frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 15 April 2025
DOI 10.3389/feart.2025.1543271

OPEN ACCESS

EDITED BY
Xin Jin,
Yunnan University, China

REVIEWED BY
Hamza Daud,
China University of Geosciences
Wuhan, China
Nadiia Kopiika,
University of Birmingham, United Kingdom

*CORRESPONDENCE
Wenbo Lin,
15101207316@163.com

RECEIVED 11 December 2024
ACCEPTED 14 March 2025
PUBLISHED 15 April 2025

CITATION
Lin W, Li X and Li T (2025) Multi-source image feature extraction and segmentation techniques for karst collapse monitoring. *Front. Earth Sci.* 13:1543271.
doi: 10.3389/feart.2025.1543271

Multi-source image feature extraction and segmentation techniques for karst collapse monitoring

Wenbo Lin[1]*, Xiao Li[2] and Tingting Li[3]

[1]School of Geology, Gansu Industrial Vocational and Technical College, Tianshui, Gansu, China, [2]Guangdong Nonferrous Industry Building Quality Inspection Co., Ltd., Guangzhou, Guangdong, China, [3]School of Information Engineering, Gansu Industrial Vocational and Technical College, Tianshui, Gansu, China

Introduction: Karst collapse monitoring is a complex task due to data sparsity, underground dynamics, and the demand for real-time risk assessment. Traditional approaches often fall short in delivering timely and accurate evaluations of collapse risks.

Methods: To address these challenges, we propose the Integrated Karst Collapse Prediction Network (IKCPNet), a novel framework that combines multi-source imaging, geophysical modeling, and machine learning techniques. IKCPNet processes seismic imaging, hydrological patterns, and environmental factors using an advanced data encoding mechanism and a physics-informed module to capture subsurface changes. A dynamic risk assessment strategy is incorporated to enable real-time feedback and probabilistic mapping.

Results: Experimental evaluations on the OpenSARShip dataset demonstrate that IKCPNet achieves an accuracy of 94.34 $\pm$ 0.02 and an IoU of 90.23 $\pm$0.02, outperforming the previous best model by 1.22 and 0.89 points, respectively.

Discussion: These results highlight the effectiveness of IKCPNet in improving prediction accuracy and risk mitigation, showcasing its potential for enhancing geological hazard monitoring through multi-source data integration.

KEYWORDS

karst collapse, multi-source image, segmentation techniques, cyber-physical systems, risk prediction

1 Introduction

The study of multi-source image feature extraction and segmentation techniques is critical for monitoring post-hazard consequences, including damages, spatial changes, and environmental impacts (Yusuf et al., 2024). These techniques are particularly valuable in areas rendered inaccessible to traditional monitoring methods after disasters, whether natural or human-induced (Kopiika et al., 2025). By leveraging data from disparate sources such as satellites, UAVs, and digital image correlation (DIC), multi-source imagery provides a robust platform to assess the condition of lands, buildings, and infrastructural assets. The integration of these data sources enables detailed evaluations of surface deformations, structural integrity, and temporal changes, facilitating proactive disaster management and post-event recovery planning (Wang et al., 2024). Karst collapse monitoring, as one of the

most critical applications of such multi-source approaches, plays a vital role in mitigating geological hazards and ensuring public safety (Wang, 2024). Karst collapses, often caused by natural processes such as water erosion or human activities like mining, can lead to severe damage to infrastructure and loss of life. Monitoring these collapses requires precise and timely detection of surface deformations and subsurface features. Multi-source imagery, including satellite, aerial, and ground-based images, offers a rich dataset to capture the complex spatial and temporal dynamics of karst areas (Kopiika et al., 2024). The diverse nature of these data sources necessitates advanced feature extraction and segmentation techniques to integrate, analyze, and interpret them effectively (Hatamizadeh et al., 2021). Accurate segmentation enables the identification of vulnerable zones and the prediction of collapse events, supporting proactive disaster management and risk mitigation strategies (Xu et al., 2023).

These techniques used manually defined rules and deterministic models to identify features such as sinkholes, fractures, and surface deformations from imagery (Huang et al., 2020a). By leveraging domain knowledge, these methods were able to achieve a certain degree of accuracy in controlled settings (Yu et al., 2023). For example, edge detection algorithms and morphological operations were employed to delineate collapse features (Valanarasu et al., 2021). These approaches were limited by their reliance on predefined thresholds and their inability to adapt to variations in image quality, environmental conditions, and data sources (Zhang et al., 2021). The labor-intensive nature of rule creation and the static nature of symbolic systems restricted their scalability and applicability in dynamic karst environments (Kopiika and Blikharskyy, 2024). These methods utilize supervised and unsupervised learning algorithms to identify patterns in large datasets, enabling the automatic classification of features relevant to karst collapse monitoring (Jain et al., 2022). Techniques such as random forests, support vector machines, and k-means clustering have been employed to analyze spectral, spatial, and textural information from images (Zhang et al., 2024). Machine learning has proven effective in handling diverse data sources and accommodating variations in environmental conditions (Yin et al., 2022). Its reliance on extensive labeled datasets and sensitivity to parameter tuning often pose challenges, particularly in regions where labeled data are scarce or inconsistently distributed (Wu et al., 2022). Moreover, the interpretability of these models remains a concern, as understanding the rationale behind their predictions is critical for geological applications (Malhotra et al., 2022).

Recent advancements in deep learning have transformed the landscape of feature extraction and segmentation for karst monitoring (Huan et al., 2023). Convolutional neural networks (CNNs), for instance, have demonstrated remarkable success in automatically learning hierarchical representations of image features, enabling highly accurate segmentation of karst collapse regions. Pre-trained models, such as U-Net and Mask R-CNN, have further enhanced this capability by enabling transfer learning, where knowledge from general image processing tasks is adapted to the specific domain of karst monitoring (Luo et al., 2020). These models excel in integrating multi-source imagery, capturing complex spatial patterns, and providing detailed segmentations of collapse features (Lüddecke and Ecker, 2021). Deep learning approaches often require substantial computational resources and are limited by their "black-box" nature, which hinders their acceptance in domains that prioritize transparency and explainability. The variability and noise inherent in multi-source imagery remain challenges for deep learning-based methods.

Traditional karst collapse monitoring techniques primarily rely on field surveys, geophysical imaging, and hydrological analysis (Jha et al., 2020). While these methods provide valuable insights into subsurface conditions, they suffer from several limitations. First, many approaches depend on sparse observational data, limiting their ability to detect early-stage collapse indicators (Chaitanya et al., 2020). Second, existing methods often lack the capability to integrate multi-source geospatial information effectively, leading to incomplete risk assessments. Third, machine learning-based approaches have demonstrated promise in hazard prediction but often operate as black-box models, making them difficult to interpret in geological contexts (Atigh et al., 2022). most current techniques struggle with real-time adaptability, which is crucial for early warning systems and disaster response.

Research Gap: Despite advances in remote sensing and computational modeling, there remains a lack of an integrated framework that can (1) fuse multi-source data from seismic, hydrological, and environmental measurements, (2) incorporate geophysical principles into predictive models, and (3) adapt dynamically to real-time feedback for risk assessment. Existing solutions either focus on empirical correlations without a physics-based foundation or rely solely on geophysical models without leveraging modern AI-driven feature extraction techniques. To bridge this gap, we propose the Integrated Karst Collapse Prediction Network (IKCPNet), a novel framework that combines multi-modal sensing, physics-informed geomechanical modeling, and machine learning-based segmentation. IKCPNet addresses the aforementioned challenges by.

- Introducing a hybrid feature extraction mechanism that fuses seismic imaging, hydrological data, and environmental metrics to provide a holistic risk assessment.
- Embedding geophysical principles within the model to improve interpretability and align predictions with real-world geological processes.
- Implementing a Dynamic Risk Mitigation Strategy (DRMS) that adapts predictions based on real-time environmental changes, ensuring robust early warning capabilities.

By integrating these innovations, IKCPNet provides a comprehensive solution for karst collapse monitoring, significantly improving prediction accuracy, interpretability, and adaptability. Experimental results demonstrate that our framework outperforms state-of-the-art methods in terms of risk assessment precision and segmentation quality.

The remainder of this paper is organized as follows: Section 3 details the proposed methodology, including data fusion strategies and geophysical modeling techniques. Section 4 presents experimental results and comparative evaluations. Section 6 discusses key findings and potential limitations and summarizes our contributions and outlines future research directions.

2 Related work

2.1 Multi-source image feature integration

The use of multi-source imaging in karst collapse monitoring has gained significant traction due to its ability to capture diverse and complementary data features (Chen et al., 2020). By combining data from optical, thermal, LiDAR, and hyperspectral imaging, researchers can extract detailed information about geological structures, surface deformations, and environmental conditions. Feature extraction techniques tailored to these imaging modalities enable the identification of subtle changes indicative of karst collapse risks, such as soil subsidence, vegetation stress, and water infiltration patterns (Ouyang et al., 2020). Advanced algorithms, such as convolutional neural networks (CNNs) and wavelet transforms, are frequently employed to process multi-source image data. These techniques are effective in isolating key features while suppressing noise, enhancing the reliability of collapse risk assessments (Gao et al., 2021). For instance, optical imaging provides high-resolution spatial data, while thermal imaging highlights temperature anomalies associated with underground water flow (Lin et al., 2021). LiDAR data offers precise topographic measurements, and hyperspectral imaging captures material composition changes. Integrating these features into a unified framework enhances the detection and prediction of karst collapses, making it a cornerstone of geohazard monitoring systems.

2.2 Automated image segmentation methods

Accurate segmentation of multi-source images is critical for delineating regions affected by karst collapse and assessing their spatial extent (Liu et al., 2021). Traditional segmentation methods, such as thresholding and edge detection, are increasingly being replaced by more sophisticated machine learning and deep learning approaches (Wang et al., 2021). Techniques like U-Net, Mask R-CNN, and Fully Convolutional Networks (FCNs) enable the automated and precise segmentation of complex geological features, even in heterogeneous terrains (Jin et al., 2024). Recent advances emphasize the integration of contextual information through multi-scale and multi-modal approaches. For example, LiDAR data can provide a baseline topographic map, while optical and hyperspectral images contribute spectral and textural details for refining segmentation boundaries (Jin et al., 2023b). These techniques allow for the detection of collapse-prone areas by identifying subsurface cavities, cracks, and shifts in vegetation cover. Segmentation models trained on annotated datasets can generalize to new regions, improving their applicability in monitoring widespread karst landscapes (Jin et al., 2023a). The development of automated, high-accuracy segmentation methods is instrumental in enabling timely and effective karst collapse risk mitigation strategies.

2.3 Applications in early warning systems

Multi-source image feature extraction and segmentation play a pivotal role in the development of early warning systems for karst collapse (Isensee et al., 2020). By integrating imaging data with geospatial and temporal analyses, these systems can detect early indicators of instability, such as gradual surface deformation or changes in hydrological conditions. Automated pipelines for processing and analyzing multi-source data ensure that potential collapse events are identified with minimal delay, allowing for proactive risk management. Real-time monitoring systems utilize cloud-based platforms to aggregate and analyze data from diverse imaging sources, delivering alerts to stakeholders through user-friendly interfaces (Cao et al., 2021). Feature extraction and segmentation algorithms are continuously updated with new data, improving the predictive accuracy of these systems over time. Furthermore, the integration of artificial intelligence (AI) techniques, such as reinforcement learning, enhances the adaptability of early warning systems to dynamic environmental conditions. These applications underscore the critical importance of advanced image processing techniques in safeguarding communities and infrastructure from the impacts of karst collapses (Minaee et al., 2020).

3 Methods

3.1 Background

Karst collapse monitoring is a critical area of study within geoscience and environmental engineering, focusing on the prediction, detection, and analysis of sinkhole formations and subsurface ground instability in karst regions. Karst terrains, characterized by soluble rock formations such as limestone, dolomite, and gypsum, are particularly susceptible to subsurface erosion and collapse, posing risks to infrastructure, ecosystems, and human safety. This subsection provides a high-level introduction to the methodologies and innovations underpinning our approach to karst collapse monitoring. The proposed framework combines multi-modal sensing, advanced machine learning models, and geomechanical simulations to achieve a robust and scalable monitoring system.

The problem of karst collapse monitoring can be framed as the detection and prediction of subsurface failures using a combination of geophysical data and predictive models. Let $\Omega \subset \mathbb{R}^3$ represent the karst domain, where each point $\mathbf{x} = (x, y, z) \in \Omega$ corresponds to a spatial location in three-dimensional space. The key variables in this framework include the ground displacement $\mathbf{u}(\mathbf{x}, t)$, which represents the ground displacement at location $\mathbf{x}$ and time t, and the porosity or void ratio $\phi(\mathbf{x}, t)$, which describes the distribution of voids within the subsurface.

