applied sciences

MDPI

Article

LGSTNet: Efficient Local–Global Collaboration for Heterogeneous Remote-Sensing Landslide Segmentation

Anhao Huang †, Hongxia Cui *,† and Ruohan Wei

Department of Software Engineering, College of Computer Science and Artificial Intelligence, Bohai University, Jinzhou 121013, China

* Correspondence: chx@qymail.bhu.edu.cn

† These authors contributed equally to this work.

Abstract

Purpose: Landslide segmentation from remote sensing imagery is challenged by scale variation, irregular morphology, blurred boundaries, and confusion with spectrally similar backgrounds. This study aims to develop an efficient framework that coordinates heterogeneous input regulation, local structural preservation, and hierarchical semantic modeling. **Methods:** LGSTNet is organized around two coupled design principles. An input-configuration-oriented local–global encoder combines input recalibration, a shallow Ghost branch, and a compact three-stage Swin Transformer. An accuracy–efficiency-oriented compressed context-fusion strategy enhances the deepest Transformer feature through compressed multi-scale, directional, and global context branches and integrates multi-level representations using a lightweight projection–fusion decoder. **Results:** Experiments were conducted on the Jiuzhai Valley, HR-GLDD, and Landslide4Sense datasets. On Landslide4Sense, LGSTNet achieved the best F1-score, foreground intersection over union (IoU), mean IoU (mIoU), and Matthews correlation coefficient (MCC) among the seven evaluated models, reaching 64.01%, 47.07%, 72.81%, and 63.39%, respectively, with 3.79 M parameters, 2.39 G FLOPs, and 63.93 frames per second. On HR-GLDD and Jiuzhai Valley, it provided competitive segmentation performance with favorable inference efficiency, although it did not outperform every baseline. A within-dataset input-configuration experiment on Landslide4Sense further showed that the F1-score of LGSTNet increased from 42.87% with RGB input to 64.01% with the full fourteen-channel input; SegFormer exhibited the same overall trend, increasing from 42.50% to 61.56%. **Conclusion:** LGSTNet provides a favorable accuracy–efficiency trade-off across the three evaluated remote-sensing configurations. The within-dataset input-configuration results provide additional evidence that the complete system can benefit from the information available in the full multi-channel input, while the cross-dataset comparison alone is not interpreted as causal evidence that channel number or modality richness by itself explains the observed performance differences.

Keywords: landslide segmentation; semantic segmentation; remote sensing images; lightweight network; feature fusion

Check for updates

Academic Editor: Atsushi Mase

Received: 3 October 2026
Revised: 20 October 2026
Accepted: 25 October 2026
Published: 4 November 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

Landslides are destructive geological hazards that are frequently triggered by intense rainfall, earthquakes, slope excavation, and other environmental disturbances. Rapid and accurate landslide mapping is therefore essential for disaster investigation, emergency response, post-disaster assessment, and long-term hazard mitigation [1,2]. Compared

with field surveys and manual interpretation, remote sensing imagery provides broader spatial coverage and higher acquisition efficiency, making it an important source of information for landslide identification and mapping in complex mountainous regions [3,4]. Recent reviews have further indicated that deep learning has become a major approach to landslide-oriented remote sensing analysis because it can automatically learn discriminative representations from large-scale imagery [5].

Semantic segmentation assigns a semantic label to each pixel and generates dense landslide masks rather than coarse image-level predictions [6]. This dense prediction paradigm facilitates the localization of landslide boundaries, estimation of affected areas, and subsequent geohazard analysis. Nevertheless, accurate landslide segmentation remains challenging. Landslide regions may occur at markedly different spatial scales, exhibit fragmented or elongated morphologies, and have indistinct boundaries with surrounding areas. Moreover, exposed soil, riverbeds, roads, shadowed regions, and construction sites may present spectral or textural characteristics similar to those of landslide surfaces, thereby increasing both false-positive and false-negative predictions.

Existing segmentation approaches provide complementary capabilities for addressing these challenges. CNN-based methods are effective at extracting local texture, edge, and structural information [6], while lightweight segmentation networks emphasize efficient inference under limited computational resources [7,8]. However, the local nature of convolution restricts its ability to model long-range dependencies. Transformer-based and hybrid CNN–Transformer architectures strengthen global contextual reasoning, multi-level feature fusion, and global–local interaction in remote sensing image segmentation [9–11], but may introduce considerable computational overhead. Consequently, simultaneously preserving fine-grained local details, modeling hierarchical semantic context, enhancing deep contextual representations, and maintaining efficient decoding remains a challenging problem.

To address these limitations, this paper formulates LGSTNet as a lightweight local–global collaborative framework for landslide segmentation under heterogeneous remote sensing input configurations. Rather than treating input attention, local feature extraction, hierarchical semantic modeling, contextual enhancement, and decoding as independent architectural additions, LGSTNet coordinates these operations according to two coupled design principles.

The first principle is heterogeneous-input-oriented local–global collaboration. Input-stage feature recalibration is coupled with a shallow Ghost branch and a compact three-stage Swin Transformer branch, allowing cross-channel response regulation, local structural preservation, and hierarchical semantic modeling to be performed within a unified encoding process. The input-recalibration configuration is selected according to the channel composition: spectral/channel attention and band mixing are used for all inputs, whereas spatial attention is additionally activated for the fourteen-channel multispectral–terrain data.

The second principle is accuracy–efficiency-oriented compressed context fusion. The deepest Swin feature is enhanced through compressed multi-scale, directional, and global context aggregation, and the resulting hierarchical features are integrated through a lightweight MLP projection–fusion decoder. Together with the compact three-stage Transformer encoder, this design limits redundant feature expansion and heavy decoding operations while preserving complementary local and global information.

The main contributions of this paper are summarized as follows:

- A heterogeneous-input-oriented local–global collaboration framework is formulated for landslide segmentation from remote sensing imagery. It couples input-configuration-specific recalibration, lightweight shallow-detail extraction, and hierarchical Transformer representation within a parallel Ghost–Swin encoder. This design

addresses cross-channel redundancy, irregular boundary preservation, and long-range semantic modeling as a coordinated representation problem rather than as separate enhancement stages.

- An accuracy–efficiency-oriented compressed context-fusion strategy is developed. The compact three-stage Swin encoder is combined with a deep context block whose multi-scale, directional, and global branches are compressed before residual fusion. The resulting features are then aligned and integrated through a common MLP projection–fusion decoder using depthwise separable operations. This strategy limits redundant feature expansion while retaining complementary shallow structural and deep contextual information.
- Experiments on RGB, four-channel, and fourteen-channel remote-sensing inputs provide both cross-dataset and within-dataset evaluations of the proposed framework. In addition to the three-dataset comparison, a controlled Landslide4Sense experiment evaluates RGB, four-channel, and full fourteen-channel input configurations on the same dataset, with SegFormer included as a representative reference model. Fine-grained ablations further isolate channel attention, spatial attention, and band mixing, while architectural replacements evaluate GhostStem, LightDSCA, and the lightweight projection–fusion decoder. These experiments support the complementary roles of input regulation, shallow local representation, compressed context enhancement, and efficient multi-level fusion without relying on cross-dataset causal attribution.

2. Related Work

Deep learning-based semantic segmentation has become an important technique for remote sensing image interpretation [6,12]. CNN-based methods such as FCN, U-Net, PSPNet, UNet++, and DeepLabv3+ have established the encoder–decoder paradigm for dense prediction and demonstrated the effectiveness of skip connections and multi-scale context modeling [13–17]. Their convolutional operators are effective at extracting local texture, edge, and shape information, which is important for recognizing landslide boundaries and exposed surface patterns. For high-resolution remote sensing imagery, these local representations are particularly useful because landslide regions often contain fine cracks, fragmented contours, and complex transitions to surrounding vegetation, soil, roads, and riverbeds. Nevertheless, the locality of convolution limits its ability to capture long-range semantic relationships. Consequently, a network relying only on local convolution may confuse landslides with bare soil or road-like regions, or recover incomplete landslide bodies when the target extends across a large spatial range.

Transformer-based vision models provide a complementary mechanism for modeling long-range dependencies through self-attention [18,19]. For dense prediction, SegFormer employs a hierarchical Transformer encoder and a lightweight MLP fusion decoder, whereas Swin Transformer reduces computational cost by restricting self-attention to local windows and introducing shifted-window connections [20,21]. SwinUNet and TransUNet further demonstrate the effectiveness of Transformer-based encoders and hybrid CNN–Transformer designs for segmentation tasks [22,23]. In remote sensing image segmentation, Transformer-based architectures have been explored to strengthen hierarchical contextual modeling and multi-level feature fusion [9–11]. Multimodal fusion and feature-enhancement architectures further integrate complementary information from heterogeneous remote sensing sources [24–28]. Recent studies have also introduced state-space models for efficient long-range representation [29], while Mamba-based multimodal fusion has been explored for heterogeneous remote sensing inputs [30]. Although these methods enhance global contextual reasoning, they may introduce considerable computational

overhead, while deep semantic representations alone may not sufficiently preserve shallow boundary details [20–23].

Recent multimedia image-analysis studies further demonstrate that segmentation and heterogeneous visual perception can benefit from coordinated local–global representation and task-oriented fusion. LET-Net combines Transformer-based global modeling with local feature alignment and a progressive refinement decoder for small-target segmentation [31]. Fusion-based deep learning has also been investigated for multimodal magnetic-resonance segmentation, where complementary image sequences are integrated at different stages of the segmentation pipeline [32]. In heterogeneous UAV perception, MCDFus fuses visible and infrared imagery through a Mamba–CNN architecture to combine complementary sensor information and support downstream visual understanding [33]. These studies support a multimedia-system perspective in which heterogeneous visual channels, local details, global context, and efficient fusion are coordinated according to the target analysis task.

Practical landslide monitoring based on remote sensing imagery requires a balance between segmentation accuracy and inference efficiency. Lightweight segmentation networks such as ESPNet, ICNet, and BiSeNet achieve efficient dense prediction through compact multi-scale modules, reduced-resolution branches, or separated spatial and semantic pathways [7,8,34]. GhostNet further indicates that many feature maps in deep networks are redundant and can be generated through inexpensive operations, providing a useful strategy for lightweight local feature extraction [35]. However, lightweight CNN-based designs remain limited in their ability to model long-range semantic dependencies.

Landslide-oriented segmentation methods have been developed from several perspectives. Domain adaptation and hybrid CNN–Transformer strategies have been employed to improve cross-domain generalization and local–global representation [36,37]. Multi-branch, attention-enhanced, foreground-aware, and multi-scale architectures have been introduced to improve landslide delineation in complex remote sensing scenes [38–41]. Recent task-specific networks have further emphasized feature aggregation, dual-branch modeling, multi-scale enhancement, and improved U-Net variants [42–45]. Although these designs improve segmentation quality, increasingly complex feature-extraction and decoding structures may increase model complexity or reduce inference speed.

Beyond architectural development, landslide mapping has evolved from traditional methods based on manually designed spectral, textural, morphological, or topographic features toward deep learning models that automatically learn task-specific representations from remote sensing data [1–5]. Recent studies further demonstrate that landslide segmentation can benefit from heterogeneous inputs, including optical imagery, multispectral bands, and terrain-derived information [37,46,47]. Compared with RGB imagery, heterogeneous multi-channel data provide complementary spectral and topographic cues that can help distinguish landslides from spectrally similar backgrounds. However, differences in channel semantics and numerical ranges also introduce more complex feature interactions and potential information redundancy.

Although prior studies have established effective mechanisms for channel recalibration, lightweight convolution, hierarchical Transformer encoding, multi-scale context aggregation, and feature decoding, these mechanisms are commonly optimized as separate architectural components. The remaining challenge is therefore not the absence of any individual operator, but how to coordinate heterogeneous input regulation, local-detail preservation, global semantic modeling, and multi-level fusion under a constrained computational budget.

This issue is particularly relevant to landslide segmentation because the input configuration may range from RGB imagery to multispectral and terrain-derived data, while the target regions exhibit irregular boundaries, fragmented structures, and substantial

scale variation. These considerations motivate a framework-level design organized around heterogeneous-input-oriented local–global collaboration and accuracy–efficiency-oriented compressed context fusion.

3. Materials and Methods

This section first describes the datasets and experimental input configurations, and then presents LGSTNet according to its two coupled design principles: heterogeneous-input-oriented local–global encoding and accuracy–efficiency-oriented compressed context fusion. The implementation includes input recalibration, parallel Ghost–Swin feature extraction, compressed deep-context enhancement, and lightweight multi-level MLP projection–fusion.

3.1. Datasets

Three landslide remote sensing datasets with different input configurations were used in this study. The first dataset is the "Jiuzhai Valley (UAV-0.5 m)" subset of the CASLandslide remote sensing dataset [48], which contains 1753 three-channel high-resolution UAV RGB image samples with a spatial size of 512 × 512 pixels and corresponding binary landslide labels. The second dataset is HR-GLDD [47], which contains four-channel remote sensing image samples with an input size of 128 × 128 pixels. Following its official split, HR-GLDD contains 1119 training samples, 284 validation samples, and 355 test samples. The third dataset is Landslide4Sense [46], which contains 4844 fourteen-channel image patches, including 3799 training samples, 245 validation samples, and 800 test samples, each with a spatial size of 128 × 128 pixels. Each patch comprises twelve Sentinel-2 multispectral bands, one ALOS PALSAR-derived slope layer, and one ALOS PALSAR-derived digital elevation model (DEM) layer.

