Remote Sensing of Environment 344 (2026) 115503

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Multi-source heterogeneous feature fusion framework for identifying retrogressive thaw slumps on the Qinghai-Tibet plateau

Taorui Zeng [a,b], Xiao Feng [c,*], Yuanming Lai [a,b], Thomas Glade [d]

[a] Institute of Future Civil Engineering Science and Technology, Chongqing Jiaotong University, Chongqing 400074, China
[b] Institute of Frontier Interdisciplinary Technology, Chongqing Jiaotong University, Chongqing 400074, China
[c] Department Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, P.O. Box 5048, 2600 GA Delft, the Netherlands
[d] ENGAGE—Geomorphic Systems and Risk Research, Department of Geography and Regional Research, University of Vienna, 1010 Vienna, Austria

ARTICLE INFO

Editor: Dr. Jing M. Chen

Keywords:
Retrogressive thaw slumps
Heterogeneous feature fusion strategy
Semantic segmentation models
Qinghai-Tibet plateau

ABSTRACT

Accurate detection of retrogressive thaw slumps (RTSs) remains a significant challenge due to their complex morphological features and subtle spectral contrasts with surrounding landscapes. This study presents a robust deep learning framework to improve RTS identification by integrating multi-source remote sensing data. Focusing on the permafrost-dominated central Qinghai-Tibet Plateau, we conducted three pioneering investigations: (i) a detailed RTS inventory integrated with topographic, environmental, spectral, and thermal analyses to uncover spatio-temporal distribution patterns; (ii) a comprehensive evaluation of twelve leading-edge semantic segmentation models for RTS identification; and (iii) the formulation of an innovative FusionSA-SegFormer model, which employs dual-level (pixel-level and feature-level) heterogeneous feature fusion of optical, spectral, thermal, and topographic remote sensing datasets. Our results highlight prominent RTS clustering within the 4700–4800 m elevation range, on moderate slopes (3–7°), in mid - slope to valley settings, and in proximity to water bodies. Temporally, the active evolution of RTSs from 2019 to 2024 was characterized by sustained degradation patterns, manifested as a continuous decline in vegetation indices and a concurrent rise in land surface temperatures. Comparative model assessments identified SA-SegFormer as the most effective baseline architecture. Building upon this, the proposed FusionSA-SegFormer demonstrated significant improvements, showing 8.8% IoU and 10.9% recall enhancements on the validation set, and achieving a superior F1-score of 0.843 (Precision: 0.838, Recall: 0.899) on the test set. Crucially, independent spatial and temporal transferability evaluations confirmed the framework's robust generalization capacity across unseen regions and varying years, maintaining consistent identification performance and high overall accuracies (>0.90). Furthermore, feature importance analysis emphasized the critical influence of spectral bands, particularly the blue band, alongside notable contributions from thermal indices. This work establishes a new benchmark for mapping permafrost disturbances and provides a valuable tool for monitoring thermokarst dynamics in warming climates.

1. Introduction

Permafrost, a fundamental component of the global cryosphere, is defined as ground that remains at or below 0 °C for at least two consecutive years, encompassing approximately 24% of the Northern Hemisphere's land surface (Dobinski, 2011; Zhang et al., 1999). Under the influence of global warming, permafrost degradation has intensified across Arctic and alpine regions, initiating a series of cascading environmental impacts. These include the release of greenhouse gases sequestered in permafrost organic carbon (Biskaborn et al., 2019a; Schuur et al., 2015) and the widespread development of thermokarst landforms, such as retrogressive thaw slumps (RTSs) (Lewkowicz and Way, 2019a; Nitze et al., 2018a). The Qinghai-Tibet Plateau, often referred to as the "Third Pole", exhibits particularly acute manifestations of these processes due to its unique high-altitude permafrost ecosystem, which plays a critical role in regulating regional water cycles, carbon emissions (Jiao et al., 2023a; Sun et al., 2024), and broader pan-Asian ecological stability (Yan et al., 2025).

RTSs are distinctive thermokarst landforms, typically characterized by a horseshoe- or wedge-shaped morphology, resulting from the

* Corresponding author.
E-mail addresses: zengtaorui@cqjtu.edu.cn (T. Zeng), F.X.Feng@tudelft.nl (X. Feng), ymlai@lzb.ac.cn (Y. Lai), thomas.glade@univie.ac.at (T. Glade).

https://doi.org/10.1016/j.rse.2026.115503
Received 31 October 2025; Received in revised form 10 May 2026; Accepted 24 May 2026
Available online 4 June 2026

exposure and subsequent melting of ground ice within ice-rich permafrost on slopes (Huang et al., 2023; Luo et al., 2024; Tian et al., 2026; Xia et al., 2024). This process induces abrupt slope failures through rapid permafrost thaw, triggered by natural mechanisms such as fluvial thermal erosion, active layer detachment, and coastal wave action, as well as anthropogenic disturbances like excavation (Balser et al., 2014; Jones et al., 2019; Sejourne et al., 2015). Spatially, RTSs are often localized in regional clusters where massive buried ice is preserved within glacial moraines or fine-grained sediments (e.g., marine deposits or loess), creating conditions conducive to thermal erosion and significant mass wasting (Coulombe et al., 2019; Lafrenière and Lamoureux, 2019). Recent observations in the central QTP indicate accelerated permafrost degradation, characterized by a contracting permafrost extent, deepening active layers (Chen et al., 2022; Zhang et al., 2022a), and the rapid proliferation of RTSs (Yang et al., 2025a; Yu et al., 2025). Rather than merely being passive geomorphic features, RTSs serve as both indicators and amplifiers of climate feedback mechanisms. For instance, Zhou et al. (2023) reported a 61% reduction in soil organic carbon associated with RTS events, while their formation disrupts permafrost thermal regimes, alters vegetation communities and releases greenhouse gases into the atmosphere (Lupachev et al., 2025; Niu et al., 2023). Amid global warming, the QTP has faced accelerated permafrost degradation, with intense RTS activity particularly evident in the central region, where topographic and climatic factors heighten permafrost vulnerability. Predominantly located in remote, high-altitude, uninhabited regions of the QTP, RTSs are situated in environments where harsh conditions and limited accessibility severely constrain long-term field monitoring. Consequently, remote sensing techniques have become indispensable for large-scale RTS identification, as satellite imagery provides the only viable means to systematically capture their spatial distribution and temporal evolution across vast, inaccessible terrains (Lu and Han, 2025; Yi et al., 2025). Despite their ecological and climatic importance, large-scale RTS identification remains a formidable challenge due to two primary factors: (i) the intricate geomorphological characteristics of RTS features, which complicate their identification, and (ii) the limitations of conventional segmentation models in addressing the spectral and spatial heterogeneity inherent in high-resolution satellite imagery.

The detection and monitoring of RTSs typically utilize three key methodologies: field surveys, UAV aerial photography, and remote sensing interpretation. Field surveys deliver high-resolution spatial data and serve as the most direct approach to documenting RTS morphology, scale, and temporal dynamics (Wang et al., 2016). For instance, Barnhart and Crosby (2013) used terrestrial laser scanning to describe RTS characteristics along the Selawik River in Alaska, while Wang et al. (2016) undertook a three-year study of 18 RTSs in northern Canada to explore their degradation patterns. Nevertheless, conducting long-term and large-scale field monitoring is exceptionally difficult due to the remote, high-altitude, and climatically harsh environments where RTSs are often located, compounded by the scarcity of human presence in these areas (Luo et al., 2019). UAV aerial photography offers a detailed perspective by capturing intricate geomorphic features of RTSs, such as headwall steepness and sediment textures on slump floors, through imagery with centimeter-level resolution (Obu et al., 2017; van der Sluijs et al., 2018). Yu et al. (2025) employed UAV photogrammetry to investigate the dynamic evolution and topographic changes of representative RTSs in the Beiluhe region. Despite these advantages, the restricted spatial coverage and significant costs associated with data collection and processing limit the suitability of UAVs for comprehensive, large-scale RTS identification. In contrast, satellite remote sensing has emerged as the most effective method for RTS identification across regional to continental scales.

Manual visual interpretation of satellite imagery remains a widely used and highly accurate method for identifying RTSs. This approach relies on expert analysis of high-resolution imagery to detect the distinct visual features of RTSs, such as the headwall, thaw slump floor, and slump bulge, which often display high spectral contrast between exposed bare soil and surrounding vegetation (typically manifesting as lighter tones), accompanied by distinct flow-like textural patterns. For instance, Ramage et al. Lee et al. (2017) mapped 287 RTSs along the Yukon Coast, Canada, using high-resolution satellite imagery, while Lewkowicz and Way (2019a, 2019b) studied the initiation and evolution of over 4000 RTSs on Banks Island, Canada, from 1984 to 2015. Similarly, Luo et al. (2022) identified 2669 active RTSs in the Qinghai-Tibet Plateau permafrost zone using imagery from 2018 to 2020. It also risks overlooking small or marginal RTS features, necessitating repeated verification, and demands significant professional expertise, limiting its scalability and transferability (Yi et al., 2025). In recent years, the rapid advancement of deep learning, particularly convolutional neural networks (CNNs), has introduced fully automated remote sensing interpretation as a promising direction for RTS identification (Huang et al., 2020; Yang et al., 2023b). Unlike traditional machine learning methods, deep learning automatically extracts hierarchical, multi-scale spatial and spectral features from high-dimensional imagery, significantly enhancing processing efficiency while delivering superior accuracy compared to conventional automated approaches (Zeng et al., 2023). In semantic segmentation tasks, these models use an end-to-end training approach to generate pixel-level classifications, making them ideal for detecting complex structures like RTSs (Feng et al., 2024). Models such as Unet (Nitze et al., 2021), Unet++ (Yang et al., 2023a), and the Deeplab series (Huang et al., 2020) have been successfully applied to RTS identification. For example, Huang et al. (2020) used DeeplabV3+ on Planet CubeSat imagery to delineate 220 RTSs in the North Ruzhou River region of the QTP, while Nitze et al. (2021) compared Unet, Unet++, and DeepLabv3 performance across polar regions in Canada and Russia. Yang et al. (2023) further evaluated variations of the Unet family, demonstrating that architectural optimization can significantly boost RTS recognition accuracy. More recently, Wu et al. (2025) introduced a hybrid CNN-Vision Transformer encoder-decoder model for RTS identification using Sentinel-2 imagery. Despite these advances, methodological challenges persist, particularly the lack of systematic benchmarking to identify optimal model architectures. Foundational model exploration for RTS identification remains limited, lacking comprehensive comparisons to identify a universally robust model with high prediction accuracy at large scales. A stable base model is essential for supporting future research in areas like multi-source data fusion (e.g., integrating InSAR, NDVI, and optical features) (Ma et al., 2025; Yang et al., 2023b), long-term change detection (G. Yang et al., 2025), and mass wasting analysis (Maier et al., 2025).

In high-altitude areas like the Qinghai-Tibet Plateau, RTSs often present irregular shapes and complex morphological features (e.g., retreating headwalls and chaotic slump floors) driven by the differential thawing of ground ice and varying terrain conditions. From a remote sensing and computer vision perspective, unlike typical rainfall- or earthquake-triggered landslides, which typically appear as distinct high-contrast patches in forested regions and are visually easier to detect in remote sensing imagery (Guo et al., 2025; Zeng et al., 2025; Zhang et al., 2024a, 2024b), RTSs pose unique challenges. Their headwalls, thaw slump floors, and bulging slump toes exhibit varied spectral reflectance due to differing material compositions, resulting in blurred boundaries and morphodynamic complexity. Additionally, RTS exposed areas can resemble aeolian scarps, alpine bare lands, or other collapse features in color and texture, complicating automatic detection (Luo et al., 2022). To address these challenges, research has progressed along two key avenues: internal model enhancement and external data integration. Internally, attention mechanisms have been introduced to mimic the cognitive focus of expert interpreters. Recent studies on landslide mapping demonstrate their potential, with Zhang et al. (2024) improving feature capture using a multi-scale attention network, Li et al. (2024a, 2024b) optimizing CNN-Transformer integration with self-attention (SA) for multi-scale fusion, and Rauf et al. (2024) boosting accuracy through a super-resolution and bottleneck self-attention CNN

framework. Externally, multi-source remote sensing data fusion has emerged as a vital strategy to resolve RTS spectral and morphological ambiguities (Ghorbanzadeh et al., 2022; Wang et al., 2024; Zhang et al., 2023). Fusion approaches are categorized as homogeneous, addressing resolution trade-offs within similar data types (e.g., spatio-spectral fusion like pansharpening), and heterogeneous, integrating diverse data such as optical imagery, elevation models, and thermal data for complementary feature representation (Li et al., 2022). Currently, heterogeneous fusion strategies for RTS identification fall into two primary categories: pixel-level and feature-level fusion. Pixel-level fusion, which directly stacks multi-source data as input channels, has been widely adopted. For instance, Yang et al. (2023) improved model generalization by stacking optical imagery with NDVI and topographic data, while Jiao et al. (2023b) integrated InSAR deformation maps with optical bands to enhance detection accuracy. This approach excels at preserving raw spatial textures and geometric alignment, which are critical for tracing the complex boundaries of RTSs. However, it often fails to effectively model the non-linear interactions between highly heterogeneous data sources (e.g., optical reflectance vs. thermal emission) at the input stage. In contrast, feature-level fusion extracts abstract representations from different branches before integration. Recently Wu et al. (2025) employed a Transformer-based framework to fuse SAR and optical features at deeper network layers. While this method effectively captures high-level semantic correlations and reduces spectral ambiguity, it often suffers from the loss of fine-grained spatial details due to successive downsampling operations. Critically, most research focuses on either pixel-level or feature-level fusion independently, ignoring their combined potential (Wang et al., 2024; Yang et al., 2025b). This limitation is particularly detrimental for RTS identification in high-altitude environments. RTSs exhibit both subtle morphological details (requiring the spatial precision of pixel-level fusion) and complex thermodynamic mechanisms (requiring the semantic logic of feature-level fusion). Therefore, relying on a single fusion strategy is insufficient to resolve the conflict between detailed boundary delineation and robust feature discrimination. To address this, there is a pressing need to develop a dual-level heterogeneous fusion framework that systematically combines the spatial fidelity of pixel-level fusion with the semantic depth of feature-level fusion.

To bridge the existing research gaps, this study undertakes three interconnected investigations in the central Qinghai-Tibet Plateau: (i) creating a detailed RTS inventory and analyzing the spatio-temporal distribution patterns of RTSs using multi-source environmental factors; (ii) performing a comprehensive benchmark of twelve leading semantic segmentation models to determine the most effective architecture for RTS identification, with a focus on the impact of attention mechanism enhancements; and (iii) developing a novel FusionSA-Segformer framework that employs heterogeneous feature fusion through dual-level integration strategies (pixel-level and feature-level). Inspired by the RGB-Depth fusion paradigm in computer vision, this framework integrates optical spectral data with topographic and thermal environmental contexts to achieve a holistic feature representation. Ultimately, this research seeks to establish an accurate, interpretable, and practical deep learning framework for RTS identification.

2. Study area and data

2.1. Study area

Permafrost constitutes about 24% of the Northern Hemisphere's land surface, with significant presence in polar, alpine, and high-altitude regions (Fig.1a). The Qinghai-Tibet Plateau, with an average elevation above 4000 m, is a critical permafrost zone defined by a cold, arid climate, alpine meadows, and grasslands. Its active layer thickness varies from 0.9 to 3.2 m, and ground ice content averages between

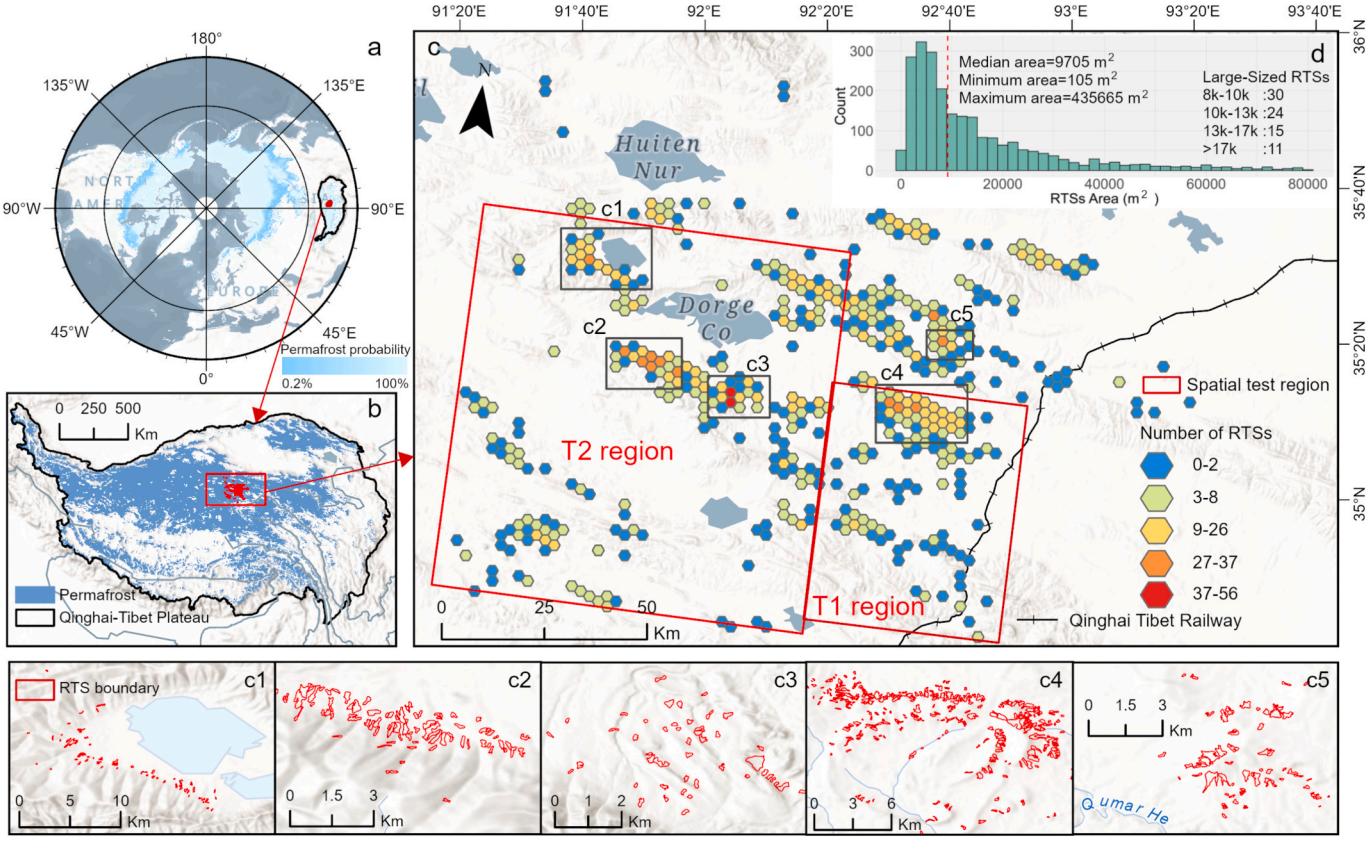

Fig. 1. Overview of the study area and distribution of RTSs in the central Qinghai-Tibet Plateau. (a) Location of the study area within permafrost zones; (b) location on the Qinghai-Tibet Plateau; (c) spatial distribution characteristics; (d) area distribution.

28.4% and 30.2% (Zou et al., 2024). The study focuses on the Beiluhe region in the central QTP, specifically within the Beiluhe Basin (34.40–35.4°N, 92.19–93.20°E) and the northwest of Kekexili National Nature Reserve (Fig.1b), covering a total area of approximately 34,094 km^2. This region was selected as the primary study site because it represents a critical "thermokarst hotspot" characterized by a high concentration of active RTSs and ice-rich permafrost that is highly sensitive to climate warming (Gao et al., 2026). Additionally, this is crossed by the Qinghai-Tibet Railway, a crucial transport link between Qinghai and Tibet (Zhao et al., 2023). The terrain features low hills and depressions shaped by alluvial and fluvial deposits from the Xiushui and Beiluhe Rivers, with wide, shallow riverbeds and slow-moving water. Spanning roughly 200 km north to south between the Fenghuo Mountains and the northern foothills of Hoh Xil, the region has an average elevation of 4500 m (Yin et al., 2017). It experiences a subarctic dry climate, with an annual mean temperature of −3.8 °C (ranging from −30 °C in January to 25 °C in July), precipitation of 300–377 mm, and evaporation exceeding 1000 mm annually. Data from the nearby Wudaoliang weather station indicate an average annual precipitation of 376.7 mm and a temperature of −4.1 °C (at 2 m above ground) between 2009 and 2021 (Luo et al., 2019; D. Yang et al., 2023). Approximately 70% of the region is underlain by continuous ice-rich permafrost, with 20% exhibiting ice contents above 50%. The permafrost layer extends 20–80 m in depth, while the active layer thickness ranges from 1.5 to 2.0 m (Luo et al., 2015). Since 2000, the area has undergone accelerated warming at a rate of 0.3 °C per decade, accompanied by a 15 mm per decade increase in precipitation, contributing to significant permafrost degradation (Shen et al., 2024).

Under the influence of global warming, permafrost degradation in the Beiluhe region has led to widespread thermokarst development, with RTSs presenting significant risks to ecosystem stability and infrastructure integrity. RTSs in this area are often distributed in clusters, predominantly on gentle slopes and fluvial terraces with high ground ice content (Fig.1c). The size of RTSs varies widely, with areas ranging from 105 to 435,665 m^2 and a median area of 9705 m^2 (Fig.1d). Among the larger RTSs, 11 exceed 17,000 m^2, and 15 fall within the 13,000 to 17,000 m^2 range, highlighting a notable prevalence of substantial RTS features in the region (Fig.1d). Field surveys and literature reviews (D. Yang et al., 2023) confirm that RTSs are commonly found on gentle slopes, where meltwater from ground ice flows across their surfaces, forming muddy deposits. These features exhibit distinct morphological traits, including vertical headwalls similar to the steep scarps of non-permafrost landslides, irregularly shaped slump floors resulting from small-scale collapses, and extensive muddy deposits created by meltwater flow. The active layer thickness in these areas typically ranges from 1 to 4 m, with frequent small-scale collapses initiating at headwall margins, contributing to their irregular shapes. Exposed ground ice melts under solar radiation, leading to water-rich scarring on active RTS surfaces. Notably, RTSs are more prevalent on shaded slopes and fluvial terraces, leaving prominent, recoverable scars that stand in stark contrast to the surrounding landscape. Critically, these degradation processes pose direct threats to infrastructure stability in the region. RTS activity has been observed encroaching on railway embankments (Fig.2a-d), developing along highway slopes (Fig.2e-f), damaging transmission line towers (Fig.2g), and causing large-scale land surface exposure (Fig.2h). These impacts underscore the urgent need for effective monitoring and mitigation strategies to protect vital infrastructure and manage environmental risks in this climate-sensitive area.

