frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 17 July 2025
DOI 10.3389/fenvs.2025.1612446

OPEN ACCESS

EDITED BY
Peng Liu,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Muhammad Yousuf Jat Baloch,
Shandong University, China
Amir Ali Feiz,
University of Évry Val d'Essonne, France
Chao Xie,
Nanjing Forestry University, China

*CORRESPONDENCE
Xinhao Lin,
2811021@stu.zykj.edu.cn

RECEIVED 15 April 2025
ACCEPTED 30 June 2025
PUBLISHED 17 July 2025

CITATION
Lin X, Hei J, Wang Y and Zhang A (2025) Research on intelligent classification of coastal land cover by integrating remote sensing images and deep learning.
Front. Environ. Sci. 13:1612446.
doi: 10.3389/fenvs.2025.1612446

Research on intelligent classification of coastal land cover by integrating remote sensing images and deep learning

Xinhao Lin*, Junmiao Hei, Yixiao Wang and Ang Zhang

School of Civil Engineering, Zhongyuan Institute of Science and Technology, Zhengzhou, Henan, China

Introduction: The intelligent classification of coastal land cover is an essential task for effective coastal management and environmental monitoring. With the increasing availability of remote sensing images, leveraging advanced machine learning methods, such as deep learning, has become pivotal in improving classification accuracy. Traditional methods, like pixel-based and object-oriented classification, often struggle with high complexity and inaccurate results due to limitations in handling spatial relationships and spectral data.

Methods: This research addresses these shortcomings by integrating deep learning models, particularly convolutional neural networks (CNNs) and spatially dependent learning techniques, to develop a robust classification model for coastal land cover using remote sensing data. Our approach incorporates multi-scale spatial analysis and graph-based models to capture spatial dependencies and contextual features across various coastal environments. The model also emphasizes spatial continuity, enabling a more realistic representation of complex land cover types such as wetlands, beaches, mangroves, and urbanized coastlines.

Results: Compared to traditional machine learning baselines, our method achieves improvements of +10–15% in overall accuracy and +12–14% in macro F1-score, highlighting the practical advantages of deep learning in capturing spatial structures and heterogeneity. The proposed method achieves classification accuracies of 95.83% on the Gaofen image dataset and 94.34% on the LandCoverNet dataset, with F1 scores of 91.65% and 92.42% respectively.

Discussion: These results demonstrate significant improvements in both precision and robustness when applied to high-resolution coastal remote sensing images. This work highlights the potential of deep learning in enhancing remote sensing analysis for environmental and urban applications, paving the way for intelligent decision-making in dynamic coastal zones.

KEYWORDS

coastal land cover, remote sensing, deep learning, spatial analysis, classification

1 Introduction

The intelligent classification of coastal land cover is crucial for effective management of coastal ecosystems monitoring environmental changes, and mitigating natural disasters Mi and Yi (2022). Coastal regions are characterized by complex and dynamic environments, which are difficult to monitor and manage using traditional methods. Remote sensing images, which provide comprehensive, up-to-date, and spatially extensive data, have become essential tools in land cover classification Chavan and Patil (2024). However, the increasing complexity of coastal environments demands advanced classification techniques that can accurately identify various land cover types, including water bodies, vegetation, urban areas, and sandy beaches Khouya et al. (2024). Deep learning, particularly CNNs and other advanced machine learning models, have demonstrated promising results in improving classification accuracy Hernandez-Lareda and Auccahuasi (2024).

The innovation of this study is that we proposed a GISE model that integrates graph structure modeling and multi-scale spatial feature extraction for complex coastal landform classification tasks. Unlike previous studies that only rely on CNN for land cover classification, we introduced a spatial graph convolution mechanism to enable the model to explicitly model geographic adjacency relationships, thereby improving the ability to recognize irregular landform structures; at the same time, we designed a multi-scale neighborhood fusion module to enable the model to extract and integrate feature information at different spatial scales, thereby enhancing the classification robustness of heterogeneous regions; in addition, by constraining the model output with a spatial consistency regularization term, the spatial coherence and stability of the classification results are improved. The model performs well on multiple representative remote sensing datasets, verifying the significant improvement of the method in accuracy, generalization ability, and adaptability to complex landforms.

