

Cross-Level Gated Dual-Attention Network for Landslide Segmentation in Remote Sensing Imagery

Jianping Liao

liaojianping@qzct.edu.cn

Quzhou College of Technology

Lihua Ye

Jiaying University

Article

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11212191/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Cross-Level Gated Dual-Attention Network for Landslide Segmentation in Remote Sensing Imagery

Jianping Liao^{1,*} and Lihua Ye²

¹Department of Information Engineering, Quzhou College of Technology, Quzhou, Zhejiang, 324000, China

²School of Computer Science, Jiaying University, Jiaxing, Zhejiang, 314001, China

*Correspondence: liaojianping@qzct.edu.cn

ABSTRACT

Accurate landslide segmentation from remote sensing imagery is essential for rapid disaster assessment and geohazard monitoring. However, complex background interference, large variations in landslide scale, irregular boundaries, and the semantic gap between shallow and deep features remain major challenges for precise segmentation. To address these issues, this study proposes CLGDA-Net, a cross-level gated dual-attention network for landslide segmentation in remote sensing imagery. Firstly, a cross-level gated context fusion module (CLGCFM) is introduced into the skip connections to adaptively integrate adjacent shallow and deep features. Specifically, the deep features are spatially aligned with shallow features and dynamically fused through an adaptive gated feature fusion mechanism, followed by receptive-field enhancement to strengthen multi-scale contextual representation. Secondly, a multi-dilation context attention module (MDCAM) is designed at the bottleneck to capture contextual information under different receptive fields using parallel dilated convolutions and to selectively emphasize informative responses. Furthermore, a multi-scale dual-attention fusion module (MSDAFM) is developed to aggregate multi-level decoder features through dual-pooling scale-aware channel attention and scale-adaptive spatial attention, thereby enhancing complementary information across both scale and spatial dimensions. Extensive experiments on Landslide4Sense and Bijie demonstrate that CLGDA-Net achieves competitive segmentation performance and effectively improves the delineation of landslide regions under complex remote sensing environments. Specifically, CLGDA-Net achieves Dice/MCC/Jaccard scores of 0.7315/0.7187/0.5774 on Landslide4Sense and 0.8316/0.8126/0.7118 on Bijie. Moreover, ablation studies verify the effectiveness of the proposed modules, highlighting the positive contributions of cross-level gated fusion and dual-attention feature enhancement to landslide segmentation. These findings suggest that CLGDA-Net has strong potential for practical applications in rapid landslide mapping, post-disaster assessment, and geohazard monitoring based on remote sensing imagery.

Introduction

Landslide events pose a serious threat to human safety, ecological systems, and regional economic development. Their formation and evolution are generally associated with a combination of triggering and conditioning factors, such as extreme rainfall, seismic activity, geological structure, slope instability, and anthropogenic disturbance. The rapid development of high-resolution Earth observation technologies has made it increasingly practical to identify and monitor landslide-affected areas over large regions. Unlike conventional ground-based surveys, remote sensing provides repeated observations over extensive and inaccessible mountainous terrain with relatively high spatial and temporal efficiency. Therefore, reliable pixel-level delineation of landslide regions from remote sensing images is of considerable importance for emergency mapping, post-event damage assessment, disaster response, and long-term geological hazard surveillance.

Early landslide mapping approaches generally depended on expert visual interpretation or manually designed descriptors derived from spectral, textural, and topographic information. Although such methods can be effective in relatively uncomplicated scenes, their robustness often decreases when applied to remote sensing images with strong appearance variability. Landslide regions may differ greatly in size, shape, and spatial distribution, while their boundaries are often indistinct or discontinuous. At the same time, several non-landslide objects, including exposed ground, roads, rock surfaces, riverbanks, and construction zones, can exhibit similar visual patterns, which increases the likelihood of missed detections and false alarms. These characteristics make reliable pixel-wise extraction of landslide areas particularly challenging for conventional approaches.

With the rapid development of deep learning techniques, an increasing number of semantic segmentation networks have been specifically explored for landslide mapping using remote sensing imagery. U-Net¹ and its improved variants²⁻⁵ have become widely used frameworks in this field because of their encoder-decoder structures, which enable the simultaneous learning of high-level semantic representations and detailed spatial information through skip connections. For example, He et al.⁶ proposed MBENet for cross-domain coseismic landslide segmentation, where boundary-sensitive feature refinement

is combined with pixel-level contrastive learning to improve both local structural preservation and semantic discrimination. Emani et al.⁷ developed SUGARFuseNet, which leverages complementary multi-temporal and multi-modal observations to improve the generalization and boundary delineation of landslide segmentation. Chai et al.⁸ introduced SWMFA-UNet to enhance landslide segmentation by jointly modeling spatial and frequency-domain information, which improves the extraction of structural details and contextual dependencies from remote sensing imagery. You et al.⁹ developed AS-UNet, which improves the representation of anisotropic landslide structures and suppresses irrelevant encoder information during feature reconstruction, while retaining a lightweight U-Net architecture. Liu et al.¹⁰ developed VFM-MoME for robust landslide extraction in difficult imaging environments. The framework exploits transferable visual priors and long-range sequence modeling to strengthen semantic discrimination while retaining detailed structural information in landslide regions. Tang et al.¹¹ developed GeoSlide-XMamba by jointly exploiting spectral and topographic cues, enabling the network to better capture terrain-related spatial dependencies and preserve landslide boundary structures. Zhong et al.¹² developed DFG-Net to exploit complementary frequency characteristics of landslide regions, enabling more effective representation of large-area semantics and detailed boundary information. Although these methods have achieved notable progress, several challenges remain. Firstly, direct skip connections may pass background noise together with useful spatial details, while the semantic mismatch between shallow and deep features is not fully addressed. Secondly, the large variation in landslide size and shape makes fixed receptive fields inadequate for capturing both small fragmented regions and large-scale structures. Thirdly, multi-level decoder features are often fused in a simple manner, which limits effective exploitation of their complementary scale and spatial information.

To better exploit hierarchical representations, the network should not only preserve fine spatial details but also make effective use of high-level semantic cues. Simple skip fusion may introduce irrelevant responses because features from different encoder stages differ markedly in representation characteristics. Meanwhile, the substantial variation in landslide size and shape requires broader and more flexible contextual modeling than that provided by fixed receptive fields. Multi-dilation processing can alleviate this limitation by capturing information at different spatial ranges. In the decoding stage, features at different resolutions also provide complementary information, but conventional addition or concatenation cannot fully distinguish their relative importance. Therefore, adaptive cross-level interaction, multi-receptive-field context extraction, and scale- and spatial-aware decoder fusion are jointly considered to improve feature discrimination and boundary delineation in complex landslide scenes. Based on these considerations, this study proposes CLGDA-Net, a cross-level gated dual-attention network for landslide segmentation in remote sensing imagery. The network adopts an encoder-decoder architecture and incorporates three key components to improve hierarchical feature interaction and multi-scale representation. Specifically, CLGCFM is introduced into the skip connections to refine cross-level feature transfer, MDCAM is placed at the bottleneck to strengthen contextual modeling under different receptive fields, and MSDAFM is employed to aggregate complementary decoder features through channel- and spatial-aware fusion. Together, these components enhance feature discrimination and boundary delineation in complex landslide scenes. The main contributions of this study can be summarized as follows:

- (1) A cross-level gated context fusion module is proposed to improve information interaction between adjacent encoder stages. It adaptively combines shallow spatial details with deeper semantic cues and further refines the fused representation through multi-scale receptive-field enhancement.
- (2) A multi-dilation context attention module is designed at the bottleneck to capture contextual information under different receptive fields. Parallel dilated convolutions enrich multi-scale representations, while attention-based recalibration emphasizes informative features.
- (3) A multi-scale dual-attention fusion module is developed to exploit complementary decoder information. It combines DPSCAM and SASAM to adaptively model scale importance and spatial responses, thereby improving multi-level feature fusion.

Methods

Network architecture of CLGDA-Net

As illustrated in Fig. 1, CLGDA-Net adopts a hierarchical encoder-decoder architecture comprising a four-stage encoder, CLGCFMs embedded in the skip pathways, an MDCAM at the bottleneck, and an MSDAFM for multi-level decoder feature aggregation. Specifically, the encoder progressively extracts hierarchical representations through four stages with channel numbers of 32, 64, 128, and 256. Each stage employs a Double Convolution block (DoubleConv), which consists of two consecutive 3×3 convolutional layers. A 2×2 max-pooling operation is subsequently applied for spatial down-sampling. The deepest encoder feature is then fed into MDCAM, where contextual information under different receptive fields is further enhanced before entering the decoding pathway. To improve feature transfer across different encoding levels, CLGDA-Net replaces direct skip connections with the proposed CLGCFM. At each skip pathway, the deeper feature is first resized and channel-aligned to match the corresponding shallow representation, after which the two features are adaptively fused and further refined before being passed to the decoder. This design enables the skip pathway to preserve detailed spatial cues while incorporating stronger semantic information, thereby reducing the representation gap between adjacent encoder stages.