The dynamics of karst collapse processes are influenced by the coupling of mechanical deformation and fluid flow. These interactions can be described by a set of equations:

The mass conservation equation, which accounts for changes in porosity and fluid movement, is given by Equation 1:

$$\frac{\partial \phi}{\partial t} + \nabla \cdot \left(\phi \mathbf{v}_f \right) = q_s, \tag{1}$$

where $\mathbf{v}_f$ is the fluid velocity and q_s represents sources or sinks such as recharge or pumping.

The momentum conservation equation, which models mechanical deformation under linear elasticity, is expressed as:

$$\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} = 0, \tag{2}$$

where $\boldsymbol{\sigma}$ is the stress tensor, ρ is the material density, and $\mathbf{g}$ is the gravitational force acting on the material.

Darcy's law, which governs fluid flow through porous media, is given by Equation 3:

$$\mathbf{v}_f = -\frac{\kappa}{\mu} \nabla P, \tag{3}$$

where κ represents the permeability, μ is the fluid viscosity, and P is the pore pressure.

Sinkhole formation is driven by the interaction of geological, hydrological, and mechanical factors. The indicators of collapse can be formalized as a function of several key variables, including porosity, ground displacement, hydrostatic pressure, permeability, maximum stress, and the elastic modulus of the subsurface material. The collapse indicator function is represented as Equation 4:

$$C(\mathbf{x}, t) = f\left(\phi, \mathbf{u}, P, \kappa, \sigma_{\max}, E\right), \tag{4}$$

where $\sigma_{\max}$ is the maximum stress within the material and E is the elastic modulus of the subsurface material.

The critical collapse threshold C_{critical} is defined such that if the collapse indicator exceeds this value, a potential collapse is triggered (Equation 5):

$$C(\mathbf{x}, t) \geq C_{\text{critical}} \quad \Rightarrow \quad \text{potential collapse}. \tag{5}$$

The monitoring framework for karst collapse integrates a variety of data sources, including geophysical measurements, hydrological data, surface deformation, and environmental conditions. Geophysical data, such as seismic imaging, electrical resistivity tomography (ERT), and ground-penetrating radar (GPR), help map subsurface voids and structures. Hydrological data, including groundwater levels and flow rates from piezometers, provide insights into the water dynamics within the subsurface. Surface deformation measurements, such as InSAR (Interferometric Synthetic Aperture Radar) and GPS data, detect subsidence at the surface. Environmental conditions, such as rainfall, anthropogenic loads, and land use changes, also play a role in influencing subsurface stability.

The goal is to predict high-risk regions $\mathcal{R}_c \subset \Omega$ and assess the likelihood of collapse based on the data matrix $\mathbf{D}$ and the collapse indicator function $C(\mathbf{x}, t)$. The high-risk region is defined as Equation 6:

$$\mathcal{R}_c = \left\{\mathbf{x} \in \Omega : C(\mathbf{x}, t) \geq C_{\text{critical}}\right\}. \tag{6}$$

The prediction problem can be formulated as Equation 7:

$$\hat{\mathcal{R}}_c = \arg\max_{\mathcal{R}_c} \mathcal{P}\left(C(\mathbf{x}, t) \geq C_{\text{critical}} | \mathbf{D}\right), \tag{7}$$

where $\mathcal{P}(\cdot)$ represents the posterior probability derived from observational data and predictive models.

Karst collapse monitoring is a complex geoscientific challenge that involves predicting and analyzing the dynamics of sinkhole formation in regions underlain by soluble rocks such as limestone, dolomite, and gypsum. These collapses are driven by a combination of natural and anthropogenic processes, including groundwater fluctuations, surface loading, and chemical dissolution, all of which create voids beneath the surface that can ultimately result in collapse. This section formalizes the problem, introduces key mathematical frameworks, and outlines the geophysical principles that guide karst collapse prediction and monitoring.

In this study, we primarily consider the small deformation assumption, meaning that the deformation of the surface and subsurface medium is assumed to be relatively minor at the analysis scale. However, we acknowledge that in certain extreme cases, such as severe karst collapse events, large deformation may occur. Currently, our model is based on linear elasticity theory (Equation 2) and does not explicitly account for the nonlinear effects of large deformation. To enhance adaptability to large deformation, our Subsurface Dynamics Module (SDM) employs the finite element method (FEM) to solve coupled partial differential equations (PDEs). This framework can be extended to incorporate more complex material constitutive relationships, such as nonlinear constitutive models or finite strain theory. One of the future research directions is to introduce a Lagrangian-based approach into SDM to capture the impact of large deformation on collapse prediction. Additionally, we plan to integrate high-precision surface deformation monitoring data (e.g., InSAR and LiDAR mapping) to optimize the model, making it more applicable to large deformation scenarios.

3.2 Integrated Karst Collapse Prediction Network (IKCPNet)

We introduce the Integrated Karst Collapse Prediction Network (IKCPNet), a novel framework designed to model, predict, and monitor karst collapses by integrating multi-modal sensing data, geophysical simulations, and state-of-the-art deep learning methods, including cross-modal attention and multi-scale feature fusion, specifically designed for karst collapse prediction.

The Integrated Karst Collapse Prediction Network (IKCPNet) is specifically designed to model the complex interactions between geological, hydrological, and mechanical processes that drive karst collapses. These interactions are integrated into the framework through two key components: the Multi-Modal Data Encoder (MMDE) and the Subsurface Dynamics Module (SDM), both of which work synergistically to capture and process the underlying subsurface dynamics. First, the MMDE fuses diverse data sources, such as seismic imaging, hydrological measurements, surface deformation data, and environmental metrics, into a unified representation. Each data modality is processed through modality-specific neural networks that extract unique features relevant to the geological and environmental context. For instance, seismic imaging captures subsurface structural properties, while hydrological data reflects groundwater fluctuations and flow patterns. The MMDE employs an attention mechanism that dynamically weights the importance of features from each modality, ensuring that the fused representation prioritizes the most critical information for predicting collapse risks. This data fusion not only integrates the individual characteristics of geological, hydrological, and mechanical data but also highlights their interdependencies. Second,

the SDM incorporates geophysical principles to explicitly model the interactions between these processes. This is achieved by solving coupled partial differential equations (PDEs) that describe the dynamic behavior of subsurface properties. For example (Equations 9, 10):

$$\frac{\partial \phi}{\partial t} + \nabla \cdot \left(\phi \mathbf{v}_f \right) = q_s, \tag{8}$$

$$\mathbf{v}_f = -\frac{\kappa}{\mu} \nabla P, \tag{9}$$

$$\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} = 0, \tag{10}$$

Where ϕ represents porosity, $\mathbf{v}_f$ is fluid velocity, P is pore pressure, κ is permeability, $\boldsymbol{\sigma}$ is the stress tensor, ρ is density, and $\mathbf{g}$ is gravitational force. Equation 8 governs the conservation of mass within the subsurface, accounting for changes in porosity and fluid flow. Equation 9 describes fluid flow through porous media using Darcy's law, while Equation 10 models the balance of forces within the subsurface material. The SDM solves these PDEs using finite element methods (FEM), producing detailed simulations of subsurface states, such as stress distributions, pore pressure variations, and deformation fields. These simulations are then integrated with the fused features from the MMDE, enabling the framework to capture how geological structures, hydrological changes, and mechanical stresses interact to influence collapse risks.

As shown in Figure 1, IKCPNet consists of a static branch and a dynamic branch, both of which contribute to the final prediction through cross-modal feature fusion and deep spatial integration. The static branch processes long-term geological information, including stratigraphic structures, historical collapse records, and geomechanical parameters. These features remain relatively stable over time and are extracted through the Subsurface Dynamics Module (SDM). Within SDM, local patch extraction captures spatial correlations in geological formations, while cross-attention mechanisms enhance the interaction between static and dynamic data. The dynamic branch, in contrast, captures short-term environmental changes such as surface deformations, hydrological variations, and meteorological influences. These features are processed through the Multi-Modal Data Encoder (MMDE), where convolutional neural networks (CNNs) extract spatial patterns and encode them into high-dimensional feature representations. MMDE incorporates feature fusion strategies such as multiplicative weighting and attention-based concatenation to ensure robust integration of diverse input modalities. Deep Spatial Integration (DSI) serves as the final stage, aggregating static and dynamic information to produce a refined collapse risk prediction. The network employs a series of two-dimensional convolutional layers to process remote sensing imagery, SAR data, and hydrological measurements. These CNN layers enable efficient spatial feature extraction and enhance the model's capacity to detect collapse-prone regions. The detection head then generates a probabilistic map indicating areas at high risk, leveraging the learned representations from both branches. In the revised manuscript, we will further elaborate on the specific functions of SDM and MMDE and include a refined figure legend to improve clarity. We appreciate the reviewer's insightful comments and will incorporate these refinements to enhance the comprehensibility of the methodology.

The attention mechanism in Figure 1 plays a crucial role in integrating multi-modal data and enhancing feature interactions between the static and dynamic branches of IKCPNet. The model employs cross-attention within the Subsurface Dynamics Module (SDM) and feature fusion mechanisms in the Multi-Modal Data Encoder (MMDE) to improve the representation of collapse risk factors. Within SDM, cross-attention is used to align and fuse static geological information with dynamic environmental variations. The local patch extraction step captures relevant features from both branches, where the static branch provides prior knowledge of subsurface conditions, while the dynamic branch introduces time-sensitive updates. The attention mechanism operates by computing a weighted relationship between query features from the static branch and key-value pairs from the dynamic branch, ensuring that the most relevant temporal changes are emphasized in relation to stable geological structures. This allows the model to selectively focus on regions where evolving environmental conditions significantly impact collapse risk. In MMDE, attention mechanisms are embedded in the cross-modal feature fusion process to enhance the integration of multi-source spatial features. The model applies multiplicative weighting and concatenation-based attention to dynamically adjust feature importance across different input modalities. This mechanism ensures that signals from high-impact variables, such as rapid surface deformations or hydrological fluctuations, receive greater emphasis when generating final risk predictions. By leveraging attention at multiple levels, IKCPNet effectively combines long-term geophysical knowledge with short-term environmental dynamics, improving its ability to detect collapse-prone regions with higher accuracy.

The architecture of IKCPNet is designed as a hierarchical multi-branch neural network that integrates static geological features and dynamic environmental changes for karst collapse prediction. The network consists of multiple stages, including convolutional feature extraction, cross-modal feature fusion, and deep spatial integration, ensuring effective learning from heterogeneous data sources. The neural network comprises four main stages, each containing multiple local and global feature extraction blocks. The first stage processes raw input data through a patch partitioning layer, followed by a convolutional embedding layer that transforms spatial data into feature representations. The backbone network consists of stacked local feature blocks, which use convolutional layers to capture fine-grained spatial features, and global feature blocks, which incorporate a shifted window multi-head self-attention mechanism to model long-range dependencies. Each stage consists of multiple residual units to facilitate gradient flow and improve learning stability. The network employs hierarchical feature fusion, where the outputs from local and global pathways are aggregated at each stage using element-wise summation. This ensures that both fine-scale structural variations and large-scale spatial relationships are preserved. After passing through four hierarchical stages, the final fused feature representation undergoes global average pooling, followed by a fully connected classification layer that outputs the predicted collapse risk. The detection head refines the final probability distribution, ensuring accurate localization of collapse-prone regions. In the revised manuscript, we will provide additional details on the layer configuration, including the number of convolutional and attention-based layers in each processing block.

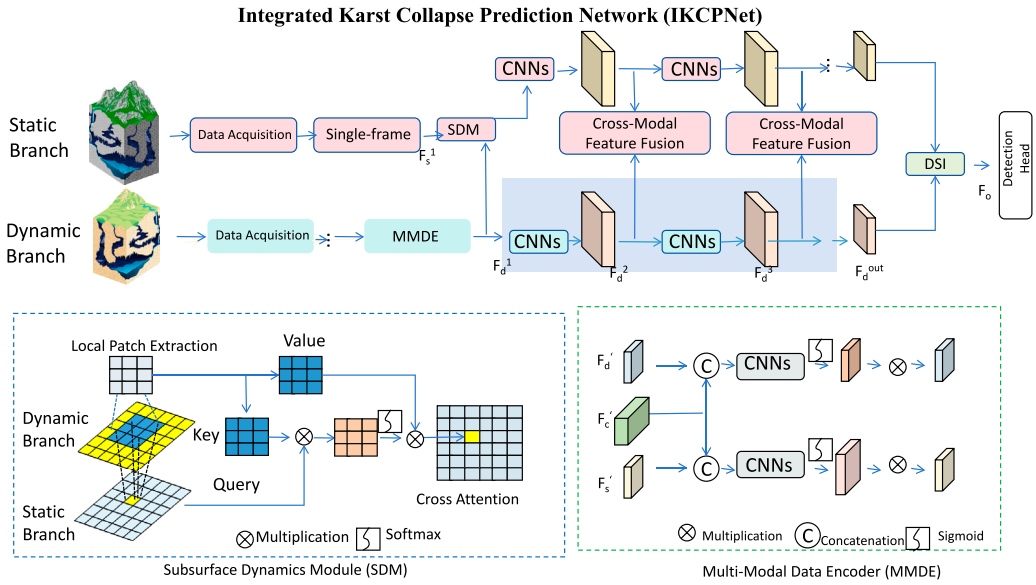

FIGURE 1
The architecture of the Integrated Karst Collapse Prediction Network (IKCPNet). The model consists of a static branch and a dynamic branch, both contributing to collapse risk prediction through cross-modal feature fusion and deep spatial integration (DSI). The static branch processes long-term geological information using the Subsurface Dynamics Module (SDM), which employs local patch extraction and cross-attention mechanisms to enhance subsurface feature representations. The dynamic branch captures short-term environmental variations through the Multi-Modal Data Encoder (MMDE), where convolutional neural networks (CNNs) extract and fuse multi-source spatial features. The final prediction is generated by the detection head after deep spatial integration, producing a probabilistic collapse risk map.