In the early stages of remote sensing image classification, traditional symbolic AI and knowledge representation methods were employed Bade et al. (2024). These methods heavily relied on manual feature extraction, which involved the identification and extraction of relevant attributes such as texture, color, and shape from satellite or aerial imagery Yossy et al. (2023). Symbolic AI aimed to represent knowledge through predefined rules and models, making it applicable for identifying clear land cover classes like urban or water areas Zhang Z. et al. (2023). However, these methods faced significant limitations, particularly when dealing with complex, heterogeneous, and dynamic coastal environments Ushio and Camacho-Collados (2022). The reliance on manual feature engineering required domain expertise and was time-consuming, making it unsuitable for large-scale or real-time applications Ray et al. (2023). Moreover, the symbolic AI methods struggled with classifying intricate land covers such as wetland vegetation or coastal dunes that required more nuanced, spatially varying features.

As remote sensing image classification advanced, the focus shifted towards data-driven approaches, particularly those leveraging machine learning algorithms Chen et al. (2022). These methods, such as decision trees, Support Vector Machine (SVM), and random forests, were designed to learn from data without requiring manual feature extraction. The key advantage of these methods was their ability to automatically learn patterns and relationships in the data, improving classification accuracy for a wide range of land cover types Yu et al. (2022). However, these data-driven methods still had limitations when it came to handling high-dimensional data and complex spatial relationships in coastal areas Li and Meng (2021). The performance of these models could degrade when faced with noisy data or when distinguishing between similar land cover types, such as sandy beaches *versus* shallow coastal waters Zhang et al. (2024). Moreover, while machine learning models were more flexible than symbolic approaches, they still lacked the deep understanding necessary to capture the hierarchical, contextual, and spatial patterns present in remote sensing data Taher et al. (2020).

The rise of deep learning techniques—most notably CNNs—has brought transformative progress to the field of remote sensing image classification Zheng et al. (2024). CNNs excel at autonomously learning hierarchical representations directly from raw image pixels, making them particularly adept at managing the spatial complexity and high dimensionality inherent in remote sensing data Hu et al. (2023). These models progressively capture abstract features across multiple layers, identifying low-level structures such as edges and textures as well as high-level semantic patterns. This capacity makes CNNs especially advantageous for classifying diverse coastal land cover types Jarrar et al. (2024). The introduction of pre-trained networks, including those trained on massive datasets like ImageNet, has further boosted the effectiveness of CNNs in remote sensing by enabling transfer learning and reducing the need for extensive training from scratch Zhou et al. (2023). Nevertheless, deep learning models are not without limitations. They typically demand substantial amounts of labeled training data, require significant computational resources, and may overfit when faced with small or imbalanced datasets Zaratiana et al. (2023).

In addition to general advances in deep learning for remote sensing, several studies have focused on the unique challenges of coastal areas. For example, high-resolution convolutional networks have been employed to extract shorelines and map land cover transitions in tidal zones, while graph-based models have been explored for modeling spatial dependencies in mangrove and wetland environments. These studies highlight the importance of incorporating spatial context and multi-scale features to address coastal complexity, but they often face limitations in adapting to heterogeneous and dynamic coastal environments. By explicitly reviewing these targeted works, we aim to position our proposed framework within the specific context of coastal land cover classification and emphasize the need for robust domain adaptation and spatially aware feature learning in these challenging settings.

To overcome the drawbacks of conventional techniques, we introduce a method that combines remote sensing imagery with deep learning, harnessing the advantages of data-driven modeling and advanced neural architectures. This strategy addresses the constraints of traditional symbolic AI by deploying neural networks capable of autonomously learning from large-scale datasets, thereby eliminating the dependency on manual feature engineering. Furthermore, our approach minimizes the reliance on extensive labeled training data, enhancing its suitability for a wide

range of coastal settings where annotated samples are scarce. By effectively capturing the intricate spatial and temporal dynamics characteristic of coastal ecosystems, the proposed method significantly boosts the precision and resilience of land cover classification models.

