MTFNet: Multimodal Fusion with Topographic Priors for Landslide Semantic Segmentation

1st Zilan Ning
Hunan University
Changsha, China
zilanning@hnu.edu.cn

2nd Juan Luo*
Hunan University
Changsha, China
juanluo@hnu.edu.cn

3rd Kexuan Feng
Hunan University
Changsha, China
kexuanfeng@hnu.edu.cn

4th Ying Qiao
Hunan University
Changsha, China
qy2020@hnu.edu.cn

5th Anping Liu
Hunan University
Changsha, China
liuanping15@hnu.edu.cn

***Abstract*—Landslide segmentation from remote sensing images is important for risk mapping and disaster response, but it remains difficult in complex scenes with unclear boundaries and confusing background. Traditional methods that rely only on RGB images often produce missed regions and false detections, since landslides can look similar to bare soil and roads. To address this, we propose Multimodal Topographic Fusion Network (MTFNet), a three branch framework that combines a frozen SAM ViT-L spatial encoder with lightweight ATL adaptation, a frequency based texture encoder that strengthens high frequency boundary details, and a Digital Elevation Model (DEM) based topographic encoder that provides geometry information such as slope change and surface shape. The multi level features from the three branches are fused with the Global Cross-attention Context Gating (GCCG) to build global interaction across branches and position wise gating, and then refined with the Cross scale Dual Residual Mixing (CDRM) to mix stable semantics with fine local details. The proposed method is evaluated on the Bijie and Landslide4Sense datasets and performs better than baseline models, with overall improvements in key metrics such as mIoU, Precision, and F1-score. These results demonstrate the effectiveness of the proposed multimodal fusion design for robust landslide segmentation in complex scenes.**

***Index Terms*—landslide semantic segmentation, remote sensing, multimodal feature fusion, DEM, vision foundation model**

I. Introduction

Landslides are a common and highly damaging natural hazard, often causing serious loss of life and property [1]. Fast and accurate landslide mapping is essential for risk assessment, post disaster surveys, and emergency response. Traditional landslide identification methods, such as field surveys [2] and manual visual interpretation [3], rely on experts' experience to judge landslide landform features. These methods usually require a lot of time and manpower, which makes them hard to use for large areas and urgent disaster response needs.

Remote sensing provides new ways to observe the ground surface. It covers large areas and has short data update cycles, which makes it more feasible to extract landslide information over large regions [4]. With the development of artificial intelligence, deep learning based semantic segmentation has become a main technique for landslide mapping. It can output pixel level masks in an end to end way and greatly improves automation [5]. Common semantic segmentation models include Fully Convolutional Networks (FCN) [6], U-Net [7], and DeepLab [8], which help improve the accuracy and efficiency of landslide mapping.

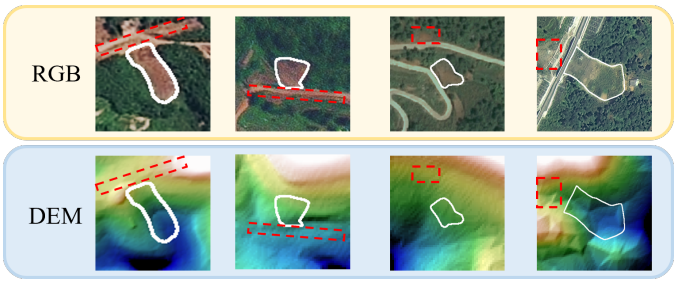

Fig. 1. White outlines indicate the ground truth landslide boundaries. Red dashed boxes mark visually confusing non-landslide regions in RGB. The DEM view provides terrain context that helps distinguish these regions and supports more reliable landslide delineation.

In existing studies, landslide segmentation usually takes optical remote sensing images as the main input. However, landslides can look similar to bare soil, roads, and other land cover types, and their boundaries are often unclear. When using only optical images, models may miss landslide areas or wrongly label non landslide regions as landslides [9]. Therefore, some works introduce DEM data to add surface geometry information, and use two branch or multi branch networks for feature fusion [10]. As shown in Fig. 1, landslides and some non landslide areas can have similar appearance in RGB images, while the DEM view highlights topographic variations that are helpful for segmentation. A DEM records surface elevation in a grid format and is widely used for topographic analysis [11]. It can also be used to derive topographic factors such as slope and curvature to describe changes in surface shape [12]. In landslide segmentation, this geometry information is often used to reduce false alarms and improve boundary localization [13].

