Self-Modulation Aggregation within Dense Skip Connections for Mapping of Retrogressive Thaw Slumps

Yi Yuan[1,2], Guiyun Zhou[1], Zhishuai Zheng[1], Minghui Chang[1,2], Andrea Maldonado[2], Binbin He[1], Martin Werner[2]

[1] School of Resources and Environment, University of Electronic Science and Technology of China, China;
[2] Big Geospatial Data Management, Technical University of Munich, Germany

Keywords: Retrogressive thaw slumps, Permafrost degradation, Remote sensing, Deep learning, Semantic segmentation.

Abstract

Accurate mapping of retrogressive thaw slumps (RTSs) in permafrost regions remains challenging due to their irregular morphology, blurred boundaries, and strong spatial correlation. This paper proposes a lightweight multi-level self-modulation (MLSM) module embedded into the UNet++ backbone to enhance non-local feature modeling for high-resolution image segmentation. The overall framework is built upon a UNet++ backbone with dense skip connections, where the proposed MLSM module adaptively fuses multi-scale contextual information to enhance feature coherence across spatially correlated regions. By incorporating a low-rank regularization term, MLSM dynamically modulates feature responses according to structural variations, allowing attention to adapt to spatially complex RTS regions. The integration of depth-wise convolution and channel recalibration further refines feature aggregation efficiency. Experimental evaluations on the Maxar dataset demonstrate that the proposed method achieves superior segmentation accuracy and smoother boundary delineation compared with existing models. The proposed framework provides a lightweight, robust, and computationally efficient solution for delineating irregular and morphologically complex RTSs.

1. Introduction

1.1 Remote Sensing of Retrogressive Thaw Slumps

RTSs are among the most widespread dynamic thermokarst landforms that develop when ice-rich permafrost thaws and ground material collapses and flows downslope (Nitze et al., 2025). Monitoring and accurately delineating RTSs are essential for understanding permafrost degradation, carbon release, and associated geomorphic changes in permafrost landscapes. High spatial resolution remote sensing imagery offers valuable means for RTS identification and mapping (Rodenhizer et al., 2024).

However, remote sensing imagery exhibits fundamental differences from both natural and medical images in terms of data acquisition, spatial structure, and semantic representation (Zhu et al., 2017). Natural photographic images, which are commonly used in conventional computer vision, are typically captured from near-ground perspectives within the visible spectrum under relatively stable illumination and scale conditions. Natural images contain visually coherent objects of interest, such as vehicles or humans, whose spatial boundaries are well defined and directly interpretable (Russakovsky et al., 2015). Medical images are acquired through highly controlled protocols with standardized intensity scales and anatomy-centred semantics. Medical images generally correspond to a single, well-defined semantic unit or region of interest that serves as the direct target for segmentation or detection tasks (Litjens et al., 2017).

Satellite remote sensing images cover extensive and heterogeneous landscapes, are characterized by complex radiometric and geometric distortions, multi-spectral or hyperspectral channels, and strong temporal variability. The same geographical object may appear at different scales or orientations, while individual patches frequently contain mixed land-cover types due to coarse spatial resolution. For natural or medical images, one image or patch usually corresponds to a complete object or lesion to be identified. In remote sensing, a single geographical feature, such as an RTS, is often divided into multiple patches for network training. The complexity of RTS spatial patterns and continuous morphological evolution poses significant challenges to reliable segmentation and interpretation (Huang et al., 2020, Nitze et al., 2021). Each patch represents only a portion of the feature, with reduced spatial resolution but high spatial correlation across adjacent patches. This inter-patch dependency challenges the assumption of sample independence and necessitates models that can learn spatial continuity and contextual relations at multiple scales. These distinctive characteristics highlight the need for specialized architectures and training strategies tailored to capture long-range spatial dependencies and structural continuity of remote sensing data.

1.2 Related Work

RTS delineation from optical remote sensing imagery has advanced from manual inventories toward automated learning-based pipelines capable of handling irregular outlines, subtle boundaries, and strong spatial clustering. Early applications on the Tibetan Plateau demonstrated that encoder–decoder segmentation networks can produce reliable RTSs inventories from high-resolution CubeSat data while contributing standardized annotation resources (Huang et al., 2020). Subsequent cross-regional studies in Siberia and Canada indicated that generalization improves when training samples are heterogeneous and labeling practices are harmonized (Yang et al., 2023). End-to-end frameworks have further incorporated elevation-derived constraints to reduce confusion with spectrally similar landforms and to strengthen boundary stability in complex terrain (Nitze et al., 2021). Within this evolving methodological landscape, lightweight attention-enhanced architectures have been introduced to strengthen feature representation without relying on large-scale pretraining (Yuan et al., 2024). Complementary work has leveraged change cues to suppress false positives in cluttered landscapes and to capture rapid geomorphic

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-687-2026

evolution (Wu et al., 2025). Recent syntheses emphasize the need for architectures that encode non-local spatial context, integrate morphological priors, and remain data-efficient (Nitze et al., 2025). High-resolution optical delineation frameworks have also been developed to produce decimeter-scale products that enable consistent regional pattern analysis across multiple subregions of the Tibetan Plateau (Yuan et al., 2025). These approaches illustrate a broader trend toward models that jointly exploit spectral, structural, and contextual information while maintaining scalability across different subregions of permafrost landscapes. Building on these advances, this study focuses on improving non-local feature aggregation and cross-scale representation in a fully automated segmentation framework.

