remote sensing

MDPI

Article

BiFusion-LDSeg: A Latent Diffusion Framework with Bi-Directional Attention Fusion for Landslide Segmentation in Satellite Imagery

Bingxin Shi [1,2,3], Hongmei Guo [3], Yin Sun [3,4,*], Jianyu Long [3], Li Yang [3], Yadong Zhou [3], Jingjing Jiao [5], Jingren Zhou [6], Yusen He [7] and Huajin Li [2,8]

1 Institute of Engineering Mechanics, China Earthquake Administration, Harbin 150080, China
2 State Key Laboratory of Geohazard Prevention and Geoenvironmental Protection, Chengdu University of Technology, Chengdu 610059, China; lihuajin@cdu.edu.cn
3 Sichuan Earthquake Administration, Chengdu 610041, China
4 Geophysical Exploration Center, China Earthquake Administration, Zhengzhou 450002, China
5 BGI Engineering Consultants Ltd., Beijing 100038, China
6 State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065, China
7 Department of Industrial and System Engineering, University of Iowa, Iowa, IA 52242, USA
8 Major Hazard Monitoring and Emergency Response Key Laboratory of Sichuan Province, Chengdu University, Chengdu 610106, China
* Correspondence: sunyin@scdzj.gov.cn

Highlights

What are the main findings?

- BiFusion-LDSeg, a latent diffusion framework with bi-directional CNN-Transformer fusion, achieves superior landslide segmentation accuracy across three earthquake-triggered datasets in Sichuan Province, China.
- The framework enables rapid and reliable inference, with ablation studies confirming the independent contribution of each proposed component.

What is the implication of the main finding?

- Latent diffusion models are computationally practical for large-scale satellite-based hazard mapping, extending their applicability beyond medical imaging.
- Built-in uncertainty quantification and strong cross-dataset generalization support direct operational deployment for post-earthquake emergency response.

Abstract

Rapid and accurate mapping of earthquake-triggered landslides from satellite imagery is critical for emergency response and hazard assessment, yet remains challenging due to irregular boundaries, extreme size variations, and atmospheric noise. This paper proposes BiFusion-LDSeg, a novel bi-directional fusion enhanced latent diffusion framework that synergistically combines CNN-Transformer architectures with generative diffusion models for robust landslide segmentation. The framework introduces three key innovations: (1) a dual-encoder with Bi-directional Attention Gates (Bi-AG) enabling sophisticated cross-modal feature calibration between local CNN textures and global Transformer context; (2) a conditional latent diffusion process operating in learned low-dimensional landslide shape manifolds, reducing computational complexity by 100× while enabling inference with only 10 sampling steps versus 1000+ in standard diffusion models; and (3) a boundary-aware progressive decoder employing multi-scale reverse attention mechanisms for precise boundary delineation. Comprehensive experiments on three earthquake datasets from

Academic Editor: Michele Saroli

Received: 28 January 2026
Revised: 22 February 2026
Accepted: 26 February 2026
Published: 27 February 2026

Sichuan Province, China (Lushan Mw 7.0, Jiuzhaigou Mw 6.5, Luding Mw 6.8) demonstrate superior performance, outperforming state-of-the-art methods by 7–13% in IoU and 5–7% in DSC across all three datasets. The framework exhibits exceptional noise robustness, strong cross-dataset generalization, and inherent uncertainty quantification, enabling reliable deployment for post-earthquake landslide inventory mapping at regional scales.

Keywords: landslide segmentation; latent diffusion model; bi-directional attention fusion; CNN-transformer hybrid encoder; boundary-aware decoding

1. Introduction

Earthquake-triggered landslides are among the most persistent natural catastrophes associated with seismic events. They often destroy houses, roads, farmland and even cause severe loss of life and casualties to humans [1]. Major seismic events can trigger thousands of landslides across vast mountainous regions. For example, the Mw 7.9 Wenchuan (China) earthquake resulted in nearly 200,000 landslides while the Mw 7.9 Gorkha (Nepal) earthquake induced >25,000 landslides and debris flows [2]. The rapid and comprehensive mapping of coseismic landslides is critically important for several reasons: guiding emergency response and rescue operations, assessing secondary hazard risks such as landslide dams and debris flows, supporting post-disaster recovery planning, and advancing the understanding of earthquake-landslide relationships [3,4]. Traditional field-based surveys are often impractical immediately after major seismic events, particularly in remote or inaccessible mountainous terrain, necessitating automated approaches for rapid landslide detection and mapping at regional scales.

Satellite remote sensing has emerged as a critical technology for earthquake-triggered landslide detection and mapping. Modern platforms such as Sentinel-2, Landsat, and high-resolution commercial satellites provide diverse spectral information enabling comprehensive landslide detection [5–7]. However, automatic segmentation of coseismic landslides presents formidable challenges including irregular boundaries, extreme size variations, and visual ambiguity arising from spectral similarities with surrounding terrain features (Figure 1). Moreover, landslide pixels constitute less than 5% of imagery and cause severe class imbalances, which further complicates model training. In addition, image quality degradation from atmospheric noise, cloud cover, and sensor artifacts limits detection accuracy. These challenges necessitate robust methods capable of handling irregular boundaries, multi-scale variations, and noisy imagery while maintaining computational efficiency for rapid post-earthquake deployment.

Deep learning methods have achieved remarkable progress in semantic segmentation tasks, with various architectures being adapted for landslide detection from satellite imagery. Traditional convolutional neural networks (CNNs) such as U-Net [8], ResUNet [9], and DeepLabv3+ [10] excel at extracting local spatial features and have demonstrated strong performance in landslide segmentation. However, their inherently limited receptive fields constrain their ability to capture large-scale contextual information critical for landslide detection [11]. Attention mechanisms including Squeeze-and-Excitation Networks (SE-Net) [12,13] and Convolutional Block Attention Modules (CBAM) [14,15] partially address this but remain constrained by their convolutional backbones [16]. Recently, Transformer-based architectures such as Vision Transformer (ViT) [17] and Swin Transformer [18] have been introduced to capture long-range dependencies through self-attention mechanisms, showing promise in modeling global context. However, these models often sacrifice fine-grained spatial details essential for precise boundary delineation,

while incurring substantial computational costs that limit operational deployment [19,20]. Hybrid CNN-Transformer architectures have emerged to combine local and global features [21,22], yet most employ simplistic fusion strategies that fail to fully exploit their complementary strengths [23]. Furthermore, these deterministic models provide binary predictions without uncertainty quantification and exhibit vulnerability to atmospheric noise and sensor artifacts prevalent in satellite imagery [24,25]. These limitations underscore the need for advanced architectures that can effectively integrate multi-scale local and global features while providing robust predictions in noisy conditions.

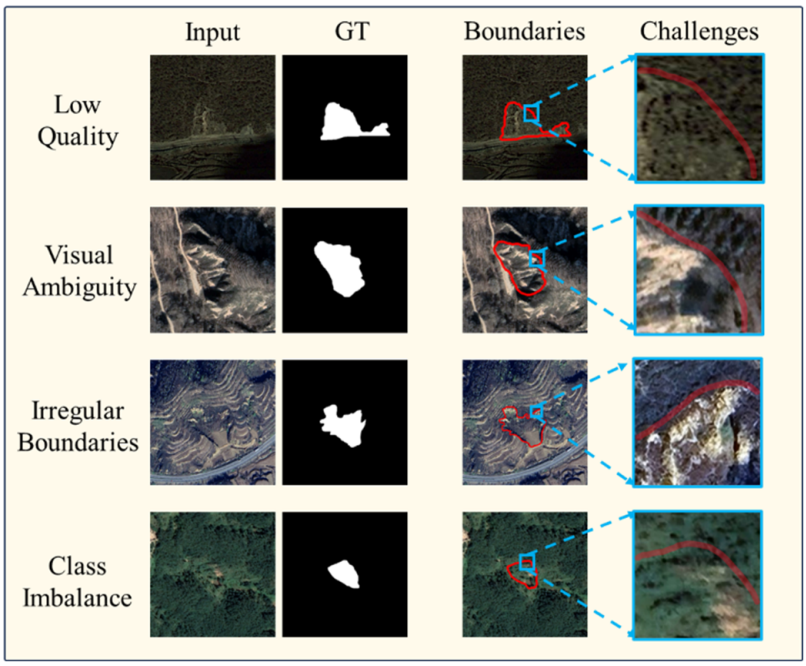

Figure 1. Motivation and challenge illustration.

While these discriminative models have achieved notable success, they fundamentally lack the ability to model the uncertainty inherent in landslide boundaries and are susceptible to noisy satellite imagery. Recently, generative models have shown promise in addressing these limitations by learning the underlying data distribution rather than direct input–output mappings [26,27]. Among generative approaches, Denoising Diffusion Probabilistic Models (DPMs) have emerged as state-of-the-art methods for image generation, demonstrating superior performance and more stable training compared to Generative Adversarial Networks (GANs) [28,29]. DPMs naturally provide uncertainty quantification through their iterative denoising process and have shown remarkable success in medical image segmentation, particularly for handling ambiguous boundaries and maintaining robustness to image noise [30,31]. However, diffusion models remain largely unexplored for remote sensing applications, despite their potential benefits for landslide detection where boundary uncertainty and atmospheric noise are prevalent challenges [32–34].

This research proposes a novel Bi-directional Fusion Enhanced Latent Diffusion Landslide Segmentation framework (BiFusion-LDSeg) that synergistically combines the strengths of CNN-Transformer architectures with the generative power of diffusion models, specifically designed for landslide segmentation in satellite imagery. The framework introduces four core innovations. First, it performs diffusion in a learned low-dimensional latent space capturing landslide shape manifolds, reducing computational complexity by over 100× and enabling inference with only 10 sampling steps versus 1000+ in standard diffusion models. Second, it applies a Bi-directional Attention Gate (Bi-AG) mechanism enabling sophisticated CNN-Transformer fusion that captures both fine-grained local

textures and large-scale landscape context as rich conditional information for diffusion. Third, this paper introduces a boundary-aware progressive decoder employing multi-scale reverse attention mechanisms to refine irregular landslide boundaries iteratively. Moreover, unlike two-stage approaches, the proposed framework adopts end-to-end learning that jointly optimizes all components ensuring latent representations are specifically optimized for segmentation. These innovations enable BiFusion-LDSeg to achieve accurate multi-scale landslide boundary delineation, maintain robustness to atmospheric noise and sensor artifacts, provide uncertainty quantification, and operate efficiently for rapid post-earthquake deployment.

The main contributions of this work are summarized as follows:

This research proposes the first latent diffusion model for landslide segmentation from satellite imagery, addressing computational challenges through efficient latent space operations.

This paper designs a novel Bi-AG module that enables effective bi-directional information exchange between CNN and Transformer encoders, capturing both local terrain textures and landscape-level patterns critical for landslide detection.

In this paper, reverse attention modules within a progressive decoder are integrated to explicitly enhance landslide boundary accuracy, which is critical for damage assessment and risk analysis

This paragraph should be revised to “The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning for landslide detection and diffusion models for segmentation. Section 3 presents the proposed methodology in detail, including the dual encoder architecture, latent diffusion process, and boundary-aware decoder. Section 4 describes the experimental setup, datasets, and discusses comprehensive experimental results and analysis. Section 5 discusses the findings, limitations, and future directions. Finally, Section 6 concludes this research paper.

2. Related Works

2.1. Deep Learning for Landslide Segmentation

Convolutional neural networks have become the dominant approach for landslide segmentation from satellite imagery. U-Net and its variants demonstrate robust performance [7,8], with ResUNet architectures particularly effective through residual connections [9], and DeepLabv3+ employing atrous spatial pyramid pooling for multi-scale context [10]. Attention mechanisms including SE-Net [12,13] and CBAM [15] have been integrated to enhance feature discrimination. However, CNN-based methods remain fundamentally constrained by limited receptive fields, restricting their ability to model long-range dependencies critical for distinguishing landslides across extensive mountainous regions [11].

Transformer-based architectures have emerged to capture global context through self-attention mechanisms. ViT [17] and Swin Transformer [18] demonstrate improved capability in modeling landscape-level patterns [20,32], computing attention across all spatial positions to capture long-range terrain relationships. However, pure Transformers sacrifice fine-grained spatial details essential for precise boundaries while incurring high computational costs [32]. Hybrid CNN-Transformer architectures [6,21–23] attempt to combine local and global features, but most employ simplistic fusion strategies that fail to fully exploit complementary strengths [23].

Current approaches exhibit critical limitations: inadequate multi-scale feature integration, poor handling of irregular boundaries, vulnerability to atmospheric noise, and lack of uncertainty quantification. These challenges necessitate novel architectures that effectively

fuse CNN and Transformer features through sophisticated attention mechanisms while maintaining computational efficiency for large-scale satellite imagery processing.

2.2. Feature Fusion Strategies in Hybrid Architectures

Hybrid architectures typically employ straightforward fusion strategies to combine CNN and Transformer features. Simple concatenation along the channel dimension [22,23] or element-wise addition [21] are commonly used but treat features equally without considering their complementary characteristics. Skip connections in U-Net-style architectures [35] directly merge multi-scale features from encoder to decoder, yet fail to selectively emphasize relevant information or suppress noise, particularly critical when integrating heterogeneous CNN and Transformer representations [36].

Attention-based fusion mechanisms have been proposed to improve feature integration through selective weighting. Single-directional attention gates filter encoder features using decoder queries [37,38], successfully applied in medical image segmentation to highlight salient regions. Cross-attention mechanisms [39] enable information exchange between modalities but typically operate unidirectionally or require complex multi-stage processing. Existing methods lack sophisticated bi-directional fusion that simultaneously calibrates both CNN and Transformer features using mutual guidance, limiting their capacity to fully exploit complementary local–global information [40]. This gap motivates the proposed Bi-AG module, which enables reciprocal feature refinement through dual attention pathways operating in parallel.

2.3. Diffusion Models for Image Segmentation

Denoising Diffusion Probabilistic Models (DPMs) have emerged as powerful generative frameworks, demonstrating superior performance in image synthesis compared to GANs [28,29]. DPMs learn data distributions through iterative denoising processes, gradually transforming Gaussian noise into structured outputs [33]. Recently, diffusion models have been adapted for medical image segmentation tasks, showing remarkable capability in handling ambiguous boundaries and providing uncertainty quantification [30,41]. MedSegDiff [30] performs diffusion in image space for medical segmentation but requires 1000+ sampling steps, incurring substantial computational costs. These standard approaches operate in high-dimensional pixel space, limiting their applicability to large-scale remote sensing imagery.

Latent diffusion models address computational challenges by performing diffusion in learned low-dimensional latent spaces [42]. LDSeg [31] demonstrates that latent space diffusion with end-to-end training achieves accurate medical image segmentation with only 2–10 sampling steps versus 1000+ in standard DPMs, reducing inference time by over 100×. Despite their success in medical imaging, diffusion models remain largely unexplored for remote sensing applications, particularly for landslide detection where boundary uncertainty and atmospheric noise are prevalent [32]. This gap motivates the proposed latent diffusion approach specifically designed for efficient landslide segmentation from large-scale satellite imagery.

3. Methodology

3.1. Overview of the Framework

This research proposes BiFusion-LDSeg, which is a novel framework that integrates bi-directional dual-encoder fusion with conditional latent diffusion for landslide segmentation in satellite imagery. As shown in Figure 2, the proposed framework consists of four major components: (a) a dual image encoder combining CNN and Transformer branches with Bi-directional Attention Gates (Bi-AG) to extract multi-scale local and

global features; (b) a label encoder that learns standardized low-dimensional representations of landslide shape manifolds; (c) a conditional denoiser operating in latent space that iteratively refines segmentation predictions guided by image embeddings; and (d) a boundary-aware progressive decoder with reverse attention modules for multi-scale feature restoration.