For the three-channel Jiuzhai Valley dataset, the 1753 image patches were shuffled using a fixed random seed of 42 and divided into training, validation, and test subsets at a ratio of 7:2:1. The split was performed at the patch level rather than at the scene or geographic region level. No explicit spatial-independence constraint was imposed; therefore, spatial independence among the three subsets cannot be guaranteed, and patches originating from the same or nearby UAV acquisition areas may share similar spatial context. Accordingly, the Jiuzhai Valley results should be interpreted as performance on randomly held-out patches rather than as an evaluation of geographically independent scenes. For HR-GLDD and Landslide4Sense, the official dataset splits were adopted. Based on the characteristics of each dataset, the input size was set to 512 × 512 for the Jiuzhai Valley RGB dataset and 128 × 128 for HR-GLDD and Landslide4Sense.

Before being fed into the network, all input samples were converted from $H \times W \times C$ to $C \times H \times W$. Dataset-specific normalization was used. For Landslide4Sense and HR-GLDD, min–max normalization was independently applied to each input channel within each sample, with a small constant of 10^{-8} added to the denominator to avoid numerical instability. For the Jiuzhai Valley RGB images, pixel values were scaled to $[0, 1]$ by division by 255.

Dataset-specific online augmentation was used during training, and the same augmentation pipeline was applied to all compared models within a given dataset. For all three datasets, random rotations selected from 0°, 90°, 180°, and 270°, together with random horizontal and vertical flipping, were used. For Landslide4Sense, random crop-and-resize augmentation was applied with a probability of 0.8, followed by random erasing of the input image with a probability of 0.35; bilinear interpolation was used for images and nearest-neighbor interpolation for masks. For HR-GLDD and Jiuzhai Valley, random crop-and-resize augmentation was applied with a probability of 0.5. In addition, multiplicative

channel-wise intensity jitter was applied with a probability of 0.3 using independently sampled factors in $[0.9, 1.1]$. All geometric transformations were synchronously applied to the input image and its corresponding mask, whereas intensity perturbations were applied only to the image.

3.2. Methods

3.2.1. Overall Architecture of LGSTNet

Given a heterogeneous multi-channel remote sensing image $\mathbf{X} \in \mathbb{R}^{C_0 \times H \times W}$, where C_0, H, and W denote the number of input channels, height, and width, respectively, the objective of landslide segmentation is to predict a dense semantic label map $\hat{\mathbf{Y}} \in \mathbb{R}^{K \times H \times W}$. In this study, $K = 2$, corresponding to the background and landslide classes. The number of input channels is dataset-dependent: $C_0 = 3$ for the Jiuzhai Valley dataset, $C_0 = 4$ for HR-GLDD, and $C_0 = 14$ for Landslide4Sense.

As illustrated in Figure 1, LGSTNet is organized according to two coupled architectural principles rather than as a collection of independent enhancement modules.

The first principle is heterogeneous-input-oriented local–global encoding. The input feature is first recalibrated according to its channel configuration and is then processed in parallel by a shallow Ghost branch and a compact three-stage Swin Transformer branch. The Ghost branch retains boundary-sensitive local structures, whereas the Swin branch extracts hierarchical semantic representations. This parallel organization enables local and global information to be preserved without requiring a heavy convolutional encoder.

The second principle is accuracy–efficiency-oriented compressed context fusion. The deepest Swin feature is enhanced by LightDSCA, which aggregates multi-scale, directional, and global context using compressed branches. The shallow local feature and the hierarchical Transformer features are subsequently projected to a common embedding dimension, spatially aligned, and fused through lightweight depthwise separable and residual refinement operations. The complete design therefore forms a continuous representation path from heterogeneous input regulation to local–global encoding and efficient multi-level fusion.

The complete forward process of LGSTNet is summarized in Equations (1)–(7):

$$\mathbf{X}_0 = \mathcal{M}_{\mathrm{SSBM}}(\mathbf{X}), \tag{1}$$

$$\mathbf{F}_1 = \mathcal{G}_2(\mathcal{G}_1(\mathbf{X}_0)), \tag{2}$$

$$\left(\mathbf{F}_2, \mathbf{F}_3, \widetilde{\mathbf{F}}_4\right) = \mathcal{S}(\mathbf{X}_0), \tag{3}$$

$$\mathbf{F}_4 = \mathcal{D}\left(\widetilde{\mathbf{F}}_4\right), \tag{4}$$

$$\mathbf{F}_f = \mathcal{F}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4), \tag{5}$$

$$\mathbf{F}_r = \mathbf{F}_f + \mathcal{R}(\mathbf{F}_f), \tag{6}$$

$$\hat{\mathbf{Y}} = \mathcal{H}(\mathbf{F}_r). \tag{7}$$

In Equations (1)–(7), $\mathcal{M}_{\mathrm{SSBM}}(\cdot)$ denotes the SSBM input-enhancement operation, and $\mathbf{X}_0$ denotes the enhanced input feature. The functions $\mathcal{G}_1(\cdot)$ and $\mathcal{G}_2(\cdot)$ in Equation (2) represent the two cascaded Ghost modules, which generate the shallow local feature $\mathbf{F}_1$. The function $\mathcal{S}(\cdot)$ in Equation (3) denotes the compact three-stage Swin Transformer encoder, which produces the hierarchical semantic features $\mathbf{F}_2$, $\mathbf{F}_3$, and $\widetilde{\mathbf{F}}_4$.

As formulated in Equation (4), the raw Stage-3 Swin feature $\widetilde{\mathbf{F}}_4$ is enhanced by the LightDSCA operation $\mathcal{D}(\cdot)$ to obtain the final deep encoder feature $\mathbf{F}_4$. According to Equations (5)–(7), the multi-level features are subsequently fused, residually refined, and

converted into the final segmentation logits. In the implementation, $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, and $\mathbf{F}_4$ correspond to the intermediate features `c1`, `c2`, `c3`, and `c4`, respectively.

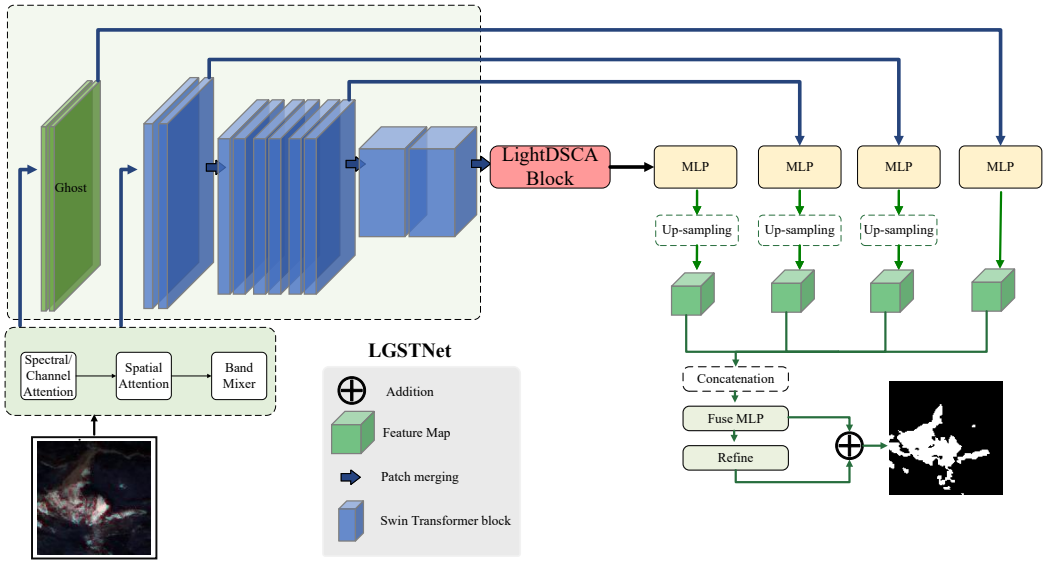


Figure 1. Overall architecture of LGSTNet. The label *img* denotes the input image. An input-configuration-oriented interface regulates heterogeneous channel responses before parallel local–global encoding by the Ghost and compact Swin branches. The deepest semantic feature is enhanced through compressed context aggregation. In the decoder, the four *MLP* blocks denote feature projection, Up-sampling denotes spatial alignment, *Fuse MLP* denotes lightweight feature fusion, and *Refine* denotes residual refinement.

3.2.2. Spectral-Spatial Attention and Band Mixing Module

Heterogeneous multi-channel remote sensing images contain complementary but potentially redundant information from multispectral bands, vegetation-related responses, and terrain-derived variables. Directly feeding all input channels into the feature encoder may introduce irrelevant background responses and weaken landslide-related cues. This problem is particularly evident for multispectral–terrain data, where different channels may make unequal contributions to foreground recognition.

Within the proposed local–global collaborative framework, SSBM serves as an input-configuration-specific recalibration interface. Drawing on channel-wise feature recalibration in SENet [49] and sequential channel–spatial attention in CBAM [50], it combines spectral/channel response weighting with explicit lightweight cross-band mixing. Spatial attention is additionally activated for the fourteen-channel multispectral–terrain input, where heterogeneous spectral and terrain responses introduce stronger spatial and cross-channel redundancy.

Accordingly, the contribution of SSBM lies in its role as an input-configuration-specific interface for the subsequent Ghost–Swin encoder, rather than in treating channel or spatial attention as an isolated enhancement operation.

The complete SSBM configuration shown in Figure 2 is adopted for the fourteen-channel Landslide4Sense dataset. For the three-channel Jiuzhai Valley dataset and the four-channel HR-GLDD dataset, the spatial-attention operation is bypassed, while spectral/channel attention and band mixing remain active.

Given an input feature $\mathbf{X} \in \mathbb{R}^{C_0 \times H \times W}$, the importance of individual spectral channels is first estimated using the channel-attention operation defined in Equation (8):

$$\mathbf{M}_c = \sigma(f_c(\mathrm{GAP}(\mathbf{X})) + f_c(\mathrm{GMP}(\mathbf{X}))), \tag{8}$$

where GAP$(\cdot)$ and GMP$(\cdot)$ denote global average pooling and global max pooling, respectively. The function $f_c(\cdot)$ is implemented using two 1×1 convolutions with GELU activation; this shared transformation is labeled *MLP* in Figure 2. The symbol $\sigma(\cdot)$ denotes the Sigmoid function.

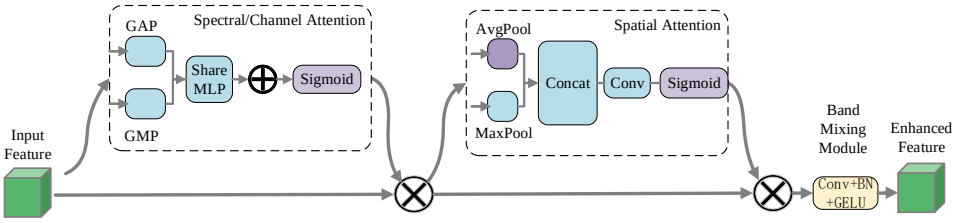


Figure 2. Structure of SSBM as the input-recalibration interface. Spectral/channel attention regulates informative channel responses, spatial attention is activated for the fourteen-channel input, and band mixing provides lightweight cross-channel interaction.

Based on the channel attention map $\mathbf{M}_c$, the channel-recalibrated feature is obtained using Equation (9):

$$\mathbf{X}_c = \mathbf{M}_c \odot \mathbf{X}, \tag{9}$$

where $\odot$ denotes element-wise multiplication. Through the operations in Equations (8) and (9), informative spectral responses are enhanced, while less relevant channels are suppressed.

For the fourteen-channel Landslide4Sense input, spatial attention is further employed to highlight landslide-related regions, including scarps, boundaries, and texture-discontinuous areas. The generation of the spatial attention map and the corresponding spatial feature refinement are formulated in Equations (10) and (11):

$$\mathbf{M}_s = \sigma(f_s([\text{AvgPool}_c(\mathbf{X}_c); \text{MaxPool}_c(\mathbf{X}_c)])), \tag{10}$$

$$\mathbf{X}_s = \mathbf{M}_s \odot \mathbf{X}_c. \tag{11}$$

In Equations (10) and (11), $\text{AvgPool}_c(\cdot)$ and $\text{MaxPool}_c(\cdot)$ denote average pooling and max pooling along the channel dimension, respectively. The notation $[\cdot;\cdot]$ represents channel-wise concatenation, and $f_s(\cdot)$ denotes a convolution with a kernel size of 7.