1.3 Attention-based Aggregation

Attention-based aggregation has emerged as a fundamental mechanism in deep learning for image analysis. It enables networks to compute content-dependent weighted combinations of feature elements rather than relying on fixed convolutional kernels (Vaswani et al., 2017). Attention mechanisms are commonly implemented through a key-query-value formulation followed by a softmax normalization, which assigns adaptive weights to spatial, channel, or temporal features. As a result, pixels or regions that carry more relevant information contribute more significantly to the final representation (Vaswani et al., 2017).

The principal advantages of attention include an adaptive receptive field, improved global contextual reasoning, and enhanced interpretability through visualizable attention maps. In practical applications, attention is realized through several architectural forms. Self-attention modules perform intra-feature aggregation, while cross-attention modules condition feature learning on external information such as text or another modality. Non-local attention structures are designed to capture long-range dependencies that cannot be modeled by conventional local convolutions (Wang et al., 2018).

The conventional dense attention mechanism suffers from quadratic computational complexity with respect to input size. To address this limitation, various efficient variants have been developed. These include local or shifted window attention used in hierarchical vision transformers, sparse and dilated attention patterns, low-rank factorization methods, kernel-based linearized attention, and hierarchy-based aggregation schemes that reduce computational cost while maintaining contextual information (Liu et al., 2021, Krzysztof et al., 2021). Attention mechanisms yield significant improvements when tasks require global reasoning, long-range dependency modeling, or cross-modal feature alignment.

However, their advantage is relatively limited in tasks dominated by local texture discrimination. Compact attention modules have also been proposed to reweight feature channels, as seen in squeeze-and-excitation and efficient channel attention designs (Hu et al., 2018, Wang et al., 2020). These modules often combine convolutional inductive biases with attention mechanisms to preserve sample efficiency and improve stability during training. Additional design considerations involve the use of positional encoding, normalization, residual connections, and appropriate regularization to prevent overfitting and ensure convergence. Efficient and adaptive attention mechanisms are still needed to better model long-range spatial dependencies and structural continuity in earth science applications.

1.4 Overview of the Proposed Method

To address the limitations of existing convolution-based segmentation networks in capturing long-range spatial dependencies and structural continuity, this study introduces a lightweight self-modulation module that equips convolutional architectures with Transformer-like global receptive fields while maintaining computational efficiency. The proposed module integrates a low-rank constraint to achieve dynamic self-modulation of feature responses, enabling the network to adapt its attention to complex structural variations commonly observed in RTS mapping. By embedding the MLSM module within the dense skip connections of the UNet++ framework, the method enhances non-local contextual interaction between encoder and decoder layers, thereby improving the structural consistency and spatial coherence of segmentation results. The design incorporates depth-wise and channel-wise operations that provide an effective balance between computational cost and representational power, allowing global dependency modeling without sacrificing efficiency.

2. Materials

2.1 Study Area

The Beiluhe region is located in the central part of the Tibet Plateau (TP) and serves as one of the most active permafrost areas in the Northern Hemisphere, as shown in Fig. 1. It lies along the Qinghai–Tibet Highway and Railway corridor, where continuous permafrost is widely distributed and highly sensitive to climate warming and human disturbance. The region has an average elevation above 4,500 m, with low annual temperatures and pronounced seasonal freeze–thaw cycles. Rich ground ice and fine-grained sediments have led to the development of extensive thermokarst landforms, where RTSs are the dominant and most dynamic type. Over recent decades, these RTSs have expanded rapidly, forming steep headwalls and lobate flow features that significantly modify the surface morphology. The Beiluhe area has therefore become a representative site for investigating permafrost degradation and geomorphic processes across the Third Pole, where numerous field and remote sensing studies have been conducted to analyze thaw-induced landscape changes and related environmental impacts. The Tibetan Plateau boundary dataset was obtained from the Third Pole Environment Data Center.