The decoder progressively restores the spatial resolution through four decoding stages. Each decoder stage first upsamples the feature from the preceding deeper stage using a transposed convolution and then concatenates it with the corresponding CLGCFM-enhanced skip feature. The concatenated feature is subsequently processed by a DoubleConv block for feature reconstruction. In addition to the main decoding pathway, CLGDA-Net introduces a progressive cross-scale aggregation path. Specifically, multi-level encoder features are first transformed by 1×1 convolutions to match the channel dimensions of the corresponding decoder features and are then combined with them through element-wise addition. The fused features are progressively propagated toward higher spatial resolutions through successive up-sampling and addition operations. In this manner, information from different encoding levels can continuously complement the decoding process and contribute to the final prediction. To further exploit complementary information from different decoder levels, three intermediate decoder features are additionally fed into MSDAFM. These features are aligned to a common resolution and jointly refined through scale-aware channel attention and scale-adaptive spatial attention. The final segmentation result is obtained by combining the output of the final decoder stage, the progressive cross-scale aggregation branch, and the MSDAFM output, followed by a sigmoid activation to generate the pixel-wise landslide prediction map. Overall, the proposed architecture facilitates effective interaction among hierarchical, contextual, and multi-scale representations, thereby improving the delineation of landslide regions in complex remote sensing imagery.

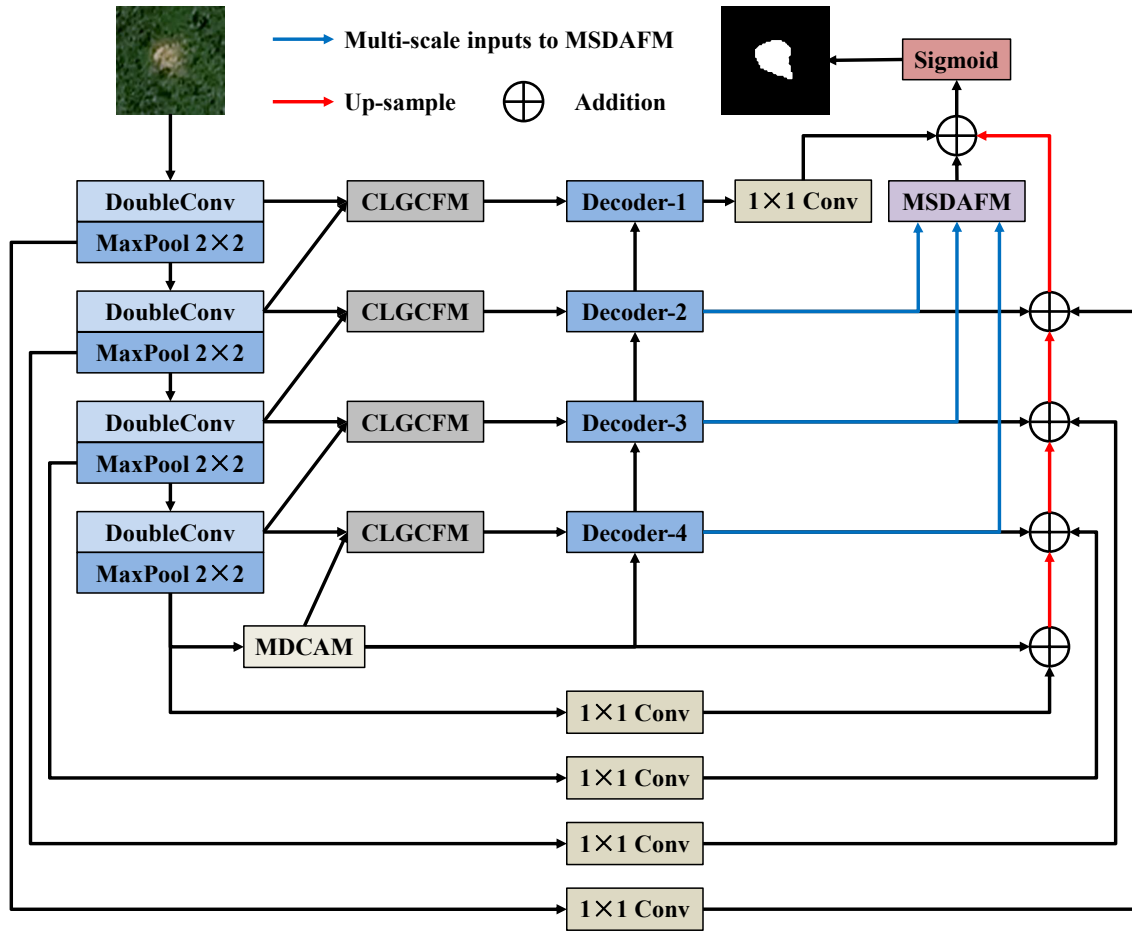


Figure 1. Network architecture of CLGDA-Net.

Multi-dilation context attention module

Landslide regions in remote sensing imagery often exhibit substantial variations in size, shape, and spatial distribution, making features extracted with a single receptive field insufficient to simultaneously characterize local details and large-scale structures. To address this issue, MDCAM combines multi-dilation feature extraction with attention-based feature recalibration. As shown in Fig. 2, the proposed multi-dilation context attention module is placed at the bottleneck of CLGDA-Net to strengthen contextual representation across multiple receptive fields. Given an input feature, a 1×1 convolution is first applied to adjust the channel dimension. The transformed feature is then divided into two paths. One path is preserved as the identity feature for

subsequent feature recalibration, while the other is processed by a 3×3 convolution for spatial feature extraction. The resulting feature is further fed into three parallel branches with dilation rates of 1, 2, and 3. Each branch contains two consecutive 3×3 dilated convolutions (DConv) with the same dilation rate, enabling the module to capture complementary contextual information at different spatial ranges. The outputs of the three dilated-convolution branches are first combined through element-wise addition and then added to the feature generated by the preceding 3×3 convolution. The fused representation is subsequently passed through a 1×1 convolution and a sigmoid activation to generate an attention map. Finally, the attention map is multiplied element-wise with the preserved identity feature to produce the output of MDCAM. Through this mechanism, MDCAM enhances informative contextual responses while suppressing less relevant features, thereby providing a more discriminative bottleneck representation for subsequent decoding.

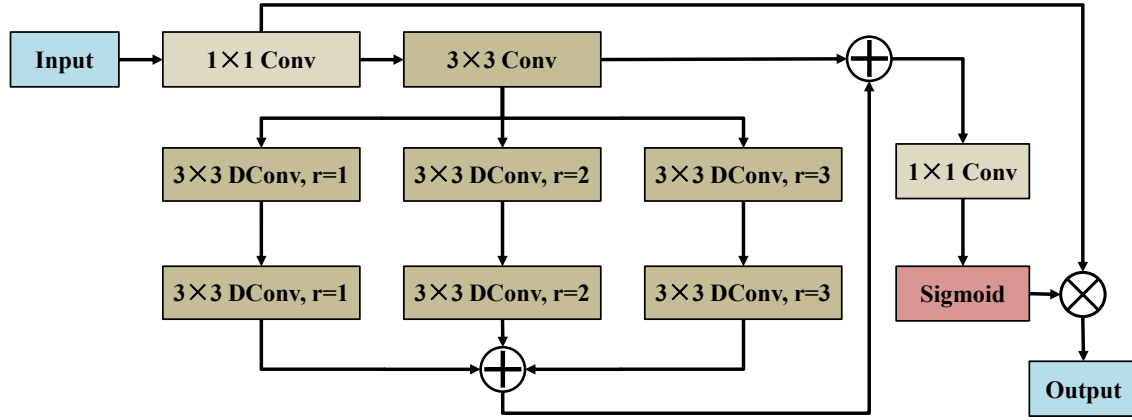


Figure 2. Structure of the multi-dilation context attention module.