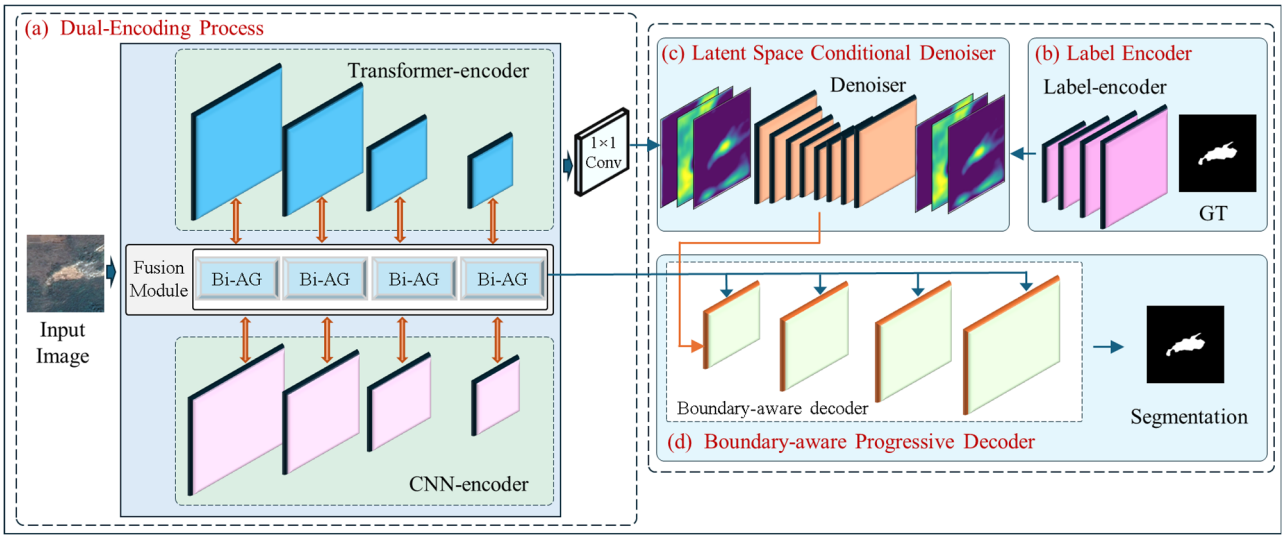

Figure 2. Overall framework architecture.

Given an input satellite image $X \in \mathbb{R}^{H\times W\times 3}$, the dual encoder generates a fused image embedding $z_i \in \mathbb{R}^{h\times w\times d}$ through four cascaded Bi-AG modules. During training, the label encoder maps the ground-truth landslide mask $Y \in \{0,1\}^{H\times W}$ to a standardized latent representation $z_l(0) \in N(0,I)$, which is progressively corrupted by Gaussian noise. The conditional denoiser learns to predict and remove this noise conditioned on z_i, enabling the framework to sample accurate segmentation masks during inference.

Unlike standard segmentation models that directly predict pixel labels, the framework learns to sample from the conditional distribution of landslide boundaries, given satellite imagery. This probabilistic formulation provides natural uncertainty quantification and robustness to noisy inputs. Compared to traditional diffusion models that operate in high-dimensional image space, the proposed latent diffusion approach reduces computational complexity by 16× while enabling efficient sampling with as few as 10 timesteps. The bi-directional fusion mechanism addresses key challenges in landslide detection: CNN branches capture fine-grained terrain textures and local landslide patterns while Transformer branches model landscape-level spatial relationships and contextual dependencies. The Bi-AG modules enable cross-attention between these complementary representations at multiple scales, ensuring robust feature extraction for landslides with irregular boundaries, extreme size variations, and visual ambiguity commonly observed in satellite imagery.

3.2. Bi-Directional Dual Encoder with Attention Fusion

3.2.1. Dual-Branch Encoding Structure

The proposed framework utilizes a dual image encoder to extract complementary local and global features from satellite imagery through parallel CNN and Transformer branches.

CNN Branch: The CNN encoder branch adopts ResNet-50 [43] as the backbone, producing multi-scale feature maps $\{X_C^i,\ i = 1, \ldots, 4\}$ at progressively reduced spatial resolutions. It consists of four consecutive convolution blocks that decrease the spatial resolution of feature maps, but the channel number increases. Given an input satellite image $X \in \mathbb{R}^{H\times W\times 3}$, the CNN encoder branch produces $X_C^1 \in \mathbb{R}^{H/4\times W/4\times 256}$,

$X_C^2 \in \mathbb{R}^{H/8\times W/8\times 512}$, $X_C^3 \in \mathbb{R}^{H/16\times W/16\times 256}$, $X_C^4 \in \mathbb{R}^{H/32\times W/32\times 2048}$ feature activation maps. These convolutional features capture local terrain textures, edge information, and fine-grained landslide patterns through hierarchical receptive fields.

Transformer Branch: The Transformer branch employs a Vision Transformer (ViT) architecture [44] that divides the input image into non-overlapping patches and processes them through 12 Transformer blocks organized into four stages $\{T_i,\ i = 1,\ \ldots, 4\}$. Each stage consists of three consecutive blocks with Multi-head Self-Attention (MSA) and feed-forward layers. The MSA block serves as an essential component of Transformers that employs multiple attention heads to calculate attention.

In detail, the input satellite image $X \in \mathbb{R}^{H\times W\times 3}$ is segmented into non-overlapping windows of size $M \times M$. In each window, linear projections are applied to get the matrices Q, K, and V following Equation (1):

$$Q = XW^Q, K = XW^K, V = XW^V \tag{1}$$

where W^Q, W^K, and W^V are learnable projection matrices that will be split into multiple heads. Suppose there are h heads, the dimension of each head will be $d_k = \frac{c}{h}$ and the split will follow Equations (2) and (3) to obtain the self-attention:

$$Q_i = split(Q, h), K_i = split(K, h), V_i = split(V, h) \tag{2}$$

$$Attention_i(Q_i, K_i, V_i) = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + B\right)V_i \tag{3}$$

where B denotes the relative position bias. Outputs of all heads are concatenated and projected back to the original dimension. Overall, the Transformer features capture long-range spatial dependencies and landscape-level context, which are crucial for distinguishing landslides from visually similar terrain features such as shadows, bare soil, or vegetation clearings [45].

3.2.2. Bi-Directional Attention Gate (Bi-AG)

To effectively fuse CNN and Transformer features at multiple scales, this research employs Bi-directional Attention Gates (Bi-AG) after each encoding stage (Figure 3). Unlike conventional fusion methods that simply concatenate or add features, Bi-AG performs cross-modal calibration through parallel attention mechanisms with two inputs and two outputs. At stage i, the module accepts CNN features X_C^i and Transformer features X_T^i and aligns them to the same spatial and channel dimensions through projection layers. Next, the CNN and Transformer features are transformed by Equations (4) and (5) as:

$$h_C^i = LN\left(Reshape\left(Down\left(\varphi^c\left(X_C^i\right)\right)\right)\right) \tag{4}$$

$$h_T^i = BN\left(Up\left(\varphi^c\left(Reshape\left(X_T^i\right)\right)\right)\right) \tag{5}$$

where φ^c denotes 1×1 convolution; $Down()$ represents spatial downsampling; $LN()$ is layer normalization; $Up()$ denotes upsampling; and $BN()$ denotes batch normalization. Then, the two attention gates generate calibration weights W_C^i & W_T^i via Equations (6) and (7) and the calibrated outputs Y_C^i & Y_T^i can be computed by Equations (8) and (9):

$$W_C^i = \sigma\left(\varphi^c\left(\mathrm{ReLU}\left(\varphi^c\left(X_C^i\right) + \varphi^c\left(h_T^i\right)\right)\right)\right) \tag{6}$$

$$W_T^i = \sigma\left(\varphi^c\left(\mathrm{ReLU}\left(\varphi^c\left(h_C^i\right) + \varphi^c\left(X_T^i\right)\right)\right)\right) \tag{7}$$

$$F_C^i = W_C^i \cdot h_T^i + X_C^i \tag{8}$$

$$F_T^i = W_T^i \cdot h_C^i + X_T^i \tag{9}$$

where σ is the sigmoid activation function. The calibrated outputs F_C^i and F_T^i are fed back to their respective encoding branches, enabling bi-directional information exchange across four cascaded stages. This mechanism allows CNN features to guide Transformer attention toward relevant spatial locations while Transformer features provide global context to refine local CNN representations [46].

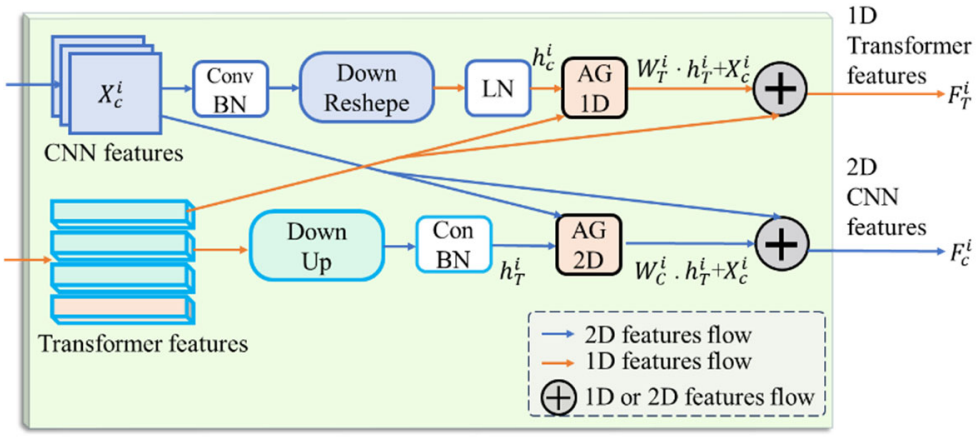

Figure 3. Bi-directional Attention Gate (Bi-AG) module.

3.2.3. Label Encoder

The dual-branch image encoder combines the calibrated features from the final Bi-AG stage through concatenation and projection layers to generate the image embedding expressed as $f_{\text{image}-\text{enc}}(\cdot)$. To enable efficient diffusion in latent space, this study designed a label encoder $f_{\text{label}-\text{enc}}(\cdot)$ that maps discrete landslide masks Y to continuous standardized representations. Here, the image encoder and the label encoder can be expressed in Equations (10) and (11) as:

$$z_i = f_{\text{image}-\text{enc}}(X) \tag{10}$$

$$z_{l(0)} = f_{\text{label}-\text{enc}}(Y) \tag{11}$$

where the label encoder $f_{\text{label}-\text{enc}}(\cdot)$ adopts a ResUNet-style architecture with four down-sampling layers but without skip connections, progressively reducing the spatial resolution from $H \times W$ to $h \times w$ (typically from 512×512 inputs to 32×32). Critically, a normalization layer in the final stage ensures the latent embeddings follow a standardized distribution with $\mu = 0$ and $\sigma = 1$. This standardized latent space $z_{l(0)} \in \mathbb{R}^{h \times d \times d}$ (where $h = 32$, $w = 32$, $d = 512$) serves as a continuous representation of landslide shape manifolds, where d is the latent dimension (typically 512). Unlike discrete binary masks, this continuous representation enables smooth transitions between landslide and non-landslide regions, naturally handling irregular boundaries and partial occlusions. The learned shape manifolds encode geometric priors about typical landslide morphology, including elongated downslope patterns, curved scarps, and fragmented debris zones, which act as strong regularization during the subsequent diffusion process.

3.3. Conditional Latent Diffusion Process

Having obtained the standardized label embedding $z_{l(0)}$ and image embedding z_i from the dual encoders, here the conditional latent diffusion process is introduced (Figure 4), which enables probabilistic segmentation generation. The diffusion process consists of two stages: a forward process systematically corrupts $z_{l(0)}$ with Gaussian noise, and a

reverse process that teaches to denoise and recover accurate landslide segmentations conditioned on z_i.

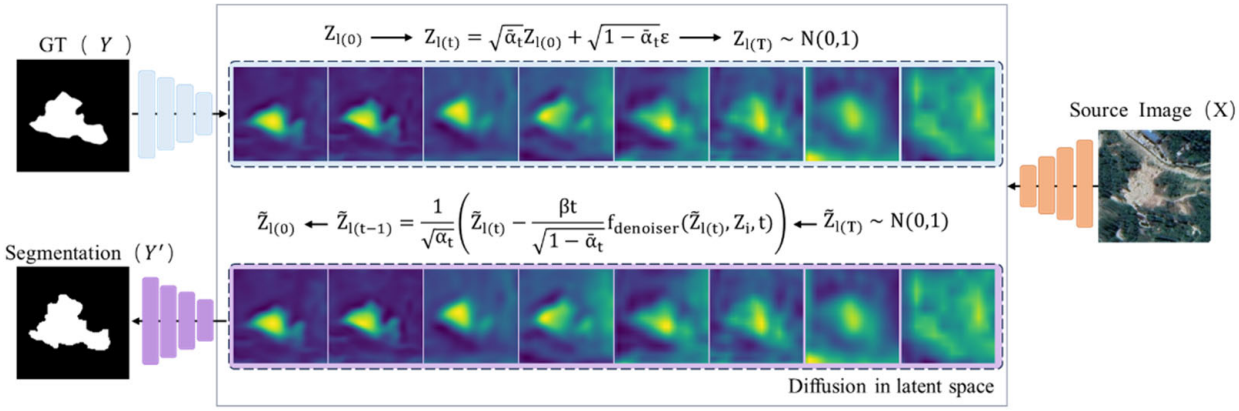

Figure 4. Latent diffusion process illustration.

Forward Process: The forward diffusion process gradually corrupts the label embedding $z_{l(0)}$ by adding Gaussian noise over T timesteps, creating a Markov chain of increasingly noisy latent representations. The forward process is defined as:

$$q\left(z_{l(1)}, \ldots, z_{l(T)} \mid z_{l(0)}\right) := \prod_{t=1}^{T}\left(z_{l(t)} \mid z_{l(t-1)}\right), \tag{12}$$

where each transition follows a Gaussian distribution:

$$q\left(z_{l(t)} \middle| z_{l(t-1)}\right) := \mathcal{N}\left(z_{l(t-1)}; \sqrt{1-\beta_t} z_{l(t-1)}, \beta_t \mathrm{I}\right), \tag{13}$$

where $\beta_t \in [\beta_1, \beta_T]$ defines the noise variance schedule at timestep t. It adopts a cosine scheduling strategy for β_t, which provides smoother noise injections and better preserves latent structure compared to linear schedules. The forward process admits a closed-form solution for sampling $z_{l(t)}$ directly from $z_{l(0)}$:

$$q\left(z_{l(t)} \middle| z_{l(0)}\right) = \mathcal{N}\left(z_{l(t)}; \sqrt{\hat{\alpha}_t} z_{l(0)}, (1-\hat{\alpha}_t)\mathrm{I}\right), \tag{14}$$

where $\alpha_t = 1 - \beta_t$ and $\hat{\alpha}_t = \prod_{t=1}^{T} \alpha_t$. Using the reparameterization trick, it can sample $z_{l(t)}$ efficiently according to Equation (15):

$$z_{l(t)} = \sqrt{\hat{\alpha}_t} z_{l(0)} + \sqrt{1-\hat{\alpha}_t}\varepsilon, \varepsilon \sim \mathcal{N}(0, \mathrm{I}) \tag{15}$$

The forward process gradually transforms the structured landslide representation into pure Gaussian noise at timestep T, where $z_{l(T)} \sim \mathcal{N}(0, \mathrm{I})$. Operating in latent space rather than image space provides critical advantages: First, the standardized distribution ($\mu = 0, \sigma = 1$) of $z_{l(0)}$ matches the Gaussian noise distribution, enabling natural corruption without distribution mismatch. Second, the low-dimensional representation $h \times w \times d$ vs. $H \times W$ reduces computational complexity by 16× and the continuous latent space ensures smooth state transitions without the artifacts that arise when adding Gaussian noise to discrete binary masks.