2.2 Satellite Data

The Maxar imagery used in this study was acquired from the Google Earth platform, provided by Maxar Technologies as part of its Open Data Program (https://www.maxar.com/open-data). Maxar WorldView data provide superior boundary detail for small slumps (Rodenhizer et al., 2024). Their high spatial and spectral fidelity allows for the accurate delineation of retrogressive thaw slump boundaries and enhances the capability of deep learning models to capture subtle morphological transitions within complex permafrost landscapes on the Tibet Plateau. The Maxar dataset we used consists of satellite imagery with a spatial resolution of approximately 2.5 m, offering detailed representation of surface features suitable for geomorphological and environmental applications. The imagery includes three visible bands, which enable effective discrimination of surface materials and vegetation conditions. Maxar satellites provide high radiometric quality and precise geolocation accuracy, making them widely adopted in remote sensing

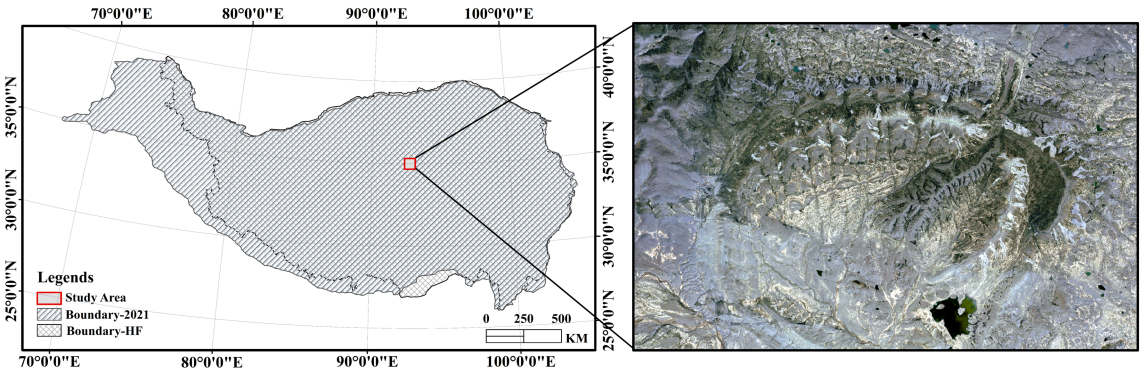

Figure 1. Location of Beiluhe study area on the Tibet Plateau.

studies of landform dynamics and natural hazard assessment. The selected images cover the Beiluhe permafrost region under clear-sky conditions, minimizing the influence of seasonal snow and cloud contamination.

2.3 Sample Generation

In this study, we used the RTS vector polygon inventory obtained from the Science Data Bank (Yuan et al., 2025), as shown in Fig. 2. Among these, to reduce subjectivity and ensure consistent delineation, three trained analysts independently digitized each slump under a standardized labeling protocol, after which a senior interpreter adjudicated the annotations and reconciled discrepancies through consensus. A subset of the inventory was corroborated by field observations to verify positional accuracy and boundary validity. For supervised learning, the vetted polygons were co-registered to the target imagery and rasterized on the native grid to produce binary masks, assigning a value of 1 to RTS pixels and 0 to background. This quality-controlled reference dataset was partitioned into training, validation, and test splits.

3. Method

3.1 UNet++

UNet++ is a standard nested encoder–decoder architecture that extends the classical U-Net by redesigning the skip connections as a set of densely connected, nested convolutional blocks, which progressively narrow the semantic gap between encoder and decoder features (Zhou et al., 2019). Here, $i \in \{0, \ldots, I\}$ denotes the depth level and $j \in \{0, \ldots, J\}$ the stage index along the skip pathway. The feature map at level i and stage j is denoted by $X^{i,j}$. For $j = 0$, encoder features are obtained by successive down-sampling:

$$X^{i,0} = f_{i,0}\big(D(X^{i-1,0})\big), \quad i = 1, \ldots, I, \tag{1}$$

where $D(\cdot)$ represents a down-sampling operator (max pooling or strided convolution), and $f_{i,0}(\cdot)$ denotes a convolutional mapping at level i consisting of convolution, normalization, and non-linear activation.

For $j \geq 1$, the nested decoder nodes are constructed by aggregating encoder and intermediate decoder features through dense skip connections. A basic intermediate node can be written as

$$X^{i,j} = f_{i,j}\big(\mathrm{Concat}\big(X^{i-1,j},\, X^{i,j-1}\big)\big), \tag{2}$$

where $\mathrm{Concat}(\cdot,\cdot)$ denotes channel-wise concatenation. In the full UNet++ design, Eq. (2) is generalized to include all preceding nodes along the same skip pathway:

$$X^{i,j} = f_{i,j}\Big(\mathrm{Concat}\big(X^{i-1,j},\, X^{i,0},\, X^{i,1}, \ldots, X^{i,j-1}\big)\Big), \tag{3}$$

which enhances feature reuse and facilitates gradient flow across different resolutions.