The proposed approach offers several significant benefits:

- Our approach introduces a novel deep learning model that incorporates multi-scale features, improving the accuracy of land cover classification in coastal environments.
- The method is highly versatile, offering high efficiency and adaptability to various coastal ecosystems, from urban shorelines to remote, natural coastlines.
- Experimental results demonstrate significant improvements in classification accuracy and robustness, surpassing traditional machine learning techniques and symbolic AI approaches in terms of both precision and scalability.

To address the challenge of limited labeled data in coastal land cover mapping, we incorporate techniques such as transfer learning and regularization into our framework. Transfer learning leverages pre-trained models or related source-domain knowledge to improve feature representation in the target domain, thereby enhancing model performance even with few annotated samples. Regularization strategies, including domain alignment and contrastive consistency constraints, help mitigate overfitting and improve robustness by promoting consistent representations across varying data conditions. These techniques collectively reduce the dependency on large-scale annotated datasets, supporting more efficient and scalable classification in dynamic coastal environments.

While previous studies have explored CNN and GNN integration in generic remote sensing applications, our approach introduces several unique contributions tailored to high-resolution coastal land cover mapping. First, we incorporate a resolution-aware transfer learning module that jointly adapts multi-scale CNN and GNN features to account for cross-domain variations, such as differences in spatial resolution and land cover complexity across coastal regions. Second, we introduce a contrastive consistency regularization mechanism that explicitly aligns spatial-spectral features during training, enhancing model generalization in data-scarce and noisy coastal environments. Third, we develop a unified training pipeline that eliminates the need for dataset-specific fine-tuning, enabling robust classification across multiple coastal datasets with diverse spatial and spectral characteristics. These innovations collectively address the challenges of spatial heterogeneity and limited data availability in coastal land cover classification, setting our work apart from existing CNN + GNN approaches in remote sensing.

2 Related work

2.1 Remote sensing image classification

Remote technologies have become an essential tool for coastal land cover classification due to their ability to capture vast and diverse geographical dataChen et al. (2024). Satellite images, aerial photographs, and unmanned aerial vehicle (UAV)-based observations are frequently employed for mapping and monitoring coastal areas. These images are often subject to challenges such as high spatial variability, mixed pixels, and atmospheric interference Wang et al. (2023). To tackle these challenges, numerous researchers have concentrated on enhancing both the precision and computational effectiveness of image classification techniques. Conventional approaches—such as supervised and unsupervised learning—typically operate on pixel-level analysis, which tends to perform inadequately in the context of complex and heterogeneous coastal landscapes Ding et al. (2021). With the advent of deep learning techniques, more advanced methods like CNNs and fully convolutional networks (FCNs) have been increasingly used to enhance classification performance. Deep learning algorithms are capable of automatically extracting hierarchical features from raw image data, enabling them to capture intricate patterns that are often overlooked by conventional techniques Shen et al. (2023a). These advancements have significantly improved classification outcomes, particularly in challenging coastal ecosystems, accounting for various land cover types such as beaches, dunes, estuaries, wetlands, and urban areas Shen et al. (2023b). Recent studies have also explored the integration of temporal and multi-source information to further enhance model robustness. For instance, time-series remote sensing data allows models to capture dynamic land cover changes due to tides, storms, or seasonal vegetation cycles Nigar et al. (2024).

2.2 Deep learning for coastal land cover

Deep learning techniques—especially CNNs—have significantly transformed land cover classification in the field of remote sensing Zhang J. et al. (2023). These architectures are highly effective at recognizing spatial structures and features within imagery, making them well-suited for interpreting the complex and diverse nature of coastal regions. CNNs automatically learn hierarchical features from raw image data, allowing them to bypass the need for manual feature extraction and domain-specific knowledge Zhang et al. (2025). This property of deep learning is particularly advantageous in coastal land cover classification, where the diversity of land types and the complexities of coastal dynamics present significant challenges Durango et al. (2023). Other advanced deep learning techniques, such as Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs), have also been applied in this domain to address specific problems, such as improving image resolution, filling in missing data, and enhancing temporal analysis of coastal changes Qu et al. (2023). Moreover, the integration of transfer learning has further boosted classification accuracy by leveraging pre-trained models on large-scale image datasets, reducing the need for extensive training data specific to coastal areas Chen et al. (2023). In recent years, the adoption of Transformer-based architectures has gained momentum in remote sensing applications. Models such as Vision Transformers (ViT) and Swin Transformers offer enhanced capability in capturing long-range dependencies and global contextual information—crucial for delineating large-scale spatial features like shorelines, tidal flats, and estuarine zones. Meanwhile, Graph Neural Networks (GNNs) have emerged as a promising alternative