2.2. Data sources

2.2.1. PlanetScope optical imagery and RTS inventory

The primary dataset for identifying and mapping RTSs in this study consists of high-resolution optical imagery from PlanetScope, provided by Planet Labs through the Planet Education and Research Program. This imagery was chosen for its exceptional combination of high spatial resolution (3–5 m) and frequent revisit capability, which are crucial for accurately capturing the distinct morphological characteristics of thaw

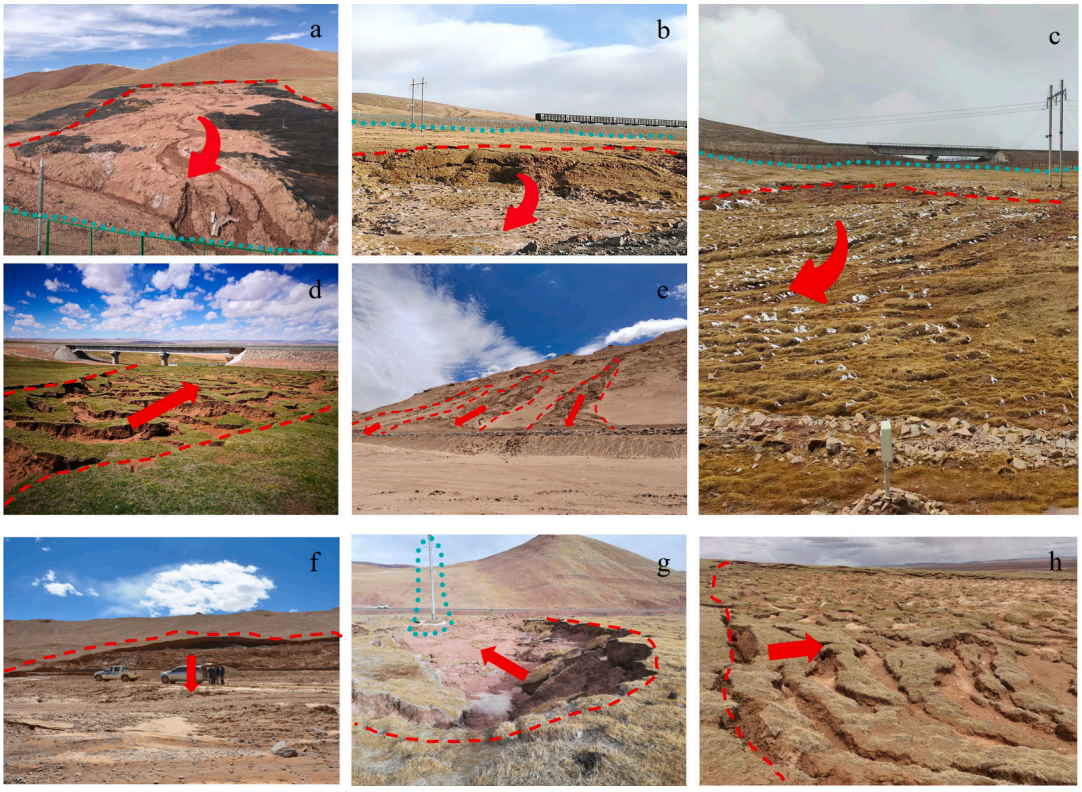

Fig. 2. Impacts of RTSs on engineering infrastructure and lands surface in the Central Qinghai-Tibet Plateau. (a-d) RTSs threatening railway infrastructure; (e-f) cluster of RTSs on highway slopes; (g) RTS damage to transmission line towers; (h) Large-scale RTS-induced exposed land surface patches.

slumps. A total of 190 PlanetScope PSScene images were acquired during the peak growing season (August to October) of 2024. The selection adhered to strict quality criteria, ensuring minimal cloud cover (≤5%) in each scene. This temporal focus was strategically designed to maximize spectral contrast between the exposed soil and disturbed vegetation of active slumps and the surrounding healthy tundra, thereby improving the accuracy of visual interpretation.

The development of a comprehensive RTS inventory began by integrating existing datasets, including thaw slump inventories for 2018–2020 by Luo et al. (2022) and for 2016–2022 by Xia et al. (2024), which provided a foundational baseline. Building on these prior efforts, the inventory was systematically updated and expanded through expert-led manual interpretation using ArcGIS Pro. This process using high-resolution remote sensing imagery from multiple sources. While PlanetScope (2024, 3 m) served as the primary input data source, sub-meter resolution imagery from Jilin-1 (0.75 m) and Beijing-2 (0.8 m) acquired in 2024 (accessed via Omap) was employed as auxiliary reference data. These higher-resolution images were specifically used to cross-validate RTS candidates and refine boundaries in areas where PlanetScope imagery exhibited spectral ambiguity or insufficient textural detail, ensuring high interpretation accuracy. The workflow involved a meticulous review of existing RTS polygons to verify their activity status and refine boundaries, as well as the identification and delineation of newly developed thaw slumps that had initiated or evolved since previous mappings (Fig.3). The resulting inventory includes 2429 individual RTS polygons (Fig.1) and serves as the critical ground-truth dataset for subsequent model training, validation, and accuracy assessment in this study.

2.2.2. *Multi-source satellite data and high resolution DEM*

To develop a robust multi-dimensional feature space for the deep learning model, this study utilized a synergistic combination of open-access satellite data and high-resolution topographic information. The primary data source for spectral index calculation was Sentinel-2 Multispectral Imagery, provided by the European Space Agency's Copernicus Program. Sentinel-2's spectral configuration, featuring red-edge and short-wave infrared bands at spatial resolutions of 10 and 20 m, is particularly well-suited for assessing vegetation health, soil moisture, and surface composition. The study employed the Level-2 A Bottom-of-Atmosphere reflectance product spanning 2019 to 2024, which includes scene-based atmospheric correction via the SEN2COR processor. To maintain data quality, a strict cloud and shadow masking procedure was applied using the QA60 quality assessment band. The high temporal revisit frequency of the Sentinel-2 constellation (approximately 5 days at mid-latitudes) proved essential for creating cloud-free seasonal and annual median composites, effectively reducing atmospheric noise and emphasizing the persistent spectral characteristics of dynamic RTS features.

To enhance thermal time-series analysis and ensure longitudinal consistency, thermal data from Landsat 8 and Landsat 9 were integrated into the feature dataset. The combined Landsat archive offers a continuous and stable record of land surface temperature (LST) at a 30 m resolution from 2019 to the present. The standard Level-2 Science Products, which include surface temperature estimates derived from the Thermal Infrared Sensor bands, were utilized for this purpose. Additionally, for regional-scale analysis of the thermal state of the permafrost environment, MODIS Land Surface Temperature Data (NASA LP DAAC Product: MODIS/061/MOD11A1) was acquired. Despite its coarser 1-km spatial resolution, the MODIS sensor's daily global coverage provided the critical temporal density needed to compute annual cumulative temperature indices fundamental to cryospheric studies. Topographic context, a primary factor influencing RTS initiation and distribution, was derived from a high-resolution 5 m DEM obtained from the local government. All auxiliary datasets were carefully co-registered to the PlanetScope base layer. A unified coordinate reference system was established, and all data were resampled to a consistent 3 m grid using a cubic convolution algorithm to maintain spatial fidelity and ensure precise pixel-level alignment across all input features for the deep learning model. The integration of these multi-temporal and multi-scale datasets resulted in a comprehensive suite of nine spectral and thermal indices, forming the foundation for subsequent analysis and modeling.

3. Methodology

This study introduces an advanced deep learning framework aimed at enhancing the recognition of RTSs through multi-source feature fusion and model optimization. The methodology, detailed in Fig.4, is structured into three sequential phases to systematically tackle the challenges associated with RTS identification in permafrost regions.

(i) In Phase I, the focus is on acquiring and processing diverse datasets to characterize the topographic, environmental, and dynamic conditions linked to RTS development. Multi-source features are extracted from a range of datasets, including high-resolution DEM, open-source resources such as the JRC Global Surface Water (Pekel et al., 2016) and Annual Vegetation Maps (Zhou et al., 2025), and satellite imagery from Sentinel-2, Landsat, and MODIS. These datasets are processed using computational algorithms to derive topographic, environmental, spectral, and thermal indices. Simultaneously, RTS ground truth boundaries are delineated using

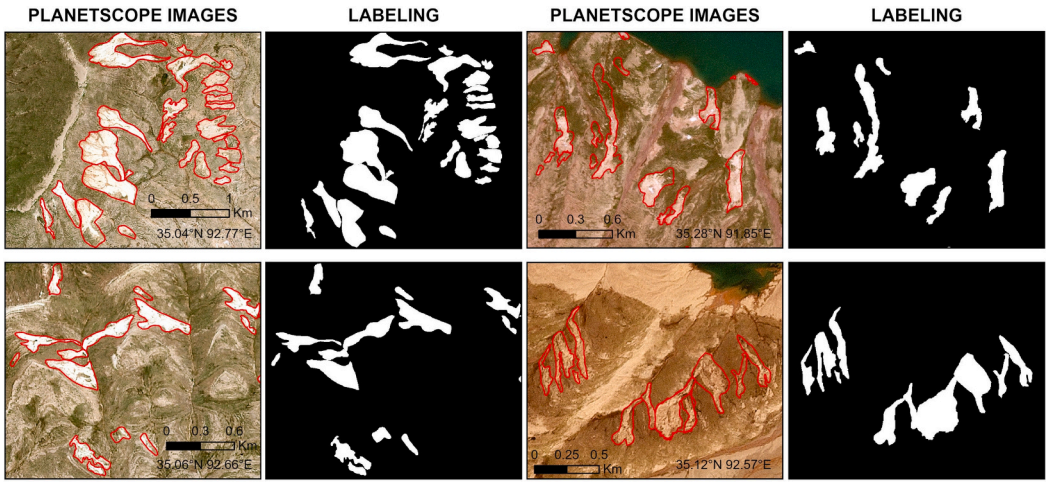

Fig. 3. PlanetScope imagery and labeling of RTSs in our study.

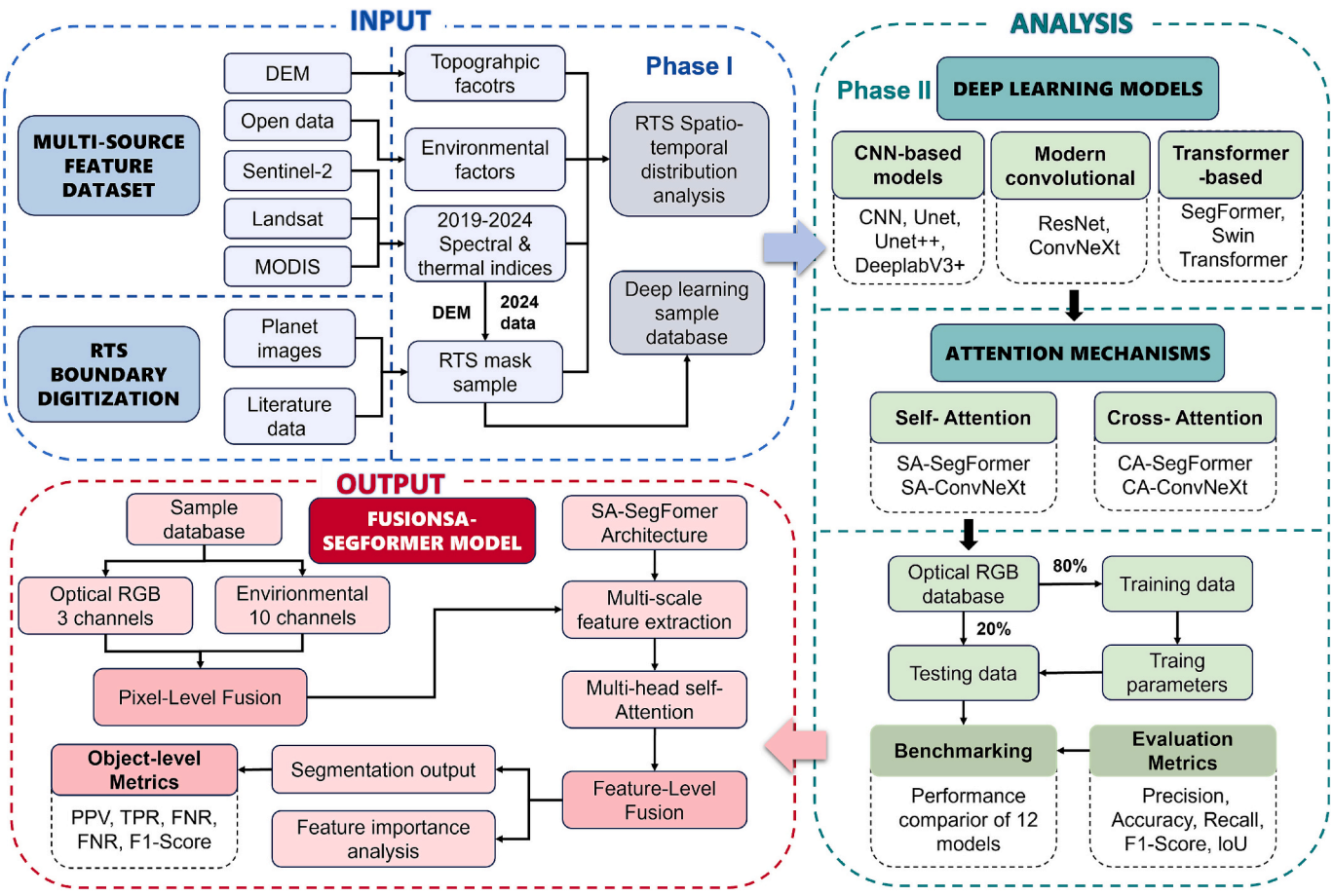

Fig. 4. Framework of the adopted methodology.

high-resolution PlanetScope imagery and existing literature data (Luo et al., 2022; Xia et al., 2024). These integrated datasets serve two key purposes: analyzing the spatiotemporal distribution characteristics of RTSs and constructing a comprehensive sample database for subsequent deep learning applications. (See Fig. 5.)

(ii) The second phase involves a systematic benchmarking of contemporary deep learning architectures for RTS identification. We evaluated twelve representative models spanning the evolution of segmentation networks, including established baselines widely used in RTS identification (e.g., Unet, DeepLabV3+) and state-of-the-art computer vision models (e.g., Swin Transformer, SegFormer, ConvNeXt). Particular emphasis was placed on optimizing SegFormer (Xie et al., 2021) and ConvNeXt (Liu et al., 2022), as both have demonstrated exceptional performance on standard benchmarks, yet their potential for permafrost feature extraction remains underexplored. This comparative analysis employed optical RGB imagery exclusively to establish baseline performance metrics, consistent with conventional RTS identification approaches. The benchmarking protocol assessed model efficacy across multiple performance indicators including accuracy, precision, recall, F1-score, and IoU. Model selection and hyperparameter tuning were conducted based on comprehensive validation set evaluation, while final performance assessment was performed on a held-out test set using object-level metrics derived from confusion matrix analysis, including True Positive Rate, False Discovery Rate, and False Negative Rate, to ensure robust evaluation of practical deployment capabilities.

(ii) The Phase III introduces the development of a novel FusionSA-SegFormer model, which implements heterogeneous feature fusion through dual-level integration strategies. The framework combines three channels of optical RGB imagery with ten spectral and thermal channels, employing both pixel-level and feature-level fusion techniques. Pixel-level fusion involves the direct concatenation of multi-source data into a unified input tensor, while feature-level fusion utilizes a multi-head self-attention mechanism within the SegFormer encoder to effectively extract and integrate multi-scale features. This approach ensures that complementary information from diverse data sources is optimally combined, resulting in superior RTS identification accuracy and providing interpretable insights into the contributions of multi-source features.

3.1. Phase I: Multi-source feature analysis and sample library development

3.1.1. Multi-source feature calculation

This phase concentrated on the acquisition and generation of a comprehensive set of features that represent the topographic, environmental, and thermal conditions contributing to the development of RTSs. These features were organized into two distinct groups: (i) topographic and environmental factors and (ii) dynamic spectral and thermal indices extracted from satellite time series. A detailed summary of all features, including their descriptions, sources, and purposes, is presented in Table 1, with the corresponding computational algorithms elaborated in Appendix A.

(i) **Topographic and environmental factors**

The primary data source for quantifying terrain morphology was a

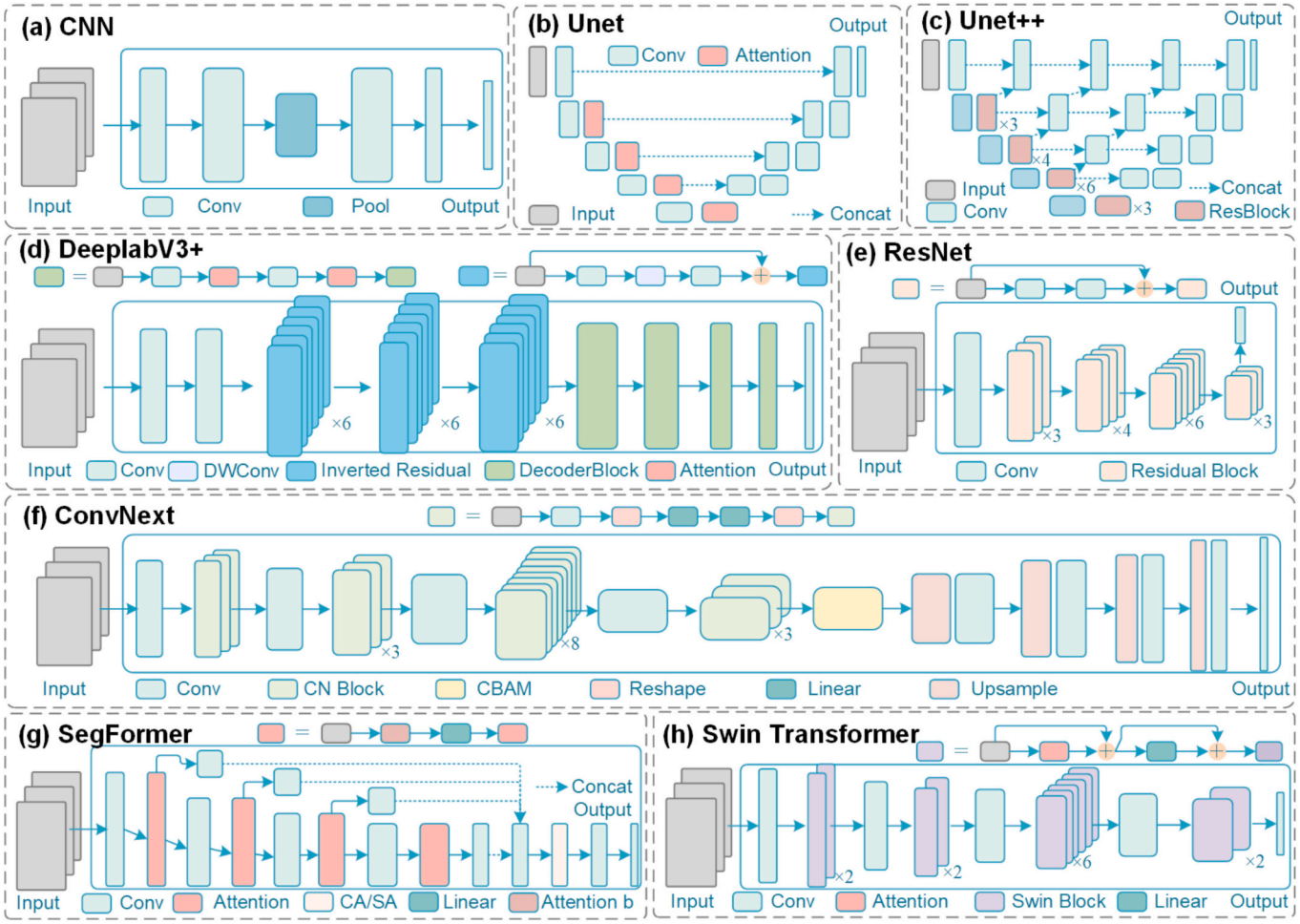

Fig. 5. Schematic diagrams of frameworks for different models in RTS identification. (a) CNN; (b) Unet; (c) Unet++; (d) DeeplabV3+; (e) ResNet; (f) ConvNeXt; (g) SegFormer; (h) Swin Transformer.

high-resolution (5 m) DEM, which was obtained from the local government. Using standard surface analysis tools, a suite of primary and secondary topographic metrics was derived from this DEM. These included slope and aspect, which are critical in influencing mechanical stability and solar insolation patterns; plan and profile curvature, which regulate subsurface water flow and sediment transport; and the Topographic Position Index (TPI), which quantitatively identifies local topographic concavities and convexities (e.g., valleys and ridges), representing key geomorphic settings for RTS initiation.

To achieve a synthesized and physiographically meaningful representation of the landscape, the SRTM Landform dataset (Theobald et al., 2015) was incorporated. This globally consistent classification, derived from the SRTM DEM, enables the assessment of geomorphic predisposition to RTS at broader scales. Hydrologically, the Distance to Water Body was calculated based on the JRC Global Surface Water Maximum Extent layer (1984–2021) (Pekel et al., 2016). This metric acts as a proxy for evaluating the potential of fluvial and thermal erosion, which are recognized as primary triggers for bank instability and slump initiation in ice-rich permafrost. Lastly, the Alpine Vegetation Type map for the Qinghai-Tibet Plateau was utilized to account for varying degrees of thermal insulation and root reinforcement provided by different plant communities, which play a significant role in buffering the underlying permafrost from atmospheric influences (Zhou et al., 2025).