Besides using DEM data, vision foundation models have also drawn attention in segmentation. Segment Anything Model (SAM) provides strong general segmentation features [14]. However, due to domain gaps in remote sensing data and the high training cost, directly transferring such models to remote sensing landslide segmentation is often not ideal. In practice, transfer learning or parameter efficient adaptation is usually needed to reduce cost and improve performance [15].

To address these issues, we propose MTFNet, a multimodal

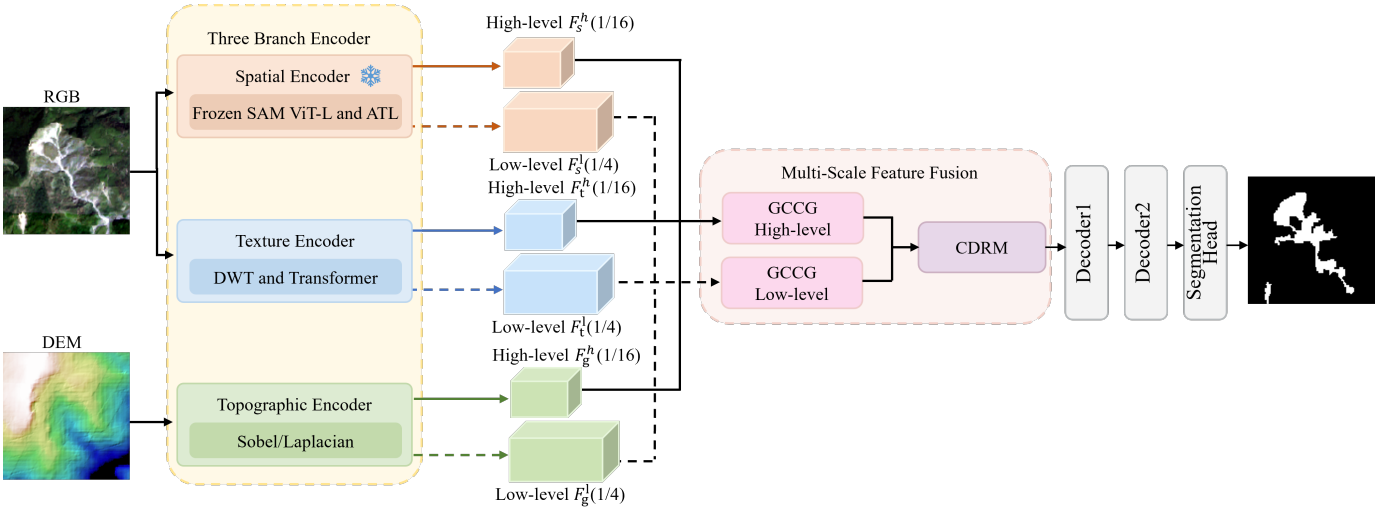

Fig. 2. Overall architecture of MTFNet. A three branch encoder extracts spatial features from RGB with frozen SAM ViT-L and ATL, texture features from the texture branch, and topographic features from the DEM branch. High level and low level features are fused by GCCG and then refined by CDRM. The decoder and segmentation head output the landslide mask.

three branch collaborative learning network for landslide segmentation. The network contains a spatial branch, a texture branch, and a topographic branch. The spatial branch provides the main semantic features. The texture branch extracts high frequency information in the frequency domain from RGB images to add boundary related details. The topographic branch encodes DEM derived topographic information, such as slope changes and local surface shape, to provide geometry constraints in regions with similar appearance. All three branches produce high level semantic features and low level detail features, which serve as layered inputs for later fusion.

For feature fusion, we use a two stage multimodal fusion scheme to make better use of the complementarity among branches. In the first stage, we introduce the Global Cross-attention Context Gating (GCCG). It lets spatial features interact globally with texture and topographic information, and produces position related gating weights to adjust the fusion degree in different regions. In the second stage, we apply the Cross scale Dual Residual Mixing (CDRM), to fuse high level semantic features with low level detail features, allowing the final features to keep both overall structure and boundary details. Considering the domain gap in remote sensing data and the training cost, we adopt a parameter efficient light adaptation in the spatial branch. We freeze the main parameters and add a conditioned tuning module guided by texture and topographic information, enabling the spatial features to receive multimodal guidance during encoding.