Although channel and spatial attention can select discriminative bands and spatial regions, these operations mainly perform feature reweighting and do not explicitly model interactions among different input channels. Therefore, lightweight cross-band interaction is performed using the band-mixing operation defined in Equation (12):

$$\mathbf{X}_0 = \delta\Big(\text{BN}\Big(\mathbf{W}_b^{1\times1} * \mathbf{X}_s\Big)\Big), \tag{12}$$

where $\mathbf{W}_b^{1\times1}$ denotes a 1×1 convolution kernel, BN$(\cdot)$ denotes batch normalization, $\delta(\cdot)$ denotes GELU activation, and $*$ denotes convolution. The 1×1 convolution mixes information along the channel dimension while maintaining the original spatial resolution and channel number. Therefore, the SSBM output satisfies $\mathbf{X}_0 \in \mathbb{R}^{C_0 \times H \times W}$.

The complete SSBM operation for the fourteen-channel Landslide4Sense input is summarized in Equation (13):

$$\mathbf{X}_0 = \mathcal{B}(\mathcal{A}_s(\mathcal{A}_c(\mathbf{X}))), \tag{13}$$

where $\mathcal{A}_c(\cdot)$, $\mathcal{A}_s(\cdot)$, and $\mathcal{B}(\cdot)$ denote channel/spectral attention, spatial attention, and band mixing, respectively.

For Jiuzhai Valley and HR-GLDD, the spatial-attention component is replaced by the identity mapping given in Equation (14):

$$\mathcal{A}_s(\mathbf{X}_c) = \mathbf{X}_c. \tag{14}$$

Therefore, Equations (13) and (14) describe the complete and simplified SSBM configurations, respectively.

3.2.3. Ghost–Swin Parallel Feature Encoder

Accurate landslide segmentation requires both fine-grained local details and high-level semantic context. Shallow texture, edge, and boundary features are important for delineating irregular landslide contours, whereas deeper contextual features are necessary for distinguishing landslides from spectrally similar background regions. To capture these complementary representations, LGSTNet employs a Ghost–Swin parallel feature encoder.

As shown in Figure 3, the SSBM-enhanced feature $\mathbf{X}_0$ is simultaneously fed into a two-module Ghost branch and a compact three-stage Swin Transformer branch. The Stage-3 Swin feature is subsequently enhanced by LightDSCA within the encoder. The complete parallel feature encoding process is summarized in Equation (15):

$$\begin{aligned} \mathbf{F}_1 &= \mathcal{G}_2(\mathcal{G}_1(\mathbf{X}_0)), \\ \left(\mathbf{F}_2, \mathbf{F}_3, \widetilde{\mathbf{F}}_4\right) &= \mathcal{S}(\mathbf{X}_0), \\ \mathbf{F}_4 &= \mathcal{D}\left(\widetilde{\mathbf{F}}_4\right). \end{aligned} \tag{15}$$

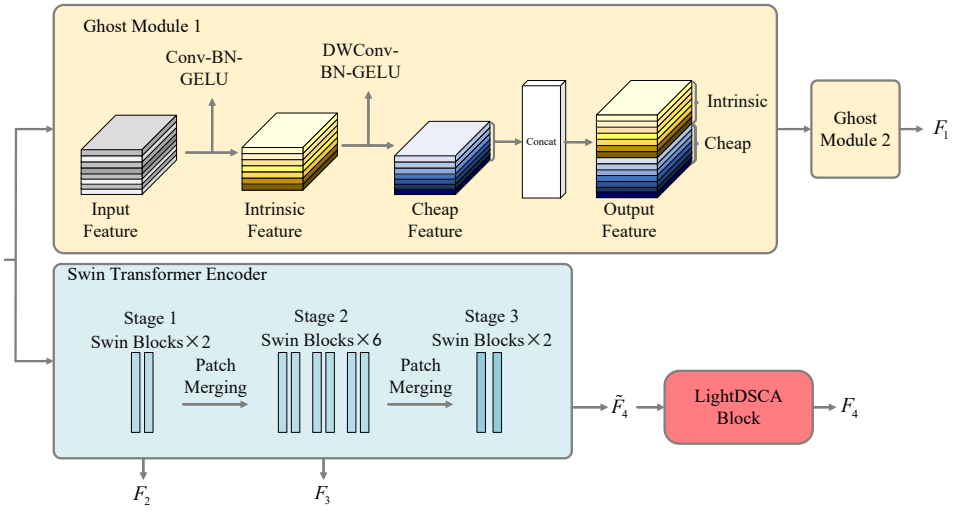


Figure 3. Structure of the Ghost–Swin parallel feature encoder. Two cascaded Ghost modules preserve shallow local details. The three-stage Swin Transformer extracts hierarchical semantic representations, and the Stage-3 output is enhanced by LightDSCA to generate the final deep encoder feature $\mathbf{F}_4$.

The dimensions of the four encoder output features are defined in Equation (16):

$$\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}, \qquad i = 1, 2, 3, 4. \tag{16}$$

As indicated in Equations (15) and (16), the encoder simultaneously provides a shallow local representation and three hierarchical semantic representations for subsequent multi-level decoding.

Ghost branch.

The Ghost branch is introduced to reduce redundant convolutional computation while retaining useful low-level spatial structures. The feature-generation process of the j-th Ghost module is described in Equations (17)–(19).

First, a standard convolution is used to generate a set of intrinsic features, as formulated in Equation (17):

$$\mathbf{I}^{(j)} = \delta\Big(\mathrm{BN}\Big(\mathbf{W}_p^{(j)} * \mathbf{U}^{(j-1)}\Big)\Big), \qquad j = 1, 2, \tag{17}$$

where $\mathbf{U}^{(0)} = \mathbf{X}_0$ and $\mathbf{W}_p^{(j)}$ denotes the standard convolution kernel of the j-th Ghost module.

Computationally inexpensive features are then generated from the intrinsic features using the depthwise convolution in Equation (18):

$$\mathbf{Q}^{(j)} = \delta\Big(\mathrm{BN}\Big(\mathbf{W}_q^{(j)} \circledast \mathbf{I}^{(j)}\Big)\Big), \tag{18}$$

where $\mathbf{W}_q^{(j)}$ denotes the depthwise convolution kernel, and $\circledast$ denotes depthwise convolution.

Finally, the intrinsic and cheap features are concatenated according to Equation (19):

$$\mathbf{U}^{(j)} = \Big[\mathbf{I}^{(j)}; \mathbf{Q}^{(j)}\Big], \qquad \mathbf{F}_1 = \mathbf{U}^{(2)}. \tag{19}$$

As summarized in Equations (17)–(19), the Ghost branch generates additional feature maps through inexpensive depthwise operations instead of standard convolutions. In the implementation, the two Ghost modules produce 32 and 64 output channels, respectively. The resulting feature $\mathbf{F}_1$ preserves shallow texture, edge, and boundary information with limited computational overhead.

Swin Transformer branch.

In parallel with the Ghost branch, a compact three-stage Swin Transformer is employed to extract hierarchical semantic representations. Unlike global self-attention, which has quadratic complexity with respect to the number of input tokens, the Swin Transformer restricts self-attention computation to non-overlapping local windows and introduces shifted windows between consecutive blocks to establish cross-window connections.

The input feature is first converted into a sequence of patch tokens using the patch-embedding operation defined in Equation (20):

$$\mathbf{Z}_0 = \mathcal{P}_{\mathrm{emb}}(\mathbf{X}_0), \tag{20}$$

where $\mathcal{P}_{\mathrm{emb}}(\cdot)$ denotes patch partition and linear embedding.

The hierarchical feature-extraction process of the compact three-stage Swin Transformer is formulated in Equation (21):

$$\begin{aligned} \mathbf{Z}_1 &= \mathcal{S}_1(\mathbf{Z}_0), & \mathbf{F}_2 &= \mathcal{T}(\mathbf{Z}_1), \\ \mathbf{Z}_2 &= \mathcal{S}_2\big(\mathcal{P}_{\mathrm{merge}}(\mathbf{Z}_1)\big), & \mathbf{F}_3 &= \mathcal{T}(\mathbf{Z}_2), \\ \mathbf{Z}_3 &= \mathcal{S}_3\big(\mathcal{P}_{\mathrm{merge}}(\mathbf{Z}_2)\big), & \widetilde{\mathbf{F}}_4 &= \mathcal{T}(\mathbf{Z}_3), \end{aligned} \tag{21}$$

where $\mathcal{S}_i(\cdot)$ denotes the i-th Swin Transformer stage, $\mathcal{P}_{\mathrm{merge}}(\cdot)$ denotes patch merging, and $\mathcal{T}(\cdot)$ rearranges the token sequence into a channel-first feature map.

Between two adjacent stages, the spatial resolution is reduced and the channel dimension is increased using the patch-merging operation defined in Equation (22):

$$\mathcal{P}_{\mathrm{merge}}(\mathbf{Z}) = \mathbf{W}_m \mathrm{LN}\Big(\Big[\mathbf{Z}^{0,0}; \mathbf{Z}^{1,0}; \mathbf{Z}^{0,1}; \mathbf{Z}^{1,1}\Big]\Big), \tag{22}$$

where $\mathbf{Z}^{a,b}$ denotes the token subset sampled with the spatial offset (a, b), $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathbf{W}_m$ denotes a learnable linear projection. The patch-merging operation in Equation (22) halves the feature height and width while increasing its channel capacity.

Within each Swin stage, window-based multi-head self-attention (W-MSA) and shifted-window multi-head self-attention (SW-MSA) are alternately applied. The computation of two consecutive Swin Transformer blocks is formulated in Equation (23):

$$\begin{aligned} \widehat{\mathbf{Z}}^{l} &= \mathrm{W{-}MSA}\left(\mathrm{LN}(\mathbf{Z}^{l-1})\right) + \mathbf{Z}^{l-1}, \\ \mathbf{Z}^{l} &= \mathrm{MLP}\left(\mathrm{LN}\left(\widehat{\mathbf{Z}}^{l}\right)\right) + \widehat{\mathbf{Z}}^{l}, \\ \widehat{\mathbf{Z}}^{l+1} &= \mathrm{SW{-}MSA}\left(\mathrm{LN}(\mathbf{Z}^{l})\right) + \mathbf{Z}^{l}, \\ \mathbf{Z}^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\widehat{\mathbf{Z}}^{l+1}\right)\right) + \widehat{\mathbf{Z}}^{l+1}. \end{aligned} \tag{23}$$

According to Equation (23), layer normalization is applied before each attention or MLP operation, and residual connections are used after each operation. W-MSA computes self-attention independently within non-overlapping local windows, whereas SW-MSA shifts the window partition in the next block to enable information exchange across neighboring windows.

For each local window, the scaled dot-product attention is computed according to Equation (24):

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}} + \mathbf{B} + \mathbf{M}\right)\mathbf{V}, \tag{24}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, respectively. The variable d denotes the feature dimension of each attention head, $\mathbf{B}$ denotes the learnable relative position bias, and $\mathbf{M}$ denotes the attention mask used in the shifted-window configuration. For W-MSA, $\mathbf{M}$ is set to zero.

Together, Equations (20)–(24) describe the complete hierarchical feature-extraction process of the Swin Transformer branch. The adopted encoder contains three stages with block depths $[2, 6, 2]$ and attention-head numbers $[2, 4, 8]$. The initial embedding dimension is set to 64, resulting in channel dimensions $C_2 = 64$, $C_3 = 128$, and $C_4 = 256$. Compared with a standard four-stage Swin Transformer, the compact three-stage configuration reduces the parameter count and computational overhead while retaining hierarchical contextual modeling capability.

LightDSCA deep-context enhancement.

Although the Stage-3 Swin feature $\widetilde{\mathbf{F}}_4$ contains rich high-level semantic information, a single deep receptive field may still be insufficient for landslides with irregular boundaries, elongated structures, and large variations in spatial extent. Therefore, LightDSCA is incorporated into the final stage of the encoder to enhance the deepest semantic feature before decoding.

The design emphasis of LightDSCA is compressed joint context aggregation rather than dilated convolution alone. Motivated by the spatial-context aggregation strategy of DSSA in SCDUNet++ [37], LightDSCA operates only on the deepest Stage-3 Swin feature and jointly models three complementary forms of context: multi-scale receptive fields, height–width directional dependencies, and image-level semantics.

Unlike directly retaining full-width parallel context branches, each LightDSCA branch is compressed to one quarter of the input channel dimension before concatenation and residual projection. This compressed organization limits channel expansion and allows

contextual enhancement to be inserted into the compact three-stage Transformer encoder without introducing a heavy additional encoding or decoding stage.

The detailed structure of LightDSCA is illustrated in Figure 4, and its complete enhancement process is described in Equations (25)–(28). Given the raw Stage-3 Swin feature $\widetilde{\mathbf{F}}_4 \in \mathbb{R}^{C_4 \times H_4 \times W_4}$, each contextual branch is compressed to an intermediate channel dimension $C_d = C_4/4$. In this study, $C_4 = 256$ and $C_d = 64$.

First, multi-scale contextual features are extracted using three dilated convolution branches with dilation rates $r \in \{1, 2, 3\}$, as formulated in Equation (25):

$$\mathbf{B}_r = \delta\left(\mathrm{BN}\left(\mathbf{W}_r^{3\times 3} *_r \widetilde{\mathbf{F}}_4\right)\right), \qquad r \in \{1, 2, 3\}, \tag{25}$$

where $*_r$ denotes a 3×3 convolution with dilation rate r, and $\mathbf{B}_r \in \mathbb{R}^{C_d \times H_4 \times W_4}$.