(ii) **Dynamic spectral and thermal indices**

In addition to static landscape context, a suite of spectral and thermal indices was computed annually from 2019 to 2024 using the Google Earth Engine platform to capture dynamic surface processes associated with RTS activity. Spectral indices were derived from the Sentinel-2 MultiSpectral Instrument (MSI) Level-2 A surface reflectance collection, while thermal indices were calculated using Landsat 8/9 and MODIS data. A standardized preprocessing workflow was applied, involving cloud masking with quality assessment bands and the generation of annual median composites to ensure data quality and phenological relevance. The calculated indices address four fundamental aspects of land surface characteristics: i) Vegetation vigor, cover, and condition: The Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index provide robust measures of photosynthetic activity and green biomass. The Tasseled Cap Greenness (TCG) component offers a stable measure of vegetation density. A sharp decline in these values serves as a direct indicator of vegetation destruction caused by slump activity. ii) Surface moisture conditions: The Normalized Difference Moisture Index (NDMI) and Tasseled Cap Wetness (TCW) are highly sensitive to liquid water content in vegetation and soil. iii) Land disturbance and soil exposure: The Normalized Burn Ratio (NBR), repurposed for detecting surface disturbance, and Tasseled Cap Brightness (TCB), which indicates overall soil exposure, are effective for mapping the bare, disturbed ground characteristic of fresh slump scars. iv) Surface thermal regime: Land Surface Temperature (LST) derived from Landsat provides a direct measure of the thermal forcing driving ground ice thaw. The Net Degree-Days (NDD), calculated as the annual net result of Thawing and Freezing Degree-Days from MODIS data, quantifies the net surface energy balance.

Table 1
Multi-source feature description for RTS identification.

Category	Feature Name	Description	Data Source & Resolution	Temporal Scope
Topographic and environmental	DEM	Digital Elevation Model, providing the foundational altitude data.	Government-provided dataset (5 m)	Static
	Slope	Rate of maximum change in elevation	Derived from DEM in ArcGIS Pro	
	Aspect	Directional orientation of the slope		
	Curvature	Convergence/divergence of flow across a surface		
	TPI	Identifying local ridges and valleys.		
	Landforms	Synthetic classification of ecological landforms	SRTM Landform (ERGo) dataset (90 m)	
	Distance to Water Body	Euclidean distance to historical water body	JRC Global Surface Water (v1.4)	Static (1984–2022)
	Vegetation Types	Alpine vegetation classification	Annual Vegetation Maps of QTP (500 m)	2022
Spectral	EVI	Enhanced Vegetation Index, measuring canopy structure and vigor.	Sentinel-2 MSI (10 m)	2019–2024 (Analysis), 2024 (Model Input)
	NDVI	Normalized Difference Vegetation Index, measuring vegetation greenness.	Sentinel-2 MSI (10 m)	
	NBR	Normalized Burn Ratio, sensitive to surface disturbance and soil exposure.	Sentinel-2 MSI (20 m)	
	NDMI	Normalized Difference Moisture Index, measuring vegetation/soil moisture content.	Sentinel-2 MSI (20 m)	
	TCB	Tasseled Cap Brightness, indicating soil exposure and surface brightness.	Sentinel-2 MSI (10 m)	
	TCG	Tasseled Cap Greenness, indicating vegetation density and health.	Sentinel-2 MSI (10 m)	
	TCW	Tasseled Cap Wetness, indicating surface and canopy moisture.	Sentinel-2 MSI (10 m)	
Thermal	LST	Land Surface Temperature, the primary driver of ground thaw.	Landsat 8/9 TIRS (30 m)	
	NDD	Net Degree-Days, net annual surface energy balance (TDD - FDD).	MODIS MOD11A1 (1000 m)	

The analysis of thaw slump spatiotemporal patterns was supported by integrating static geomorphological parameters (e.g., DEM, Slope, Aspect) with the 2019–2024 time series of spectral and thermal indices. Subsequently, the semantic segmentation model utilized the DEM and the 2024 indices as its input features for further analysis and modeling.

3.1.2. Sample library development

Building upon the high-fidelity RTS inventory established in Section 2.2.1, this phase focused on constructing a machine-learning-ready sample library. The core process involved integrating binary RTS labels with a multi-dimensional stack of input features designed to provide the model with comprehensive contextual information. The input to the semantic segmentation model comprised a synergistic combination of three data types, co-registered to a common 3 m spatial resolution: (i) High-resolution optical imagery (PlanetScope): This provided primary visual evidence for identifying characteristic texture, color, and shape features of RTS. (ii) Topographic feature: The high-resolution DEM overlay offered essential topographic context, capturing the terrain characteristics critical to RTS initiation and distribution. (iii) Spectral and thermal features: A suite of spectral indices (EVI, NDVI, NBR, NDMI, TCB, TCG, TCW) and thermal indices (LST, NDD) for the year 2024 was incorporated. These features provided enhanced information on vegetation vigor, soil exposure, surface moisture, and land surface thermal regime, complementing the visual and topographic data.

Through the systematic integration of these datasets, a comprehensive sample library was generated, containing binary classification masks where each pixel was assigned to either RTS (Class 1) or background (Class 0). The complete pixel stack, encompassing all input features, along with the corresponding label masks, was partitioned into smaller image chips. A tile size of 256*256 pixels was selected as an optimal trade-off to balance computational efficiency with adequate receptive fields for capturing RTS morphological context. This dimension was used with a corresponding stride of 256 pixels, resulting in non-overlapping tiles that collectively covered the entire study area. This process yielded a total of 1195 training chips, which were systematically partitioned into independent subsets for model development and evaluation. Specifically, the dataset was divided into training (70%, 837 chips), validation (15%, 179 chips), and testing (15%, 179 chips) sets to ensure robust training, tuning, and assessment of the deep learning model in subsequent phases.

3.2. Phase II: Semantic segmentation model selection and benchmarking

The second phase involved a systematic benchmarking of contemporary deep learning architectures for the identification of RTS. We evaluated twelve state-of-the-art semantic segmentation models that have been previously employed in geoscientific applications. Particular emphasis was placed on optimizing SegFormer and ConvNeXt architectures, both recognized as cutting-edge transformer and convolutional networks, respectively, due to their exceptional performance in remote sensing segmentation tasks. This comparative analysis initially utilized optical RGB imagery exclusively to establish baseline performance metrics, mirroring conventional approaches in current RTS identification methodologies.

3.2.1 Semantic segmentation models.

(i) **CNN-based models**. This group commenced with a custom-designed CNN model, characterized by a streamlined encoder-decoder architecture with a lightweight framework. The encoder consists of two convolutional layers with ReLU activation, followed by max-pooling for downsampling, while the decoder employs a transposed convolution for upsampling and a final convolutional layer to produce a single-channel output. Additionally, we implemented the Unet (Ronneberger et al., 2015), renowned for its symmetric encoder-decoder structure and skip connections that enable precise localization by preserving spatial details across scales. Its enhanced variant, Unet++ (Zhou et al., 2018) wwas also included, featuring nested, dense skip pathways to mitigate the semantic gap between encoder and decoder features. Furthermore, DeeplabV3+ (Chen et al., 2018) was incorporated, representing architectures that leverage atrous spatial pyramid pooling to effectively capture multi-scale contextual information through parallel

convolutional operations with varying dilation rates.

(ii) **Modern convolutional models**. This group leverages advanced convolutional backbones within established encoder-decoder frameworks to enhance feature extraction capabilities. The ResNet (Targ et al., 2016) introduced residual connections to effectively mitigate the vanishing gradient problem, enabling the training of significantly deeper and more powerful feature encoders. ConvNeXt (Liu et al., 2022), a modern pure-CNN architecture, systematically modernizes a standard ResNet by incorporating design principles inspired by Vision Transformers, achieving state-of-the-art performance with a purely convolutional approach while maintaining efficiency and robustness in feature representation.

(iii) **Transformer-based models**. To evaluate the performance of self-attention mechanisms, we incorporated SegFormer (Xie et al., 2021), which employs a hierarchical encoder and a lightweight all-MLP (Multi-Layer Perceptron) decoder to efficiently combine local attention with global context, demonstrating strong potential in remote sensing applications. The Swin Transformerr (Liu et al., 2021) was also selected, introducing a hierarchical feature representation and computing self-attention within shifted windows to improve computational efficiency while capturing long-range dependencies in spatial data.

3.2.1. Optimization with attention mechanisms

Building upon the comprehensive model selection in Phase II, SegFormer and ConvNeXt were identified as the primary candidates for architectural enhancement due to their established efficacy in various complex visual recognition tasks beyond remote sensing. SegFormer represents the cutting-edge in transformer-based segmentation, renowned for its ability to model long-range dependencies through hierarchical encoder structures, which are particularly beneficial for capturing spatially extensive patterns. Meanwhile, ConvNeXt exemplifies modern convolutional architecture design, achieving remarkable performance through systematic refinements of classical CNN paradigms, balancing efficiency and robustness in feature extraction. These two models collectively embody the most promising architectural directions in contemporary computer vision, making them ideal foundations for investigating advanced attention mechanisms in permafrost disturbance mapping, specifically for RTS identification.

To further augment their capabilities for RTS identification, we integrated two distinct attention paradigms into both architectures: self-Attention and cross-attention (CA). This systematic approach yielded four optimized variants: SA-SegFormer, CA-SegFormer, SA-ConvNeXt, and CA-ConvNeXt. The SA mechanism enhances intra-feature contextualization by computing spatial dependencies across all positions within individual feature maps, enabling improved identification of geographically dispersed but morphologically connected thaw slump components. This is particularly useful for recognizing fragmented or irregular RTS features across large spatial extents. Conversely, the CA mechanism facilitates inter-scale feature alignment, allowing high-resolution spatial details to dynamically interact with semantically rich low-resolution contexts, thereby refining boundary delineation in topographically complex terrain where RTS features often exhibit subtle transitions and intricate edge characteristics.

3.2.2. Evaluation metrics

To comprehensively evaluate the semantic segmentation models, we adopted a dual-framework evaluation strategy. During the training and validation phases, pixel-wise metrics such as Accuracy, Precision, Recall, F1-Score, and IoU were used to monitor performance. These metrics guided model selection and served as crucial indicators to monitor and identify potential overfitting by offering detailed insights into segmentation quality at the pixel level. For the final evaluation on the test set, we incorporated object-level analysis using metrics derived from a confusion matrix, treating each RTS site as a distinct entity. The confusion matrix was constructed based on the following criteria (Y. Yang et al., 2023): a True Positive (TP) was recorded when a predicted RTS segment showed any spatial overlap with a ground truth polygon, acknowledging the value of partial detections in locating RTSs; a False Positive (FP) was assigned to a predicted segment with no overlap with any ground truth polygon; and a False Negative (FN) was attributed to a ground truth RTS polygon that remained completely undetected. Using these classifications, we calculated key object-level metrics: Precision (Positive Predictive Value, PPV), which evaluates the reliability of positive predictions; Recall (True Positive Rate, TPR), which measures the proportion of actual RTSs correctly identified and reflects detection sensitivity; and F1-Score, the harmonic mean of Precision and Recall, providing a balanced measure of model performance. Additionally, the False Discovery Rate (FDR) was computed to quantify the proportion of predicted RTSs that were incorrect, indicating the frequency of false alarms among positive predictions. The False Negative Rate (FNR) was used to assess the proportion of actual RTSs missed by the model, highlighting gaps in detection coverage.

3.3. Phase III: Multi-source feature heterogeneous fusion for enhanced identification

In the third phase, we introduce the novel FusionSA-SegFormer model, designed to integrate multi-source features for improved RTS detection and feature importance analysis. This phase tackles the complex, heterogeneous nature of RTS characteristics in remote sensing imagery by combining diverse environmental data to boost identification accuracy. Inspired by the RGB-Depth fusion paradigm in computer vision—which enhances scene understanding by augmenting visual appearance (RGB) with geometric structure (Depth)—our approach adopts a parallel strategy. We treat the heterogeneous environmental data as the auxiliary context layer to complement the primary optical imagery. Consequently, the framework employs a dual-level fusion strategy, incorporating both pixel-level and feature-level fusion. It uses three channels of optical RGB imagery alongside ten spectral and thermal channels as input. Additionally, we integrate interpretability mechanisms to quantify the contribution of each feature, providing valuable insights into the environmental drivers of RTS dynamics.

3.3.1. Dual-level fusion strategy

To effectively harness multi-source data for RTS segmentation, we developed a dual-level fusion strategy that combines pixel-level and feature-level approaches, ensuring optimal synergy of heterogeneous inputs from optical RGB imagery (3 channels) and spectral/thermal features (10 channels) (Fig. 6a).

At the pixel level, fusion occurs during input preparation by concatenating optical RGB imagery with ten raster-based environmental features along the channel dimension, forming a unified 13-channel input tensor. This early integration captures diverse RTS characteristics, spanning terrain, vegetation, and thermal conditions, from the onset of the network's encoding process. Similar to RGB-Depth fusion, where appearance and geometric data are merged at the input stage, our pixel-level fusion creates a rich multi-modal representation. Input data undergo preprocessing via normalization and spatial alignment for consistency across heterogeneous types. Optical channels are normalized to a 0–1 range and standardized using established mean and standard deviation values for natural imagery, while raster features are scaled via a min-max approach. Missing data are handled with placeholder arrays of zero values to maintain input structure. The resulting tensor, formatted to fixed spatial dimensions (256*256 pixels), serves as the foundation for further processing.

The core of our strategy lies in feature-level fusion, implemented within the tailored SA-SegFormer architecture to refine and integrate intermediate representations. Beyond simple pixel concatenation, this approach enables dynamic feature interaction and reweighting. After processing the 13-channel input tensor, multi-level feature maps are extracted through a Transformer-based hierarchical encoder, projected

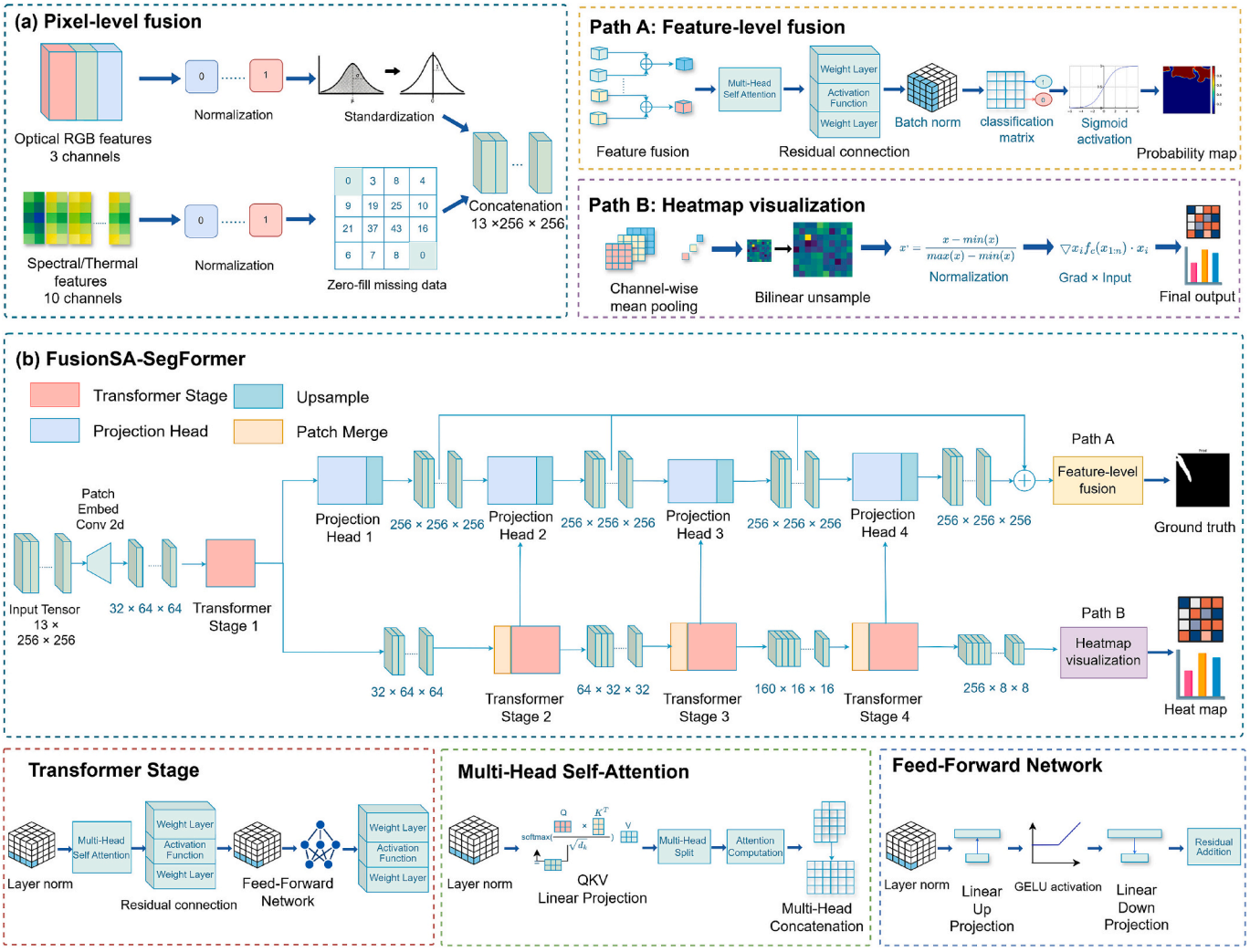

Fig. 6. Schematic illustrations of data fusion and model architecture for RTS Identification, (a) Pixel-level fusion; (b) Architecture of FusionSA-SegFormer model.

to a consistent channel dimension (default 256), and upscaled to match the input resolution (256*256 pixels). These maps are combined via summation across hierarchy levels to create a unified representation. A multi-head self-attention mechanism (configured with 4 attention heads) within the encoder models global spatial dependencies, ensuring effective integration of complementary information from diverse sources. This advanced fusion enhances the model's focus on contextually relevant RTS regions by capturing long-range dependencies across multi-source features, ultimately improving segmentation accuracy.

3.3.2. *FusionSA-SegFormer model*

Building on the dual-level fusion strategy, we developed the FusionSA-SegFormer model by adapting the SA-SegFormer architecture, which draws from the SegFormer framework with a custom multi-head self-attention module (Fig. 6b). Utilizing a Transformer-based encoder (Xie et al., 2021), the model is configured to process the 13-channel pre-fused input tensor through its hierarchical structure. Feature maps extracted at multiple levels are aligned to a uniform dimension (256 channels) using projection heads, upscaled to input resolution via bilinear interpolation, and summed to form a fused representation. This representation is enhanced through a custom multi-head self-attention operation (4 heads) for contextual integration, followed by batch normalization and a residual connection to ensure training stability. The refined features pass through a lightweight classification head with convolutional layers, normalization, and ReLU activation, producing binary segmentation maps for RTS identification.

To improve robustness, data augmentation techniques such as random resized cropping (scale 0.5 to 1.0), horizontal/vertical flips, and 90-degree rotations are applied during training. Training parameters align with optimizations from Phase II, including a batch size of 8 and a learning rate schedule ranging from 3e-4 to a peak of 1e-3. To address class imbalance and optimize boundary delineation, we employed a composite objective function defined as:

$$\zeta = \lambda_{BCE}\zeta_{BCE} + \lambda_{Dice}\zeta_{Dice} + \lambda_{Focal}\zeta_{Focal} \quad (1)$$

where the component weights were empirically set to $\lambda_{BCE} = 0.5$, $\lambda_{Dice} = 1.0$, and $\lambda_{Focal} = 1.0$. Specifically, the Binary Cross-Entropy term (ζ_{BCE}) utilized a positive-class weight of $\omega_{pos} = 5.0$ to penalize false negatives; the Focal Loss term (ζ_{Focal}) employed a focusing parameter of $\gamma = 2.0$and a class balance factor of $\alpha = 0.8$; and the IoU-related term was implemented using a soft Dice loss (ζ_{Dice}). The model is trained for 200 epochs using an adaptive optimization algorithm with weight decay regularization. A key innovation of this phase is the feature importance analysis, which quantifies the contribution of individual input features to RTS segmentation. Using a gradient-based approach (Grad×Input), saliency maps are computed during validation by scaling input gradients for each of the 13 channels by their input values. Aggregated per-channel importance scores for optical and environmental features (e.g., DEM, EVI, NDVI, LST) are recorded and visualized as bar charts to highlight relative contributions. Spatially resolved saliency maps, upscaled to

original dimensions, are displayed alongside ground truth, predicted outputs, and probability heatmaps for detailed spatial analysis. This interpretability framework validates the effectiveness of our heterogeneous fusion framework and identifies critical feature combinations driving accurate RTS identification.