Reverse Process and Conditional Denoiser: The reverse process iteratively denoises the corrupted latent representation to recover the clean landslide segmentation, conditioned on

the image embedding z_i. It models the reverse process as a learned Markov chain expressed in Equation (16):

$$p_\theta\left(z_{l(0:T)} \mid z_i\right) \coloneqq p(z_{l(T)}) \prod_{t=1}^{T} p_\theta\left(z_{l(t-1)} \mid z_{l(t)}, z_i\right), \tag{16}$$

where $p\left(z_{l(T)}\right) \sim \mathcal{N}(0, \mathrm{I})$ is the prior distribution and each reverse transition is parameterized as Equation (17):

$$p_\theta\left(z_{l(t-1)} \mid z_{l(t)}, z_i\right) \coloneqq \mathcal{N}\left(z_{l(t-1)}; \mu_\theta\left(z_{l(t)}, z_i, t\right), \sigma_t^2 \mathrm{I}\right), \tag{17}$$

In this research, a conditional denoiser $f_{\text{denoiser}}(\cdot)$ is trained to predict the noise ε added at each timestep t. The denoiser adopts a ResUNet architecture with time-embedding blocks and multi-head self-attention layers at deeper levels. At each timestep, the noisy latent $z_{l(t)}$ is concatenated with the image embedding z_i along the channel dimension, forming a two-channel input $\left[z_{l(t)} \oplus z_i\right] \in \mathbb{R}^{h \times w \times 2d}$. The timestep t is encoded through sinusoidal positional embeddings and injected into the denoiser via adaptive normalization layers. The denoiser predicts the noise component expressed in Equation (18):

$$\hat{\varepsilon}\left(z_{l(t)}, z_i, t\right) = f_{\text{denoiser}}\left(\left[z_{l(t)} \oplus z_i\right], t\right), \tag{18}$$

Following the DDPM formulation, the reverse sampling process at timestep t is given by:

$$z_{l(t-1)} = \frac{1}{\sqrt{\alpha_t}}\left(z_{l(t)} - \frac{\beta_t}{\sqrt{1-\alpha_t}} \hat{\varepsilon}\left(z_{l(t)}, z_i, t\right)\right) + \sigma_t, \tag{19}$$

where $z \sim \mathcal{N}(0, \mathrm{I})$ and σ_t is the noise variance. The image embedding z_i guides the denoising process by providing spatial context about terrain features, vegetation patterns, and topographic characteristics that distinguish landslides from surrounding areas. During training, the denoiser learns to predict noise across all timesteps $t \sim Uniform(\{1, 2, \ldots, T\})$ by minimizing the simplified objective, as defined in Equation (20):

$$\mathcal{L}_{\mathrm{D}iff}(\theta) = \mathbb{E}_{t, z_{l(0)}, \varepsilon}\left[\left\| \varepsilon - \hat{\varepsilon}_\theta\left(\sqrt{\bar{\alpha}_t} z_{l(0)} + \sqrt{1-\bar{\alpha}_t}\varepsilon, z_i, t\right)\right\|^2\right] \tag{20}$$

This objective enables the denoiser to model the gradient $\nabla_{z_{l(t)}} \log p\left(z_{l(t)} \mid z_i\right)$ at different noise levels. For inference, it employs the Denoising Diffusion Implicit Model (DDIM) sampling strategy, which enables deterministic sampling with significantly fewer steps. Starting from pure Gaussian noise $\widetilde{z}_{l(T)} \sim \mathcal{N}(0, I)$, it iteratively applies the denoiser for $t = T, \ldots, 1$ using evenly spaced timesteps. The denoised latent representation z_{dn} is obtained from the final step and passed to the decoder as shown in Equation (21):

$$z_{dn} = z_{l(t)} - \hat{\varepsilon}_\theta\left(z_{l(t)}, z_i, t\right) \tag{21}$$

Remarkably, the experiments show that only 10 sampling steps achieve the same segmentation accuracy as using all $T = 1000$ steps, reducing inference time significantly.

3.4. Boundary-Aware Progressive Decoder

With the denoised latent representation z_{dn} obtained from the reverse diffusion process, it introduces the boundary-aware progressive decoder (Figure 5) that restores full-resolution segmentation maps with refined landslide boundaries through multi-stage upsampling and Reverse Attention (RA) modules.

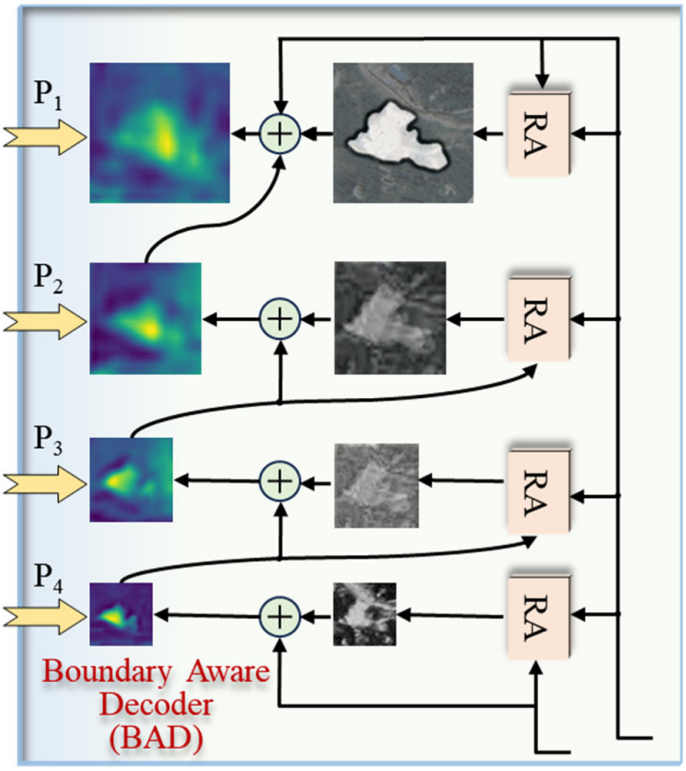

Figure 5. Boundary-aware progressive decoder.

Progressive Upsampling Strategy: The decoder progressively upsamples z_{dn} from $h \times w$ to $H \times W$ through four stages, generating intermediate feature maps F_D^i with increasing resolutions: $F_D^1 \in \mathbb{R}^{H/16 \times W/16 \times 512}$, $F_D^2 \in \mathbb{R}^{H/8 \times W/8 \times 256}$, $F_D^3 \in \mathbb{R}^{H/4 \times W/4 \times 128}$, and $F_D^4 \in \mathbb{R}^{H/2 \times W/2 \times 64}$. This gradual resolution increase enables recovery of fine-grained spatial details while maintaining semantic consistency.

Reverse Attention Modules: To guide the decoder toward uncertain boundary areas, RA modules are introduced at each decoding stage i, as displayed in Figure 6 below. The reverse attention weight a_i is computed by inverting and upsampling the coarse prediction P_{i+1} from the previous stage:

$$a_i = 1 - (\sigma(Up(P_{i+1}))) \tag{22}$$

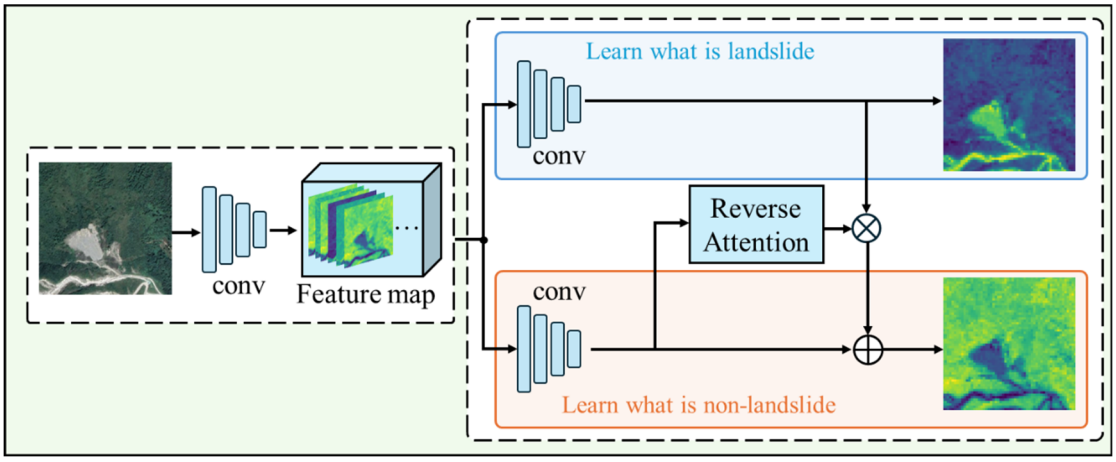

Figure 6. Reverse attention module.

This assigns higher values to uncertain regions near boundaries. The reverse attention features are obtained by element-wise multiplication:

$$r_i = F_D^i \bigodot a_i \tag{23}$$

This mechanism forces the decoder to focus on refining boundary pixels rather than redundantly processing high-confidence regions, which is critical for landslides with irregular shapes and partial occlusions.

Integration with Encoder Features: The decoder receives skip connections from both encoder branches at corresponding scales. At stage i, it concatenates reverse attention features r_i with calibrated encoder features F_C^i and F_T^i:

$$G_i = Concat\left(r_i, F_C^i, F_T^i\right) \tag{24}$$

This fusion combines low-level spatial details from CNN, high-level semantics from Transformer, and boundary-focused features from RA. A residual block processes G_i to produce refined features:

$$F_D^{i+1} = ResBlock(G_i) + Up\left(F_D^i\right) \tag{25}$$

Multi-Scale Supervision: It generates intermediate predictions $\{P_i,\ i = 1,\ \ldots, 4\}$ at each stage by applying 1×1 convolution and softmax:

$$P_i = Softmax\left(Conv_{1\times1}\left(F_D^i\right)\right) \tag{26}$$

These are supervised using downsampled ground-truth masks. The final prediction $\hat{Y} = P_4 \in \mathbb{R}^{H\times W\times C}$ is obtained after the last stage. This multi-scale supervision provides strong gradient signals at intermediate layers, accelerating convergence and improving boundary detail capture across different landslide scales.

3.5. Loss Function and Optimization

To train the BiFusion-LDSeg framework end to end, this research designs a composite loss function that jointly optimizes segmentation accuracy, denoising capability, and multi-scale feature learning. The total loss combines segmentation loss, denoiser loss, and deep supervision across intermediate predictions.

Segmentation Loss: Given an input image X, the final prediction $\hat{Y}$, and intermediate predictions $\{P_i, i = 1, \ldots, 4\}$, this framework employs a combination of cross-entropy loss $\mathcal{L}_{CE}$ and Dice Similarity Coefficient (DSC) loss $\mathcal{L}_{DSC}$. The cross-entropy loss ensures accurate pixel-wise classification:

$$\mathcal{L}_{CE}\left(\hat{Y}, Y\right) = -\sum_{j=1}^{H\times W} Y_j \log\left(\hat{Y}_j\right) \tag{27}$$

where Y_j and $\hat{Y}_j$ denote the ground truth and predicted probabilities at pixel j. The Dice loss addresses class imbalance, which is critical for landslide segmentation where landslide pixels often occupy less than 20% of the image:

$$\mathcal{L}_{DSC}\left(\hat{Y}, Y\right) = 1 - \frac{2\sum_j \hat{Y}_j Y_j}{\sum_j \hat{Y}_j + \sum_j Y_j} \tag{28}$$

The combined segmentation loss for a single prediction is expressed in Equation (29):

$$\mathcal{L}_{\text{Seg}}(\hat{Y}, Y) = \mathcal{L}_{\text{CE}}(\hat{Y}, Y) + \gamma \mathcal{L}_{\text{DSC}}(\hat{Y}, Y) \tag{29}$$

where γ is a weighting coefficient (typically $\gamma = 2$) that balances the two terms.

Denoiser Loss: As introduced in Section 3.3, the denoiser is trained to predict the Gaussian noise added during the forward diffusion process. The denoiser loss minimizes the mean squared error between predicted and true noise in latent space, as expressed in Equation (30):

$$\mathcal{L}_{Diff}(\theta) = \mathbb{E}_{t, z_{l(0)}, \varepsilon}\left[\left\| \varepsilon - \hat{\varepsilon}_\theta\left(\sqrt{\bar{\alpha}_t} z_{l(0)} + \sqrt{1-\bar{\alpha}_t}\varepsilon, z_i, t\right)\right\|^2\right] \tag{30}$$

Operating in latent space rather than image space reduces computational cost while maintaining denoising effectiveness. This loss enables the denoiser to learn score functions at different noise levels, facilitating accurate sampling during inference.

Deep Supervision Loss: To enhance gradient flow and improve multi-scale boundary learning, this research applies supervision to all intermediate predictions $\{P_i, i = 1, \ldots, 4\}$. Each intermediate prediction is supervised using the ground-truth mask Y downsampled to the corresponding resolution as expressed in Equation (31):

$$\mathcal{L}_{Deep} = \sum_{i=1}^{4} w_i \mathcal{L}_{\text{Seg}}(P_i, Y) \tag{31}$$

where w_i are stage-specific weights that emphasize later stages with finer resolutions. In this research, $w_1 = 0.5$, $w_2 = 0.7$, $w_3 = 0.9$, and $w_4 = 1.0$, gradually increasing supervision strength as features approach full resolution. This strategy provides strong learning signals at early decoder stages, preventing gradient vanishing and accelerating convergence.

Total Loss Function: The complete training objective combines all loss components:

$$\mathcal{L}_{Total} = \mathcal{L}_{\text{Seg}}(\hat{Y}, Y) + \lambda \mathcal{L}_{Diff}(\theta) + \mathcal{L}_{\text{Deep}} \tag{32}$$

where λ balances the segmentation and denoising objectives. Here, it empirically sets $\lambda = 1.0$ to give equal importance to both tasks.

3.6. Training Strategy

The training strategy of the proposed framework is simple and straight forward. Unlike two-stage approaches that train VAE and denoiser separately, BiFusion-LDSeg employs joint optimization of all components—dual encoder, label encoder, denoiser, and decoder—via unified backpropagation. Gradients from the combined loss $\mathcal{L}_{Total}$ (Equation (32)) flow through the entire pipeline: segmentation loss $\mathcal{L}_{Seg}$ updates the decoder, denoiser loss $\mathcal{L}_{Diff}$ optimizes noise prediction, and both losses jointly refine the encoders. This enables the label encoder to learn segmentation-specific latent representations rather than generic reconstructions, the image encoder to extract features optimal for denoiser conditioning, and the decoder to adapt to denoised latent characteristics.

To enhance model robustness and prevent overfitting on limited landslide datasets, extensive data augmentation is applied, including random rotation ($\pm 180^\circ$), horizontal and vertical flipping, and random scaling (0.8–1.2×). For satellite imagery, it additionally applies color jittering (brightness ± 20%, contrast ± 20%, saturation ± 15%) to account for seasonal variations, different acquisition times, and atmospheric conditions. Random Gaussian blur ($\sigma \in [0.5, 1.5]$) simulates varying image quality. All augmentations are applied consistently to both input images and ground-truth masks.

The framework is optimized using the Adam optimizer with an initial learning rate of 7×10^{-5} and exponentially decaying learning rate following a cosine schedule. The model is trained for 600 epochs with a batch size of 8. All images are resized to 512 × 512 pixels. During training, timestep t is uniformly sampled from $\{1, \ldots, T\}$ for each batch. Gradient clipping with a maximum norm of 1.0 is applied to stabilize training. Weight decay of 1×10^{-4} is used for regularization to prevent overfitting on the relatively small landslide datasets.