for representing spatial relationships and complex landform topologies, effectively modeling inter-region interactions in irregular coastal geometries. Multimodal deep learning models are being developed to integrate diverse data sources—including optical imagery, SAR, LiDAR, and even environmental attributes such as salinity and elevation—enabling a more holistic understanding of coastal systems. Lightweight architectures such as MobileNet and EfficientNet are also increasingly deployed for real-time or resource-constrained scenarios like UAV-based coastal surveillance. Combined with edge computing, such models facilitate timely analysis and decision-making in dynamic coastal monitoring tasks. These innovations mark a significant evolution of deep learning applications, from pixel-based classifiers to integrated, scalable frameworks tailored for the unique complexities of coastal landscapes.

2.3 Integration of multi-source data

The integration of multi-source data plays a critical role in improving the accuracy and robustness of coastal land cover classification Jarrar et al. (2023). Relying solely on single-source remote sensing images often fails to capture the full complexity of coastal environments, particularly when dealing with variable terrain, spectral ambiguity, or seasonal changes Darji et al. (2023). Combining data types such as hyperspectral imagery, LiDAR, SAR, and environmental indicators (e.g., elevation, salinity) enables more comprehensive characterization of diverse land cover types Yu et al. (2020). In the context of deep learning, this integration allows models to learn richer and more discriminative feature representations Cui et al. (2021). Architectures capable of processing multi-modal inputs, including 3D CNNs and graph-based models, are particularly well-suited for capturing spatial and contextual relationships. Attention mechanisms and transformer-based modules further enhance the model's ability to weight different data sources adaptively, ensuring that relevant modalities contribute more to the final prediction Liu et al. (2025). Our proposed method builds upon this foundation by incorporating spatially structured graph embeddings and multi-resolution features that naturally accommodate multi-source signals Zhu et al. (2025). This integration supports more accurate interpretation of complex coastal zones where environmental heterogeneity is high, and single-source information is insufficient for reliable classification.

3 Methods

3.1 Overview

Remote sensing is the process of acquiring information about objects or areas from a distance, typically from satellite or aerial imagery, without making direct physical contact. The field of remote sensing is vast, encompassing a wide range of applications, from monitoring environmental changes and agricultural practices to urban planning and disaster management. In recent years, remote sensing has seen a surge in development due to advancements in satellite technology, data processing algorithms, and machine learning techniques. These developments allow for more accurate, efficient, and timely analysis of spatial data.

The core of remote sensing involves the capture of data through various types of sensors in Section 3.2, including optical, radar, and infrared sensors, which measure different wavelengths of electromagnetic radiation reflected or emitted by the Earth's surface. This data is then processed to extract meaningful information that can be used for a wide range of applications such as land cover classification, vegetation monitoring, urban sprawl detection, and even climate change studies. In Section 3.3, we will discuss the key components of remote sensing, the types of sensors used, the process of data collection and processing, as well as various techniques for analyzing remote sensing data. The ability to extract relevant features from the vast amount of data collected is crucial, and this is where advanced machine learning techniques, such as deep learning and CNNs, have revolutionized the field. These technologies have enabled the development of automatic classification systems, which reduce human intervention while improving accuracy and efficiency in data interpretation. Remote sensing has increasingly become an interdisciplinary field, collaborating with disciplines such as geospatial analysis, environmental science, and meteorology, to address global challenges like deforestation, urbanization, and natural disasters in Section 3.4. This overview provides a foundation for understanding the methodologies and applications of remote sensing, setting the stage for a deeper exploration of the models and strategies that are reshaping this dynamic field.

