applied sciences

MDPI

Article

DEGC-TransUNet: A Dual-Encoder TransUNet with Global Context Enhancement for Mountaintop Area Extraction from Grid DEMs

Fangbin Zhou [1,2], Junwei Bian [1,*] and Jiamin Huang [1]

[1] School of Aeronautic Engineering, Changsha University of Science & Technology, Changsha 410114, China; arthur1975@126.com (F.Z.); 24201050174@csust.edu.cn (J.H.)
[2] Engineering Laboratory of Special Information Technology of Highway Geological Disaster Early Warning in Hunan Province, Changsha University of Science & Technology, Changsha 410114, China
* Correspondence: 23101020005@stu.csust.edu.cn; Tel.: +86-182-73445238

Abstract

Accurate extraction of mountaintop areas from grid digital elevation models (DEMs) is essential for terrain analysis, geomorphological research, hydrological modeling, natural disaster monitoring, and emergency communication site selection. However, existing deep-learning-based methods often suffer from inadequate representation of local details and limited global contextual awareness, leading to blurred boundaries and reduced segmentation accuracy in complex mountainous terrains. To address these limitations, this study proposes a dual-encoder and global-context-enhanced TransUNet framework, named DEGC-TransUNet, for automated mountaintop delineation. The architecture integrates a convolutional encoder to capture fine-grained local terrain features and a MaxViT-based encoder to model multi-scale global context by encoding low-dimensional topographic attributes such as slope and curvature. A dedicated feature fusion module harmonizes complementary representations from both encoding paths, while a BiFormer-based strategy is introduced at the bottleneck to strengthen long-range dependencies and enhance convergence. The experimental results demonstrate that DEGC-TransUNet significantly outperforms baseline models such as TransUNet, DE-TransUNet, and GC-TransUNet, with relative improvements of 19.8% in Intersection over Union (IoU), 10.4% in overall accuracy (ACC), and 10.9% in F1-score. These findings provide a robust solution for mountaintop extraction, with significant potential in analyzing geomorphological evolution, simulating soil erosion, modeling species distribution in "sky island" ecosystems, and optimizing strategic placements for communication base stations and wind energy infrastructures.

Keywords: grid DEMs; mountaintop area extraction; dual-encoder network; TransUNet; vision transformer; feature fusion; global context modeling

Check for updates

Academic Editor: Nathan J. Moore

Received: 1 April 2026
Revised: 5 May 2026
Accepted: 6 May 2026
Published: 8 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital Elevation Models (DEMs) provide a fundamental digital representation of the Earth's surface and are widely used in terrain analysis, geomorphological mapping, hydrological modeling, and environmental assessment [1–3]. As a structured form of terrain representation, grid DEMs describe surface elevation using regularly spaced grid cells, making them especially well-suited for quantitative terrain analysis and spatial modeling [4]. As a key terrain element, mountaintop areas reflect the morphological characteristics and spatial structure of mountainous landscapes and play an important role in mountain

partitioning, ridge–valley structure analysis, and landform evolution studies [5–7]. Accurate extraction of mountaintop areas from grid DEMs is therefore of significant importance for both theoretical geomorphology and practical geospatial applications.

Mountaintop areas are typically characterized by local elevation maxima and their surrounding regions, exhibiting strong spatial continuity, gradual transitions, and often indistinct boundaries influenced by adjacent terrain [8]. These characteristics make mountaintop area extraction a challenging task, especially in regions with complex topography or subtle elevation variations [9,10]. Despite decades of methodological evolution—moving from traditional rule-based geometric analysis [11–14] to machine-learning statistical classifiers [15–19], and recently to deep-learning-based semantic segmentation [20–24]—effectively balancing local detail representation with global contextual understanding remains a key challenge. Specifically, traditional and CNN-based methods often struggle with boundary blurring or structural discontinuities in complex spatial organizations. Meanwhile, pure Transformer-based models face hurdles in computational efficiency and fine-scale feature preservation [25–27].

To address these gaps, this study proposes a Dual-Encoder TransUNet with Global Context Enhancement (DEGC-TransUNet) for mountaintop area extraction from grid DEMs. The proposed model integrates a convolutional encoder for fine-scale local feature extraction with a Multi-Axis Vision Transformer (MaxViT) encoder for multi-scale global context modeling [28]. To effectively reconcile local granularity and global semantics, we employ a fusion module incorporating coordinate attention and Squeeze-and-Excitation (SaE) mechanisms [29,30], supplemented by a BiFormer-based enhancement module at the bottleneck layer [31,32]. The principal contributions of this study are summarized as follows:

Innovative Hybrid Dual-Path Architecture: By synergistically integrating CNNs and MaxViT, our framework captures both fine-grained local textures and long-range spatial dependencies, overcoming the receptive field limitations of traditional convolutional models.

Efficient Global Context Enhancement: The introduction of a BiFormer-based module at the bottleneck enables the model to explicitly reinforce broad geomorphological relationships while maintaining high computational efficiency through dynamic sparse attention.

Enhanced Boundary Fidelity and Precision: Our approach specifically addresses the challenge of indistinct mountaintop boundaries in complex terrains, achieving superior segmentation accuracy and structural continuity compared to current State-of-the-Art (SOTA) methods.

To evaluate the effectiveness of the proposed model, comprehensive experiments were conducted across diverse mountainous regions with varying geomorphological characteristics. The results suggest that DEGC-TransUNet provides measurable improvements in segmentation accuracy and boundary fidelity compared to the baseline TransUNet model [33], demonstrating its potential for robust generalization in complex terrains.

The remainder of this paper is structured to provide a comprehensive exploration of our proposed framework. Section 2 reviews the relevant literature on traditional and deep-learning-based terrain analysis. Section 3 details the architectural design and implementation of the DEGC-TransUNet. Section 4 presents the experimental setup, results, and comparative analysis. Section 5 provides an in-depth discussion of the experimental findings and model limitations, followed by Section 6, which offers concluding remarks and outlines potential future research directions.

2. Related Work

2.1. Traditional and Machine-Learning-Based Extraction

The identification of mountaintop elements traditionally relies on rule-based geometric analysis of grid DEMs. Early researchers proposed methods such as the extreme value method based on terrain profile convexity [34] and the window comparison method based on raster neighborhood analysis for rapid peak point extraction [10]. Other approaches introduced inverse DEM hydrological simulation and spatial partitioning strategies [12,35]. Although these methods are conceptually intuitive, they are highly sensitive to parameters such as analysis window size and elevation thresholds. Studies by Gu et al. [36] and Liu et al. [37] demonstrated that improper threshold settings lead to significant deviations in both the number and location of extracted peaks, particularly in complex terrains where boundaries remain poorly defined [11,12].

With the advancement of machine learning, statistical classifiers such as Support Vector Machines (SVM) and Random Forests (RF) were introduced to overcome the limitations of explicit rules. Researchers utilized these models to classify micro-landforms by learning from multiple terrain factors, such as slope, curvature, and elevation [38,39]. Li et al. [18] reported that RF-based ensemble learning significantly outperformed traditional rule-based methods in hilly regions. However, these statistical approaches rely heavily on feature engineering. As noted by Du et al. [40], handcrafted features often struggle to maintain stable performance in terrains with high geomorphological heterogeneity, leading to decreased generalization across different landscapes [15,16].

2.2. Deep Learning and Convolutional Neural Networks

Deep learning, particularly Convolutional Neural Networks (CNNs), has shifted the paradigm toward automated feature learning from high-resolution DEMs. CNN-based architectures, such as U-Net, FCN, and ResNet, have been widely applied to terrain classification and mountaintop recognition [41–43]. For instance, Zhang [20] utilized CNNs combined with terrain factors for landform classification, while Li et al. [44] proposed an improved Faster R-CNN for peak point identification using grayscale contour overlays. Zhou et al. [22] further demonstrated the effectiveness of CNNs in classifying specific mountain components, such as shoulders and peaks.

Despite their success in capturing local textures, CNNs are inherently limited by the locality of convolutional operations. The restricted receptive field prevents these models from effectively capturing long-range spatial dependencies and global contextual information [24,25,45]. In complex, mountainous regions characterized by interlocking ridges and valleys, this limitation often results in "salt-and-pepper" noise, fragmented structures, and blurred boundaries for areal entities, such as mountaintop regions [16,36].

2.3. Transformer and Hybrid Architectures

To address the limitations of local receptive fields, Transformer-based frameworks leveraging self-attention mechanisms have been introduced. These architectures can explicitly model relationships between all feature units, capturing large-scale geomorphological structures. Chen et al. [5] proposed a terrain knowledge-guided Transformer for geomorphon mapping, and Wang et al. [26] enhanced landform encoding by incorporating topographic priors into the attention mechanism.

Recent trends have focused on hybrid CNN–Transformer models to synergize local detail extraction with global context modeling. Examples include the integration of Transformers into U-Net-like encoders [32,33] and the use of dual-attention mechanisms to fuse spatial and channel features [46,47]. Wang et al. [48] successfully applied such hybrid

models to seismic fault recognition, demonstrating their capacity for extracting linear and structural geographic features.

However, hybrid models encounter significant challenges regarding computational efficiency and adaptive feature fusion. The quadratic computational complexity of standard Transformers makes them inefficient for processing large-scale, high-resolution grid DEMs [25,47]. Furthermore, balancing the fine-grained granularity of local terrain with broad geomorphological semantics remains difficult. Excessive global attention may smooth out critical local peaks, while insufficient integration leads to structural fragmentation [49]. Most existing methods also struggle with cross-regional generalization and are heavily dependent on large labeled datasets [23,27]. This study seeks to bridge these gaps by proposing the DEGC-TransUNet, which utilizes dynamic sparse attention and multi-scale fusion to achieve robust and efficient mountaintop area extraction.

3. Materials and Methods

3.1. Overall Framework of DEGC-TransUNet

The DEGC-TransUNet adopts a U-Net-style encoder–decoder backbone [42], leveraging its multi-scale feature aggregation and spatial detail preservation capabilities. To overcome the limitations of traditional convolutional architectures in modeling long-range spatial dependencies, we introduce a dual-path encoding strategy (Figure 1).

