

A Hybrid CNN–Transformer Model for Terrace Extraction from Remote Sensing Imagery

Fengying Jin , Zhihui Wang , Yinan Wang , Jintao Zhao, Chunjing Zhao, Qingfeng Xu and Kai Guo * 

Key Laboratory of Soil and Water Conservation on the Loess Plateau, Yellow River Institute of Hydraulic Research, Zhengzhou 450003, China; jinfengying@whu.edu.cn (F.J.); wzh8588@aliyun.com (Z.W.); wongyinan@foxmail.com (Y.W.); zjt_try@163.com (J.Z.); zhaocjhky@163.com (C.Z.); xqflol@163.com (Q.X.)

* Correspondence: guokai@hhu.edu.cn

Highlights

What are the main findings?

- As a hybrid CNN–Transformer architecture with an encoder–dual-decoder design, the proposed TerraceNet achieves accurate terrace extraction in the Wuding River Basin, attaining an IoU of 77.61% and an F1-Score of 87.39% on a 2 m resolution GF-1 satellite imagery dataset.
- A boundary–segmentation joint optimization mechanism is introduced, in which automatically generated boundary supervision signals are incorporated to enhance the model’s perception of terrace geometric boundaries.

What are the implications of the main findings?

- The hybrid architecture demonstrates that integrating CNN-based local detail preservation with Transformer-based global context modeling provides an effective solution for extracting morphologically diverse terraces from high-resolution remote sensing imagery.
- This study offers a practical technical pathway for precision agricultural remote sensing monitoring under complex terrain conditions, with potential applicability to terrace resource surveys, soil and water conservation monitoring, and other scale-variant agricultural feature extraction tasks.

Abstract

Terraces constitute a critical form of land surface modification, serving both as essential agricultural resources and effective soil and water conservation measures. However, their morphological diversity, pronounced scale variation, and ambiguous geometric boundaries pose substantial challenges for automated extraction from high-resolution remote sensing imagery. Existing CNN- and Transformer-based methods still face difficulties in simultaneously preserving fine spatial details and modeling long-range contextual information. To address these limitations, this study proposes TerraceNet, a hybrid CNN–Transformer architecture with an encoder–dual-decoder design. The framework employs a ConvNeXt encoder to extract multi-scale features, which are subsequently aggregated by a dynamic fusion feature pyramid network (DF-FPN) and fed into two parallel decoders: a boundary decoder dedicated to fine-grained edge localization and a multi-scale Transformer decoder that incorporates boundary priors for global context modeling and final segmentation. Experimental results on a 2 m resolution GF-1 satellite imagery dataset from the Wuding River Basin in the Loess Plateau demonstrate that TerraceNet an IoU of 77.61% and an F1-Score of 87.39%. These results validate the effectiveness of the proposed architecture for extracting morphologically diverse terraces in complex terrain.



Academic Editors: Qi Wang, Shiqing Wei, Jin Liu and Muhammad Yasir

Received: 21 May 2026

Revised: 9 June 2026

Accepted: 15 June 2026

Published: 9 July 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: terraces; remote sensing; CNN–Transformer hybrid; terrace extraction

1. Introduction

Terraces, as an important form of land surface modification, are extensively distributed in mountainous and hilly regions worldwide [1,2]. They not only serve as vital agricultural resources for food production but also function as effective engineering measures for soil and water conservation, playing critical roles in reducing surface runoff, preventing soil erosion, improving slope stability, and maintaining agro-ecosystem sustainability [3,4]. Many countries and regions, including China (e.g., the Loess Plateau and Yunnan–Guizhou Plateau), Southeast Asian nations, and Mediterranean countries, possess large-scale terrace systems with long cultivation histories [5–7]. Accurate mapping and dynamic monitoring of the spatial distribution of terraces are therefore highly significant for agricultural management, ecological assessment, soil erosion control, and land degradation prevention [8,9].

However, terraces exhibit pronounced morphological diversity (Figure 1), ranging from narrow, elongated strip-type terraces to broad, contiguous terrace patches. This morphological complexity, together with the ambiguity of geometric boundaries, poses substantial challenges for automated extraction [10]. Firstly, substantial scale discrepancies across terrace patches require models to capture both fine-grained local details and global semantic context; conventional segmentation models often struggle to represent these cross-scale characteristics simultaneously. Secondly, imaging conditions, spatial resolution, shadowing, seasonal differences, and adjacency to other land cover types can blur terrace boundaries and cause feature adhesion, making accurate boundary delineation particularly difficult in complex terrain.

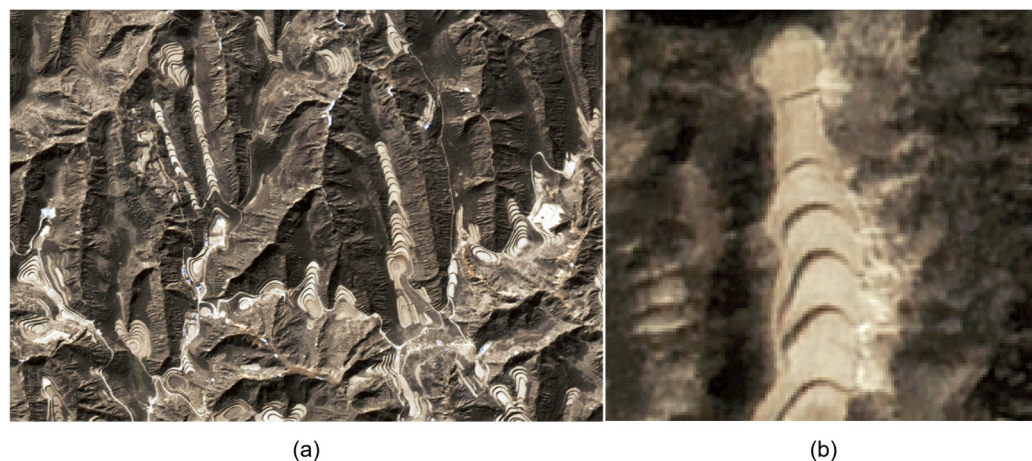


Figure 1. Representative examples of terraces. (a) Morphological diversity across terrace types, ranging from narrow strip-type to broad patch-type and small fragmented patches. (b) A representative case of boundary ambiguity and feature adhesion.

In terms of terrace extraction methods, existing approaches can be broadly categorized into machine learning- and deep learning-based methods. Machine learning-based methods, such as random forest [11], XGBoost [12], support vector machine [13], and multi-layer perceptron [14], generally rely on handcrafted spectral, textural, geometric, and topographic features. These methods offer relatively strong interpretability and stable performance in small-sample or homogeneous regions [15]. For instance, Cao et al. [16] produced the first 30 m resolution terrace map of China using random forest and the Google Earth Engine platform. Ding et al. [15] evaluated three machine learning methods for object-based artificial terrace mapping and found that random forest achieved the

best performance among the tested models. Nevertheless, handcrafted features are often insufficient for describing the complex spatial structures and diverse morphologies of terraces, which limits the generalization ability of traditional machine learning methods across regions and imaging conditions.

In recent years, deep learning-based methods have gradually become the mainstream paradigm for terrace extraction. Convolutional neural network (CNN)-based methods [17], exemplified by UNet [18], DeepLabV3 [19], and their variants, can automatically learn multi-scale spatial representations through end-to-end learning. Lu et al. [5] developed a deep learning-based terrace extraction model using UNet++ and produced a 1.89 m resolution terrace distribution map for the Loess Plateau. Yu et al. [20] proposed an automatic terrace identification method for this area based on deep transfer learning. Zhao et al. [21] proposed NLDF-Net, which combines an asymmetric non-local block with linear attention distribution and dual fusion modules, achieving improved extraction performance in Guangdong Province. Despite these advances, CNN-based methods remain constrained by the local receptive field of convolutional kernels; as such, they may have limited capacity for modeling long-range dependencies and maintaining semantic consistency across large terrace patches.