Our main contributions are as follows:

- We propose a three branch encoder that combines topographic information and frequency domain texture. In the topographic branch, DEM data provides elevation variation information to distinguish landslides from background regions with boundaries that look similar.
- We introduce a two stage fusion strategy with GCCG and CDRM. GCCG builds global cross modal interaction and generates spatial gates for DEM and texture refinement, and CDRM mixes high level semantics with low level details across scales. It strengthens DEM guidance in hard regions and avoids fusion in easy regions.
- Based on a frozen SAM, we introduce a parameter efficient, multimodal conditioned light adaptation together with a topo aware positional prior, which guides the spatial branch with texture and topographic information during encoding.

II. Method

A. Overall Architecture

As shown in Fig. 2, MTFNet uses a three branch encoder with multi scale fusion for landslide semantic segmentation. The spatial branch is the main stream and provides reliable semantics from RGB by using a frozen SAM ViT-L with lightweight adaptation. The texture branch focuses on high frequency appearance changes that often happen near landslide borders, which helps reduce broken edges. The topographic branch encodes DEM based geometry, such as slope changes and surface shape, which helps distinguish regions with similar RGB but different landform. Each branch outputs two feature levels, a low level feature with higher resolution for local details and a high level feature with lower resolution for stable semantics. We fuse the three branches at both levels with GCCG module, and then refine cross scale information with CDRM module. Finally, a lightweight decoder and a segmentation head output the landslide mask.

B. Multi Branch Encoder

As shown in Fig. 3, given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a DEM map $D \in \mathbb{R}^{H\times W}$, we build three branches $m \in \{s, t, g\}$ for spatial, texture, and topographic encoding. Each branch outputs low level and high level features, denoted as F_m^l and F_m^h. The low level feature keeps fine patterns

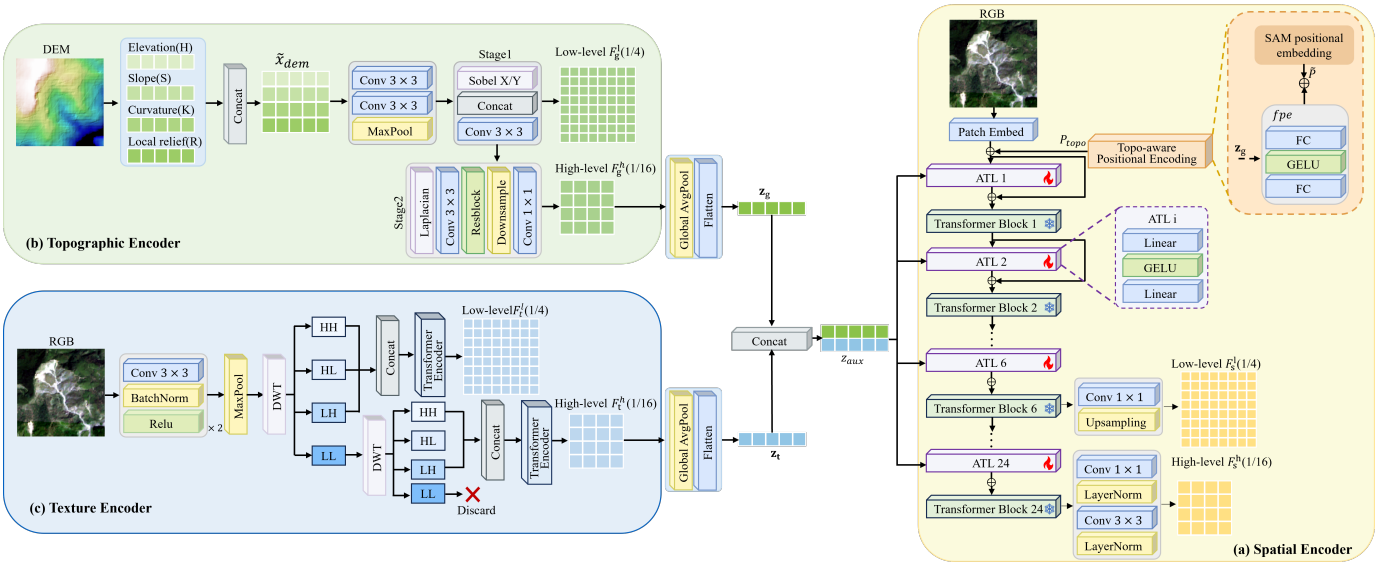

Fig. 3. Structure of the three branch encoder. (a) Spatial encoder uses frozen SAM and inserts ATL. Topo aware positional encoding P_{topo} is added to obtain F_s^l and F_s^h. (b) Topographic encoder extracts DEM features to obtain F_g^l, F_g^h, and z_g. (c) Texture encoder extracts frequency based features to obtain F_t^l, F_t^h, and z_t. The auxiliary vector $z_{\text{aux}} = [z_t|z_g]$ is used as the condition for ATL.

and edges, while the high level feature is more stable for region understanding. These features are fused by GCCG and CDRM to combine semantics, boundaries, and topographic information.