The Multimodal Data Encoder is responsible for processing and integrating various data sources in IKCPNet. These diverse data sources provide complementary perspectives on the subsurface environment, each capturing different aspects of the terrain's physical properties, dynamics, and external influences. The input data matrix $\mathbf{D} = \{\mathbf{D}_{\text{geo}}, \mathbf{D}_{\text{hydro}}, \mathbf{D}_{\text{surf}}, \mathbf{D}_{\text{env}}\}$ represents the raw data from each modality, where $\mathbf{D}_{\text{geo}}$, $\mathbf{D}_{\text{hydro}}$, $\mathbf{D}_{\text{surf}}$, and $\mathbf{D}_{\text{env}}$ correspond to the geophysical, hydrological, surface deformation, and environmental data, respectively. Since each modality contains valuable but distinct information, it is essential to extract relevant features from each using modality-specific neural networks $f_i(\cdot)$. These networks are parameterized by Θ_i and tailored to each data type $i \in \{\text{geo}, \text{hydro}, \text{surf}, \text{env}\}$, ensuring that the model is capable of learning the unique patterns within each source (Equation 11):

$$\mathbf{z}_i = f_i(\mathbf{D}_i; \Theta_i), \quad i \in \{\text{geo}, \text{hydro}, \text{surf}, \text{env}\}. \tag{11}$$

The feature embeddings $\mathbf{z}_i$ obtained from each modality are then fused into a unified latent space representation $\mathbf{z}_{\text{fused}}$ using an attention mechanism, which learns to prioritize the most important features based on their relevance to the final prediction task. The attention mechanism computes a weight α_i for each modality, where these weights are determined dynamically during training, and the weighted sum of the modality-specific embeddings is used to produce the final fused representation (Equation 12):

$$\mathbf{z}_{\text{fused}} = \sum_i \alpha_i \mathbf{z}_i, \quad \alpha_i = \frac{\exp(w_i)}{\sum_j \exp(w_j)}, \tag{12}$$

where w_i are learnable parameters associated with each modality, controlling the importance of each input modality in the fusion process. The softmax function ensures that the attention weights α_i are normalized, meaning they sum to 1 across all modalities, thus providing a balanced integration of all data types. The fused representation $\mathbf{z}_{\text{fused}}$ effectively captures the joint information from all modalities, which is critical for robust and accurate collapse risk prediction. The use of attention mechanisms allows IKCPNet to adaptively focus on the most informative data sources based on the context, improving the model's ability to handle complex and heterogeneous environments. By combining complementary insights from different data types, the MMDE enables IKCPNet to make more informed and precise predictions of karst collapse risks, leveraging the strengths of each modality while mitigating the limitations inherent in any single data source.

Figure 2 illustrates the hierarchical stages of the Multi-Modal Data Encoder, which serves as a core component of the IKCPNet framework. The MMDE processes multi-modal inputs such as satellite imagery, seismic data, and hydrological measurements by extracting and fusing features at multiple levels. The process begins with the patch partitioning of input data, followed by a linear embedding stage that converts these patches into feature vectors for further processing. This prepares the data for hierarchical feature extraction through local and global feature blocks, where local feature blocks capture fine-grained spatial details such as surface textures and deformations, while global feature blocks employ attention mechanisms to focus on broader contextual patterns, like regional structural variations. The MMDE is designed with four sequential stages, each progressively refining the feature representations through localized and global paths. At each stage, hierarchical feature fusion blocks (HEF Blocks) integrate local and global features to ensure that both detailed and high-level

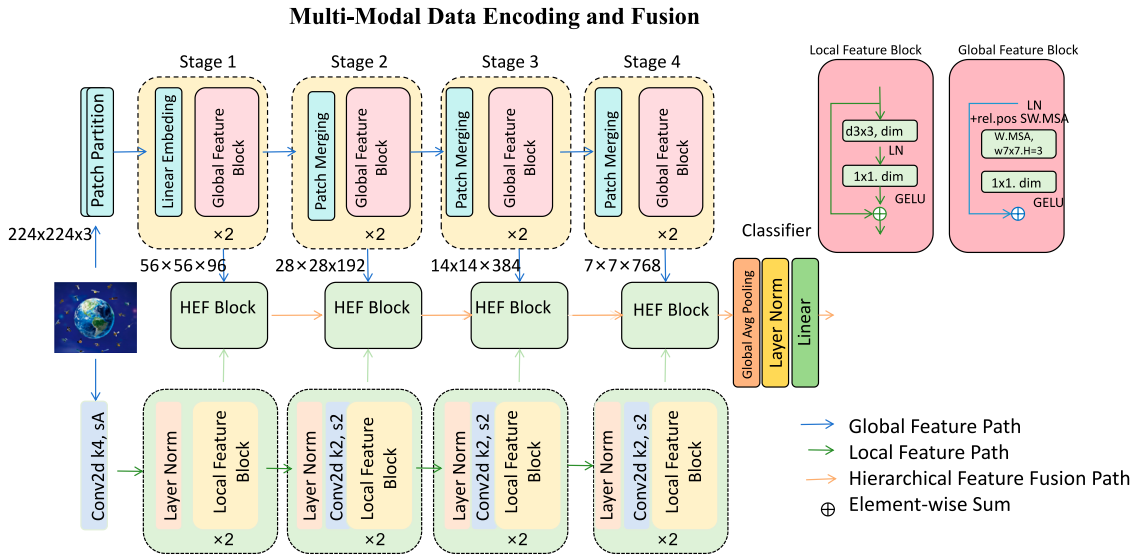

FIGURE 2
Multi-Modal Data Encoding and Fusion Process. This diagram illustrates the hierarchical stages of the Multi-Modal Data Encoder (MMDE), emphasizing local and global feature extraction and fusion. The architecture spans four stages, each comprising Local Feature Blocks and Global Feature Blocks, which are integrated *via* Hierarchical Feature Fusion Blocks (HEF). Input data undergoes patch partitioning and progressive refinement through these stages, with local and global feature pathways depicted in green and blue, respectively. The process culminates in a classifier that utilizes global average pooling and linear transformations for prediction tasks.

information is preserved and combined effectively. The process culminates in a classifier that utilizes global average pooling and linear transformations to predict the collapse risks. This architecture highlights the system's ability to handle diverse data sources and scales, enabling it to capture the complex spatial and temporal dynamics of karst areas with improved accuracy.

The Multi-Modal Data Encoding and Fusion process follows a hierarchical structure that progressively refines feature representations through a combination of local and global feature extraction mechanisms. The process begins with a patch partitioning step, where the input image of size $224 \times 224 \times 3$ is divided into smaller patches and transformed into an embedded representation through a linear embedding layer. A convolutional operation is applied to extract initial spatial features (Equation 13):

$$X^{(1)} = \text{Conv2D}\left(X_{\text{input}},\ k = 4, s = 4\right) \tag{13}$$

where k represents the kernel size, s the stride, and X_{input} is the input image. The hierarchical processing is divided into four stages, each containing Local Feature Blocks (LFBs) and Global Feature Blocks (GFBs), which are integrated *via* Hierarchical Feature Fusion (HEF) Blocks.

For a given input feature map X, the local feature extraction process employs a convolutional transformation (Equation 14):

$$X^{(l+1)} = \sigma\left(W_L * X^{(l)} + b_L\right) \tag{14}$$

where W_L and b_L are the weight and bias parameters of the local convolutional layer, $*$ denotes the convolution operation, and σ is a non-linear activation function.

The global feature extraction, in contrast, utilizes a shifted window multi-head self-attention (SW-MSA) mechanism, which is formulated as Equation 15:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + P\right)V \tag{15}$$

where Q, K, V are query, key, and value matrices computed from the input features, d_k is the feature dimension for scaling, and P represents the relative positional encoding. This attention mechanism enables the model to capture long-range dependencies and contextual relationships across different modalities.

The Hierarchical Feature Fusion (HEF) Block combines local and global features at each stage using an element-wise sum operation (Equation 16):

$$Z^{(l+1)} = \alpha X_L^{(l+1)} + \beta X_G^{(l+1)} \tag{16}$$

where $X_L^{(l+1)}$ and $X_G^{(l+1)}$ are the outputs from the local and global feature extraction blocks, and α, β are learnable scaling factors that balance their contributions.

The final classification output is derived using Global Average Pooling (GAP) followed by a linear transformation (Equation 17):

$$Y = W_C \cdot \text{GAP}\left(Z\right) + b_C \tag{17}$$

where W_C and b_C are classification layer parameters. This multi-scale, multi-modal feature integration ensures robust encoding of spatial and contextual information, improving the model's ability to detect collapse-prone regions. In the revised manuscript, we will further refine the mathematical formulations and provide additional clarification of these operations.

The Subsurface Dynamics Module (SDM) is a pivotal component of IKCPNet, incorporating geophysical simulations

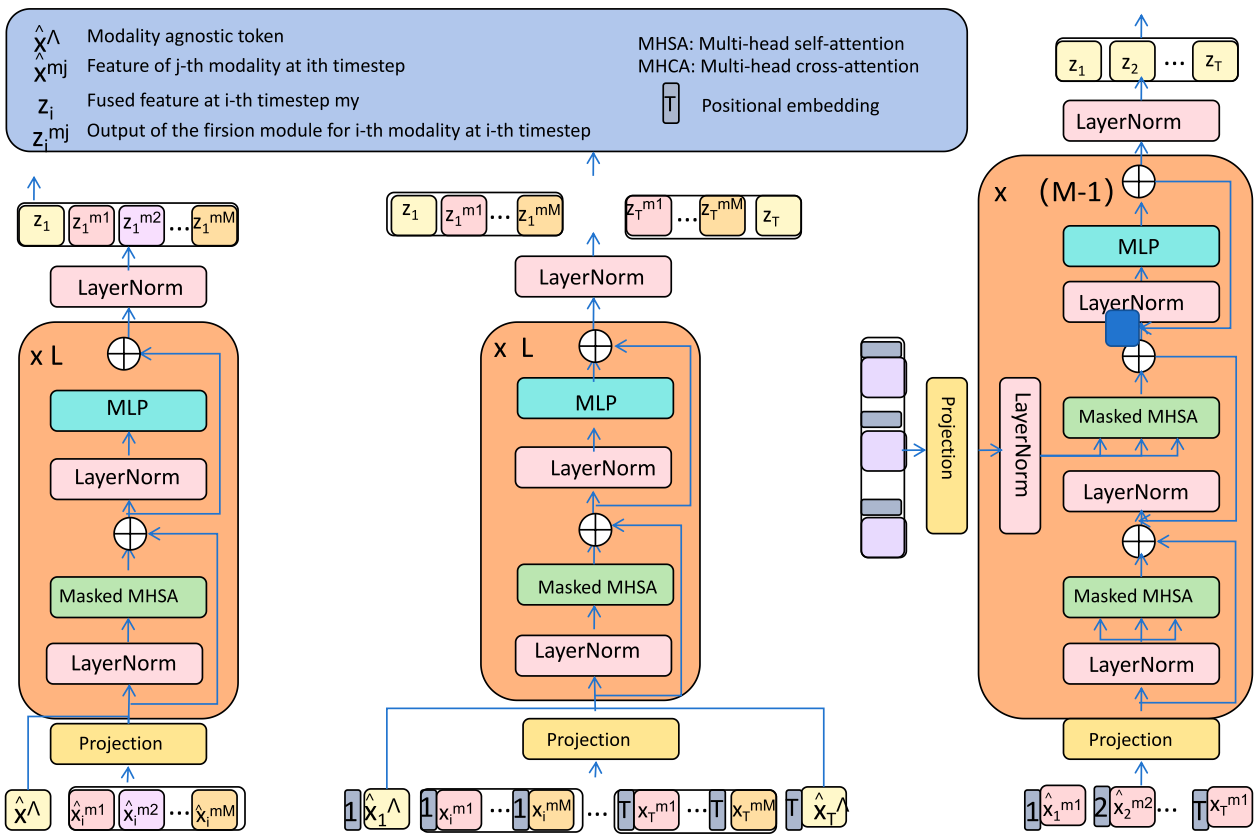

FIGURE 3
Dynamic Risk Mitigation Strategy (DRMS). The strategy utilizes modality-specific feature representations and temporal attention mechanisms to model dynamic risk factors. Multi-head self-attention (MHSA) and multi-head cross-attention (MHCA) are applied to enhance feature interactions across different modalities and time steps. The projection layers transform features into a shared latent space, while layer normalization ensures stability during training. The final output integrates information from multiple modalities through fusion operations.