UNet++ further employs deep supervision by attaching segmentation heads to several decoder outputs at the top level.

Here, $f_k(\cdot)$ denote the prediction head associated with node $X^{0,k}$, and $\hat{Y}^{(k)} = f_k(X^{0,k})$ the corresponding segmentation map. The final prediction is obtained as a weighted combination:

$$Y = \sum_{k=1}^{K} w_k\, \hat{Y}^{(k)}, \tag{4}$$

where w_k is a learnable coefficient satisfying $w_k \geq 0$. This nested and densely supervised design improves information flow between encoder and decoder and strengthens multi-scale contextual modeling. In the present work, UNet++ serves as the backbone, and the proposed MLSM module is embedded along the skip connections to further enhance non-local context modeling for RTS segmentation, as shown in Fig. 3.

3.2 Multi-Level Self-Modulation Module (MLSM)

The proposed lightweight multi-level self-modulation module is a plug-in block that enhances non-local context modeling while preserving the efficiency of the convolutional backbone, as shown in Fig. 4. It operates on feature maps from multiple encoder levels and is inserted into the dense skip connections of UNet++. MLSM combines a local branch, which captures compact texture information, with a low-rank self-modulation branch that encodes global structural correlations. The two branches are fused to generate an adaptive modulation mask, which selectively emphasizes structurally consistent regions before multi-level aggregation.

Let $\{X^{(l)}\}_{l=1}^{L}$ denote the feature maps from L encoder levels, where $X^{(l)} \in \mathbb{R}^{B\times C\times H_l\times W_l}$ and B, C, H_l, and W_l are the batch size, number of channels, height, and width at level l, respectively. All feature maps are first projected to a common resolution (H, W) by bilinear upsampling:

$$\tilde{X}^{(l)} = \mathrm{Up}\big(X^{(l)}\big), \quad \tilde{X}^{(l)} \in \mathbb{R}^{B\times C\times H\times W}, \tag{5}$$

where $\mathrm{Up}(\cdot)$ denotes the upsampling operator.

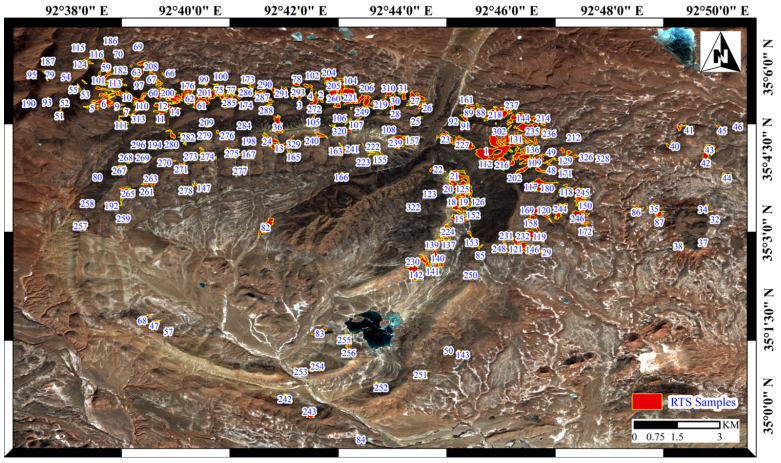

Figure 2. Distribution of the RTS polygon inventory in the Beiluhe area (Yuan et al., 2025).

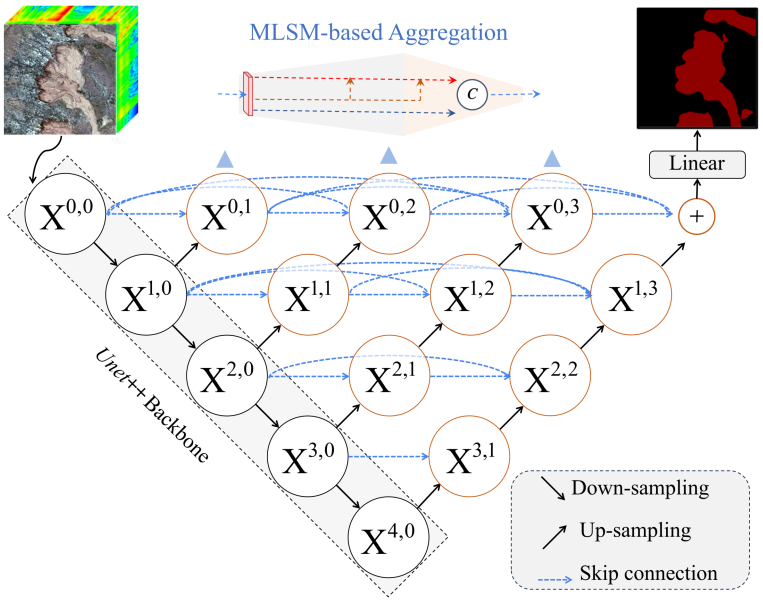

Figure 3. Overview of the proposed UNet++-based framework with multi-level self-modulation aggregation for RTS mapping.