Topographic Encoder. Landslides are strongly related to landform, and DEM provides geometry information that can be hard to infer from RGB. We build a four channel DEM derived input $\tilde{X}_{\text{dem}}$ from elevation and its local changes. Let $H(x, y)$ be elevation and (x, y) be pixel coordinates. We compute slope S, curvature term K, and local relief R, then stack H, S, K, and R as a four channel map $\tilde{X}_{\text{dem}}$:

$$S = \sqrt{\left(\frac{\partial H}{\partial x}\right)^2 + \left(\frac{\partial H}{\partial y}\right)^2} \tag{1}$$

$$K = \nabla^2 H = \frac{\partial^2 H}{\partial x^2} + \frac{\partial^2 H}{\partial y^2} \tag{2}$$

$$R = H - \text{Mean}_\Omega(H) \tag{3}$$

We feed $\tilde{X}_{\text{dem}}$ into a lightweight CNN. In the first step, we use a Sobel operator to capture first order changes and edges, followed by a 3×3 convolution to produce F_g^l. In the second step, we use a Laplacian operator to capture second order shape changes, followed by a residual block and one downsampling step to produce F_g^h. We apply global average pooling on F_g^h to obtain a global vector z_g. This vector provides geometry guidance for the spatial branch, helping the model follow topographic changes and reducing confusion between flat background and landslide regions.

Texture Encoder. Landslide borders often show clear high frequency patterns, and we extract frequency based texture features from RGB to improve edge quality. We use shallow convolutions and max pooling for efficient feature extraction, then apply DWT to split features into one low frequency part LL and three high frequency parts LH, HL, and HH. Since border clues mainly lie in the high frequency parts, we concatenate LH, HL, and HH and use a lightweight Transformer to produce the low level feature F_t^l. We then downsample LL with DWT, keep the high frequency parts again, and use a second lightweight Transformer to produce the high level feature F_t^h. Global average pooling on F_t^h gives a global texture vector z_t, which summarizes border related patterns and supports later fusion.

Spatial Encoder. The spatial branch uses the SAM image encoder as the semantic backbone and freezes its parameters to reduce training cost and keep strong general features. To introduce topographic guidance early, we adjust the SAM positional encoding using a bias derived from DEM features:

$$P_{\text{topo}} = \tilde{P} + f_{pe}(\text{GAP}(F_g^h)) \tag{4}$$

Here $\tilde{P}$ is the interpolated SAM positional encoding and f_{pe} maps the DEM global summary into a positional bias. This makes the spatial tokens more sensitive to topographic changes that often align with landslide shapes. To further use multi modal context while keeping the backbone frozen, we apply ATL guided by a condition vector z_{aux} formed by concatenating z_t and z_g:

$$X_s' = X_s + \text{ATL}(X_s; z_{\text{aux}}) \tag{5}$$

This design lets the model adapt spatial tokens using global texture and geometry summaries, improving task fit without heavy fine tuning.

C. Multi Scale Feature Fusion

Global Cross-attention Context Gating (GCCG). GCCG takes the spatial branch as the main stream and uses texture

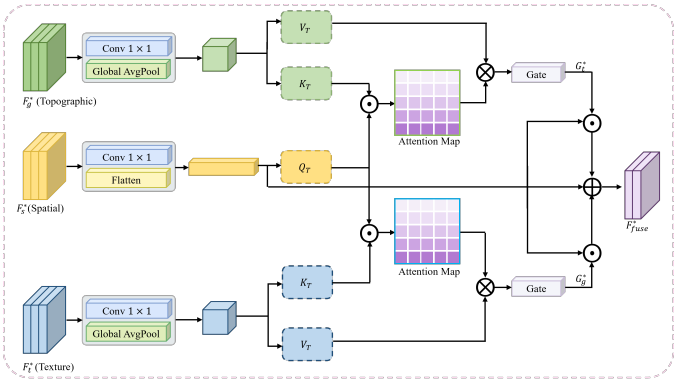

Fig. 4. Structure of GCCG. Spatial features are used as Query, and global context from texture and topography is used as Key and Value. Gates are used for residual fusion.