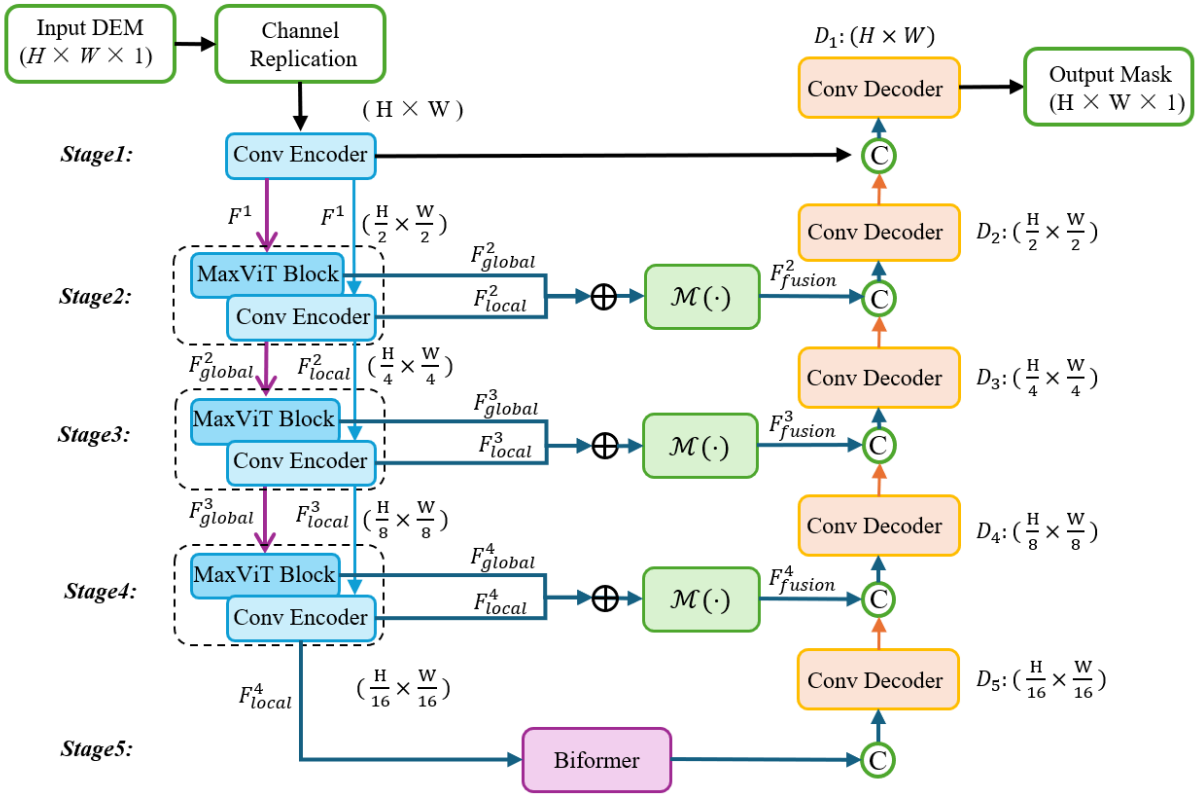


Figure 1. Architecture of the proposed DEGC-TransUNet for mountaintop area extraction from grid DEMs. The arrows indicate the direction of data flow, and different colors are used to distinguish various functional modules and various feature map scales.

For an input single-channel DEM $X \in \mathbb{R}^{H\times W\times 1}$, a Channel Replication step is first performed to transform it into a three-channel representation $\hat{X} \in \mathbb{R}^{H\times W\times 3}$, where the elevation values are identical across all channels. This ensures compatibility with the Transformer-based backbone while preserving the original topographic information [50].

In the initial stage (Stage 1), the Convolutional Encoder (Conv Encoder) processes $\hat{X}$ to capture fundamental topographic primitives, such as high-resolution edges and initial elevation gradients. Starting from Stage 2 to Stage 4, the architecture transitions into a dual-path structure where the Conv Encoder concurrently works with a MaxViT-based Encoder [28] to extract complementary features (Figure 1). The feature extraction at these dual-path stages is formulated as:

$$F^l_{local} = ConvEnc^l\left(F^{l-1}\right) \tag{1}$$

$$F^l_{global} = MaxVit^l\left(F^{l-1}\right) \tag{2}$$

where $l \in \{2, 3, 4\}$. F^l_{local} preserves local structural details essential for accurate extraction, while F^l_{global} captures multi-scale global contextual information through multi-axis attention mechanisms (Figure 1).

To harmonize these representations, a dedicated Feature Fusion Module (FFM), denoted as $\mathcal{M}$, is implemented at each dual-path encoding stage (Stages 2–4, Figure 1):

$$F^l_{fusion} = \mathcal{M}\left(F^l_{local} \oplus F^l_{global}\right) \tag{3}$$

where $\oplus$ represents the channel-wise concatenation. The FFM effectively reorganizes the fused features, reducing redundancy while enhancing the synergy between local details and global contextual cues.

At the structural bottleneck, a BiFormer module [47] is incorporated to further strengthen the global representation capability (Figure 1). Finally, the decoder reconstructs the segmentation map through symmetric upsampling stages. High-resolution skip features—from the Conv Encoder in Stage 1 and the FFM in Stages 2–4—are merged with the upsampled decoder features at the concatenation nodes (denoted as 'C' in Figure 1). This fusion process, formulated as $F_{out} = Concat\left(F_{decoder}, F_{fused}\right)$, ensures that the network leverages both macroscopic contextual cues and fine-grained structural fidelity for precise mountaintop delineation.

3.2. Dual-Encoder Architecture and Feature Fusion

The dual-encoder architecture is designed to synergize local topographic primitives with global contextual representations for improved mountaintop delineation (Figure 2).

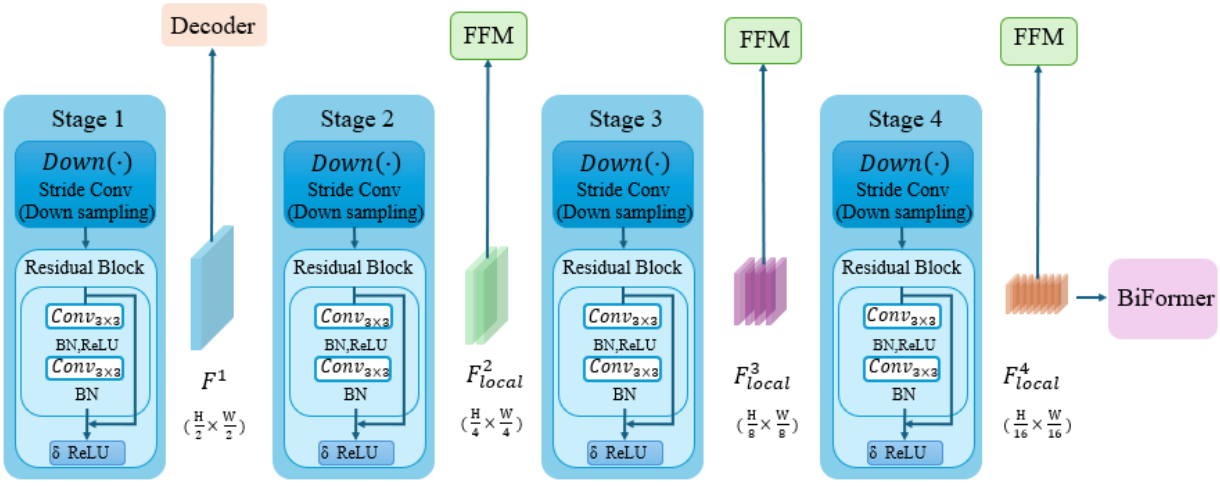


Figure 2. Schematic diagram of the convolutional encoder architecture.

3.2.1. Convolutional and MaxViT Encoders

The convolutional encoder utilizes a hierarchical downsampling structure of stacked residual convolutional layers to capture local spatial patterns [43], as illustrated in Figure 2. At each stage l, the local feature extraction is expressed as:

$$F_{local}^{l} = SC\left(F_{local}^{l-1}\right) = \delta\left(Conv_{3\times3}\left(Down\left(F_{local}^{l-1}\right)\right)\right) \tag{4}$$

where $Down(\cdot)$ denotes the strided convolution for downsampling and δ represents the ReLU activation function.

In parallel, the MaxViT-based encoder introduces Transformer-style attention mechanisms to model long-range spatial dependencies, as illustrated in Figure 3. As mentioned in Section 3.1, the three-channel input $\hat{X}$ ensures compatibility with the Transformer backbone. The MaxViT block captures global relationships through Multi-Axis Attention [28], decomposing the process into block-wise and grid-wise components:

$$F_{global}^{l} = GSA\left(\mathrm{BSA}\left(F_{global}^{l-1}\right)\right) \tag{5}$$

where BSA and GSA denote Block-wise and Grid-wise Self-Attention, respectively. This mechanism allows the model to distinguish true mountaintop areas from locally elevated regions that are contextually irrelevant by considering the macroscopic terrain structure [49].

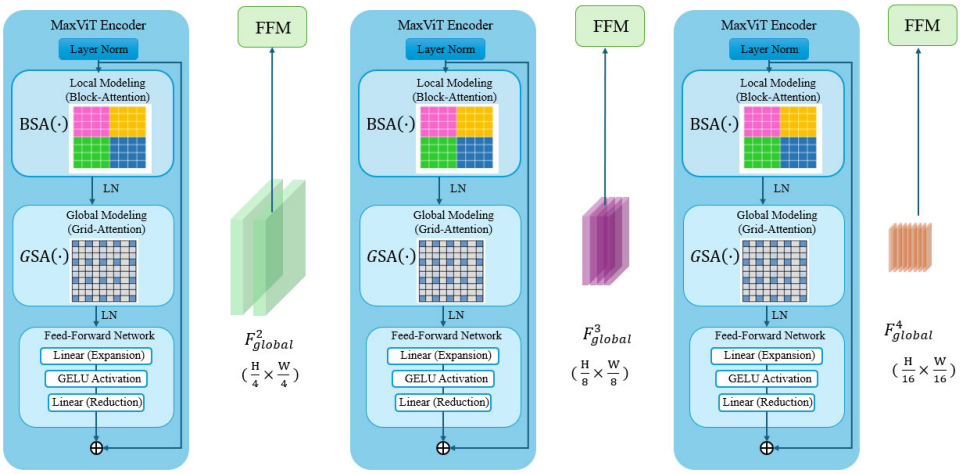


Figure 3. Schematic diagram of the MaxViT-based encoder architecture.

3.2.2. Att-SaE Feature Fusion Module

To integrate the heterogeneous features, the Att-SaE Module is employed to reconcile representations through a spatial-channel refinement process, as detailed in Figure 4. First, Coordinate Attention (CA) [29] is applied to the concatenated features:

$$F_{CA} = CA\left(F_{local}^{l} \oplus F_{global}^{l}\right)) \tag{6}$$

where CA recalibrates the feature maps by aggregating spatial information along horizontal and vertical axes. Subsequently, the features are refined by a Squeeze-and-Excitation (SaE) [30] layer:

$$F_{fusion}^{l} = \sigma(\mathrm{g}(\mathrm{Conv}(F_{CA}))) \bigotimes \mathrm{Conv}(F_{CA}) \tag{7}$$

where $\mathrm{g}(\cdot)$ denotes global average pooling, σ is the sigmoid function, and $\bigotimes$ represents element-wise multiplication. The SaE layer adaptively reweights channels to boost the discriminative power of the fused representations before they are passed to the decoder.

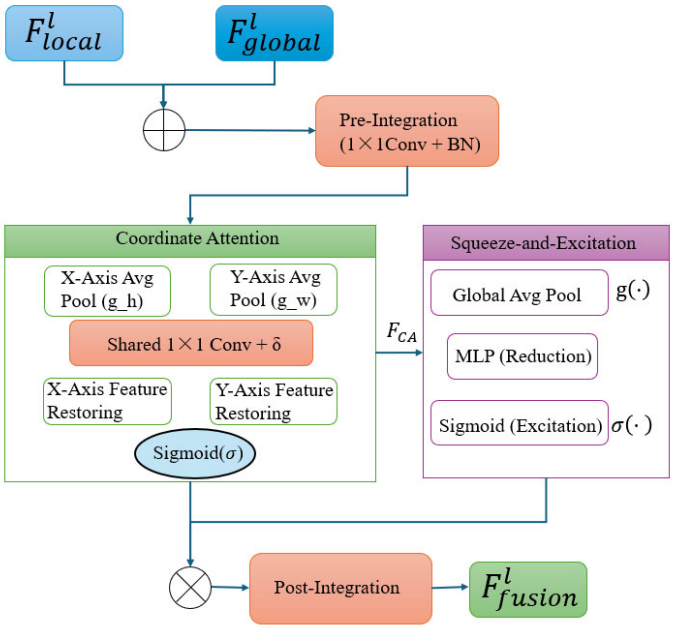


Figure 4. Detailed structure of the Att-SaE feature fusion module.