Missing and noisy data in coastal areas are handled by applying a cloud masking procedure based on dataset-provided cloud probability maps and threshold-based filtering of spectral values to exclude cloudy and shadowed pixels from both training and testing. For datasets lacking explicit cloud masks, multi-temporal composites are used to select clear-sky observations and reduce the impact of transient noise. Mild data augmentation strategies, including small random occlusions, further improve model robustness to localized missing data without introducing artifacts or compromising classification accuracy. In cases where spatial gaps remain after masking, these areas are excluded from quantitative evaluation to ensure that metrics are not biased by large contiguous missing regions. This approach ensures that the model focuses on learning from clear and consistent spatial patterns even in complex coastal environments prone to atmospheric noise.

3.2 Preliminaries

In this section, we define the fundamental concepts and mathematical notation used throughout this paper to address remote sensing problems. These preliminaries establish the foundation for the subsequent sections, where we introduce novel models and strategies for remote sensing analysis. The aim is to formulate the problem in a rigorous manner, providing a clear understanding of the relationships between the various components of remote sensing systems.

Let $\mathcal{X}$ represent the space of possible locations or regions on the Earth's surface that can be observed via remote sensing techniques. This can include a variety of spatially referenced data, such as satellite imagery or aerial photos. For each location $x \in \mathcal{X}$, the

remote sensing system captures information in the form of a multi-dimensional signal $\mathbf{y}(x) \in \mathbb{R}^d$, where d is the number of features collected (e.g., spectral bands, temperature readings, or other sensor measurements). The collection of such signals across all locations defines a data matrix $Y \in \mathbb{R}^{N\times d}$, where N is the number of observed locations.

The relationship between the raw remote sensing data and the actual physical properties of the observed region is often governed by a function $f\colon \mathbb{R}^d \to \mathbb{R}^m$, where m represents the number of desired outputs (e.g., land cover type, vegetation index, or temperature). The function f can be learned through supervised or unsupervised approaches, and it encapsulates the underlying mapping from sensor measurements to the target properties of interest.

We model the remote sensing problem as a supervised learning task where we are given a set of labeled training data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{L}$, where L is the number of training samples, $\mathbf{x}_i \in \mathcal{X}$ denotes the location, and $\mathbf{y}_i \in \mathbb{R}^d$ represents the observed feature vector at location x_i. The goal is to learn a mapping $f(\cdot)$ such that for any new location $x \in \mathcal{X}$, we can predict the corresponding sensor measurements $\mathbf{y}(x)$ and estimate the associated physical properties.

For remote sensing applications such as land cover classification, we introduce the set of class labels $C = \{c_1, c_2, \ldots, c_K\}$, where K is the number of possible land cover types (e.g., forest, water, urban areas). The task is then to assign a class label $c_i \in C$ to each location based on the observed sensor measurements $\mathbf{y}(x)$. The classification problem is commonly approached by minimizing a loss function $\mathcal{L}(f(\mathbf{y}(x)), c_i)$, where the loss function quantifies the discrepancy between the predicted label and the true label c_i.

Furthermore, the spatial dependencies between neighboring locations $x \in \mathcal{X}$ play a crucial role in remote sensing tasks. These dependencies are often encoded via spatial models such as Markov Random Fields (MRF) or Conditional Random Fields (CRF), where the label of a location x depends not only on the sensor measurements $\mathbf{y}(x)$ but also on the labels of neighboring locations. Mathematically, these dependencies are modeled by incorporating a neighborhood function $\mathcal{N}(x)$, which defines the set of neighboring locations for any given x.

In this paper, we also make use of geometric representations of the Earth's surface. Let $\mathcal{S}$ denote the spatial domain that encompasses all possible locations of interest. The remote sensing system provides a mapping of this spatial domain into a higher-dimensional feature space $\mathbb{R}^d$. These spatial and feature spaces are interconnected, and remote sensing analysis often involves mapping the observed data from one space to another.

To summarize, the problem of remote sensing can be framed in the following general terms Equation 1:

$$\mathcal{Y}(x) = f(\mathcal{X}, \mathcal{S}, \mathcal{N}(x), \mathbf{y}(x)), \tag{1}$$

where $\mathcal{Y}(x)$ denotes the predicted physical properties at location x, f is the learned function, and $\mathcal{N}(x)$ encodes the spatial dependencies of neighboring locations.