To address the limitations of CNNs in capturing global dependencies, Transformer-based methods (e.g., SegFormer [22], Mask2Former [23], and Swin-Unet [24]) have been introduced into remote sensing segmentation tasks. These methods effectively capture global feature dependencies through self-attention mechanisms [25]. For instance, Kan et al. [26] proposed a network that integrates a Swin Transformer architecture with a dual-branch attention mechanism and boundary-aware auxiliary supervision, improving terrace boundary delineation and multi-source feature fusion under complex terrain conditions. Li et al. [27] developed a dual-modal Ω -like super-resolution Transformer network (Ω SFormer) for cross-scale, high-precision vectorized terrace extraction, providing a novel technical pathway from pixel-level segmentation to vector output. However, pure Transformer methods may exhibit deficiencies in preserving local detail features, especially for narrow, fragmented terraces with indistinct boundaries.

To overcome these issues, this study proposes a hybrid architecture, TerraceNet, which integrates CNNs' local detail modeling capability with Transformers' global context awareness for the high-precision extraction of morphologically diverse terraces from remote sensing imagery. The network adopts an encoder–dual-decoder architecture. Specifically, the ConvNeXt encoder [28] extracts multi-scale features, which are then fused by a dynamic fusion feature pyramid network (DF-FPN) [29]. The fused features are subsequently fed into two collaborative decoders: the boundary decoder focuses on fine-grained localization of terrace boundaries and generates structural priors, while the Transformer segmentation decoder incorporates boundary features and models long-range contextual dependencies through self-attention mechanisms to enhance the semantic consistency of terrace regions. The two decoders are jointly optimized by a combined segmentation and boundary supervision loss.

The main contributions of this study are threefold:

- (1) A CNN–Transformer hybrid extraction framework is proposed. Unlike pure CNN methods, which are limited in long-range dependency modeling, and pure Transformer methods, which may lose fine local details, TerraceNet combines the complementary advantages of both architectures to jointly capture the fine-grained local details and global semantic context of terraces.
- (2) A boundary–segmentation joint optimization mechanism is designed. Boundary labels are automatically derived from segmentation masks and introduced as auxil-

ary supervision, enabling the model to explicitly enhance its perception of terrace geometric boundaries.

- (3) Improved extraction accuracy is achieved. Experimental results on the Wuding River Basin terrace extraction dataset show that the proposed method outperforms several representative semantic segmentation models, demonstrating the effectiveness and advantages of the proposed architecture.

The remainder of this study is structured as follows. Section 2 presents the methodology, Section 3 reports the experimental results, Section 4 provides a discussion, and Section 5 concludes the study.

2. Methods

2.1. Overall Model Architecture

This study proposes a hybrid CNN–Transformer architecture, TerraceNet, for automated terrace extraction from high-resolution remote sensing imagery. As illustrated in Figure 2a, the framework consists of four sequential and collaborative modules: (1) a ConvNeXt-based encoder for multi-scale feature extraction; (2) a DF-FPN module for adaptive multi-level feature fusion; (3) a boundary decoder for boundary detection and structural prior generation; (4) a multi-scale Transformer decoder for boundary-guided semantic segmentation. Through this design, TerraceNet aims to jointly preserve local terrace details, enhance boundary perception, and model global contextual relationships.

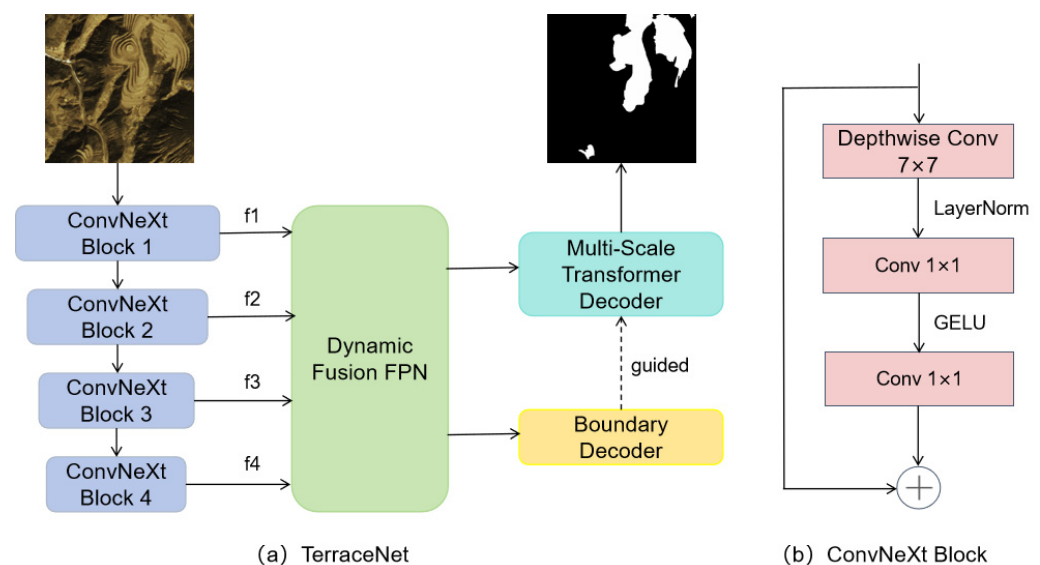


Figure 2. Framework of the proposed model. (a) Overall framework of TerraceNet; (b) structure of the ConvNeXt block.

2.2. ConvNeXt Encoder

The encoder employs the ConvNeXt-Tiny architecture as the backbone network (Figure 2b). ConvNeXt is a modernized convolutional neural network that systematically integrates design principles inspired by vision transformers while retaining the inductive bias and computational efficiency of convolutional architectures. Compared with conventional backbones such as ResNet or VGG, ConvNeXt provides stronger hierarchical feature representation with relatively moderate computational overhead, making it suitable for high-resolution remote sensing applications where both local detail preservation and efficient processing are required.

The encoder is divided into four hierarchical stages, producing a set of multi-scale deep features:

$$\{f_1, f_2, f_3, f_4\} = \text{ConvNeXt}(I) \quad (1)$$

where I denotes the input image, and f_1, f_2, f_3, f_4 denote the output feature maps with channel dimensions of 96, 192, 384, and 768, respectively. These feature maps correspond to spatial down-sampling factors of 4, 8, 16, and 32. The shallow-level features retain the fine-grained textural details and structural information of terraces, whereas the deep-level features encode global semantic context and long-range dependencies.

2.3. Dynamic Fusion Feature Pyramid Network

To effectively integrate multi-scale features and unify channel dimensions, a DF-FPN is introduced after the encoder. The proposed DF-FPN architecture consists of two primary components: lateral connections and a top-down pathway. In contrast to conventional FPNs that rely on fixed addition or concatenation operations, this module incorporates a dynamic fusion gate mechanism that learns adaptive fusion weights for lateral features and up-sampled high-level features. This design allows the network to flexibly balance low-level spatial details and high-level semantic information for terraces with different scales and shapes.