3.7. Measurement Metrics

To assess the segmentation performance, four measurement metrics are selected in this study, including the Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Average Symmetric Surface Distance (ASSD), and Hausdorff Distance (HD) [47]. All four measurement metrics are expressed in Equations (33)–(36) as follows:

$$IoU = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|} \tag{33}$$

$$DSC = 1 - 2\frac{|\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} \tag{34}$$

$$ASSD = \frac{\sum_{\hat{y} \in \hat{Y}} \min_{y \in Y} d(\hat{y}, y) + \sum_{y \in Y} \min_{\hat{y} \in \hat{Y}} d(\hat{y}, y)}{|\hat{Y}| + |Y|} \tag{35}$$

$$HD(\hat{Y}, Y) = \max(\mathrm{hd}(\hat{Y}, Y), \mathrm{hd}(Y, \hat{Y})) \tag{36}$$

where Y denotes the ground-truth mask and $\hat{Y}$ denotes the segmentation output; y and $\hat{y}$ represents the pixels within ground-truth mask and segmentation contour; and $\mathrm{hd}(\hat{Y}, Y)$ represents the Hausdorff distance between the two masks, which can be expressed in Equation (37) below:

$$\mathrm{hd}(\hat{Y}, Y) = \max_{\hat{y} \in Y} \min_{y \in Y} \|\hat{y} - y\|_2 \tag{37}$$

4. Experimental Results

4.1. Datasets and Study Areas

To comprehensively evaluate the proposed BiFusion-LDSeg framework, experiments were conducted on three earthquake-triggered coseismic landslide datasets from different seismic events in Sichuan Province, China: Jiuzhaigou (2017, Mw 6.5), Luding (2022, Mw 6.8), and Lushan (2013, Mw 7.0) earthquakes, as illustrated in Figure 7. Here, Figure 7a displays the Digital Elevation Model (DEM) of Sichuan Province, China. Figure 7b reflects the landslide distribution in the Luding earthquake region (pink diamonds). Figure 7c shows the landslide distribution in the Lushan earthquake region (yellow diamonds), and Figure 7d shows the landslide distribution in the Jiuzhaigou earthquake region (blue triangles). These regions are characterized by complex mountainous terrain with elevations ranging from 500 to 7000 m, steep slopes exceeding 30°, and diverse geological conditions including fractured rock masses and active fault zones. The selected study areas represent varying landslide densities, sizes, and morphological characteristics, providing robust evaluation scenarios for testing model generalization across different geological and seismic settings. All three datasets were collected from high-resolution satellite imagery (0.5–2 m spatial resolution) obtained through Google Earth Platform, capturing both pre- and post-earthquake conditions. Ground-truth landslide boundaries were meticulously delineated by domain experts following established interpretation protocols and validated through field surveys where accessible.

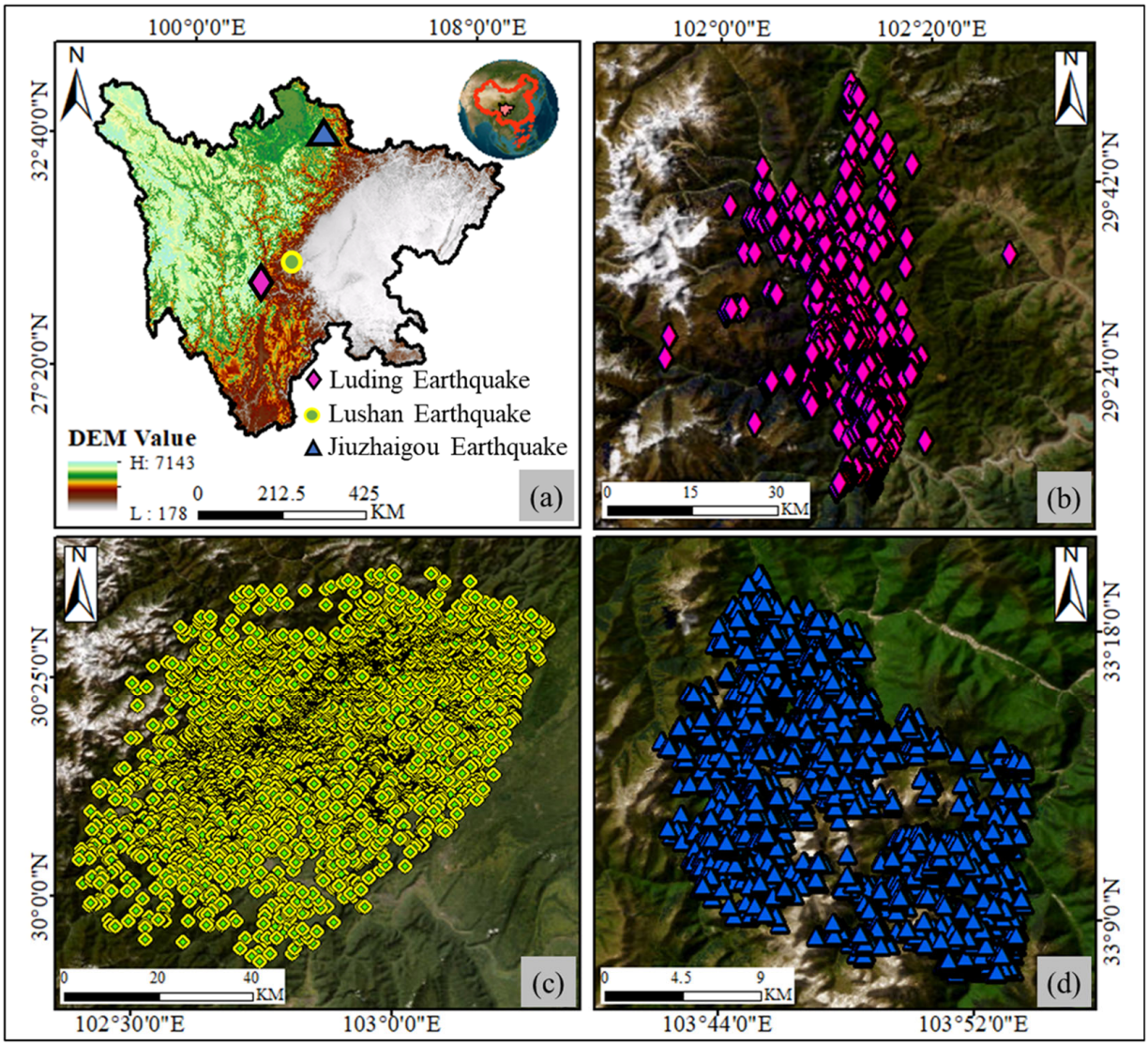

Figure 7. Geographical distribution of study areas and landslide inventories. (**a**) Location of study case; (**b**) Landslide distribution in the Luding earthquake region; (**c**) Landslide distribution in the Lushan earthquake region; (**d**) Landslide distribution in the Jiuzhaigou earthquake region.

As summarized in Table 1, the complete dataset comprises 24,800 high-resolution satellite image patches: 4800 from Jiuzhaigou, 5000 from Luding, and 15,000 from Lushan regions. Each dataset is randomly partitioned following an 80%-10%-10% split strategy for training, validation, and testing, respectively, ensuring sufficient samples for model optimization while maintaining independent evaluation sets. The Jiuzhaigou dataset contains landslides predominantly triggered by the 2017 Mw 6.5 earthquake in the eastern Tibetan Plateau, characterized by shallow soil slides and rockfalls with areas ranging from 100 to 50,000 m^2. The Luding dataset captures landslides induced by the 2022 Mw 6.8 earthquake along the Xianshuihe fault zone, featuring larger-scale deep-seated failures and debris flows (500–100,000 m^2). The Lushan dataset, derived from the 2013 Mw 7.0 earthquake in the Longmenshan fault system, exhibits the most diverse landslide inventory with sizes spanning three orders of magnitude (50–200,000 m^2) and includes various failure types such as rock avalanches, translational slides, and complex slope movements.

Table 1. Statistical summary of coseismic landslide datasets.

Dataset	Total Images	Training (80%)	Validation (10%)	Testing (10%)	Earthquake Event
Jiuzhaigou	4800	3840	480	480	Mw 6.5 (2017)
Luding	5000	4000	480	480	Mw 6.8 (2022)
Lushan	15,000	12,000	480	480	Mw 7.0 (2013)
Total	24,800	19,840	2480	2480	

4.2. Implementation Details

The proposed BiFusion-LDSeg framework is implemented using PyTorch 1.13.0 and trained on NVIDIA A100 GPUs (80 GB memory) with CUDA 11.7. The dual-encoder architecture employs ResNet-50 as the CNN backbone and a Vision Transformer (ViT) with 12 blocks organized into four stages (3 blocks per stage) for the Transformer branch. The label encoder and decoder follow symmetric ResUNet-style architectures with four downsampling and upsampling stages, respectively, producing latent representations of size $32 \times 32 \times 512$ for all three datasets. The conditional denoiser adopts a U-Net architecture with time-embedding blocks and multi-head self-attention layers at resolutions 16×16 and 8×8, resulting in approximately 87.3 million trainable parameters for the complete framework.

To ensure statistical reliability, all experiments were repeated five times with different random seeds. Reported results show mean $\pm$ standard deviation across these runs. Statistical significance between methods was assessed using paired t-tests with significance threshold $p < 0.05$. All models are trained for 600 epochs using the Adam optimizer with an initial learning rate of 7×10^{-5} with a cosine annealing schedule, and a batch size of 8, with the total diffusion timesteps set to $T = 1000$ using cosine noise scheduling. Loss function weights are configured as $\lambda = 1.0$ for balancing segmentation and denoiser losses, $\gamma = 2$ for the Dice loss, and intermediate supervision weights $\{w_1, w_2, w_3, w_4\} = \{0.5,\ 0.7,\ 0.9,\ 1.0\}$ for multi-scale predictions, while early stopping with patience of 50 epochs is applied based on validation IoU. During inference, this research employs the DDIM sampling algorithm with only 10 evenly spaced sampling steps (versus 1000 during training), achieving optimal segmentation accuracy within approximately 0.317 s per 512×512 image without requiring any post-processing operations, as the boundary-aware decoder naturally produces clean landslide boundaries.

Figure 8 illustrates the training convergence behavior, where the top row (a, c, e) displays training loss components (total loss, segmentation loss, diffusion loss, and deep supervision loss) over 600 epochs for the Lushan, Jiuzhaigou, and Luding datasets. The bottom row (b, d, f) illustrates performance metrics (Dice coefficients and mean IoU) demonstrating consistent improvement and stabilization across all datasets. All loss components (total, segmentation, diffusion, and deep supervision) decrease steadily and stabilize after approximately 400 epochs across all three datasets, while validation metrics DSC consistently improve and plateau, demonstrating effective model learning and generalization without overfitting.

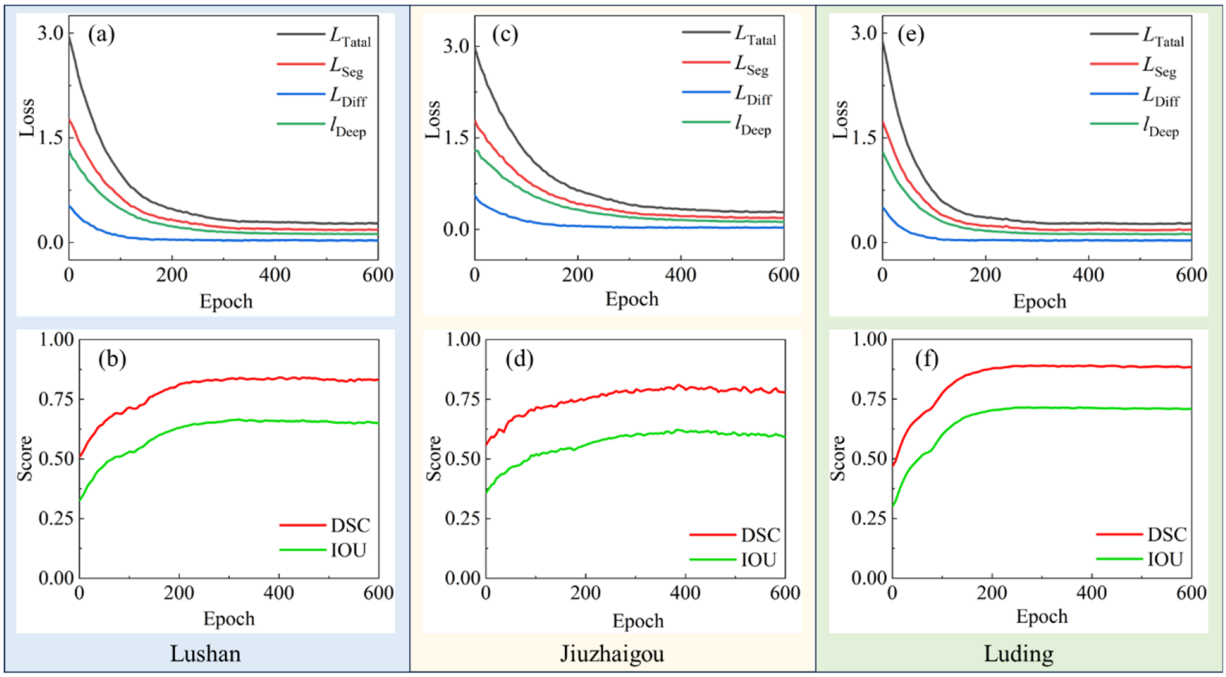

Figure 8. Training convergence and validation performance curves for three datasets. (**a**) training loss components for the Lushan datasets; (**b**) training loss components for the Jiuzhaigou datasets; (**c**) training loss components for the Luding datasets; (**d**) performance metrics for the Lushan datasets; (**e**) performance metrics for the Jiuzhaigou datasets; (**f**) performance metrics for the Luding datasets.

Figure 9 presents the performance–efficiency tradeoff across varying DDIM sampling steps. A clear three-phase behavior is observed across all datasets: rapid performance improvement from steps 1 to 5, stabilization at step 10 representing the elbow point, and negligible further gains beyond step 10. Additionally, average IoU improves by less than 0.3% from step 10 to step 100, while inference time increases by 10× (0.317 s to 3.171 s). Step 10 achieves an average IoU of 69.2% and DSC of 81.8%, confirming it as the optimal efficiency–accuracy operating point for rapid post-earthquake deployment.

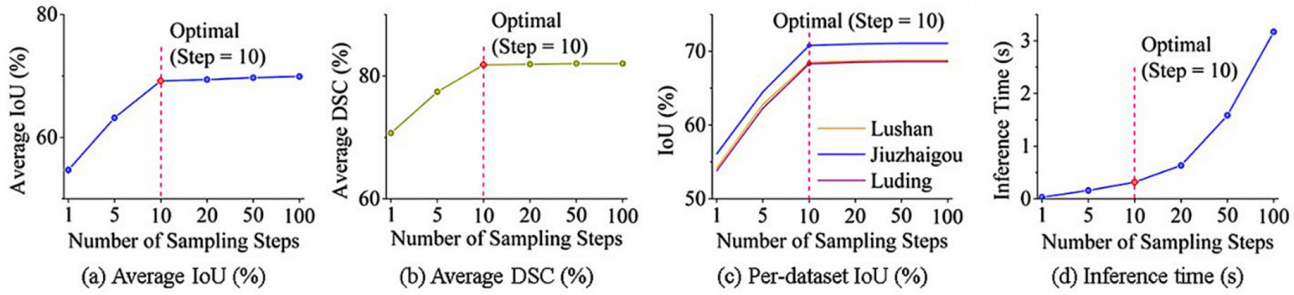

Figure 9. Segmentation performance and inference efficiency of BiFusion-LDSeg across varying DDIM sampling steps.