(1) Local branch. A depth-wise convolution is applied to a spatially pooled version of $\tilde{X}^{(l)}$ to extract compact local texture descriptors:

$$x_s^{(l)} = f_{\mathrm{dw}}\big(\mathrm{Pool}(\tilde{X}^{(l)})\big), \tag{6}$$

where $\mathrm{Pool}(\cdot)$ is adaptive max pooling that reduces the spatial size to $(H/s, W/s)$ with a downscale factor s, and $f_{\mathrm{dw}}(\cdot)$ denotes a depth-wise convolutional mapping (kernel size $k = 3$) applied independently to each channel.

(2) Low-rank self-modulation branch. To capture non-local structural correlations, each feature map $\tilde{X}^{(l)}$ is flattened along the spatial dimension. For the c-th channel, we obtain

$$\tilde{X}_c^{(l)} \in \mathbb{R}^{B\times(H\cdot W)}, \tag{7}$$

and compute a soft low-rank descriptor using a combination of Frobenius and L_1 norms:

$$r^{(l)} = \big\|\tilde{X}^{(l)}\big\|_F + \lambda\big\|\tilde{X}^{(l)}\big\|_1, \tag{8}$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius and L_1 norms, respectively, and λ controls the strength of the low-rank regularization. The descriptor $r^{(l)}$ is reshaped to $\mathbb{R}^{B\times C\times 1\times 1}$ and used to modulate channel-wise responses.

(3) Multi-level aggregation. The local and low-rank branches are fused through learnable self-modulation parameters $\alpha, \beta \in \mathbb{R}^{1\times C\times 1\times 1}$:

$$T^{(l)} = \alpha \cdot x_s^{(l)} + \beta \cdot r^{(l)}, \tag{9}$$

$$M^{(l)} = \sigma\big(f_{\mathrm{m}}(T^{(l)})\big), \tag{10}$$

where $f_{\mathrm{m}}(\cdot)$ is a convolutional mapping with kernel size $k = 1$ and $\sigma(\cdot)$ denotes the GELU activation. The modulation mask $M^{(l)}$ is then upsampled to the full resolution and applied to the corresponding feature map:

$$\hat{M}^{(l)} = \mathrm{Up}(M^{(l)}), \qquad \hat{X}^{(l)} = \tilde{X}^{(l)} \odot \hat{M}^{(l)}, \tag{11}$$

where $\odot$ denotes element-wise multiplication. The multi-level modulated features are concatenated and projected by a convolutional layer to form the aggregated output:

$$X^{\mathrm{mlsm}} = f_{\mathrm{agg}}\big(\mathrm{Concat}(\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(L)})\big), \tag{12}$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-687-2026 | © Author(s) 2026. CC BY 4.0 License.

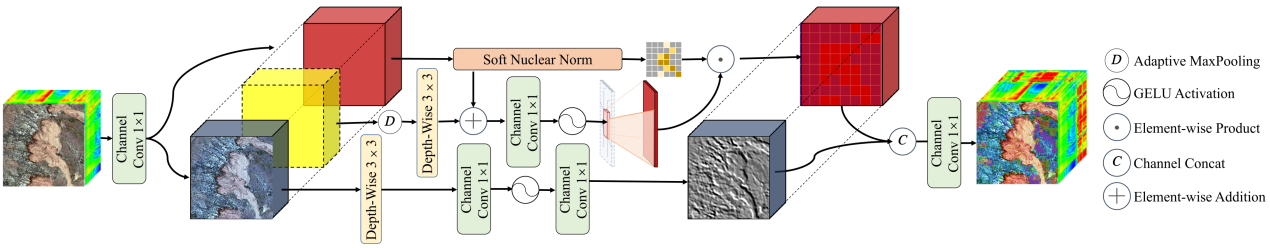

Figure 4. Lightweight multi-level self-modulation module (MLSM).

where $\mathrm{Concat}(\cdot)$ denotes channel-wise concatenation and $f_{\mathrm{agg}}(\cdot)$ is a convolutional mapping with kernel size $k = 1$. The resulting feature map $X^{\mathrm{mlsm}} \in \mathbb{R}^{B\times C\times H\times W}$ encodes multi-level, globally modulated representations with enhanced structural consistency.

3.3 Integration of MLSM

The MLSM module is integrated into the UNet++ backbone through the dense skip pathways, while the main encoder and decoder streams remain purely convolutional. Let $X^{i,j}$ denote the feature map at depth level i and skip stage j in the UNet++ topology. For each node $X^{i,j}$ that participates in a skip connection, the original concatenated feature is first processed by an MLSM block, which performs global self-modulation and low-rank-aware aggregation, and the resulting modulated feature is then forwarded to the subsequent decoder layers.