Cross-level gated context fusion module

Features generated at different encoding depths contain complementary information. Shallow features retain richer spatial details and boundary cues, whereas deeper features provide stronger semantic representations. Directly transferring shallow features through conventional skip connections may therefore introduce redundant background responses and does not sufficiently exploit the semantic information available at deeper levels. To improve information exchange between adjacent encoder stages, we design the cross-level gated context fusion module, as illustrated in Fig. 3. CLGCFM receives a shallow feature and its adjacent deeper feature as inputs. Since the two features differ in spatial resolution and channel dimension, the deeper feature is first up-sampled by a factor of two and then processed by a 1×1 convolution to match the spatial and channel dimensions of the shallow feature. The aligned deep feature and the original shallow feature are subsequently fed into the adaptive gated feature fusion module (AGFFM). Instead of directly adding or concatenating the two inputs, AGFFM dynamically adjusts their contributions according to the feature content, allowing useful semantic information from the deeper layer to complement the detailed representation retained in the shallow layer. After adaptive fusion, the resulting feature is further processed by the receptive field block (RFB). The RFB extracts contextual information at different spatial ranges through multiple convolutional branches, thereby refining the fused representation and improving its sensitivity to landslide structures with varying shapes and sizes. Finally, the enhanced feature is delivered to the corresponding decoder stage. In this way, CLGCFM combines cross-level semantic interaction with multi-scale contextual refinement, providing more informative skip features for subsequent feature reconstruction.

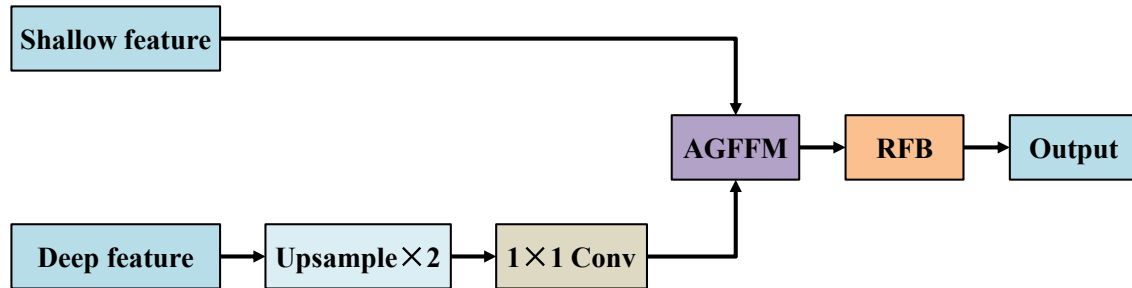


Figure 3. Structure of the cross-level gated context fusion module.

The adaptive gated feature fusion module is designed to dynamically integrate two feature representations with different semantic characteristics. In the cross-level fusion process, shallow features generally contain detailed spatial information but may also include background noise, whereas deeper features provide stronger semantic discrimination but lack fine-grained structural details. Therefore, assigning fixed weights to different feature sources may limit the effectiveness of feature interaction. To address this issue, AGFFM introduces an adaptive gating mechanism to learn the contribution of each input feature. As illustrated in Fig. 4, AGFFM receives two input features with the same spatial resolution and channel dimension. The two features are first concatenated along the channel dimension to construct a joint representation. The concatenated feature is then processed by two independent 1×1 convolutions to generate two attention responses. These responses are concatenated and normalized through a softmax operation, producing two complementary spatial weighting maps. The obtained weights are subsequently applied to the corresponding input features through element-wise multiplication, enabling the module to adaptively emphasize informative regions from each feature source. Finally, the weighted features are combined through element-wise addition to obtain the fused representation. Compared with direct addition or concatenation, AGFFM can dynamically adjust the contribution of different-level features according to their local characteristics, thereby reducing irrelevant background interference and enhancing the complementary interaction between shallow spatial details and deep semantic information.

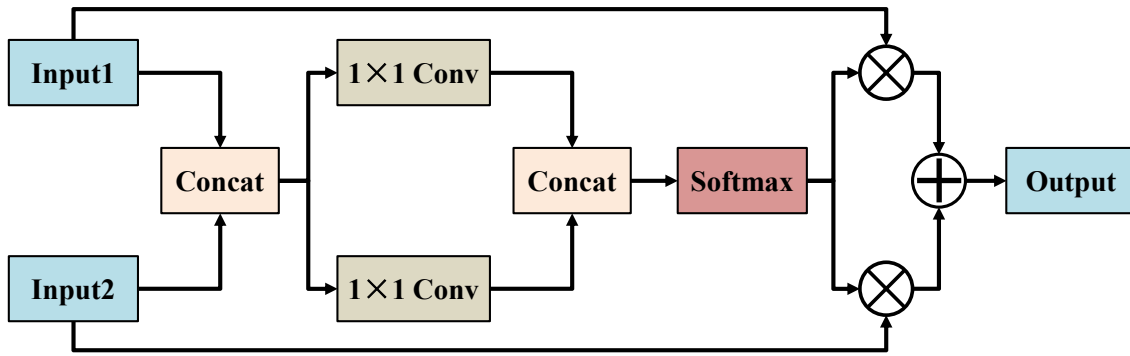


Figure 4. Structure of the adaptive gated feature fusion module.

To further enhance the multi-scale contextual representation of the fused features, the receptive field block is introduced into CLGCFM, as illustrated in Fig. 5. RFB was originally proposed by Liu et al.¹³ and is adopted in this study to refine the features generated by AGFFM. Specifically, the input feature is processed through four parallel branches with different receptive-field configurations. The first branch employs a 1×1 convolution, while the remaining three branches use asymmetric convolutions with kernel sizes of $1 \times 3/3 \times 1$, $1 \times 5/5 \times 1$, and $1 \times 7/7 \times 1$, followed by 3×3 dilated convolutions with dilation rates of 1, 2, and 3. The outputs of the four branches are concatenated and further integrated by a 3×3 convolution. Meanwhile, the input feature is transformed through a 1×1 convolution to form a residual connection, which is added to the multi-branch feature to obtain the final output. By introducing RFB into CLGCFM, contextual information at different spatial scales can be captured more effectively, thereby strengthening the representation of landslide regions with diverse sizes and irregular structures.

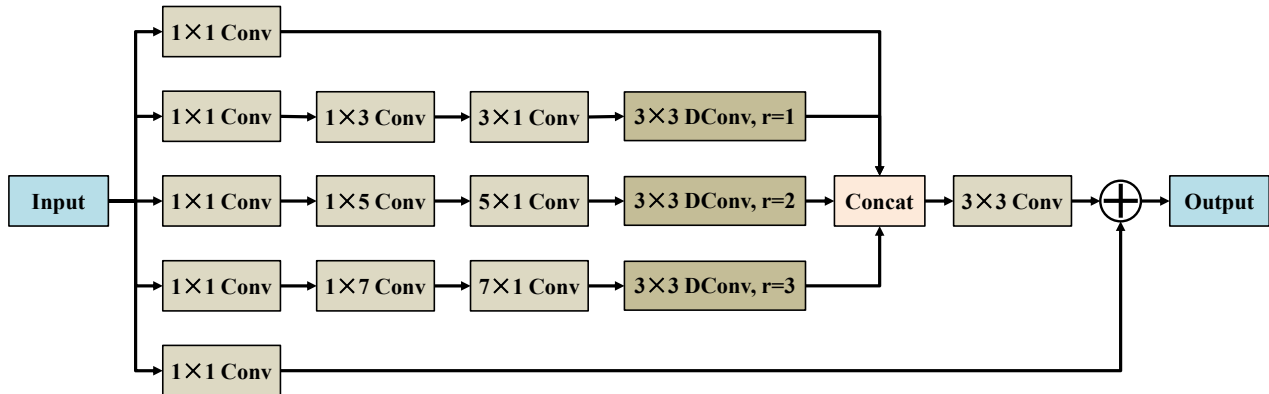


Figure 5. Structure of the receptive field block.