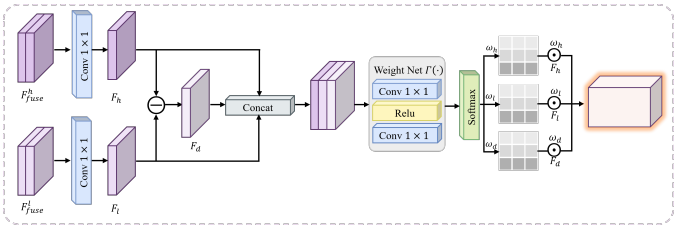

Fig. 5. Structure of CDRM. 1×1 convolutions align the high level and low level features to obtain F_h and F_l. The residual F_d is computed and weights are predicted by $\Gamma(\cdot)$. Softmax produces pixel wise weights to fuse the three features.

and topographic branches to adjust it. As shown in Fig. 4, given F_s^*, F_t^*, F_g^* with $* \in \{h, l\}$, we first align channels by 1×1 convolutions. We use global average pooling on texture and DEM features to form compact context tokens, and apply standard cross attention with spatial tokens as Query. We compute interaction features from the texture and topographic branches via cross attention:

$$A_m^* = \mathrm{CrossAttn}(Q_s^*, K_m^*, V_m^*) \tag{6}$$

where $Q_s^* = \mathrm{Flatten}(\tilde{F}_s^*)$ and $K_m^* = V_m^* = \mathrm{GAP}(\tilde{F}_m^*)$, with $m \in \{t, g\}$ ($m = t$ for texture and $m = g$ for topography). A lightweight gating head converts interaction features into position-wise gates:

$$G_m^* = \sigma\big(W_m^h \,\mathrm{LN}(A_m^*)\big), \quad m \in \{t, g\} \tag{7}$$

Finally, we fuse by gated residual addition:

$$F_{\text{fuse}}^* = S^* + S^* \odot G_t^* + S^* \odot G_g^* \tag{8}$$

This keeps the spatial features stable and allows texture and topographic information to refine difficult areas. It avoids excessive fusion in easy regions and focuses more on boundaries and areas with clear topographic information.

Cross Scale Dual Residual Mixing (CDRM). The CDRM module merges high level semantics with low level details in a scale aware way. Direct addition often causes scale mismatch, which can blur borders or introduce local errors. As shown in Fig. 5, CDRM uses the residual between two scales as an extra correction signal. We align channels and denote the high level feature as F_h and the low level feature as F_l. We build a cross scale residual feature F_d and concatenate F_d, F_h, and F_l to predict mixing weights:

$$[\omega_d, \omega_h, \omega_l] = Softmax\left(\Gamma(Concat(F_d, F_h, F_l))\right) \tag{9}$$

$$F_{mix} = \omega_h \odot F_h + \omega_l \odot F_l + \omega_d \odot F_d \tag{10}$$

We then apply a lightweight refinement Φ and add it back:

$$F_{mid} = F_{mix} + \Phi(F_{mix}) \tag{11}$$

CDRM improves the balance between region correctness and edge sharpness, allowing the decoder to receive features that are stable and locally accurate.

III. Experiments

A. Experimental Setup

Dataset. We train and test our model on two public landslide semantic segmentation datasets, Bijie landslide dataset and Landslide4Sense dataset. The Bijie landslide dataset provides TripleSat high resolution optical images and DEM data with landslide boundaries annotated in shapefile format [16]. It contains 770 landslide samples and 2003 non landslide samples. In our experiments, we split the dataset into train : validation : test = 7 : 2 : 1, and resize all inputs to 256×256.

Landslide4Sense dataset is a global benchmark released for the Landslide4Sense 2022 challenge by IARAI [17]. It includes multispectral Sentinel-2 imagery and topographic layers, including DEM and slope, with pixel level binary masks. We use the official train, validation, test split with 3799, 245, and 800 patches, respectively, and the provided 128×128 patches. For preprocessing, we select RGB bands for training and apply standard data augmentation on the training set, including random flips and rotations.

Implementation details. All experiments are conducted on a single RTX 4090 GPU. We train the model for 100 epochs with a batch size of 4 using AdamW, weight decay of 5×10^{-4}, and a cosine annealing learning rate schedule with an initial learning rate of 2×10^{-4}. Landslide detection is formulated as binary semantic segmentation. We evaluate model performance using mean-IoU (mIoU), Landslide-IoU, precision (P), recall (REC), F1-score, and overall accuracy (OA).