This integration strategy is consistent with the original motivation of UNet++, where nested skip connections are introduced to progressively reduce the semantic gap between encoder and decoder. By placing MLSM exclusively on the skip pathways, the global context modeling and structural consistency of cross-scale feature fusion are strengthened, while the local inductive bias and training stability of the convolutional backbone are preserved. Applying MLSM to all skip levels ensures that information exchange between encoder and decoder is consistently regularized across resolutions, which is particularly beneficial for segmenting RTSs with irregular shapes and strong spatial correlation. Owing to the lightweight design of MLSM, the additional computational overhead introduced by full-skip integration remains moderate and suitable for high-resolution remote sensing applications.

3.4 Loss Integration

The proposed network is trained in a supervised manner using RTS label masks. Let $I \in \mathbb{R}^{H\times W\times C}$ denote an input image patch and $Y^* \in \{0,1\}^{H\times W}$ the corresponding ground-truth map, where $Y_p^* = 1$ indicates RTS pixels and $Y_p^* = 0$ denotes background at pixel p. The network outputs a probability map $P \in [0,1]^{H\times W}$ after a sigmoid activation.

(1) Cross-entropy loss. To enforce pixel-wise supervision under class imbalance between foreground and background, we adopt binary cross-entropy as the segmentation loss. Let Ω be the set of pixels in the patch. The segmentation loss is defined as

$$\mathcal{L}_{\mathrm{seg}} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\left[-Y_p^*\log P_p - (1-Y_p^*)\log(1-P_p)\right], \tag{13}$$

(2) Dice loss. To improve the segmentation robustness, we also use the Dice loss, which directly measures the similarity between the prediction and the ground truth. Let P_p and Y_p^* denote the predicted probability and ground-truth label at pixel $p \in \Omega$, respectively. The Dice loss is defined as

$$\mathcal{L}_{\mathrm{dice}} = 1 - \frac{2\sum_{p\in\Omega} P_p Y_p^* + \epsilon}{\sum_{p\in\Omega} P_p + \sum_{p\in\Omega} Y_p^* + \epsilon}, \tag{14}$$

where ϵ is a small constant introduced for numerical stability. This term encourages greater overlap between the predicted RTS regions and the reference mask, and is particularly effective for handling class imbalance in binary segmentation.

(3) Overall objective. The total training objective combines the segmentation loss with the Dice term:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \alpha\,\mathcal{L}_{\mathrm{dice}}, \tag{15}$$

where $\alpha \geq 0$ controls the contribution of the Dice loss.

4. Experiments

4.1 Experimental Setup

All experiments used high-resolution Maxar imagery with a ground sampling distance of 2.5 m. Before patch extraction, each image underwent a linear contrast stretch to 0-255. Since this study focuses on RTS mapping under RGB imagery, all methods were implemented and evaluated using the same RGB input to ensure a consistent comparison setting. Patches were generated with a sliding window. Squares of 384×384 pixels were extracted with an overlap of 128 pixels, yielding 2,318 patches. A total of 263 representative patches were selected for training, and repeated experiments confirmed that this training set was sufficient to achieve stable convergence and consistent performance in this study. Of the remaining patches, a random 30% subset was drawn and split evenly into validation and test sets, which were never used for model updates. Ground-truth masks were derived from vetted RTS polygons and co-registered to the imagery.

The UNet++ backbone followed a standard nested encoder–decoder design with five resolution levels $(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16})$. Each convolutional block consisted of two 3×3 convolutions, each followed by Batch Normalization and ReLU activation. The channel width started at 32 and doubled at each downsampling stage. Downsampling was performed by 2×2 max-pooling, and upsampling was implemented by bilinear interpolation. Deep supervision heads were attached to the top-level decoder nodes, and their predictions were fused by learnable softmax-normalized weights to produce the final output. The proposed MLSM blocks were inserted along the dense skip

Table 1. Mean and standard deviation of mIoU, F1-score, and OA for different mapping methods over repeated experiments.

Method/Metric	mIoU(%)	F1-score(%)	OA(%)
DeepLab V3+	50.3 ± 3.1	57.8 ± 3.8	84.2 ± 0.9
LessNet	63.9 ± 2.2	75.7 ± 2.1	87.7 ± 0.8
AmRTSNet (RGB)	72.1 ± 1.6	82.3 ± 1.3	93.9 ± 0.4
MLSM (ours)	79.2 ± 1.3	87.6 ± 0.9	95.3 ± 0.5

pathways to enhance non-local context aggregation. The benchmark methods included the well-established DeepLab V3+, LessNet, and AmRTSNet. For a fair comparison, all models were trained using the same fixed training subset, the same batch size of 9, and an identical data order for each epoch. The training subset remained unchanged throughout the experiments. The learning schedule, stopping criteria, and all other hyperparameter settings were kept the same across runs. All experiments were conducted under the same software and hardware conditions. For all compared methods, the reported quantitative metrics are given as the mean and standard deviation computed from repeated experiments under identical settings.