3.3. Global Context Enhancement via Bottleneck MixStructureBlock

To strengthen long-range dependency modeling while maintaining training efficiency, a global context enhancement module, termed MixStructureBlock, is incorporated at the structural bottleneck (Figure 5). This module integrates an improved BiFormer [47] mechanism with Fast Fourier Convolution (FFC) [51] to capture macroscopic terrain structures through a multi-stage residual refinement process.

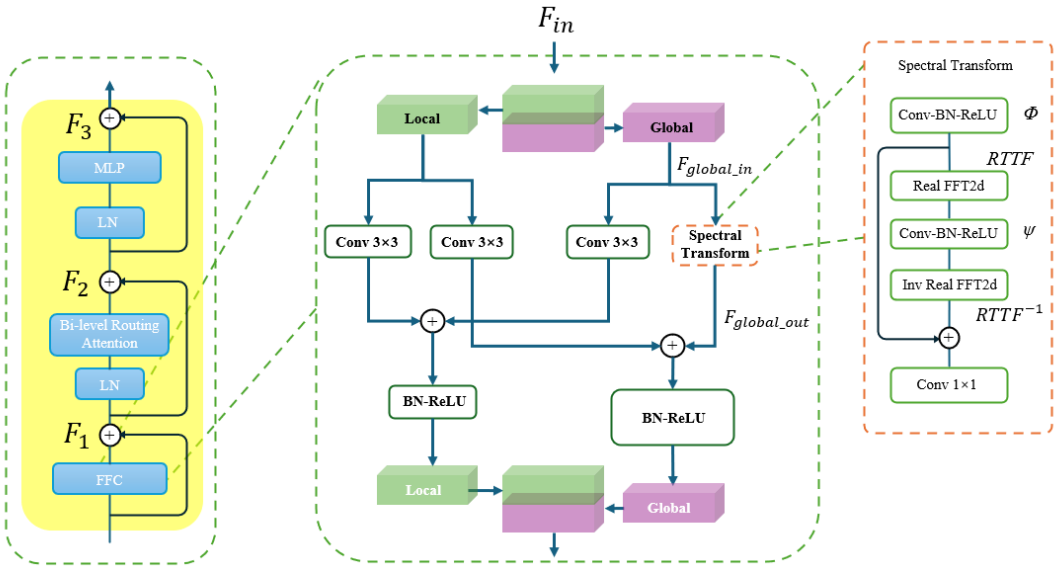


Figure 5. The structure of the BiFormer-based global context enhancement module integrating Fast Fourier Convolution (FFC), which models long-range spatial dependencies via sparse attention routing and introduces frequency-domain feature enhancement to complement spatial representations.

3.3.1. Hierarchical Residual Refinement

As illustrated in the primary backbone of the MixStructureBlock, the input bottleneck feature F_{in} undergoes three successive stages of transformation. To ensure training stability

and prevent gradient degradation, each stage is implemented with Layer Normalization (LN) and element-wise addition (denoted by + in Figure 5) for residual learning:

$$F_1 = F_{in} + FFC(LN(F_{in})) \tag{8}$$

$$F_2 = F_1 + BRA(LN(F_1)) \tag{9}$$

$$F_3 = F_2 + MLP(LN(F_2)) \tag{10}$$

where BRA denotes the Bi-level Routing Attention, and MLP represents the multi-layer perceptron. This sequential design allows the model to progressively integrate spectral global cues and sparse-attentional spatial relationships.

3.3.2. Fast Fourier Convolution (FFC) with Spectral Transform

The FFC module within the MixStructureBlock employs a dual-branch architecture that splits the feature map into a Local path and a Global path. While the Local path utilizes conventional convolutions to preserve spatial details, the Global path captures features in the frequency domain using a Spectral Transform [51] (Figure 5). Based on the properties of real-valued topographic signals, the spectral branch is formulated as:

$$F_{global_out} = Conv_{1\times1}\left(F_{global_in} + RFFT^{-1}\left(\psi\left(RTTF\left(\varnothing\left(F_{global_in}\right)\right)\right)\right)\right) \tag{11}$$

where $RTTF$ and $RFFT^{-1}$ represent the Real 2D Fast Fourier Transform and its inverse, respectively. $\varnothing$ denotes the spatial preprocessing (Conv-BN-ReLU) and ψ represents the convolution in the frequency domain. By accounting for the global context across the entire DEM grid, FFC effectively identifies consistent geomorphological structures.

3.3.3. Bi-Level Routing Attention (BRA)

To complement the spectral features, the BRA component facilitates content-aware sparse attention by routing queries to the most relevant key-value pairs. The routing index matrix A_{route} is determined by calculating the affinity between region-level queries Q_g and keys K_g:

$$A_{route} = Top_k(\text{Softmax}(\frac{Q_g\left(K_g\right)^T}{\sqrt{d}})) \tag{12}$$

By restricting attention computations to the Top_k routed regions, the BiFormer mechanism enhances the global representation capability for mountaintop delineation with significantly reduced computational cost compared to dense self-attention mechanisms. This dual-path modeling at the bottleneck ensures that both macroscopic terrain context and fine-grained structural fidelity are preserved.

3.4. Decoder, Loss Functions, and Inference Strategy

The decoder utilizes a symmetric hierarchical upsampling structure based on bilinear interpolation, adhering to the classical U-Net design [42] to reconstruct high-resolution segmentation maps. At each decoding stage l, the feature map D^l is generated by upsampling the preceding lower-resolution features and integrating spatially aligned features from the encoder through skip connections.

Specifically, the fused encoder features F^l_{fusion} from the FFM are incorporated to enhance spatial detail recovery and boundary reconstruction accuracy. The decoding process at each stage can be formulated as:

$$D^l = Conv\left(Up\left(D^{l-1}\right) \oplus F^l_{fusion}\right) \tag{13}$$

where $Up(\cdot)$ denotes the bilinear upsampling operation, and $\oplus$ represents the channel-wise concatenation used to bridge the semantic gap between the encoder and decoder.

By concatenating the low-level spatial primitives with high-level semantic features, the decoder effectively recovers the fine-grained topographic boundaries of mountaintops. The final output layer employs a 1×1 convolution followed by a Softmax function to produce the pixel-wise classification map for mountaintop areas.

To address class imbalance between mountaintop areas and background regions, a hybrid loss function combining Binary Cross-Entropy (BCE) Loss and Dice Loss is used during training [30,31]. Dice-based loss functions [52] are commonly applied in segmentation tasks with imbalanced class distributions and have proven effective in improving region-level consistency and boundary delineation. During inference, whole-image prediction is used to avoid boundary artifacts commonly introduced by sliding-window-based inference strategies [53]. This approach is particularly critical for preserving the spatial continuity of mountaintop regions.

The total objective function is formulated as follows:

$$Loss = \lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{Dice}} \tag{14}$$

The Binary Cross-Entropy loss is defined as:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{15}$$

where N denotes the total number of samples, $y_i \in \{0, 1\}$ represents the ground-truth label for sample i, and p_i is the predicted probability for the corresponding class. Additionally, the Dice Loss is formulated as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{M} y_i p_i + \epsilon}{\sum_{i=1}^{M} y_i + \sum_{i=1}^{M} p_i + \epsilon} \tag{16}$$

where M is the number of pixels (or data points).y_i is the ground truth value for pixel i.p_i is the predicted value for pixel i. The smoothing term ϵ is added to ensure numerical stability and avoid division by zero during training.

Specifically, the weighting parameters, λ_1 and λ_2, in Equation (14) are both set to 1. This configuration was finalized after systematic experimental testing. Specifically, we conducted a sensitivity analysis by varying the ratio of λ_1 and λ_2 (e.g., 1:0.5, 1:1, and 1:2). The results indicated that the 1:1 ratio effectively balances the gradient contributions of L_{BCE} and L_{Dice}, which exhibit similar numerical magnitudes. This balance prevents either loss term from dominating the optimization process, thereby ensuring the most stable convergence and superior segmentation performance."

4. Experiments

4.1. Study Area and Dataset

4.1.1. Study Area Overview

Experiments were conducted in the Jiuyuangou watershed (110°11′52.8″–110°25′19.2″ E, 37°31′18″–37°38′16.8″ N), situated in the middle reaches of the Wuding River basin on the northern Loess Plateau, Shaanxi Province, China. The study area covers approximately 250 km^2 and is characterized by typical loess hilly–gully landforms [16]. This region features a highly dissected terrain with an average ground-surface slope of 27° and a gully density of 6.15 km/km^2. The elevation ranges from 807.632 to 1198.60 m, with a maximum relative relief of 390 m between hilltops and gully bottoms. The extreme topographic

complexity—marked by severe soil erosion, fully developed gullies, and broken terrain-makes this area a representative site for evaluating the robustness and applicability of mountaintop extraction methods under complex morphological conditions.

4.1.2. Dataset Generation and Partitioning

The primary data source utilized in this study consists of high-resolution DEMs with a spatial resolution of 5 m, obtained from the National Fundamental Geographic Information Database of China [54]. To construct a reliable supervision signal, representative mountaintop areas were manually delineated through expert visual interpretation, producing high-quality binary reference masks.