Multi-scale dual-attention fusion module

Decoder features generated at different stages contain complementary information: deeper features provide stronger semantic representations, whereas shallower features retain richer spatial and structural details. However, directly adding or concatenating these features treats different scales uniformly and may fail to fully exploit their complementary characteristics. To achieve more effective multi-level decoder feature interaction, we design the multi-scale dual-attention fusion module, as illustrated in Fig. 6. MSDAFM takes the features from Decoder-2, Decoder-3, and Decoder-4 as inputs. First, each decoder feature is processed by a 1×1 convolution to adjust its channel dimension. The three features are then up-sampled by factors of 2, 4, and 8 to obtain a common spatial resolution. After spatial and channel alignment, the resulting features are concatenated along the channel dimension to construct a unified multi-scale representation. This operation allows semantic information and detailed structural cues from different decoding levels to be jointly preserved. The concatenated feature is subsequently refined by two successive attention mechanisms. The dual-pooling scale-aware channel attention module (DPSCAM) first evaluates the relative importance of information from different feature scales and adaptively reweights their responses. The resulting feature is then fed into the scale-adaptive spatial attention module (SASAM), which further emphasizes informative spatial regions according to the responses of different scales. To preserve the original multi-scale information and reduce excessive feature transformation, a residual connection is introduced between the concatenated feature and the output of the dual-attention pathway. The enhanced feature is finally processed by a 3×3 convolution for feature refinement and a 1×1 convolution to generate the output. Through the combination of multi-scale alignment, scale-aware channel weighting, and spatial attention refinement, MSDAFM enables complementary decoder information to be selectively integrated rather than uniformly fused. Consequently, the module enhances both semantic consistency and spatial detail representation, providing a more discriminative prediction feature for landslide regions with complex shapes, varying scales, and indistinct boundaries.

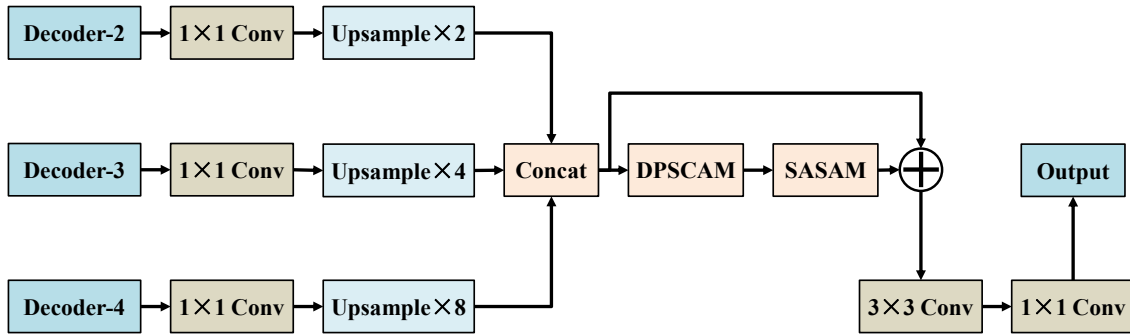


Figure 6. Structure of the multi-scale dual-attention fusion module.

Inspired by recent studies on channel attention modules^{14,15}, we design a dual-pooling scale-aware channel attention module to adaptively evaluate the relative importance of multi-scale decoder features. As shown in Fig. 7, the input feature is first fed into two parallel pooling branches. The upper branch employs Lp pooling to summarize the overall magnitude distribution of feature responses, whereas the lower branch adopts Soft Pooling to preserve informative high-response regions through weighted feature aggregation. The resulting descriptors are subsequently processed by shared MLP transformations, each consisting of a Flatten operation, two Linear layers, and an intermediate ReLU activation. The outputs of the two branches are then combined through element-wise addition, thereby integrating complementary global information provided by the two pooling strategies. The fused descriptor is further processed by the channel attention computation shown in the right part of Fig. 7. Specifically, it is first reshaped according to the number of feature scales and the number of channels assigned to each scale. A mean operation is then performed over the channels within each scale to obtain a compact scale-level descriptor. Subsequently, softmax normalization is applied across different scales to generate normalized scale importance weights. These weights are successively expanded and reshaped to recover the corresponding channel arrangement and are further extended to the spatial dimensions of the original input feature. Finally, the generated scale-aware weights are multiplied element-wise with the original input feature to obtain the output of DPSCAM. In this way, DPSCAM enables the network to selectively emphasize more informative decoder-scale features while suppressing less relevant responses, thereby improving multi-scale feature representation.

Inspired by conventional spatial attention mechanisms^{16,17}, we design a scale-adaptive spatial attention module to characterize the spatial importance of features from different decoder scales. As shown in Fig. 8, the input feature is first transformed by a 3×3 convolution to extract local spatial information, followed by a 1×1 convolution to generate scale-specific spatial responses. Softmax normalization is then performed across the scale dimension, allowing different scales to compete for importance at each spatial position. The normalized responses are subsequently expanded along the channel dimension and

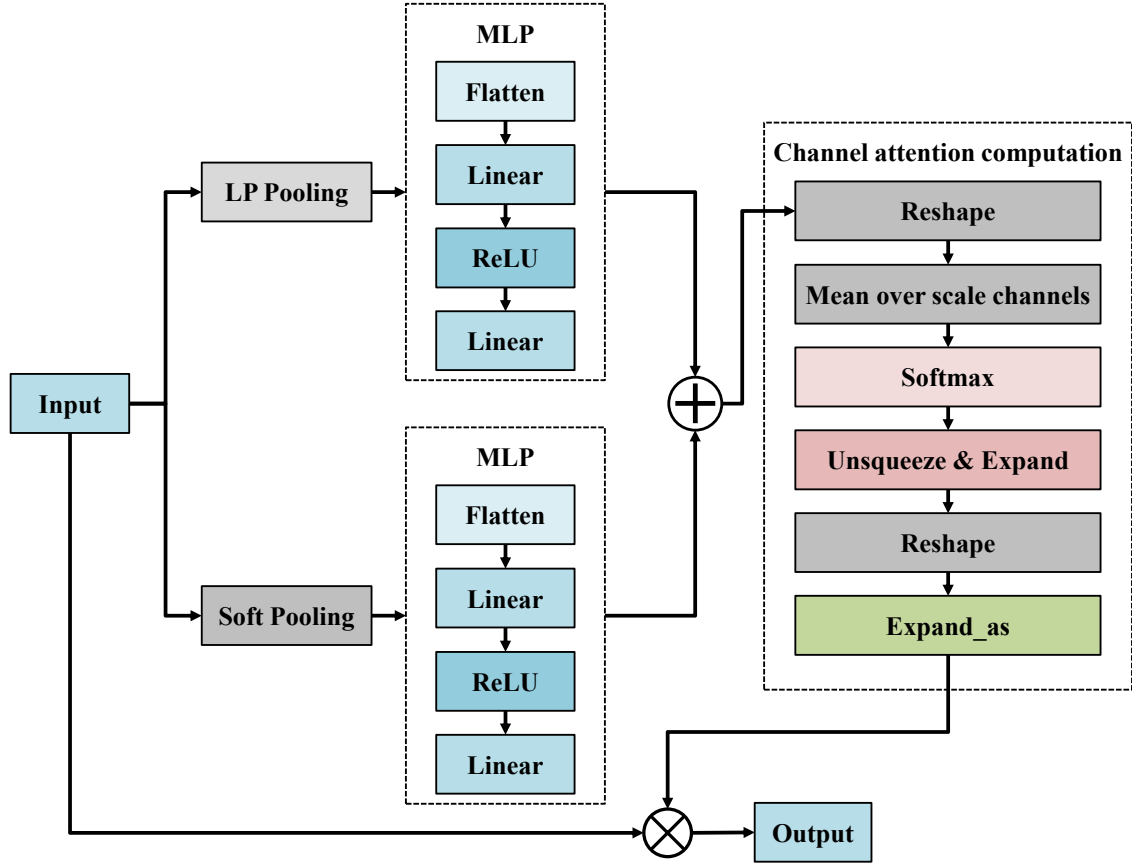


Figure 7. Structure of the dual-pooling scale-aware channel attention module.

reshaped to conform to the organization of the original multi-scale feature. Finally, the generated spatial weights are applied to the input feature through element-wise multiplication. In this manner, SASAM assigns scale-dependent responses to different spatial locations, enabling the network to emphasize informative landslide regions while reducing interference from irrelevant background areas.

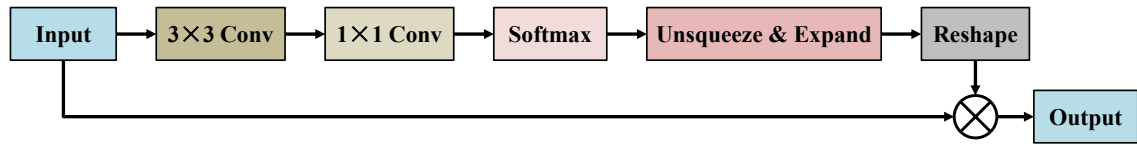


Figure 8. Structure of the scale-adaptive spatial attention module.