4.3. Comparison with the SOTA Methods

This subsection comprehensively evaluates the proposed BiFusion-LDSeg against eleven state-of-the-art segmentation methods spanning five categories: traditional CNNs (U-Net, U-Net++, DeepLabv3+), attention-based methods (Attention U-Net, DANet), Transformer-based architectures (Swin-UNet, SegFormer), hybrid CNN-Transformer models (TransUNet), and generative approaches (MedSegDiff, LSegDiff). To ensure statistical rigor, all experiments were repeated five times with different random seeds, and results are reported as mean ± standard deviation. Statistical significance of performance differences

was assessed using paired t-tests ($p < 0.05$). Tables 2–4 present quantitative comparisons across the Lushan, Jiuzhaigou, and Luding datasets respectively, evaluated using four metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Average Symmetric Surface Distance (ASSD), and Hausdorff Distance (HD). BiFusion-LDSeg consistently achieves superior performance across all three datasets and metrics, demonstrating robust generalization capability. Notably, on the challenging Lushan dataset with extreme size variations, the proposed method achieves IoU of 68.5 ± 1.2% and DSC of 81.2 ± 0.9%, while maintaining significantly better boundary accuracy with ASSD of 3.0 ± 0.3 pixels versus 4.2 ± 0.5 pixels. The performance gains are consistent across the Jiuzhaigou (IoU: 70.8 ± 1.0%, DSC: 83.1 ± 0.8%) and Luding (IoU: 68.3 ± 1.1%, DSC: 81.1 ± 0.9%) datasets, validating the effectiveness of bi-directional fusion enhanced latent diffusion for earthquake-triggered landslide segmentation under diverse geological conditions.

Table 2. Quantitative comparison on the Lushan dataset.

Method	Reference	IoU	DSC	ASSD	HD
Traditional CNNs					
U-Net	Ronneberger et al. 2015 [35]	58.2 ± 1.8	73.5 ± 1.5	5.4 ± 0.7	38.6 ± 3.8
U-Net++	Zhou et al. 2018 [48]	59.4 ± 1.6	74.5 ± 1.3	5.1 ± 0.6	36.5 ± 3.5
DeepLabv3+	Chen et al. 2018 [49]	59.8 ± 1.7	74.8 ± 1.4	4.9 ± 0.6	35.4 ± 3.6
Attention-based					
Attention U-Net	Oktay et al. 2018 [37]	61.0 ± 1.5	75.8 ± 1.2	4.7 ± 0.5	33.8 ± 3.2
DANet	Fu et al. 2019 [50]	61.5 ± 1.4	76.2 ± 1.1	4.5 ± 0.5	32.7 ± 3.0
Transfomer-based					
Swin-UNet	Cao et al. 2022 [40]	62.1 ± 1.3	76.7 ± 1.0	4.3 ± 0.4	31.6 ± 2.8
SegFormer	Xie et al. 2021 [51]	62.6 ± 1.2	77.1 ± 0.9	4.2 ± 0.4	30.8 ± 2.7
Hybrid TransUNet	Chen et al. 2021 [23]	63.1 ± 1.3	77.4 ± 1.0	4.0 ± 0.4	30.0 ± 2.6
Generative					
MedSegDiff	Wu et al. 2024 [30]	61.9 ± 1.4	76.4 ± 1.1	4.4 ± 0.5	32.2 ± 2.9
LSegDiff	Vu et al. 2023 [52]	64.6 ± 1.3	78.5 ± 1.0	4.2 ± 0.5	30.4 ± 2.7
Proposed BiFusion-LDSeg		68.5 ± 1.2	81.2 ± 0.9	3.3 ± 0.3	21.0 ± 2.1

Table 3. Quantitative comparison on the Jiuzhaigou dataset.

Method	Reference	IoU	DSC	ASSD	HD
Traditional CNNs					
U-Net	Ronneberger et al. 2015 [35]	60.1 ± 1.6	75.1 ± 1.3	5.0 ± 0.6	36.5 ± 3.5
U-Net++	Zhou et al. 2018 [48]	61.2 ± 1.5	75.9 ± 1.2	4.7 ± 0.6	34.6 ± 3.2
DeepLabv3+	Chen et al. 2018 [49]	61.6 ± 1.4	76.2 ± 1.1	4.6 ± 0.5	33.6 ± 3.1
Attention-based					
Attention U-Net	Oktay et al. 2018 [37]	62.7 ± 1.3	77.1 ± 1.0	4.3 ± 0.5	32.2 ± 2.9
DANet	Fu et al. 2019 [50]	63.2 ± 1.2	77.4 ± 0.9	4.2 ± 0.4	31.2 ± 2.7
Transfomer-based					
Swin-UNet	Cao et al. 2022 [40]	63.8 ± 1.2	77.8 ± 0.9	4.0 ± 0.4	30.2 ± 2.6
SegFormer	Xie et al. 2021 [51]	64.2 ± 1.1	78.2 ± 0.8	3.9 ± 0.4	29.4 ± 2.5
Hybrid TransUNet	Chen et al. 2021 [23]	64.8 ± 1.2	78.6 ± 0.9	3.7 ± 0.4	28.7 ± 2.4
Generative					
MedSegDiff	Wu et al. 2024 [30]	63.5 ± 1.3	77.6 ± 1.0	4.1 ± 0.5	30.8 ± 2.8
LSegDiff	Vu et al. 2023 [52]	66.9 ± 1.2	80.0 ± 0.9	3.8 ± 0.4	29.1 ± 2.5
Proposed BiFusion-LDSeg		70.8 ± 1.0	83.1 ± 0.8	2.5 ± 0.3	19.8 ± 2.0

Table 4. Quantitative comparison on the Luding dataset.

Method	Reference	IoU	DSC	ASSD	HD
Traditional CNNs					
U-Net	Ronneberger et al. 2015 [35]	59.4 ± 1.7	74.5 ± 1.4	5.2 ± 0.7	37.4 ± 3.6
U-Net++	Zhou et al. 2018 [48]	60.5 ± 1.6	75.3 ± 1.3	4.9 ± 0.6	35.4 ± 3.4
DeepLabv3+	Chen et al. 2018 [49]	60.9 ± 1.5	75.7 ± 1.2	4.8 ± 0.6	34.3 ± 3.3
Attention-based					
Attention U-Net	Oktay et al. 2018 [37]	62.1 ± 1.4	76.6 ± 1.1	4.5 ± 0.5	32.8 ± 3.1
DANet	Fu et al. 2019 [50]	62.6 ± 1.3	77.0 ± 1.0	4.4 ± 0.5	31.8 ± 2.9
Transfomer-based					
Swin-UNet	Cao et al. 2022 [40]	63.2 ± 1.3	77.5 ± 1.0	4.2 ± 0.4	30.8 ± 2.7
SegFormer	Xie et al. 2021 [51]	63.7 ± 1.2	77.9 ± 0.9	4.0 ± 0.4	30.0 ± 2.6
Hybrid TransUNet	Chen et al. 2021 [23]	64.2 ± 1.2	78.3 ± 0.9	3.9 ± 0.4	29.3 ± 2.5
Generative					
MedSegDiff	Wu et al. 2024 [30]	63.0 ± 1.4	77.2 ± 1.0	4.3 ± 0.5	31.4 ± 2.8
LSegDiff	Vu et al. 2023 [52]	64.4 ± 1.2	78.4 ± 0.9	4.0 ± 0.4	29.7 ± 2.6
Proposed BiFusion-LDSeg		68.3 ± 1.1	81.1 ± 0.9	2.9 ± 0.5	20.4 ± 2.4

Qualitative comparisons illustrated in Figures 10–12 provide deeper insights into model capabilities across varying difficulty levels and boundary precision requirements. Figure 10 demonstrates that for easy-to-medium cases (Cases 1–9) with clear landslide signatures and relatively regular boundaries, the proposed BiFusion-LDSeg produces segmentation masks nearly identical to ground truth across all three datasets, with prediction uncertainty heatmaps showing high confidence (red regions) concentrated within true landslide areas and minimal false activations in background regions. Figure 11 reveals critical advantages in challenging scenarios (Cases 10–18) involving small landslides, irregular shapes, partial occlusions, and spectral ambiguities (<3% of total cases), where the proposed methods exhibit substantial over-segmentation (Cases 13, 15) or miss fragmented landslide components (Cases 10, 11, 17). Figure 12 provides detailed boundary accuracy analysis for six representative cases (Cases 19–24). Here, the zoomed-in boundary overlays demonstrate that predicted boundaries (blue contours) closely align with ground-truth boundaries (red contours) even for complex curved scarps and narrow debris tails. The uncertainty quantification capability inherent to the proposed diffusion-based approach enables reliability assessment for operational deployment, with high-confidence predictions suitable for automated mapping and medium-confidence regions flagged for expert review.

Figure 13 illustrates the progressive boundary refinement mechanism through the multi-stage boundary-aware decoder, validating the design principles discussed in Section 3.4. The intermediate predictions P_1 through P_4 demonstrate systematic improvement at each decoder stage, where coarse predictions from earlier stages P_1 and P_2 capture overall landslide extent but exhibit smoothed boundaries, while later stages P_3 and P_4 progressively restore fine-grained boundary details through reverse attention mechanisms that explicitly focus on uncertain boundary pixels, ultimately producing final segmentations that accurately delineate irregular landslide boundaries including concave scarps, narrow runout zones, and fragmented debris deposits across all three seismic regions.

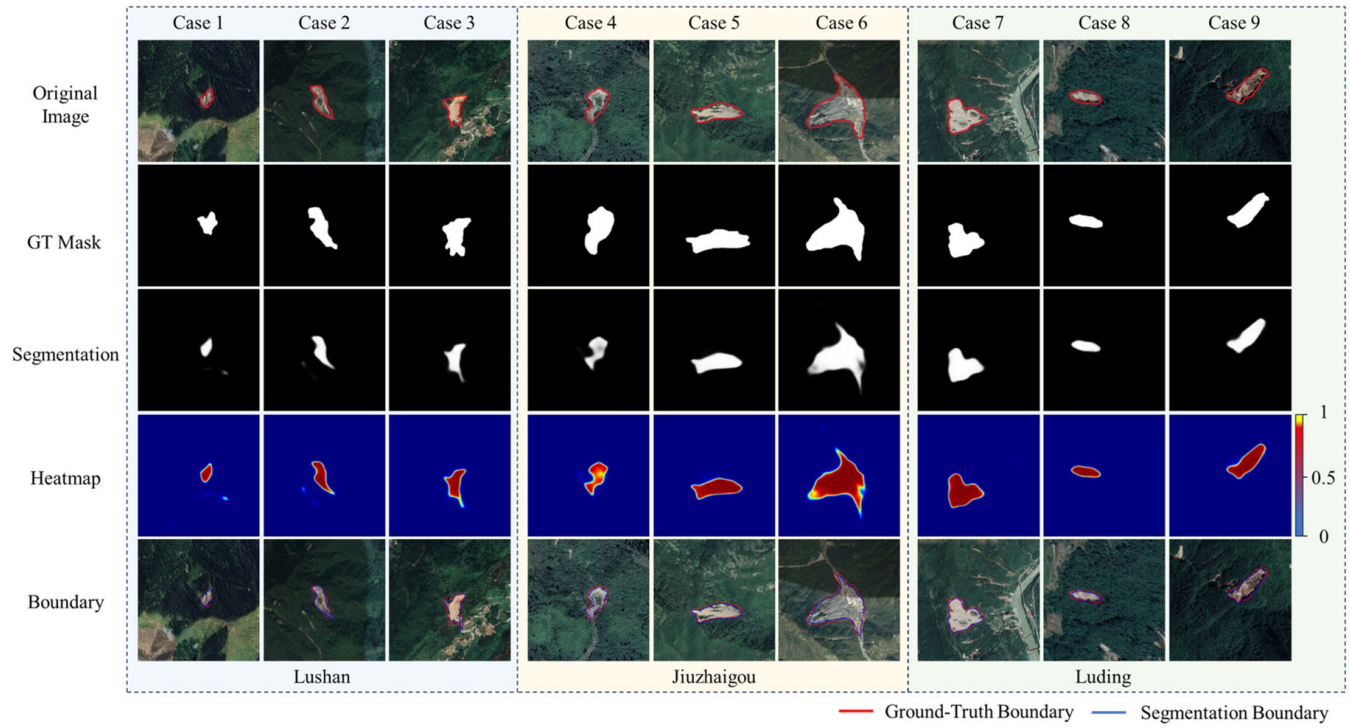

Figure 10. Qualitative segmentation review for easy-to-medium cases across three datasets.

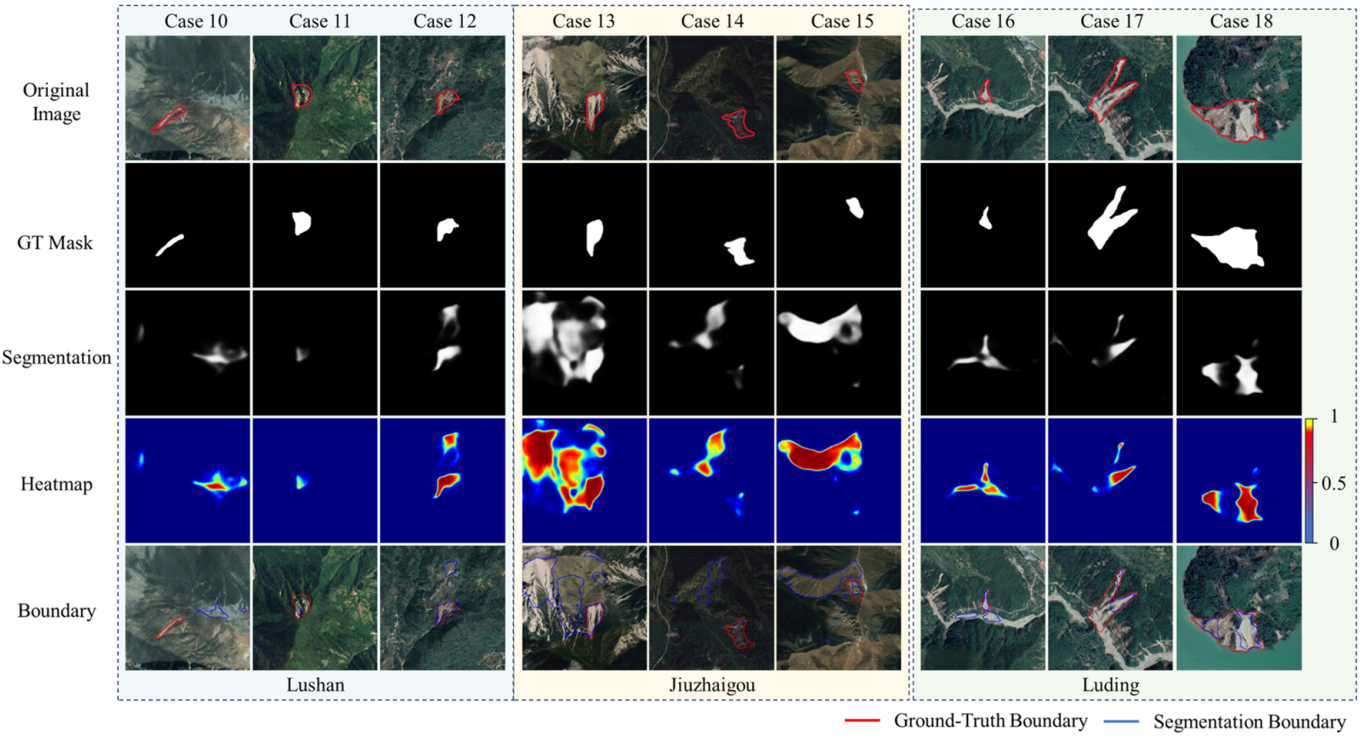

Figure 11. Qualitative segmentation review for challenging cases across three datasets.