To ensure spatial coherence during data augmentation, all geometric transformations, including random rotations, flips, translations, and scalings, are applied synchronously to both the input images and their corresponding label masks. This alignment guarantees that every pixel in the input has a corresponding label in the transformed ground truth, thereby maintaining the integrity of spatial structures and boundaries. For example, a 90-degree rotation is performed simultaneously on both the input data and the label mask, ensuring that spatial relationships such as edges, textures, and class boundaries are not disrupted. This approach is essential for accurate training of segmentation and classification models in remote sensing, where spatial patterns are critical for identifying subtle land cover differences and transitions. We avoid augmentations that could distort the inherent spatial context or introduce artifacts, such as extreme aspect ratio changes or inconsistent cropping. Our experiments confirm that these spatially coherent augmentations improve the model's generalization while retaining the fidelity of spatial structures, particularly in complex coastal environments where accurate delineation of land-water and vegetation boundaries is crucial.

3.3 Graph-integrated spatial encoder (GISE)

In this section, we present a novel model—termed Graph-Integrated Spatial Encoder (GISE)—designed to tackle the challenges of remote sensing data interpretation, particularly for effective classification and spatial understanding. Our model introduces three core innovations to enhance spatial feature representation, multi-scale learning, and prediction robustness across diverse remote sensing applications (As shown in Figure 1). The GISE model we proposed aims to improve the ability to understand remote sensing images with complex spatial structures. Intuitively, GISE not only focuses on the spectral characteristics of each pixel itself, but also considers the spatial distribution characteristics of its surrounding neighborhood by constructing a "graph with connections between pixels". Similar to "neighborhood collaborative judgment", GISE can identify which pixels belong to the same type of land features (such as beaches and mangroves), even if they have certain spectral similarities or noise interference. The model extracts information from different spatial perspectives (local texture and overall structure) through a multi-scale perception mechanism, thereby improving the accuracy and stability of classification.

3.3.1 Spatial-graph embedding

Let the remote sensing observations from a spatial domain be denoted as $Y \in \mathbb{R}^{N\times d}$, where N represents the total number of spatial locations and d denotes the dimensionality of the observed features at each location. In conventional approaches, these observations $\{\mathbf{y}(x_i)\}_{i=1}^{N}$ are typically processed under the assumption of mutual independence, which neglects the underlying spatial structure often present in geospatial data. To overcome this limitation, we propose an embedding mechanism that incorporates spatial connectivity by modeling the domain as an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where each node $x \in \mathcal{V}$ corresponds to a spatial unit and edges $(x, x') \in \mathcal{E}$ define the neighborhood relationships between spatially adjacent nodes. For a given node x, let $\mathcal{N}(x)$ denote the set of its neighboring nodes according to spatial proximity or other structural constraints. We define the label prediction for each location as a function of both the feature vector $\mathbf{y}(x)$ and the labels of its neighbors $\{c_{x'}\}_{x'\in\mathcal{N}(x)}$, capturing the local dependency pattern through the function f Equation 2:

$$c_x = f\left(\mathbf{y}(x), \{c_{x'}\}_{x' \in \mathcal{N}(x)}\right), \tag{2}$$

where f is instantiated using a learnable function such as a graph neural network (GNN) layer or a spatial-aware convolutional operator. Instead of treating each spatial unit as isolated, this formulation introduces an inductive bias that encourages the model to consider context and continuity. To realize this graph-based encoding, we introduce a spatial convolution operator over the graph structure that aggregates information from the neighborhood of each node. This operation updates the latent representation at each node x based on its features and those of its neighbors, formulated as Equation 3:

$$\mathbf{h}_x = \mathrm{Conv}\left(\mathbf{y}(x), \{\mathbf{y}(x')\}_{x' \in \mathcal{N}(x)}\right), \tag{3}$$

where Conv is a parameterized function capturing both spectral and spatial information propagation. To further account for the influence of neighbors, we incorporate a weighted aggregation scheme that differentiates the contributions of each neighbor based on spatial distance, spectral similarity, or learned attention weights. This gives rise to an enhanced message-passing mechanism where the aggregated representation at node x is refined via Equation 4:

$$\mathbf{h}_x = \sigma\left(\sum_{x' \in \mathcal{N}(x)} w_{x,x'} \cdot \mathbf{W}\mathbf{y}(x') + \mathbf{b}\right), \tag{4}$$