Loss function

In remote sensing images, landslide pixels often account for a much smaller proportion than background pixels, which may cause the optimization process to be dominated by the majority class. To alleviate this effect, CLGDA-Net is trained using Dice loss^{18,19}, which evaluates the consistency between the predicted segmentation and the reference mask from the perspective of region overlap. For a prediction map containing N pixels, the Dice loss can be written as:

$$L_{dice}(y, p) = 1 - \frac{2 \sum_{i=1}^N p_i y_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N p_i} \quad (1)$$

where p_i and y_i are the ground-truth value and the corresponding predicted probability at pixel i .

Experiments

Dataset

To evaluate the segmentation performance and generalization ability of CLGDA-Net, experiments were conducted on two publicly available landslide datasets, namely Landslide4Sense²⁰ and Bijie²¹. These two datasets differ in geographical environment, landslide appearance, and image characteristics, providing complementary conditions for assessing the robustness of the proposed method. Representative images and their corresponding annotations are shown in Fig. 9.

Landslide4Sense dataset. Landslide4Sense is a remote sensing benchmark developed for landslide mapping and contains imagery collected from areas with diverse terrain conditions and landslide characteristics. The dataset provides pixel-level annotations of landslide regions, making it suitable for evaluating semantic segmentation models under complex geographical environments. In this study, 1980 images were used for the experiments, including 1385 images for training, 396 for validation, and 199 for testing.

Bijie dataset. The Bijie dataset consists of remote sensing images acquired from the Bijie region of Guizhou Province, China. It contains multiple types of landslide-related phenomena, including rock falls, rock slides, and debris flows, and therefore exhibits considerable variation in landslide morphology and surrounding background. A total of 770 images were used in our experiments, of which 464 were assigned to the training set, 153 to the validation set, and 153 to the test set.

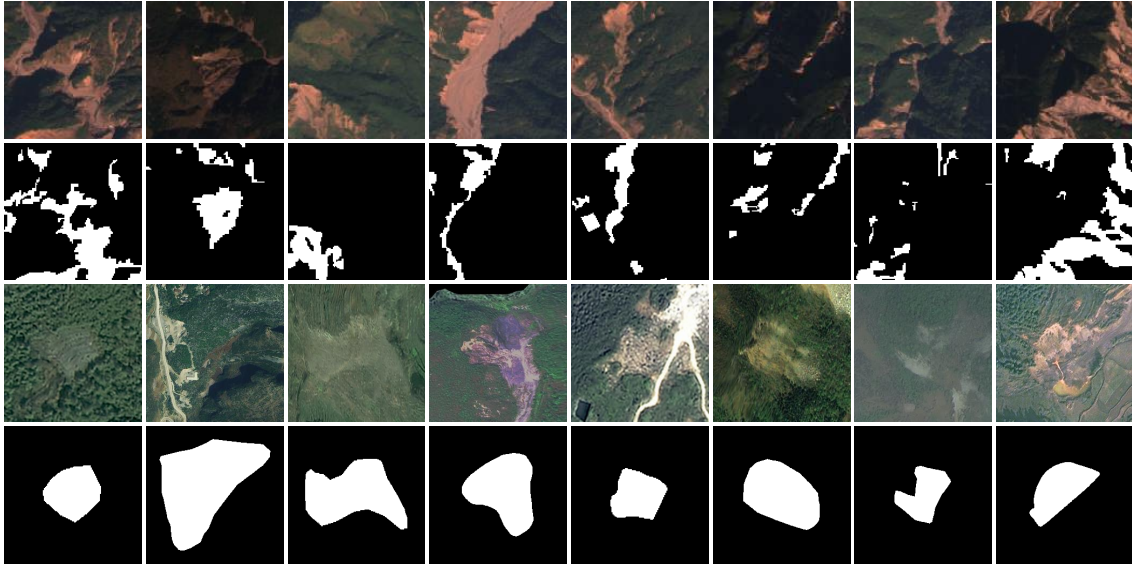


Figure 9. Representative samples and corresponding ground-truth annotations from the Landslide4Sense and Bijie datasets. The first two rows correspond to Landslide4Sense, whereas the last two rows correspond to Bijie.

Evaluation metrics

To quantitatively assess the segmentation performance of different methods, three commonly used evaluation metrics are adopted in this study, including the Dice coefficient (Dice)^{22,23}, Matthews correlation coefficient (MCC)^{24,25}, and Jaccard index (Jaccard)^{26,27}. Their definitions are given as follows:

$$Dice = \frac{2TP}{2TP + FN + FP} \quad (2)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (3)$$

$$Jaccard = \frac{TP}{TP + FN + FP} \quad (4)$$

Experimental settings

All experiments were carried out using the PyTorch 2.0 framework on a workstation equipped with an NVIDIA RTX A6000 GPU. To maintain a unified input format and reduce differences caused by image resolution, all images and their corresponding masks were resized to 256×256 pixels before being fed into the network. CLGDA-Net was optimized using the Adam optimizer²⁸ for 200 epochs with a batch size of 32. The initial learning rate was set to 0.001. No additional data augmentation was introduced, so that the performance differences could be evaluated under the original data distribution. All comparative and ablation experiments followed the same training protocol to provide a fair basis for performance comparison. The corresponding training and validation curves of Dice, MCC, and Jaccard are presented in Fig. 10. The first row shows the results obtained on Landslide4Sense, whereas the second row corresponds to the Bijie dataset. As training progresses, the three training metrics on both datasets increase steadily, indicating that CLGDA-Net gradually learns more discriminative representations for landslide regions. The validation curves exhibit relatively noticeable fluctuations during the early training stage but become progressively more stable as the number of epochs increases. In particular, after approximately 100 epochs, the validation metrics on both datasets tend to remain within a relatively stable range, demonstrating that the optimization process has largely converged. Overall, the evolution of the three metrics shows a consistent trend across the two datasets, indicating that CLGDA-Net can achieve stable learning behavior under different landslide image distributions.

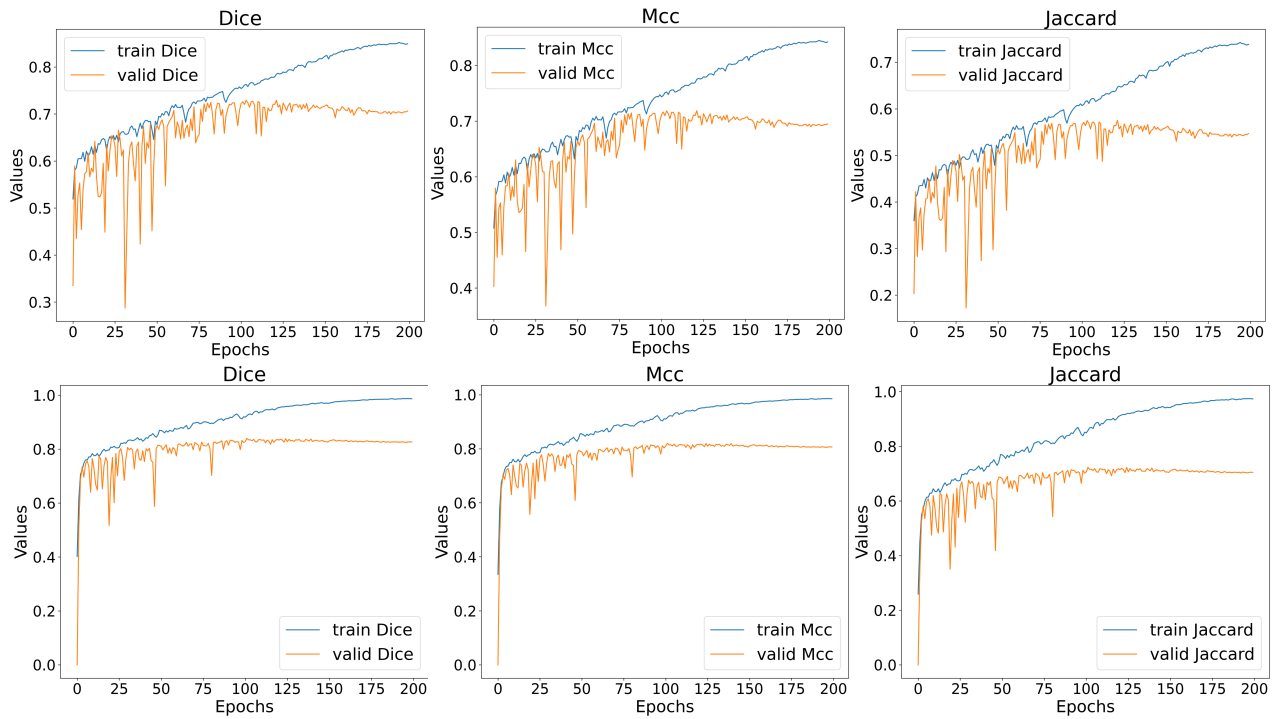


Figure 10. Training and validation curves of Dice, MCC, and Jaccard for CLGDA-Net on the Landslide4Sense and Bijie datasets. The first row shows the results on Landslide4Sense, while the second row presents the results on Bijie.