To ensure the geomorphological rigor of the dataset and mitigate the limitations of traditional threshold-based segmentation, we implemented an expert-knowledge-guided stepwise annotation protocol. As illustrated in Figure 6, this hierarchical workflow transitions from discrete point identification to continuous physical entity delineation through three integrated stages:

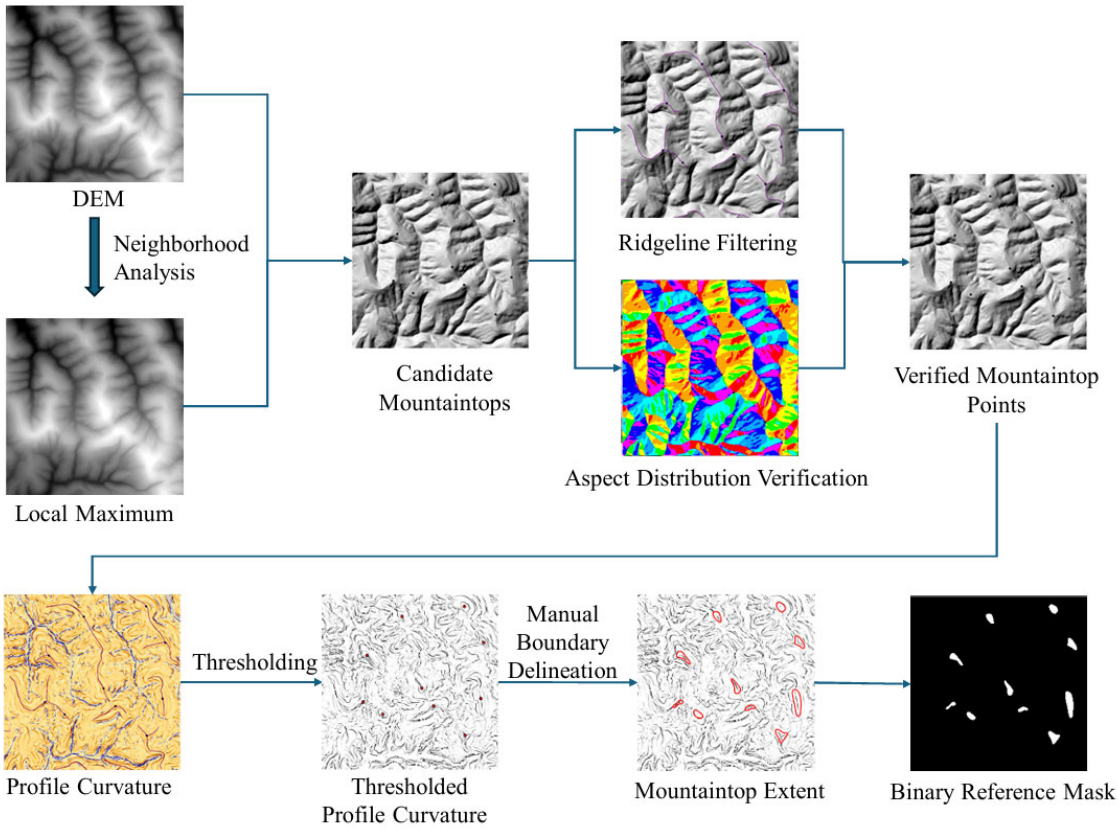


Figure 6. The expert-knowledge-guided stepwise annotation protocol for mountaintop reference mask generation. The arrows in this figure indicate the logical sequence of the workflow, connecting consecutive processing steps.

Geometric Localization: Initial candidate mountaintops are identified by detecting local elevation maxima through a neighborhood analysis within a sliding window. This stage establishes the potential centroids of the target features.

Morphological Verification via Multi-feature Constraints: To eliminate pseudo-peaks and noise, we perform a dual-verification process. First, we analyze the aspect distribution; a true mountaintop must exhibit a radial-divergent pattern, acting as a center from which all aspects slope away. Second, spatial ridgeline constraints are applied to ensure that

the candidate points are topographically aligned with the terrain skeleton, effectively suppressing isolated noise and false extrema.

Adaptive Boundary Delineation: Unlike arbitrary fixed-radius labeling, the physical boundaries of the reference masks are determined using thresholded profile curvature. By isolating curvature values sensitive to the summit-to-slope transition, we can identify the geomorphological break-points where the terrain transitions from the convex summit to the stable side-slope. This ensures that the final binary reference masks capture the authentic physical extent of mountaintops across diverse terrains and resolutions.

This self-adaptive approach allows the model to learn from high-fidelity supervision signals that inherently encode complex terrain characteristics without requiring manual normalization of raw topographic factors.

In total, 1264 valid samples were generated, each comprising a localized DEM patch and its corresponding binary mountaintop mask. To enhance the validity and generalization of the model, the dataset was partitioned into training and testing subsets at an 8:2 ratio. Unlike purely random partitioning, a spatial systematic sampling strategy with a fixed step size was employed to select the test samples. This approach ensures that the test set is uniformly distributed across the entire study area and encompasses diverse terrain units (as shown in Figure 7b), thereby providing a more rigorous and geographically representative assessment of the model's performance.

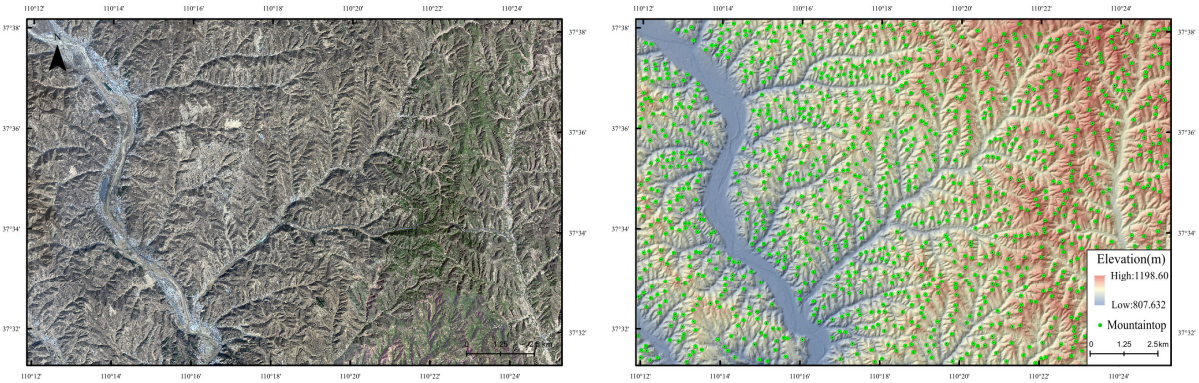


Figure 7. Overview of the study area and sample distribution: (**a**) high-resolution remote sensing imagery of the Jiuyuangou region; (**b**) spatial distribution of manually delineated mountaintop area samples. (Base data provided by the National Geomatics Center of China).

4.2. Experimental Setup

The experimental platform was equipped with high-performance computing resources, including an AMD EPYC 7325 CPU, 256 GB of system memory, and two NVIDIA GeForce RTX 3090 GPUs. All models, including the proposed DEGC-TransUNet, baseline models, and ablation variants, were trained using this dual-GPU setup under identical hyperparameter configurations to ensure a fair and rigorous comparison.

All experiments were implemented using the PyTorch (version 1.10.0+cu113) deep-learning framework [55] on a Windows-based platform. The Adam optimizer [56] was employed for network optimization, with an initial learning rate set to 0.001. To enhance convergence stability and prevent the model from falling into local optima during the training process, a polynomial learning rate decay strategy [57] was adopted, which is formulated as:

$$l = lr_{initial} \times \left(1 - \frac{iter}{\max_iter}\right)^{power} \quad (17)$$

where $lr_{initial}$ is the initial learning rate, *iter* and max_*iter* represent the current and maximum number of iterations, respectively. The training process followed an iteration-based scheme with a total of 10,000 iterations, ensuring the network achieved full convergence.

4.3. Evaluation Metrics

To rigorously assess model performance, several standard quantitative metrics for semantic segmentation are used, including Precision, Recall, F1-score, Kappa, Overall Accuracy (ACC), and Intersection over Union (IoU) [58]. These metrics offer complementary perspectives, enabling a comprehensive evaluation of the model's predictive capabilities across various terrains.

The specific formulations are as follows:

Precision quantifies the proportion of correctly identified positive pixels relative to all pixels predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

Recall (or Sensitivity) measures the model's ability to capture all actual positive pixels:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

F1-score is the harmonic mean of Precision and Recall, providing a balanced metric for classification performance:

$$\text{F1} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{20}$$

Kappa [59] coefficient evaluates the agreement between predicted and observed labels while accounting for random chance:

$$\text{Kappa} = \frac{2 \times (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)} \tag{21}$$

ACC measures the proportion of correctly classified pixels among the total sample size:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

IoU reflects the geometric overlap between the predicted and ground-truth regions:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{23}$$

In these equations, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Collectively, these metrics facilitate a holistic assessment of segmentation fidelity across physiographically diverse landscapes.

In addition to the aforementioned accuracy metrics, we introduce two indicators to evaluate the computational complexity and efficiency of the models:

Floating-point operations (FLOPs) are employed to quantify the theoretical computational cost of the model. It represents the total number of floating-point arithmetic operations required during a single forward pass of the network, serving as a hardware-independent measure of model complexity.

Inference Latency is defined as the average time required for the model to process a single input sample during the inference phase. In this study, latency is calculated as the mean time across all samples in the test set to ensure a stable and representative measurement of real-world deployment efficiency.

4.4. Training Convergence Analysis

All models exhibit a rapid decrease in training loss during the initial training stage (Figure 8a), indicating effective optimization at early iterations. The baseline TransUNet demonstrates a relatively slow convergence rate and maintains a higher loss value throughout the training process. In contrast, both DE-TransUNet and GC-TransUNet achieve lower loss values than the baseline, suggesting that introducing either the dual-encoder architecture or the global context enhancement module can improve optimization efficiency to some extent. Notably, the proposed DEGC-TransUNet converges significantly faster than the other models and reaches the lowest final loss value, reflecting more stable optimization dynamics and improved learning capacity. This behavior indicates that the combination of dual-encoder feature learning and BiFormer-based global context modeling effectively facilitates gradient propagation and reduces optimization difficulty.

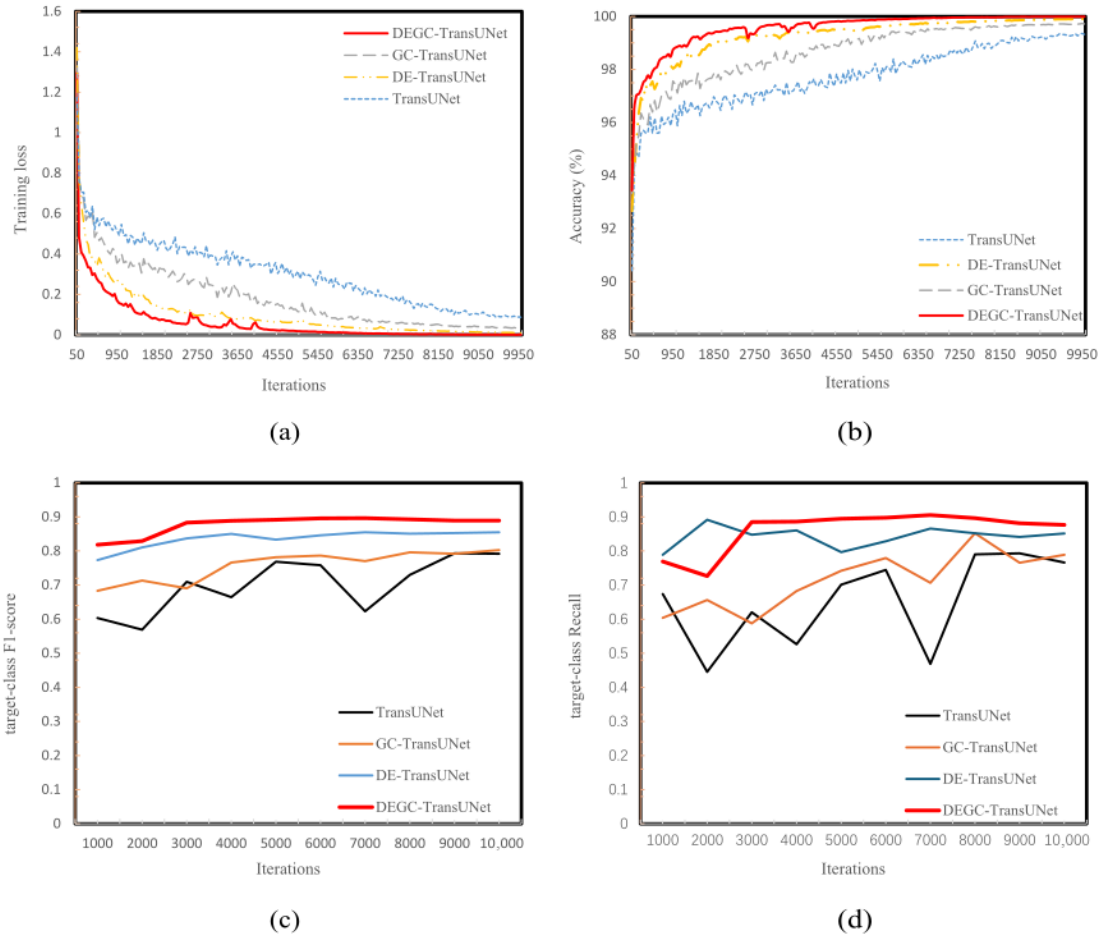


Figure 8. Convergence and validation performance analysis for mountaintop extraction using four architectural variants: (**a**) training loss, (**b**) training accuracy, (**c**) target-class F1-score on validation set, and (**d**) target-class Recall on validation set.