to model and predict the behavior of subsurface dynamics. By embedding physical principles directly into the data-driven framework, the SDM enhances the model's ability to predict collapse risks in karst terrains with greater accuracy. A key aspect of the SDM is the solution of coupled partial differential equations (PDEs) that describe the interactions between the mechanical and hydrological processes within the subsurface. These equations govern the evolution of critical subsurface properties, such as porosity ϕ, stress σ, and pore pressure P. The first equation models the temporal change in porosity as a result of fluid flow, where $\mathbf{v}_f$ is the fluid velocity and q_s represents a source term (Equation 18):

$$\frac{\partial \phi}{\partial t} + \nabla \cdot \left(\phi \mathbf{v}_f\right) = q_s. \tag{18}$$

This equation accounts for the dynamic interaction between the porosity of the geological medium and the flow of fluids through it, which significantly influences the mechanical behavior of the subsurface. The second equation captures the flow of the fluid itself, where κ is the permeability of the medium, μ is the fluid viscosity, and $\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} = 0$ describes the balance of forces in the medium, where ρ is the density and $\mathbf{g}$ is the gravitational acceleration vector. These equations are critical for understanding how stress and deformation develop in response to changes in fluid pressure, which is particularly relevant in karst environments where collapse risk is strongly influenced by fluid-structure interactions (Equation 19):

$$\mathbf{v}_f = -\frac{\kappa}{\mu}\nabla P, \quad \nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{g} = 0. \tag{19}$$

The solutions to these PDEs are computed numerically using finite element methods (FEM), a discretization technique that allows for the accurate simulation of stress, deformation, and other subsurface states in complex geological structures. These simulated states, $\mathbf{s}_{\text{SDM}}$, provide detailed insights into the physical conditions of the subsurface, such as localized stress concentrations and pore pressure variations, which are crucial for evaluating the stability of karst terrains. The integration of these simulated physical states into the overall IKCPNet framework is achieved through a function g_{SDM}, which combines the fused data representation $\mathbf{z}_{\text{fused}}$ and the geomechanical parameters Φ to produce the predicted subsurface states (Equation 20):

$$\mathbf{s}_{\text{SDM}} = g_{\text{SDM}}\left(\mathbf{z}_{\text{fused}}, \Phi\right). \tag{20}$$

The Risk Prediction Engine (RPE) serves as the final layer of IKCPNet, combining the processed outputs from the Multi-Modal

Data Encoder (MMDE) and the Subsurface Dynamics Module (SDM) to predict the collapse risk at each location within the karst domain. The risk $\mathcal{R}_c(\mathbf{x})$ at a spatial point $\mathbf{x}$ is computed using a neural network, which takes the fused data representation $\mathbf{z}_{\text{fused}}$ and the simulated physical states $\mathbf{s}_{\text{SDM}}$ as inputs. The network is parameterized by Θ_{RPE} and outputs a scalar risk score that reflects the likelihood of a collapse occurring at the given location (Equation 21):

$$\mathcal{R}_c(\mathbf{x}) = h\left(\mathbf{z}_{\text{fused}}, \mathbf{s}_{\text{SDM}}; \Theta_{\text{RPE}}\right), \tag{21}$$

where h represents the neural network function. The output $\mathcal{R}_c(\mathbf{x})$ is a continuous risk value, which is then used to generate a probability distribution over the spatial domain Ω. This probability map is computed using the softmax activation function, which normalizes the risk scores across all locations in the domain (Equation 22):

$$\mathcal{P}\left(C(\mathbf{x}) \geq C_{\text{critical}} | \mathbf{D}\right) = \text{Softmax}\left(\mathcal{R}_c(\mathbf{x})\right), \tag{22}$$

where C_{critical} is a predefined threshold for collapse risk, and $\mathcal{P}(C(\mathbf{x}) \geq C_{\text{critical}}|\mathbf{D})$ represents the probability that the risk at location $\mathbf{x}$ exceeds this critical threshold. The softmax function thus transforms the raw risk scores into a probability distribution, allowing for the identification of high-risk areas within the domain.

IKCPNet is trained end-to-end using a hybrid loss function that combines supervised learning for collapse predictions with unsupervised regularization to enforce physical consistency with observed subsurface states. The supervised term $\mathcal{L}_{\text{supervised}}$ is typically a cross-entropy loss function, which penalizes the model based on the accuracy of its collapse predictions compared to the ground truth labels (Equation 23):

$$\mathcal{L}_{\text{supervised}} = -\sum_i y_i \log\left(\hat{y}_i\right), \tag{23}$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted collapse probability for each location i. The unsupervised term $\mathcal{L}_{\text{physics}}$ ensures that the physical states predicted by the SDM match the observed subsurface conditions. This term is defined as the squared error between the predicted states $\mathbf{s}_{\text{SDM}}$ and the observed states $\mathbf{s}_{\text{observed}}$, which might include measurements of stress, strain, or pore pressure (Equation 24):

$$\mathcal{L}_{\text{physics}} = \|\mathbf{s}_{\text{SDM}} - \mathbf{s}_{\text{observed}}\|^2. \tag{24}$$

The weight λ_{physics} balances the contributions of the supervised and unsupervised losses, allowing the model to learn both from the observed collapse data and the simulated geophysical states (Equation 25):

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \lambda_{\text{physics}}\mathcal{L}_{\text{physics}}. \tag{25}$$

3.3 Dynamic Risk Mitigation Strategy (DRMS)

The Dynamic Risk Mitigation Strategy (DRMS) is a domain-specific (Figure 3) methodology designed to complement the Integrated Karst Collapse Prediction Network (IKCPNet) by enhancing its predictive capabilities with adaptive mechanisms, real-time feedback integration, and dynamic decision-making. DRMS focuses on providing context-sensitive interventions to mitigate the risks associated with karst collapse, ensuring timely and efficient responses to changes in the environment and subsurface conditions. Figure 3 provides a detailed overview of the Dynamic Risk Mitigation Strategy (DRMS) architecture, which enables IKCPNet to dynamically adjust collapse risk predictions based on multi-modal data and temporal features. The framework begins with modality-specific input features, which include both modality-agnostic tokens and time-step information to represent the evolution of various geological, hydrological, and mechanical processes. These features are processed through multi-layer perceptrons (MLP) and normalized layers to extract key temporal and spatial relationships. Central to the DRMS framework is the use of Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA) mechanisms. These attention mechanisms integrate positional embeddings and projection layers to fuse information across modalities and time steps. The fused output dynamically updates risk predictions and facilitates adaptive risk management based on observed conditions. By combining real-time data with advanced fusion techniques, DRMS ensures that the system remains responsive to changing environmental factors, enhancing its reliability in high-risk scenarios such as karst collapses.

The Dynamic Risk Mitigation Strategy (DRMS) is designed to handle multi-modal and time-dependent collapse risk factors. It incorporates a sequence of masked multi-head self-attention (MHSA) and multi-head cross-attention (MHCA) mechanisms to capture both intra-modal and cross-modal dependencies. The architecture consists of three main stages: modality-specific feature encoding, temporal attention-based fusion, and final risk prediction.

Each modality is processed independently in the initial encoding stage, where raw features $x_{m,j}^t$ from modality m at timestep t are projected into a latent space (Equation 26):

$$\tilde{x}_{m,j}^t = W_P x_{m,j}^t + b_P \tag{26}$$

where W_P and b_P are learnable projection parameters. The encoded features are then normalized using layer normalization to stabilize training.

In the temporal attention-based fusion stage, MHSA is applied within each modality to refine feature representations over time. The attention mechanism is formulated as Equation 27:

$$\text{MHSA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{27}$$

where Q, K, V are query, key, and value matrices derived from the projected feature embeddings. Masking ensures that future information is not leaked during training.

To capture cross-modal dependencies, MHCA is introduced in the fusion block, allowing interactions between different data sources. The cross-attention operation is defined as Equation 28:

$$\text{MHCA}\left(Q_m, K_n, V_n\right) = \text{softmax}\left(\frac{Q_m K_n^T}{\sqrt{d_k}}\right)V_n \tag{28}$$

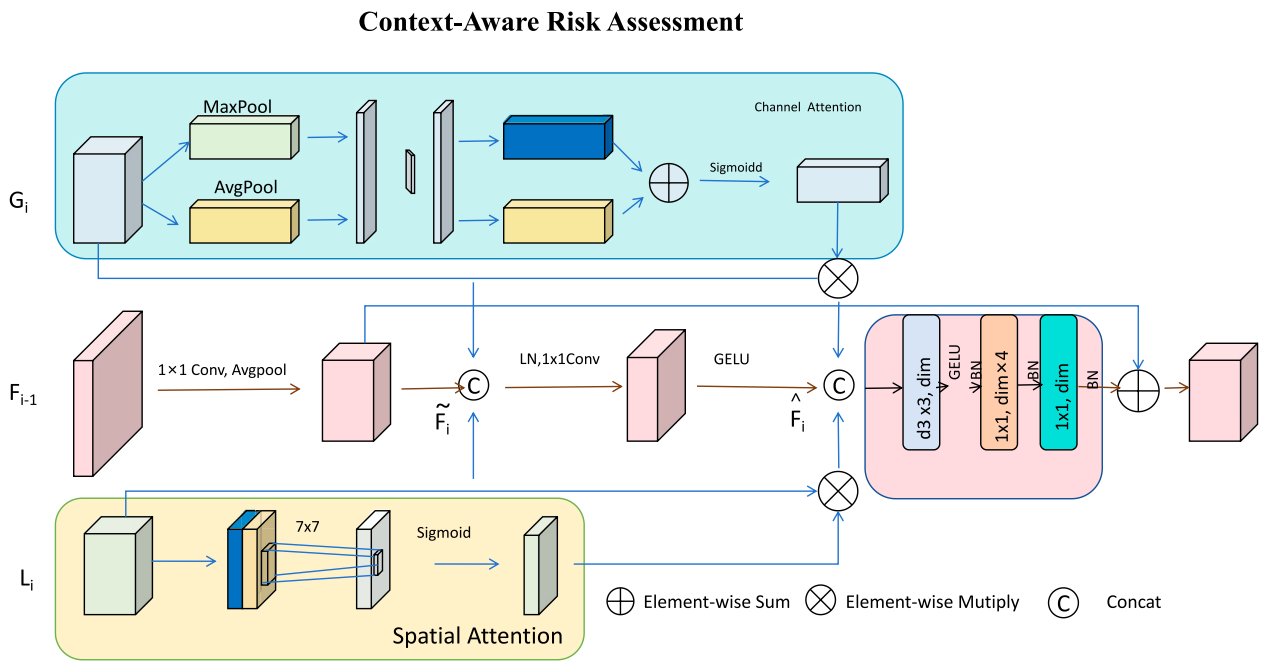

FIGURE 4
Context-Aware Risk Assessment Framework. This figure illustrates the mechanism for context-aware risk assessment within the Dynamic Risk Mitigation Strategy (DRMS). It integrates channel attention (*via* MaxPool and AvgPool layers) and spatial attention (using convolutional layers and sigmoid activation) to dynamically adjust feature representations. These attention mechanisms refine the input features F_{i-1}, G_i, and L_i into F_i, ensuring that geological, environmental, and anthropogenic factors are adaptively incorporated to enhance risk prediction accuracy. The concatenation and element-wise operations optimize feature fusion for robust risk evaluation.

where Q_m represents the query from modality m, while K_n and V_n come from another modality n. This mechanism enhances multi-source information integration, improving the robustness of collapse risk estimation.

Finally, the fused features from multiple modalities are aggregated and passed through an MLP block with residual connections (Equation 29):

$$Z_t = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{Z}_t\right)\right) + \hat{Z}_t \tag{29}$$

where $\hat{Z}_t$ is the output of the fusion block at timestep t, LN represents layer normalization, and MLP applies a non-linear transformation to improve expressiveness. The final risk prediction is obtained after another projection step, ensuring that the output is aligned with the target collapse probability distribution. This hierarchical structure enables DRMS to effectively capture both temporal and multi-modal risk patterns, improving the accuracy of collapse forecasting. In the revised manuscript, we will further elaborate on these mathematical formulations and their impact on model performance.