To capture anisotropic spatial dependencies along terrain structures and landslide boundaries, height-wise and width-wise directional pooling are employed. The directional spatial context is calculated according to Equation (26):

$$\begin{aligned} \mathbf{B}_h &= \mathrm{Up}_{H_4,W_4}\left(\mathbf{W}_h^{1\times 1} * \mathrm{Pool}_h\left(\widetilde{\mathbf{F}}_4\right)\right), \\ \mathbf{B}_w &= \mathrm{Up}_{H_4,W_4}\left(\mathbf{W}_w^{1\times 1} * \mathrm{Pool}_w\left(\widetilde{\mathbf{F}}_4\right)\right), \\ \mathbf{B}_s &= \delta(\mathrm{BN}(\mathbf{B}_h \oplus \mathbf{B}_w)). \end{aligned} \tag{26}$$

In Equation (26), $\mathrm{Pool}_h(\cdot)$ and $\mathrm{Pool}_w(\cdot)$ denote directional average pooling along the height and width dimensions, respectively. The operator $\mathrm{Up}_{H_4,W_4}(\cdot)$ denotes bilinear interpolation to the spatial size $H_4 \times W_4$, while $\oplus$ denotes element-wise addition.

In addition to multi-scale and directional spatial information, an image-level semantic representation is captured by the global context branch defined in Equation (27):

$$\mathbf{B}_g = \mathrm{Expand}_{H_4,W_4}\left(\delta\left(\mathrm{BN}\left(\mathbf{W}_g^{1\times 1} * \mathrm{GAP}\left(\widetilde{\mathbf{F}}_4\right)\right)\right)\right), \tag{27}$$

where $\mathrm{Expand}_{H_4,W_4}(\cdot)$ denotes spatial broadcasting to $H_4 \times W_4$.

Finally, the multi-scale, directional, and global contextual features are concatenated and projected back to C_4 channels. The final deep encoder feature is obtained through the residual fusion operation defined in Equation (28):

$$\mathbf{F}_4 = \widetilde{\mathbf{F}}_4 \oplus \delta\left(\mathrm{BN}\left(\mathbf{W}_d^{1\times 1} * \left[\mathbf{B}_1; \mathbf{B}_2; \mathbf{B}_3; \mathbf{B}_s; \mathbf{B}_g\right]\right)\right). \tag{28}$$

As shown in Equations (25)–(28), LightDSCA combines multi-scale, directional, and global contextual information. The residual connection in Equation (28) preserves the original Stage-3 semantic representation while incorporating complementary contextual cues from different receptive fields and spatial directions. As a result, the enhanced feature $\mathbf{F}_4$ provides richer semantic and spatial-context information for subsequent multi-level decoding.

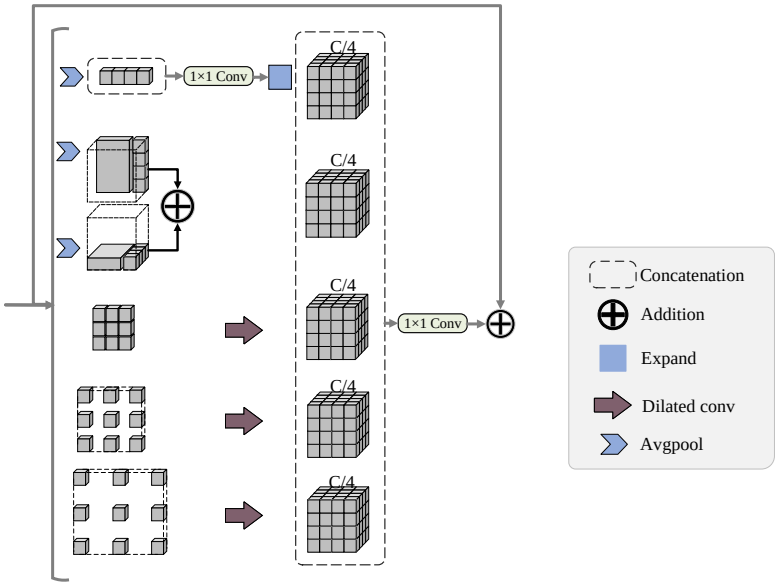


Figure 4. Structure of the LightDSCA block incorporated into the Ghost–Swin parallel encoder. Multi-scale dilated convolution, directional pooling, and global context branches are concatenated and fused, followed by residual addition with the original Stage-3 Swin feature.

3.2.4. Lightweight Projection–Fusion Decoder

After local–global encoding and compressed deep-context enhancement, LGSTNet employs a lightweight projection–fusion decoder to integrate the four encoder features. The decoding path combines common-dimensional feature projection, spatial alignment, depthwise-separable fusion, and residual refinement.

Specifically, the four blocks labeled *MLP* in Figure 1 implement the feature projections defined in Equation (29). The labels Up-sampling and *Concatenation* correspond to the spatial alignment and feature assembly in Equation (30). The block labeled *Fuse MLP* denotes the depthwise-separable fusion in Equation (31), whereas *Refine* denotes the residual refinement in Equation (32).

Compared with progressive or nested convolutional decoders, this design avoids repeated high-dimensional feature reconstruction. Each encoder feature is first projected to a shared embedding dimension, after which all features are aligned at a common spatial scale and jointly fused. Depthwise separable operations are then used to refine spatial and cross-channel information with limited computational overhead.

The decoder receives the four encoder features $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, and $\mathbf{F}_4$. Here, $\mathbf{F}_1$ is generated by the Ghost branch, $\mathbf{F}_2$ and $\mathbf{F}_3$ are generated by Swin Transformer Stages 1 and 2, and $\mathbf{F}_4$ is the LightDSCA-enhanced Stage-3 feature. The complete decoding process is formulated in Equations (29)–(33).

First, each encoder feature is projected into a common embedding dimension C_e using a 1×1 convolution, batch normalization, and GELU activation, as defined in Equation (29):

$$\mathbf{Z}_i = \delta\left(\mathrm{BN}\left(\mathbf{W}_i^{1\times 1} * \mathbf{F}_i\right)\right), \qquad i = 1, 2, 3, 4, \tag{29}$$

where $\mathbf{Z}_i \in \mathbb{R}^{C_e \times H_i \times W_i}$. In this study, $C_e = 128$.

Since the projected features have different spatial resolutions, $\mathbf{Z}_2$, $\mathbf{Z}_3$, and $\mathbf{Z}_4$ are upsampled to the spatial resolution of $\mathbf{Z}_1$ and concatenated according to Equation (30):

$$\mathbf{F}_{\mathrm{cat}} = \left[\mathbf{Z}_1; \mathrm{Up}_{H_1,W_1}(\mathbf{Z}_2); \mathrm{Up}_{H_1,W_1}(\mathbf{Z}_3); \mathrm{Up}_{H_1,W_1}(\mathbf{Z}_4)\right], \tag{30}$$

where $\mathbf{F}_{\text{cat}} \in \mathbb{R}^{4C_e \times H_1 \times W_1}$. The alignment operation in Equation (30) allows shallow structural details and deep semantic information to be aggregated at a common spatial scale.

After spatial alignment, the concatenated feature is efficiently fused using the depthwise separable operation defined in Equation (31):

$$\mathbf{F}_f = \mathcal{D}_p\Big(\delta\Big(\mathrm{BN}\Big(\mathbf{W}_{f,pw}^{1\times1} * \Big(\mathbf{W}_{f,dw}^{3\times3} \circledast \mathbf{F}_{\text{cat}}\Big)\Big)\Big)\Big), \tag{31}$$

where $\mathbf{W}_{f,dw}^{3\times3}$ denotes the depthwise convolution kernel, $\mathbf{W}_{f,pw}^{1\times1}$ denotes the pointwise convolution kernel, $\circledast$ denotes depthwise convolution, and $\mathcal{D}_p(\cdot)$ denotes dropout with probability $p = 0.2$. The operation in Equation (31) first aggregates spatial information independently within each channel and then performs cross-channel feature fusion.

The fused feature is subsequently enhanced using the residual refinement operation in Equation (32):

$$\mathbf{F}_r = \mathbf{F}_f + \mathbf{W}_{r,pw}^{1\times1} * \delta\Big(\mathrm{BN}\Big(\mathbf{W}_{r,dw}^{3\times3} \circledast \mathbf{F}_f\Big)\Big), \tag{32}$$

where the first term preserves the original fused representation and the second term further refines its local spatial details and channel responses.

Finally, the refined feature is converted into segmentation logits using a 1×1 classification layer and bilinearly upsampled to the original image resolution according to Equation (33):

$$\hat{\mathbf{Y}} = \mathrm{Up}_{H,W}\Big(\mathbf{W}_o^{1\times1} * \mathbf{F}_r\Big), \tag{33}$$

where $\mathrm{Up}_{H,W}(\cdot)$ denotes bilinear interpolation to the original spatial size $H \times W$.

Overall, Equations (29)–(33) describe how the decoder progressively projects, aligns, fuses, and refines the multi-level encoder features before generating the final landslide segmentation result. This lightweight decoder combines the boundary-sensitive local representation of the Ghost branch with the hierarchical contextual representations of the Swin Transformer branch.

3.3. Loss Function

Landslide segmentation is commonly affected by severe foreground–background imbalance because landslide pixels usually occupy only a small proportion of remote sensing images. When the model is optimized solely using a conventional pixel-wise classification loss, the optimization process may be dominated by background pixels, leading to incomplete landslide extraction, fragmented foreground regions, and weak boundary recovery.

To alleviate this problem, LGSTNet is optimized using a compound segmentation loss that combines Lovász-Softmax loss and Tversky loss [51,52]. The overall training objective is defined in Equation (34):

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Lovasz}} + \lambda_2 \mathcal{L}_{\text{Tversky}}, \tag{34}$$

where λ_1 and λ_2 denote the weights of the Lovász-Softmax and Tversky loss terms, respectively. In this study, both weights are set to 1.

The Lovász-Softmax loss is a differentiable surrogate for directly optimizing the intersection-over-union criterion. Compared with a purely pixel-wise classification loss, it is more consistent with the IoU-based evaluation protocol used in semantic segmentation. It therefore improves the region-level overlap between the predicted landslide regions and the ground-truth masks, particularly when foreground regions are small, spatially fragmented, or irregularly shaped.

The Tversky loss is further introduced to address class imbalance and control the trade-off between false positives and false negatives. The calculation of the Tversky loss is described in Equations (35)–(36). The Tversky index is first defined in Equation (35):

$$\mathrm{TI} = \frac{\mathrm{TP} + \epsilon}{\mathrm{TP} + \alpha \mathrm{FP} + \beta \mathrm{FN} + \epsilon}, \tag{35}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The constant ϵ is used to avoid numerical instability.

The corresponding Tversky loss is obtained from the Tversky index according to Equation (36):

$$\mathcal{L}_{\mathrm{Tversky}} = 1 - \mathrm{TI}. \tag{36}$$

Following the setting in SCDUNet++ [37], α and β are set to 0.3 and 0.7, respectively. Since β is larger than α, false negatives receive a higher penalty than false positives during optimization. This setting encourages the network to recover more complete landslide regions and reduces the omission of foreground pixels.

Therefore, Equations (34)–(36) define the complete optimization objective used to train LGSTNet. By combining Lovász-Softmax loss and Tversky loss, the proposed training objective jointly improves region-level overlap, foreground completeness, and segmentation robustness under highly imbalanced foreground–background distributions.

4. Results and Discussion

4.1. Experimental Settings

4.1.1. Implementation Details and Experimental Environment

All experiments were conducted on the AutoDL cloud platform. The software environment consisted of PyTorch 2.0.0, Python 3.8, Ubuntu 20.04, and CUDA 11.8. The hardware platform included an NVIDIA RTX 3090 GPU with 24 GB of memory, a 20-core vCPU, 90 GB of RAM, and a 30 GB system disk. To ensure consistent efficiency evaluation, all FPS values for each dataset were measured using the same hardware platform, input resolution, and evaluation pipeline. FPS was evaluated with a batch size of 1. Before timing, each model was warmed up for 10 iterations. For each selected checkpoint, the complete test-set evaluation was timed three times, with CUDA synchronization performed immediately before and after each timed run. The reported FPS was obtained by averaging the timing results over the three independently trained checkpoints. For the fourteen-channel Landslide4Sense comparison, parameter counts and FLOPs were additionally measured for all seven models under the same $14 \times 128 \times 128$ input configuration and are reported together with FPS.

To provide a controlled comparison, all baseline models, LGSTNet, and its replacement variants were retrained under a unified experimental protocol within each dataset. Specifically, for a given dataset, all models used the same data split, input resolution, dataset-specific preprocessing and augmentation pipeline, loss function, optimizer, learning-rate schedule, number of training epochs, and model-selection criterion. The preprocessing and augmentation settings differed across datasets only to accommodate their different data formats and channel configurations, and were held fixed across all compared models on the same dataset. This protocol improves consistency across the evaluated models, although it may not represent the individually optimized performance of every baseline. For the Jiuzhai Valley dataset, the random seed used for dataset splitting was fixed at 42 to improve reproducibility.