Ablation study

To investigate the contribution of each proposed component in CLGDA-Net, ablation experiments were conducted on the Landslide4Sense dataset. The results are summarized in Table 1. In these experiments, the Baseline denotes the basic architecture of CLGDA-Net without the proposed CLGCFM, MDCAM, and MSDAFM modules, which is constructed to evaluate the effectiveness of each individual component. Starting from the original U-Net, the Baseline achieves Dice, MCC, and Jaccard scores of 0.7026, 0.6916, and 0.5483, indicating that the basic encoder-decoder structure already provides a reasonable capability for landslide segmentation. After introducing CLGCFM, the performance increases to 0.7145/0.7022/0.5568 in terms of Dice, MCC, and Jaccard. This improvement demonstrates that cross-level feature interaction can effectively reduce the semantic discrepancy between encoder features and provide more discriminative representations for decoding. By further incorporating MDCAM, the three metrics increase to 0.7242, 0.7127, and 0.5681. The improvement indicates that multi-dilation contextual modeling enables the network to capture landslide regions with diverse scales and complex spatial structures. Finally, the complete model equipped with MSDAFM achieves the best performance, reaching Dice, MCC, and

Jaccard values of 0.7315, 0.7187, and 0.5774. This result verifies that multi-scale decoder feature recalibration provides complementary information and further enhances segmentation accuracy. Overall, the ablation results demonstrate that each proposed module contributes positively to CLGDA-Net, and the progressive integration of CLGCFM, MDCAM, and MSDAFM leads to continuous performance improvements.

Method	Dice	MCC	Jaccard
U-Net	0.6731	0.6592	0.5096
Baseline	0.7026	0.6916	0.5483
Baseline+CLGCFM	0.714	0.7022	0.5568
Baseline+CLGCFM+MDCAM	0.7242	0.7127	0.5681
Baseline+CLGCFM+MDCAM+MSDAFM (CLGDA-Net)	0.7315	0.7187	0.5774

Table 1. Ablation study on the Landslide4Sense dataset.

Comparison experiments

To evaluate the segmentation capability of CLGDA-Net, comparative experiments were conducted on the Landslide4Sense dataset against several representative segmentation networks, including U-Net¹, MBSNet²⁹, LOANet³⁰, FBSNet³¹, ELANet³², CCBANet³³, C2F-Net³⁴, and ABCNet³⁵. All methods were evaluated using Dice, MCC, and Jaccard under the same experimental settings, and the quantitative results are summarized in Table 2. As shown in Table 2, CLGDA-Net achieves the highest performance across all three evaluation metrics, with Dice, MCC, and Jaccard scores of 0.7315, 0.7187, and 0.5774. Among the compared methods, ABCNet obtains relatively competitive results, reaching 0.7128/0.7006/0.5548, while MBSNet achieves 0.7088/0.6968/0.5501. Compared with ABCNet, CLGDA-Net improves Dice, MCC, and Jaccard by 0.0187, 0.0181, and 0.0226. In comparison with the conventional U-Net, the corresponding improvements reach 0.0584, 0.0595, and 0.0678. These results indicate that the proposed cross-level gated fusion, multi-dilation contextual enhancement, and multi-scale dual-attention aggregation enable CLGDA-Net to obtain more discriminative feature representations and more accurate landslide delineation under complex remote sensing conditions.

Method	Dice	MCC	Jaccard
U-Net ¹	0.6731	0.6592	0.5096
MBSNet ²⁹	0.7088	0.6968	0.5501
LOANet ³⁰	0.6669	0.6540	0.5027
FBSNet ³¹	0.6733	0.6625	0.5115
ELANet ³²	0.6831	0.6743	0.5221
CCBANet ³³	0.6972	0.6853	0.5364
C2F-Net ³⁴	0.6823	0.6687	0.5202
ABCNet ³⁵	0.7128	0.7006	0.5548
CLGDA-Net	0.7315	0.7187	0.5774

Table 2. Comparison experiments on the Landslide4Sense dataset.

As shown in Fig. 11, the segmentation results produced by different methods exhibit noticeable differences in both regional completeness and boundary preservation. U-Net, LOANet, FBSNet, and ELANet tend to miss some small or fragmented landslide regions, particularly when the target occupies only a limited portion of the image or presents an irregular shape. MBSNet, CCBANet, and C2F-Net improve the continuity of the predicted regions to some extent, but several samples still contain incomplete boundaries or local false responses. ABCNet provides relatively competitive predictions and preserves more landslide structures; however, some fine-scale regions are still not fully recovered. In comparison, CLGDA-Net produces segmentation maps that are generally more consistent with the ground-truth annotations. It better preserves small and discontinuous landslide areas while maintaining more complete regional structures in larger targets. At the same time, the predicted boundaries are more closely aligned with the annotated contours, with fewer obvious omissions and spurious responses in complex background regions. These qualitative observations are consistent with the quantitative results in Table 2, further demonstrating the effectiveness of CLGCFM, MDCAM, and MSDAFM in improving cross-level feature interaction, multi-scale contextual representation, and decoder feature fusion.

As shown in Table 3, CLGDA-Net demonstrates strong segmentation performance on the Bijie dataset. Among all compared methods, CLGDA-Net achieves the highest Dice, MCC, and Jaccard scores of 0.8316, 0.8126, and 0.7118. CCBANet obtains

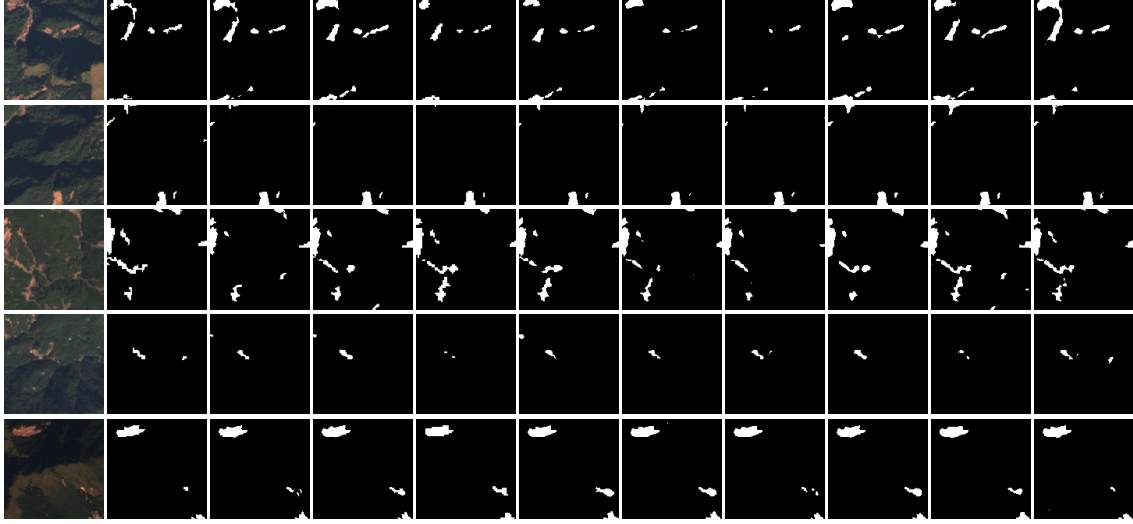


Figure 11. Visual comparison of different segmentation methods on the Landslide4Sense dataset. From left to right: original images, ground-truth masks, U-Net, MBSNet, LOANet, FBSNet, ELANet, CCBANet, C2F-Net, ABCNet, and CLGDA-Net.

the second-best results, reaching 0.8285/0.8094/0.7075, followed by C2F-Net with 0.8229/0.8025/0.6995. Compared with CCBANet, CLGDA-Net improves Dice, MCC, and Jaccard by 0.0031, 0.0032, and 0.0043. Relative to the basic U-Net, the corresponding gains reach 0.0717, 0.0790, and 0.0988. These results indicate that the proposed network maintains competitive performance under the different image characteristics and landslide distributions of the Bijie dataset. The improvements further suggest that cross-level gated fusion, multi-dilation contextual modeling, and multi-scale dual-attention aggregation can jointly enhance the representation of landslide regions and improve segmentation accuracy.