4. Results

4.1. Phase I: Environmental and spectral characteristics of RTSs

Distinct topographic and environmental factors influencing the distribution of RTSs across the study area were revealed through multi-source feature analysis (Fig. 7). RTSs exhibited a strong altitudinal constraint, with the majority located within a narrow elevation range of 4700 to 4800 m above sea level (Fig.7a). Slope analysis indicated a preference for gentle to moderate inclines, with a dominant frequency peak at approximately 5° (mostly concentrated between 3° and 7°) (Fig.7b), reflecting terrain conducive to thaw-driven mass wasting. Surface curvature assessments, including plan curvature (affecting water flow convergence and divergence) and profile curvature (influencing flow acceleration and deceleration), showed a near-normal distribution with peak frequencies centered near zero. Most RTSs occurred within a curvature range of −0.5 to 0.5 for both metrics (Fig.7c, d), suggesting a prevalence on relatively planar slope segments with minimal topographic variability. The Topographic Position Index, distinguishing ridges (positive values) from valleys (negative values), highlighted a concentrated RTS distribution skewed toward negative values, with a dominant frequency peak around −3 and the majority falling between −10 and 3 (Fig.7e). This pattern indicates a strong association with valley depressions (negative TPI) and mid-slope transitions (near-zero TPI), while effectively excluding prominent ridges (high positive values). Hydrological proximity to historical water bodies emerged as a key control, with over 50% of RTSs located within 1000 m of a water body and a dominant concentration confined within 2000 m. While the frequency shows a sharp decline beyond this distance, a notable distribution tail extends up to approximately 6000 m (Fig.7f). Aspect analysis revealed a notable directional asymmetry, with a significant proportion of RTSs oriented toward the Northwest (NW), North (N), and Northeast (NE) sectors (Fig.7g). Landform classification underscored the dominance of valley morphology, with narrow valleys hosting 46.1% of mapped RTSs and broader valley settings accounting for 27.9%, collectively comprising nearly three-quarters of inventoried

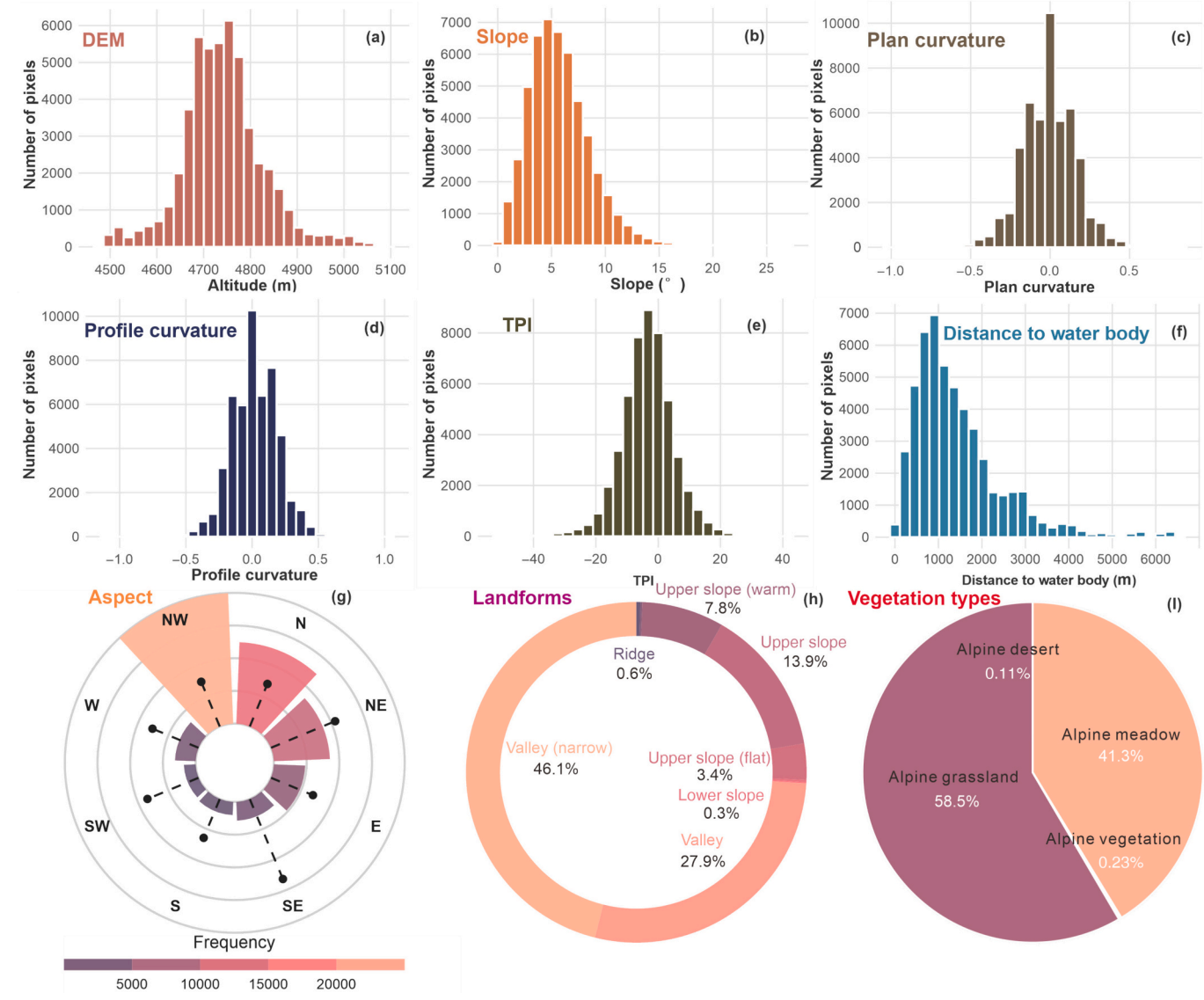

Fig. 7. Geomorphological and environmental factors of RTSs. Histograms display the frequency distribution of RTS polygons with respect to (a) elevation; (b) slope; (c) plan curvature; (d) profile curvature; (e) TPI, (f) distance to water body; (g) aspect; (h) landforms; (i) vegetation types.

RTSs. Upper slope environments and warm upper slopes contributed 13.9% and 7.8%, respectively, indicating limited occurrence in elevated or thermally distinct positions (Fig.7h). Vegetation cover analysis displayed a bimodal distribution, dominated by alpine grassland (58.5% of RTS locations) and alpine meadow (41.3%), highlighting a strong linkage between RTS distribution and vegetation patterns influenced by elevation and moisture availability (Fig.7i).

Distinct signatures of surface change were uncovered through an interannual analysis of spectral and thermal indices within RTS footprints from 2019 to 2024 (Fig. 8). To validate the statistical significance of these temporal variations, we employed the Kruskal-Wallis test for inter-annual differences and linear trend analysis for directional consistency. The results confirm that the observed trends across all indices are statistically significant (***, $p < 0.001$). Vegetation indices consistently declined over the study period. The median EVI dropped from 0.122 in 2019 to a low of 0.095 in 2023, with a slight recovery to 0.103 in 2024 (Trend Slope: −0.006/yr, $p < 0.001$; Fig.8a), while its interquartile range shrank from 0.101 in 2019 to 0.051 in 2023, indicating reduced spatial heterogeneity in vegetation cover within RTS-affected areas. Similarly, the median NDVI decreased from 0.190 in 2019 to 0.160 in 2024, reaching a minimum of 0.135 in 2023 (Slope: −0.012/yr, $p < 0.001$; Fig.8b), confirming progressive degradation of vegetation health. Indices sensitive to surface disturbance and moisture content also showed significant shifts. The NBR fell from 0.123 in 2019 to 0.092 in 2024, with a notable low of 0.044 in 2022 (Slope: −0.007/yr, $p < 0.001$; Fig.8c), while the NDMI declined from 0.023 in 2019 to 0.004 in 2024, dipping below zero to −0.02 in 2022 (Slope: −0.004/yr, $p < 0.001$; Fig.8d), pointing to a persistent drying trend linked to permafrost thaw and hydrological changes. The TCB index, associated with soil exposure, rose sharply from 0.605 in 2019 to a peak of 0.767 in 2022, before stabilizing at 0.662 in 2024 (0.016/yr, $p < 0.001$; Fig.8e). Conversely, the TCG index, a proxy for vegetation cover, dropped from

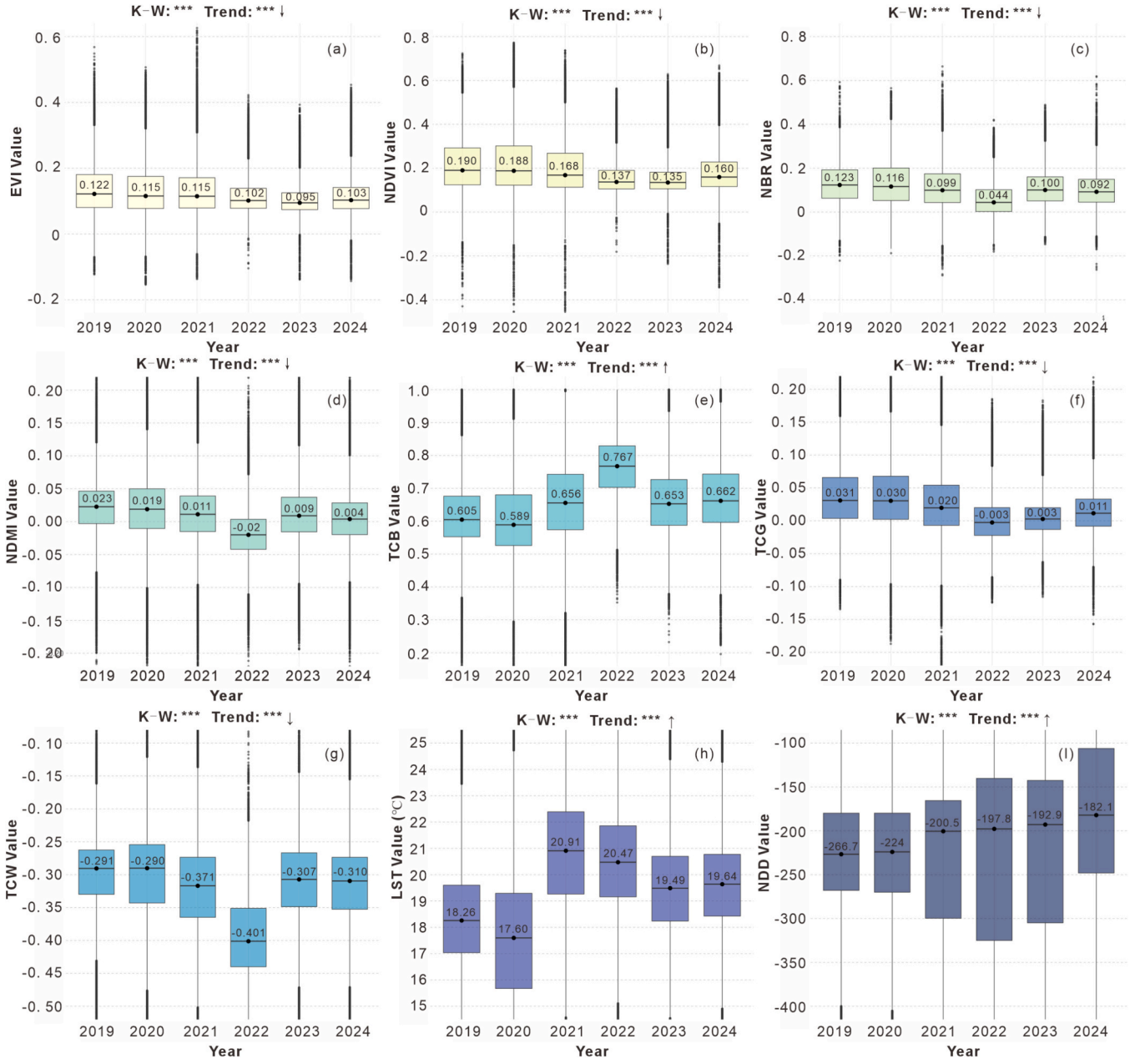

Fig. 8. Temporal trends of spectral and thermal indices within RTS areas. Boxplots show the annual distribution (2019–2024) of (a) EVI; (b) NDVI; (c) NBR; (d) NDMI; (e) TCB; (f) TCG; (g) TCW; (h) LST; (i) NDD.

0.031 in 2019 to a low of −0.003 in 2022, reflecting ongoing vegetation loss (Slope: −0.006/yr, $p < 0.001$; Fig.8f). The TCW index mirrored this decline, falling from −0.291 in 2019 to a minimum of −0.401 in 2022 (Slope: −0.005/yr, $p < 0.001$; Fig.8g), consistent with the drying pattern observed in NDMI. Thermal indices revealed a clear warming signal, with mean summer Land Surface Temperature rising from 18.26 °C in 2019 to a peak of 20.91 °C in 2021, and remaining elevated at 19.64 °C in 2024 (Slope: +0.36 °C/yr, $p < 0.001$; Fig.8h). The NDD, indicative of net annual surface energy balance, shifted toward less negative values, increasing from −266.7 in 2019 to −182.1 in 2024 (Slope: 1.783/yr, $p < 0.001$; Fig.8i), signaling a reduction in net cooling and a transition to a warmer surface energy regime, consistent with accelerated permafrost degradation in RTS-affected landscapes.

4.2. *Phase II: Comparative analysis of semantic segmentation models*

The training progression of the evaluated semantic segmentation models exhibited rapid convergence, with distinct performance hierarchies emerging across key metrics. All models exhibited rapid convergence in accuracy (Fig.9a), with most achieving stable performance early in the training process. While most architectures maintained accuracy above 0.9, the Swin Transformer displayed fluctuations within the 0.85 to 0.95 range. Precision metrics (Fig.9b) stabilized around epoch 10 across all models, with minimal oscillations thereafter. ResNet, Unet++, SA-SegFormer, and CA-SegFormer stood out as top performers, consistently maintaining superior precision throughout training. Recall metrics (Fig.9c) revealed significant disparities: the conventional CNN architecture performed poorly, fluctuating around 0.05, whereas SA-SegFormer demonstrated exceptional stability and high performance,

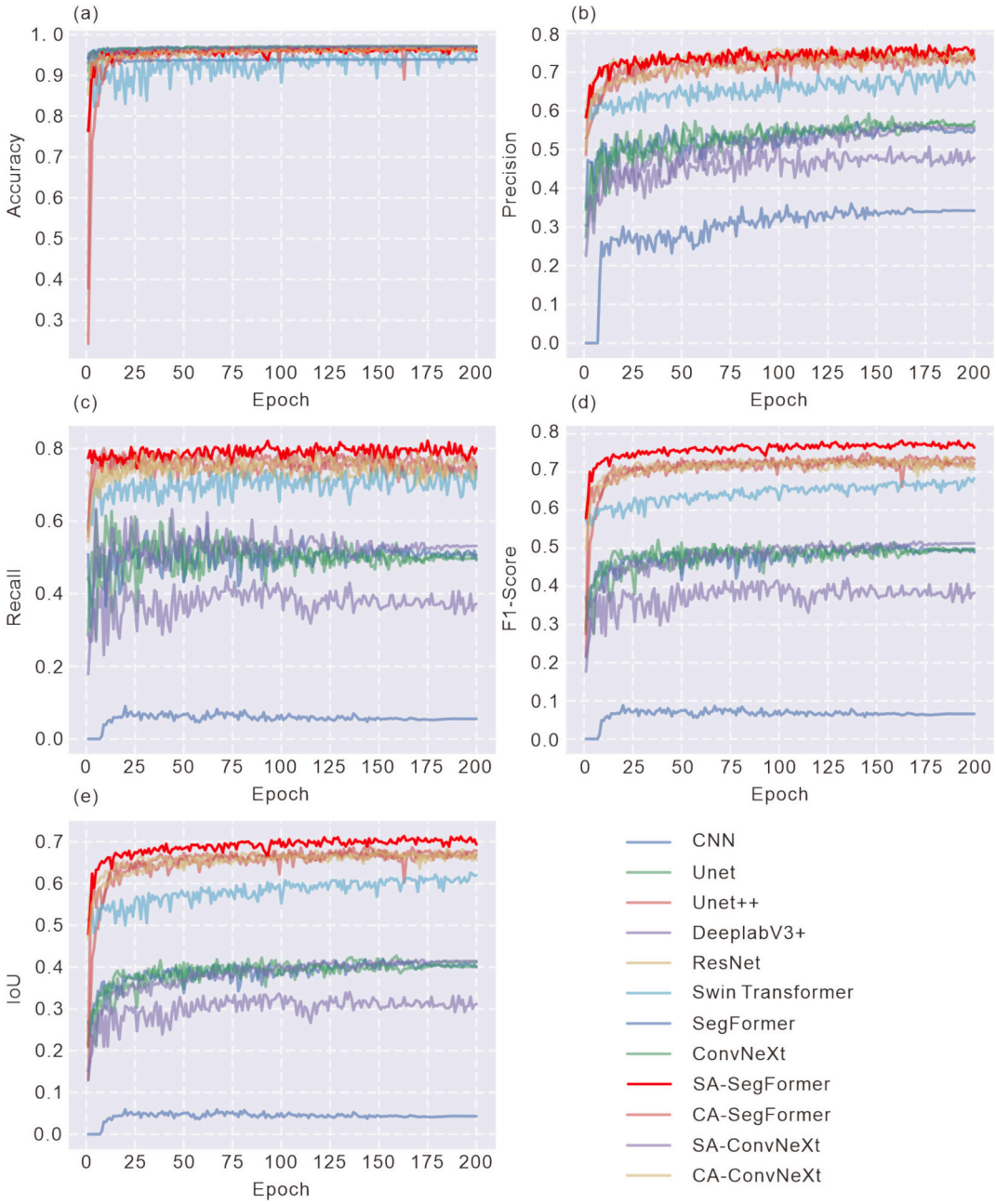

Fig. 9. Comparative analysis of evaluation metrics across semantic segmentation models. (a) Accuracy comparison; (b) precision comparison; (c) recall comparison; (d) F1-Score comparison; (e) IoU comparison.

achieving a recall of approximately 0.8. Other models generally scored below 0.8 in recall, with DeeplabV3+ and SegFormer exhibiting notable variability. Comprehensive metrics such as F1-Score (Fig.9d) and IoU) (Fig.9e) further delineated performance hierarchies. A top-performing group emerged, including CA-SegFormer, ResNet, SA-SegFormer, Unet++, CA-ConvNeXt, and Swin Transformer. An intermediate group consisted of ConvNeXt, SegFormer, Unet, SA-ConvNeXt, and DeeplabV3+, while the traditional CNN model consistently underperformed across all metrics.

A multi-dimensional performance assessment via radar charts reveals distinct capability profiles across the five key evaluation metrics. In Fig. 10a, the comparison of CNN, Unet, and Unet++ reveals a clear gradation. Unet++ exhibits a well-balanced pentagonal profile, excelling particularly in Recall and F1-Score. Unet demonstrates moderate performance with a weaker Recall, while CNN shows severely limited results across all dimensions, especially in Recall (0.057) and IoU (0.043). Fig. 10b compares DeeplabV3+, Swin Transformer, and ResNet, showcasing varied performance characteristics. ResNet displays a robust and balanced profile with high scores in Precision (0.740), Accuracy (0.964), and Recall (0.734). Swin Transformer presents a moderately sized pentagon, strong in Accuracy but weaker in Recall and IoU. DeeplabV3+ has the smallest coverage area in this group, reflecting lower overall performance, particularly in Recall (0.375) and IoU (0.302). Fig. 10c highlights significant performance differences among SegFormer, CA-SegFormer, and SA-SegFormer. SA-SegFormer stands out with an exceptionally large and balanced pentagonal shape, achieving the highest overall performance score (0.786, calculated as the arithmetic mean of the five evaluated metrics) among all models, with outstanding Recall (0.790) and F1-Score (0.759). CA-SegFormer shows strong competitive performance with a slightly smaller but similarly balanced profile, while SegFormer exhibits notably weaker performance. Lastly, Fig. 10d compares ConvNeXt, SA-ConvNeXt, and CA-ConvNeXt, revealing distinct variations. CA-ConvNeXt demonstrates a balanced profile with strengths in Accuracy and Precision, achieving an overall score of 0.757. In contrast, SA-ConvNeXt and ConvNeXt display smaller pentagonal shapes with comparable strengths and weaknesses, scoring 0.572 and 0.575, respectively.

Quantitative assessment on the test set, further delineates the object-level detection capabilities of each model (Table 2). Among the top-performing models, SA-SegFormer showcased exceptional detection capabilities with the highest F1-score of 0.817 and achieving a precision of 76.92%. CA-SegFormer and ResNet also delivered strong results, with F1-Scores of 0.7814 and 0.7583. In contrast, the lowest-performing models included CNN (F1-Score, 0.1356), DeeplabV3+ (F1-Score, 0.4248), and SegFormer (F1-Score, 0.5015). These models struggled with both detection completeness and reliability, with CNN notably

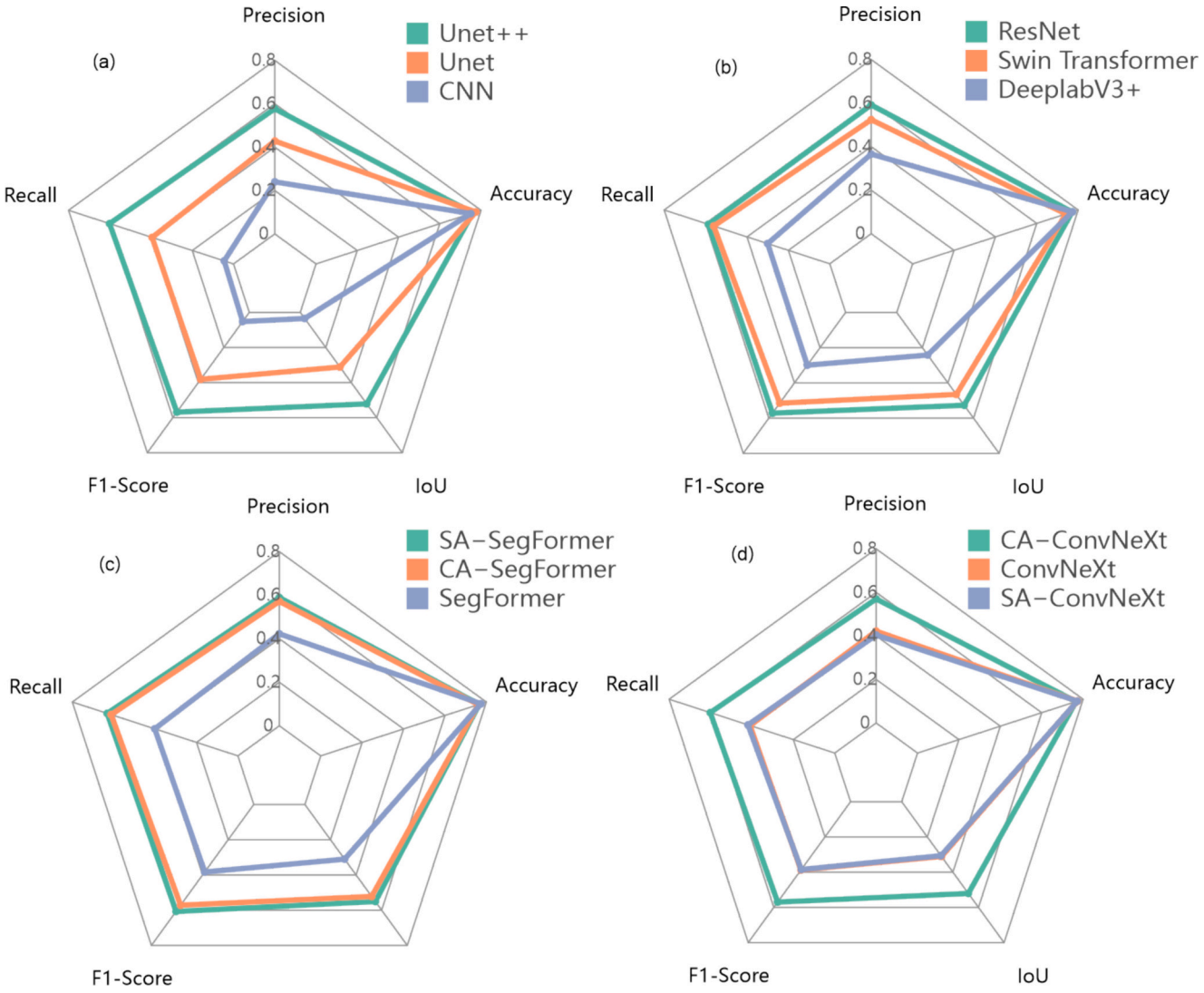

Fig. 10. Radar chart comparison of different models. (a) CNN, Unet, Unet++; (b) DeeplabV3+, Swin Transformer, ResNet; (c) SegFormer, SA-SegFormer, CA-SegFormer, (d) ConvNeXt, SA-ConvNeXt, CA-ConvNeXt.