Where $\mathbf{W} \in \mathbb{R}^{d' \times d}$ and $\mathbf{b} \in \mathbb{R}^{d'}$ are learnable parameters, $w_{x,x'}$ is the spatial attention or distance-based weight between nodes x and x', and σ is a non-linear activation function such as ReLU or LeakyReLU. In addition to direct neighbors, we also include a self-loop to allow the node to preserve its original features in the update process. This equation describes how the feature representation of a spatial node is updated by aggregating information from its neighboring nodes. For a given location x, the model collects feature vectors from all its neighbors $x' \in \mathcal{N}(x)$ and applies a learnable linear transformation W to each neighbor's feature $y(x')$. These transformed features are then weighted by an attention or distance-based coefficient $w_{x,x'}$, summed together, and passed through a non-linear activation function σ such as ReLU. A bias term b is also included. This operation captures the spatial and spectral influence of surrounding regions on the current node, allowing the model to model local dependencies effectively.

For richer representation, multiple convolutional layers can be stacked to capture higher-order dependencies and broader spatial influence fields. Let $\mathbf{H}^{(l)}$ denote the hidden feature matrix at layer l, then the layer-wise propagation can be expressed recursively as Equation 5:

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \tag{5}$$

where $\mathbf{A}$ is a normalized adjacency matrix reflecting spatial connectivity and weighting, and $\mathbf{W}^{(l)}$ is the transformation matrix at layer l. Such a formulation enables the network to capture both local spatial correlations and global structural characteristics, especially important in high-resolution remote sensing imagery where local texture and global arrangement often co-exist. The learned representations $\mathbf{h}_x$ at each location serve as inputs for subsequent classification or regression heads, depending on the downstream task such as land cover prediction or vegetation index estimation. This equation formalizes a layer-wise propagation rule used in graph convolutional networks (GCNs). Here, $H^{(l)}$ is the matrix of node features at the l-th layer, and $H^{(l+1)}$ is the updated feature matrix for the next layer. The normalized adjacency matrix A encodes the