No external pretrained weights were used for the baseline networks in the reported comparisons. Instead, the models were trained from scratch under the unified protocol

described above. For the multi-channel datasets, the input layers of the applicable baseline models were instantiated directly with the native channel dimensionality of each dataset, namely $C_{in} = 3$, 4, and 14 for Jiuzhai Valley, HR-GLDD, and Landslide4Sense, respectively. Thus, the four- and fourteen-channel inputs were not reduced to RGB, and no RGB-pretrained input layer was expanded by channel replication or weight averaging. Because the models were not initialized from RGB pretrained weights, no separate initialization rule was required for additional channels; all input-channel weights were initialized together with the remaining model parameters according to the default or implementation-specific initialization scheme of each network.

The baseline configurations used in the experiments were as follows. SegFormer employed a four-stage hierarchical Transformer encoder with embedding dimensions [64, 128, 320, 512], stage depths [3, 4, 6, 3], attention heads [1, 2, 5, 8], spatial-reduction ratios [8, 4, 2, 1], and a decoder embedding dimension of 256. SwinUNet used a patch size of 4, an embedding dimension of 96, encoder depths [2, 2, 6, 2], attention heads [3, 6, 12, 24], a window size of 7, and a drop-path rate of 0.1. STUNet used a base channel width of 64, encoder widths of 64, 128, 256, and 512, a 1024-channel bottleneck, bilinear upsampling, skip connections, and sequential channel and spatial recalibration at the bottleneck. SCDUNet++ used a Swin Transformer branch with an embedding dimension of 64, depths [2, 6, 2], attention heads [2, 4, 8], and a window size of 8, together with a nested UNet++ decoder and a spatial–spectral aggregation block composed of dilated convolutional branches and spatial/spectral pooling. TransUNet used an R50–ViT-B/16 configuration with a ResNetV2 hybrid encoder, a 768-dimensional Transformer representation, an MLP dimension of 3072, 12 attention heads, 12 Transformer layers, and ResNet block depths (3, 4, 9). UNet++ used a base feature width of 16, a four-level nested dense decoder, GroupNorm-based residual blocks, shallow CBAM modules, a lightweight PVT-SRA bottleneck with latent dimension 128, one PVT block, four attention heads, a spatial-reduction ratio of 2, a detail-aggregation branch, and learnable softmax-weighted output fusion.

All models were optimized using AdamW 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, with the minimum learning rate set to 9×10^{-6}. The models were trained for 80 epochs with a batch size of 8. Each model was independently trained three times from scratch under the same experimental protocol. Model performance was evaluated on the validation set after each epoch, and the checkpoint with the lowest validation loss from each independent run was retained for final test-set evaluation. Accuracy, Recall, F1-score, and mIoU were also monitored during training but were not used as the primary checkpoint-selection criterion. For the main quantitative comparisons, segmentation metrics are reported as mean $\pm$ standard deviation (%) across the three independent training runs, while FPS is reported under the same inference setting.

Regarding input recalibration, the complete SSBM configuration, consisting of channel/spectral attention, spatial attention, and band mixing, was adopted for the fourteen-channel Landslide4Sense dataset. For the three-channel Jiuzhai Valley dataset and the four-channel HR-GLDD dataset, the spatial-attention branch was bypassed, while channel/spectral attention and band mixing remained active.

4.1.2. Evaluation Metrics

Accuracy, Precision, Recall, F1-score, IoU, mIoU, and MCC are used to evaluate the segmentation performance. The definitions of these metrics are given in Equations (37)–(43).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{37}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{38}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{39}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{40}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{41}$$

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k + \text{FN}_k}, \tag{42}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{43}$$

In Equations (37)–(43), TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, and K denotes the number of classes.

In this study, IoU refers to the landslide-class IoU, whereas mIoU is computed by averaging the IoU values of the background and landslide classes. Since landslide pixels usually occupy only a small proportion of remote sensing images, F1-score, landslide-class IoU, and MCC are considered more informative than overall accuracy for evaluating foreground segmentation quality.

4.2. Quantitative Comparison with Representative Segmentation Networks

To comprehensively evaluate segmentation performance and inference efficiency, six representative models were selected for comparison: UNet++ [16], SegFormer [20], SwinUNet [22], STUNet [53], SCDUNet++ [37], and TransUNet [23]. These baselines cover nested convolutional segmentation, lightweight Transformer segmentation, Swin-based segmentation, CNN–Transformer hybrid segmentation, and enhanced U-Net-style designs. The comparison therefore evaluates both segmentation accuracy and inference efficiency.

4.2.1. Comparative Results on the Fourteen-Channel Landslide4Sense Dataset

Table 1 shows that LGSTNet achieves the best F1, IoU, mIoU, and MCC on Landslide4Sense, reaching 64.01%, 47.07%, 72.81%, and 63.39%, respectively, at 63.93 FPS. Compared with SCDUNet++, LGSTNet improves F1, IoU, mIoU, and MCC by 3.12, 3.30, 1.74, and 3.10 percentage points, respectively, while FPS increases from 57.24 to 63.93. Compared with TransUNet, LGSTNet has slightly lower Acc. and Precision, but achieves higher Recall, F1, IoU, mIoU, and MCC. Since landslide pixels account for a small proportion of the image, F1, IoU, and MCC are more informative than overall accuracy alone. Table 2 further compares model complexity under the same fourteen-channel 128 × 128 input configuration. SCDUNet++ has the smallest complexity, with 1.93 M parameters and 0.92 G FLOPs. LGSTNet requires 3.79 M parameters and 2.39 G FLOPs, which is lower than TransUNet (93.23 M, 8.05 G), STUNet (34.66 M, 16.52 G), SwinUNet (34.26 M, 3.32 G), SegFormer (26.84 M, 3.51 G), and UNet++ (9.16 M, 9.39 G). Together with the FPS results in Table 1, these results indicate a favorable accuracy–complexity–speed trade-off rather than absolute superiority in every efficiency metric. The strong Landslide4Sense result is interpreted together with the within-dataset input-configuration experiment reported below; the cross-dataset comparison by itself is not used to attribute the advantage solely to channel number or modality richness.

Table 1. Comparison results on the fourteen-channel Landslide4Sense dataset. Segmentation metrics are reported as mean ± standard deviation (%) over three independent training runs. Bold values indicate the best result for each reported metric.

	Efficiency	**Segmentation Metrics (%)**						
Method	**FPS**	**Acc.**	**Precision**	**Recall**	**F1**	**IoU**	**mIoU**	**MCC**
UNet++	76.09	98.25 ± 0.05	53.26 ± 1.26	59.38 ± 1.53	56.15 ± 0.33	39.04 ± 0.32	68.63 ± 0.16	55.35 ± 0.31
SegFormer	52.17	98.49 ± 0.06	59.42 ± 1.54	63.87 ± 1.78	61.56 ± 0.57	44.47 ± 0.60	71.47 ± 0.30	60.84 ± 0.67
SwinUNet	46.45	98.35 ± 0.06	55.69 ± 1.64	61.97 ± 1.47	58.66 ± 0.15	41.51 ± 0.17	69.92 ± 0.11	57.91 ± 0.20
STUNet	**103.15**	98.38 ± 0.02	56.32 ± 0.52	64.42 ± 0.87	60.10 ± 0.33	42.96 ± 0.34	70.66 ± 0.17	59.42 ± 0.32
SCDUNet++	57.24	98.39 ± 0.02	56.22 ± 0.53	66.41 ± 0.88	60.89 ± 0.31	43.77 ± 0.32	71.07 ± 0.16	60.29 ± 0.18
TransUNet	21.85	**98.57 ± 0.04**	**62.62 ± 1.25**	60.84 ± 1.70	61.72 ± 0.53	44.63 ± 0.55	71.60 ± 0.28	61.00 ± 0.53
LGSTNet	63.93	98.56 ± 0.04	60.46 ± 1.02	**68.01 ± 1.35**	**64.01 ± 0.42**	**47.07 ± 0.45**	**72.81 ± 0.22**	**63.39 ± 0.45**

Table 2. Model complexity comparison on the fourteen-channel Landslide4Sense dataset. Parameter counts and FLOPs are measured using the same 14 × 128 × 128 input configuration for all models. Bold values indicate the lowest parameter count and FLOPs.

Method	**Params (M)**	**FLOPs (G)**
UNet++	9.16	9.39
SegFormer	26.84	3.51
SwinUNet	34.26	3.32
STUNet	34.66	16.52
SCDUNet++	**1.93**	**0.92**
TransUNet	93.23	8.05
LGSTNet	3.79	2.39

4.2.2. Comparative Results on the Four-Channel HR-GLDD Dataset

Table 3 reports the quantitative comparison on HR-GLDD. LGSTNet achieves 93.44% Acc., 72.33% F1, 56.65% IoU, 74.74% mIoU, and 68.91% MCC. It should be noted that LGSTNet does not achieve absolute dominance on this dataset. TransUNet obtains the highest F1, IoU, mIoU, and MCC, while STUNet achieves both higher segmentation metrics and a higher FPS than LGSTNet. Therefore, the HR-GLDD results indicate competitive segmentation performance rather than overall superiority. Compared with SCDUNet++, LGSTNet shows slightly lower F1 and IoU but improves speed from 53.20 FPS to 63.63 FPS. This indicates that the proposed lightweight fusion design can reduce part of the computational burden while maintaining competitive segmentation quality. The relatively limited advantage on HR-GLDD is therefore treated as dataset-dependent behavior, since the cross-dataset comparison does not isolate channel composition from other differences in data characteristics and experimental conditions.

Table 3. Comparison results on the four-channel HR-GLDD dataset. Segmentation metrics are reported as mean ± standard deviation (%) over three independent training runs. Bold values indicate the best result for each reported metric.

	Efficiency	**Segmentation Metrics (%)**						
Method	**FPS**	**Acc.**	**Precision**	**Recall**	**F1**	**IoU**	**mIoU**	**MCC**
UNet++	72.92	93.26 ± 0.19	67.93 ± 1.12	72.46 ± 1.35	70.12 ± 0.41	53.99 ± 0.49	73.33 ± 0.24	66.37 ± 0.48
SegFormer	48.98	91.86 ± 0.34	59.90 ± 1.45	77.05 ± 1.62	67.40 ± 0.55	50.83 ± 0.63	70.97 ± 0.35	63.49 ± 0.63
SwinUNet	46.70	93.27 ± 0.31	66.74 ± 1.28	76.57 ± 1.48	71.32 ± 0.38	55.42 ± 0.45	74.04 ± 0.18	67.73 ± 0.42
STUNet	**96.39**	93.71 ± 0.25	**68.64 ± 0.95**	78.11 ± 1.22	73.07 ± 0.32	57.57 ± 0.38	75.35 ± 0.21	69.71 ± 0.35
SCDUNet++	53.20	93.45 ± 0.20	66.73 ± 1.05	79.71 ± 1.40	72.65 ± 0.45	57.05 ± 0.56	74.93 ± 0.26	69.31 ± 0.51
TransUNet	22.14	**93.80 ± 0.29**	68.44 ± 1.22	**80.29 ± 1.67**	**73.89 ± 0.49**	**58.60 ± 0.62**	**75.90 ± 0.39**	**70.69 ± 0.63**
LGSTNet	63.63	93.44 ± 0.22	67.05 ± 1.18	78.50 ± 1.55	72.33 ± 0.43	56.65 ± 0.53	74.74 ± 0.29	68.91 ± 0.46

4.2.3. Results on the Three-Channel Jiuzhai Valley Dataset

As shown in Table 4, on the Jiuzhai Valley dataset, LGSTNet achieves 96.07% Acc., 83.66% Precision, 83.09% Recall, 83.37% F1, 71.49% IoU, 83.56% mIoU, and 81.15% MCC at 31.53 FPS. Compared with UNet++, SegFormer, SwinUNet, and STUNet, LGSTNet obtains better or comparable performance in most major segmentation metrics, particularly in F1, IoU, mIoU, and MCC. Compared with SCDUNet++, LGSTNet achieves the same accuracy and a close F1-score, while improving the inference speed from 16.04 to 31.53 FPS. Although TransUNet achieves the best F1, IoU, mIoU, and MCC on this dataset, its inference speed is lower than that of LGSTNet. These results indicate that LGSTNet does not achieve absolute segmentation superiority on Jiuzhai Valley, but provides a competitive accuracy–efficiency trade-off for high-resolution RGB landslide segmentation.

Table 4. Comparison results on the three-channel Jiuzhai Valley dataset. Segmentation metrics are reported as mean ± standard deviation (%) over three independent training runs. Bold values indicate the best result for each reported metric.