The training accuracy curves in Figure 8b further confirm these observations. While all models show a steady increase in accuracy with training progress, the baseline TransUNet exhibits slower accuracy improvement and lower final accuracy. Both DE-TransUNet and GC-TransUNet achieve higher accuracy levels than the baseline, demonstrating that enhancing either local–global feature representation or long-range dependency modeling contributes positively to segmentation performance. The proposed DEGC-TransUNet consistently outperforms all comparison models across the entire training process, reaching near-saturated accuracy at an earlier stage and maintaining superior performance thereafter.

This indicates faster convergence and stronger representation learning capability. To further justify the quality of the test process and evaluate the generalization capability on unseen data, the validation F1-score and Recall for the target class were recorded every 1000 iterations (Figure 8c,d). Unlike continuous training curves, these validation checkpoints provide periodic assessments of the models' practical segmentation performance.

For instance, the baseline TransUNet and the ablation models show significant performance drops at 2000 and 7000 iterations, suggesting a high sensitivity to specific validation samples or optimization instability. While DE-TransUNet achieves a high Recall peak at 2000 iterations, it subsequently experiences a decline, indicating transient overfitting during the early phase. In contrast, the proposed DEGC-TransUNet consistently maintains the highest F1-score and Recall throughout the entire validation process. Despite minor oscillations, its performance trajectory remains overall upward and reaches superior peaks (e.g., a F1-score of 0.8968 and a Recall of 0.9054 at 7000 iterations). More importantly, in the final stages (8000 to 10,000 iterations), DEGC-TransUNet exhibits the most stable performance plateau, effectively mitigating the drastic oscillations observed in the baseline. This alignment between steady training convergence and superior validation stability provides strong evidence of the robustness and exceptional generalization capability of the DEGC-TransUNet framework.

Overall, the convergence analysis demonstrates that the proposed DEGC-TransUNet not only achieves higher segmentation accuracy but also exhibits improved training stability and efficiency. The synergistic integration of dual-encoder feature extraction and global context enhancement enables more effective optimization compared with single-module improvements, providing a solid foundation for the superior quantitative and qualitative results reported in subsequent sections.

4.5. Comparison with Baseline Methods

To evaluate the comprehensive performance of the proposed DEGC-TransUNet, a series of comparative experiments was conducted against several well-established baseline models, ranging from lightweight architectures to complex transformer-based networks. The baselines include FCN [42], FPN [60], FCN-ResNet18 [43], PSPNet t [61], DeepLabV3+ [62], and the standard TransUNet. All models were trained and tested under identical dataset partitions and evaluation protocols to ensure a fair comparison.

The quantitative results across multiple dimensions, including segmentation accuracy, model complexity, and computational efficiency, are summarized in Table 1.

Table 1. Quantitative evaluation of mountaintop extraction performance, model complexity, and computational efficiency among different methods.

Method	Precision	Recall	F1-Score	Kappa	ACC	IoU	Flops (G)	Params (M)	Time	Latency (ms)
TransUNet	83.27	79.16	81.16	0.633	81.62	68.29	43.4	109.07	1:18:00	219.51
Fcn	88.89	81.28	84.92	0.711	85.56	73.79	4.455	13.047	0:36:35	1.88
Fpn	88.04	81.44	84.61	0.704	85.19	73.33	4.455	13.047	0:45:28	1.86
Fcn-ResNet18	86.20	81.57	83.82	0.685	84.26	72.15	2.53	13.54	0:13:48	1.86
Pspnet	90.31	80.82	85.30	0.722	86.08	74.37	1.45	11.334	0:11:30	1.90
Deeplabv3plus	89.01	78.52	83.43	0.688	84.41	71.58	4.58	12.33	0:34:16	1.87
DEGC-TransUNet	90.97	89.07	90.01	0.802	90.12	81.84	57.95	90.54	2:22:58	211.38

4.5.1. Segmentation Accuracy Analysis

As shown in Table 1, DEGC-TransUNet consistently achieves the highest performance across all accuracy-related metrics. Specifically, it reaches an IoU of 81.84% and an F1-score

of 90.01%, representing a significant leap compared to the second-best baseline model (PSPNet with 74.37% IoU).

Precision and Recall: The proposed model achieves a Precision of 90.97% and a Recall of 89.07%, indicating that the dual-encoder structure effectively reduces both false positives (misidentified terrain) and false negatives (missed peaks).

Boundary and Region Continuity: While lightweight models like FCN-ResNet18 and FCN exhibit lower IoU (ranging from 72.15% to 73.79%), they often struggle with fragmented predictions in complex terrain. In contrast, DEGC-TransUNet demonstrates superior boundary delineation and region continuity, confirming that the integration of global context enhancement is crucial for capturing the macro-geomorphological patterns of mountaintops.

4.5.2. Model Complexity and Efficiency

A critical aspect of this comparison is the trade-off between extraction accuracy and computational overhead.

Computational Cost (FLOPs): DEGC-TransUNet involves 57.95 GFLOPs, which is higher than the lightweight ResNet18-based models (e.g., FCN-ResNet18 at 2.53 GFLOPs). However, compared to the standard TransUNet (43.4 GFLOPs), the increased computational cost is justifiable given the substantial 13.55% improvement in IoU.

Parameter Count: Notably, despite its superior accuracy, DEGC-TransUNet maintains a lower parameter count (90.54 M) compared to the original TransUNet (109.07 M). This suggests that the proposed architectural modifications, such as the DEGC module, provide a more efficient parameter utilization for terrain feature learning.

Inference Latency: In terms of real-world deployment, DEGC-TransUNet achieves an average latency of 211.38 ms per image, which is slightly faster than the baseline TransUNet (219.51 ms). Although lightweight baselines like FCN and PSPNet offer near-real-time performance (approx. 1.8–1.9 ms), their significantly lower extraction accuracy limits their utility in precision-demanding geomorphological research.

Overall, although DEGC-TransUNet is heavier than traditional CNN-based baselines in terms of FLOPs and latency, it strikes an optimal balance by delivering State-of-the-Art accuracy with manageable computational costs. The results confirm that for complex mountaintop extraction, the integration of long-range spatial dependencies and multi-scale feature fusion provides a level of precision that lightweight models cannot currently achieve.

4.5.3. Summary of Comparison

The results highlight that while lightweight models offer advantages in inference speed, they fail to provide the necessary precision for delicate mountaintop extraction. DEGC-TransUNet strikes an optimal balance by delivering State-of-the-Art accuracy with manageable computational costs, outperforming the original TransUNet in both extraction quality and model efficiency (lower parameters and latency). This proves that jointly modeling multi-scale terrain details and long-range dependencies is highly effective for processing grid DEM data.

4.6. Performance Comparison and Ablation Study

To quantify the individual and synergistic contributions of the proposed architectural components, a structure-level ablation study was conducted. We evaluated the baseline TransUNet, the dual-encoder version (DE-TransUNet), the global context version (GC-TransUNet), and the complete DEGC-TransUNet [63]. The results, encompassing extraction accuracy, model complexity, and computational efficiency, are summarized in Table 2.

Table 2. Quantitative ablation study of architectural components in terms of extraction accuracy, model complexity, and computational efficiency.

Method	Precision	Recall	F1-Score	Kappa	ACC	IoU	Flops (G)	Params (M)	Time	Latency (ms)
TransUNet	83.27	79.16	81.16	0.633	81.62	68.29	43.4	109.07	1:18:00	219.51
DE-TransUNet	86.88	86.84	86.86	0.737	86.86	76.78	74.64	170.76	2:26:26	268.29
GC-TransUNet	83.11	81.21	82.15	0.647	82.35	69.71	48.56	114.14	1:11:37	203.25
DEGC-TransUNet	90.97	89.07	90.01	0.802	90.12	81.84	57.95	90.54	2:22:58	211.38

The integration of the dual-encoder mechanism (DE-TransUNet) resulted in a substantial performance leap, with the IoU increasing from 68.29% to 76.68%. This improvement confirms that the dual-path design significantly enhances the model's ability to jointly capture fine-scale terrain details and multi-scale contextual features from DEM data. However, this enhancement comes at the cost of increased model complexity; the parameter count rose to 170.76 M, and the FLOPs reached 74.64 G, leading to a higher inference latency of 268.29 ms. These results suggest that while the dual-encoder mechanism is critical for accuracy, a naive implementation introduces significant parameter redundancy.

The GC-TransUNet, which incorporates only the global context enhancement module, showed a moderate improvement in accuracy, reaching an IoU of 69.71%. A key observation is the module's high efficiency: it maintains a lower latency (203.25 ms) and lower computational cost (48.56 GFLOPs) compared to the baseline TransUNet. This demonstrates that the GC module effectively models long-range dependencies and mitigates local misclassifications in complex terrains without imposing a significant computational burden.

The complete DEGC-TransUNet achieves the highest performance across all metrics, with an IoU of 81.84% and an F1-score of 90.01%. Most notably, the final model achieves a superior balance between performance and efficiency:

Complexity Optimization: Through optimized architectural integration, the parameters of DEGC-TransUNet (90.54 M) are significantly lower than those of both the DE-TransUNet (170.76 M) and the baseline TransUNet (109.07 M).

Computational Efficiency: The model exhibits a reduced FLOPs (57.95 G) compared to the intermediate DE-TransUNet, and its inference latency (211.38 ms) is even faster than the original TransUNet (219.51 ms).

These findings indicate that the joint modeling of local terrain details and global context not only maximizes extraction precision but also—through our proposed integration strategy—optimizes the model's structural efficiency, making it highly suitable for high-precision geomorphological tasks.

The experimental results demonstrate highly consistent performance across all experimental runs through a 5-fold cross-validation framework. The model achieved a mean F1-score of 90.01% with a marginal standard deviation of $\pm$0.42%, and a mean Accuracy (ACC) of 90.12% $\pm$ 0.38%. Specifically, for the representative fold (the results of which are detailed in Table 2), the IoU and Precision for the target mountaintop areas reached 81.84% and 90.97%, respectively. The extremely low standard deviation observed across the different data partitions confirms that the model is highly robust and effectively mitigates potential concerns regarding overfitting or random performance fluctuations. Such a high level of consistency demonstrates the model's superior generalization capability in capturing invariant geomorphological patterns across diverse and complex terrain units in the study area.