4.2 Mapping Results of Image Patches

Fig. 5 shows the qualitative RTS mapping results of the four methods in representative patches of the study area. DeepLab V3+ tends to under-segment RTSs. Only the central part of the disturbed area is detected, while marginal zones and narrow downslope extensions are missed, leading to incomplete objects, broken shapes, and false negatives. LessNet improves the overall segmentation of RTS bodies and provides more stable regional coverage, but it remains conservative in narrow lobes and along headwall boundaries. Small RTS parts are often eroded, and some predicted regions contain interior gaps. AmRTSNet produces relatively large response regions, but its predictions are less reliable under the limited spectral information setting used in this dataset. In patches with bright bedrock or exposed sediment, non-RTS surfaces are more likely to be confused with RTS targets, while some shadowed or elongated active areas remain only partially extracted. By comparison, the proposed MLSM generates masks that are more complete and more consistent with the ground truth. It better preserves elongated protrusions and narrow branches, while maintaining smoother and more reasonable object geometry. The boundary shape, especially around curved headwall sections and downslope tongues, is also closer to the reference labels. These visual results suggest that MLSM offers stronger structural integrity and better separation from spectrally similar background areas.

The quantitative results in Table 1 further support the visual comparisons and show clear advantages of the proposed method. MLSM achieves the best performance across all three metrics, reaching 79.2% mIoU, 87.6% F1-score, and 95.3% OA. Compared with AmRTSNet under the same input setting, MLSM still yields consistent improvements, increasing mIoU by 7.1 percentage points, F1-score by 5.3 percentage points, and OA by 1.4 percentage points. Compared with DeepLab V3+ and LessNet, MLSM also achieves consistently higher overall accuracy values. These results indicate that MLSM provides more accurate region overlap and more reliable foreground delineation, while maintaining the highest overall classification accuracy.

LessNet narrows the gap via attention and multi-level fusion but still exhibits conservative behavior at fine structures, which lowers F1. AmRTSNet, whose architecture targets multispectral inputs, experiences separability loss when restricted to RGB, leading to increased confusion between RTS textures and spectrally similar units, which constrains mIoU. MLSM with low-rank regularization introduces lightweight global context that enhances long-range dependency modeling, limits fragmented predictions across tiles, and stabilizes region continuity. Moreover, placing the module along UNet++ dense skips improves encoder–decoder reconciliation, so multi-scale cues are aggregated with better spatial agreement and fewer contradictory activations at transitions. Omission errors for MLSM are concentrated in very low-contrast or heavily shadowed areas under RGB imagery, where the signal-to-noise ratio is limited and boundaries are weak. Nonetheless, under identical training data and optimization settings, MLSM provides the most robust and stable boundary agreement on this subset, as reflected by consistent improvements in mIoU and F1 across scenes.

5. Conclusion

This study develops MLSM, a lightweight module for improving non-local feature aggregation in RTS mapping under RGB imagery. Embedded into UNet++, the proposed design strengthens structural continuity and multi-scale contextual interaction while keeping the model efficient. Experiments show that MLSM achieves more accurate and spatially consistent RTS delineation than DeepLab V3+, LessNet, and AmRTSNet. These results indicate that MLSM is an effective and practical solution for segmenting RTSs with complex and irregular morphologies. In future work, we will further evaluate the ablation effects of individual modules and extend the framework to multi-modal data, such as the integration of digital elevation models and optical imagery, to better exploit spatial heterogeneity for more precise RTS mapping.

6. Acknowledgments

This work was supported by the Second Tibetan Plateau Scientific Expedition and Research (2022QZKK0101) and the National Natural Science Foundation of China (42271427).

References

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, L., Luo, J., Lin, Z., Niu, F., Liu, L., 2020. Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images. *Remote Sensing of Environment*, 237, 111534.