Lateral connections. Each encoder output feature map f_i is processed by a 1×1 convolutional layer to project its channel dimension to 256:

$$c_i = \text{Conv}_{1 \times 1}(f_i), i = 1, 2, 3, 4 \quad (2)$$

Top-down pathway. Starting from the highest-level feature c_4 , the top-down pathway progressively up-samples and fuses features with corresponding lateral connections. To achieve adaptive feature fusion, a dynamic fusion module is designed. Unlike a conventional FPN that employs fixed summation, the dynamic fusion module learns spatial gating weights to balance the contributions of up-sampled high-level semantic features and lateral low-level detail features:

$$p_i = \text{DynamicFusion}(c_i, \text{Upsample}(p_{i+1})) \quad (3)$$

The gating weights are generated by a lightweight subnetwork consisting of two 1×1 convolutional layers followed by a softmax activation function. This learnable mechanism enables the network to adaptively emphasize either semantic information from deeper layers or fine-grained details from shallower layers according to the image content, which is particularly beneficial for handling terraces with diverse spatial morphologies.

Following top-down fusion, each output feature map is refined by a 3×3 convolutional layer with batch normalization and a ReLU activation function, yielding the final multi-scale representations $\{p_1, p_2, p_3, p_4\}$, each with 256 channels. These unified feature maps provide a consistent feature basis for the subsequent boundary and segmentation decoders.

2.4. Boundary Decoder

The boundary decoder is designed to extract geometric terrace boundary information. It serves two primary objectives: (1) producing boundary prediction maps for supervised boundary learning and (2) generating structural priors that guide the segmentation decoder toward more boundary-consistent predictions. This auxiliary branch is particularly important for terrace extraction because terrace boundaries are easily confused with adjacent land cover features.

The boundary ground truth used for supervising the boundary decoder was automatically derived from the binary terrace segmentation masks, rather than manually annotated independently. More specifically, the inner-boundary labels were generated by subtracting the eroded mask from the original binary mask through a morphological erosion operation. The generated boundary maps were then used as ground-truth labels for boundary supervision.

The boundary decoder operates on the first three DF-FPN outputs $\{p_1, p_2, p_3\}$, which exhibit progressively coarser spatial resolutions of $H/4$, $H/8$, and $H/16$, respectively. These feature levels are selected because they preserve sufficient spatial detail while still containing useful semantic information.

As illustrated in Figure 3, the boundary decoder employs a multi-scale feature fusion strategy. Each input feature map is first projected to a compact 64-channel representation through a 3×3 convolutional layer followed by batch normalization and a ReLU activation function (ConvBNReLU block):

$$\hat{p}_i = \text{ConvBNReLU}(p_i), \quad i = 1, 2, 3 \quad (4)$$

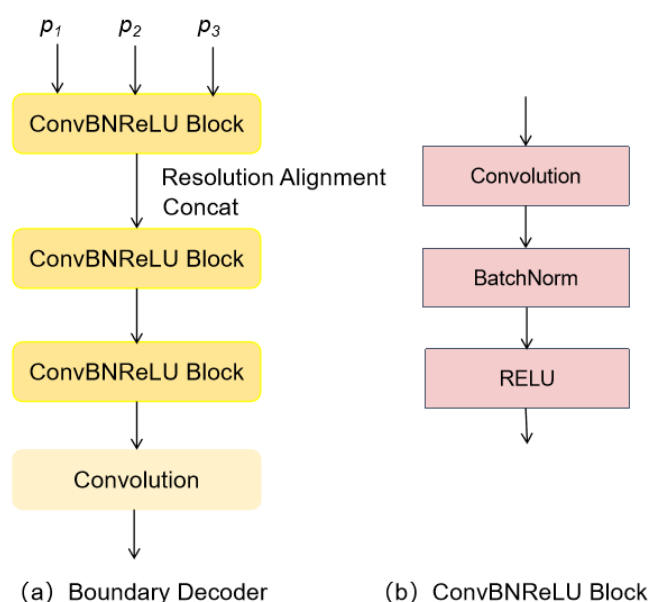


Figure 3. Architecture of the boundary decoder. (a) Overall architecture of the boundary decoder. (b) Architecture of the ConvBNReLU block.

To enable element-wise fusion, using bilinear interpolation, the lower-resolution features $\hat{p}_2$ and $\hat{p}_3$ are up-sampled to match the spatial resolution of $\hat{p}_1$ (i.e., $H/4 \times W/4$). The three spatially aligned feature maps are subsequently concatenated along the channel dimension and fed into a fusion module consisting of two consecutive ConvBNReLU blocks with 64 channels. This design enables the decoder to aggregate multi-scale information while maintaining computational efficiency.

The fused feature $F_{edge} \in \mathbb{R}^{64 \times H/4 \times W/4}$ serves two purposes. Firstly, it is processed by a 1×1 convolutional layer to generate the boundary logits at a $1/4$ resolution, which are subsequently up-sampled by a factor of 4 via bilinear interpolation to match the original input resolution ($H \times W$), yielding the final boundary prediction map E_{logits} . Secondly, the fused feature itself is used as a structural prior F_{edge} and provided to the segmentation decoder for boundary-aware modulation.

2.5. Multi-Scale Transformer Segmentation Decoder

The multi-scale Transformer segmentation decoder is the core component of the TerraceNet framework (Figure 4). It takes the multi-scale features produced by the ConvNeXt encoder and DF-FPN as input and further models global contextual dependencies through a Transformer architecture. The decoder comprises five modules: Transformer stage, feature fusion block, boundary guidance, feature refinement, and a segmentation head.

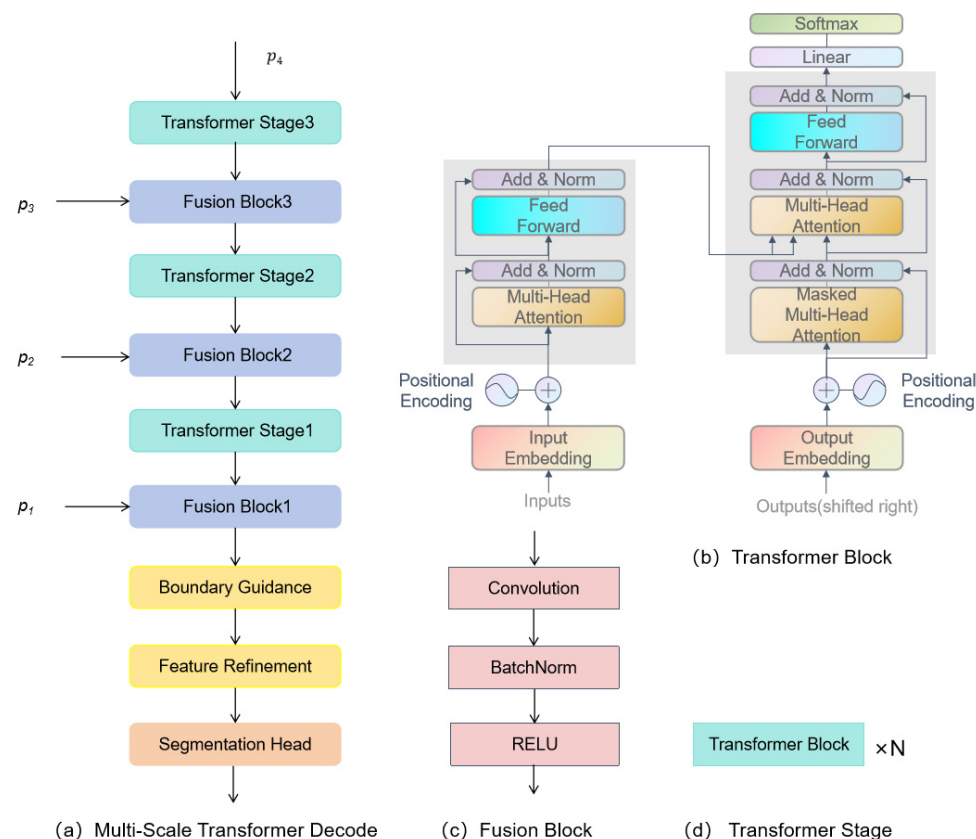


Figure 4. Architecture of the multi-scale Transformer decoder. (a) Overall architecture of the multi-scale Transformer decoder and (b) architecture of the Transformer block, (c) fusion block, and (d) Transformer stage.