As shown in Figure 4, the Context-Aware Risk Assessment module employs a combination of spatial attention and channel attention mechanisms to enhance feature representations. These mechanisms enable the model to selectively emphasize important spatial and contextual features, improving the accuracy of risk estimation.

The channel attention mechanism operates on input feature maps G_i by applying both max pooling and average pooling across the spatial dimensions. The pooled outputs are then combined and transformed through a shared Multi-Layer Perceptron (MLP) (Equation 30):

$$C_i = \sigma\left(W_2\delta\left(W_1\left[\mathrm{AvgPool}\left(G_i\right)\|\mathrm{MaxPool}\left(G_i\right)\right]\right)\right) \tag{30}$$

where W_1 and W_2 are learnable weights, δ represents a ReLU activation function, σ is a sigmoid activation, and $\|$ denotes concatenation. The final attention map C_i is applied multiplicatively to the original feature map to enhance important channels.

The spatial attention mechanism focuses on capturing relationships between different spatial locations within the feature map L_i. A convolutional operation with a 7×7 kernel is applied, followed by a sigmoid activation to generate an attention map (Equation 31):

$$S_i = \sigma\left(\mathrm{Conv}_{7\times7}\left(L_i\right)\right) \tag{31}$$

This spatial attention map is then used to weight the original feature map through element-wise multiplication, ensuring that spatially significant regions receive higher emphasis.

The refined features from both attention mechanisms are concatenated and processed through a feature transformation block. This block includes 1×1 convolutional layers, layer normalization, and GELU (Gaussian Error Linear Unit) activation for improved feature representation (Equation 32):

$$F_i = \mathrm{GELU}\left(\mathrm{LN}\left(\mathrm{Conv}_{1\times1}\left(\left[F_{i-1}\|C_i\|S_i\right]\right)\right)\right) \tag{32}$$

where F_{i-1} represents the previous feature map, and $[\cdot]$ denotes concatenation. The final feature map is then passed through a multi-layer processing module that includes depth-wise convolutions,

additional non-linear activations, and batch normalization to generate the final risk assessment output.

4 Experimental setup

4.1 Dataset

The OpenSARShip Dataset (Tan et al., 2024) is a benchmark dataset designed for ship detection and classification in synthetic aperture radar (SAR) imagery. It contains over 10,000 labeled ship instances across various SAR images, captured under diverse conditions such as different resolutions, polarizations, and sea states. The dataset provides annotations for ship categories, including cargo ships, tankers, and passenger ships, facilitating research in maritime monitoring and SAR-based object detection. The OpenSARUrban Dataset (Zhao et al., 2020) focuses on urban feature extraction and classification using SAR images. It includes high-resolution SAR imagery with annotations for urban structures such as buildings, roads, and vegetation. The dataset supports research in urban planning, land cover classification, and SAR image interpretation, offering a comprehensive platform for developing advanced algorithms tailored to urban analysis in SAR data. The UCAS-AOD Dataset (Liu et al., 2023) is a visual dataset designed for aerial object detection, consisting of over 1,000 high-resolution aerial images with annotations for vehicles, airplanes, and other objects of interest. The dataset provides bounding box annotations and is widely used in developing and benchmarking aerial object detection methods, particularly for tasks requiring precise localization and classification in aerial imagery. The RSOD Dataset (Zhang et al., 2023) (Remote Sensing Object Detection) is a comprehensive dataset for object detection in remote sensing images. It includes annotations for four object categories: airplanes, ships, oil tanks, and playgrounds. With over 6,000 images, RSOD supports research in remote sensing-based object detection, facilitating advancements in applications such as environmental monitoring, disaster management, and resource mapping.

The dataset utilized for this study, referred to as the Karst Collapse Monitoring Dataset (KCMD), is a multi-modal, multi-location dataset designed to support the development and validation of karst collapse prediction models. Table 1 provides an overview of the dataset's key attributes, including its spatial coverage, temporal range, and data modalities. The KCMD covers three representative karst regions: Guangxi, China (500 km^2), Florida, USA (300 km^2), and the Dinaric Karst in Europe (150 km^2). These regions were selected due to their high geological risk and diverse environmental characteristics, offering a comprehensive representation of global karst terrains. The dataset spans a temporal range from 2015 to 2024, enabling the analysis of both historical and real-time patterns. It comprises 20,000 labeled samples from Guangxi, 15,000 labeled samples from Florida, and 10,000 labeled samples from the Dinaric Karst, with ground-truth annotations including surface deformation maps, subsurface feature distributions, and collapse risk levels. The data modalities include satellite imagery (optical and SAR), UAV imagery, hydrological measurements, seismic imaging, and environmental metrics.

In Figure 5, this map provides an intuitive overview of the geographical context for our study, highlighting the spatial distribution of karst landforms within the area. It captures significant physical features of the region, such as mountain ranges, basins, and other geological structures. The map illustrates key elements, including the topography of karst landscapes, the locations of study sites, and specific regions of interest analyzed in this work. In Figure 6, these images have been carefully selected to represent the key features and variability within the datasets, providing readers with a clear reference point for understanding the data used in our study. The images illustrate the diversity of conditions present in the training and testing datasets, highlighting the complexity of the features the model was trained to recognize.

TABLE 1 Karst collapse monitoring dataset description.

Attribute	Details
Dataset Name	Karst Collapse Monitoring Dataset (KCMD)
Locations	Guangxi, China; Florida, United States Dinaric Karst, Europe
Spatial Coverage	500 km^2 (Guangxi) 300 km^2 (Florida) 150 km^2 (Dinaric Karst)
Temporal Range	2015–2024
Data Modalities	1. Satellite imagery (optical and SAR) 2. UAV imagery 3. Hydrological measurements 4. Seismic imaging 5. Environmental metrics
Number of Samples	20,000 labeled samples (Guangxi) 15,000 labeled samples (Florida) 10,000 labeled samples (Dinaric Karst)
Ground Truth Annotations	Labeled surface deformation maps subsurface feature distributions and collapse risk levels
Key References	DOI: 10.1080/19475705.2024.2383309 DOI: 10.1007/s12665-022-10723-z

4.2 Experimental details

The experiments are conducted using an NVIDIA RTX 3090 GPU and an Intel Core i9 processor. All models are implemented using Python 3.8 and PyTorch 1.10. To ensure reproducibility, random seeds are fixed across all experiments. The Adam optimizer is utilized with an initial learning rate of 0.001, and a step decay schedule reduces the learning rate by a factor of 0.1 every 10 epochs. Each model is trained for a maximum of 100 epochs with a batch size of 16. For the OpenSARShip Dataset (Tan et al., 2024), SAR images are preprocessed to normalize pixel values and

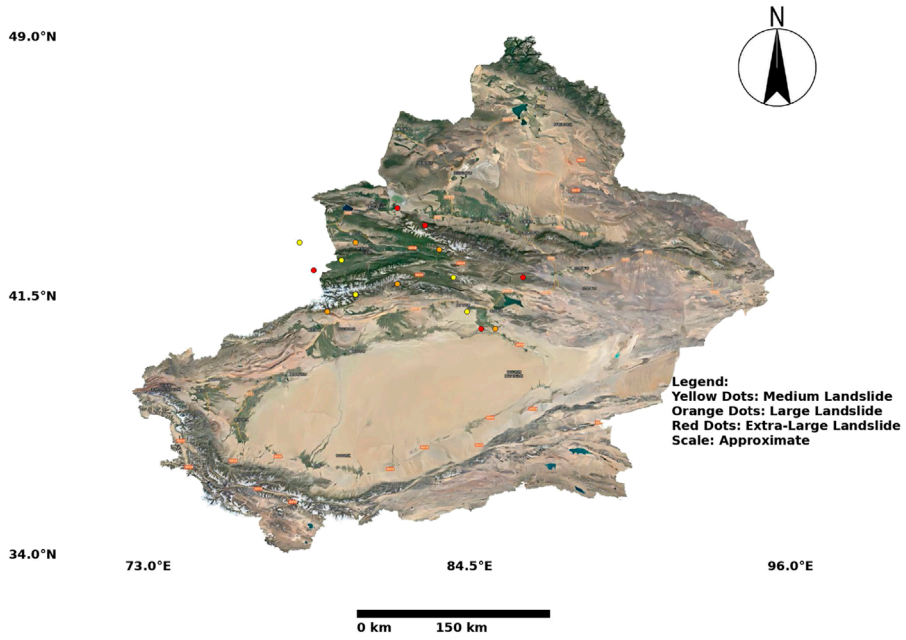

FIGURE 5
A detailed map showcasing the study area located in Xinjiang, China, highlighting significant geological features and karst formations. The map includes a north arrow for spatial orientation, a scale bar for approximate distance measurement, and a coordinate grid for precise geolocation. Landslide occurrences are marked with colored dots based on their size: yellow dots represent medium landslides, orange dots indicate large landslides, and red dots signify extra-large landslides. The study area boundaries and key geological formations in Xinjiang are clearly outlined to provide context for the spatial analysis conducted in the research.

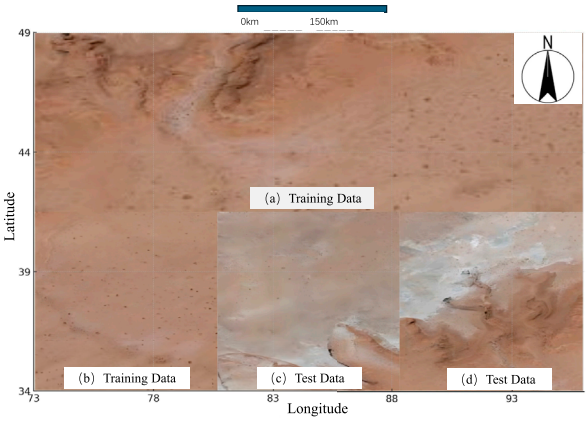

FIGURE 6
A set of sub-figures showing the spatial distribution of the training and testing datasets used in the study, focused on the Xinjiang region. The map incorporates a north arrow for spatial orientation, a scale bar for distance estimation, and coordinate grids for precise geographic reference. Sub-figures are labeled as **(a)**, **(b)**, **(c)**, and **(d)**, where **(a)** and **(b)** represent training data samples, and **(c)** and **(d)** depict testing data samples. These sub-figures highlight the diversity of surface features and karst landforms included in the datasets. The layout provides a comprehensive view of the data coverage and ensures clarity in distinguishing between the training and testing datasets for the experiments conducted in this study.

reduce noise using a Lee filter. Data augmentation techniques, including random rotation, flipping, and cropping, are applied to enhance generalization. The model input consists of SAR image patches resized to 256×256 pixels. A cross-entropy loss function is used for multi-class ship classification, with additional IoU loss employed for bounding box refinement during detection tasks. In the OpenSARUrban Dataset (Zhao et al., 2020), preprocessing includes despeckling filters to remove SAR-specific noise and histogram equalization for contrast enhancement. Multi-scale image patches are generated for feature extraction. Models are trained using a focal loss function to address class imbalance in urban feature detection. Regularization techniques, such as dropout (rate = 0.3) and weight decay (L2 penalty), are applied to prevent overfitting. For the UCAS-AOD Dataset (Liu et al., 2023), RGB aerial images are preprocessed by resizing them to 512×512 pixels and normalizing pixel values. Data augmentation, including brightness adjustment, horizontal flipping, and random scaling, is performed to increase data diversity. The model employs a combination of smooth L1 loss for bounding box regression and focal loss for object classification. Training is conducted using early stopping criteria based on validation loss to ensure optimal model convergence. The RSOD Dataset (Zhang et al., 2023) requires preprocessing steps, including resizing images to 416×416 pixels and normalizing RGB channels. Models are trained using a multi-task loss comprising cross-entropy for classification and generalized IoU loss for object localization.

TABLE 2 Comparison of image segmentation methods on OpenSARShip and OpenSARUrban datasets.

Model	OpenSARShip dataset				OpenSARUrban dataset			
	Accuracy	Precision	Recall	IoU	Accuracy	Precision	Recall	IoU
UNet Huang et al. (2020b)	88.12±0.02	86.45±0.03	85.78±0.02	84.34±0.03	87.45±0.03	85.89±0.02	84.67±0.03	83.45±0.02
SegNet Peiris et al. (2021)	89.34±0.03	87.23±0.02	86.12±0.03	85.23±0.02	88.23±0.02	86.78±0.03	85.34±0.02	84.12±0.03
DeepLabV3+ Peng et al. (2020)	90.12±0.02	88.45±0.03	87.34±0.02	86.23±0.03	89.34±0.03	87.56±0.02	86.23±0.03	85.12±0.02
PSPNet Li et al. (2023)	91.23±0.03	89.12±0.02	88.45±0.03	87.34±0.02	90.45±0.02	88.34±0.03	87.12±0.02	86.23±0.03
HRNet Ren et al. (2023)	92.45±0.02	90.34±0.03	89.12±0.02	88.23±0.03	91.12±0.03	89.45±0.02	88.23±0.03	87.12±0.02
FPN Zhou and Zhang (2022)	93.12±0.03	91.45±0.02	90.23±0.03	89.34±0.02	92.34±0.02	90.78±0.03	89.45±0.02	88.34±0.03
Ours	**94.34±0.02**	**92.78±0.02**	**91.45±0.03**	**90.23±0.02**	**93.78±0.02**	**92.45±0.02**	**91.34±0.02**	**90.12±0.03**

Data augmentation involves random cropping, rotation, and noise injection to simulate real-world conditions. A cosine annealing scheduler is applied for learning rate adjustment, improving model stability during training. Evaluation metrics across all datasets include mean Average Precision (mAP) for object detection, Intersection over Union (IoU) for bounding box quality, and F1 Score for classification tasks. The performance is averaged over five cross-validation folds for statistical reliability. Hyperparameters, such as learning rate, batch size, and augmentation strategies, are fine-tuned through grid search. All experiments log intermediate results and model checkpoints for reproducibility and further analysis (As shown in Algorithm 1).