Table 2
Object-level evaluation of baseline models on test set.

Model	TP	FP	FN	TPR/Recall $\frac{TP}{TP+FN}$	FDR $\frac{FP}{TP+FN}$	FNR $\frac{FN}{TP+FN}$	PPV/Precision $\frac{TP}{TP+FP}$	F1-Score $\frac{2\times(PPV\times TPR)}{(PPV+TPR)}$
CNN	35	185	263	0.117	0.840	0.882	0.159	0.135
Unet	172	198	126	0.577	0.535	0.422	0.464	0.514
Unet++	235	125	63	0.788	0.347	0.211	0.652	0.714
DeeplabV3+	142	228	156	0.476	0.616	0.523	0.383	0.424
Swin	225	145	73	0.755	0.391	0.245	0.608	0.673
ResNet	248	108	50	0.832	0.303	0.167	0.696	0.758
Segformer	165	195	133	0.553	0.541	0.446	0.458	0.501
SA-Segformer	260	78	38	0.872	0.230	0.127	0.769	0.817
CA-Segformer	252	95	46	0.845	0.273	0.154	0.726	0.781
Convnext	182	188	116	0.610	0.508	0.389	0.491	0.544
SA-Convnext	195	172	103	0.654	0.468	0.345	0.531	0.586
CA-Convnext	238	132	60	0.798	0.356	0.201	0.643	0.712

failing to detect 263 out of 298 true RTS features and generating 185 false positives. A consistent trend across all models was the higher occurrence of false positives compared to false negatives. Even the best-performing SA-SegFormer recorded 78 false positives against only 38 false negatives, suggesting that while models are generally effective at identifying potential RTS regions, they often struggle to accurately distinguish true RTS features from spectrally similar terrain.

Qualitative assessments of segmentation results for representative RTS examples highlight significant disparities in model efficacy under varying conditions (Fig. 11). In the first case, a large RTS area is analyzed, with clear disparities in model performance (Fig.11a). The CNN model fails entirely to identify any meaningful RTS regions, exposing the shortcomings of basic convolutional architectures for complex geological feature detection. Most advanced models, however, successfully capture the core morphological characteristics of the RTS. Nevertheless, several models, including Unet++, DeeplabV3+, ResNet, SegFormer, ConvNeXt, and SA-ConvNeXt, exhibit over-segmentation tendencies, misclassifying adjacent non-RTS areas as part of the target region. Swin Transformer and CA-ConvNeXt, on the other hand, struggle with blurred and uncertain boundaries around the RTS perimeter. CA-SegFormer and SA-SegFormer emerge as the top performers in this scenario, achieving the most accurate segmentation of the RTS morphology. Their attention-enhanced mechanisms enable precise boundary delineation while ensuring comprehensive coverage of the actual RTS area, demonstrating the value of specialized attention modules in geological feature extraction.

The second case presents a more complex scenario involving multiple interconnected RTS areas, further challenging the models under intricate terrain conditions (Fig.11b). Once again, the CNN model shows severely limited performance, reinforcing its inadequacy for advanced RTS identification tasks. Additionally, segmentation results from Unet, DeeplabV3+, Swin Transformer, and SA-ConvNeXt exhibited distinct false positive patterns, where non-RTS areas were incorrectly classified as RTS, potentially due to over-sensitivity or misinterpretation of topographic and textural features resembling RTS signatures. In this challenging multi-RTS scenario, ConvNeXt and SA-SegFormer stand out as the most effective models, achieving competent boundary segmentation that closely aligns with the true RTS extents. Alongside Unet++, SegFormer, and CA-SegFormer, SA-SegFormer particularly excels in maintaining boundary precision across multiple connected regions.

4.3. Phase III: Multi-source heterogeneous feature fusion analysis

In Phase III of our study, we conducted a comprehensive assessment of the proposed FusionSA-SegFormer model, which integrates multi-source feature fusion to enhance segmentation performance, compared to the baseline SA-SegFormer model. The results reveal consistent and significant improvements across all evaluation metrics (Fig.12). FusionSA-SegFormer achieved substantial performance gains, with precision rising from 0.757 to 0.822 (8.6% improvement), recall increasing from 0.747 to 0.829 (10.9% improvement), and F1-score improving from 0.734 to 0.796 (8.5% improvement). Most notably, the IoU metric exhibited a remarkable enhancement, advancing from 0.679 to 0.739 (8.8% improvement), underscoring the model's superior boundary delineation capabilities.

Object-level evaluation on the test set further substantiated these advancements, particularly in reducing classification errors (Table 3). FusionSA-SegFormer achieved an F1-score of 0.843, marking a 3.1% absolute improvement over the best optical baseline SA-SegFormer (0.818), and a 6.2% improvement over CA-SegFormer (0.781). Compared to the baseline SA-SegFormer, the multi-source fusion approach reduced false positives by 33.3% and false negatives by 21.1%, while boosting true positives by 3.1%. The performance improvement over the original SegFormer was even more pronounced, with a 62.4% increase in true positives and a 73.3% reduction in false positives. These consistent gains across both pixel-level and object-level metrics affirm the efficacy of integrating multi-source remote sensing features into the segmentation framework. The notable improvements in recall and IoU highlight the fusion strategy's ability to capture complete RTS extents while preserving precise boundary definitions. This performance elevation positions FusionSA-SegFormer as a state-of-the-art approach for RTS segmentation, emphasizing the critical role of comprehensive feature representation in addressing complex geological analysis tasks.

To validate the independence of the input features and ensure the reliability of the subsequent importance analysis, we first examined the inter-feature dependencies using Pearson correlation coefficients (Fig. 13). The correlation matrix reveals that the selected environmental, spectral, and thermal features maintain a high degree of orthogonality. Most feature pairs exhibit weak correlations, with the maximum observed coefficient reaching only 0.45 (between TCW and TCG). This value is well below the threshold typically indicative of multicollinearity (0.7), confirming that the chosen heterogeneous features provide complementary rather than redundant information to the deep learning model. Following this independence validation, a distinct hierarchy in feature contributions to RTS identification is revealed through the importance analysis of the FusionSA-SegFormer model (Fig. 14). The ranking demonstrates that RGB spectral channels dominate the top positions in importance, with the blue band exhibiting the highest mean importance (0.176 ± 0.032), followed by the red (0.122 ± 0.034) and green bands (0.102 ± 0.019). Thermal and environmental indices form a middle tier of significance, including Land Surface Temperature (0.096 ± 0.018), Enhanced Vegetation Index (0.084 ± 0.015), and Net Degree-Days (0.083 ± 0.019). Conventional vegetation indices such as TCB, NBR, and NDMI display moderate importance, with values ranging from 0.072 to 0.076, while NDVI (0.049 ± 0.011) shows a relatively lower contribution. Topographic and other spectral features, including DEM, TCG, and TCW, exhibit the lowest importance scores, all below 0.030.

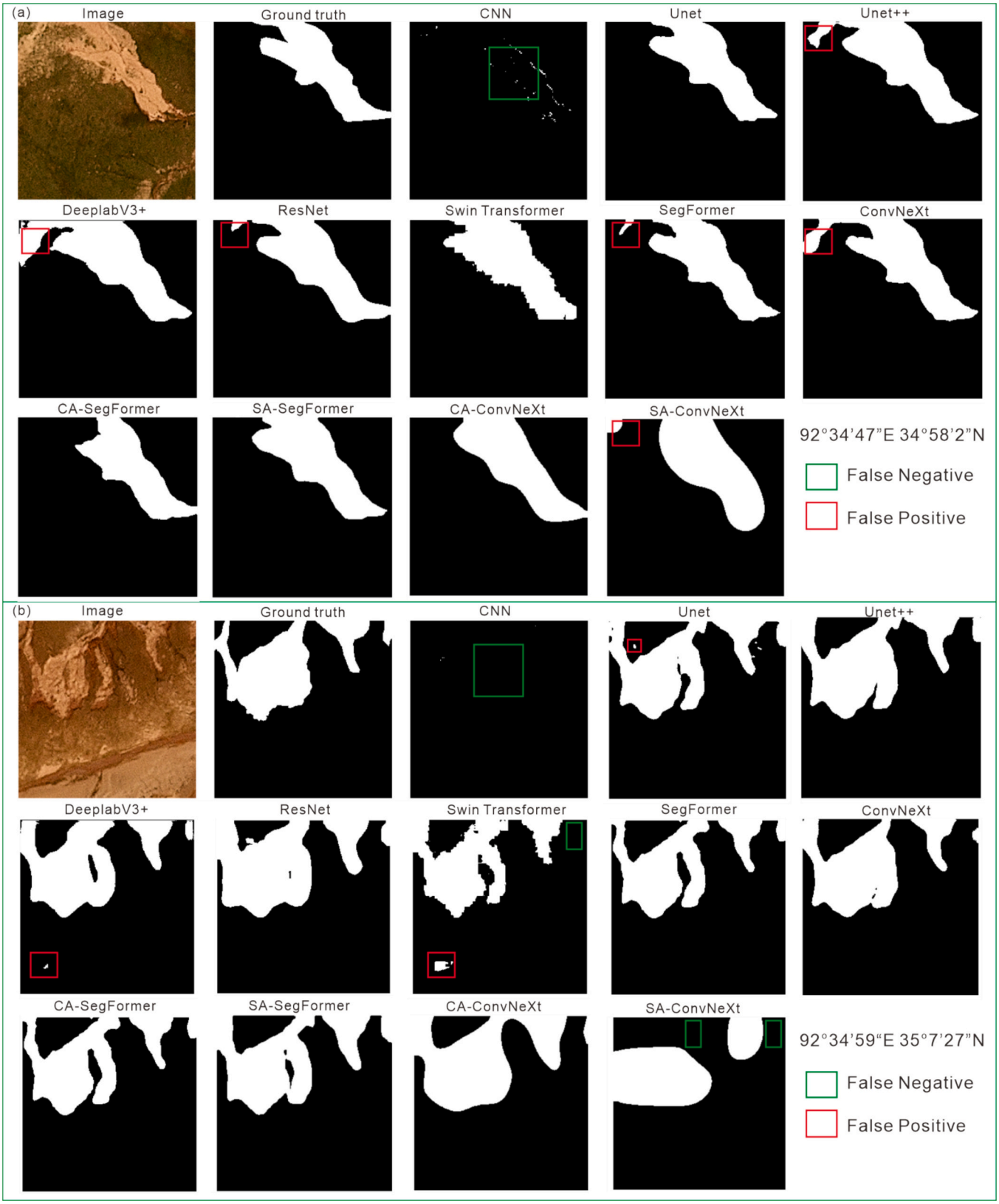

Fig. 11. Comparative visualization of RTS segmentation results across different models. (a) A substantial RTS area; (b) several interconnected RTS areas.

Following the evaluation of overall model accuracy and feature importance rankings, we further analyzed the performance of the FusionSA-SegFormer model in detecting RTSs across a variety of challenging scenarios by comparing it directly with the baseline SA-SegFormer (Fig. 15). In scenarios with clear boundaries and high contrast (Fig. 15a-c), both models demonstrated effective identification capabilities. FusionSA-SegFormer accurately delineates the location and morphology, comparable to SA-SegFormer, verifying that the fusion

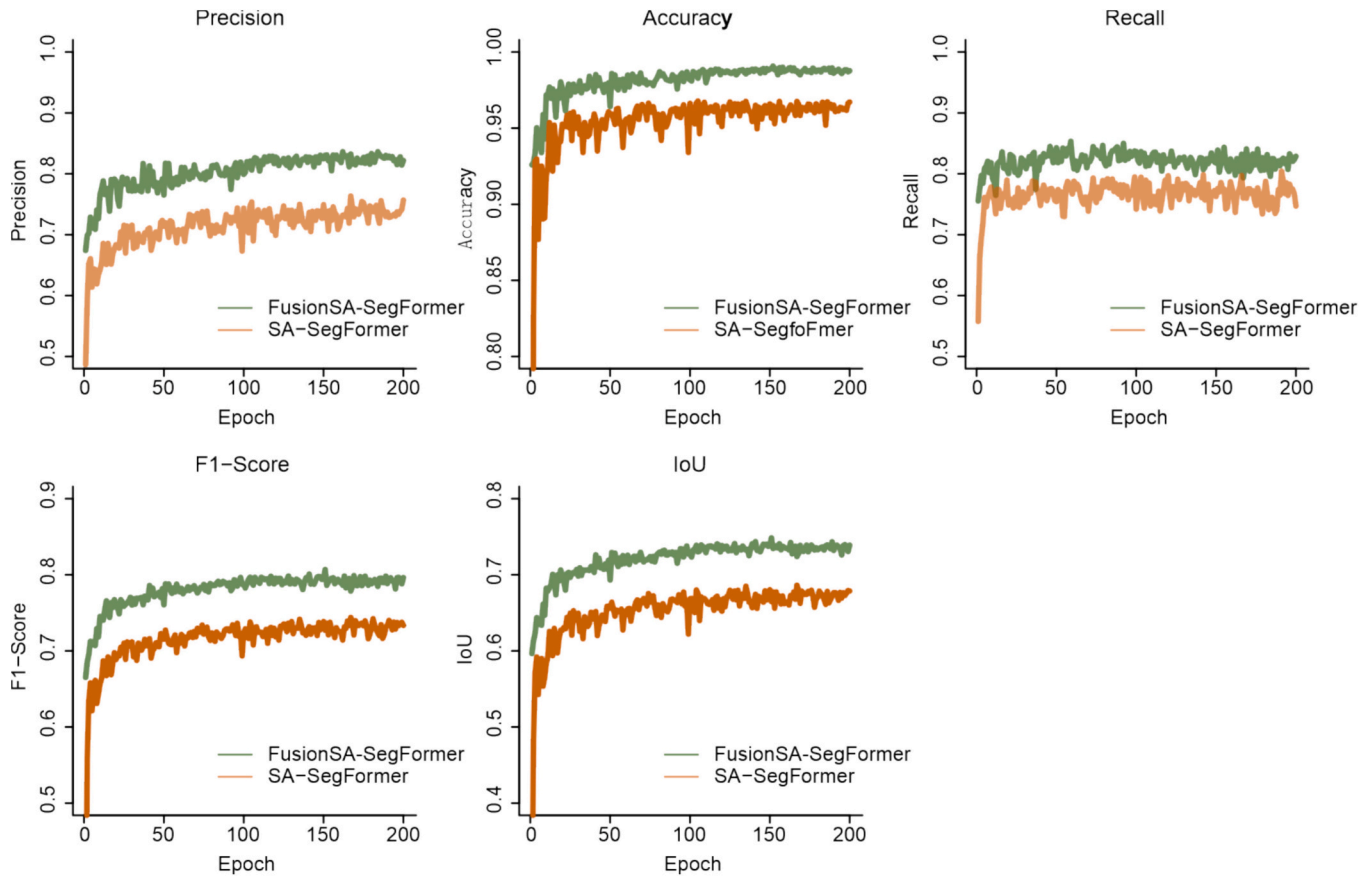

Fig. 12. Performance comparison of FusionSA-SegFormer model and SA-SegFormer across multiple metrics.

Table 3
Comparative object-level evaluation of FusionSA-SegFormer on test set.

Model	TP	FP	FN	TPR/Recall	FDR	FNR	PPV/Precision	F1-Score
FusionSA-SegFormer	268	52	30	0.899	0.162	0.101	0.838	0.843
SA-Segformer	260	78	38	0.872	0.230	0.127	0.769	0.817
CA-Segformer	252	95	46	0.845	0.273	0.154	0.726	0.781
Segformer	165	195	133	0.553	0.541	0.446	0.458	0.501

architecture retains the high precision of the baseline in simple conditions. However, the superiority of the heterogeneous fusion strategy becomes evident in complex environments. In Fig. 15d, characterized by extensive exposed areas and micro-topographic variations, the SA-SegFormer failed to detect a smaller, subtle RTS feature (omission error) due to spectral confusion. In contrast, FusionSA-SegFormer successfully captured all RTS targets, illustrating its adaptability to heterogeneous landscapes and resilience against environmental noise by leveraging thermal and topographic cues. Furthermore, Figs. 15e-f showcase performance in detecting complex RTS clusters. While SA-SegFormer could generally localize the RTS positions, it struggled with precise boundary delineation, often producing smoothed or fragmented contours. FusionSA-SegFormer, conversely, accurately recognized both the locations and the fine-grained boundary details, effectively separating adjacent features.

Spatially explicit visualizations and importance rankings provide an in-depth exploration of feature contributions within the FusionSA-SegFormer model (Fig.16). Fig. 16a showcases the precise identification of a single RTS at a specified location, with feature saliency maps revealing prominent hotspots of model attention within the RTS region. The corresponding feature importance ranking for this location identifies optical features as the primary contributors, with RGB—B, RGB-G, and LST ranking highest. This indicates that spectral and thermal characteristics play a central role in RTS identification under relatively uniform terrain conditions. In contrast, Fig. 16b examines RTS identification in a complex terrain environment, where diverse land cover and topographic variability introduce significant challenges. The feature saliency maps illustrate how critical information from various features enables the model to boost confidence in RTS identification by effectively filtering out environmental noise. Notably, distinct differences in feature attention are observed between the two RTS structures in the image. For instance, features such as NDD, NDMI, and TCW exhibit limited attention to the RTS on the right side, while DEM, EVI, NBR, and TCB effectively capture the distribution of both RTS features. This complementary attention across features highlights their collective role in achieving comprehensive RTS recognition in specific imagery. The feature importance ranking for this location further reveals a shift in dominant contributors, with RGB—B, RGB-R, and EVI emerging as the most influential, suggesting that the relative importance of features varies with environmental context.

It is important to distinguish between global feature importance ranking, which aggregates influence scores across the entire dataset to establish a general hierarchy (Fig. 14), and local feature contributions, which describe the spatially varying attention of the model for specific

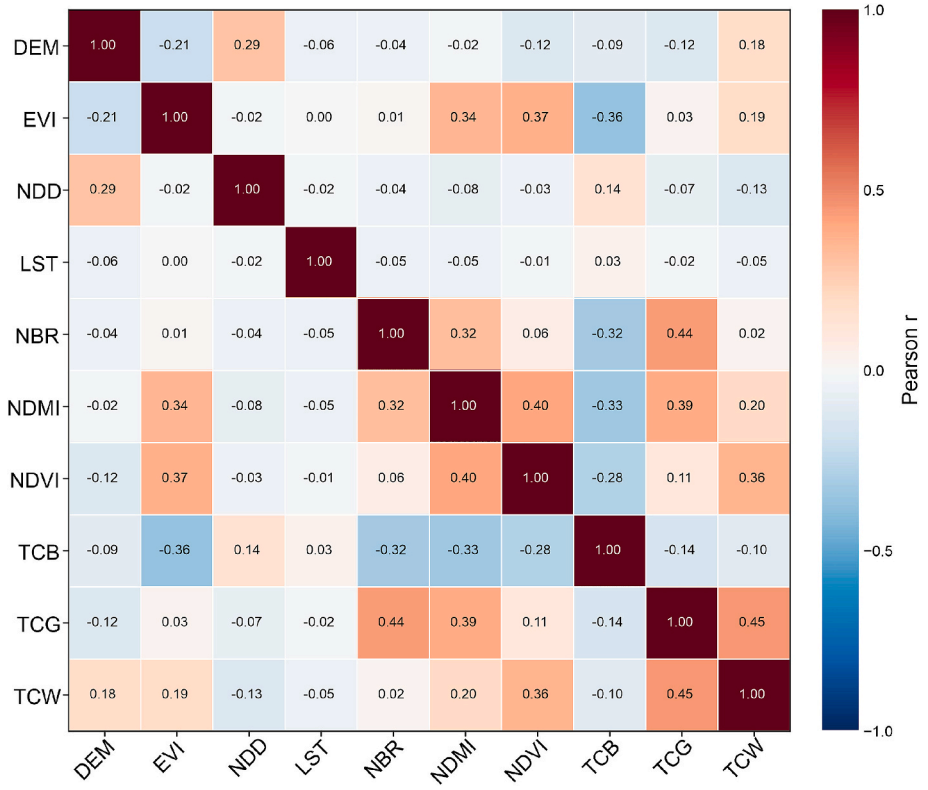

Fig. 13. Pearson correlation matrix of the ten features.

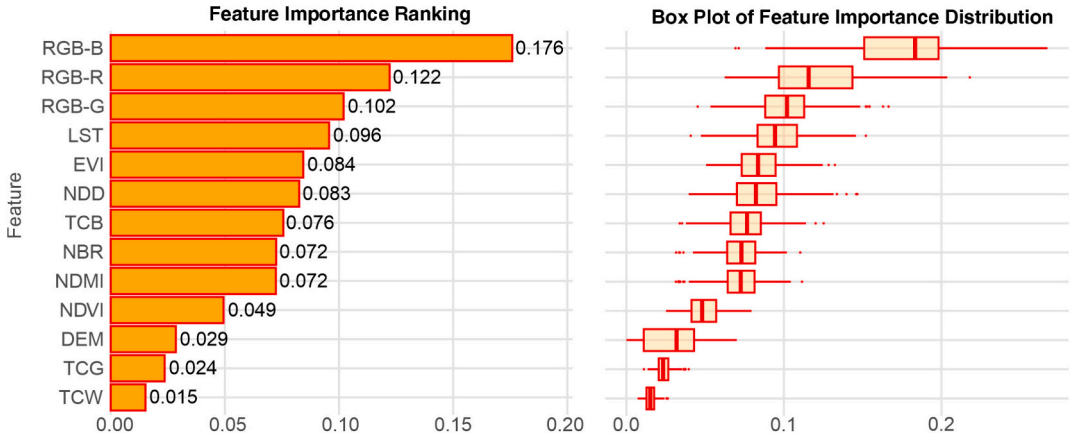

Fig. 14. Ranking of feature importance for RTS identification.