Baseline methods. We compare MTFNet with UNet [7], DeepLabV3+ [18], SegFormer [19], TransLandSeg [20], DCT-Unet++ [21], and CAFNet [22]. All baselines are trained and tested under the same data split, input setting, and evaluation metrics. We use RGB and DEM as inputs for all methods; when a model does not natively support multi modal input, we concatenate RGB and DEM along the channel dimension.

B. Performance Comparison on the Bijie Landslide Dataset

Quantitative evaluation. Table I shows that MTFNet achieves the best performance on the Bijie landslide dataset, with 89.06% mIoU and 80.43% Landslide-IoU. It also reaches 93.03% precision and 95.35% recall. Compared with the second-best TransLandSeg, MTFNet improves mIoU and Landslide-IoU by 0.91% and 1.48%, respectively, indicating

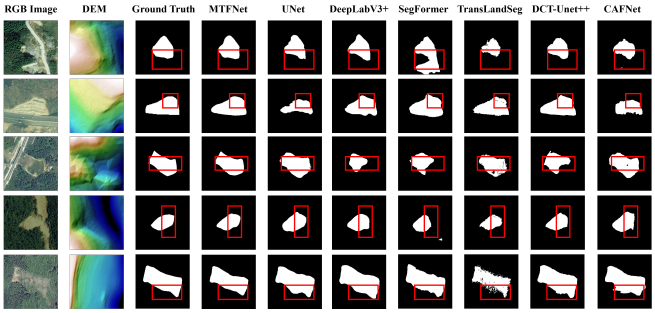

Fig. 6. Visualization results of different methods on selected images from the Bijie landslide dataset. White pixels indicate landslide regions, and red boxes mark error prone areas. The fourth column shows MTFNet (ours).

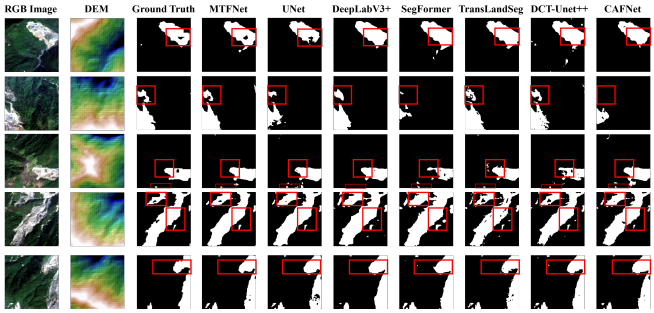

Fig. 7. Visualization results of different methods on selected images from the Landslide4Sense dataset. White pixels indicate landslide regions, and red boxes mark error prone areas. The fourth column shows MTFNet (ours).

TABLE I
QUANTITATIVE COMPARISON RESULTS ON THE BIJIE LANDSLIDE DATASET.

Method	mIoU (%)	Landslide-IoU (%)	P (%)	REC (%)	F1-score (%)	OA (%)
UNet	85.43	74.27	89.98	93.75	91.75	96.89
DeepLabv3+	84.11	71.85	90.41	91.38	90.89	96.68
SegFormer	76.06	58.23	83.40	87.35	85.22	94.37
TransLandSeg	88.15	78.95	92.52	94.43	93.45	97.59
DCT-Unet++	85.73	74.69	84.09	86.99	85.51	97.05
CAFNet	82.42	69.03	88.25	91.46	89.77	96.16
MTFNet (ours)	**89.06**	**80.43**	**93.03**	**95.35**	**93.84**	**97.70**

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE LANDSLIDE4SENSE DATASET.

Method	mIoU (%)	Landslide-IoU (%)	P (%)	REC (%)	F1-score (%)	OA (%)
UNet	68.63	39.15	74.77	81.37	77.66	98.12
DeepLabv3+	64.89	32.51	68.77	**83.30**	73.84	97.30
SegFormer	66.06	34.39	71.32	80.43	75.01	97.75
TransLandSeg	71.05	43.50	79.12	80.85	79.96	98.61
DCT-Unet++	66.25	34.68	72.01	79.56	75.20	97.85
CAFNet	67.67	37.09	76.53	76.70	76.61	98.26
MTFNet (ours)	**74.05**	**49.22**	**84.43**	81.15	**82.70**	**98.89**

more accurate localization and fewer false alarms. Overall, CNN and generic Transformer baselines tend to produce coarser or fragmented boundaries in blurry transition areas, while MTFNet better preserves complete landslide regions.