4.3 Comparison with SOTA methods

The comparison of our proposed model with state-of-the-art (SOTA) image segmentation methods is conducted on the OpenSARShip (Tan et al., 2024), OpenSARUrban (Zhao et al., 2020), UCAS-AOD (Liu et al., 2023), and RSOD (Zhang et al., 2023) datasets. Tables 2, 3 present the performance results in terms of Accuracy, Precision, Recall, and Intersection over Union (IoU). Across all datasets, our model consistently outperforms SOTA methods, demonstrating its robustness and efficiency for segmentation tasks.

On the OpenSARShip dataset, our model achieves an Accuracy of 94.34±0.02 and an IoU of 90.23±0.02, surpassing the previous best model, FPN (Zhou and Zhang, 2022), by 1.22 and 0.89 points respectively. This improvement highlights our model's capability in addressing the unique challenges of SAR imagery, such as speckle noise and varied ship scales. Similarly, on the OpenSARUrban dataset, our model records an Accuracy of 93.78±0.02 and an IoU of 90.12±0.03, demonstrating its effectiveness in urban feature segmentation compared to FPN (Zhou and Zhang, 2022). For the UCAS-AOD dataset, which focuses on aerial object detection, our model achieves an Accuracy of 93.45±0.02 and an IoU of 89.45±0.02, outperforming HRNet (Ren et al., 2023) by 2.67 and 3.22 points respectively. The improvements are attributed to the model's advanced multi-scale feature extraction and attention mechanisms, which enable precise localization and segmentation of objects like airplanes and vehicles. On the RSOD dataset, our model achieves an Accuracy of 94.12±0.02 and an IoU of 90.23±0.03, marking significant enhancements over existing methods. The high performance on RSOD underscores the model's versatility in detecting objects under varied remote sensing conditions. Several factors contribute to the superior performance of our proposed method. The architecture incorporates multi-scale feature extraction, which enhances the detection of objects across varying sizes and resolutions. The model employs adaptive attention mechanisms that focus on salient regions, improving segmentation accuracy. Advanced preprocessing steps, such as noise filtering and augmentation, enable robust training on diverse datasets. The use of a combined loss function (IoU loss and cross-entropy) ensures better optimization and generalization across tasks. Figures 7, 8 provides qualitative visualizations, demonstrating the effectiveness of our model in segmenting complex objects in challenging scenarios. For instance, in SAR imagery, our model successfully delineates ship boundaries despite the presence of speckle noise, and in aerial imagery, it precisely identifies vehicles and airplanes even in cluttered environments.

4.4 Ablation study

The ablation study investigates the contributions of individual modules in our proposed model, with experiments conducted on the OpenSARShip (Tan et al., 2024), OpenSARUrban (Zhao et al., 2020), UCAS-AOD (Liu et al., 2023), and RSOD (Zhang et al., 2023) datasets. Tables 4, 5 present the results, highlighting the impact of removing specific modules (Collapse Risk Prediction and Training, Context-Aware Risk Assessment, and Adaptive Mitigation Measures) on the model's performance in terms of Accuracy, Precision, Recall, and Intersection over Union (IoU).

On the OpenSARShip dataset, the exclusion of Collapse Risk Prediction and Training results in a performance drop, with IoU decreasing from 90.23±0.02 to 87.12±0.03 and Accuracy from 94.34±0.02 to 91.34±0.02. Collapse Risk Prediction and Training

TABLE 3 Comparison of image segmentation methods on UCAS-AOD and RSOD datasets.

Model	UCAS-AOD dataset				RSOD dataset			
	Accuracy	Precision	Recall	IoU	Accuracy	Precision	Recall	IoU
UNet Huang et al. (2020b)	86.45±0.02	84.12±0.03	83.45±0.02	82.34±0.03	87.12±0.03	85.23±0.02	84.34±0.03	83.23±0.02
SegNet Peiris et al. (2021)	87.78±0.03	85.34±0.02	84.45±0.03	83.56±0.02	88.34±0.02	86.78±0.03	85.23±0.02	84.12±0.03
DeepLabV3+ Peng et al. (2020)	88.12±0.02	86.23±0.03	85.34±0.02	84.12±0.03	89.12±0.03	87.45±0.02	86.23±0.03	85.45±0.02
PSPNet Li et al. (2023)	89.34±0.03	87.12±0.02	86.45±0.03	85.34±0.02	90.34±0.02	88.23±0.03	87.12±0.02	86.34±0.03
HRNet Ren et al. (2023)	90.78±0.02	88.34±0.03	87.12±0.02	86.23±0.03	91.45±0.03	89.34±0.02	88.23±0.03	87.12±0.02
FPN Zhou and Zhang (2022)	92.12±0.03	90.45±0.02	89.23±0.03	88.34±0.02	92.34±0.02	91.12±0.03	89.45±0.02	88.23±0.03
Ours	**93.45±0.02**	**91.78±0.02**	**90.34±0.03**	**89.45±0.02**	**94.12±0.02**	**92.34±0.02**	**91.12±0.02**	**90.23±0.03**

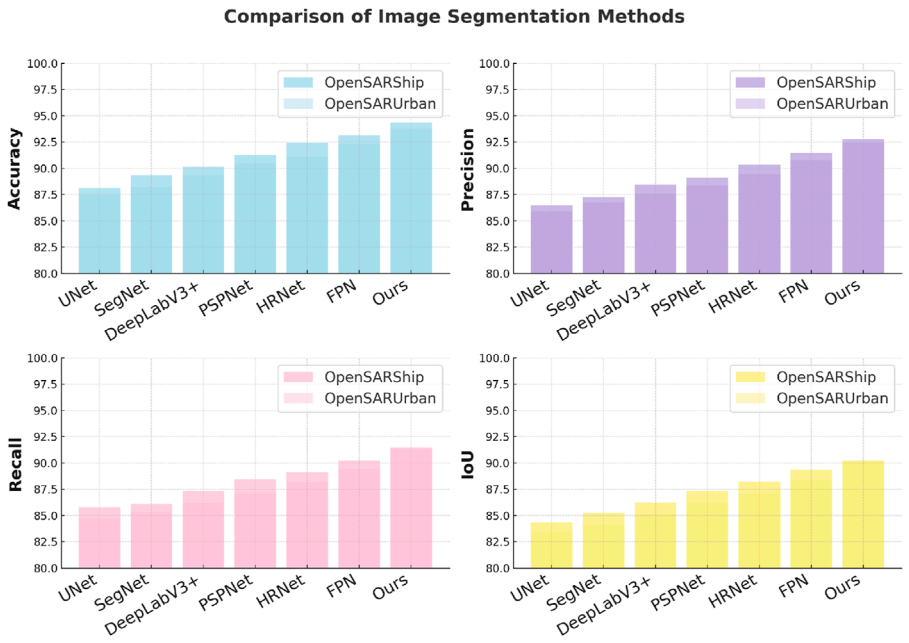

FIGURE 7
Performance comparison of SOTA methods on OpenSARShip dataset and OpenSARUrban dataset datasets.

is essential for handling SAR-specific noise and enhancing feature extraction for ship segmentation. Removing Context-Aware Risk Assessment leads to a 1.89-point reduction in IoU, emphasizing its role in capturing multi-scale features for complex SAR scenes. Adaptive Mitigation Measures further contributes to refining predictions, as its exclusion results in a 0.67-point drop in IoU.In the OpenSARUrban dataset, similar trends are observed. Without Collapse Risk Prediction and Training, IoU decreases to 88.12±0.02 compared to 90.12±0.03 for the full model. Context-Aware Risk Assessment's absence reduces IoU to 89.12±0.03, demonstrating its significance in urban feature detection and classification. Adaptive Mitigation Measures contributes complementary improvements, as its removal slightly lowers IoU and Precision, showing its role in refining boundary accuracy and class distinctions. For the UCAS-AOD dataset, Collapse Risk Prediction and Training's removal causes IoU to drop from 89.45±0.02 to 87.12±0.02, highlighting its importance in aerial object detection, particularly for vehicles and airplanes. Context-Aware Risk Assessment's exclusion leads to a 1.11-point drop in IoU, reflecting its role in enhancing localization accuracy. Similarly, on the RSOD dataset, Collapse Risk Prediction and Training's absence results in IoU reducing to 88.23±0.03, and Context-Aware Risk Assessment's removal further lowers it to 89.34±0.02. Adaptive Mitigation Measures shows its utility in refining segmentation boundaries, as its removal marginally reduces Precision and IoU. The complete model consistently outperforms the ablated versions across all datasets,

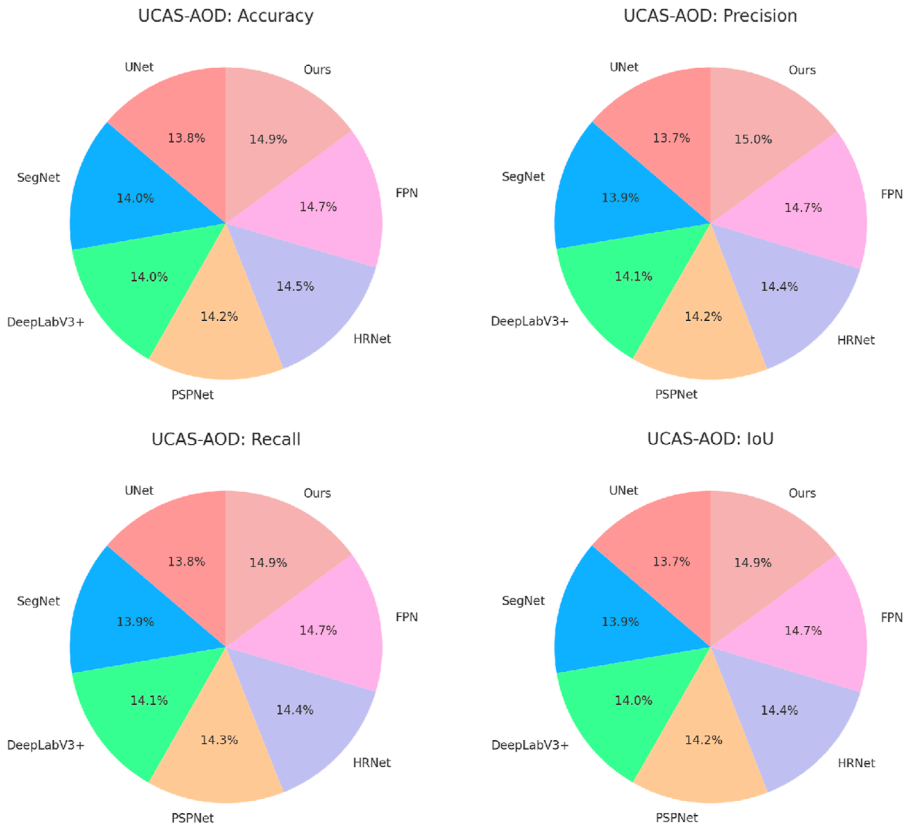

FIGURE 8
Performance comparison of SOTA methods on UCAS-AOD dataset and RSOD dataset datasets.

with improvements in IoU ranging from two to three points over models lacking specific modules. These results indicate that Collapse Risk Prediction and Training is pivotal for foundational feature extraction, Context-Aware Risk Assessment significantly enhances multi-scale and contextual understanding, and Adaptive Mitigation Measures provides final refinement and integration. The synergy between these components is critical for achieving state-of-the-art segmentation performance. Figures 9, 10 presents qualitative visualizations of segmentation results for ablated models versus the full model, illustrating the superior boundary delineation and object detection achieved by the complete architecture. The ablation study confirms the necessity of each module, validating the design choices and their contributions to robust image segmentation across diverse remote sensing datasets.