RTS instances (Fig. 16). The analysis in Fig. 16 underscores the nuanced differences in feature contributions and the value of multi-source feature fusion for RTS identification. While optical features consistently dominate importance rankings, the variability in attention and ranking across features, such as the prominence of spectral indices like EVI in complex terrains versus thermal indices like LST in simpler settings, demonstrates that integrating a diverse set of features enhances the robustness and adaptability of the FusionSA-SegFormer model.

5. Discussion

The identification and monitoring of RTSs in high-altitude permafrost regions have long been hindered by the intricate interplay of thermokarst dynamics, heterogeneous landscape responses, and the inherent limitations of conventional remote sensing approaches (Dai et al., 2025). Traditional methodologies often fail to capture the subtle yet progressive nature of RTS development, particularly during the early stages when timely intervention could yield the most impactful mitigation outcomes. Moreover, prior studies have typically focused on either the geomorphic predispositions or the thermal and spectral signatures of RTSs in isolation, without employing an integrated analytical framework that bridges predisposing factors with dynamic process indicators. This research gap is particularly pressing in the context of accelerating climate change, where a holistic understanding of the complete trajectory of RTS evolution—from initial predisposition through active development to eventual stabilization—is critical for advancing predictive modeling and enhancing risk assessment (Tao et al., 2025). To overcome these challenges, our study introduces a novel multi-stage methodology that systematically integrates multi-source feature analysis with state-of-the-art deep learning techniques.

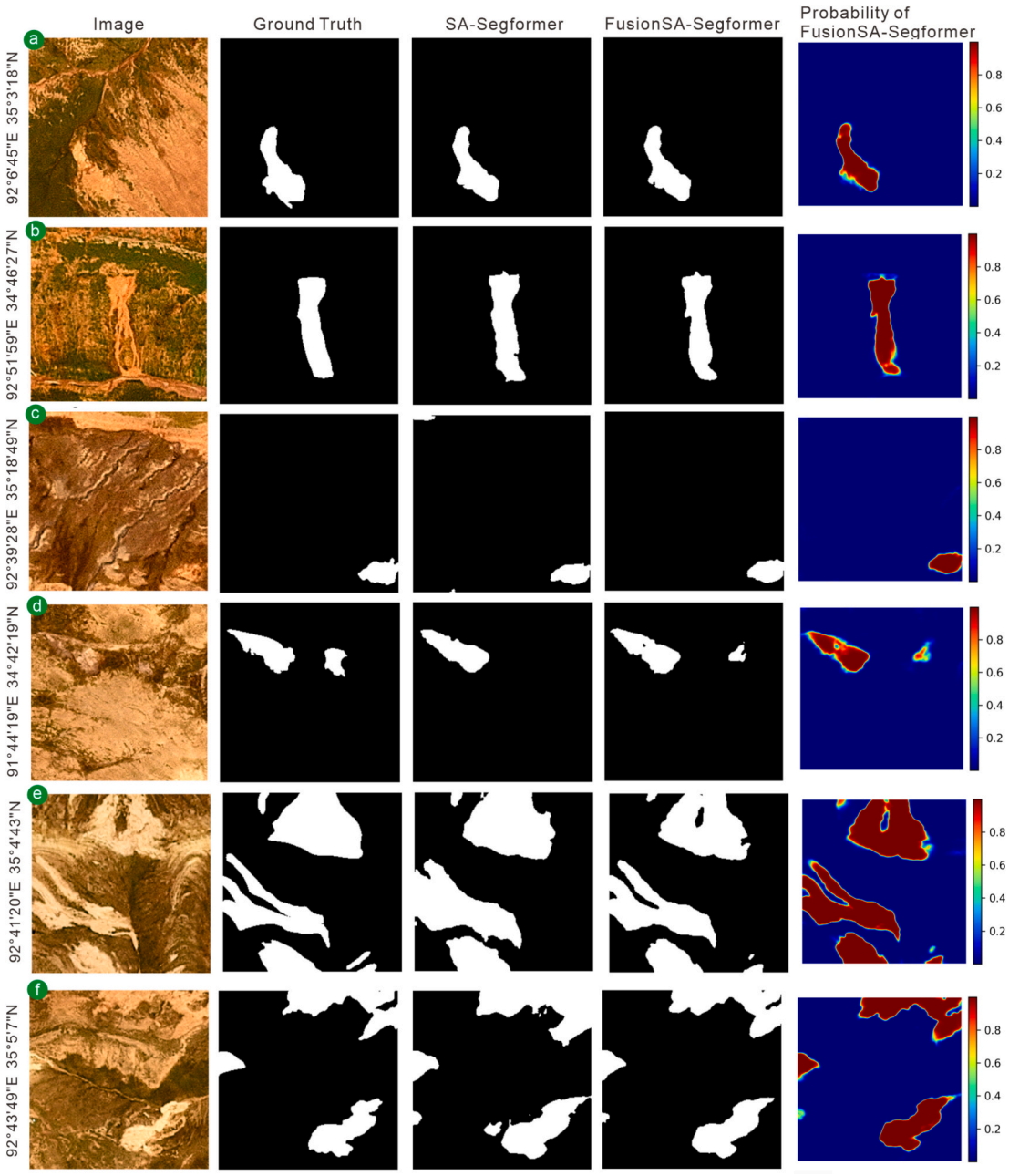

Fig. 15. RTS identification performance by FusionSA-SegFormer model. (a, b) single RTS identification; (c) RTS identification at image edge; (d) RTS identification in complex terrain; (e, f) RTSs group identification.

We first conducted a comprehensive environmental characterization of RTS distribution patterns through meticulous topographic, hydrological, and vegetation analyses. This foundational step provided critical insights into the contextual factors influencing RTS formation and progression. Subsequently, we developed optimized segmentation models, including the FusionSA-SegFormer, capable of detecting and mapping RTS features with unprecedented accuracy. By combining diverse data sources and leveraging advanced computational approaches, our methodology addresses the limitations of traditional methods, offering a robust framework for understanding and monitoring thermokarst processes in complex permafrost environments.

5.1. *Spatio-temporal patterns of RTSs*

Our comprehensive analysis of multi-source features uncovers distinct spatiotemporal patterns in the distribution and evolution of RTSs. The pronounced concentration of RTSs within the 4700–4800 m elevation band indicates a critical altitudinal thermal threshold for permafrost stability in this region (Luo et al., 2022). The preference for gentle to moderate slopes (5°-10°) suggests that this gradient strikes an ideal balance between thaw-induced soil weakening and material transport, enabling RTSs to sustain their retrogressive characteristics over time. Additionally, the dominance of mid-slope to valley settings, as evidenced by the TPI distribution, underscores the role of hydrological connectivity and sediment accumulation as key drivers of thermal erosion and slope instability (Kokelj and Jorgenson, 2013). The sharp decline in RTS frequency beyond 1000 m from water bodies further highlights a strong hydrological control; while direct thermal erosion triggers instability for RTSs immediately along shorelines, the broader association likely reflects the geomorphic concentration of ice-rich deposits in valley bottoms and enhanced soil thermal conductivity driven by higher moisture content. Aspect asymmetry also emerges as a

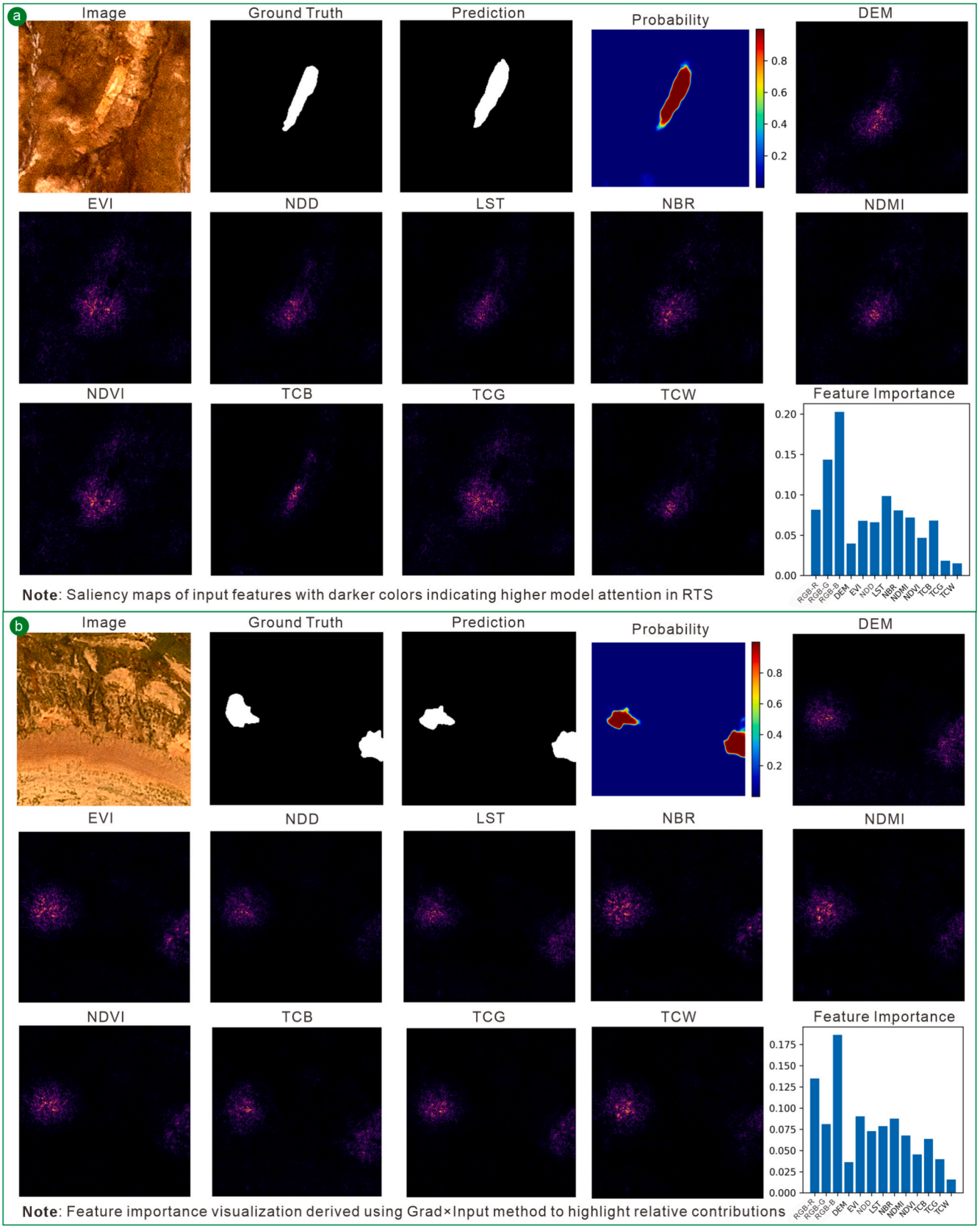

Fig. 16. Detailed visualization of RTS identification using FusionSA-SegFormer model with feature saliency and importance rankings at (a) 92°50′36″E, 35°26′43”N and (b) 92°24′19″E, 35°18′23”N.

significant factor, with RTSs predominantly oriented toward the Northwest (NW), North (N), and Northeast (NE). This pattern points to microclimatic effects driven by reduced solar insolation, shadowing, or wind-driven snow deposition, which foster episodic thawing on shaded, north-facing slopes. Such findings align with observations from other permafrost regions, including the Canadian Arctic and Siberian tundra, where pole-facing slopes often retain higher ice content and exhibit greater susceptibility to thermokarst activity (Kokelj and Jorgenson, 2013; Nitze et al., 2018a, 2018b). Moreover, the prevalence of valley-dominated landforms (approximately 75% of RTSs) concentrates moisture and sediment, amplifying the vulnerability of ice-rich permafrost to rapid thaw and mass wasting. Vegetation cover, predominantly Alpine grassland (58.5%) and meadow (41.3%), provides limited thermal buffering due to shallow root systems and minimal organic accumulation (Zhou et al., 2025; Luo et al., 2018). This renders permafrost in this region more susceptible to thermal disturbances compared to areas with thicker organic layers.

Collectively, these patterns underpin a conceptual model wherein RTS formation hinges on the convergence of three critical conditions: (i) ice-rich permafrost situated within altitudinal and aspect-controlled thermal regimes that facilitate degradation; (ii) hydrological connectivity that supplies liquid water to induce thermal erosion or enhance heat transfer in saturated soils; and (iii) geomorphic settings that promote slope instability and facilitate the exposure of massive ground ice. The temporal progression of RTS development involves vegetation loss, surface alteration, and altered energy fluxes, which perpetuate permafrost destabilization. Furthermore, emerging studies highlight engineering activities as additional catalysts for RTS progression in such environments (Luo et al., 2018).

5.2. Performance analysis of semantic segmentation models for RTS identification

A detailed evaluation of twelve semantic segmentation models reveals key relationships between architectural design and the efficacy of RTS identification, with distinct performance trends emerging across validation metrics, test set results, and boundary detection capabilities. Validation results highlight that attention-enhanced transformer architectures consistently outperform conventional convolutional neural networks (CNNs) across all metrics. Notably, SA-SegFormer achieved the highest overall score (0.786) with exceptional recall (0.790), demonstrating superior capability in comprehensive RTS identification. The rapid convergence and stable performance plateaus of top-tier models (Fig. 9) indicate their ability to effectively learn discriminative RTS features without significant overfitting. The performance hierarchy reveals three distinct tiers: attention-enhanced transformers (e.g., CA-SegFormer, SA-SegFormer) and advanced CNNs (e.g., ResNet, Unet++) form the top tier; standard transformers and modern CNNs (e.g., Swin Transformer, ConvNeXt variants) occupy the middle tier; and basic convolutional architectures lag significantly behind (Nitze et al., 2021). While ResNet's strong performance underscores the value of residual learning for feature preservation, its lower recall compared to attention-enhanced models suggests limitations in capturing long-range spatial dependencies crucial for RTS identification. Object-level evaluation on the test set offers critical insights into practical deployment. SA-SegFormer (F1-score, 0.817) and CA-SegFormer (F1-score, 0.781) confirm their dominance from validation, while also revealing key error patterns. A consistent trend of higher false positives relative to false negatives across all models points to a fundamental challenge in distinguishing true RTS features from spectrally similar terrain. This error distribution has notable practical implications: while models effectively identify most genuine RTS features (low false negatives), they struggle with specificity, often misclassifying other permafrost disturbances as RTS. Basic CNNs, with extreme error counts (e.g., 185 false positives and 263 false negatives), prove entirely inadequate for this complex task. Even top-performing models like SA-SegFormer (78 false positives) and CA-SegFormer (95 false positives) exhibit significant commission errors, suggesting the need for additional post-processing or multi-temporal verification in operational applications.

Qualitative analysis of segmentation results (Fig. 11) and feature visualizations (Fig. 17) highlights critical differences in boundary detection and feature learning mechanisms. The specific selection of Self-Attention (SA) and Cross-Attention (CA) over other variants was grounded in the unique morphological challenges of RTSs. Unlike simple channel or spatial attention mechanisms (e.g., SE-Block) that process dimensions independently, SA captures the global long-range dependencies required to semantically link distant RTS components (e.g., headwall and toe). Conversely, CA facilitates robust feature refinement by filtering background noise through cross-level correlation modeling. Consequently, Attention-enhanced models (CA-SegFormer, SA-SegFormer) exhibit precise boundary delineation and coherent probability distributions, reflecting their ability to model long-range dependencies and contextual relationships effectively (Jia et al., 2023). In contrast, models like Swin Transformer and CA-ConvNeXt produce blurred boundaries and fragmented predictions, indicating limitations in spatial localization despite reasonable feature extraction (Wang et al., 2022). Intermediate feature visualizations further underscore the superiority of attention-enhanced architectures, with CA-SegFormer's progressive feature refinement and SA-SegFormer's coherent attention patterns enabling both boundary precision and comprehensive RTS coverage. Conversely, chaotic intermediate patterns in Unet++ and dispersed activations in ResNet, despite their strong quantitative results, suggest less targeted feature organization, manifesting as boundary uncertainties and false positives.

This analysis yields two key conclusions regarding Phase II objectives: (i) Architectural selection: Transformer-based architectures with integrated attention mechanisms provide the most effective foundation for RTS identification, surpassing convolutional alternatives in both detection completeness and boundary precision; (ii) error characteristics: the consistent prevalence of commission errors across all models underscores the inherent challenge of RTS identification, indicating that optimal model selection must align with specific application needs, prioritizing high-recall models for comprehensive detection or high-precision models for accurate mapping (Zhang et al., 2022b).

5.3. Multi-source heterogeneous feature fusion strategy and feature importance analysis

The core innovation of this study centers on the development of the FusionSA-SegFormer model, which introduces a novel heterogeneous fusion strategy through dual-level integration of pixel-level and feature-level fusion within a unified deep learning framework for RTS identification. The efficacy of this approach is clearly evidenced in the object-level evaluation, where FusionSA-SegFormer achieved a 3.1% absolute improvement in F1-score over the baseline SA-SegFormer, attaining a score of 0.843 on the test set. This performance enhancement is particularly noteworthy given the model's concurrent reduction in both false positives and false negatives—commission errors decreased by 33.3% and omission errors by 21.1% compared to the baseline. Drawing inspiration from the RGB-Depth fusion paradigm in computer vision, our methodology adapts this concept to remote sensing by treating optical imagery as the “RGB” component and environmental features as the “Depth” channel (Lee et al., 2017; Zhang et al., 2024c). This architectural strategy proved highly effective in mitigating the persistent issue of false positives that affected other models, as demonstrated by SegFormer's high false discovery rate (54.2%) compared to FusionSA-SegFormer's significantly lower rate (16.2%). The pixel-level fusion integrates optical RGB imagery with carefully curated environmental features into a comprehensive 13-channel input stack, preserving raw data integrity while ensuring spatial correspondence across modalities. Meanwhile, the feature-level fusion, facilitated by multi-head self-attention, enables dynamic, context-aware integration of multi-scale

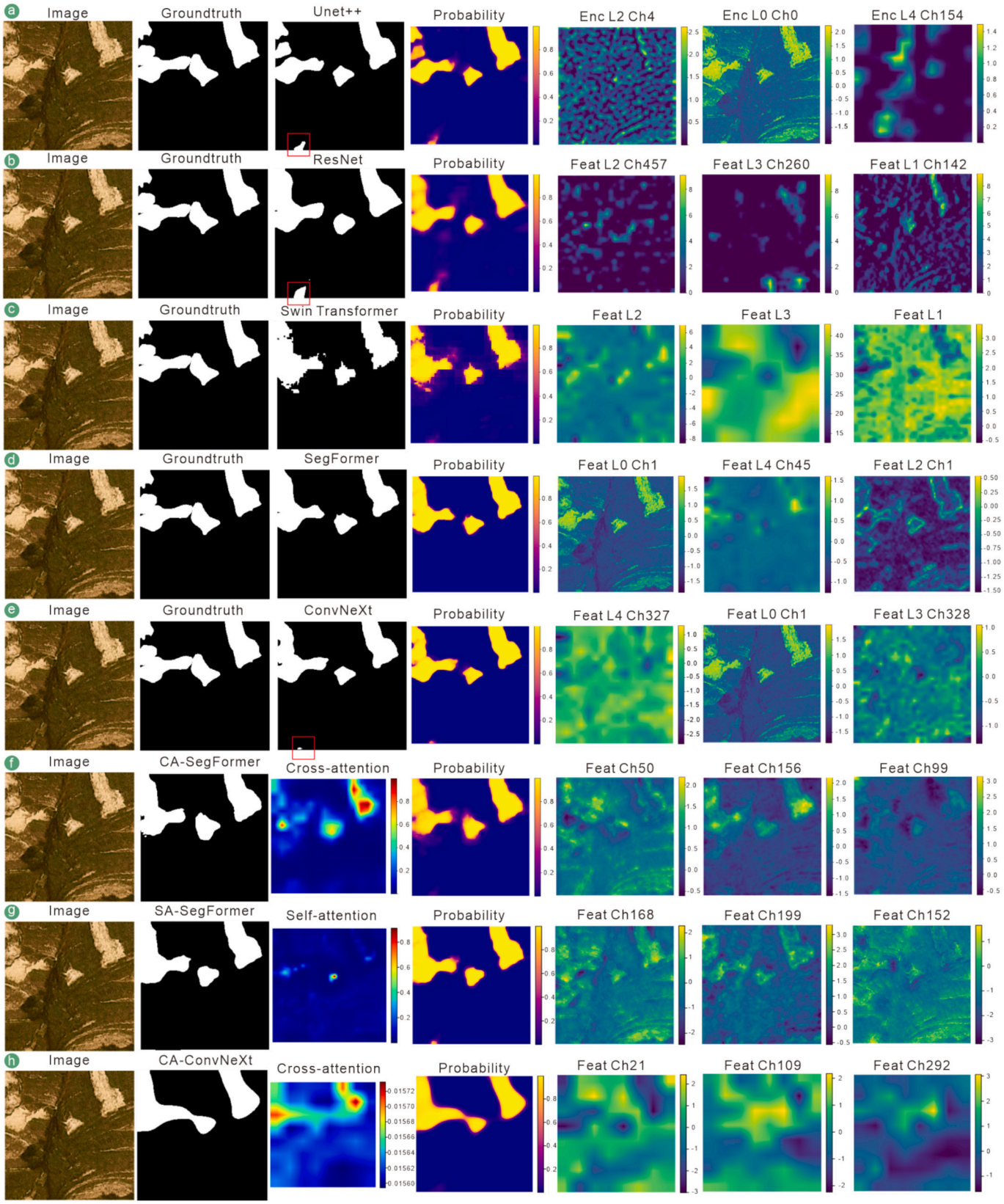

Fig. 17. Visual comparison of model predictions across probability maps and feature maps for: (a) Unet++, (b) ResNet, (c) Swin Transformer, (d) SegFormer, (e) ConvNeXt, (f) CA-SegFormer, (g) SA-SegFormer, and (h) CA-ConvNeXt. The visualization includes a probability map indicating prediction confidence and three channel-specific feature maps (RTS Loc: 92°35′46E, 34°48′56”N. Enc: encoder; L: layer; Ch: channel; Feat: feature).

features. This dual-level approach resulted in substantial gains in recall (10.9% improvement) while also enhancing precision (8.6% improvement), showcasing its ability to boost detection completeness without sacrificing reliability. The synergistic interplay of both fusion levels generates emergent capabilities unattainable through single-level methods, with object-level results confirming superior performance across diverse topographic settings and complex boundary conditions.