	Efficiency	**Segmentation Metrics (%)**						
Method	**FPS**	**Acc.**	**Precision**	**Recall**	**F1**	**IoU**	**mIoU**	**MCC**
UNet++	27.88	93.56 ± 0.12	75.73 ± 0.85	67.23 ± 1.25	71.23 ± 0.62	55.31 ± 0.74	74.16 ± 0.31	67.77 ± 0.52
SegFormer	**37.74**	94.77 ± 0.25	75.80 ± 1.15	82.10 ± 1.42	78.82 ± 0.65	65.04 ± 0.89	79.62 ± 0.35	75.92 ± 0.61
SwinUNet	34.12	95.17 ± 0.21	78.45 ± 0.92	81.70 ± 1.35	80.04 ± 0.58	66.73 ± 0.81	80.69 ± 0.28	77.32 ± 0.55
STUNet	29.08	94.82 ± 0.30	74.86 ± 1.05	84.84 ± 1.50	79.54 ± 0.69	66.03 ± 0.95	80.14 ± 0.33	76.78 ± 0.62
SCDUNet++	16.04	96.07 ± 0.18	82.78 ± 0.75	84.40 ± 1.28	83.59 ± 0.71	71.80 ± 1.05	83.72 ± 0.51	81.36 ± 0.48
TransUNet	22.02	**96.29 ± 0.15**	83.49 ± 0.88	**85.70 ± 1.55**	**84.58 ± 0.72**	**73.28 ± 1.08**	**84.58 ± 0.53**	**82.48 ± 0.65**
LGSTNet	31.53	96.07 ± 0.24	**83.66 ± 0.95**	83.09 ± 1.40	83.37 ± 0.68	71.49 ± 1.00	83.56 ± 0.49	81.15 ± 0.58

Overall, LGSTNet exhibits dataset-dependent but complementary performance. On HR-GLDD, it provides competitive segmentation performance but does not surpass all baselines. On Jiuzhai Valley, it achieves segmentation quality close to SCDUNet++ and TransUNet while improving inference speed over these two models. On Landslide4Sense, it obtains the strongest foreground segmentation performance in terms of F1, IoU, mIoU, and MCC. However, the three datasets differ simultaneously in channel composition, spatial resolution, sample size, sensor characteristics, geographic context, and splitting strategy. Therefore, the cross-dataset comparison alone does not isolate the effect of input modality and should not be used to attribute the observed Landslide4Sense advantage solely to the additional multispectral and terrain-derived channels.

4.3. Controlled Input-Configuration Experiment on Landslide4Sense

To reduce the confounding effect of comparing different datasets, an additional within-dataset experiment was conducted exclusively on Landslide4Sense. Three input configurations were evaluated using the same Landslide4Sense train/validation/test split, preprocessing pipeline, training schedule, model-selection criterion, and evaluation procedure: (1) a three-channel RGB subset, (2) a four-channel RGB–NIR subset, and (3) the complete fourteen-channel Landslide4Sense input. LGSTNet and SegFormer were both evaluated under these three settings. For SegFormer, the architecture was unchanged apart from the dimensionality of the input projection. For LGSTNet, the input-recalibration policy followed the channel-configuration rule defined in Section 3.2.2: channel attention and band mixing were retained in all settings, whereas spatial attention was additionally active for the fourteen-channel configuration. Therefore, this experiment provides a controlled same-dataset comparison of the complete input configurations rather than isolating the causal effect of channel count alone.

Table 5 shows a consistent improvement as more informative Landslide4Sense input channels are made available. For LGSTNet, F1 increases from 42.87% under the RGB setting to 46.06% under the four-channel setting and 64.01% under the full fourteen-channel setting. The corresponding IoU values increase from 27.28% to 29.92% and 47.07%. SegFormer exhibits the same overall trend, with F1 increasing from 42.50% to 45.80% and 61.56%, and IoU increasing from 26.99% to 29.71% and 44.47%.

Table 5. Controlled input-configuration experiment on Landslide4Sense. All settings use the same data split, training protocol, and evaluation procedure. Results are reported as mean ± standard deviation (%) over three independent training runs. Bold values indicate the best result for each model across the three input configurations.

Input Configuration	C_{in}	LGSTNet				SegFormer			
		F1 (%)	IoU (%)	mIoU (%)	MCC (%)	F1 (%)	IoU (%)	mIoU (%)	MCC (%)
RGB	3	42.87 ± 1.30	27.28 ± 1.06	62.12 ± 0.80	43.32 ± 0.73	42.50 ± 1.30	26.99 ± 1.05	61.50 ± 0.60	42.30 ± 0.80
RGB–NIR	4	46.06 ± 1.10	29.92 ± 0.93	63.65 ± 0.45	45.95 ± 0.65	45.80 ± 1.10	29.71 ± 0.93	63.20 ± 0.50	45.60 ± 0.70
Full input	14	**64.01 ± 0.42**	**47.07 ± 0.45**	**72.81 ± 0.22**	**63.39 ± 0.45**	**61.56 ± 0.57**	**44.47 ± 0.60**	**71.47 ± 0.30**	**60.84 ± 0.67**

The relative difference between LGSTNet and SegFormer is small under the reduced-input settings but becomes larger for the full fourteen-channel configuration. LGSTNet exceeds SegFormer in F1 by 0.37, 0.26, and 2.45 percentage points under the three-, four-, and fourteen-channel settings, respectively; the corresponding IoU differences are 0.29, 0.21, and 2.60 percentage points. These results provide within-dataset evidence that both

models benefit from the fuller Landslide4Sense input and that LGSTNet maintains a larger advantage under the complete fourteen-channel configuration. Because the input content and, for LGSTNet, the associated input-recalibration configuration change together, the experiment is interpreted as evidence for input-configuration effectiveness rather than as proof that channel number alone causes the improvement.

4.4. Qualitative Comparison

To further evaluate segmentation behavior, qualitative comparisons were conducted on representative samples from the HR-GLDD, Jiuzhai Valley, and Landslide4Sense datasets. In addition to quantitative metrics, qualitative visual comparisons provide complementary evidence of boundary recovery, regional completeness, small-target preservation, and background suppression.

As shown in Figure 5, on Landslide4Sense, the advantage of LGSTNet becomes more evident. Landslide targets in this dataset are often small, fragmented, and easily confused with bare soil, riverbeds, and shadowed regions. UNet++ and STUNet can detect part of the landslide regions but may produce noisy predictions or miss thin branches. SegFormer and TransUNet show stronger semantic modeling ability, but local discontinuity may still appear when landslide regions are narrow or fragmented. In contrast, LGSTNet tends to produce more complete landslide masks and better preserve curved, strip-shaped, or scattered landslide structures in these representative examples. These visual observations are consistent with the higher Recall, F1, IoU, and MCC reported in Table 1.

As shown in Figure 6, on HR-GLDD, the visual differences among models are relatively moderate, consistent with the quantitative results. Most methods can identify the main landslide bodies, but their predictions differ in local completeness and noise suppression. UNet++ tends to preserve local responses but may introduce scattered false positives in background regions with similar spectral characteristics. SegFormer and TransUNet show stable semantic discrimination, but some boundaries are incomplete or over-smoothed. In the selected representative cases, LGSTNet tends to produce compact and continuous regions, especially for landslides with regular shapes or moderate spatial extent.

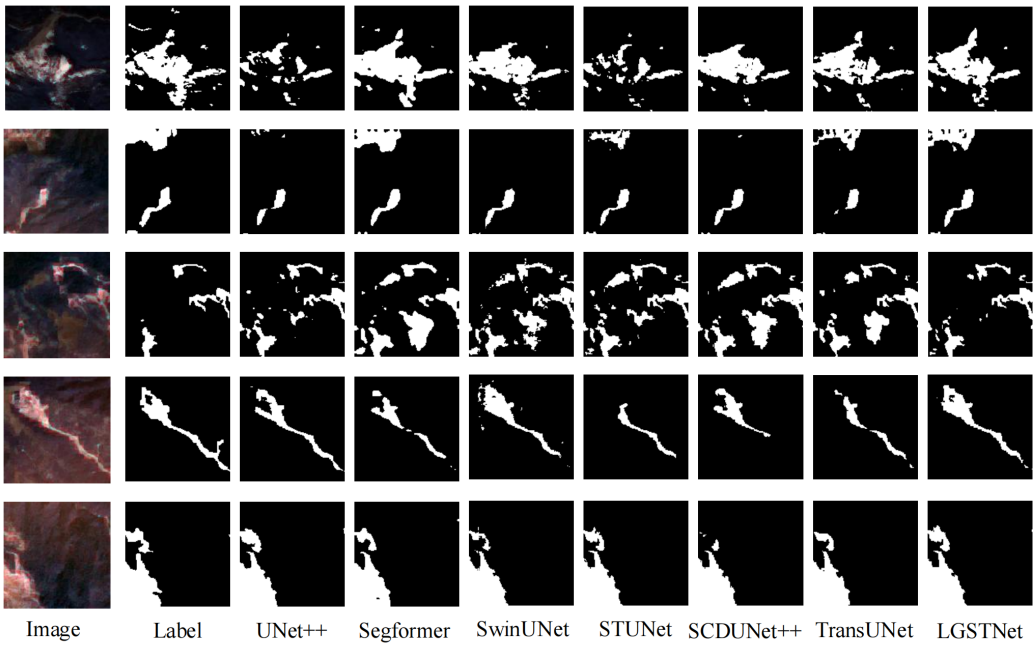


Figure 5. Examples of semantic segmentation results on the Landslide4Sense dataset.

As shown in Figure 7, on Jiuzhai Valley, landslide regions often present larger spatial extent, clearer texture variation, and more complex boundaries. LGSTNet tends to preserve the overall contour of landslide bodies in the selected cases. Compared with UNet++ and SwinUNet, it reduces over-segmentation in road-adjacent or exposed-surface areas. Compared with SegFormer, its predictions are less fragmented and better aligned with ground truth boundaries. GhostStem and multi-level feature fusion help maintain edge continuity and local structural details, which are important for high-resolution RGB landslide delineation. However, the quantitative results show that SCDUNet++ and TransUNet still achieve slightly higher segmentation metrics on this dataset, indicating that LGSTNet mainly provides a balanced accuracy–efficiency trade-off rather than absolute segmentation superiority.

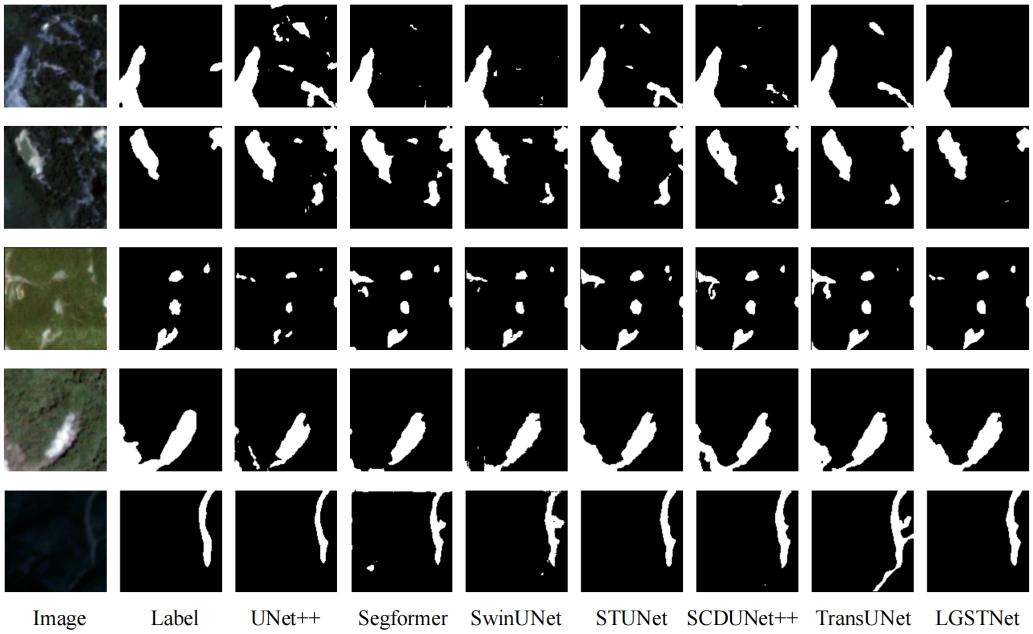


Figure 6. Examples of semantic segmentation results on the HR-GLDD dataset.