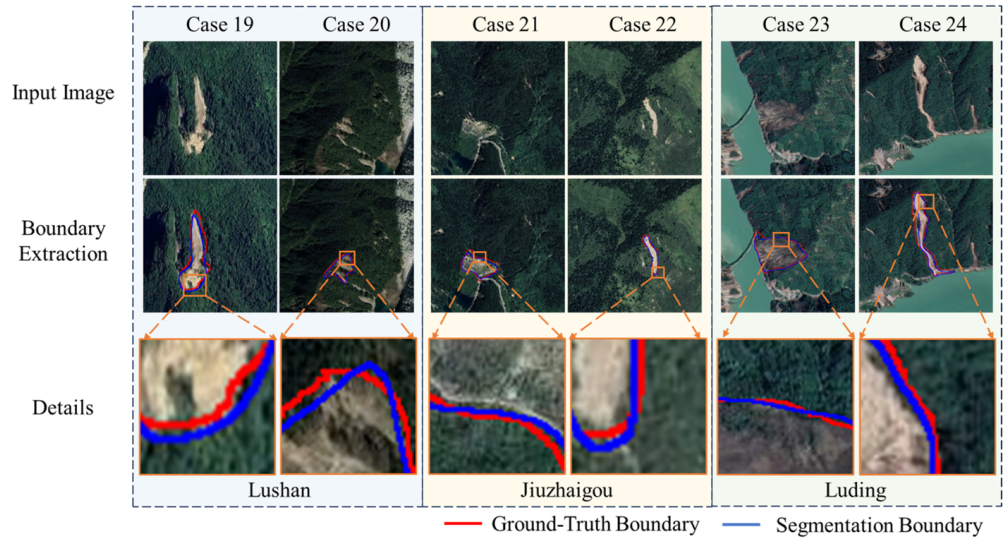

Figure 12. Boundary accuracy visualization for six representative cases.

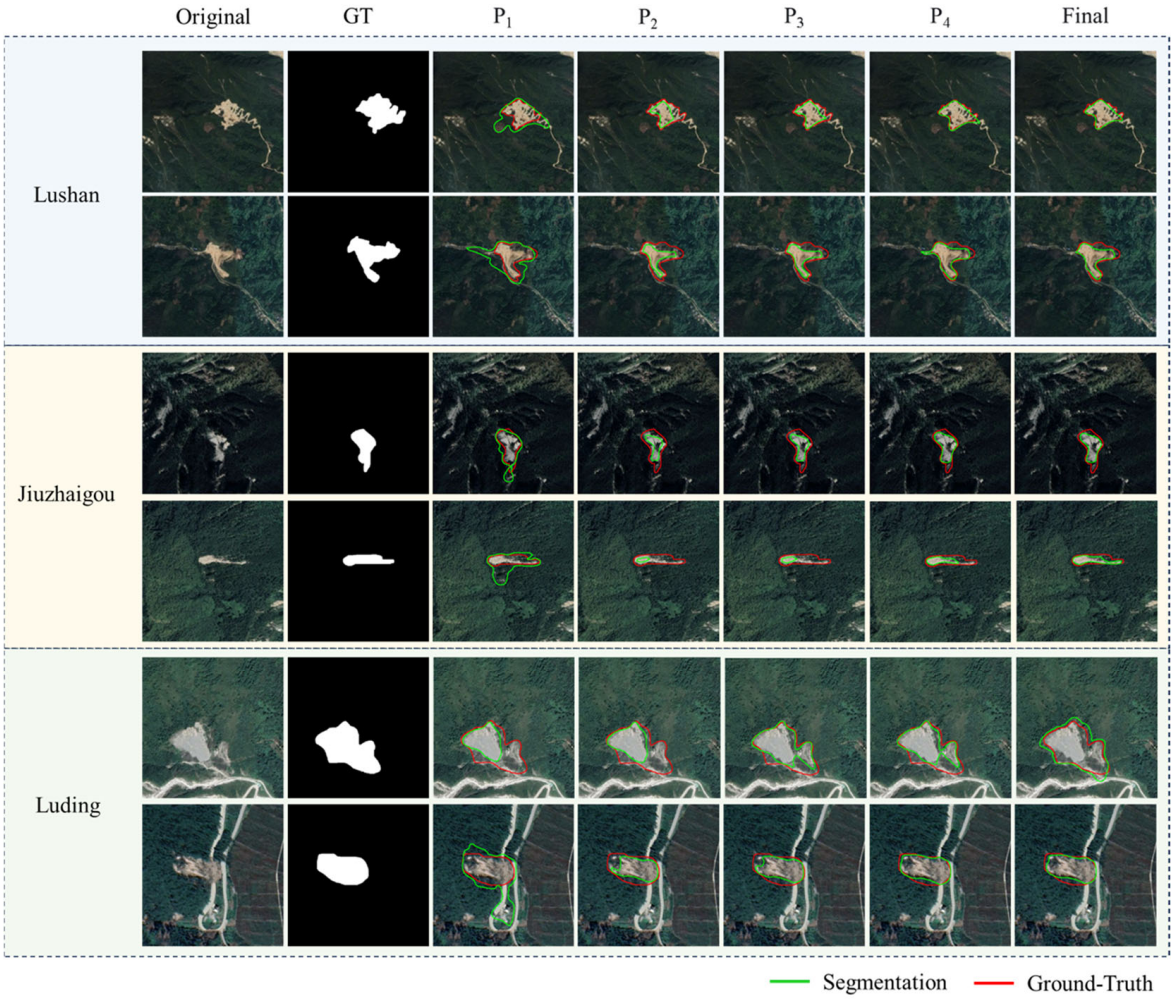

Figure 13. Progressive boundary refinement through multi-stage boundary-aware decoder.

4.4. Ablation Study

To validate the effectiveness of each proposed component, comprehensive ablation studies were conducted by systematically removing or replacing key modules and evaluating their impact on segmentation performance. To provide a comprehensive view

of each component's contribution, Table 5 presents ablation results averaged across all three datasets (Lushan, Jiuzhaigou, and Luding), while dataset-specific baseline comparisons are detailed in Tables 2–4. As illustrated in Figure 14 and Table 5, the baseline model employing only the CNN encoder with the standard U-Net decoder achieves 76.6% DSC and 63.0% IoU. The model then progressively adds proposed components and yields substantial improvements: bi-directional attention fusion contributes +3.8% DSC; latent diffusion adds +0.3% DSC; boundary-aware decoder provides +0.3% DSC; and end-to-end training contributes +0.6% DSC. The full framework achieves 81.8% DSC and 69.2% IoU by adding all the above components cumulatively. Notably, the Bi-AG module with bi-directional information exchange (80.4% DSC) significantly outperforms single-direction attention variants (C→T: 79.5%, T→C: 80.1%) and simple concatenation (79.2%), validating the effectiveness of mutual feature calibration between the CNN and Transformer branches. The boundary-aware decoder with reverse attention modules improves ASSD from 3.4 pixels (standard decoder) to 2.8 pixels (proposed decoder), demonstrating superior boundary delineation capability while maintaining reasonable inference efficiency at 0.31 s per 512 × 512 image.

Table 5. Component-wise ablation study results (average across three datasets).

Model Variant	IoU (%)	DSC (%)	ASSD	Inference Time (s)	Description
Baseline	62.1	76.6	4.8	0.022	Standard U-Net decoder
Encoder Ablations					
CNN Encoder only	63.8	77.9	4.3	0.028	Without Transformer branch
Transformer only	62.9	77.2	4.5	0.038	Without CNN branch
Simple Concatenation	64.3	78.3	4.0	0.045	CNN + Trans, no Bi-AG
One-way Attention (C→T)	65.6	79.2	3.6	0.052	CNN calibrates Trans
One-way Attention (T→C)	65.3	79.0	3.7	0.051	Trans calibrates CNN
Bi-AG Fusion (Proposed)	66.7	80.0	3.3	0.058	Proposed bi-directional fusion
Diffusion Ablations					
Without Latent Diffusion	66.8	80.1	3.4	0.025	Deterministic decoder
Image-space Diffusion	67.2	80.4	3.2	0.282	Traditional diffusion
Latent Diffusion (Proposed)	67.6	80.7	3.1	0.313	Proposed latent approach
Decoder Ablations					
Standard Decoder	67.4	80.5	3.3	0.284	Without RA modules
Boundary-Aware Decoder (Proposed)	68.2	81.1	3.1	0.305	With RA modules
Training Strategy Ablations					
Two-stage Training	66.9	80.2	3.4	0.311	Encoder→Denoiser
End-to-End (Proposed)	68.2	81.1	2.9	0.317	Joint optimization
Full Framework (Proposed)	69.2	81.8	2.8	0.317	All components

Note: Results represent mean performance across Lushan, Jiuzhaigou, and Luding datasets.

To verify that BiFusion-LDSeg's performance gains stem from the diffusion mechanism rather than increased model capacity, this study designed a parameter-matched non-diffusion baseline (BiFusion-Det) by replacing the latent diffusion component with a scaled-up deterministic decoder while keeping all other components identical, yielding approximately 87.1 M parameters. As shown in Table 6, BiFusion-Det achieves 67.8% IoU and 80.9% DSC—a modest improvement over the standard non-diffusion baseline (66.8% IoU/80.1% DSC) attributable solely to increased capacity. BiFusion-LDSeg nonetheless outperforms BiFusion-Det by a further 1.4% IoU and 0.9% DSC, confirming that the latent diffusion mechanism contributes genuine and distinct performance gains beyond parameter scaling. Additionally, the diffusion mechanism uniquely provides uncertainty quantification and noise robustness capabilities that parameter scaling alone cannot replicate.

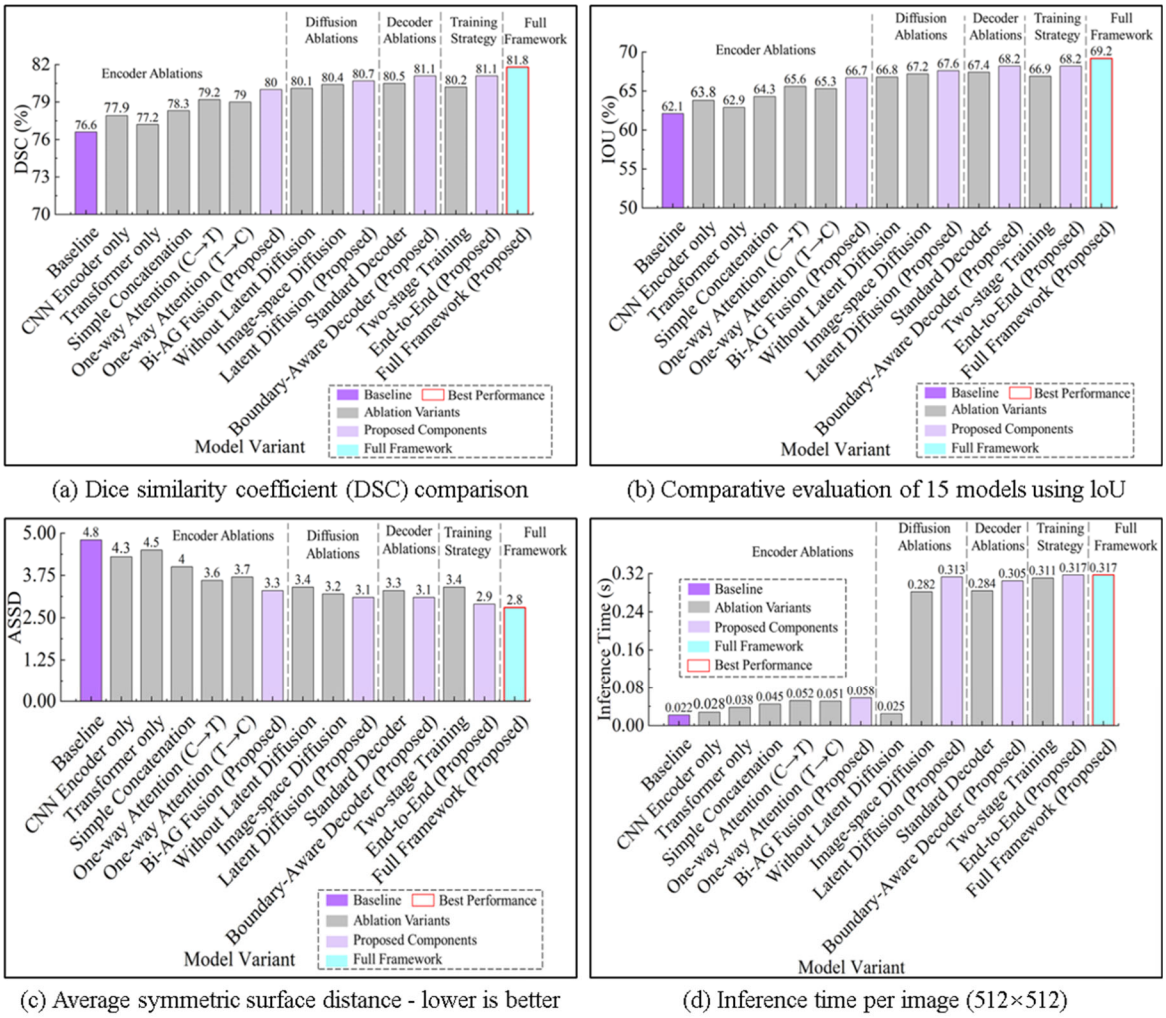

Figure 14. Comprehensive ablation study results across three datasets.

Table 6. Parameter-matched baseline Comparison for Validating the Latent Diffusion Contribution.

Model Variant	IoU (%)	DSC (%)	ASSD	Inference Time (s)	Parameters (M)
Without Latent Diffusion (standard)	66.8	80.1	3.4	0.025	52.1
BiFusion-Det (parameter-matched)	678	80.9	3.2	0.038	87.1
Image-Space Diffusion	67.2	80.4	3.2	0.282	87.3
BiFusion-LDSeg (proposed)	69.2	81.8	2.8	0.317	87.3

Note: Results represent mean performance across Lushan, Jiuzhaigou, and Luding datasets. BiFusion-Det denotes the parameter-matched non-diffusion baseline.

Table 7 presents the loss function component analysis, demonstrating the synergistic effects of combining multiple loss terms for optimal performance. The combined cross-entropy and Dice loss with γ = 2.0 outperforms using either loss alone (CE only: 78.6% DSC, Dice only: 78.9% DSC) by achieving 80.4% DSC, effectively addressing class imbalance in landslide segmentation. The diffusion loss with λ = 1.0 contributes an additional 0.8% DSC improvement and accelerates convergence from 425 to 395 epochs, while deep supervision with progressive weights {0.5, 0.7, 0.9, 1.0} further enhances boundary accuracy by 0.6% DSC. The complete loss configuration achieves 81.8% DSC with the fastest convergence at 385 epochs, validating the proposed end-to-end joint optimization strategy.

Table 7. Loss function component analysis (average across three datasets).