2.5.1. Transformer Stage

The four DF-FPN outputs are first projected to a unified embedding dimension of 256 via 1×1 convolutions:

$$\tilde{p}_i = \text{Conv}_{1 \times 1}(\hat{p}_i), i = 1, 2, 3, 4 \quad (5)$$

Subsequently, according to the feature hierarchy, each projected feature map is either directly fed into a Transformer stage or fused with features from adjacent levels before being forwarded to the Transformer stage. Each Transformer stage consists of stacked Transformer blocks, and each block sequentially applies a multi-head self-attention (MHSA) mechanism and a feed-forward network (FFN). This structure enables the decoder to capture long-range semantic relationships within elongated strip-like terraces and broad contiguous terrace patches while maintaining multi-scale feature representation. The computation within a single Transformer block can be formally expressed as follows:

$$q = \text{LN}(\tilde{p}) + \text{MHSA}(\text{LN}(\tilde{p})) \quad (6)$$

$$q' = q + \text{FFN}(\text{LN}(q)) \quad (7)$$

where LN is layer normalization.

To balance computational efficiency with global modeling capacity, a lightweight configuration is adopted, where the number of blocks per stage is set to (2, 2, 2). The hidden dimension of the FFN is set to four times the embedding dimension (i.e., 1024), and the number of attention heads is set to 8. This configuration provides sufficient contextual modeling capacity while avoiding excessive computational cost.

2.5.2. Feature Fusion Block

A top-down fusion strategy is employed to progressively integrate high-level semantic information with low-level spatial details:

$$x_4 = q'_4 \quad (8)$$

$$x_i = \text{FusionBlock}(\text{Upsample}(x_{i+1}), q'_i), \quad i = 3, 2, 1 \quad (9)$$

Each fusion block performs channel-wise concatenation, followed by a 3×3 convolutional layer with batch normalization and ReLU activation. This design preserves the semantic abstraction derived from high-level features and the spatial precision inherent to low-level features, enabling the decoder to gradually recover fine terrace structures during top-down reconstruction.

2.5.3. Boundary Guidance

To effectively integrate structural information derived from the boundary decoder, this study proposes a boundary guidance module that employs a residual learning mechanism to fuse decoder features with auxiliary boundary information. More specifically, boundary features are first extracted from the boundary decoder and projected into a unified embedding space using a 1×1 convolution. The projected boundary features are then concatenated with the decoder features and jointly transformed through a 3×3 convolutional layer. Finally, the transformed features are added element-wise to the original decoder features through a residual connection. This design introduces boundary structural information while preserving the decoder features' primary semantic content, thereby enhancing the model's capability to perceive terrace boundaries.

2.5.4. Refinement and Segmentation Head

The boundary-enhanced feature map is further refined through a depth-wise separable convolutional module. This module sequentially applies a 3×3 depth-wise followed by a 1×1 pointwise convolution and incorporates a residual connection. This refinement strategy enhances local feature consistency and representational capacity while maintaining relatively low computational complexity. Subsequently, the segmentation head consists of a 3×3 convolution that reduces the channel dimension from 256 to 128, followed by batch normalization, ReLU activation, and a 1×1 convolution layer to produce class logits. The final output is up-sampled by a factor of 4 using bilinear interpolation, restoring the prediction map from one-quarter resolution to the original input resolution.

2.6. Loss Function

The model is optimized using a joint loss function that provides simultaneous supervision for both segmentation and boundary branches. The total loss is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \mathcal{L}_{seg_dice} + \lambda_{edge} \cdot \mathcal{L}_{edge} \quad (10)$$

where λ_{edge} is a hyper-parameter used to balance the contribution of the boundary task, and it is set to 0.3 in this study.

2.6.1. BCE Loss

Binary cross-entropy (BCE) loss is employed as the primary pixel-wise classification objective. Given the predicted logits $S_{pred,i}$ and the ground-truth label $S_{gt,i} \in \{0, 1\}$ for the i -th pixel, the BCE loss is defined as follows:

$$\mathcal{L}_{bce} = \frac{1}{N} \sum_{i=1}^N [S_{gt,i} \log(\sigma(S_{pred,i})) + (1 - S_{gt,i}) \log(1 - \sigma(S_{pred,i}))] \quad (11)$$

where σ denotes the sigmoid function and N is the total pixel number.

2.6.2. Dice Loss

Dice loss is incorporated to further alleviate class imbalance between terrace pixels and the background:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum(S_{pred} \cap S_{gt}) + \epsilon}{\sum S_{pred} + \sum S_{gt} + \epsilon} \quad (12)$$

where S_{pred} and S_{gt} denote sigmoid-normalized predictions and ground truth, respectively, and ϵ is a smoothing term to prevent division by zero.

2.6.3. Boundary Loss

For the highly imbalanced boundary detection task, a weighted BCE loss is adopted, where the positive weight is automatically calculated based on the ratio of non-boundary to boundary pixels:

$$w_{pos} = \frac{N_{neg}}{N_{pos} + \epsilon} \quad (13)$$

where N_{pos} and N_{neg} denote the number of boundary and non-boundary pixels, respectively, and ϵ is a small constant. The positive weight w_{pos} is clipped to the range $[1, 10]$ to avoid excessively large weights. The boundary loss is then computed as follows:

$$\mathcal{L}_{edge} = -\frac{1}{N} \sum_{i=1}^N [E_{gt,i} \cdot \log(\sigma(E_{pred,i})) \cdot w_{pos} + (1 - E_{gt,i}) \cdot \log(1 - \sigma(E_{pred,i}))] \quad (14)$$

where $E_{gt,i} \in \{0, 1\}$ is the ground-truth boundary label of the i -th pixel, and $E_{pred,i}$ is the corresponding predicted boundary logits output by the model.

3. Experiments

3.1. Datasets

The dataset used in this study was constructed from GF-1 satellite imagery (China Centre for Resources Satellite Data and Application, Beijing, China) covering the Wuding River Basin in the Loess Plateau, China. The imagery was acquired in May 2024, with a spatial resolution of 2 m. The Wuding River Basin contains representative terrace landscapes with diverse morphologies, including narrow strip-like terraces, fragmented terrace patches, and broad contiguous terraces, making it suitable for evaluating terrace extraction methods under complex terrain conditions.

The ground-truth labels were produced through a manual visual interpretation workflow. More specifically, terrace regions were first visually interpreted from the GF-1 imagery and manually delineated as vector polygons using QGIS software. The vectorized terrace polygons were then checked and refined by domain experts to ensure annotation accuracy and consistency. After expert review, the verified terrace polygons were rasterized into binary segmentation masks, where terrace and non-terrace background pixels were assigned values of 1 and 0, respectively. Finally, the images and corresponding binary masks were cropped into 512×512 pixel samples. The complete dataset contains 1538 samples, including 1127 training and 411 test samples, with no spatial overlap between the two subsets

(Table 1, Figure 5). This workflow ensured that the image patches and labels were spatially aligned and suitable for pixel-level semantic segmentation.

Table 1. Sample distribution of the Wuding River Basin terrace dataset.