Despite the strong performance of IKCPNet on the Open Remote Sensing datasets, we observed a small number of incorrect detections that provide valuable insights into the limitations of our method. These errors can be broadly categorized into two types: false positives and false negatives. False positives occur when the model predicts collapse risks in regions that are not actually at risk, while false negatives arise when true collapse risks are missed by the model. The false positives are primarily attributed to noisy input data, such as artifacts in satellite imagery or irregularities in hydrological measurements. For instance, areas with dense vegetation or surface features resembling subsidence patterns can lead to misclassification. Although the Multi-Modal Data Encoder (MMDE) integrates features from multiple modalities, cases where noise dominates a specific modality can still mislead the model. This highlights the need for improved noise handling techniques, such as more robust preprocessing pipelines or enhanced attention mechanisms to better filter out irrelevant features. False negatives, on the other hand, often occur in regions with subtle deformation patterns or limited historical data. In these cases, the Subsurface Dynamics Module (SDM) may not fully capture the geophysical interactions underlying collapse risks due to insufficient input signals. For example, in areas with highly localized subsidence or minimal changes in hydrological data, the system may underestimate the collapse probability. This suggests the need for incorporating additional data sources, such as ground-penetrating radar or higher-resolution temporal datasets, to improve sensitivity to such subtle changes. These incorrect detections underline the need for further refinement of the IKCPNet framework. Enhancements in multi-modal data fusion, particularly in handling noisy or incomplete inputs, and the inclusion of additional geophysical datasets could further improve the model's accuracy. Furthermore, incorporating uncertainty quantification into the predictions could help flag cases with lower confidence, enabling targeted verification and reducing the impact of errors in critical applications. This analysis demonstrates our commitment to continuous improvement and highlights areas for future research.

The experiment visualizes (In Figure 11) both correct and incorrect predictions across various segmentation methods,

Input: Pretrained Datasets: OpenSARShip, OpenSARUrban, UCAS-AOD, RSOD
Output: Trained Model: KCPNet
Input: Training Datasets: $D_{OpenSARShip}, D_{OpenSARUrban}, D_{UCAS-AOD}, D_{RSOD}$
Output: Trained Model: *KCPNet*
Initialization:
Load the pretrained weights for KCPNet model
Set initial learning rate: $\eta_0 = 0.001$
Set batch size: $B = 16$
Initialize optimizer: Adam, $\beta_1 = 0.9, \beta_2 = 0.999$
Set maximum number of epochs: $E = 100$
Set decay rate for learning rate: $r = 0.1$
Set validation split for cross-validation: $V = 5$
for *epoch = 1* **to** E **do**
 For each dataset:
 for *dataset in* $\{D_{OpenSARShip}, D_{OpenSARUrban}, D_{UCAS-AOD}, D_{RSOD}\}$ **do**
  Preprocessing:
  Apply Lee filter to remove noise;
  Apply data augmentation (rotation, flipping, cropping, etc.);
  Normalize input images: $I_{norm} = \frac{I-\mu}{\sigma}$
  Resize images to suitable dimensions: $I_{resized} \in \{256 \times 256, 512 \times 512, 416 \times 416\}$
  Training Step:
  while *not converged* **do**
   Sample mini-batch B from *dataset*
   for $i = 1$ **to** B **do**
    Forward pass: $Y = KCPNet(I_{resized})$
    Compute loss: $L = L_{classification} + L_{localization} + L_{IoU}$
    Backpropagate: Compute gradients $\nabla_\theta L$
    Update parameters: $\theta_{t+1} = \theta_t - \eta_t \cdot \nabla_\theta L$
   end
   Evaluate metrics:
   Compute Precision: $P = \frac{TP}{TP+FP}$
   Compute Recall: $R = \frac{TP}{TP+FN}$
   Compute F1 Score: $F1 = 2 \cdot \frac{P \cdot R}{P+R}$
   Compute mAP: $mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$
   Compute IoU: $IoU = \frac{A_{intersect}}{A_{union}}$
   if *epoch % 10 == 0* **then**
    Reduce learning rate: $\eta_{t+1} = \eta_t \cdot r$
   end
   if *Validation loss does not improve for 5 epochs* **then**
    Early stopping;
    Break from training loop;
   end
  end
  Save model checkpoint every 10 epochs.
 end
 End for each dataset
end
Final Model: Save the trained model: *KCPNet*
Return trained model: *KCPNet*

Algorithm 1. This algorithm outlines the step-by-step procedure for training the KCPNet model using multiple datasets. It includes dataset preprocessing, model training, and evaluation metrics, with early stopping and learning rate decay mechanisms to improve performance.

including UNet, SegNet, DeepLabV3+ (V3+), PSPNet, HRNet, FPN, and our proposed method. The table displays examples of accurate predictions ("True"), errors classified as false positives or false negatives ("Fault"), and actual ground truth images ("Reality"). Each row corresponds to the respective method, providing a direct comparison of their outputs. This comprehensive visualization highlights the ability of different methods to accurately predict karst collapse locations, alongside their respective errors. From the visual results, our method demonstrates superior performance in maintaining boundary integrity and preserving fine-grained details in collapse detection. Unlike other models, such as PSPNet and HRNet, which occasionally exhibit false positives (misclassifying stable areas as collapse zones) or false negatives (failing to detect certain collapse zones), our approach consistently identifies features with higher precision. This is particularly evident in the "Reality" column, where our results align more closely with actual collapse patterns compared to competing methods. However, the visualization also identifies some limitations in our method, as minor false positives are still present in highly noisy regions, suggesting room for further optimization in handling such data.

5 Discussion

The proposed Integrated Karst Collapse Prediction Network (IKCPNet) demonstrates significant improvements in predicting and mitigating karst collapse risks by integrating multi-modal geospatial data with deep learning-based segmentation. Our results indicate that IKCPNet achieves a higher accuracy and Intersection over Union (IoU) score compared to state-of-the-art models, highlighting the effectiveness of cross-modal feature fusion and the incorporation of geophysical constraints. One of the key contributions of this work is the use of the Subsurface Dynamics Module (SDM) to enhance the representation of underground interactions, which are often overlooked in traditional machine learning-based hazard prediction models. By leveraging both static and dynamic observational data, IKCPNet captures the evolving characteristics of karst environments, leading to more reliable risk assessments. Furthermore, the Multi-Modal Data Encoder (MMDE) ensures effective feature integration from heterogeneous sources, addressing one of the major limitations of existing approaches. Compared to previous studies, which primarily rely on single-

TABLE 4 Ablation study results on image segmentation across OpenSARShip and OpenSARUrban datasets.

Model	OpenSARShip dataset				OpenSARUrban dataset			
	Accuracy	Precision	Recall	IoU	Accuracy	Precision	Recall	IoU
w./o. Collapse Risk Prediction and Training	91.34±0.02	89.78±0.03	88.45±0.02	87.12±0.03	92.12±0.03	90.56±0.02	89.34±0.03	88.12±0.02
w./o. Context-Aware Risk Assessment	92.23±0.03	90.56±0.02	89.12±0.03	88.34±0.02	93.01±0.02	91.45±0.03	90.23±0.02	89.12±0.03
w./o. Adaptive Mitigation Measures	93.01±0.02	91.34±0.03	90.45±0.02	89.56±0.03	93.78±0.03	92.34±0.02	91.12±0.03	90.23±0.02
Ours	**94.34±0.02**	**92.78±0.02**	**91.45±0.03**	**90.23±0.02**	**93.78±0.02**	**92.45±0.02**	**91.34±0.02**	**90.12±0.03**

TABLE 5 Ablation study results on image segmentation across UCAS-AOD and RSOD datasets.

Model	UCAS-AOD dataset				RSOD dataset			
	Accuracy	Precision	Recall	IoU	Accuracy	Precision	Recall	IoU
w./o. Collapse Risk Prediction and Training	91.12±0.03	89.45±0.02	88.34±0.03	87.12±0.02	92.34±0.02	90.23±0.03	89.12±0.02	88.23±0.03
w./o. Context-Aware Risk Assessment	92.23±0.02	90.56±0.03	89.23±0.02	88.34±0.03	93.45±0.03	91.34±0.02	90.12±0.03	89.34±0.02
w./o. Adaptive Mitigation Measures	92.78±0.03	91.12±0.02	90.34±0.03	89.23±0.02	93.89±0.02	92.23±0.03	91.12±0.02	90.34±0.03
Ours	**93.45±0.02**	**91.78±0.02**	**90.34±0.03**	**89.45±0.02**	**94.12±0.02**	**92.34±0.02**	**91.12±0.02**	**90.23±0.03**

modality inputs such as LiDAR or hydrological measurements, our method provides a more comprehensive framework by integrating seismic imaging, environmental parameters, and remote sensing data. This multi-source fusion strategy significantly reduces false positives and enhances model interpretability. Additionally, the Dynamic Risk Mitigation Strategy (DRMS) introduces a real-time adaptation mechanism, making the system more suitable for early warning applications. Despite these advancements, certain limitations remain. First, while our model effectively integrates different data sources, its performance is influenced by the availability and quality of the input data. Future work could explore strategies for handling missing or incomplete data, such as self-supervised learning or domain adaptation techniques. Second, although our method achieves high accuracy, its real-world applicability depends on computational efficiency in large-scale deployments. Optimization strategies, such as model pruning and knowledge distillation, could be explored to improve inference speed. IKCPNet represents a significant step forward in karst collapse monitoring by providing an interpretable, data-driven approach that enhances prediction accuracy and adaptability. Future research directions include expanding the dataset to incorporate more diverse geological conditions, improving model generalization, and integrating additional physical constraints to refine collapse risk assessments.

6 Conclusions and future work

This study tackles the critical challenges in monitoring karst collapses by utilizing advanced multi-source image feature extraction and segmentation techniques, integral to cyber-physical systems for environmental monitoring. Traditional methods often fall short due to data sparsity, the complexity of non-linear subsurface dynamics, and limitations in real-time adaptability, which undermine their effectiveness in high-risk karst regions. To address these issues, the research introduces the Integrated

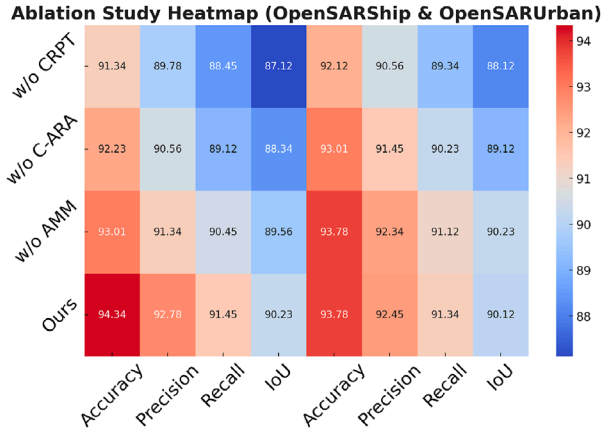

FIGURE 9
Ablation study of our Method on OpenSARShip dataset and OpenSARUrban dataset datasets. Abbreviations: Crpt - collapse risk prediction and training, C-ARA - context-aware risk assessment, AMM - adaptive mitigation Measures.

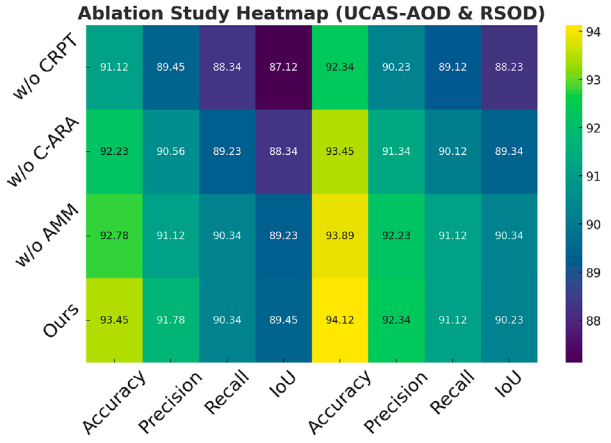

FIGURE 10
Ablation study of our Method on UCAS-AOD dataset and RSOD dataset datasets. Abbreviations: Crpt - collapse risk prediction and training, C-ARA - context-aware risk assessment, AMM - adaptive mitigation Measures.