Krzysztof, C., Valerii, L., David, D., Xingyou, S., Andreea, G., Tamas, S., Peter, H., Jared, D., Afroz, M., Lukasz, K. et al., 2021. Rethinking attention with performers. *Proceedings of ICLR*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., Sánchez, C. I., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

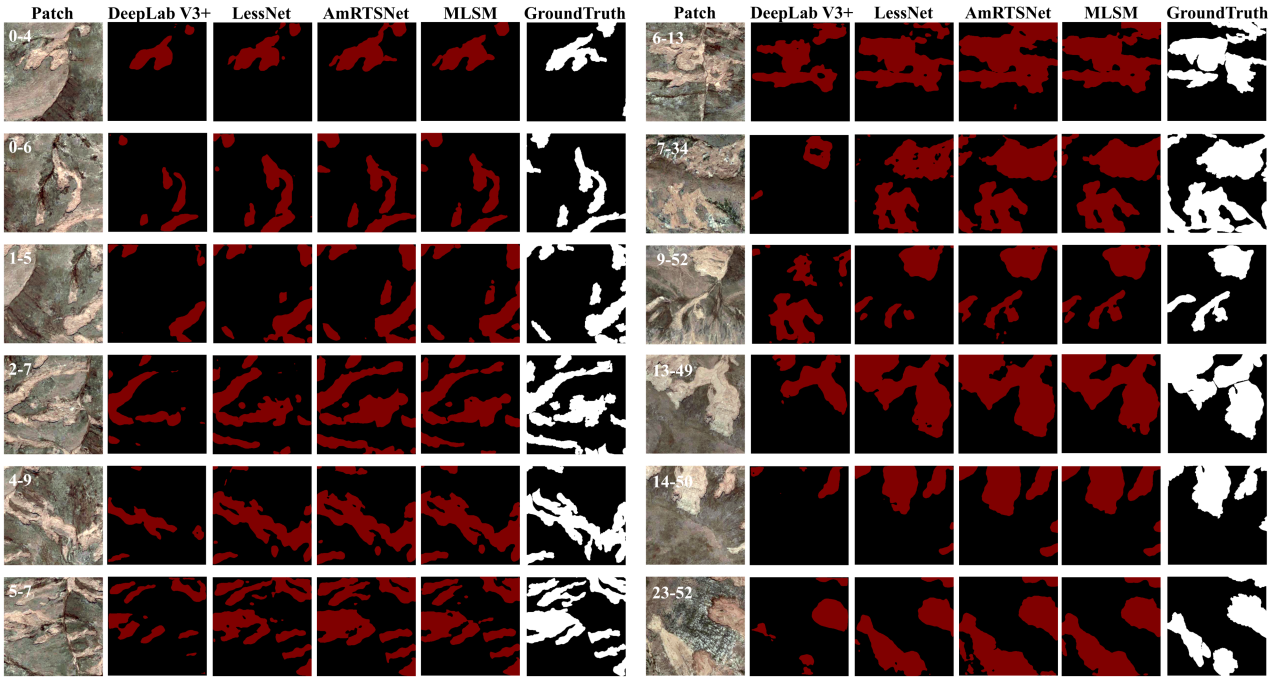

Figure 5. Comparison of RTS mapping results produced by different methods in the Beiluhe area.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Nitze, I., Heidler, K., Barth, S., Grosse, G., 2021. Developing and testing a deep learning approach for mapping retrogressive thaw slumps. *Remote Sensing*, 13(21), 4294.

Nitze, I., Heidler, K., Nesterova, N., Küpper, J., Schütt, E., Hölzer, T., Barth, S., Lara, M. J., Liljedahl, A. K., Grosse, G., 2025. DARTS: Multi-year database of AI-detected retrogressive thaw slumps in the circum-arctic permafrost region. *Scientific Data*, 12(1), 1512.

Rodenhizer, H., Yang, Y., Fiske, G., Potter, S., Windholz, T., Mullen, A., Watts, J. D., Rogers, B. M., 2024. A comparison of satellite imagery sources for automated detection of retrogressive thaw slumps. *Remote Sensing*, 16(13), 2361.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11534–11542.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wu, F., Jiang, C., Wang, C., Zou, L., Li, T., Guan, S., Tang, Y., 2025. Retrogressive thaw slumps recognition and occurrence analysis using deep learning with satellite remote sensing in the central Qinghai-Tibet Plateau. *Geomorphology*, 471, 109581.

Yang, Y., Rogers, B. M., Fiske, G., Watts, J., Potter, S., Windholz, T., Mullen, A., Nitze, I., Natali, S. M., 2023. Mapping retrogressive thaw slumps using deep neural networks. *Remote Sensing of Environment*, 288, 113495.

Yuan, Y., Zhou, G., Ding, J., Li, S., Liu, Z., He, B., 2025. Automatic mapping and pattern analysis of retrogressive thaw slumps on the central Tibetan Plateau using deep learning. *Journal of Geographical Sciences*, 35(10), 2248–2270.

Yuan, Y., Zhou, G., Su, Z., Liu, Z., Sun, W., Meng, X., Wen, J., Wu, Z., 2024. A lightweight and enhanced semantic segmentation network for mapping of retrogressive thaw slumps from Sentinel-2 images. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 130–133.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., Liang, J., 2019. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6), 1856–1867.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.