Qualitative evaluation. As shown in Fig. 6, on the Bijie dataset, MTFNet produces more complete and stable landslide regions and keeps boundaries more consistent, with fewer missing parts and spurious edges. In contrast, other methods more often show fragmented predictions or boundary shifts, especially in transition areas and along blurry edges.

C. Performance Comparison on the Landslide4Sense Dataset

Quantitative evaluation. As shown in Table II, on the Landslide4Sense dataset, the overall performance is slightly lower than that on the Bijie dataset because the images provide fewer clear details and the landslide pixels are relatively scarce in many scenes. As shown in Table II, MTFNet achieves the best results, reaching 74.05% mIoU and 49.22% Landslide-IoU, and it also leads in precision, F1-score, and OA. Compared with the second best TransLandSeg, MTFNet improves mIoU and Landslide-IoU by 3.00% and 5.72%, which indicates more accurate and more complete landslide foreground under cross-region settings.

Qualitative evaluation. As shown in Fig. 7, on the Landslide4Sense dataset, MTFNet better preserves thin and complex landslide structures and yields fewer false detections in diverse backgrounds. Many baselines tend to produce coarse boundaries or scattered false positives, which is consistent with the quantitative results.

D. Ablation Study

We conduct seven ablation experiments under the same data split and training settings to evaluate how the multi modal branch design and fusion strategy contribute to landslide semantic segmentation. Quantitative results are reported in Table III. The ablations validate the texture and topographic branches, compare different fusion designs, and assess the effect of lightweight ATL when the SAM image encoder is frozen.

To measure the contribution of each branch, we set the first three groups. Group 0 uses only the SAM spatial branch and achieves mIoU of 86.60% and Landslide-IoU of 76.30% on Bijie, and mIoU of 67.72% on Landslide4Sense dataset. Group 1 adds the texture branch and improves overall performance, especially on Landslide4Sense dataset where Landslide-IoU increases to 40.71%, indicating that high frequency texture signals complement spatial features. Group 2 replaces the texture branch with the topographic branch and further improves results, with precision on Landslide4Sense dataset reaching 84.96%, showing that DEM based geometric priors help suppress false detections.

To compare fusion strategies, we set Group 3 to Group 5. Group 3 uses all three branches with simple fusion and increases mIoU on Landslide4Sense dataset to 72.68%, but the gains are limited. Group 4 introduces GCCG and improves results on both datasets, indicating that hierarchical gating better selects useful texture and topographic information at different scales. Group 5 further adds CDRM, raising Landslide-IoU on Bijie to 78.20% and precision on Landslide4Sense dataset to 85.59%, suggesting that cross scale interaction helps combine high level semantics with low level details for more complete regions and fewer false alarms.

Finally, Group 6 enables ATL on top of the full fusion model and achieves the best performance on both datasets, with mIoU of 89.06% on Bijie and 74.05% on Landslide4Sense dataset. This indicates that bringing multi modal guidance into the encoding stage yields more effective spatial features than

TABLE III
ABLATION STUDY ON THE BIJIE AND LANDSLIDE4SENSE DATASET.

Group	Spatial Encoder	Texture Encoder	Topographic Encoder	GCCG	CDRM	ATL	Bijie Landslide Dataset						Landslide4Sense Dataset					
							mIoU (%)	Landslide-IoU (%)	P (%)	REC (%)	F1-score (%)	OA (%)	mIoU (%)	Landslide-IoU (%)	P (%)	REC (%)	F1-score (%)	OA (%)
0	✓						86.60	76.30	92.19	93.35	92.23	97.05	67.72	36.93	78.72	74.78	76.60	98.53
1	✓	✓					86.97	76.97	92.59	93.28	92.43	97.10	69.68	40.71	80.87	76.64	78.59	98.66
2	✓		✓				87.07	77.10	92.10	93.90	92.51	97.14	70.01	41.24	84.96	74.67	78.89	98.79
3	✓	✓	✓				87.26	77.37	92.03	94.15	92.64	97.24	72.68	46.53	84.02	79.27	81.46	98.84
4	✓	✓	✓	✓			87.34	77.46	92.09	94.20	92.70	97.28	72.92	47.03	83.02	80.46	81.69	98.81
5	✓	✓	✓	✓	✓		87.72	78.20	92.31	94.35	92.96	97.34	73.03	47.18	**85.59**	78.72	81.77	98.89
6	✓	✓	✓	✓	✓	✓	**89.06**	**80.43**	**93.03**	**95.35**	**93.84**	**97.70**	**74.05**	**49.22**	84.43	**81.15**	**82.70**	**98.89**

fusion only at the decoder stage.