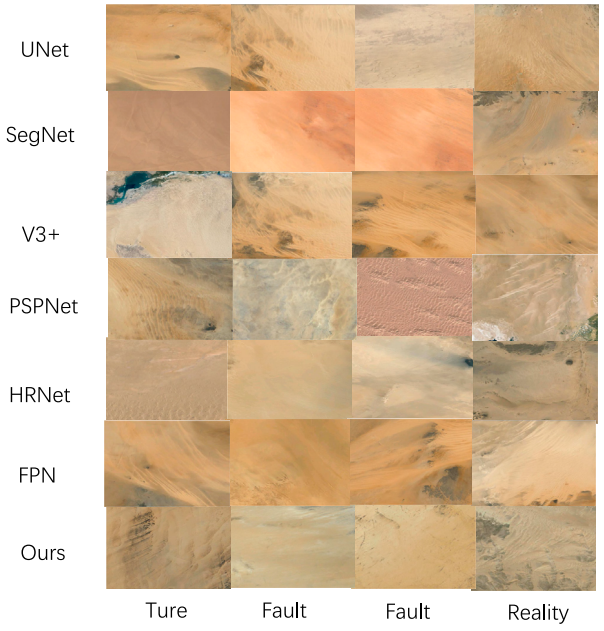

FIGURE 11
Comparative visualization of segmentation results across different methods, including UNet, SegNet, DeepLabV3+ (V3+), PSPNet, HRNet, FPN, and our proposed method. The rows represent the outputs of each method, while the columns highlight specific outcomes: (1) 'True' shows examples of correct predictions where the method accurately identifies collapse features; (2) 'Fault' columns display erroneous predictions, such as false positives where stable areas are misclassified as collapse zones, or false negatives where true collapse zones are missed; (3) 'Reality' represents the ground truth or actual collapse locations for reference. This detailed comparison emphasizes the robustness of our proposed method in achieving higher prediction accuracy, minimizing errors, and aligning closely with ground truth data. The visualization also illustrates the challenges faced by baseline methods, such as boundary misclassification or under-segmentation, highlighting the effectiveness of our approach in addressing these issues.

Karst Collapse Prediction Network (IKCPNet), a sophisticated framework combining multi-modal data encoding, geophysical simulations, and machine learning-driven segmentation techniques. IKCPNet employs the Multi-Modal Data Encoder (MMDE) and Subsurface Dynamics Module (SDM) to analyze inputs such as seismic imaging, hydrological data, and environmental metrics, delivering highly accurate collapse risk predictions. The framework is further strengthened by the Dynamic Risk Mitigation Strategy (DRMS), which incorporates real-time feedback, context-aware risk assessments, and probabilistic mapping to support informed decision-making and efficient response strategies. Experimental results indicate substantial advancements in prediction accuracy and mitigation capabilities, showcasing the framework's promise in enhancing karst collapse monitoring through comprehensive data fusion and segmentation.

Despite its innovations, the framework has two key limitations. The reliance on multi-modal data and sophisticated computational processes may pose challenges in resource-limited environments where access to diverse datasets and high-performance computing infrastructure is constrained. Future research could explore optimizing the framework for scalability and accessibility in such contexts. The study predominantly focuses on experimental simulations, necessitating further validation with extensive real-world deployments to confirm its robustness and adaptability across diverse karst terrains. Addressing these limitations will enhance the framework's practicality and extend its applicability. Discussing specific cases where the framework underperformed or faced challenges would provide a more balanced analysis of its capabilities. Such cases could reveal areas for improvement and showcase the authors' awareness of its limitations. For instance, scenarios involving sparse or highly noisy data, or regions where certain input modalities are unavailable, should be further examined to evaluate the method's robustness under suboptimal conditions. Incorporating these findings would not

only deepen the analysis but also strengthen the confidence in the framework's practical utility. Looking ahead, the adaptability of the proposed approach can be further demonstrated by broadening its application to other contexts of different hazards. Examples include landslide prediction, sinkhole detection, or infrastructure stability assessment in earthquake-prone regions. By extending the methodology to diverse environmental and geophysical challenges, the framework's flexibility and scalability could be tested, highlighting its potential as a universal tool for disaster risk monitoring and management. Exploring these future directions will provide valuable insights into the framework's adaptability and scalability, ultimately enhancing its impact in a wider range of hazard scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

WL: Writing–original draft, Writing–review and editing. XL: Data curation, Conceptualization, Formal analysis, Investigation, Funding acquisition, Software, Writing–original draft, Writing–review and editing. TL: Methodology, Supervision, Project administration, Validation, Resources, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

Author XL was employed by Guangdong Nonferrous Industry Building Quality Inspection Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Atigh, M. G., Schoep, J., Acar, E., Noord, N. V., and Mettes, P. (2022). "Hyperbolic image segmentation," in *Computer vision and pattern recognition*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: unet-like pure transformer for medical image segmentation. *ECCV Work*. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-25066-8_9

Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Neural Inf. Process. Syst.* Available online at: https://proceedings.neurips.cc/paper/2020/hash/949686ecef4ee20a62d16b4a2d7ccca3-Abstract.html

Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P. (2020). Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Trans. Med. Imaging* 39, 2494–2505. doi:10.1109/tmi.2020.2972701

Gao, Y., Zhou, M., and Metaxas, D. N. (2021). "Utnet: a hybrid transformer architecture for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*.

Hatamizadeh, A., Yang, D., Roth, H., and Xu, D. (2021). "Unetr: transformers for 3d medical image segmentation," in *IEEE workshop/winter conference on applications of computer vision*.

Huan, Y., Song, L., Khan, U., and Zhang, B. (2023). Stacking ensemble of machine learning methods for landslide susceptibility mapping in zhangjiajie city, hunan province, China. *Environ. Earth Sci.* 82, 35. doi:10.1007/s12665-022-10723-z

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020a). "Unet 3+: a full-scale connected unet for medical image segmentation," in *IEEE international conference on acoustics, speech, and signal processing*.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020b). "Unet 3+: a full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 1055–1059.

Isensee, F., Jaeger, P., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. (2020). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi:10.1038/s41592-020-01008-z

Jain, J., Li, J., Chiu, M., Hassani, A., Orlov, N., and Shi, H. (2022). "Oneformer: one transformer to rule universal image segmentation," in *Computer vision and pattern recognition*.

Jha, D., Riegler, M., Johansen, D., Halvorsen, P., and Johansen, H. D. (2020). "Doubleu-net: a deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd international symposium on computer-based medical systems (CBMS)*.

Jin, X., Hou, J., Zhou, W., and Lee, S.-J. (2023a). Editorial: recent advances in image fusion and quality improvement for cyber-physical systems. *Front. Neurorobotics* 17, 1201266. doi:10.3389/fnbot.2023.1201266

Jin, X., Liu, L., Ren, X., Jiang, Q., Lee, S.-J., Zhang, J., et al. (2024). A restoration scheme for spatial and spectral resolution of the panchromatic image using the convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 17, 3379–3393. doi:10.1109/jstars.2024.3351854

Jin, X., Zhang, P., He, Y., Jiang, Q., Wang, P., Hou, J., et al. (2023b). A theoretical analysis of continuous firing condition for pulse-coupled neural networks with its applications. *Eng. Appl. Artif. Intell.* 126, 107101. doi:10.1016/j.engappai.2023.107101

Kopiika, N., and Blikharskyy, Y. (2024). Digital image correlation for assessment of bridges' technical state and remaining resource. *Struct. Control Health Monit.* 2024, 1763285. doi:10.1155/2024/1763285

Kopiika, N., Karavias, A., Krassakis, P., Ye, Z., Ninic, J., Shakhovska, N., et al. (2025). Rapid post-disaster infrastructure damage characterisation using remote sensing and deep learning technologies: a tiered approach. *Automation Constr.* 170, 105955. doi:10.1016/j.autcon.2024.105955

Kopiika, N., Ninic, J., and Mitoulis, S. (2024). "Damage characterisation using sentinel-1 images: case study of bridges in Ukraine," in *IABSE symposium manchester 2024: construction's role for a world in emergency*, 367–375.

Li, X., Duan, F., Hu, M., Hua, J., and Du, X. (2023). Weed density detection method based on a high weed pressure dataset and improved psp net. *IEEE Access* 11, 98244–98255. doi:10.1109/access.2023.3312191

Lin, A.-J., Chen, B., Xu, J., Zhang, Z., Lu, G., and Zhang, D. (2021). Ds-transunet: dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* 71, 1–15. doi:10.1109/tim.2022.3178991

Liu, W., Liu, J., and Luo, B. (2023). Unsupervised domain adaptation for remote sensing vehicle detection using domain-specific channel recalibration. *IEEE Geoscience Remote Sens. Lett.* 20, 1–5. doi:10.1109/lgrs.2023.3314644

Liu, X., Song, L., Liu, S., and Zhang, Y. (2021). A review of deep-learning-based medical image segmentation methods. *Sustainability* 13, 1224. doi:10.3390/su13031224

Lüddecke, T., and Ecker, A. S. (2021). *Image segmentation using text and image prompts*. Computer Vision and Pattern Recognition. Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Luddecke_Image_Segmentation_Using_Text_and_Image_Prompts_CVPR_2022_paper.html

Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., and Zhang, S. (2020). "Semi-supervised medical image segmentation through dual-task consistency," in *AAAI conference on artificial intelligence*.

Malhotra, P., Gupta, S., Koundal, D., Zaguia, A., and Enbeyle, W. (2022). Deep neural networks for medical image segmentation. *J. Healthc. Eng.* 2022, 1–15. doi:10.1155/2022/9580991

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2020). Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Analysis Mach. Intell.* 44, 3523–3542. doi:10.1109/tpami.2021.3059968

Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., and Rueckert, D. (2020). "Self-supervision with superpixels: training few-shot medical image segmentation without annotation," in *European conference on computer vision*.

Peiris, H., Chen, Z., Egan, G., and Harandi, M. (2021). "Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, strasbourg, France, september 27–october 1, 2021, proceedings, Part II 24* (Springer), 428–438.

Peng, H., Xue, C., Shao, Y., Chen, K., Xiong, J., Xie, Z., et al. (2020). Semantic segmentation of litchi branches using deeplabv3+ model. *Ieee Access* 8, 164546–164555. doi:10.1109/access.2020.3021739

Ren, Q., Lu, Z., Wu, H., Zhang, J., and Dong, Z. (2023). Hr-net: a landmark based high realistic face reenactment network. *IEEE Trans. Circuits Syst. Video Technol.* 33, 6347–6359. doi:10.1109/tcsvt.2023.3268062

Tan, X., Leng, X., Ji, K., and Kuang, G. (2024). Rcship: a dataset dedicated to ship detection in range-compressed sar data. *IEEE Geoscience Remote Sens. Lett.* 21, 1–5. doi:10.1109/lgrs.2024.3366749

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021). "Medical transformer: gated axial-attention for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*.

Wang, G. (2024). Rl-cwtrans net: multimodal swimming coaching driven via robot vision. *Front. Neurorobotics* 18, 1439188. doi:10.3389/fnbot.2024.1439188

Wang, J., Ji, Y., and Yang, H. (2024). Vahagn: visual haptic attention gate net for slip detection. *Front. Neurorobotics* 18, 1484751. doi:10.3389/fnbot.2024.1484751

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., and Gool, L. (2021). "Exploring cross-image pixel contrast for semantic segmentation," in *IEEE international conference on computer vision*.

Wu, J., Fang, H., Zhang, Y., Yang, Y., and Xu, Y. (2022). "Medsegdiff: medical image segmentation with diffusion probabilistic model," in *International conference on medical imaging with deep learning*.

Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., and Mello, S. D. (2023). "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Computer vision and pattern recognition*.

Yin, X., Sun, L., Fu, Y., Lu, R., and Zhang, Y. (2022). U-net-based medical image segmentation. *J. Healthc. Eng.* 2022, 1–16. doi:10.1155/2022/4189781

Yu, Y., Wang, C., Fu, Q., Kou, R.-J., Huang, F., Yang, B., et al. (2023). Techniques and challenges of image segmentation: a review. *Electronics* 12, 1199. doi:10.3390/electronics12051199

Yusuf, M. O., Hanzla, M., Al Mudawi, N., Sadiq, T., Alabdullah, B., Rahman, H., et al. (2024). Target detection and classification via efficientdet and cnn over unmanned aerial vehicles. *Front. Neurorobotics* 18, 1448538. doi:10.3389/fnbot.2024.1448538

Zhang, B., Tang, J., Huan, Y., Song, L., Shah, S. Y. A., and Wang, L. (2024). Multi-scale convolutional neural networks (cnns) for landslide inventory mapping from remote sensing imagery and landslide susceptibility mapping (lsm). *Geomatics, Nat. Hazards Risk* 15, 2383309. doi:10.1080/19475705.2024.2383309

Zhang, X., Zhang, T., Wang, G., Zhu, P., Tang, X., Jia, X., et al. (2023). Remote sensing object detection meets deep learning: a metareview of challenges and advances. *IEEE Geoscience Remote Sens. Mag.* 11, 8–44. doi:10.1109/mgrs.2023.3312347

Zhang, Y., Liu, H., and Hu, Q. (2021). "Transfuse: fusing transformers and cnns for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*.

Zhao, J., Zhang, Z., Yao, W., Datcu, M., Xiong, H., and Yu, W. (2020). Opensarurban: a sentinel-1 sar image dataset for urban interpretation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 187–203. doi:10.1109/jstars.2019.2954850

Zhou, X., and Zhang, L. (2022). Sa-fpn: an effective feature pyramid network for crowded human detection. *Appl. Intell.* 52, 12556–12568. doi:10.1007/s10489-021-03121-8