Dataset Split	Number of Samples	Proportion
Train	1127	73.3%
Test	411	26.7%
Total	1538	100%

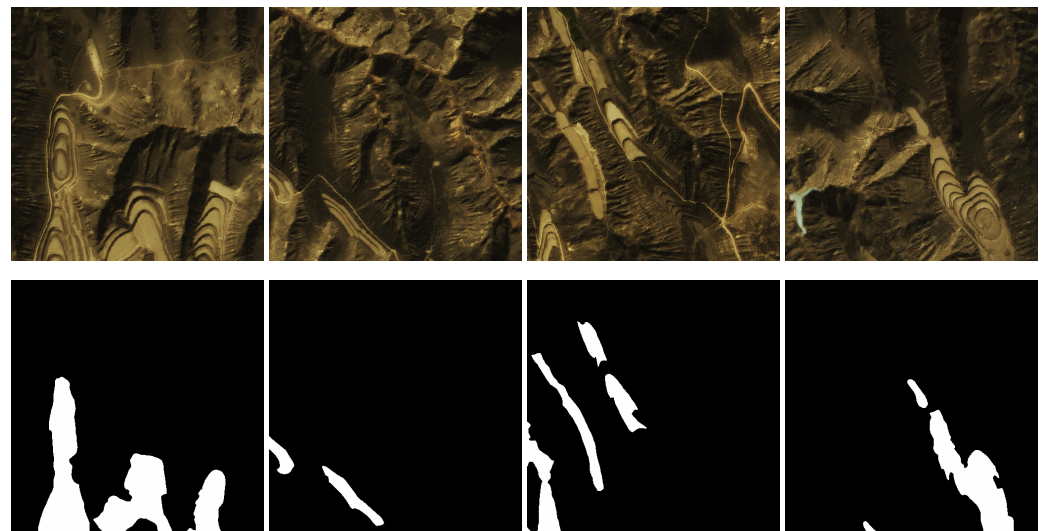


Figure 5. Sample images and corresponding annotation masks from the Wuding River Basin terrace dataset.

3.2. Evaluation Metrics

We adopted four complementary metrics to comprehensively evaluate the performance of the proposed TerraceNet and compare it with other methods: Intersection over Union (IoU), F1-Score, Precision, and Recall. These metrics are widely used in semantic segmentation tasks and are derived from the confusion matrix, where true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) denote correctly classified terrace pixels, non-terrace pixels incorrectly classified as terrace, correctly classified non-terrace pixels, and terrace pixels incorrectly classified as non-terrace, respectively.

The four metrics are formally defined as follows:

Precision measures the proportion of correctly predicted terrace pixels among all pixels predicted as terrace:

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

Recall measures the proportion of correctly predicted terrace pixels among all ground-truth terrace pixels:

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

F1-Score is the harmonic mean of Precision and Recall, providing a balanced assessment of model performance:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

IoU, which is also referred to as the Jaccard index, quantifies the overlap between the predicted segmentation and the ground truth:

$$IoU = \frac{TP}{TP + FP + FN} \quad (18)$$

Among these metrics, IoU and F1-Score are considered the most critical for segmentation tasks, jointly accounting for both over- (FP) and under-segmentation (FN). Precision reflects the reliability of terrace pixel prediction, whereas Recall reflects the completeness of extracted terrace regions. All metrics are reported as percentages ranging from 0% to 100%, with higher values indicating better performance.

3.3. Comparison Methods

To comprehensively evaluate the performance of the proposed TerraceNet, we compare it against five representative semantic segmentation methods covering both CNN- and Transformer-based architectures. All comparison methods were implemented using publicly available codebases and trained under the same basic experimental settings, including the same dataset split, input size, optimizer, learning rate schedule, and number of training epochs. Since these methods were trained without data augmentation, TerraceNet (without augmentation) was used as the fair counterpart for comparison. TerraceNet (with augmentation) was additionally evaluated to investigate the effect of the data augmentation strategy.

UNet [18]: U-Net is a classical encoder–decoder architecture originally proposed for biomedical image segmentation. It uses symmetric skip connections between the encoder and decoder paths to preserve spatial details while capturing multi-scale semantic information. In this study, UNet serves as a strong CNN-based reference model for evaluating the benefit of the proposed hybrid design.

DeepLabV3+ [30]: DeepLabV3+ is a CNN-based segmentation model that incorporates atrous spatial pyramid pooling (ASPP) to capture multi-scale contextual information. It also employs a decoder module to refine segmentation boundaries, making it suitable for fine-structured targets such as terrace boundaries.

KNet [31]: K-Net is a recent CNN-based unified segmentation framework that performs semantic, instance, and panoptic segmentation using learnable kernel representations. Its dynamic kernel update mechanism enables adaptive feature aggregation, providing a useful comparison for evaluating TerraceNet’s representation capacity.

SegFormer [22]: SegFormer is an efficient Transformer-based semantic segmentation framework that comprises a hierarchical Transformer encoder and a lightweight MLP decoder. It removes the need for positional encodings and achieves robust performance across various object scales, making it a representative lightweight Transformer-based comparison method.

Mask2Former [23]: Mask2Former is a universal image segmentation model that unifies semantic, instance, and panoptic segmentation tasks using masked attention. Its Transformer-based architecture captures long-range dependencies through masked attention mechanisms and provides a strong comparison method for evaluating global context modeling in terrace extraction.

3.4. Implementation Details

Training settings: TerraceNet was implemented using PyTorch 2.4.1 and trained on a single NVIDIA GeForce RTX 3090 GPU for 100 epochs with a batch size of four. The AdamW optimizer was employed with an initial learning rate of 1×10^{-4} and a weight

decay of 1×10^{-4} . A cosine annealing learning rate schedule was adopted, decreasing the learning rate from 1×10^{-4} to 1×10^{-6} over 100 epochs.

Data augmentation: Geometric augmentations include random horizontal flipping, vertical flipping, and 90° rotation, each with a probability of 0.5. These geometric transformations are applied synchronously to both images and masks to preserve pixel-level alignment. Spectral augmentations, including contrast adjustment, brightness adjustment, and Gaussian noise injection, are only applied to images to improve illumination and imaging variation robustness. No data augmentation is applied to the test set.

Pre-training: The ConvNeXt-Tiny encoder is initialized with ImageNet-1K pre-trained weights and subsequently fine-tuned during training, which helps accelerate convergence and improve feature representation for the relatively limited terrace dataset.

Post-processing: No post-processing operations were applied to the model outputs during evaluation. More specifically, morphological filtering, small-object removal, hole filling, and vectorization were not used. The reported quantitative results therefore reflect the intrinsic segmentation performance of each model.

Random seed: A fixed random seed (42) was used for all experiments to improve reproducibility and reduce the influence of random initialization and data shuffling.

3.5. Computational Complexity Analysis

To further evaluate the computational cost of the proposed model, TerraceNet was compared with the representative comparison methods in terms of the parameter number, FLOPs, and inference time. FLOPs were calculated using an input size of $1 \times 3 \times 512 \times 512$, which is consistent with the image patch size used in the experiments. Inference time was measured on a single NVIDIA GeForce RTX 3090 GPU with a batch size of one. For each model, 50 warm-up iterations were first performed, followed by 200 repeated forward passes, and the average inference time was reported.