4.7. Qualitative Analysis Based on Landform Types

To facilitate a geomorphologically coherent interpretation of the ablation results, a qualitative analysis was conducted by categorizing mountaintop samples according to their morphological attributes [45]. Adopting the taxonomic framework proposed by Yang et al. [6] regarding landform evolution, mountaintop areas were grouped into four representative types based on shape-related indices: quasi-circular, elliptical, linear, and irregular mounds. These morphologies were quantitatively characterized using a suite of geometric descriptors widely used in landscape spatial analysis [64,65], including circularity index, perimeter–area ratio, and boundary fractal dimension (Table 3), which represent geometric regularity, boundary compactness, and morphological complexity, respectively.

Table 3. Classification criteria of loess landform units based on morphological indices.

Landform Type	Circularity Index	Perimeter–Area Ratio	Boundary Fractal Dimension
Quasi-circular mound	1.05–1.10	23.0–45.0	1.26–1.30
Elliptical mound	1.11–1.20	22.0–45.0	1.27–1.31
Linear mound	1.21–1.35	24.0–40.0	1.28–1.33
Irregular mound	1.36–1.53	18.0–45.0	1.28–1.34

Circularity Index (D_i): This index, also known as the patch compactness index, reflects the degree to which a patch shape approximates a circle:

$$D_i = \frac{P_i}{2\sqrt{\pi A_i}} \tag{24}$$

where D_i is the circularity index, P_i is the patch perimeter, and A_i is the patch area. The minimum value is 1, indicating a perfect circle; higher values signify increasing morphological irregularity.

Perimeter-Area Ratio (C_i): This metric, often referred to as the patch roundness index, represents the edge-to-area relationship and the intensity of the boundary effect:

$$C_i = \frac{P_i}{A_i} \tag{25}$$

where C_i is the perimeter–area ratio. This index effectively captures the boundary compactness, as larger patches or those with smoother margins exhibit lower C_i values. The Fractal Dimension (FD_i) parameter quantifies the shape complexity and spatial filling capacity of the patch boundary:

$$FD_i = \frac{2\ln P_i}{\ln A_i} \tag{26}$$

where FD_i denotes the fractal dimension. Higher FD_i values correspond to more convoluted and complex mountaintop boundaries, reflecting high levels of terrain dissection.

Figure 9 presents qualitative comparisons of different models across mountaintop samples with decreasing circularity and increasing morphological complexity. Samples (a) and (b) correspond to quasi-circular mounds, characterized by compact shapes and smooth boundaries. Samples (c) and (d) represent elliptical mounds, exhibiting moderate elongation and boundary irregularity. Sample (e) corresponds to a linear mound, while sample (f) represents an irregular mound with highly complex boundaries.

For quasi-circular mounds (Figure 9a,b), all models are able to identify the general mountaintop region; however, the baseline TransUNet tends to produce fragmented predictions or minor false positives around the boundary. The dual-encoder architecture improves local feature sensitivity, while the global context enhancement further stabi-

lizes region integrity. The proposed DEGC-TransUNet achieves the most compact and boundary-consistent results, closely matching the ground-truth labels.

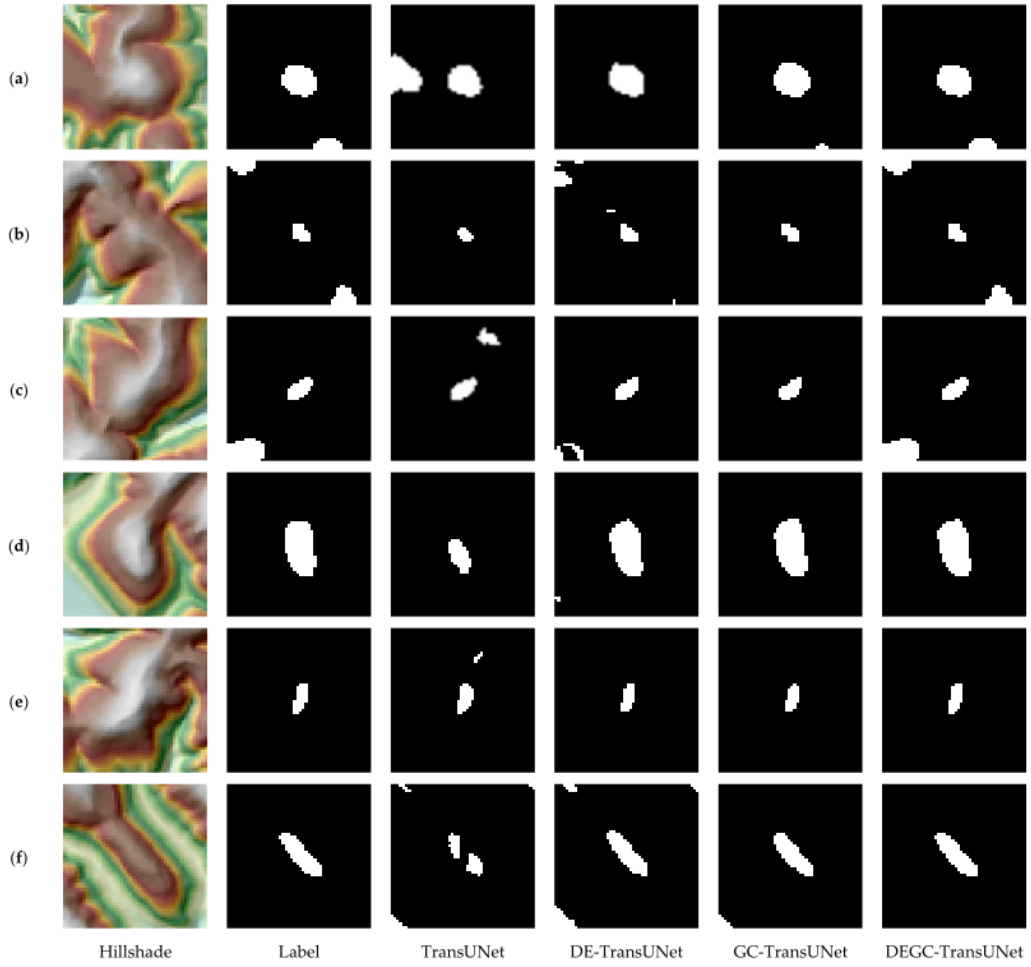


Figure 9. Qualitative comparison of mountaintop area extraction results across different morphological samples: (**a**) quasi-circular mound 1; (**b**) quasi-circular mound 2; (**c**) elliptical mound 1; (**d**) elliptical mound 2; (**e**) linear mound; (**f**) irregular mound. From left to right: DEM with color-shaded relief, ground truth, and predictions by TransUNet (baseline), DE-TransUNet (dual-encoder variant), GC-TransUNet (global-context variant), and the proposed DEGC-TransUNet.

In the case of elliptical mounds (Figure 9c,d), boundary delineation becomes more challenging due to elongated geometry. The baseline TransUNet exhibits noticeable under-segmentation and boundary distortion. The Dual-Encoder TransUNet (DE-TransUNet) improves boundary continuity by enhancing local feature representation, whereas the Global-Context-enhanced TransUNet (GC-TransUNet) better preserves overall shape consistency. DEGC-TransUNet effectively integrates both advantages, producing predictions that accurately capture the elongated structure while maintaining smooth boundaries.

For linear mounds (Figure 9e), long-range spatial dependency becomes critical. Models relying primarily on local convolutional features tend to break the structure into disconnected segments. The introduction of BiFormer-based global context modeling significantly alleviates this issue by capturing large-scale terrain continuity. As a result, DEGC-TransUNet produces a continuous and structurally coherent prediction that aligns well with the reference label.

In the most challenging irregular mound case (Figure 9f), characterized by high boundary complexity and irregular geometry, all ablation variants exhibit varying degrees of misclassification. The baseline and single-module variants suffer from fragmented

predictions and boundary noise. In contrast, DEGC-TransUNet demonstrates superior robustness, effectively suppressing false positives while preserving the overall mound structure. This result highlights the complementary roles of dual-encoder feature learning and global context enhancement in handling complex geomorphological patterns.

Overall, the qualitative results indicate that the proposed DEGC-TransUNet consistently outperforms baseline and ablation models across different landform types. The performance gap becomes more pronounced as mountaintop morphology transitions from regular to irregular forms, confirming that joint modeling of local terrain details and global spatial context is essential for accurate mountaintop area extraction in complex geomorphological settings.

Although the baseline models can generally identify the main mountaintop structures, their predictions often suffer from boundary fragmentation and incomplete region coverage, especially in areas with complex terrain morphology (Figure 10).

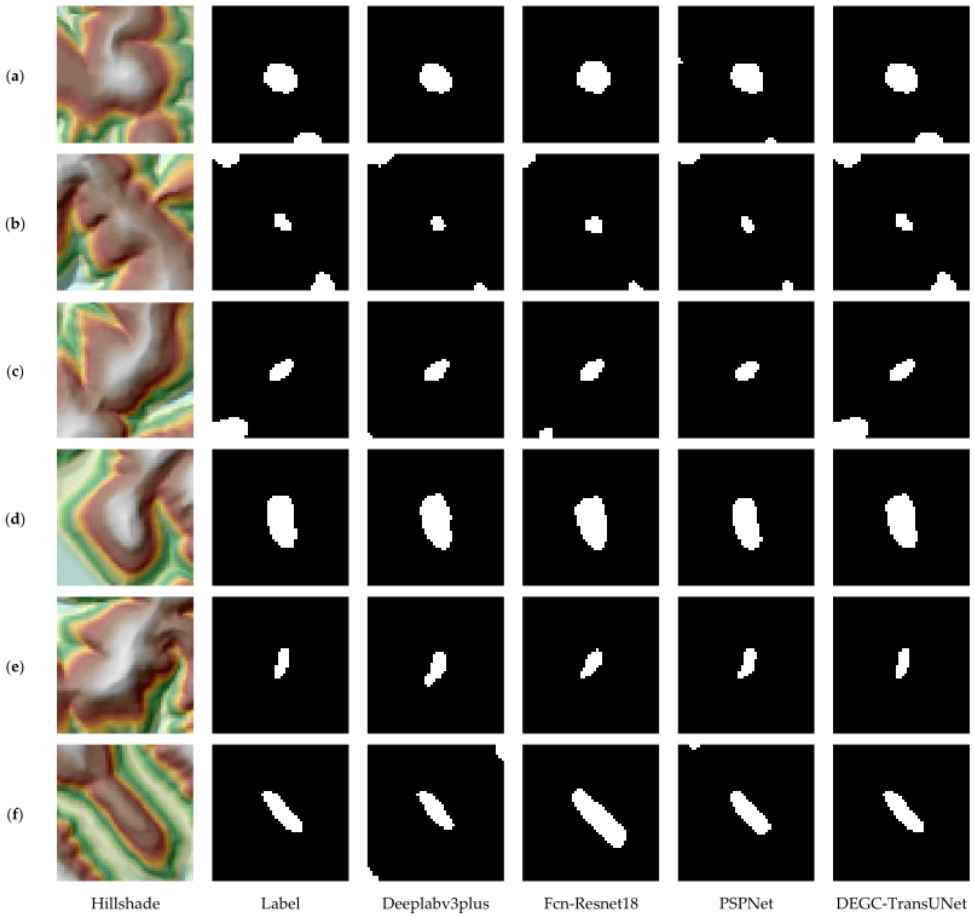


Figure 10. Qualitative comparison of mountaintop area extraction results across different morphological samples: (**a**) quasi-circular mound 1; (**b**) quasi-circular mound 2; (**c**) elliptical mound 1; (**d**) elliptical mound 2; (**e**) linear mound; (**f**) irregular mound. From left to right: DEM with color-shaded relief, ground truth, and predictions by Deeplabv3plus, Fcn-ResNet18, Pspnet, and the proposed DEGC-TransUNet.