Method	Dice	MCC	Jaccard
U-Net ¹	0.7599	0.7336	0.6130
MBSNet ²⁹	0.8145	0.7940	0.6874
LOANet ³⁰	0.8077	0.7873	0.6777
FBSNet ³¹	0.8162	0.7956	0.6896
ELANet ³²	0.8038	0.7816	0.6722
CCBANet ³³	0.8285	0.8094	0.7075
C2F-Net ³⁴	0.8229	0.8025	0.6995
ABCNet ³⁵	0.8168	0.7960	0.6904
CLGDA-Net	0.8316	0.8126	0.7118

Table 3. Comparison experiments on the Bijie dataset.

As illustrated in Fig. 12, the Bijie dataset contains landslide regions with relatively diverse shapes and compact spatial distributions, which places higher demands on contour reconstruction and region consistency. Several comparison methods show obvious discrepancies from the ground-truth masks. In some cases, U-Net, LOANet, and ELANet produce incomplete target regions, while FBSNet and CCBANet occasionally introduce excessive responses around the landslide boundary. C2F-Net and ABCNet generally preserve the main target structure more effectively, but local shape distortion and boundary deviation can still be observed in several samples. By contrast, CLGDA-Net provides predictions with more coherent region shapes and better contour agreement with the annotations. Its segmentation results more consistently retain the overall morphology of the landslide regions while reducing both under-segmentation and unnecessary expansion into surrounding background areas. This advantage is particularly evident for samples with irregular outlines and varying target sizes. The qualitative results in Fig. 12 therefore complement the quantitative comparison in Table 3, indicating that CLGDA-Net can maintain reliable segmentation behavior on the Bijie dataset under different target appearances and background conditions.

Analysis of computational efficiency

In addition to segmentation accuracy, the computational cost of different methods was further evaluated using GFLOPs, parameter count (Params), and frames per second (FPS). The corresponding results on the Landslide4Sense dataset are

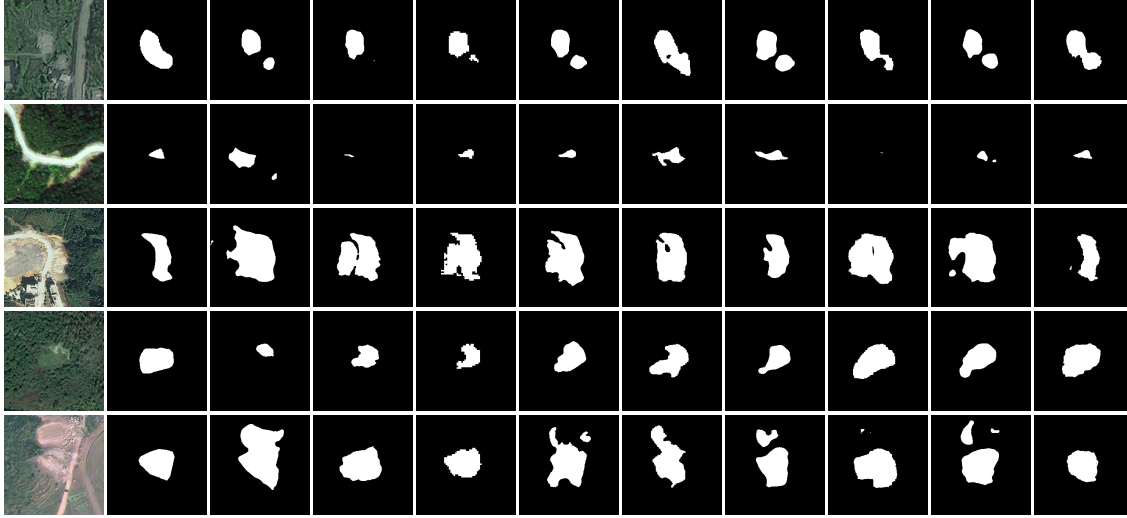


Figure 12. Visual comparison of different segmentation methods on the Bijie dataset. From left to right: original images, ground-truth masks, U-Net, MBSNet, LOANet, FBSNet, ELANet, CCBANet, C2F-Net, ABCNet, and CLGDA-Net.

reported in Table 4. As shown in Table 4, lightweight models such as LOANet, FBSNet, and ELANet require relatively low computational costs, with GFLOPs of 1.39, 2.87, and 1.50. U-Net also maintains a relatively small model size and achieves the highest inference speed of 218.78 FPS. In contrast, CLGDA-Net requires 46.11 GFLOPs and contains 30.73 M parameters, indicating that the introduction of cross-level feature interaction, multi-dilation contextual modeling, and multi-scale dual-attention fusion increases the computational burden of the network. Nevertheless, CLGDA-Net achieves an inference speed of 43.16 FPS, which is higher than those of CCBANet, C2F-Net, and ABCNet, whose FPS values are 38.46, 34.43, and 12.81. Combined with the segmentation results in Table 2, these findings indicate that CLGDA-Net currently places greater emphasis on feature representation and segmentation accuracy rather than model compactness. Although the proposed network achieves improved landslide segmentation performance, its computational complexity and parameter size remain relatively high. Therefore, reducing model complexity while preserving segmentation accuracy will be an important direction of our future work. In subsequent studies, lightweight convolution, channel compression, and more efficient feature-fusion strategies will be investigated to reduce the number of parameters and computational operations, thereby improving the deployment efficiency of CLGDA-Net on resource-constrained platforms.

Method	GFLOPs	Params(M)	FPS
U-Net ¹	3.49	1.95	218.78
MBSNet ²⁹	7.60	3.81	98.78
LOANet ³⁰	1.39	1.39	70.12
FBSNet ³¹	2.87	0.60	13.60
ELANet ³²	1.50	0.74	63.85
CCBANet ³³	13.97	31.58	38.46
C2F-Net ³⁴	7.06	26.49	34.43
ABCNet ³⁵	6.34	9.44	12.81
CLGDA-Net	46.11	30.73	43.16

Table 4. Analysis of computational efficiency on the Landslide4Sense dataset.

Optimizer selection

To examine the influence of different optimization strategies on the training of CLGDA-Net, seven commonly used optimizers were evaluated on the Landslide4Sense dataset, including Adam, Adamax, AdamW, NAdam, RMSprop, Rprop, and SGD. The corresponding Dice, MCC, and Jaccard results are reported in Table 5. As shown in Table 5, Adam achieves the best overall performance, with Dice, MCC, and Jaccard scores of 0.7315, 0.7187, and 0.5774. NAdam ranks second, obtaining 0.7246/0.7114/0.5687, which is slightly lower than Adam across all three metrics. AdamW and Adamax produce similar results, while RMSprop, SGD, and Rprop show comparatively weaker segmentation performance. Among them, Rprop yields the

lowest Dice and Jaccard values of 0.6948 and 0.5337. The results indicate that the choice of optimizer has a noticeable effect on the convergence behavior and final segmentation accuracy of CLGDA-Net. Compared with the other tested methods, Adam provides a more suitable optimization process for the proposed network on Landslide4Sense. Therefore, Adam is selected as the optimizer for all subsequent experiments in this study.

Optimizer	GFLOPs	Params(M)	FPS
Adam	0.7315	0.7187	0.5774
Adamax	0.7041	0.6904	0.5443
AdamW	0.7053	0.6929	0.5459
NAdam	0.7246	0.7114	0.5687
RMSprop	0.7022	0.6891	0.5417
Rprop	0.6948	0.6805	0.5337
SGD	0.7002	0.6877	0.5398

Table 5. Optimizer selection on the Landslide4Sense dataset.

Conclusion

This study proposes CLGDA-Net, a cross-level gated dual-attention network for landslide segmentation in remote sensing imagery. The network is designed to improve hierarchical feature interaction, contextual representation, and multi-scale decoder fusion. Specifically, CLGCFM enhances information exchange between adjacent encoder stages through adaptive gated fusion and receptive-field refinement. MDCAM enlarges the contextual range at the bottleneck by combining multiple dilation rates with attention-based recalibration. In addition, MSDAFM integrates multi-level decoder features through DPSCAM and SASAM, allowing the network to model complementary information from both scale-channel and spatial perspectives. Experiments on the Landslide4Sense and Bijie datasets demonstrate the effectiveness of the proposed method. CLGDA-Net achieves Dice/MCC/Jaccard scores of 0.7315/0.7187/0.5774 on Landslide4Sense and 0.8316/0.8126/0.7118 on Bijie, outperforming the comparison methods under the adopted experimental settings. The ablation results further confirm that CLGCFM, MDCAM, and MSDAFM each contribute positively to segmentation performance, while the qualitative comparisons show improved region integrity and boundary delineation for landslides with different shapes and scales. Although CLGDA-Net provides improved segmentation accuracy, its computational complexity remains relatively high, with 46.11 GFLOPs and 30.73 M parameters. Therefore, future work will focus on reducing the computational burden while maintaining segmentation performance. Lightweight convolutional operations, channel compression, and more efficient multi-scale feature fusion strategies will be investigated to improve the practical deployment capability of the network. In addition, further validation on larger and more geographically diverse landslide datasets will be considered to examine the robustness and generalization of CLGDA-Net in complex remote sensing environments.