Table 2 presents the computational complexity of different methods. TerraceNet has 40.05 M parameters and 103.45 G FLOPs, making it comparable to DeepLabV3+ in terms of parameter count but lower in FLOPs. Although TerraceNet introduces a boundary decoder and a multi-scale Transformer decoder, its FLOPs are lower than those of UNet, DeepLabV3+, and KNet. However, its inference time is higher than that of the comparison methods, which may be due to the additional multi-branch decoder structure, repeated feature interpolation and concatenation, and boundary-guided fusion operations. These results indicate that TerraceNet achieves improved segmentation accuracy with acceptable computational complexity, although further acceleration through lightweight design, efficient attention, or model compression remains a potential direction for future work.

Table 2. Computational complexity comparison of different methods.

Method	Params (M)	FLOPs (G)	Inference Time (ms)
UNet	31.04	193.18	21.00
DeepLabV3+	40.35	183.08	18.12
KNet	49.18	407.28	12.40
SegFormer	3.71	6.75	4.34
Mask2Former	15.75	30.36	8.95
TerraceNet	40.05	103.45	28.43

3.6. Comparison Results

Table 3 presents the quantitative comparison results on the test set. The proposed TerraceNet achieves the highest performance among the compared methods across all four evaluation metrics, demonstrating its effectiveness for terrace extraction in the Wuding

River Basin dataset. Notably, we report on two TerraceNet variants: TerraceNet (without augmentation) trained without data augmentation, and TerraceNet (with augmentation) trained with the full data augmentation pipeline described in Section 3.4. This setting separates fair comparison with existing methods from the additional data augmentation analysis.

Table 3. Comparison of different methods on the test set.

Method	IoU (%)	F1 (%)	Precision (%)	Recall (%)
UNet	72.94	84.35	85.95	82.81
DeepLabV3+	74.32	85.27	85.33	85.22
KNet	74.87	85.63	86.55	84.72
SegFormer	72.93	84.35	85.27	83.44
Mask2Former	75.41	85.98	86.70	85.27
TerraceNet (without augmentation)	76.44	86.65	87.98	85.35
TerraceNet (with augmentation)	77.61	87.39	87.88	86.92

Comparison with CNN-based methods: Among the CNN-based methods, K-Net achieves the best performance with an IoU of 74.87%, followed by DeepLabV3+ (74.32%) and UNet (72.94%). TerraceNet (without augmentation) outperforms KNet by 1.57 percentage points in IoU (76.44% vs. 74.87%) and by 1.02 percentage points in F1-Score (86.65% vs. 85.63%). This improvement can be attributed to two key factors. Firstly, the Transformer decoder captures long-range contextual dependencies, which is particularly beneficial for the accurate segmentation of large-scale terrace fields. Secondly, the boundary decoder provides explicit boundary supervision, effectively compensating for the inherent limitation of pure Transformer methods in fine-grained boundary delineation—a task with which the latter struggle considerably when handling narrow, strip-like terraces.

Comparison with Transformer-based methods: Among the Transformer-based methods, Mask2Former achieves the highest segmentation accuracy with an IoU of 75.41%, whereas SegFormer achieves 72.93%. TerraceNet (without augmentation) outperforms Mask2Former by 1.03 percentage points in IoU (76.44% vs. 75.41%) and by 0.67 in F1-Score (86.65% vs. 85.98%). From a quantitative perspective, TerraceNet achieves higher Recall (85.35%) and Precision (87.98%), indicating a balanced capability to identify terrace regions while suppressing false positives.

Effect of data augmentation: Comparing the two TerraceNet variants (with or without augmentation), applying data augmentation leads to consistent performance improvements across most evaluation metrics. More specifically, augmentation increases IoU from 76.44% to 77.61% (an improvement of 1.17 percentage points), F1-Score from 86.65% to 87.39% (+0.74 percentage points), and Recall from 85.35% to 86.92% (+1.57 percentage points), while Precision shows only a slight decrease (87.98% vs. 87.88%). These results suggest that data augmentation enhances the model's generalization capability, particularly by improving terrace extraction completeness under diverse conditions.

A visualization of the prediction results is shown in Figure 6. In the first row, TerraceNet and U-Net produce relatively complete segmentation results, while DeepLabV3+, KNet, SegFormer, and Mask2Former generate predictions with noticeable holes.

In the second row, TerraceNet achieves the best segmentation performance, while U-Net, DeepLabV3+, and Mask2Former fail to identify the gaps between adjacent terrace blocks. K-Net incorrectly recognizes a small isolated terrace patch, while SegFormer misidentifies the terrace intervals.

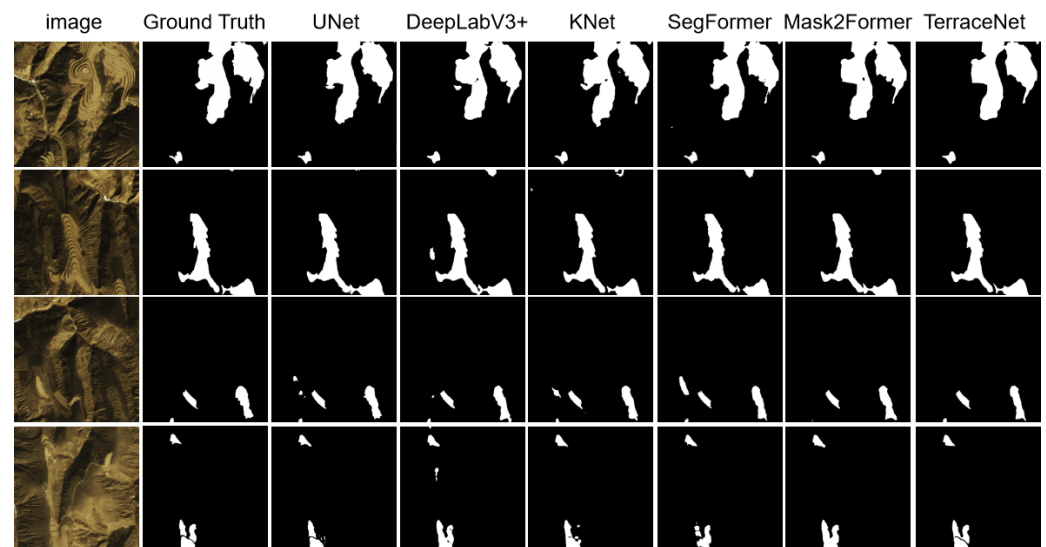


Figure 6. Visual comparison of the terrace extraction results among different methods.

In the third row, TerraceNet again achieves the best performance. U-Net, DeepLabV3+, K-Net, and SegFormer produce false positive predictions, incorrectly identifying non-terrace regions as terraces, whereas Mask2Former exhibits false negatives and misses a small portion of the terrace area.

In the fourth row, TerraceNet provides a more complete extraction result, while U-Net, K-Net, and SegFormer suffer from under-segmentation, missing parts of the terrace regions. DeepLabV3+ exhibits over-segmentation and fails to identify terrace intervals, with small holes appearing within the predicted regions, and Mask2Former also misclassifying non-terrace areas as terraces.

Overall, compared with the other methods, TerraceNet generates more accurate and spatially coherent segmentation results across the four representative test samples, demonstrating its effectiveness in extracting both narrow strip-like terraces and broad contiguous terraces under challenging conditions.

3.7. Ablation Study

To systematically evaluate the contribution of each core component in TerraceNet, we conducted ablation experiments on the test set (Table 4). Three key components were investigated: the DF-FPN for multi-scale feature fusion, the boundary decoder for boundary supervision and structural prior generation, and the multi-scale Transformer decoder for global context modeling.

Effect of DF-FPN: Compared with the baseline model, adding the DF-FPN module improves IoU from 76.89% to 77.23% (+0.34%) and the F1-Score from 86.94% to 87.15% (+0.21%). This improvement confirms the effectiveness of adaptive multi-scale feature fusion for terrace extraction, particularly in capturing narrow terrace boundaries and maintaining the overall structure of large contiguous terrace patches.