In contrast, DEGC-TransUNet generates more spatially continuous and morphologically consistent mountaintop regions. The extracted results exhibit improved boundary integrity and reduced local misclassification, which is crucial for reliable geomorphological interpretation. This indicates that the proposed model is more capable of capturing both global terrain context and local elevation-related features, thereby providing more stable and accurate extraction results.

To further assess the cross-regional generalization ability of the proposed DEGC-TransUNet framework, the model was applied to several independent mountainous regions using 1 m resolution DEMs from the OpenTopography facility [66]. These test areas, entirely excluded from the training phase, cover diverse geomorphological settings across multiple geographic locations, including Yosemite National Park and Alaska (United States), New Zealand, Pakistan, and China.

The proposed model demonstrates remarkable adaptability across a wide range of heterogeneous terrains (Figure 11). Notably, predictions from the Gansu Loess Plateau exhibit the highest consistency and accuracy, with clearly defined and continuous boundaries. This strong performance can be attributed to the geomorphological similarity between the training and test regions. In contrast, test areas in New Zealand and Pakistan, which feature complex lithological and tectonic structures, show slightly increased boundary uncertainty, highlighting the impact of spatial heterogeneity [67] on model generalization [68].

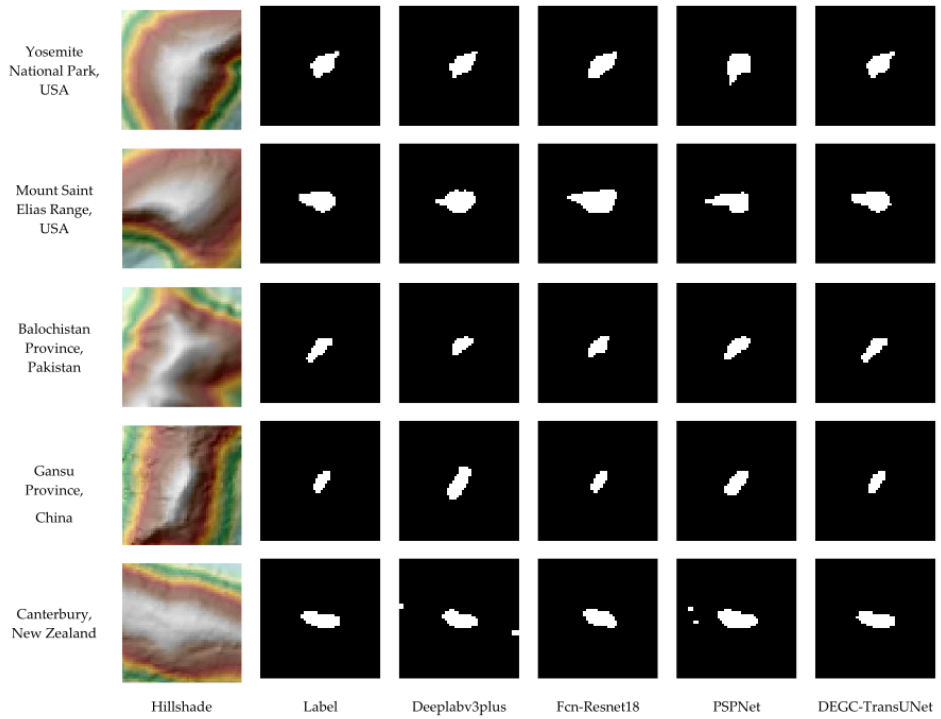


Figure 11. Generalization Performance of DEGC-TransUNet and Classic Semantic Segmentation Models on External Mountainous Regions Using 1 m Resolution DEM Data from Different Geographic Locations (the colors represent relative elevation values for terrain visualization and readability).

For regions outside the Loess Plateau, with more distinct geomorphic features, DEGC-TransUNet successfully captures integral mountaintop structures, despite minor boundary deviations in highly rugged and dissected terrains. Overall, the model's predictions align well with hillshade-derived morphology, indicating that it effectively captures intrinsic topographic signatures, avoiding overfitting to region-specific patterns [69]. In comparison, the classic models, while still functional, exhibit lower consistency and precision in these cross-regional tasks, particularly in more complex terrain. These results emphasize that the dual-encoder architecture, integrated with global context enhancement, provides robust generalization capabilities across varied regions and spatial resolutions, further supporting its applicability to large-scale, cross-regional terrain analysis.

5. Discussion

5.1. Effects of Dual-Encoder Design and Global Context Enhancement

The experimental results demonstrate that the proposed DEGC-TransUNet benefits significantly from the complementary strengths of its dual-encoder architecture and global context enhancement strategy. Compared to baseline models, such as TransUNet and DE-TransUNet, which primarily rely on local feature extraction or partial global context, DEGC-TransUNet effectively integrates both local and global information, leading to better boundary delineation and reduced misclassification, especially in complex terrain with irregular boundaries [42,43,60,61].

Meanwhile, the MaxViT-based encoder introduces multi-scale global contextual information at an early stage through efficient attention mechanisms. This capability enables the network to better distinguish true mountaintop areas from surrounding high-elevation terrain and reduce misclassification caused by purely local feature ambiguity. Similar observations have been reported in recent CNN–Transformer hybrid segmentation studies, where global context modeling significantly improves structural completeness and spatial consistency [28,42,69,70].

The incorporation of the BiFormer-based global context enhancement module further strengthens long-range dependency modeling. Compared with standard self-attention mechanisms, BiFormer achieves a favorable balance between global representation capability and computational efficiency through sparse attention routing [71]. This design is particularly advantageous for mountaintop area extraction, where elongated structures and large-scale terrain relationships must be captured. The superior performance of DEGC-TransUNet compared with its ablation variants confirms that global context enhancement and dual-encoder feature learning provide complementary benefits rather than redundant information.

5.2. Qualitative Interpretation, Limitations, and Future Perspectives

While baseline models, including PSPNet and DeepLabV3+, perform well in standard segmentation tasks, they show limitations when handling complex mountaintop boundaries, as observed in the comparison (Figure 10). DEGC-TransUNet consistently outperforms these models, particularly in the accurate delineation of boundaries in irregular and fragmented terrains. These improvements are consistent with prior studies indicating that combining local feature refinement with global contextual modeling leads to more robust segmentation performance [28–30,33,70].

Despite these advantages, several limitations remain. First, the ground-truth mountaintop areas are manually delineated, which may introduce subjectivity and uncertainty, especially in regions with gradual terrain transitions. Second, although Transformer-based components enhance representation capability, they also increase computational complexity compared with purely convolutional approaches, as reported in previous studies [33,42]. Although multi-country generalization experiments were conducted in this study, the number of external test samples remains limited. Broader validation across more geomorphological settings and climatic zones is still necessary to comprehensively evaluate large-scale operational applicability.

Future research will focus on extending the proposed framework to multi-resolution DEM datasets, exploring lightweight Transformer variants for improved efficiency, and integrating additional terrain derivatives or multi-source geospatial data. These directions are consistent with current trends in DEM-based deep-learning research and are expected to further enhance the applicability of the proposed approach [20,21,28].

6. Conclusions

This study proposed a dual-encoder TransUNet framework with global context enhancement (DEGC-TransUNet) for mountaintop area extraction from grid DEMs. By jointly modeling fine-scale local terrain details and multi-scale global contextual information, the proposed method effectively addresses the challenges of blurred boundaries and limited long-range dependency modeling commonly encountered in complex mountainous regions.

The experimental results demonstrate that the synergy between the convolutional encoder, the MaxViT-based global context modeling, and the BiFormer-based enhancement module yields significant performance gains. Specifically, DEGC-TransUNet consistently outperforms the baseline TransUNet and its ablation variants, achieving superior segmentation accuracy and boundary delineation.

In comparison with other classic semantic segmentation models, including PSPNet and DeepLabV3+, DEGC-TransUNet consistently shows superior performance in terms of boundary delineation and Intersection over Union (IoU), further confirming the advantage of integrating both local and global feature learning. This improvement underscores the effectiveness of the dual-encoder architecture combined with global context enhancement, which provides more robust generalization across diverse geomorphological regions.

The high-precision extraction capability of DEGC-TransUNet provides a robust scientific foundation for diverse fields. It can be directly integrated into geomorphological evolution analysis, soil erosion simulation, and biodiversity modeling. Furthermore, the framework offers a reliable decision-support tool for engineering tasks, such as the strategic site selection of communication base stations and wind energy infrastructures in high-altitude regions.

Despite the promising results, this study has several limitations. The experiments focused on grid DEMs within specific geomorphological settings and at a fixed resolution. In addition, the increased model complexity introduces higher computational costs. Future work will focus on extending the framework to multi-resolution datasets, exploring lightweight variants, and conducting broader cross-regional validation to further enhance its operational applicability for large-scale terrain analysis.

Future work will focus on extending the proposed framework to multi-resolution DEM datasets, exploring lightweight model variants for improved computational efficiency, and evaluating cross-regional generalization performance. Furthermore, integrating additional geomorphological constraints and extending the method to other terrain element extraction tasks represent promising directions for further research.

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

Funding: This research was funded by the National Natural Science Foundation of China (grant number 41671446), the Changsha Municipal Natural Science Foundation (grant number KQ2402013), and the 2024 Changsha University of Science and Technology Academic Degree Graduate Student Research Innovation Project (grant number CLKYCX24112).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The high-resolution Digital Elevation Model (DEM) data used in this study are openly available from OpenTopography at https://opentopography.org (accessed on 20 November 2025) and the National Geomatics Center of China at http://www.ngcc.cn (accessed on 15 September 2025). The experimental results and the code for the DEGC-TransUNet framework are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

DEM	Digital Elevation Model
CNN	Convolutional Neural Network
DEGC-TransUNet	Dual-Encoder and Global-Context-enhanced TransUNet
MaxViT	Multi-Axis Vision Transformer
Conv Encoder	Convolutional Encoder
SaE	Squeeze-and-Excitation
FFM	Feature Fusion Module
CA	Coordinate Attention
FFC	Fast Fourier Convolution
BRA	Bi-level Routing Attention
LN	Layer Normalization
BCE	Binary Cross-Entropy
IoU	Intersection over Union
ACC	Overall Accuracy