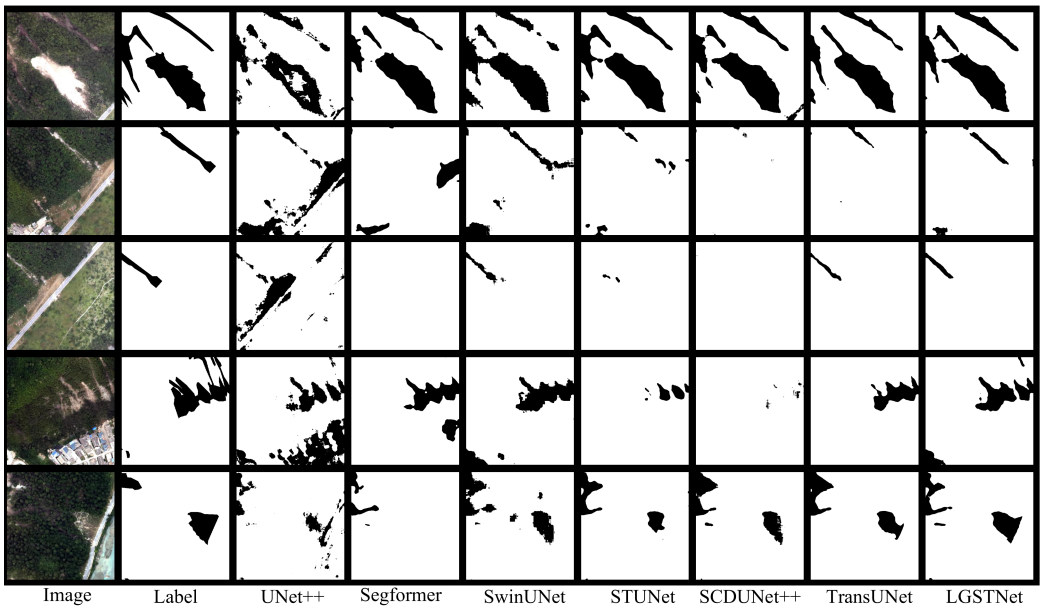


Figure 7. Examples of semantic segmentation results on the Jiuzhai Valley dataset. Black pixels denote landslide foreground and white pixels denote background, following the original annotation convention of the Jiuzhai Valley dataset.

Overall, the qualitative results indicate that LGSTNet can improve structural completeness, boundary continuity, and background suppression in many challenging cases. Strong qualitative performance is observed on the evaluated Landslide4Sense examples; however, these cross-dataset visual comparisons do not isolate the contribution of input modality from other dataset-specific factors.

4.5. Fine-Grained Component Ablation and Architectural Replacement Analysis

To clarify the contribution of the individual input-recalibration operations and the principal architectural components of LGSTNet, fine-grained ablation and controlled replacement experiments were conducted on Landslide4Sense and HR-GLDD. For each experiment, only one component was removed or replaced at a time while the remaining architecture, data split, preprocessing pipeline, training strategy, and evaluation protocol were kept unchanged.

For Landslide4Sense, the complete SSBM configuration contains channel attention, spatial attention, and band mixing. These three operations were independently removed to evaluate their individual effects. For HR-GLDD, spatial attention is not activated in the complete configuration; therefore, only channel attention and band mixing were independently ablated. GhostStem, LightDSCA, and the MLP projection–fusion decoder were additionally replaced by their corresponding alternatives to evaluate the contribution of the major architectural stages.

4.5.1. Results on the Fourteen-Channel Landslide4Sense Dataset

Table 6 shows that independently removing any of the three SSBM operations degrades the overall segmentation quality on Landslide4Sense. Removing channel attention decreases F1, IoU, mIoU, and MCC by 4.85, 5.07, 2.70, and 3.95 percentage points, respectively. Although Recall increases from 68.01% to 73.50%, Precision drops from 60.46% to 49.50%, indicating that channel attention mainly helps suppress irrelevant channel responses and reduce false-positive predictions rather than simply increasing foreground coverage.

Table 6. Fine-grained ablation and architectural replacement results on the fourteen-channel Landslide4Sense dataset. Segmentation metrics are reported as mean ± standard deviation (%) over three independent training runs.

Variant	Params (M)	FLOPs (G)	FPS	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	IoU (%)	mIoU (%)	MCC (%)
Full LGSTNet	3.79	2.39	63.93	98.56 ± 0.04	60.46 ± 1.02	68.01 ± 1.35	64.01 ± 0.42	47.07 ± 0.45	72.81 ± 0.22	63.39 ± 0.45
w/o Channel Attention	3.79	2.39	65.50	98.15 ± 0.05	49.50 ± 0.85	73.50 ± 0.96	59.16 ± 0.62	42.00 ± 0.63	70.11 ± 0.32	59.44 ± 0.68
w/o Spatial Attention	3.79	2.39	65.55	97.85 ± 0.07	46.80 ± 0.98	75.50 ± 1.05	57.80 ± 0.75	40.65 ± 0.74	69.23 ± 0.37	58.45 ± 0.82
w/o Band Mixing	3.79	2.39	65.51	98.06 ± 0.05	48.96 ± 0.79	74.02 ± 0.91	58.93 ± 0.68	41.77 ± 0.68	69.90 ± 0.34	59.31 ± 0.71
ConvStem	3.80	2.57	62.22	98.50 ± 0.02	60.06 ± 0.72	60.82 ± 0.88	60.44 ± 0.55	43.31 ± 0.56	70.89 ± 0.28	59.67 ± 0.62
w/o LightDSCA	3.22	2.36	64.41	98.22 ± 0.04	52.36 ± 0.82	65.82 ± 0.94	58.33 ± 0.71	41.17 ± 0.71	69.69 ± 0.36	57.82 ± 0.78
Conv Decoder	4.35	2.84	57.48	98.44 ± 0.03	57.65 ± 0.76	65.40 ± 0.89	61.28 ± 0.58	44.17 ± 0.60	71.30 ± 0.30	60.61 ± 0.65

Removing spatial attention produces the largest degradation among the three SSBM ablations. F1 decreases from 64.01% to 57.80%, IoU from 47.07% to 40.65%, mIoU from

72.81% to 69.23%, and MCC from 63.39% to 58.45%. Meanwhile, Recall increases to 75.50%, whereas Precision decreases sharply to 46.80%. This behavior indicates that spatial attention is important for controlling spatial false positives and improving the balance between landslide-region completeness and background suppression under the fourteen-channel configuration.

Removing band mixing also results in clear performance degradation. Compared with the complete model, F1, IoU, mIoU, and MCC decrease by 5.08, 5.30, 2.91, and 4.08 percentage points, respectively. This result supports the role of the 1×1 band-mixing operation in enabling lightweight cross-channel interaction after attention-based recalibration.

The architectural replacement experiments further demonstrate complementary contributions of the remaining components. Replacing GhostStem with a conventional convolutional stem increases FLOPs from 2.39 G to 2.57 G and decreases FPS from 63.93 to 62.22, while F1 decreases by 3.57 percentage points. Replacing LightDSCA with an identity mapping reduces the parameter count and FLOPs but causes F1, IoU, mIoU, and MCC to decrease by 5.68, 5.90, 3.12, and 5.57 percentage points, respectively. Replacing the MLP projection–fusion decoder with a convolutional decoder increases the model complexity to 4.35 M parameters and 2.84 G FLOPs, reduces FPS to 57.48, and decreases F1 from 64.01% to 61.28%.

4.5.2. Results on the Four-Channel HR-GLDD Dataset

As shown in Table 7, both active SSBM operations contribute positively under the four-channel HR-GLDD configuration. Removing channel attention decreases F1 from 72.33% to 70.61%, IoU from 56.65% to 54.56%, mIoU from 74.74% to 73.50%, and MCC from 68.91% to 67.00%. Removing band mixing also decreases F1, IoU, mIoU, and MCC by 1.50, 1.81, 1.06, and 1.71 percentage points, respectively. Spatial attention is not included in the complete HR-GLDD configuration and is therefore not separately ablated on this dataset.

The architectural replacement results on HR-GLDD show a similar but more moderate trend. Replacing GhostStem with ConvStem increases FLOPs from 2.35 G to 2.51 G and reduces FPS from 63.63 to 58.20, while F1 decreases by 1.13 percentage points. Removing LightDSCA decreases F1 by 1.07 percentage points, and replacing the MLP projection–fusion decoder with a convolutional decoder decreases F1 by 1.04 percentage points while increasing the model complexity to 4.34 M parameters and 2.80 G FLOPs.

Table 7. Fine-grained ablation and architectural replacement results on the four-channel HR-GLDD dataset. Segmentation metrics are reported as mean ± standard deviation (%) over three independent training runs.

Variant	Params (M)	FLOPs (G)	FPS	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	IoU (%)	mIoU (%)	MCC (%)
Full LGSTNet	3.77	2.35	63.63	93.44 ± 0.22	67.05 ± 1.18	78.50 ± 1.55	72.33 ± 0.43	56.65 ± 0.53	74.74 ± 0.29	68.91 ± 0.46
w/o Channel Attention	3.77	2.35	63.75	93.10 ± 0.16	65.80 ± 0.82	76.20 ± 0.98	70.61 ± 0.65	54.56 ± 0.78	73.50 ± 0.47	67.00 ± 0.72
w/o Band Mixing	3.77	2.35	63.78	93.14 ± 0.15	66.09 ± 0.79	76.32 ± 0.92	70.83 ± 0.61	54.84 ± 0.73	73.68 ± 0.45	67.20 ± 0.69
ConvStem	3.79	2.51	58.20	93.14 ± 0.17	65.77 ± 0.85	77.61 ± 0.95	71.20 ± 0.63	55.28 ± 0.76	73.90 ± 0.46	67.64 ± 0.71
w/o LightDSCA	3.21	2.32	62.84	93.20 ± 0.15	66.13 ± 0.76	77.25 ± 0.89	71.26 ± 0.59	55.35 ± 0.71	73.96 ± 0.43	67.69 ± 0.66
Conv Decoder	4.34	2.80	55.09	93.23 ± 0.15	66.37 ± 0.81	77.02 ± 0.93	71.29 ± 0.64	55.39 ± 0.77	74.00 ± 0.38	67.72 ± 0.68

Overall, the fine-grained ablation results show that channel attention, spatial attention, and band mixing provide distinguishable contributions rather than functioning as an inseparable attention block. Together with GhostStem, LightDSCA, and the lightweight MLP projection–fusion decoder, these components form a coordinated architecture in which input recalibration, local-detail preservation, hierarchical semantic modeling, contextual enhancement, and efficient multi-level fusion provide complementary functions. The ablation effects that were observed to be larger on Landslide4Sense than on HR-GLDD are reported as dataset-dependent empirical differences and are not attributed solely to the number of input channels.

4.6. Limitations and Future Work

Although LGSTNet demonstrates competitive performance across the three datasets and achieves its strongest comparative results on Landslide4Sense, several limitations remain. First, the proposed model does not achieve absolute superiority on all datasets. In particular, HR-GLDD, STUNet and TransUNet obtain higher values on several accuracy-related metrics, and STUNet also achieves a higher FPS. The added within-dataset Landslide4Sense experiment reduces the confounding inherent in the original cross-dataset comparison and shows that both LGSTNet and SegFormer improve as the input configuration expands from RGB to the complete fourteen-channel representation. Nevertheless, this experiment does not isolate the contribution of every individual spectral or terrain-derived channel. Moreover, LGSTNet intentionally changes its input-recalibration configuration by activating spatial attention for the fourteen-channel setting. The observed gains should therefore be interpreted as evidence for the effectiveness of the complete input configuration rather than as proof that channel number alone causes the improvement. Future work should further evaluate individual modality groups and band-level contribution using strictly fixed architectural settings.

Second, the Jiuzhai Valley dataset was split into training, validation, and test subsets using a random patch-level split with a fixed seed, without explicitly enforcing scene-level or geographic-region-level independence. As a result, patches assigned to different subsets may share similar spatial context when they originate from the same or nearby UAV acquisition areas. Consequently, the reported Jiuzhai Valley results should be interpreted as performance on randomly held-out patches rather than as an evaluation of geographically independent scenes. Future work should further evaluate LGSTNet under scene-level, event-level, or cross-region splits to better assess geographic robustness.

Third, the current model does not introduce explicit boundary-aware supervision or post-processing. Therefore, highly fragmented landslides, shadow-covered regions, road-adjacent exposed surfaces, and bare-soil areas may still lead to false positives or incomplete boundary recovery. Incorporating boundary constraints, uncertainty estimation, or topology-aware refinement may further improve delineation quality.

Finally, the efficiency evaluation was conducted on an RTX 3090 GPU. Although FPS comparisons are measured under the same hardware environment, practical rapid landslide monitoring may involve edge devices or large-scene tiled inference. Future work will therefore investigate deployment-oriented optimization, large-area inference, and more adaptive heterogeneous feature fusion for real-world landslide mapping.

5. Conclusions

This paper presents LGSTNet, a lightweight local–global collaborative framework for landslide segmentation from remote sensing images. The architecture is organized around two coupled principles. The first is heterogeneous-input-oriented local–global encoding, which coordinates input-configuration-specific recalibration, shallow structural representa-

tion, and hierarchical Transformer modeling. The second is accuracy–efficiency-oriented compressed context fusion, which combines compressed deep-context enhancement with lightweight multi-level projection, alignment, and fusion. The contribution therefore lies in the coordinated representation and efficiency strategy rather than in treating the constituent operators as isolated architectural novelties.

Experiments on three public landslide datasets show dataset-dependent but complementary behavior. On HR-GLDD, LGSTNet achieves competitive segmentation performance and higher speed than several Transformer-based baselines, but it does not achieve absolute dominance over all methods. On Jiuzhai Valley, LGSTNet obtains segmentation quality close to SCDUNet++ and TransUNet while achieving higher inference speed. However, because the Jiuzhai Valley dataset was divided using a random patch-level split without explicitly enforced spatial independence, these results should be interpreted as performance on randomly held-out patches rather than as evidence of geographic or cross-region generalization. Future work will further evaluate LGSTNet under scene-level, event-level, and cross-region splits to assess geographic robustness. On Landslide4Sense, LGSTNet achieves the best F1, IoU, mIoU, and MCC among all compared methods.