Feature importance analysis further validates our fusion strategy, uncovering a distinct hierarchy among environmental factor contributions. The prominence of RGB spectral channels, especially the blue band, emphasizes the critical role of visual characteristics in RTS identification, while the significant contributions of thermal indices (LST and NDD) highlight the importance of thermal regime considerations (Li et al., 2024). Remarkably, the high relevance of coarse-resolution NDD data, despite its 1 km spatial resolution, illustrates the power of heterogeneous fusion in capturing essential thermal processes that transcend spatial scale limitations. The interpretability framework embedded within FusionSA-SegFormer provides unparalleled insights into the interactions among heterogeneous features. By achieving a balanced reduction in errors—with a low false discovery rate (16.2%) and a low false negative rate (10.1%)—the model validates the fusion approach's capacity to resolve the inherent trade-off between detection sensitivity and specificity that often challenges RTS identification systems. This transparency not only affirms the robustness of our fusion methodology but also provides quantitative evidence regarding the relative utility of different data modalities. By revealing the dominant hierarchy of spectral and thermal features over static topographic variables, the study suggests that future permafrost disturbance mapping should prioritize the synergistic integration of high-resolution optical imagery with thermal regime indicators.

To further verify the necessity of the full feature set and assess potential redundancy, we conducted a stepwise backward elimination analysis (Fig. 18). Starting with the full 13-channel input (F1-Score: 0.796), we sequentially removed features beginning with the lowest-ranked variable (TCW). The results demonstrate a continuous decline in model performance with each removal. Eliminating the two least important features (TCW and TCG) caused the F1-Score to drop to 0.789, while removing the static DEM further reduced it to 0.784. A sharp performance degradation was observed when core spectral indices (e.g., NDVI, NDMI) were excluded, culminating in a baseline F1-Score of 0.734 when using optical RGB imagery alone. This monotonic trend confirms that although some features exhibit lower individual contribution scores, they provide complementary information that is essential for maximizing segmentation accuracy. Therefore, the integration of all 13 heterogeneous features is justified to resolve the morphological ambiguity of RTSs in complex permafrost environments.

5.4. Transferability validation

The identification of RTSs presents unique challenges fundamentally different from conventional landslide detection. Unlike typical mass movements with distinct morphological expressions, RTSs exhibit gradual transitions and high spectral similarity with surrounding undisturbed terrain (Yi et al., 2025). Our FusionSA-SegFormer framework effectively addresses these ambiguities by integrating complementary modalities—optical, thermal, and topographic data. By capturing multifaceted spectral transitions and thermal anomalies specific to permafrost disturbances, our heterogeneous fusion paradigm outperforms conventional models. The transferability of this framework appears highly promising, not only for mapping RTSs across different geographical contexts but also for detecting other cryospheric hazards like thermokarst lakes and ice-wedge degradation (Qin et al., 2023). To rigorously assess the model's generalization capabilities in real-world scenarios, we conducted independent spatial and temporal transferability evaluations.

5.4.1. Spatial transferability

Testing the spatial transferability of deep learning models is crucial for large-scale permafrost disturbance mapping. To effectively demonstrate this, we designed a spatial hold-out cross-validation strategy, ensuring absolutely no spatial overlap between the training and testing datasets. Two specific, spatially independent test regions were designated (Fig. 1). The Region T1 contains 194 test chips, while Region T2 contains 436 test chips. During the evaluation, when testing on Region T1, the data from Region T2 was merged into the training and validation sets alongside the remaining dataset. Conversely, when testing on Region T2, the data from Region T1 was incorporated into the training phase. This strategy rigorously ensures that the spatial characteristics of the test regions remain completely unseen by the model during training. The quantitative results demonstrate the framework's spatial generalization capacity. Testing on Region T1 yielded a precision of 0.65, a recall of 0.63, an F1-score of 0.60, and accuracy of 0.90. For Region T2, the model achieved a precision of 0.63, a recall of 0.68, an F1-score of 0.62, and accuracy of 0.91. Representative visualization results are presented in Fig. 19. The model exhibited particularly strong identification performance in areas with clustered RTS developments. While the metrics indicate a decrease in precision compared to the intra-regional test results, the overall accuracy remains robust (>0.90), confirming that spatial transferability across heterogeneous landscapes is viable.

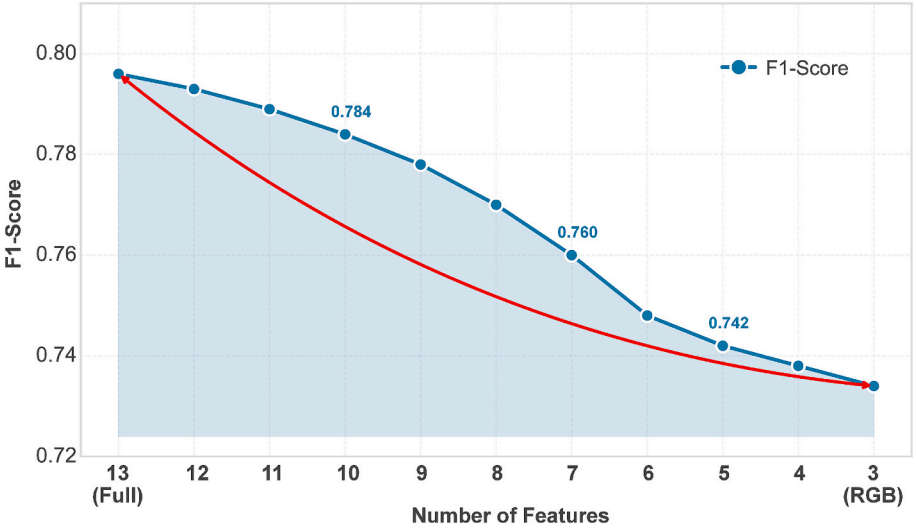

Fig. 18. Impact of stepwise feature elimination on model performance.

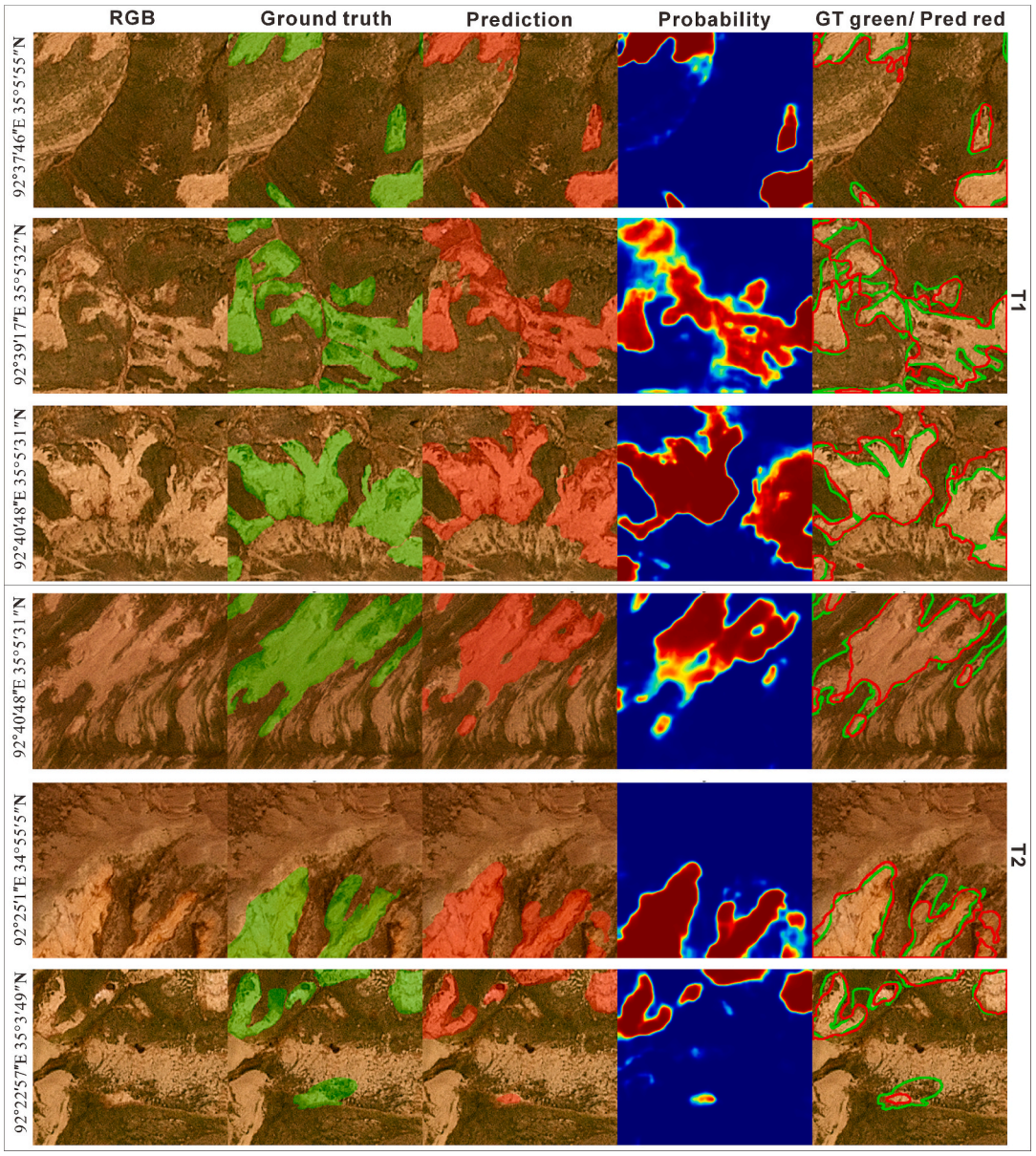

Fig. 19. Visualization of spatial transferability validation for the FusionSA-SegFormer model in Regions T1 and T2.

The performance degradation can be primarily attributed to the resolution limitations of PlanetScope imagery, which struggles to adequately resolve newly initiated, micro-scale RTSs. Furthermore, the extensive presence of bare ground in the Qinghai-Tibet Plateau introduces significant background noise, leading to spatial variation challenges.

5.4.2. Temporal transferability

Images acquired at different times pose distinct challenges for deep learning models due to temporal variations in ground conditions (e.g., vegetation phenology, soil moisture) and atmospheric effects. To evaluate temporal transferability, we constructed a comparative dataset encompassing optical imagery and corresponding spectral/thermal indices for the years 2023 and 2024. To prevent information leakage, a temporal cross-validation strategy was employed: "Train-2023 to Test-2024" (trained on 80% of the 2023 dataset and evaluated on the geographically corresponding 20% from 2024) and vice versa for "Train-2024 to Test-2023". The quantitative assessments yielded robust results across different years. The "Train-2023 to Test-2024" scenario achieved a precision of 0.656, recall of 0.629, F1-score of 0.603, and overall accuracy of 0.932. Similarly, the "Train-2024 to Test-2023" evaluation recorded a precision of 0.652, recall of 0.658, F1-score of 0.630, and accuracy of 0.937. Representative predictions (Fig. 20) demonstrate the model's consistent capability in delineating RTS boundaries across different temporal snapshots. Notably, the model successfully identified RTS boundaries even under conditions affected by partial cloud cover in the 2023 PlanetScope imagery. This resilience underscores the critical advantage of our multi-source fusion approach; the integration of robust thermal and topographic features effectively compensates for the degradation of optical spectral signals caused by atmospheric interference. Nevertheless, while temporal transferability is demonstrated, the

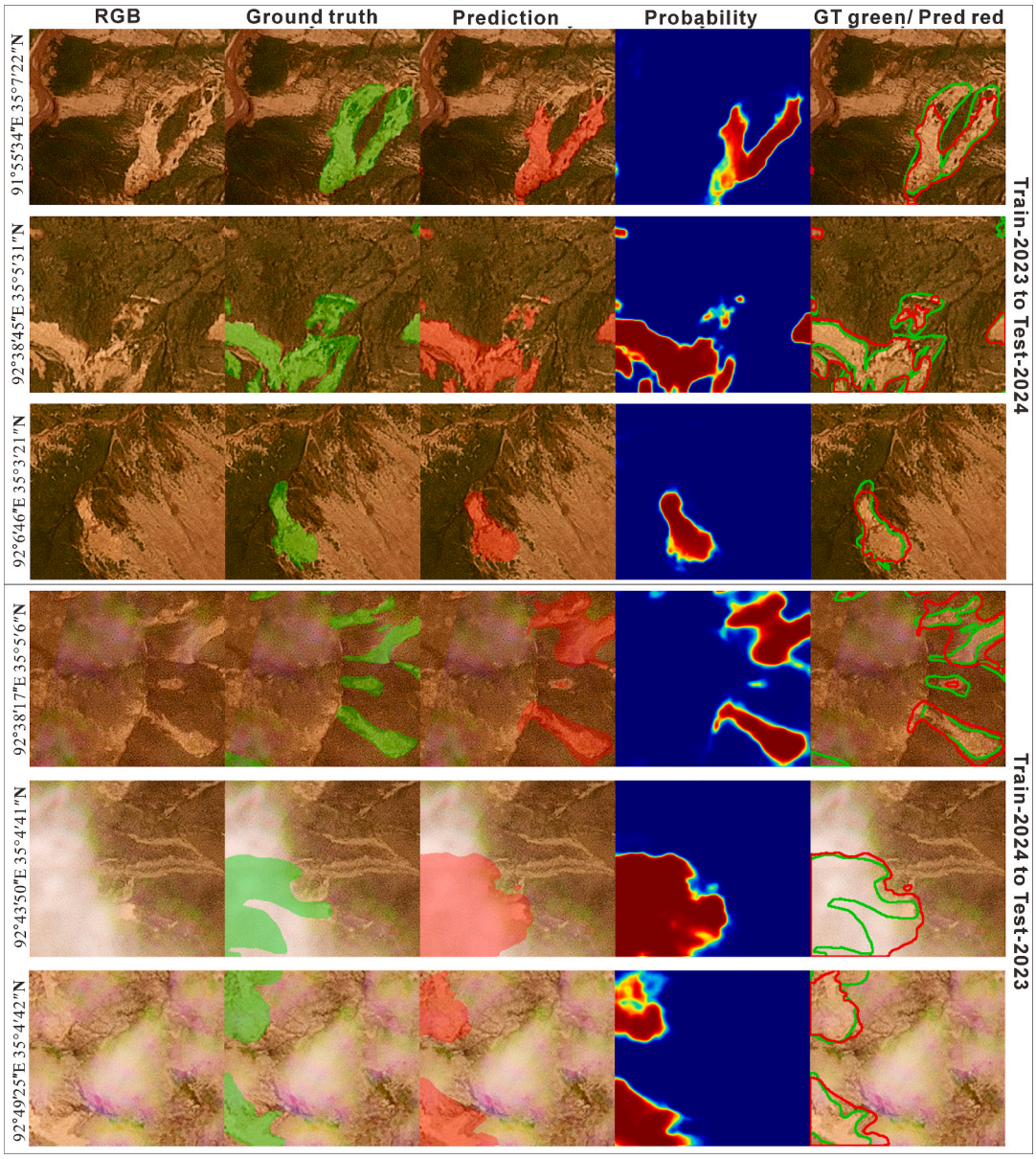

Fig. 20. Visualization of spatial transferability validation for the FusionSA-SegFormer model in Regions T1 and T2.

temporal dynamics of permafrost environments suggest that models should ideally be trained on multi-year composite datasets to achieve optimal operational robustness.

5.5. Limitations and future research

While this study presents significant advancements in RTS identification through the FusionSA-SegFormer framework, several limitations should be acknowledged to guide future research directions. The heterogeneous fusion framework, while effective, introduces specific challenges related to data compatibility and model architecture that warrant further investigation. Data limitations represent a primary constraint in our heterogeneous fusion framework. A prominent constraint relates to our reliance on commercial PlanetScope imagery. While its high spatial resolution (3–5 m) is highly advantageous for delineating fine-scale RTS morphological features, its application for large-scale or long-term operational monitoring is heavily restricted. Acquiring commercial data for vast permafrost regions incurs prohibitive costs. Furthermore, academic access (such as the Education and Research program) is typically limited to small quotas (e.g., 3000 km^2 per month) suitable only for localized studies, and recent shifts in data provision policies have further restricted access in certain regions. Additionally, the static nature of DEM data, particularly the absence of real-time topographic updates, limits the model's capacity to capture rapid surface changes associated with active RTS development. This temporal discrepancy is especially problematic for monitoring dynamic thermokarst processes where topographic alterations occur continuously throughout the thaw season (Kokelj et al., 2021). Similarly, the varying acquisition times and seasonal coverage of optical and thermal satellite data may introduce phenological biases, potentially affecting

the consistency of spectral and thermal indicators across different RTS sites (Nitze et al., 2018). The spatial resolution constraints of certain input features, particularly the moderate-resolution thermal indices from MODIS and Landsat, present additional uncertainties in fine-scale RTS delineation (Mohammadpour and Viegas, 2022). Interestingly, the success of our heterogeneous fusion framework is highlighted by the performance of coarse-resolution data, despite its 1 km resolution, the Net Degree-Days ranked as the third most important feature after optical bands, underscoring its exceptional effectiveness in characterizing RTS dynamics. This promising finding suggests that employing spatial downscaling approaches to derive higher-resolution NDD data could significantly enhance future RTS identification capabilities (Su et al., 2025). While these datasets provide valuable regional coverage, their native resolution may insufficiently capture the detailed spatial patterns of smaller RTS features, potentially leading to under-detection or boundary inaccuracies in complex terrain.

Methodologically, the attention mechanisms in our heterogeneous fusion framework, while generally enhancing model performance, introduce higher computational costs and certain computational dependencies that may affect result stability. The self-attention modules, though effective for capturing long-range dependencies across different data modalities, demonstrate variable sensitivity to input feature quality and scaling (Samadzadegan et al., 2025; Zhang et al., 2022c). We observed occasional instability in attention weight distributions when processing highly heterogeneous terrain conditions, particularly in areas with strong spectral contrasts or mixed land cover types. The dual-level fusion strategy, while innovative, presents challenges in optimally balancing contributions from diverse feature types within the heterogeneous framework. The fixed architecture design may not optimally adapt to varying environmental contexts across the study area, potentially leading to suboptimal feature weighting in regions with atypical RTS characteristics. Furthermore, the model's performance demonstrates some dependency on the quality and representativeness of the training samples, with reduced reliability in areas under-represented in the training dataset. Consequently, the model's generalizability to other permafrost regions (e.g., Arctic lowlands) with different landscape characteristics remains to be fully validated in future studies.

Future research should prioritize enhancing heterogeneous fusion capabilities through several key directions: First, developing dynamic topographic datasets that capture seasonal ground changes would significantly improve the temporal dimension of our fusion approach. Second, creating adaptive fusion architectures that can automatically adjust feature weighting based on environmental contexts would enhance model robustness across diverse permafrost landscapes (Ma et al., 2024). Third, expanding the heterogeneous fusion framework to incorporate additional data sources, including InSAR deformation measurements and higher-resolution thermal data, would provide more comprehensive feature representation. Fourth, establishing standardized benchmark datasets for RTS identification across different permafrost regions would enable more objective evaluation of heterogeneous fusion framework (Biskaborn et al., 2019a, 2019b). Fifth, integrating physical process models with deep learning approaches could help bridge the gap between statistical associations and mechanistic understanding within the fusion framework. Finally, expanding validation efforts to include diverse permafrost environments and longer temporal scales would strengthen the model's generalization capacity for climate change impact assessment (Guo et al., 2024).

6. Conclusions

This study addresses the critical challenge of automated retrogressive thaw slump identification in permafrost regions by developing an integrated methodology that combines multi-source feature analysis with advanced deep learning techniques. Through systematic implementation of three methodological phases, we have established a comprehensive framework for RTS identification and characterization that significantly advances current capabilities in permafrost disturbance monitoring.

RTSs predominantly cluster within the 4700–4800 m elevation, on gentle to moderate slopes (3°–7°), clear preference for northwest to northeast aspects, in mid-slope to valley settings, and within 1000 m of water bodies. Temporally, sustained degradation from 2019 to 2024 was evident through declining vegetation indices (e.g., EVI, NDVI), decreasing moisture (NDMI), and rising warming thermal regimes (LST and NDD), indicating a feedback loop of permafrost thaw, surface exposure, and deepening instability.

The comprehensive evaluation of twelve semantic segmentation models identifies SA-SegFormer as the optimal baseline architecture for RTS identification. On the test set, it maintains strong performance with an F1-score of 0.818, precision of 0.769, and recall of 0.872. The attention-enhanced variants consistently outperform conventional architectures, forming a clear performance hierarchy that underscores the importance of advanced attention mechanisms for complex geomorphological feature detection.

The developed FusionSA-SegFormer model, incorporating multi-source feature fusion strategy, establishes new state-of-the-art performance for RTS segmentation. On the test set, FusionSA-SegFormer achieving a superior F1-score of 0.843, with false positives decreased by 33.3% and false negatives by 21.1% compared to the baseline. Feature importance analysis reveals a clear hierarchy in feature contributions, with RGB spectral channels, particularly the blue band, demonstrating paramount importance, followed by thermal indices including LST and NDD. The dominance of spectral and thermal characteristics over topographic features highlights their critical role in RTS identification, providing valuable insights for future feature selection strategies in permafrost disturbance mapping.