Effect of boundary decoder: Adding the boundary decoder alone achieves an IoU of 77.21% and an F1-Score of 87.14%, comparable to the baseline + DF-FPN configuration. However, a Precision–Recall trade-off is observed: Recall increases from 86.89% to 87.51% (+0.62%), while Precision decreases from 87.42% to 86.77% (−0.65%). This suggests that boundary information can improve terrace extraction performance, but without global contextual modeling or an effective fusion mechanism, the additional boundary prior may also introduce false positives. Boundary supervision is therefore most effective when combined with a module capable of suppressing contextually inconsistent predictions.

Table 4. Ablation study results on the test set.

Method	IoU (%)	F1 (%)	Precision (%)	Recall (%)
baseline	76.89	86.94	86.97	86.91
baseline + DF-FPN	77.23	87.15	87.42	86.89
baseline + DF-FPN + boundary decoder	77.21	87.14	86.77	87.51
DF-FPN + multi-scale Transformer	76.97	86.99	87.36	86.63
FPN + boundary + multi-scale Transformer	77.13	87.09	87.13	87.05
DF-FPN + boundary + multi-scale Transformer	77.61	87.39	87.88	86.92

Note: The baseline model adopts the same ConvNeXt encoder and a plain top-down CNN decoder, which consists of four convolutional blocks. The first block processes the highest-level feature map, while the remaining three blocks progressively up-sample the feature maps by a factor of two using bilinear interpolation and concatenate them with the corresponding lower-level features. Each convolutional block consists of a 3×3 convolution, batch normalization, and ReLU activation. Finally, a 1×1 convolution is used to generate the binary segmentation logits, which are further up-sampled by bilinear interpolation to the original image resolution. No boundary decoder, boundary-guided fusion, or Transformer block is included in this baseline.

Effect of multi-scale Transformer decoder: The DF-FPN + multi-scale Transformer configuration leads to a slight performance decline, with IoU decreasing to 76.97% and F1-Score to 86.99%, accompanied by marginal reductions in both Precision and Recall. This indicates that directly combining the multi-scale Transformer with DF-FPN in the absence of boundary guidance does not fully exploit the advantages of global context modeling. Long-range dependency modeling may also introduce additional noise when structural boundary information is insufficient, thereby adversely affecting terrace segmentation.

When the DF-FPN, boundary decoder, and multi-scale Transformer module are jointly incorporated (i.e., the DF-FPN + boundary + multi-scale Transformer configuration), the model achieves a substantial performance improvement, reaching the highest performance among all ablation settings: IoU of 77.61% (+0.72% over the baseline), F1-Score of 87.39% (+0.45%), Precision of 87.88% (+0.91%), and Recall of 86.92% (+0.01%). Notably, replacing DF-FPN with the conventional FPN under the same configuration (FPN + boundary + multi-scale Transformer) yields a lower performance, with an IoU of 77.13% and F1-Score of 87.09%. These results confirm that the dynamic fusion mechanism contributes positively to terrace extraction when working synergistically with boundary priors and global context modeling.

4. Discussion

4.1. Interpreting Our Findings in the Context of Previous Studies

The experimental results in Sections 3.6 and 3.7 reveal several noteworthy patterns. Firstly, Transformer methods, particularly Mask2Former, outperform most CNN-based methods on the terrace extraction task, which is consistent with studies demonstrating Transformer architectures' advantages in global context modeling [32]. This finding indicates that global contextual information is beneficial for capturing long-range semantic relationships in terrace fields, especially for elongated strip-like terraces and broad contiguous terrace patches.

Secondly, the ablation study reveals a fundamental difference between CNN and Transformer decoders in utilizing boundary information. Compared with the baseline + DF-FPN configuration, introducing boundary supervision to the CNN decoder (baseline + DF-FPN + boundary decoder) improves Recall but reduces Precision, resulting

in limited net gains. This phenomenon can be attributed to the inherently strong local edge perception capabilities of the CNN decoder, which limits the marginal benefit of additional boundary information. In contrast, when boundary supervision is used in conjunction with the multi-scale Transformer decoder, the full model better resolves the Precision–Recall trade-off and achieves high levels of both. This demonstrates that Transformer-based global context is important for contextualizing boundary information and suppressing false positives introduced by boundary supervision alone.

Thirdly, the synergistic combination of DF-FPN, boundary decoder, and multi-scale Transformer decoder achieves the best overall performance. This result supports the hypothesis that adaptive multi-scale features, structural boundary priors, and global context modeling are complementary and mutually reinforcing. The success of TerraceNet demonstrates that hybrid CNN–Transformer architectures are well-suited for scale-variant features such as terraces, which require both fine-grained boundary details and large-scale contextual understanding.

4.2. Implications and Limitations

Implications: Firstly, the DF-FPN effectively enhances hierarchical feature fusion for terraces with different scales and morphologies. Although the ConvNeXt encoder generates multi-level features, shallow and deep features play different roles in terrace extraction. Shallow features preserve fine spatial details and boundary information, which are particularly important for elongated strip-like terraces, whereas deep features provide stronger semantic responses for distinguishing terrace regions from complex backgrounds. By adaptively weighting and fusing features from different levels, DF-FPN improves the model's representation of terraces with varying sizes and boundary conditions. Secondly, the Transformer decoder complements DF-FPN by further modeling long-range semantic dependencies, which helps maintain the continuity of elongated terraces and the consistency of broad contiguous terrace patches. Third, the explicit incorporation of boundary priors provides a transferable strategy for improving the segmentation of linear, narrow, and fine-structured objects, such as roads, rivers, field boundaries, and other agricultural land surface features. Finally, the proposed method can provide reliable spatial information for terrace resource investigation and soil and water conservation monitoring, thereby supporting practical decision-making in ecological restoration and land management.

Limitations: Several limitations should be acknowledged. Firstly, the dataset used in this study is limited to the Wuding River Basin in the Loess Plateau; therefore, generalizability to other terrace systems, such as Southeast Asian rice terraces and Mediterranean stone-walled terraces, requires further validation. Secondly, the 2 m spatial resolution represents a typical high-resolution satellite image setting, but performance on very high (<1 m) or lower-resolution imagery requires systematic investigation. Thirdly, although TerraceNet achieves improved accuracy, its inference time is higher than that of some compared methods, indicating that lightweight optimization remains necessary for large-scale or real-time deployment.

4.3. Future Research Directions

Several future research directions can be further explored. Firstly, cross-region generalization should be evaluated across diverse terrace systems using domain adaptation or transfer learning techniques to reduce reliance on region-specific labels. Secondly, multi-temporal analysis incorporating time-series imagery could enable dynamic monitoring of terrace abandonment, cultivation status, and conservation effectiveness. Thirdly, lightweight deployment through pruning, quantization, knowledge distillation, or efficient attention mechanisms could facilitate real-time applications on edge devices, unmanned

aerial vehicles, or operational monitoring platforms. Finally, integrating terrace extraction results with ecological and soil erosion models could transform pixel-level mapping outputs into quantitative decision-support information for land conservation planning.

5. Conclusions

This study proposes TerraceNet, a hybrid CNN–Transformer architecture for terrace extraction from high-resolution remote sensing imagery. The proposed model integrates four key components: a ConvNeXt-based encoder for feature extraction, a DF-FPN for adaptive multi-scale feature fusion, a boundary decoder for explicit boundary supervision and structural prior generation, and a multi-scale Transformer decoder for global context modeling and final segmentation.