Loss Configuration	IoU (%)	DSC (%)	ASSD (Pixels)	Training Convergence (Epochs)
CE only (γ = 0)	64.8	78.6	4.3	520
Dice only (CE removed)	65.2	78.9	4.1	495
CE + Dice (γ = 1.0)	66.5	79.8	3.7	465
CE + Dice (γ = 2.0)	67.3	80.4	3.5	425
CE + Dice (γ = 3.0)	66.9	80.1	3.6	440
w/o Diffusion Loss (λ = 0)	65.1	79.1	4.0	485
Diffusion Loss (λ = 0.5)	67.8	80.6	3.3	410
Diffusion Loss (λ = 1.0)	68.5	81.2	3.0	395
Diffusion Loss (λ = 2.0)	68.1	80.9	3.1	405
w/o Deep Supervision	66.3	79.5	3.8	470
With Supervision (w = {0.5, 0.7, 0.9, 1.0})	68.0	80.8	3.2	405
Full Loss (γ = 2.0, λ = 1.0, w = {0.5, 0.7, 0.9, 1.0})	69.2	81.8	2.8	385

5. Discussion

Having demonstrated in Section 4 superior quantitative and qualitative performance, this section provides an in-depth analysis of BiFusion-LDSeg's key advantages: noise robustness, cross-dataset generalization, uncertainty quantification, and learned attention patterns, which reveal the framework's multi-scale feature discrimination capabilities.

Figure 15 demonstrates the superior noise robustness of BiFusion-LDSeg compared to baseline methods across varying noise levels (σ = 0 to 0.3) on all three datasets. While traditional deterministic models (U-Net, ResUNet, DeepLabv3+) exhibit rapid performance degradation with increasing noise—with DSC scores dropping by 15–25% at σ = 0.3. The proposed framework maintains relatively stable performance with only 8–12% degradation under identical noise conditions. This enhanced robustness can be attributed to the learned low-dimensional image embeddings that effectively filter high-frequency noise while preserving essential landslide-discriminative features, combined with the learned shape manifolds in latent space that act as strong geometric priors during the diffusion process.

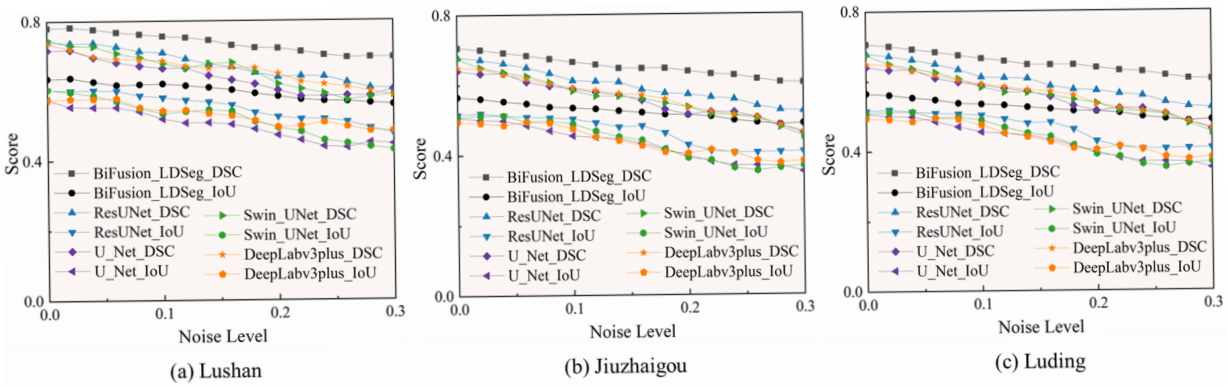

Figure 15. Robustness to atmospheric noise.

Table 8 presents size-stratified segmentation performance across four landslide area categories. A consistent performance gradient is observed across all methods and datasets, with accuracy declining sharply for smaller targets. BiFusion-LDSeg outperforms all baselines at every scale, achieving IoU of 79.8–81.5% for large landslides but dropping to 32.8–35.1% for very small landslides (<50 m^2), which at the imagery resolution of 0.5–2 m correspond to targets of fewer than 20 pixels. This steep decline is expected, as the 32 × 32 latent representation inherently suppresses fine spatial detail at sub-pixel scales. Promising directions to improve small-target detection include scale-aware loss weighting,

higher-resolution decoder pathways, and super-resolution preprocessing. These remain important directions for future work.

Table 8. Size-stratified segmentation performance (IoU%/DSC%) across landslide area categories on three datasets.

Dataset	Size	Proportion	U-Net IoU/DSC	TransUNet IoU/DSC	BiFusion-LDSeg IoU/DSC
Lushan	Very small	8%	22.4/36.6	28.1/43.9	33.5/50.2
	Small	20%	41.3/58.4	48.7/65.5	54.2/70.3
	Medium	51%	60.4/75.3	65.2/78.9	70.1/82.5
	Large	21%	71.8/83.6	75.4/86.0	79.8/88.7
Jiuzhaigou	Very small	11%	23.8/38.2	29.4/45.3	35.1/52.0
	Small	24%	43.1/60.2	50.3/67.0	56.4/72.1
	Medium	48%	62.5/76.9	66.8/80.1	72.6/84.0
	Large	17%	73.2/84.5	76.9/86.8	81.5/89.6
Luding	Very small	7%	21.9/35.8	27.6/43.1	32.8/49.5
	Small	15%	40.5/57.6	47.9/64.8	53.1/69.4
	Medium	53%	61.3/75.8	65.7/79.3	70.4/82.2
	Large	25%	72.6/83.9	76.1/86.3	80.2/88.9

Note: Very small: <50 m^2; Small: 50–500 m^2; Medium: 500–10,000 m^2; Large: >10,000 m^2.

Figure 16 presents a cross-dataset generalization analysis where models trained on one earthquake dataset are evaluated on the other two datasets without fine-tuning, revealing BiFusion-LDSeg's superior transferability across diverse geological and seismic settings. The proposed framework achieves the most balanced performance matrix with average cross-dataset DSC of 78.2% compared to U-Net (61.3%), Swin-UNet (71.5%), and MedSegDiff (72.2%), demonstrating reduced performance gaps between in-domain and out-of-domain evaluation. This robust cross-dataset performance validates that the proposed bi-directional CNN-Transformer fusion effectively learns multi-scale terrain features and landscape-level contextual patterns that transfer across different mountainous regions with varying lithological and topographic characteristics.

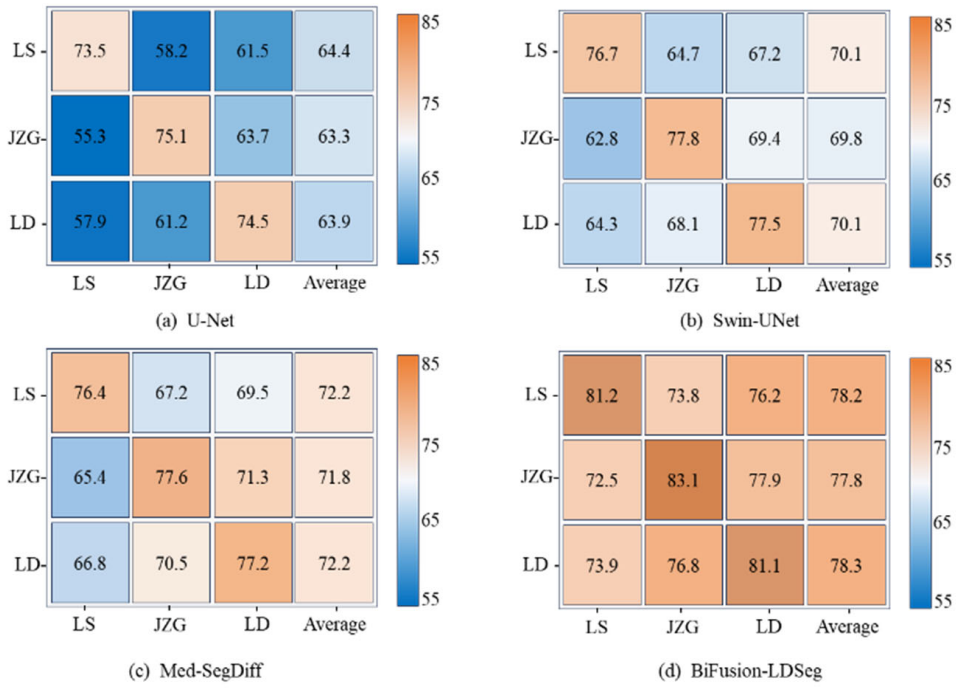

Figure 16. Cross-dataset generalization capability.

To assess cross-regional generalization beyond Sichuan Province, this research conducted additional validation experiments on the Ludian earthquake dataset from Yunnan Province, China (Mw 6.5, 2014). The Ludian region is geologically distinct from the three Sichuan datasets, characterized by deeply incised valley terrain along the Zhaotong fault system, different lithological compositions, and higher landslide density relative to total affected area. As shown in Table 9, BiFusion-LDSeg achieves IoU of 65.1 ± 1.2% and DSC of 78.6 ± 0.9% on this external dataset, outperforming all baseline methods by a consistent margin while maintaining superior boundary accuracy (ASSD: 3.6 ± 0.4 pixels). The absolute performance scores are moderately lower than those obtained on the Sichuan datasets, which is expected given the domain shift in terrain characteristics and the absence of any fine-tuning on the Ludian data. Nevertheless, the consistent performance advantage of BiFusion-LDSeg over competing methods for this geologically distinct region provides meaningful evidence that the bi-directional CNN-Transformer fusion and latent diffusion mechanism learn generalizable landslide-discriminative representations rather than region-specific patterns.

Table 9. Quantitative segmentation performance on the external validation dataset from the Ludian earthquake region, Yunnan Province, China (Mw 6.5, 2014).

Method	IoU (%)	DSC (%)	ASSD (Pixels)	HD (Pixels)
U-Net	55.3 ± 1.9	71.2 ± 1.6	6.1 ± 0.8	41.3 ± 4.1
U-Net++	56.8 ± 1.8	72.4 ± 1.5	5.8 ± 0.7	39.4 ± 3.8
DeepLabv3+	57.2 ± 1.7	72.8 ± 1.4	5.6 ± 0.7	38.2 ± 3.7
Attention U-Net	58.4 ± 1.6	73.7 ± 1.3	5.3 ± 0.6	36.8 ± 3.4
DANet	58.9 ± 1.5	74.1 ± 1.2	5.1 ± 0.6	35.7 ± 3.2
Swin-UNet	59.5 ± 1.4	74.6 ± 1.1	4.9 ± 0.5	34.6 ± 3.0
SegFormer	60.1 ± 1.3	75.1 ± 1.0	4.7 ± 0.5	33.8 ± 2.8
TransUNet	60.7 ± 1.3	75.5 ± 1.0	4.5 ± 0.5	33.1 ± 2.7
MedSegDiff	59.2 ± 1.5	74.3 ± 1.2	5.0 ± 0.6	34.9 ± 3.1
LSegDiff	61.8 ± 1.2	76.4 ± 1.0	4.4 ± 0.5	32.5 ± 2.6
BiFusion-LDSeg	65.1 ± 1.2	78.6 ± 0.9	3.6 ± 0.4	24.3 ± 2.3

Statistically significant improvement over baselines (paired t-test, $p < 0.05$).

Figure 17 visualizes the calibrated feature outputs produced by the Bi-AG module at Stage 4 for representative complex landslide cases. The calibrated CNN outputs (F_C^i, Row a) exhibit sharpened and spatially precise activations along landslide boundaries and terrain discontinuities, reflecting how Transformer global context selectively reinforces boundary-relevant local features. The calibrated Transformer outputs (F_T^i, Row b) demonstrate broader and more coherent activations across the full landslide extent and surrounding topographic context, reflecting how CNN spatial detail refines the Transformer's global attention toward landslide-discriminative regions. The complementary and non-redundant nature of these two outputs validates the design rationale of the bi-directional attention mechanism.

Figure 18 illustrates the inherent uncertainty quantification capability of BiFusion-LDSeg through pixel-wise confidence heatmaps generated via multiple stochastic forward passes during inference. This addresses a critical limitation of deterministic segmentation models for operational deployment in post-earthquake response scenarios. The uncertainty maps reveal that high-confidence predictions (bright regions) concentrate within landslide cores and clear background areas. Additionally, the elevated uncertainty (purple regions) localizes precisely along landslide boundaries and ambiguous transitional zones where spectral signatures overlap with surrounding terrain features. This spatial distribution of uncertainty provides actionable information for prioritizing expert review, where high-confidence automated predictions (>0.8) covering approximately 85% of pixels can be

directly incorporated into rapid landslide inventory maps, while medium-confidence regions (0.5–0.8) are flagged for manual verification by domain specialists.

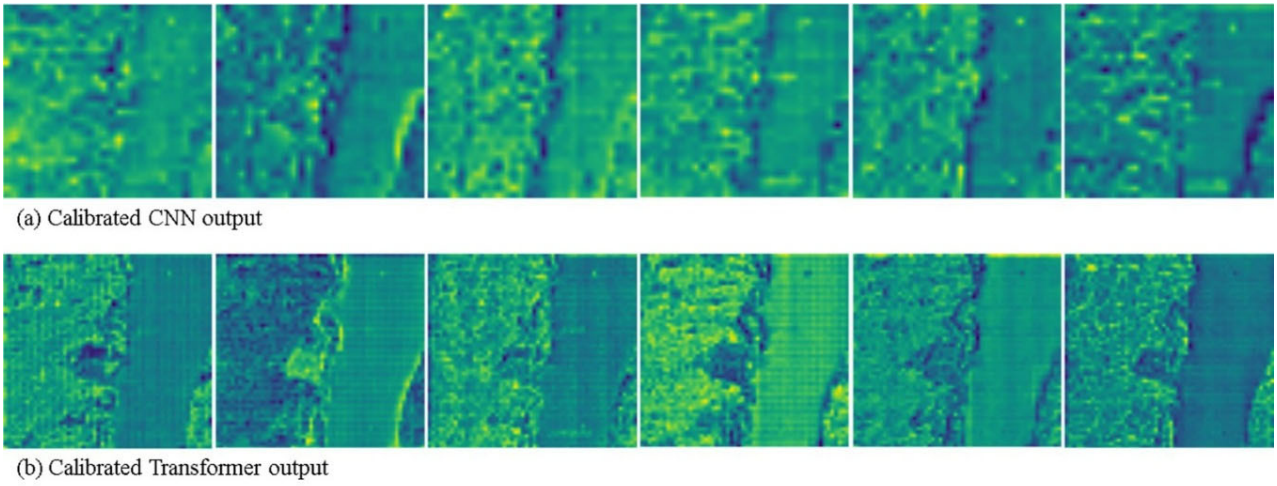

Figure 17. Visualization of calibrated feature outputs from the Bi-AG module for representative complex landslide cases.

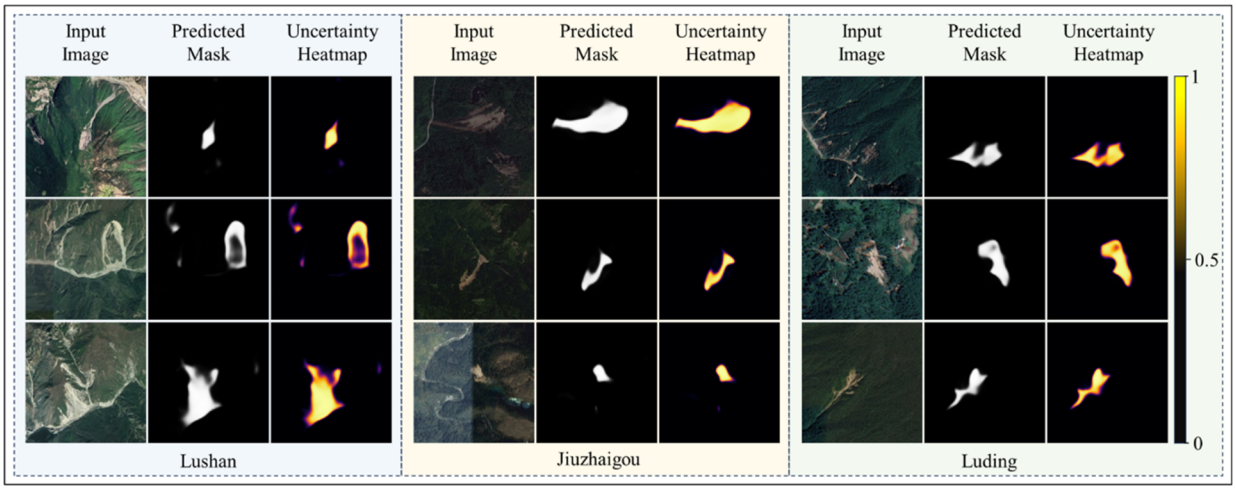

Figure 18. Uncertainty quantification for reliable deployment.