References

1. Riazanoff, S.; Corseaux, A.; Albinet, C.; Strobl, P.A.; López-Vázquez, C.; Guth, P.L.; Tadono, T. Best BiCubic Method to Compute the Planimetric Misregistration between Images with Sub-Pixel Accuracy: Application to Digital Elevation Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 96. [CrossRef]
2. Dávila-Hernández, S.; González-Trinidad, J.; Júnez-Ferreira, H.E.; Bautista-Capetillo, C.F.; Morales de Ávila, H.; Cázares Escareño, J.; Ortiz-Letechipia, J.; Robles Rovelo, C.O.; López-Baltazar, E.A. Effects of the Digital Elevation Model and Hydrological Processing Algorithms on the Geomorphological Parameterization. *Water* **2022**, *14*, 2363. [CrossRef]
3. Yan, G.; Tang, G.; Lu, D.; Ma, J.; Yang, X.; Li, F. Distinguishing the intervalley plain from the intermountain flat for landform mapping using the sightline algorithm. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 86. [CrossRef]
4. Zhu, Y.; Liu, X.; Zhao, J.; Cao, J.; Wang, X.; Li, D. Effect of DEM Interpolation Neighbourhood on Terrain Factors. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 30. [CrossRef]
5. Podobnikar, T. Detecting Mountain Peaks and Delineating Their Shapes Using Digital Elevation Models, Remote Sensing and Geographic Information Systems Using Autometric Methodological Procedures. *Remote Sens.* **2012**, *4*, 784–809. [CrossRef]
6. Yuan, S. *Evolution Process and Law of Loess Landform in Eastern Gansu, China*; Wuhan University of Technology: Wuhan, China, 2021.
7. Shirani, K.; Solhi, S.; Pasandi, M. Automatic Landform Recognition, Extraction, and Classification using Kernel Pattern Modeling. *J. Geovisualization Spat. Anal.* **2023**, *7*, 2. [CrossRef]
8. Silván-Cárdenas, J.L.; Wang, L.; Zhan, F.B. Representing geographical objects with scale-induced indeterminate boundaries: A neural network-based data model. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 295–318. [CrossRef]
9. Mokarram, M.; Sathyamoorthy, D. A review of landform classification methods. *Spat. Inf. Res.* **2018**, *26*, 647–660. [CrossRef]
10. Chen, C.; Liu, Y.; Li, Y.; Chen, D. Explainable artificial intelligence framework for urban global digital elevation model correction based on the SHapley additive explanation-random forest algorithm considering spatial heterogeneity and factor optimization. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103843. [CrossRef]
11. Zhou, F.; Liu, X. Improvement of hill-position classification decision scheme based on DTA and research on automatic micro-topographic classification. *Acta Agric. Boreali-Occident. Sin.* **2008**, *17*, 343–346.
12. Zhong, T.; Tang, G.; Zhou, Y. Automatic extraction of mountain peaks based on inverse DEM. *Bull. Surv. Mapp.* **2009**, *4*, 35–37.
13. Jasiewicz, J.; Stepinski, T.F. Geomorphons—A pattern recognition approach to classification and mapping of landforms. *Geomorphology* **2013**, *182*, 147–156. [CrossRef]
14. Chen, P.; Lu, P.; Liu, W. Automatic extraction of mountain peaks based on independent closed contour surfaces. *Geospat. Inf.* **2020**, *18*, 48–50. [CrossRef]

15. Xiao, Z. *Research on Extraction of Mountain Peak Range Based on Grid DEM*; Changsha University of Science & Technology: Changsha, China, 2022.
16. Jiang, L.; Ling, D.; Zhao, M.; Wang, C.; Liang, Q.; Liu, K. Effective Identification of Terrain Positions from Gridded DEM Data Using Multimodal Classification Integration. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 443. [CrossRef]
17. Zhou, F.; Zou, L.; Zhang, X. Micro-landform classification method of grid DEM based on BP neural network. *Bull. Surv. Mapp.* **2019**, *10*, 101–104.
18. Li, H.; Shi, Y.; Shen, A.; Ma, Y.; Liang, Y.; Li, X.; Dong, J.; Zhao, N.; She, J.; Wang, T. Research on the classification of micro-terrain in the alluvial fan of Helan Mountains east foot based on Random Forest algorithm. *Quat. Sci.* **2025**, *45*, 178–190.
19. 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]
20. Zhang, J. *Saddle and Peak Elements Recognition Technology Based on CNN*; Xi'an University of Architecture and Technology: Xi'an, China 2019.
21. Zhen, J.; Qi, J.; Wang, Y. Rapid extraction of summit points for emergency communication siting based on enhanced CNN. *Geomat. Inf. Sci. Wuhan. Univ.* **2025**, *50*, 1042–1053.
22. Zhou, F.; Zou, L.; Liu, X. Micro-landform classification of grid DEM based on convolutional neural network. *Geomat. Inf. Sci. Wuhan. Univ.* **2021**, *46*, 1186–1193.
23. Yang, J.; Xu, J.; Lv, Y.; Zhou, C.; Zhu, Y.; Cheng, W. Deep learning-based automated terrain classification using high-resolution DEM data. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103249. [CrossRef]
24. Wu, Y.; Sun, X.; Qi, L.; Xu, J.; Gao, D.; Zhu, Z. Integrating DEM and Deep Learning for Forested Terrain Analysis: Enhancing Fire Risk Assessment Through Mountain Peak and Water System Extraction in Chongli District. *Forests* **2025**, *16*, 692. [CrossRef]
25. Li, W.; Hsu, C.Y.; Wang, S.; Yang, Y.; Lee, H.; Liljedahl, A.; Witharana, C.; Rogers, B.M.; Arundel, S.T. Segment Anything Model Can Not Segment Anything: Assessing AI Foundation Model's Generalizability in Permafrost Mapping. *Remote Sens.* **2024**, *16*, 797. [CrossRef]
26. Huang, L.; Jiang, B.; Lv, S.; Liu, Y.; Fu, Y. Deep-learning-based semantic segmentation of remote sensing images: A survey. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8370–8396. [CrossRef]
27. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4.
28. Tu, Z.; Talebi, H.; Zhang, H. MaxViT: Multi-axis vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 459–479.
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 13713–13722.
30. Narayanan, M. SENetV2: Aggregated dense layer for channelwise and global representations. *arXiv* **2023**, arXiv:2311.10807. [CrossRef]
31. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
32. Cao, H. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2023.
33. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
34. Huang, P.; Liu, Z. Study of extraction of ridge and valley from DEM data. *Bull. Surv. Mapp.* **2005**, *4*, 11–13.
35. Zhou, F.; Xiao, Z.; Liu, X.; Ma, G.; Zheng, S. Mountain Peak Extraction of Grid DEM Based on Aspect Distribution Feature. *Geomat. Inf. Sci. Wuhan Univ.* **2024**, *49*, 419–425. [CrossRef]
36. Gu, L.; Wang, C.; Li, P.; Wang, J.; Wang, Z. Research on mountain top extraction accuracy based on DEM. *Geomat. Inf. Sci. Wuhan Univ.* **2016**, *41*, 131–135.
37. Cao, P.; Cheng, S.; Lin, H.; Wang, X.; Li, M.; Chen, J. DEM in quantitative analysis of structural geomorphology: Application and prospect. *J. Geomech.* **2021**, *27*, 949–962. (In Chinese)
38. Zou, L. Micro landform Classification Method of Grid Dem Based on Artificial Intelligence. Master's Thesis, Changsha University of Science & Technology, Changsha, China, 2021.
39. Zhang, X.; Guo, Y.; Zhou, F. Micro landform classification method of grid DEM based on random forest. *Geomat. Spat. Inf. Technol.* **2019**, *42*, 107–109.
40. Du, P.; Chen, N.; Wu, K.; Li, Z.; Zhang, Y. Evaluation of landslide susceptibility in southeast Tibet based on a randomforest model. *J. Chengdu Univ. Technol. (Sci. Technol. Ed.)* **2024**, *51*, 328–344.
41. 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), Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

42. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2015; pp. 3431–3440.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778.
44. Li, K.; Kong, Y.; Zhang, Y.; Gao, K. Research on mountain peak recognition based on improved convolutional neural network. *Comput. Meas. Control.* **2021**, 154–158, 164. [CrossRef]
45. Lv, J.; Shen, Q.; Lv, M. Deep learning-based semantic segmentation of remote sensing images: A review. *Front. Ecol. Evol.* **2023**, *11*, 1201125. [CrossRef]
46. Zhu, L.; Wang, X.; Ke, Z. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 10323–10333.
47. Zhang, P.; Du, P.; Lin, C.; Xin, W.; Erzhu, L.; Zhaohui, X.; Xuyu, B. A Hybrid Attention-Aware Fusion Network (HAFNet) for Building Extraction from High-Resolution Imagery and LiDAR Data. *Remote Sens.* **2020**, *12*, 3764. [CrossRef]
48. Wang, X.; Zhang, W.; Chen, T.; Zhang, A.; Zhao, D. Seismic fault identification method based on CBAM-TransUNet. *J. China Coal Soc.* **2024**, *50*, 1192–1202.
49. Wang, H.; Xiong, L.; Cao, H.; Zhao, F.; Yu, F.; Wei, H.; Tang, G.; Lane, S.N. A transformer-based landform mapping method combining topographic knowledge and digital elevation data. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *145*, 104974. [CrossRef]
50. Yang, Z.; Han, L.; Zheng, X.; Li, W.; Feng, L.; Wang, Y.; Yang, Y. Landslide identification using remote sensing images and DEM based on convolutional neural network: A case study of loess landslide. *Remote Sens. Nat. Resour.* **2022**, *34*, 224–230. [CrossRef]
51. Suvorov, R.; Logacheva, E.; Mashikhin, A. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2022; p. 2149.
52. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*; IEEE: New York, NY, USA, 2016; pp. 79–87.
53. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An object-based seamless semantic segmentation strategy for large-scale remote sensing imagery with deep learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 11–26. [CrossRef]
54. National Geomatics Center of China. National Fundamental Geographic Information Database of China; Beijing, China, 2021. Available online: http://www.ngcc.cn (accessed on 30 September 2025).
55. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
56. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
57. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579. [CrossRef]
58. Everingham, M.; Van Gool, L.; Williams, C.K.I. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
59. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
60. Lin, T.-Y.; Dollár, P.; Girshick, R.; Ghiasi, G.; Hariharan, B.; He, K. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.
61. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 2881–2890.
62. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 801–818.
63. Meyes, R.; Lu, M.; De Puiseau, C.W. Ablation studies in artificial neural networks. *arXiv* **2019**, arXiv:1901.08644. [CrossRef]
64. McGarigal, K. *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*; US Department of Agriculture, Forest Service, Pacific Northwest Research Station: Portland, OR, USA, 1995.
65. Zhou, Y.; Tang, G.; Yang, X. Positive and negative terrains on northern Shaanxi Loess Plateau. *J. Geogr. Sci.* **2010**, *20*, 64–76. [CrossRef]
66. OpenTopography. High-Resolution Topographic Data and Tools Provided by OpenTopography. San Diego, CA, USA. 2026. Available online: https://opentopography.org (accessed on 30 September 2025).
67. Zhang, Q.; Wang, T. Deep learning for exploring landslides with remote sensing and geo-environmental data: Frameworks, progress, challenges, and opportunities. *Remote Sens.* **2024**, *16*, 1344. [CrossRef]
68. Cao, J.; Tang, G.; Fang, X. Terrain relief periods of loess landforms based on terrain profiles of the Loess Plateau in northern Shaanxi Province, China. *Front. Earth Sci.* **2019**, *13*, 410–421. [CrossRef]
69. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N. UNet++: A nested U-Net architecture for medical image segmentation. *arXiv* **2018**, arXiv:1807.10165.

70. Huang, H. UNet 3+: A full-scale connected UNet for medical image segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2020; pp. 1055–1059.
71. Sun, G.; Pan, Y.; Kong, W. DA-TransUNet: Integrating spatial and channel dual attention with Transformer U-Net. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1398237. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.