Experiments on the Wuding River Basin dataset show that TerraceNet (without augmentation) achieves an IoU of 76.44% and F1-Score of 86.65%, outperforming Mask2Former by 1.03 and 0.67 percentage points, respectively, under the same non-augmented setting. With data augmentation, TerraceNet further improves the IoU and F1-Score to 77.61% and 87.39%, respectively. Ablation studies confirm that the joint use of the three components—DF-FPN, boundary decoder, and multi-scale Transformer decoder—achieves the best performance, demonstrating their complementary roles in terrace extraction.

In terms of practical applications, the proposed method provides a potential technical tool for terrace resource investigation and soil and water conservation monitoring. By accurately extracting terrace distributions from high-resolution remote sensing imagery, TerraceNet can support the rapid terrace resource mapping, terrace spatial pattern assessment, and terrace conservation status monitoring in hilly and mountainous regions. In addition, the extracted terrace information can provide useful spatial data for remote sensing-based soil and water conservation monitoring, soil erosion assessment, ecological restoration evaluation, and agricultural land management.

Author Contributions: Conceptualization, F.J., Z.W. and K.G.; methodology, F.J. and K.G.; validation, F.J. and Z.W.; formal analysis, F.J.; data curation, J.Z.; writing—original draft preparation, F.J., Y.W. and J.Z.; visualization, C.Z. and Q.X.; supervision, K.G. and Z.W.; funding acquisition, F.J. and K.G. 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 Nos. 42501525, U2443214], the Natural Science Foundation of Henan Province [Grant No. 252300420863], and the Central Nonprofit Research Institutions Basic Scientific Research Special Fund [Grant Nos. HKF202404, HKY-YF-2024-03, HKY-JBYW-2025-07].

Data Availability Statement: Data is available from the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.5 for the purposes of language polishing and grammar improvement. The authors have reviewed and edited the output and take full responsibility for the content of this publication. Additionally, we also utilized the MDPI language editing service to further refine the writing of this manuscript.

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

References

1. Wei, W.; Chen, D.; Wang, L.; Daryanto, S.; Chen, L.; Yu, Y.; Lu, Y.; Sun, G.; Feng, T. Global synthesis of the classifications, distributions, benefits and issues of terracing. *Earth-Sci. Rev.* **2016**, *159*, 388–403. [\[CrossRef\]](#)
2. Deng, C.; Zhang, G.; Liu, Y.; Nie, X.; Li, Z.; Liu, J.; Zhu, D. Advantages and disadvantages of terracing: A comprehensive review. *Int. Soil Water Conserv. Res.* **2021**, *9*, 344–359. [\[CrossRef\]](#)
3. Chen, D.; Wei, W.; Chen, L. Effects of terracing practices on water erosion control in China: A meta-analysis. *Earth-Sci. Rev.* **2017**, *173*, 109–121. [\[CrossRef\]](#)

4. Feng, J.; Wei, W.; Pan, D. Effects of rainfall and terracing-vegetation combinations on water erosion in a loess hilly area, China. *J. Environ. Manag.* **2020**, *261*, 110247. [[CrossRef](#)]
5. Lu, Y.; Li, X.; Xin, L.; Song, H.; Wang, X. Mapping the terraces on the Loess Plateau based on a deep learning-based model at 1.89 m resolution. *Sci. Data* **2023**, *10*, 115. [[CrossRef](#)] [[PubMed](#)]
6. Zhang, E.; Chen, Y.; Wei, S.; Liu, C.; Wang, H.; Deng, B.; Lin, H.; Yang, X.; Li, Y.; Duan, X. A 30 m resolution dataset of soil and water conservation terraces across China for 2000, 2010, and 2020. *Earth Syst. Sci. Data* **2025**, *17*, 6315–6330. [[CrossRef](#)]
7. Baryła, A.; Pierzgałski, E. Ridged terraces-functions, construction and use. *J. Environ. Eng. Landsc. Manag.* **2008**, *16*, 1–6. [[CrossRef](#)]
8. Shi, Y.; Wei, W. Mapping long-term terrace change in the Chinese Loess Plateau: A new approach by decision tree model with digital elevation model (DEM) and land use. *Earth Surf. Process. Landf.* **2024**, *49*, 1316–1329. [[CrossRef](#)]
9. Meng, X.; Song, Z.; Cui, X.; Shi, P. Changes of Terrace Distribution in the Qinba Mountain Based on Deep Learning. *Sustainability* **2025**, *17*, 10971. [[CrossRef](#)]
10. Luo, L.; Li, F.; Dai, Z.; Yang, X.; Liu, W.; Fang, X. Terrace extraction based on remote sensing images and digital elevation model in the loess plateau, China. *Earth Sci. Inform.* **2020**, *13*, 433–446. [[CrossRef](#)]
11. Ho, T.K. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*; IEEE: New York, NY, USA, 1995; pp. 278–282.
12. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery (ACM): New York, NY, USA, 2016; pp. 785–794.
13. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
14. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [[CrossRef](#)]
15. Ding, H.; Na, J.; Jiang, S.; Zhu, J.; Liu, K.; Fu, Y.; Li, F. Evaluation of three different machine learning methods for object-based artificial terrace mapping—A case study of the loess plateau, China. *Remote Sens.* **2021**, *13*, 1021. [[CrossRef](#)]
16. Cao, B.; Yu, L.; Naipal, V.; Ciaia, P.; Li, W.; Zhao, Y.; Wei, W.; Chen, D.; Liu, Z.; Gong, P. A 30-meter terrace mapping in China using Landsat 8 imagery and digital elevation model based on the Google Earth Engine. *Earth Syst. Sci. Data Discuss.* **2020**, *13*, 2437–2456.
17. Chen, Y.; Yao, X.; Deng, G.; Huang, W.; Tang, L.; Li, W.; Bilal, M.; Liu, Q. Crowdsourcing Data Engine Driven Deep Adaption Network for Global Urban Green Space Changes Monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 24990–25004. [[CrossRef](#)]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
19. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.0558.
20. Yu, M.; Rui, X.; Xie, W.; Xu, X.; Wei, W. Research on automatic identification method of terraces on the loess plateau based on deep transfer learning. *Remote Sens.* **2022**, *14*, 2446. [[CrossRef](#)]
21. Zhao, Y.; Zou, J.; Liu, S.; Xie, Y. Terrace Extraction Method Based on Remote Sensing and a Novel Deep Learning Framework. *Remote Sens.* **2024**, *16*, 1649. [[CrossRef](#)]
22. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 7262–7272.
23. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 1290–1299.
24. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 205–218.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
26. Kan, G.; Xiao, J.; Liu, B.; Wang, B.; He, C.; Yang, H. Remote Sensing Extraction and Spatiotemporal Change Analysis of Time-Series Terraces in Complex Terrain on the Loess Plateau Based on a New Swin Transformer Dual-Branch Deformable Boundary Network (STDBNet). *Remote Sens.* **2025**, *18*, 85. [[CrossRef](#)]
27. Li, C.; Wang, Y.; Zhang, C.; Zhang, Y. ΩFormer: Dual-Modal Ω-like Super-Resolution Transformer Network for Cross-scale and High-accuracy Terraced Field Vectorization Extraction. *arXiv* **2024**, arXiv:2411.17088.
28. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 11976–11986.
29. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. 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.

30. 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 Nature: Cham, Switzerland, 2018; pp. 801–818.
31. Zhang, W.; Pang, J.; Chen, K.; Loy, C.C. K-net: Towards unified image segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10326–10338.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

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.