To quantitatively assess uncertainty calibration, this research adopted Expected Calibration Error (ECE) [53], which measures the alignment between predicted confidence scores and actual segmentation accuracy. Theoretically, a lower ECE indicates uncertainty estimates that are better calibrated. Table 10 presents quantitative uncertainty evaluation across all three datasets. BiFusion-LDSeg achieves the lowest Expected Calibration Error (ECE: 5.2 ± 0.8%), indicating well-calibrated confidence estimates compared to all baselines. The confidence-stratified IoU analysis further confirms practical utility—high-confidence predictions (>0.8) achieve 72.4 ± 1.1% IoU while medium-confidence regions (0.5–0.8) yield 58.6 ± 1.5% IoU—validating the framework's operational deployment strategy, where high-confidence predictions covering approximately 85% of pixels are directly incorporated into landslide inventory maps and medium-confidence regions are flagged for expert review.

Table 10. Quantitative uncertainty evaluation across three datasets.

Model Variant	ECE (%)	IoU-High (%)	IoU-Med (%)	High-Conf Coverage (%)
U-Net	12.8 ± 1.4	61.3 ± 1.8	48.2 ± 2.1	71.3 ± 2.4
TransUNet	10.4 ± 1.2	65.8 ± 1.5	52.6 ± 1.9	74.8 ± 2.1
MedSegDiff	8.6 ± 1.1	67.2 ± 1.4	54.3 ± 1.8	79.2 ± 2.0
LSegDiff	7.9 ± 1.0	68.5 ± 1.3	55.8 ± 1.7	81.4 ± 1.9
BiFusion-LDSeg (Proposed)	5.2 ± 0.8	72.4 ± 1.1	58.6 ± 1.5	85.3 ± 1.7

Note: ECE—Expected Calibration Error (lower is better). IoU-High and IoU-Med denote segmentation accuracy at high (>0.8) and medium (0.5–0.8) confidence thresholds respectively. Statistically significant improvement over all baselines (paired t-test, $p < 0.05$).

Figure 19 visualizes the learned attention mechanisms across different decoder stages, revealing how BiFusion-LDSeg progressively focuses on multi-scale landslide characteristics through hierarchical feature representations. Pattern 1 primarily activates on main landslide bodies and source areas with high spectral contrast, Pattern 2 emphasizes elongated runout zones and debris deposits along downslope directions, Pattern 3 captures curved scarp boundaries and lateral margins, and Pattern 4 focuses on fragmented or small-scale landslide components that are challenging for conventional methods to detect. This decomposition of attention patterns demonstrates that the bi-directional fusion mechanism enables the network to learn complementary representations where CNN branches specialize in local texture discontinuities and sharp boundaries while Transformer branches model spatial relationships between landslide components and surrounding topographic context across the landscape.

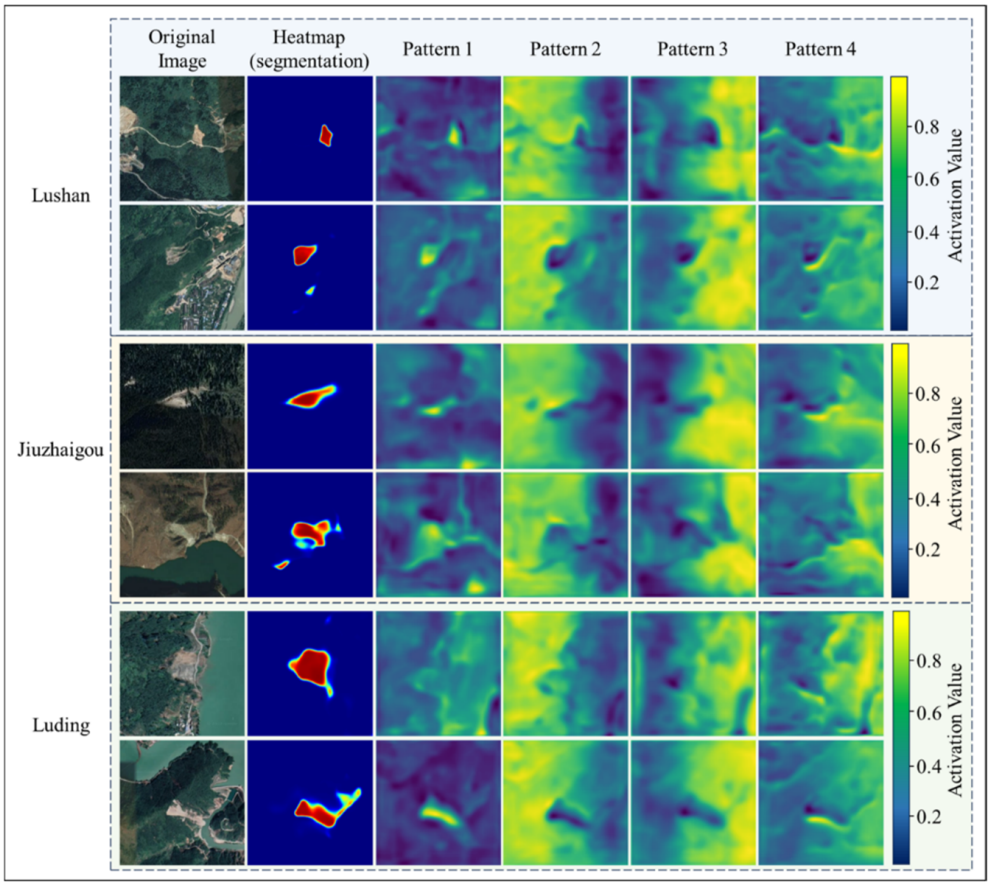

Figure 19. Learned attention patterns and feature discrimination.

6. Conclusions

This paper presents BiFusion-LDSeg, a novel bi-directional fusion enhanced latent diffusion framework for earthquake-triggered landslide segmentation from satellite imagery. The proposed framework synergistically integrates three key innovations: a dual-encoder architecture with Bi-directional Attention Gates (Bi-AG) enabling sophisticated CNN-Transformer feature fusion; a conditional latent diffusion process operating in learned low-dimensional landslide shape manifolds; and a boundary-aware progressive decoder with multi-scale reverse attention mechanisms for precise boundary delineation. Through end-to-end joint optimization, the framework achieves superior segmentation accuracy across three earthquake datasets (Lushan, Jiuzhaigou, Luding) while maintaining computational efficiency through rapid inference with only 10 sampling steps. Extensive experiments demonstrate the framework's robustness to atmospheric noise, strong cross-dataset generalization capability, and inherent uncertainty quantification for reliable operational deployment in post-earthquake emergency response scenarios.

An important limitation of the current framework is its exclusive reliance on optical satellite imagery without incorporating prior geological or geomorphological knowledge. Terrain indices such as slope, curvature, and topographic wetness index, alongside lithological and hydrological maps, carry rich physical information that could substantially enhance model interpretability in geologically complex terrain. Future work will explore integrating these as auxiliary encoder inputs or physics-informed regularization constraints within the latent diffusion process. Furthermore, a follow-up study is currently underway that extends the present framework toward a comprehensive multi-modal landslide reactivation risk assessment system, incorporating three complementary input streams: morphological information from satellite imagery, antecedent and forecasted precipitation records, and post-shock seismic activity data—collectively capturing the primary physical drivers of landslide reactivation at regional scales.

Future research directions include extending the framework to multi-temporal change detection for monitoring landslide evolution and secondary hazards, incorporating multi-modal data fusion combining optical imagery with SAR and LiDAR-derived topographic features, and developing semi-supervised learning strategies to reduce annotation requirements for rapid deployment following new seismic events. Additionally, investigating physics-informed constraints derived from geomorphological and geotechnical principles could further enhance model interpretability and generalization across diverse geological settings, while transfer learning approaches may enable effective adaptation to other mass movement types including debris flows, rockfalls, and snow avalanches with minimal fine-tuning.

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

Funding: This research was funded by the scientific research fund of the Institute of Engineering Mechanics (China earthquake administration) (Grant No. 2020EEEVL0202) and research grants from the National Institute of Natural Hazards, Ministry of Emergency Management of China (Grant No. J2219811), the Key Science and Technology Program Project of the Ministry of Emergency Management (Grant No. 2024EMST040405), the National Natural Science Foundation of China (Grant No. 42407239), the Key Laboratory Incentive Funding of Sichuan Province, China (Grant No. 2025JDPT0068-03), and Coal—Major Project (Grant No. 2025ZD1700904).

Data Availability Statement: Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

Conflicts of Interest: Jingjing Jiao is an employee of BGI Engineering Consultants Ltd., Beijing, 100038, China. No other financial interests (such as consultancies, stock ownership, grants, patents, or honoraria) directly related to the content of this manuscript are declared by this author. All other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

References

1. Hou, C.; Yu, J.; Ge, D.; Yang, L.; Xi, L.; Pang, Y.; Wen, Y. A Transfer Learning Approach for Landslide Semantic Segmentation Based on Visual Foundation Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 11561–11572. [CrossRef]
2. Fan, X.; Wang, X.; Fang, C.; Jansen, J.D.; Dai, L.; Tanyas, H.; Zang, N.; Tang, R.; Xu, Q.; Huang, R. Deep learning can predict global earthquake-triggered landslides. *Natl. Sci. Rev.* **2025**, *12*, nwaf179. [CrossRef] [PubMed]
3. Gupta, K.; Satyam, N. Optimizing seismic hazard inputs for co-seismic landslide susceptibility mapping: A probabilistic analysis. *Nat. Hazards* **2024**, *120*, 8459–8481. [CrossRef]
4. Robinson, T.R.; Rosser, N.J.; Densmore, A.L.; Williams, J.G.; Kincey, M.E.; Benjamin, J.; Bell, H.J.A. Rapid post-earthquake modelling of coseismic landslide intensity and distribution for emergency response decision support. *Nat. Hazards Earth Syst. Sci.* **2017**, *17*, 1521–1540. [CrossRef]
5. Bragagnolo, L.; Rezende, L.R.; da Silva, R.; Grzybowski, J.M.V. Convolutional neural networks applied to semantic segmentation of landslide scars. *CATENA* **2021**, *201*, 105189. [CrossRef]
6. Lu, W.; Hu, Y.; Zhang, Z.; Cao, W. A dual-encoder U-Net for landslide detection using Sentinel-2 and DEM data. *Landslides* **2023**, *20*, 1975–1987. [CrossRef]
7. Devara, M.; Maurya, V.K.; Dwivedi, R. Landslide extraction using a novel empirical method and binary semantic segmentation U-NET framework using sentinel-2 imagery. *Remote Sens. Lett.* **2024**, *15*, 326–338. [CrossRef]
8. Li, H.; He, Y.; Xu, Q.; Deng, J.; Li, W.; Wei, Y. Detection and segmentation of loess landslides via satellite images: A two-phase framework. *Landslides* **2022**, *19*, 673–686. [CrossRef]
9. Ghorbanzadeh, O.; Crivellari, A.; Ghamisi, P.; Shahabi, H.; Blaschke, T. A comprehensive transferability evaluation of U-Net and ResU-Net for landslide detection from Sentinel-2 data (case study areas from Taiwan, China, and Japan). *Sci. Rep.* **2021**, *11*, 14629. [CrossRef]
10. Liu, X.; Peng, Y.; Lu, Z.; Li, W.; Yu, J.; Ge, D.; Xiang, W. Feature-fusion segmentation network for landslide detection using high-resolution remote sensing images and digital elevation model data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
11. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sens.* **2019**, *11*, 196. [CrossRef]
12. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
13. Ma, H.; Han, G.; Peng, L.; Zhu, L.; Shu, J. Rock thin sections identification based on improved squeeze-and-Excitation Networks model. *Comput. Geosci.* **2021**, *152*, 104780. [CrossRef]
14. Pham, T.M.; Do, N.; Pham, H.T.T.; Bui, H.T.; Do, T.T.; Hoang, M.V. CResU-Net: A method for landslide mapping using deep learning. *Mach. Learn. Sci. Technol.* **2024**, *5*, 035008. [CrossRef]
15. Nigelesh, T.; Shruthik, V.S.; Reddy, V.S.; Singh, R.P. Landslide Detection in Satellite Images using InceptionU-Net and Convolutional Block Attention Module. *Procedia Comput. Sci.* **2025**, *258*, 4301–4310. [CrossRef]
16. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
17. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
18. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
19. Liu, B.; Wang, W.; Wu, Y.; Gao, X. Attention Swin Transformer UNet for Landslide Segmentation in Remotely Sensed Images. *Remote Sens.* **2024**, *16*, 4464. [CrossRef]
20. Gao, M.; Chen, F.; Wang, L.; Zhao, H.; Yu, B. Swin Transformer-based multi-scale attention model for landslide extraction from large-scale area. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14.

21. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
22. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
23. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
24. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
25. Sameen, M.I.; Sarkar, R.; Pradhan, B.; Drukpa, D.; Alamri, A.M.; Park, H.-J. Landslide spatial modelling using unsupervised factor optimisation and regularised greedy forests. *Comput. Geosci.* **2020**, *134*, 104336. [CrossRef]
26. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
27. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
28. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
29. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 8780–8794.
30. Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; Xu, Y. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*; PMLR: London, UK, 2024; pp. 1623–1639.
31. Zaman, F.A.; Jacob, M.; Chang, A.; Liu, K.; Sonka, M.; Wu, X. Latent diffusion for medical image segmentation: End to end learning for fast sampling and accuracy. *arXiv* **2024**, arXiv:2407.12952. [CrossRef]
32. Liu, Y.; Yue, J.; Xia, S.; Ghamisi, P.; Xie, W.; Fang, L. Diffusion models meet remote sensing: Principles, methods, and perspectives. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–22. [CrossRef]
33. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
34. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*; PMLR: London, UK, 2021; pp. 8162–8171.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
36. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 171–180.
37. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [CrossRef]
38. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
40. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
41. Bogensperger, L.; Narnhofer, D.; Ilic, F.; Pock, T. Score-based generative models for medical image segmentation using signed distance functions. *arXiv* **2023**, arXiv:2303.05966. [CrossRef]
42. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
45. Upadhyay, A.K.; Bhandari, A.K. MaS-TransUNet: A multi-attention swin transformer U-net for medical image segmentation. *IEEE Trans. Radiat. Plasma Med. Sci.* **2024**, *9*, 613–626. [CrossRef]
46. Yuan, F.; Peng, Y.; Huang, Q.; Li, X. A bi-directionally fused boundary aware network for skin lesion segmentation. *IEEE Trans. Image Process.* **2024**, *33*, 6340–6353. [CrossRef]

47. Li, H.; He, Y.; Xu, Q.; Deng, J.; Li, W.; Wei, Y.; Zhou, J. Sematic segmentation of loess landslides with STAPLE mask and fully connected conditional random field. *Landslides* **2023**, *20*, 367–380. [CrossRef]
48. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11.
49. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
50. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3146–3154.
51. 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.
52. Vu Quoc, H.; Tran Le Phuong, T.; Trinh Xuan, M.; Dinh Viet, S. Lsegdiff: A latent diffusion model for medical image segmentation. In Proceedings of the 12th International Symposium on Information and Communication Technology, Ho Chi Minh, Vietnam, 7–8 December 2023; pp. 456–462.
53. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017*; PMLR: London, UK, 2017; Volume 70, pp. 1321–1330.