References

1. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
2. Fu, Y., Zhong, H. & Fang, C. Attention u-mamba: A simple and efficient method for landslide segmentation. *IEEE Geosci. Remote. Sens. Lett.* **22**, 1–5 (2025).
3. Li, G. *et al.* A landslide area segmentation method based on an improved unet. *Sci. Reports* **15**, 11852 (2025).
4. Tang, M., He, Y., Aslam, M., Akpokodje, E. & Jilani, S. F. Enhanced u-net++ for improved semantic segmentation in landslide detection. *Sensors* **25**, 2670 (2025).
5. Li, H. *et al.* Ep-unet++: Advancing landslide monitoring through enhanced remote sensing semantic segmentation. *IEEE Transactions on Geosci. Remote. Sens.* (2026).
6. He, X., Wu, X., Wang, X., Ren, X. & Guo, J. Cross-domain coseismic landslide segmentation: local boundary enhancement & global pixel contrastive learning. *Geomatics, Nat. Hazards Risk* **17**, 2623104 (2026).
7. Emani, G. F. *et al.* Sugarfusenet: Diffusion-driven domain adaptation and bimodal bitemporal fusion for advancing global landslide segmentation on novel gbmt-slid dataset. *Neurocomputing* **696**, 134060 (2026).
8. Chai, C. *et al.* Swmfa-unet: Joint spatial-wavelet-fourier modeling for landslide segmentation in remote sensing images. *Adv. Space Res.* (2026).

9. You, H. *et al.* As-unet: A lightweight u-net with asymmetric strip attention and joint gating for rgb optical landslide segmentation. *Sensors* **26**, 4820 (2026).
10. Liu, J. *et al.* Vfm-mome: A remote sensing landslide image segmentation network guided by a visual foundation model and a mixture of mamba experts. *Remote. Sens.* **18**, 2293 (2026).
11. Tang, Y. *et al.* Geoslide-xmamba: A spectral-topographic boundary-aware state-space network for landslide semantic segmentation. *Sensors* **26**, 4146 (2026).
12. Zhong, D. *et al.* Dfg-net: a dual-stream frequency-guided network for landslide semantic segmentation from remote sensing imagery. *Int. J. Digit. Earth* **19**, 2628476 (2026).
13. Liu, S., Huang, D. *et al.* Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 385–400 (2018).
14. Yang, Y., Cao, J., Zhao, H., Chang, Z. & Wang, W. High frequency domain enhancement and channel attention module for multi-view stereo. *Comput. Electr. Eng.* **121**, 109855 (2025).
15. Sun, L. *et al.* Kcam-senet: Speech enhancement network with kan-based channel attention module. *Speech Commun.* **182**, 103425 (2026).
16. Gong, P. *et al.* 2.5d hau-net with gated spatial attention for automatic hippocampus segmentation in mri. *J. Neurosci. Methods* **431**, 110730 (2026).
17. Zhong, Q., Luo, J., Fang, J., others & Chen, Y. Two-stage cascaded vision transformer with spatial attention for dense settlement detection in remote sensing imagery. *npj Herit. Sci.* **13**, 211 (2025).
18. Luo, H., Zhou, D. M., Cheng, Y. J. & Wang, S. Q. Mpeda-net: A lightweight brain tumor segmentation network using multi-perspective extraction and dense attention. *Biomed. Signal Process. Control.* **91**, 106054 (2024).
19. Yuan, H. J., Chen, L. N. & He, X. F. Mmunet: Morphological feature enhancement network for colon cancer segmentation in pathological images. *Biomed. Signal Process. Control.* **91**, 105927 (2024).
20. Ghorbanzadeh, O., Xu, Y., Ghamisi, P., Kopp, M. & Kreil, D. Landslide4sense: Reference benchmark data and deep learning models for landslide detection. *IEEE Transactions on Geosci. Remote. Sens.* **60**, 5633017 (2022).
21. Ji, S., Yu, D., Shen, C., Li, W. & Xu, Q. Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides* **17**, 1337–1352 (2020).
22. Li, Z., Zhang, X., Zhao, M., Shi, F. & Zhou, W. Direction-guided network for retinal vessel segmentation in octa images. *Biomed. Signal Process. Control.* **103**, 107455 (2025).
23. Feng, R. *et al.* Clac-net: A composite medical image segmentation framework using self-attention and cross-layer asymmetric connections. *The Vis. Comput.* **41**, 3925–3955 (2025).
24. Wan, M. *et al.* Virus diffusion algorithm: A novel metaheuristic algorithm for segmenting the spun crack. *Meas. Sci. Technol.* **36**, 045601 (2025).
25. Shang, Z., Yu, C., Huang, H. & Li, R. Dcnet: A lightweight retinal vessel segmentation network. *Digit. Signal Process.* **153**, 104651 (2024).
26. Abdel-Nabi, H., Ali, M. Z. & Awajan, A. A joint multi-decoder dual-attention u-net framework for tumor segmentation in whole slide images. *J. King Saud Univ. - Comput. Inf. Sci.* **35**, 101835 (2023).
27. Zeng, Q., Yang, L., Li, Y., Xie, L. & Feng, Y. Rgvpsseg: Multimodal information fusion network for retinogeniculate visual pathway segmentation. *Med. & Biol. Eng. & Comput.* **63**, 1397–1411 (2025).
28. Bai, L. *et al.* Msanlfnet: Semantic segmentation network with multiscale attention and nonlocal filters for high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* **19**, 6512405 (2022).
29. Jin, S., Yu, S., Peng, J., Wang, H. & Zhao, Y. A novel medical image segmentation approach by using multi-branch segmentation network based on local and global information synchronous learning. *Sci. Reports* **13**, 6762 (2023).
30. Han, X., Liu, Y., Liu, G., Lin, Y. & Liu, Q. Loanet: A lightweight network using object attention for extracting buildings and roads from uav aerial remote sensing images. *PeerJ Comput. Sci.* **9**, e1467 (2023).
31. Gao, G. *et al.* Fbsnet: A fast bilateral symmetrical network for real-time semantic segmentation. *IEEE Transactions on Multimed.* **25**, 3273–3283 (2023).
32. Yi, Q., Dai, G., Shi, M., Huang, Z. & Luo, A. Elanet: Effective lightweight attention-guided network for real-time semantic segmentation. *Neural Process. Lett.* **55**, 6425–6442 (2023).

33. Nguyen, T. C. *et al.* Ccbanet: Cascading context and balancing attention for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 633–643 (Springer, Cham, Switzerland, 2021).
34. Sun, Y., Chen, G., Zhou, T., Zhang, Y. & Liu, N. Context-aware cross-level fusion network for camouflaged object detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 1025–1031 (IJCAI, Montreal, Canada, 2021).
35. Pan, P., Wang, H., Wang, C. & Nie, C. Abc: Attention with bilinear correlation for infrared small target detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2381–2386 (IEEE, Brisbane, Australia, 2023).

Acknowledgements

Thanks to the editorial team and all the anonymous reviewers who helped us improve the quality of this paper.

Author contributions statement

Conceptualization and supervision by J.L.; methodology and validation by J.L. and L.Y.; software, formal analysis, investigation, visualization, and writing by L.Y.; writing review and editing by J.L. All authors have read and agreed to the published version of the manuscript.

Data availability

The datasets used in this study are publicly available. The Landslide4Sense dataset is accessible via its public repository at <https://www.kaggle.com/datasets/niyarrbarman/landslide-divided?select=dataset>. The Bijie dataset is available from its public repository at http://gpcv.whu.edu.cn/data/Bijie_pages.html.

Funding

This work was supported in part by the Science and Technology Major Projects of Quzhou under Grant 2025K250.

Conflict of Interests

The authors have no conflict of interest to declare.

Additional information

Correspondence and requests for materials should be addressed to J.L.