IV. CONCLUSION

This work focuses on landslide semantic segmentation in remote sensing images, where backgrounds are confusing and boundaries are often unclear. We propose MTFNet, which combines RGB based spatial semantics, frequency based texture details, and DEM based geometry information in a three branch design. We use DEM derived topographic priors, such as slope changes and surface shape, to help distinguish landslides from visually similar regions. We also design a two stage fusion scheme. The GCCG module builds global cross branch interaction and produces position wise gating, and the CDRM module mixes stable semantics with fine details across scales. Experiments on the Bijie and Landslide4Sense datasets show that MTFNet performs better than baseline methods, and ablation studies verify the effect of each branch and each fusion module. Future work will explore stronger generalization under larger region shifts and more efficient deployment for large scale mapping.

REFERENCES

[1] L. M. Highland and P. Bobrowsky, *The landslide handbook-A guide to understanding landslides*. US Geological Survey, 2008, no. 1325.

[2] H. Wang, L. Zhang, K. Yin, H. Luo, and J. Li, "Landslide identification using machine learning," *Geoscience Frontiers*, vol. 12, no. 1, pp. 351–364, 2021.

[3] A. Mohan, A. K. Singh, B. Kumar, and R. Dwivedi, "Review on remote sensing methods for landslide detection using machine and deep learning," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 7, p. e3998, 2021.

[4] C. J. Van Westen, E. Castellanos, and S. L. Kuriakose, "Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview," *Engineering geology*, vol. 102, no. 3-4, pp. 112–131, 2008.

[5] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[9] L. Wu, R. Liu, N. Ju, A. Zhang, J. Gou, G. He, and Y. Lei, "Landslide mapping based on a hybrid cnn-transformer network and deep transfer learning using remote sensing images with topographic and spectral features," *International Journal of Applied Earth Observation and Geoinformation*, vol. 126, p. 103612, 2024.

[10] Y. Jin, X. Li, S. Zhu, B. Tong, F. Chen, R. Cui, and J. Huang, "Accurate landslide identification by multisource data fusion analysis with improved feature extraction backbone network," *Geomatics, Natural Hazards and Risk*, vol. 13, no. 1, pp. 2313–2332, 2022.

[11] E. Guilbert and B. Moulin, "Data structures for 2d representation of terrain models," *Encyclopedia*, vol. 5, no. 3, p. 98, 2025.

[12] I. D. Moore, R. Grayson, and A. Ladson, "Digital terrain modelling: a review of hydrological, geomorphological, and biological applications," *Hydrological processes*, vol. 5, no. 1, pp. 3–30, 1991.

[13] X. Liu, Y. Peng, Z. Lu, W. Li, J. Yu, D. Ge, and W. Xiang, "Feature-fusion segmentation network for landslide detection using high-resolution remote sensing images and digital elevation model data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[15] C. Hou, J. Yu, D. Ge, L. Yang, L. Xi, Y. Pang, and Y. Wen, "Translandseg: a transfer learning approach for landslide semantic segmentation based on vision foundation model," *arXiv preprint arXiv:2403.10127*, 2024.

[16] S. Ji, D. Yu, C. Shen, W. Li, and Q. Xu, "Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks," *Landslides*, vol. 17, no. 6, pp. 1337–1352, 2020.

[17] O. Ghorbanzadeh, Y. Xu, H. Zhao, J. Wang, Y. Zhong, D. Zhao, Q. Zang, S. Wang, F. Zhang, Y. Shi *et al.*, "The outcome of the 2022 landslide4sense competition: Advanced landslide detection from multisource satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9927–9942, 2022.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[20] C. Hou, J. Yu, D. Ge, L. Yang, L. Xi, Y. Pang, and Y. Wen, "A transfer learning approach for landslide semantic segmentation based on visual foundation model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[21] J. Wang, Q. Zhang, H. Xie, Y. Chen, and R. Sun, "Enhanced dual-channel model-based with improved unet++ network for landslide monitoring and region extraction in remote sensing images," *Remote Sensing*, vol. 16, no. 16, p. 2990, 2024.

[22] X. Cai, C. Song, Z. Li, Y. Chen, B. Chen, J. Du, C. Yu, W. Zhu, and J. Peng, "A lightweight context-aware adaptive fusion network for automatic identification of active landslides," *International Journal of Applied Earth Observation and Geoinformation*, vol. 144, p. 104882, 2025.