Crucially, rigorous evaluations of spatial and temporal transferability confirmed the robustness and generalization capacity of the FusionSA-SegFormer framework. Employing a spatial hold-out cross-validation strategy across two independent regions, the model maintained high overall accuracy (>0.90) and achieved F1-scores of 0.60 to 0.62, demonstrating its viability for cross-regional mapping despite the complex bare-ground background of the Qinghai-Tibet Plateau. Furthermore, temporal cross-validation between 2023 and 2024 yielded consistent F1-scores (0.603–0.630) and exceptional accuracies (>0.93). The integration of multi-source features effectively mitigated the impact of atmospheric interferences (e.g., cloud cover) and temporal ground variations, underscoring the model's resilience for long-term, dynamic monitoring tasks.

This research contributes both methodological innovations and scientific insights to permafrost studies, offering an effective framework for automated RTS monitoring while advancing our understanding of the environmental factors controlling thermokarst dynamics in warming mountain environments.

CRediT authorship contribution statement

Taorui Zeng: Writing – original draft, Visualization, Data curation. **Xiao Feng:** Writing – review & editing, Methodology. **Yuanming Lai:** Investigation, Funding acquisition. **Thomas Glade:** Supervision, Methodology.

Informed consent statement

Not applicable for studies not involving humans.

Institutional review board statement

Not applicable for studies not involving humans or animals.

Funding

This research was funded by the Postdoctoral Fellowship Program of CPSF (GZC20250310), General Program of Chongqing Natural Science Foundation (CSTB2025NSCQ-GPX0837), Special Funding of Chongqing Postdoctoral Research Project (2020925020), the China Postdoctoral Science Foundation (2024MD764034). The authors declare no conflict of interest.

Declaration of Competing Interest

None.

Acknowledgments

We wish to thank the State Key Laboratory of Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources.

Appendix A. Appendix

This appendix details the comprehensive computational methodologies and data processing workflows employed for deriving topographic, environmental, spectral, and thermal features utilized in this study. All processing of satellite-derived indices was performed on the Google Earth Engine cloud computing platform to leverage its scalability and efficiency in handling large geospatial datasets.

A.1. Topographic and environmental features

The primary data source for topographic features was a high-resolution DEM. Using standard surface analysis tools within ArcGIS Pro, key terrain attributes such as Slope, Aspect, Plan Curvature, and Profile Curvature were calculated. Additionally, the TPI was computed to distinguish landscape features such as ridges (positive TPI) and valleys (negative TPI) by comparing the elevation of each cell to the mean elevation of its surrounding neighborhood, thereby providing a relative topographic context. To obtain a comprehensive and ecologically relevant landform classification, we utilized the SRTM Landform dataset (Theobald et al., 2015). This global dataset, with a resolution of 90 m, integrates the Continuous Heat-Insolation Load Index and a multi-scale TPI derived from the SRTM DEM. It categorizes the landscape into 15 distinct geomorphic classes (e.g., valley, slope, ridge, flat) based on TPI and slope parameters, offering a standardized representation of the geomorphic context critical for understanding RTS distribution and initiation. Fluvial erosion is a key trigger for the initiation of RTSs, especially in ice-rich permafrost landscapes, where it destabilizes banks and often prompts inland retrogression of slumps. Consequently, Distance to Water Body was identified as a critical predisposing factor for such mass movements. To quantify this spatial relationship, we derived the Distance to Water Body feature using the Global Surface Water dataset (Pekel et al., 2016), a high-resolution (30 m) product that maps the long-term distribution of surface water from 1984 to 2022 based on Landsat imagery. We specifically employed the max_extent layer, a binary image delineating all pixels where water has been detected over the 38-year period, capturing both permanent and intermittent water bodies. A Euclidean distance analysis was then conducted on this layer to calculate the shortest distance from each cell in the study area to the nearest potential source of fluvial erosion, thereby providing a robust hydrological proximity metric. Vegetation plays a significant role in thermally insulating underlying permafrost through canopy cover and root-mat/organic layers, which mitigate summer thaw. To represent annual vegetation status, we utilized the Annual Vegetation Maps of the Qinghai-Tibet Plateau for the year 2022 (Zhou et al., 2025). This dataset, with a spatial resolution of 500 m and an overall accuracy of 83.27%, offers a reliable classification of vegetation cover types across the plateau, facilitating an understanding of how different vegetation regimes influence permafrost stability.

A.2. Spectral features

All spectral indices utilized in this study were derived from the Sentinel-2 Level-2 A surface reflectance collection spanning 2019 to 2024. A rigorous and standardized preprocessing workflow was implemented to ensure data quality and consistency. Specifically, cloud and cirrus pixels were masked using the QA60 band, and annual median composites were generated from images with less than 30% cloud cover to minimize atmospheric interference and seasonal variability.

EVI was employed as a robust metric for assessing canopy structure and photosynthetic activity, particularly in high-biomass regions. By incorporating the blue band to correct for aerosol influences and mitigate saturation effects in dense vegetation canopies, EVI provides a more reliable indicator of vegetation structure and phenology compared to traditional indices. Low or declining EVI values may indicate sparse vegetation or degradation, rendering the underlying permafrost more vulnerable to thermal disturbances. The EVI was calculated using the following formula:

$$EVI = 2.5 \times \frac{B8 - B4}{B8 + 6 \times B4 - 7.5 \times B2 + 1} \tag{A.1}$$

Normalized Difference Vegetation Index, a fundamental indicator of vegetation greenness and density, was computed as a baseline comparison. The calculation function:

$$NDVI = \frac{B8 - B4}{B8 + B4} \tag{A.2}$$

Normalized Burn Ratio, originally designed for mapping fire severity, was repurposed to detect surface disturbance and soil exposure associated

with thermokarst processes. It highlights changes in moisture content and vegetation cover by contrasting the NIR and shortwave infrared (SWIR2) regions:

$$NBR = \frac{B8 - B12}{B8 + B12} \tag{A.3}$$

The Normalized Difference Moisture Index was calculated to sensitively monitor vegetation water content and soil moisture conditions, both critical factors influencing the thermal and mechanical properties of permafrost terrain. The index was computed using the standard formula:

$$NDMI = \frac{B8 - B11}{B8 + B11} \tag{A.4}$$

Tasseled Cap Transformation components were derived using Sentinel-2 specific coefficients (Shi and Xu, 2019). The Brightness (TCB) component was calculated as:

$$TCB = 0.2569B2 + 0.2934B3 + 0.3020B4 + 0.3740B6 + 0.4180B7 + 0.3580B8 + 0.3834B8A + 0.0896B11 + 0.0780B12 \tag{A.5}$$

The Greenness (TCG) component was computed as:

$$TCG = -0.2818B2 - 0.3020B3 - 0.4283B4 - 0.2959B5 + 0.1602B6 + 0.3127B7 + 0.3138B8 + 0.4261B8A - 0.1341B11 - 0.2538B12 \tag{A.6}$$

The Wetness (TCW) component was derived as:

$$TCW = 0.1763B2 + 0.1615B3 + 0.0486B4 + 0.0170B5 + 0.0223B6 + 0.0219B7 - 0.0755B8 - 0.0910B8A - 0.7701B11 - 0.5293B12 \tag{A.7}$$

All indices were clamped to their theoretical ranges and exported at their native spatial resolutions (10 m or 20 m).

A.3. Thermal features

The surface thermal regime was characterized using data from both Landsat and MODIS satellites. 30 m-resolution Land Surface Temperature was derived from Landsat 8/9 Collection 2 Level-2 products. After masking clouds, shadows, and snow using the QA_PIXEL band, the thermal band (ST_B10) was converted to degrees Celsius. From the time-series of valid LST observations, mean annual LST composites were generated for each year.

Net Degree-Days was derived from MODIS MOD11A1 daily land surface temperature data. After quality filtering using QC bands, daily mean LST was computed by averaging daytime and nighttime observations. Annual cumulative indices were calculated for each year 2019–2024. Thawing Degree-Days (TDD) were computed as:

$$TDD = \sum_{i=1}^{n} \max(LST_i, 0) \tag{A.8}$$

Freezing Degree-Days (FDD) were calculated as:

$$FDD = \sum_{i=1}^{n} |\min(LST_i, 0)| \tag{A.9}$$

The Net Degree-Days was then derived as:

$$NDD = TDD - FDD \tag{A.10}$$

Data availability

I have made all the codes and datasets open source. The relevant repository address can be found in the manuscript.

References

Balser, A.W., Jones, J.B., Gens, R., 2014. JOURNAL OF GEOPHYSICAL RESEARCH-EARTH SURFACE 119, 1106–1120.

Barnhart, T.B., Crosby, B.T., 2013. Remote Sens. 5, 2813–2837.

Biskaborn, B.K., Smith, S.L., Noetzli, J., Matthes, H., Vieira, G., Streletskiy, D.A., Schoeneich, P., Romanovsky, V.E., Lewkowicz, A.G., Abramov, A., Allard, M., Boike, J., Cable, W.L., Christiansen, H.H., Delaloye, R., Diekmann, B., Drozdov, D., Etzelmüller, B., Grosse, G., Guglielmin, M., Ingeman-Nielsen, T., Isaksen, K., Ishikawa, M., Johansson, M., Johannsson, H., Joo, A., Kaverin, D., Kholodov, A., Konstantinov, P., Kröger, T., Lambiel, C., Lanckman, J.-P., Luo, D., Malkova, G., Meiklejohn, I., Moskalenko, N., Oliva, M., Phillips, M., Ramos, M., Sannel, A.B.K., Sergeev, D., Seybold, C., Skryabin, P., Vasiliev, A., Wu, Q., Yoshikawa, K., Zheleznyak, M., Lantuit, H., 2019. Nat. Commun. 10, 264.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with Atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 833–851.

Chen, J., Wu, T., Zou, D., Liu, L., Wu, X., Gong, W., Zhu, X., Li, R., Hao, J., Hu, G., Pang, Q., Zhang, J., Yang, S., 2022. Remote Sens. Environ. 268, 112778.

Coulombe, S., Fortier, D., Lacelle, D., Kanevskiy, M., Shur, Y., 2019. CRYOSPHERE 13, 97–111.

Dai, C., Ward Jones, M.K., Van Der Sluijs, J., Nesterova, N., Howat, I.M., Liljedahl, A.K., Higman, B., Freymueller, J.T., Kokelj, S.V., Sriram, S., 2025. Nat. Commun. 16.

Dobinski, W., 2011. Earth-Science Reviews 108, 158–169.

Feng, X., Du, J., Wu, M., Chai, B., Miao, F., Wang, Y., 2024. Landslides 21, 2211–2226.

Gao, S., Liu, Y., Men, X., Zhao, H., Wang, L., Fu, Z., Zhang, Z., Yang, Y., Jiang, G., Wu, Q., 2026. Remote Sens. Environ. 335, 115262.

Ghorbanzadeh, O., Xu, Y., Ghamisi, P., Kopp, M., Kreil, D., 2022. IEEE Trans. Geosci. Remote Sensing 60, 1–17.

Guo, H., Zhu, W., Xiao, C., Zhao, C., Chen, L., 2024. Int. J. Appl. Earth Obs. Geoinf. 133, 104114.

Guo, Zizheng, Zeng, T., Zhang, Y., Yu, W., Wang, L., Guo, Zhanxu, Glade, T., 2025. Geomorphology 486, 109886.

Huang, L., Luo, J., Lin, Z., Niu, F., Liu, L., 2020. Remote Sens. Environ. 237.

Huang, L., Willis, M.J., Li, G., Lantz, T.C., Schaefer, K., Wig, E., Cao, G., Tiampo, K.F., 2023. ISPRS J. Photogramm. Remote Sens. 205, 301–316.

Jia, Z., You, K., He, W., Tian, Y., Feng, Y., Wang, Y., Jia, X., Lou, Y., Zhang, J., Li, G., Zhang, Z., 2023. IEEE Trans. on Image Process. 32, 1829–1842.

Jiao, C., Wang, Y., Shan, Y., He, P., He, J., 2023a. Land Degrad. Dev. 34, 2573–2588.

Jiao, Z., Xu, Z., Guo, R., Zhou, Z., Jiang, L., 2023b. Int. J. Disaster Risk Sci. 14, 523–538.

Jones, M.K.W., Pollard, W.H., Jones, B.M., 2019. Environ. Res. Lett. 14.

Kokelj, S.V., Jorgenson, M.T., 2013. Permafrost & Periglacial 24, 108–119.

Kokelj, S.V., Kokoszka, J., Van Der Sluijs, J., Rudy, A.C.A., Tunnicliffe, J., Shakil, S., Tank, S.E., Zolkos, S., 2021. Cryosphere 15, 3059–3081.

Lafrenière, M.J., Lamoureux, S.F., 2019. Earth Sci. Rev. 191, 212–223.

Lee, S., Park, S.-J., Hong, K.-S., 2017. RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation. In: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp. 4990–4999.

Lewkowicz, A.G., Way, R.G., 2019. Nat. Commun. 10.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. International Journal of Applied Earth Observation and Geoinformation 112, 102926.
Li, P., Wang, Y., Si, T., Ullah, K., Han, W., Wang, L., 2024a. Eng. Appl. Artif. Intell. 127, 107337.
Li, X.-L., Zhang, Z., Lu, J.-X., Brouchkov, A., Yan, Q.-K., Yu, Q.-H., Zhang, S.-R., Melnikov, A., 2024b. Adv. Clim. Chang. Res. 15, 113–123.
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, in: 2021 IEEE/CVF International conference on computer vision (ICCV). In: Montreal, Q.C. (Ed.), Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. Canada, pp. 9992–10002.
Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Lu, P., Han, J., 2025. Permafr. Periglac. Process. 36, 329–342.
Luo, J., Niu, F., Lin, Z., Liu, M., Yin, G., 2015. Sci. Bull. 60, 556–564.
Luo, L., Ma, W., Zhuang, Y., Zhang, Y., Yi, S., Xu, J., Long, Y., Ma, D., Zhang, Z., 2018. Ecol. Indic. 93, 24–35.
Luo, J., Niu, F., Lin, Z., Liu, M., Yin, G., 2019. GEOMORPHOLOGY 341, 79–85.
Luo, J., Niu, F., Lin, Z., Liu, M., Yin, G., Gao, Z., 2022. Geophys. Res. Lett. 49.
Luo, J., Yin, G.-A., Niu, F.-J., Dong, T.-C., Gao, Z.-Y., Liu, M.-H., Yu, F., 2024. Adv. Clim. Chang. Res. 15, 253–264.
Lupachev, A.V., Tananaev, N.I., Murton, J.B., Kalinin, P.I., Malyshev, V.V., Danilov, P.P., 2025. Quatern. Res.
Ma, X., Zhang, X., Pun, M.-O., Liu, M., 2024. IEEE Trans. Geosci. Remote Sensing 62, 1–15.
Ma, J., Wang, G., Sun, S., Song, C., Li, J., Guo, L., Li, K., Huang, P., Lin, S., 2025. Int. J. Appl. Earth Obs. Geoinf. 137.
Maier, K., Bernhard, P., Ly, S., Volpi, M., Nitze, I., Li, S., Hajnsek, I., 2025. Int. J. Appl. Earth Obs. Geoinf. 137.
Mohammadpour, P., Viegas, C., 2022. Applications of multi-source and multi-sensor data fusion of remote sensing for Forest species mapping. In: Pandey, P.C., Arellano, P. (Eds.), Advances in Remote Sensing for Forest Monitoring. Wiley, pp. 255–287.
Nitze, I., Grosse, G., Jones, B.M., Romanovsky, V.E., Boike, J., 2018. Nat. Commun. 9.
Nitze, I., Heidler, K., Barth, S., Grosse, G., 2021. Remote Sens. 13.
Niu, F., Jiao, C., Luo, J., He, J., He, P., 2023. INTERNATIONAL JOURNAL OF DISASTER RISK SCIENCE 14, 566–585.
Obu, J., Lantuit, H., Grosse, G., Guenther, F., Sachs, T., Helm, V., Fritz, M., 2017. Geomorphology 293, 331–346.
Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. Nature 540, 418–422.
Qin, Y., Lu, P., Han, J., Wang, Q., Li, Z., Wu, J., Li, R., 2023. CATENA 231, 107309.
Rauf, F., Khan, M.A., Bhatti, M.K., Hamza, A., Aleryani, A., Alouane, M.T.-H., AlHammadi, D.A., Nam, Y., 2024. IEEE j. Sel. Top. Appl. Earth Observations Remote Sensing 17, 18622–18634.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.
Samadzadegan, F., Toosi, A., Dadrass Javan, F., 2025. Int. J. Remote Sens. 46, 1327–1402.
Schuur, E.A.G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven, C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K., Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Nature 520, 171–179.
Sejourne, A., Costard, F., Fedorov, A., Gargani, J., Skorve, J., Masse, M., Mege, D., 2015. GEOMORPHOLOGY 241, 31–40.
Shen, T., Jiang, P., Ju, Q., Zhao, J., Chen, X., Lin, H., Yang, B., Tan, C., Zhang, Y., Fu, X., Yu, Z., 2024. J. Hydrol. 628, 130501.
Shi, T., Xu, H., 2019. IEEE j. Sel. Top. Appl. Earth Observations Remote Sensing 12, 4038–4048.
Su, Q., Meng, X., Sun, L., Guo, Z., 2025. Remote Sens. 17, 2350.
Sun, Z., Gao, Z., Wang, Y., Liu, G., 2024. CATENA 247.
Tao, J., Liljedahl, A.K., Burn, C.R., Grosse, G., Noetzli, J., Goetz, S.J., Douglas, T.A., Yang, Y., 2025. Environ. Res. Lett. 20, 100201.
Targ, S., Almeida, D., Lyman, K., 2016. Resnet in Resnet: Generalizing Residual Architectures.
Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., Albano, C.M., 2015. PloS One 10, e0143619.
Tian, Y., Zeng, T., Lü, Q., Jiang, H., Yang, S., Cao, H., Yu, W., 2026. Remote Sens. 18, 380.
van der Sluijs, J., Kokelj, S.V., Fraser, R.H., Tunnicliffe, J., Lacelle, D., 2018. Remote Sens. 10.
Wang, B., Paudel, B., Li, H., 2016. LANDSLIDES 13, 1–8.
Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S., 2022. IEEE Geosci. Remote Sensing Lett. 19, 1–5.
Wang, H., Liu, J., Zeng, S., Xiao, K., Yang, D., Yao, G., Yang, R., 2024. Landslides 21, 901–917.
Wu, F., Jiang, C., Wang, C., Zou, L., Li, T., Guan, S., Tang, Y., 2025. Geomorphology 471.
Xia, Z., Liu, L., Mu, C., Peng, X., Zhao, Z., Huang, L., Luo, J., Fan, C., 2024. Geophys. Res. Lett. 51.
Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Adv. Neural Inf. Process. Syst. 34, 12077–12090.
Yan, X., Zhang, X., Liu, B., Mithan, H.T., Hellstrom, J., Nuber, S., Drysdale, R., Wu, J., Lin, F., Zhao, N., Zhang, Y., Kang, W., Liu, J., 2025. Nat. Commun. 16, 290.
Yang, D., Qiu, H., Ye, B., Liu, Y., Zhang, J., Zhu, Y., 2023a. J. Geophys. Res. Earth Surf. 128.
Yang, Y., Rogers, B.M., Fiske, G., Watts, J., Potter, S., Windholz, T., Mullen, A., Nitze, I., Natali, S.M., 2023b. Remote Sens. Environ. 288.
Yang, G., Qiu, H., Wang, N., Yang, D., Liu, Y., 2025a. Remote Sens. Environ. 325.
Yang, C., Zhu, Y., Zhang, J., Wei, X., Zhu, H., Zhu, Z., 2025b. Landslides 22, 471–483.
Yi, Y., Wu, T., Wu, M., Jiang, H., Yang, Y., Rogers, B.M., 2025. Earth-Science Reviews 261, 105020.
Yin, G., Niu, F., Lin, Z., Luo, J., Liu, M., 2017. Sci. Total Environ. 581–582, 472–485.
Yu, F., Luo, J., Niu, F., Lin, Z., Li, B., Mu, Y., Ju, X., Liu, M., Yin, G., Gao, Z., Zhao, X., Zhang, C., 2025. Landslides 22, 3351–3363.
Zeng, T., Glade, T., Xie, Y., Yin, K., Peduto, D., 2023. International Journal of Disaster Risk Reduction 94, 103820.
Zeng, T., Guo, Z., Wang, L., Xu, C., Wu, F., Jin, B., Peduto, D., 2025. Bull. Eng. Geol. Environ. 84, 356.
Zhang, T., Barry, R.G., Knowles, K., Heginbottom, J.A., Brown, J., 1999. Polar Geogr. 23, 132–154.
Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C., 2022a. IEEE Trans. Geosci. Remote Sensing 60, 1–20.
Zhang, G., Nan, Z., Hu, N., Yin, Z., Zhao, L., Cheng, G., Mu, C., 2022b. Earth's Future 10, e2022EF002652.
Zhang, Q., Xu, Y., Zhang, J., Tao, D., 2022c. ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and beyond.
Zhang, H., Wang, H., Zhang, J., Luo, J., Yin, G., 2023. Int. J. Disaster Risk Sci. 14, 539–548.
Zhang, X., Li, L., Han, L., 2024a. Landslides 21, 2913–2925.
Zhang, R., Lv, J., Yang, Y., Wang, T., Liu, G., 2024b. Landslides 21, 1849–1864.
Zhang, L., Zeng, T., Wang, L., Li, L., 2024c. Earth Sci Inform 17, 3547–3566.
Zhao, Z., Xia, Z., Liu, L., 2023. Decadal evolution of retrogressive thaw slumps retrieved from Landsat imagery via Heatmap regression: A case study of the Beiluhe region in Central Tibet. In: IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Pasadena, CA, USA, pp. 94–97.
Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. UNet++: A nested U-net architecture for medical image segmentation. In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 3–11.
Zhou, W., Ma, T., Yin, X., Wu, X., Li, Q., Rupakheti, D., Xiong, X., Zhang, Q., Mu, C., de Foy, B., Rupakheti, M., Kang, S., Qin, D., 2023. Environ. Sci. Technol. 57, 6910–6921.
Zhou, G., Ren, H., Zhang, L., Lv, X., Zhou, M., 2025. Earth Syst. Sci. Data 17, 773–797.
Zou, D., Zhao, L., Hu, G., Du, E., Liu, G., Wang, C., Li, W., 2024. Permafrost Temperature Baseline at 15 Meters Depth in the Qinghai-Tibet Plateau (2010–2019).