Candel, J.H.J.; Maas, G.J. Comparing geomorphological maps made manually and by deep learning. *Earth Surf. Process. Landf.* **2022**, *47*, 1089–1107. [CrossRef]
23. Li, S.; Xiong, L.; Tang, G.; Strobl, J. Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery. *Geomorphology* **2020**, *354*, 107045. [CrossRef]
24. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
25. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
26. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [CrossRef]
27. Hatamizadeh, A.; Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025.
28. Ma, J.; Li, F.; Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv* **2024**, arXiv:2401.04722.
29. Xing, Z.; Ye, T.; Yang, Y.; Cai, D.; Gai, B.; Wu, X.-J.; Gao, F.; Zhu, L. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv* **2024**, arXiv:2401.13560.
30. Wang, Z.; Zheng, J.Q.; Zhang, Y.; Cui, G.; Li, L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv* **2024**, arXiv:2402.05079.
31. Maxwell, A.E.; Farhadpour, S.; Reed, M.M. A deep learning architecture and workflow for geomorphic feature extraction using digital terrain model-derived land surface parameters. *Earth Surf. Process. Landf.* **2026**, *51*, e70295. [CrossRef]
32. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2405.10530.
33. Liu, Y.; Cheng, G.; Sun, Q.; Tian, C.; Wang, L. CWmamba: Leveraging CNN-Mamba Fusion for Enhanced Change Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 2501505. [CrossRef]
34. Pires de Lima, R.; Marfurt, K. Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
35. Kuettel, D.; Guillaumin, M.; Ferrari, V. Segmentation Propagation in ImageNet. In *Computer Vision—ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012.

36. Chen, Z.; Zheng, Y. Research on Surface Information Extraction Based on Deep Learning and Transfer Learning. *J. Geosci. Environ. Prot.* **2023**, *11*, 67–78. [CrossRef]
37. Liu, F.; Lin, G.; Shen, C. CRF learning with CNN features for image segmentation. *Pattern Recognit.* **2015**, *48*, 2983–2992. [CrossRef]
38. Maxwell, A.E.; Farhadpour, S.; Ali, M. Exploring Transfer Learning for Anthropogenic Geomorphic Feature Extraction from Land Surface Parameters Using UNet. *Remote Sens.* **2024**, *16*, 4670. [CrossRef]
39. Tuggener, L.; Schmidhuber, J.; Stadelmann, T. Is it enough to optimize CNN architectures on ImageNet? *Front. Comput. Sci.* **2022**, *4*, 1041703. [CrossRef]
40. He, K.; Girshick, R.; Dollár, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019.
41. Maxwell, A. *terraceDL: A Geomorphology Deep Learning Dataset of Agricultural Terraces in Iowa, USA*; Figshare: London, UK, 2023. [CrossRef]
42. Maxwell, A.; Farhadpour, S.; Ali, M. *MineBenchDL: A Geomorphology Deep Learning Dataset of Historic Surface Coal Mine Benches in West Virginia, USA*; Figshare: London, UK, 2024. [CrossRef]
43. Maxwell, A. *vfillDL: A Geomorphology Deep Learning Dataset of Valley Fill Faces Resulting from Mountaintop Removal Coal Mining (Southern West Virginia, Eastern Kentucky, and Southwestern Virginia, USA)*; Figshare: London, UK, 2023. [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
45. Hijmans, R.J. *Terra: Spatial Data Analysis R Package*; Version 1.7-78; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: https://CRAN.R-project.org/package=terra (accessed on 15 April 2026).
46. Ilich, A.; Lecours, V.; Misiuk, B.; Murawski, S. *MultiscaleDTM: Multi-Scale Geomorphometric Terrain Attributes*; (v0.8.3); Zenodo: Geneva, Switzerland, 2024.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016.
48. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8009205. [CrossRef]
49. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
50. Liu, N.; He, T.; Tian, Y.; Wu, B.; Gao, J.; Xu, Z. Common-azimuth seismic data fault analysis using residual UNet. *Interpretation* **2020**, *8*, SM25–SM37. [CrossRef]
51. Cao, K.; Zhang, X. An Improved Res-UNet Model for Tree Species Classification Using Airborne High-Resolution Images. *Remote Sens.* **2020**, *12*, 1128. [CrossRef]
52. PyTorch. Available online: https://www.pytorch.org (accessed on 31 December 2020).
53. Welcome to Python.org. Available online: https://www.python.org/ (accessed on 5 January 2021).
54. Yakubovskiy, P. Segmentation Models. GitHub Repository. Available online: https://github.com/qubvel-org/segmentation_models.pytorch. (accessed on 15 April 2026).
55. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2009.
56. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [CrossRef]
57. Farhadpour, S.; Warner, T.A.; Maxwell, A.E. Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sens.* **2024**, *16*, 533. [CrossRef]
58. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192.