The within-dataset Landslide4Sense input-configuration experiment provides additional evidence beyond the original cross-dataset comparison. For LGSTNet, F1 and IoU increase from 42.87% and 27.28% under RGB input to 64.01% and 47.07% under the complete fourteen-channel configuration, respectively, while SegFormer shows the same overall trend. LGSTNet's F1 advantage over SegFormer grows from 0.37 percentage points under RGB input to 2.45 percentage points under the full fourteen-channel configuration. Because the input content and LGSTNet's input-recalibration configuration change together, these results support the effectiveness of the complete heterogeneous-input configuration without establishing channel count alone as the causal source of the gain.

Fine-grained ablation further shows that channel attention, spatial attention, and band mixing make distinguishable contributions to the SSBM input interface. The architectural replacement experiments also support complementary roles for GhostStem, LightDSCA, and the MLP projection–fusion decoder. Taken together, these results indicate that LGSTNet's behavior is better explained by the coordination of input recalibration, local-detail preservation, hierarchical semantic modeling, contextual enhancement, and efficient multi-level fusion than by any single component in isolation.

Author Contributions: A.H.: Conceptualization, Methodology, Software, Data curation, Formal analysis, Validation, Visualization, Writing–original draft. H.C.: Conceptualization, Supervision, Funding acquisition, Project administration, Writing–review and editing. R.W.: Resources, Software, Validation, Visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development Program of Liaoning Province under Grant [2024JH2/102500071] and the Liaoning Provincial Science and Technology Commissioner Team Project under Grant [2026JH5/10400152].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CASLandslide, HR-GLDD, and Landslide4Sense datasets analyzed in this study are publicly available from the original sources cited in Section 3.1. Landslide4Sense is also available at https://huggingface.co/datasets/ibm-nasa-geospatial/Landslide4sense, accessed on 24 October 2026. No new dataset was generated. The fixed Jiuzhai Valley split file and the experimental materials supporting the reported results are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the providers of the public landslide remote sensing datasets used in this study.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Mohan, A.; Singh, A.K.; Kumar, B.; Dwivedi, R. Review on Remote Sensing Methods for Landslide Detection Using Machine and Deep Learning. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e3998. https://doi.org/10.1002/ett.3998.
2. Wang, H.; Zhang, L.M.; Yin, K.; Luo, H.; Li, J. Landslide Identification Using Machine Learning. *Geosci. Front.* **2021**, *12*, 351–364. https://doi.org/10.1016/j.gsf.2020.02.012.
3. Mondini, A.C.; Guzzetti, F.; Chang, K.T.; Monserrat, O.; Martha, T.R.; Manconi, A. Landslide Failures Detection and Mapping Using Synthetic Aperture Radar: Past, Present and Future. *Earth-Sci. Rev.* **2021**, *216*, 103574. https://doi.org/10.1016/j.earscirev.2021.103574.
4. Zhang, Q.; Wang, T. Deep Learning for Exploring Landslides with Remote Sensing Images: A Review. *Remote Sens.* **2024**, *16*, 1344. https://doi.org/10.3390/rs16081344.
5. Yang, Z.Q.; Qi, W.W.; Xu, C.; Shao, X.Y. Exploring Deep Learning for Landslide Mapping: A Comprehensive Review. *China Geol.* **2024**, *7*, 330–350. https://doi.org/10.31035/cg2024032.
6. Lv, J.; Shen, Q.; Lv, M.; Li, Y.; Shi, L.; Zhang, P. Deep Learning-Based Semantic Segmentation of Remote Sensing Images: A Review. *Front. Ecol. Evol.* **2023**, *11*, 1201125. https://doi.org/10.3389/fevo.2023.1201125.
7. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11214, pp. 561–580. https://doi.org/10.1007/978-3-030-01249-6_34.
8. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11207, pp. 418–434. https://doi.org/10.1007/978-3-030-01219-9_25.
9. Wang, R.; Ma, L.; He, G.; Johnson, B.A.; Yan, Z.; Chang, M.; Liang, Y. Transformers for Remote Sensing: A Systematic Review and Analysis. *Sensors* **2024**, *24*, 3495. https://doi.org/10.3390/s24113495.
10. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. https://doi.org/10.1109/LGRS.2022.3143368.
11. 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. https://doi.org/10.1016/j.isprsjprs.2022.06.008.
12. Li, J.; Cai, Y.; Li, Q.; Kou, M.; Zhang, T. A Review of Remote Sensing Image Segmentation by Deep Learning Methods. *Int. J. Digit. Earth* **2024**, *17*, 2328827. https://doi.org/10.1080/17538947.2024.2328827.
13. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
15. 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 (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 6230–6239. https://doi.org/10.1109/CVPR.2017.660.
16. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1.
17. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11211, pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. Available online: https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need (accessed on 24 October 2026).

19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), 2021. Available online: https://openreview.net/forum?id=YicbFdNTTy (accessed on 24 October 2026).
20. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090. Available online: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html (accessed on 24 October 2026).
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.
22. 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 *Computer Vision–ECCV 2022 Workshops*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13803, pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9.
23. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers. *Med. Image Anal.* **2024**, *97*, 103280. https://doi.org/10.1016/j.media.2024.103280.
24. He, S.; Yang, H.; Zhang, X.; Li, X. MFTransNet: A Multi-Modal Fusion with CNN-Transformer Network for Semantic Segmentation of HSR Remote Sensing Images. *Mathematics* **2023**, *11*, 722. https://doi.org/10.3390/math11030722.
25. Liu, Y.; Zhang, Y.; Wang, Y.; Mei, S. BiTSRS: A Bi-Decoder Transformer Segmentor for High-Spatial-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 840. https://doi.org/10.3390/rs15030840.
26. Ma, X.; Zhang, X.; Pun, M.O.; Liu, M. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. https://doi.org/10.1109/TGRS.2024.3373033.
27. Liu, Y.; Gao, K.; Wang, H.; Yang, Z.; Wang, P.; Ji, S.; Huang, Y.; Zhu, Z.; Zhao, X. A Transformer-Based Multi-Modal Fusion Network for Semantic Segmentation of High-Resolution Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104083. https://doi.org/10.1016/j.jag.2024.104083.
28. Li, K.; Ji, H.; Li, Z.; Cui, Z.; Liu, C. AFNE-Net: Semantic Segmentation of Remote Sensing Images via Attention-Based Feature Fusion and Neighborhood Feature Enhancement. *Remote Sens.* **2025**, *17*, 2443. https://doi.org/10.3390/rs17142443.
29. Zhu, Q.; Cai, Y.; Fang, Y.; Yang, Y.; Chen, C.; Fan, L.; Nguyen, A. RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *Heliyon* **2024**, *10*, e38495. https://doi.org/10.1016/j.heliyon.2024.e38495.
30. Wang, Y.; Cao, L.; Deng, H. MFMamba: A Mamba-Based Multi-Modal Fusion Network for Remote Sensing Semantic Segmentation. *Sensors* **2024**, *24*, 7266. https://doi.org/10.3390/s24227266.
31. Ta, N.; Chen, H.; Liu, X.; Jin, N. LET-Net: Locally Enhanced Transformer Network for Medical Image Segmentation. *Multimed. Syst.* **2023**, *29*, 3847–3861. https://doi.org/10.1007/s00530-023-01165-z.
32. Torres, H.R.; Oliveira, B.; Morais, P.; Fritze, A.; Hahn, G.; Rüdiger, M.; Fonseca, J.C.; Vilaça, J.L. Infant Head and Brain Segmentation from Magnetic Resonance Images Using Fusion-Based Deep Learning Strategies. *Multimed. Syst.* **2024**, *30*, 71. https://doi.org/10.1007/s00530-024-01267-2.
33. Zhang, Y.; Cha, H.; Zhang, H.; Li, X.; Chen, Z.; Bai, J.; Fu, T.; Du, Y.C.; Zhang, X. MCDFus: Visible and Infrared Image Fusion via Mamba-CNN for Enhanced UAV Perception. *Multimed. Syst.* **2026**, *32*, 32. https://doi.org/10.1007/s00530-025-02080-1.
34. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11217, pp. 325–341. https://doi.org/10.1007/978-3-030-01261-8_20.
35. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1580–1589. https://doi.org/10.1109/CVPR42600.2020.00165.
36. Zhang, X.; Yu, W.; Pun, M.O.; Shi, W. Cross-Domain Landslide Mapping from Large-Scale Remote Sensing Images Using Prototype-Guided Domain-Aware Progressive Representation Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 1–17. https://doi.org/10.1016/j.isprsjprs.2023.01.018.
37. Wu, L.; Liu, R.; Ju, N.; Zhang, A.; Gou, J.; He, G.; Lei, Y. Landslide Mapping Based on a Hybrid CNN-Transformer Network and Deep Transfer Learning Using Remote Sensing Images with Topographic and Spectral Features. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *126*, 103612. https://doi.org/10.1016/j.jag.2023.103612.
38. Wang, K.; He, D.; Sun, Q.; Yi, L.; Yuan, X.; Wang, Y. A Novel Network for Semantic Segmentation of Landslide Areas in Remote Sensing Images with Multi-Branch and Multi-Scale Fusion. *Appl. Soft Comput.* **2024**, *158*, 111542. https://doi.org/10.1016/j.asoc.2024.111542.
39. Liu, B.; Wang, W.; Wu, Y.; Gao, X. Attention Swin Transformer UNet for Landslide Segmentation in Remotely Sensed Images. *Remote Sens.* **2024**, *16*, 4464. https://doi.org/10.3390/rs16234464.

40. Yu, C.C.; Chen, Y.D.; Cheng, H.Y.; Jiang, C.L. Semantic Segmentation of Satellite Images for Landslide Detection Using Foreground-Aware and Multi-Scale Convolutional Attention Mechanism. *Sensors* **2024**, *24*, 6539. https://doi.org/10.3390/s24206539.
41. Zhou, N.; Hong, J.; Cui, W.; Wu, S.; Zhang, Z. A Multiscale Attention Segment Network-Based Semantic Segmentation Model for Landslide Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1712. https://doi.org/10.3390/rs16101712.
42. Zhu, X.; Zhang, Z.; He, Y.; Wang, W.; Yang, S.; Hou, Y. LandslideNet: A Landslide Semantic Segmentation Network Based on Single-Temporal Optical Remote Sensing Images. *Adv. Space Res.* **2024**, *74*, 4616–4638. https://doi.org/10.1016/j.asr.2024.07.038.
43. Li, Y.; Zhu, W.; Wu, J.; Zhang, R.; Xu, X.; Zhou, Y. DBSANet: A Dual-Branch Semantic Aggregation Network for Landslide Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 807. https://doi.org/10.3390/rs17050807.
44. Li, C.; Zou, Q.; Li, G.; Yu, W. Landslide Segmentation in High-Resolution Remote Sensing Images: The Van–UPerAttnSeg Framework with Multi-Scale Feature Enhancement. *Remote Sens.* **2025**, *17*, 1265. https://doi.org/10.3390/rs17071265.
45. Tang, M.; He, Y.; Aslam, M.; Akpokodje, E.; Jilani, S.F. Enhanced U-Net++ for Improved Semantic Segmentation in Landslide Detection. *Sensors* **2025**, *25*, 2670. https://doi.org/10.3390/s25092670.
46. Ghorbanzadeh, O.; Xu, Y.; Ghamisi, P.; Kopp, M.; Kreil, D.P. Landslide4Sense: Reference Benchmark Data and Deep Learning Models for Landslide Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. https://doi.org/10.1109/TGRS.2022.3215209.
47. Meena, S.R.; Nava, L.; Bhuyan, K.; Puliero, S.; Soares, L.P.; Dias, H.C.; Floris, M.; Catani, F. HR-GLDD: A Globally Distributed Dataset Using Generalized Deep Learning (DL) for Rapid Landslide Mapping on High-Resolution (HR) Satellite Imagery. *Earth Syst. Sci. Data* **2023**, *15*, 3283–3298. https://doi.org/10.5194/essd-15-3283-2023.
48. Xu, Y.; Ouyang, C.; Xu, Q.; Wang, D.; Zhao, B.; Luo, Y. CAS Landslide Dataset: A Large-Scale and Multisensor Dataset for Deep Learning-Based Landslide Detection. *Sci. Data* **2024**, *11*, 12. https://doi.org/10.1038/s41597-023-02847-z.
49. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
50. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
51. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-over-Union Measure in Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 4413–4421. https://doi.org/10.1109/CVPR.2018.00464.
52. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *Machine Learning in Medical Imaging (MLMI 2017)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10541, pp. 379–387. https://doi.org/10.1007/978-3-319-67389-9_44.
53. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. https://doi.org/10.1109/TGRS.2022.3144165.

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.