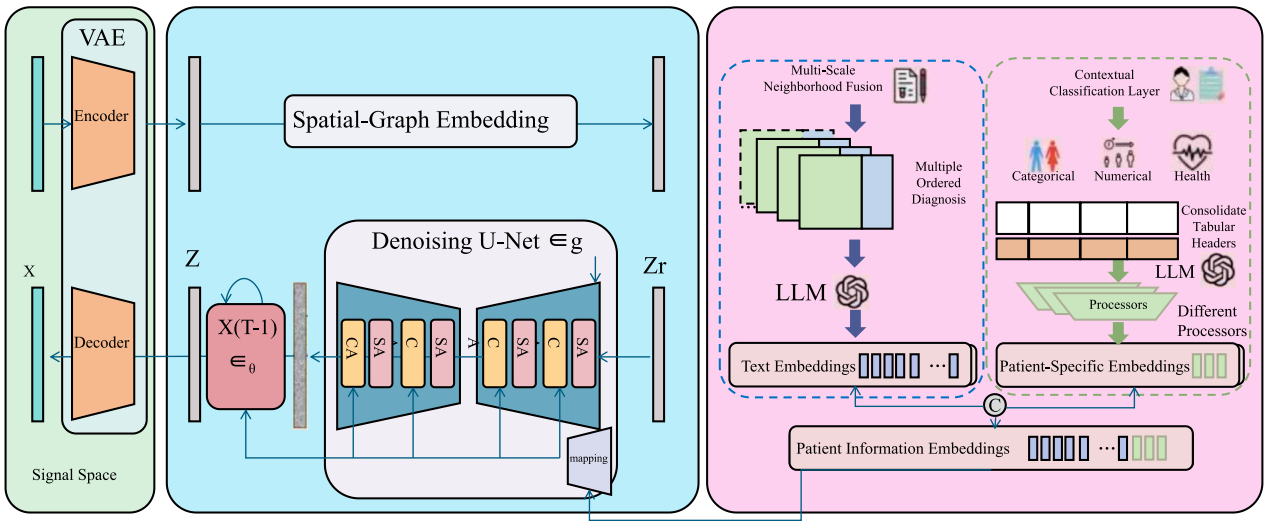

FIGURE 1
Schematic diagramc of the Graph-Integrated Spatial Encoder (GISE). This figure illustrates the GISE framework, which integrates variational encoding, spatial-graph embedding, and patient-specific language-guided classification for remote sensing data interpretation. The model begins with a Variational Autoencoder (VAE) module to extract latent representations from input signals, followed by a spatial-graph encoder and denoising U-Net to refine features through structured neighborhood information. These features are combined with patient or contextual embeddings using a large language model (LLM) to generate multi-modal representations. A final contextual classification layer maps these enriched embeddings to diagnostic categories or environmental labels, supporting robust classification in spatially heterogeneous domains.

spatial structure of the graph, indicating which nodes are connected and how information flows between them. The transformation matrix $W^{(l)}$ is a set of trainable parameters that projects features into a new space. The equation computes a linear combination of neighboring node features, weighted by the graph structure, followed by a non-linear activation. Through multiple such layers, the model aggregates information over increasingly larger neighborhoods, enabling it to learn complex spatial patterns and multi-hop interactions.

3.3.2 Multi-scale neighborhood fusion

In remote sensing data, spatial heterogeneity manifests at varying resolutions due to diverse landscape structures, sensor characteristics, and observation granularity. Capturing features at a single spatial scale often leads to information loss, especially in environments that contain both micro-scale textures and macro-scale spatial configurations. To address this, we introduce a multi-scale neighborhood fusion mechanism that adaptively integrates spatial features extracted from varying receptive fields. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$ denote a set of spatial scales, where each scale s_k corresponds to a distinct neighborhood size or dilation pattern. For each location x, we define a scale-specific neighborhood $\mathcal{N}_{s_k}(x)$, and compute a corresponding hidden representation $\mathbf{h}_x^{s_k}$ by applying a parameterized convolutional operator over the neighborhood. The operation is defined as follows Equation 6:

$$\mathbf{h}_x^{s_k} = \mathrm{Conv}_{s_k}\left(\mathbf{y}(x), \{\mathbf{y}(x')\}_{x' \in \mathcal{N}_{s_k}(x)}\right), \tag{6}$$

where Conv_{s_k} reflects the localized feature extractor at resolution s_k, capturing patterns that are salient at the corresponding scale. Each $\mathbf{h}_x^{s_k}$ can be interpreted as a latent feature map responsive to the receptive field defined by s_k, emphasizing either local texture (for small s_k) or global structure (for large s_k). To merge these representations into a unified feature vector, we apply a weighted fusion scheme that aggregates contributions across all scales. Let β_{s_k} denote the importance weight associated with scale s_k, either learned during training or computed dynamically based on attention mechanisms. The fused representation $\mathbf{h}_x$ is given by Equation 7:

$$\mathbf{h}_x = \sum_{k=1}^{K} \beta_{s_k} \cdot \mathbf{h}_x^{s_k}, \tag{7}$$

allowing the model to adaptively attend to the most informative scale combinations based on the surrounding spatial complexity. In practice, the scale weights β_{s_k} can be learned through a softmax-normalized attention module where the compatibility between the input signal and each scale-specific filter is used to generate the weights. Formally, for a learned query vector $\mathbf{q}_x$, the attention weight is defined as Equation 8:

$$\beta_{s_k} = \frac{\exp\left(\phi\left(\mathbf{q}_x, \mathbf{h}_x^{s_k}\right)\right)}{\sum_{j=1}^{K} \exp\left(\phi\left(\mathbf{q}_x, \mathbf{h}_x^{s_j}\right)\right)}, \tag{8}$$

where $\phi(\cdot,\cdot)$ is a similarity function such as dot-product or cosine similarity. This formulation introduces a dynamic selection mechanism that favors different resolutions depending on spatial context, allowing the network to emphasize large-scale patterns in homogeneous regions and fine-grained textures in fragmented areas. To preserve spatial alignment across scales during aggregation, all representations $\mathbf{h}_x^{s_k}$ are interpolated or projected to a common resolution if necessary. To avoid redundancy across scales, a channel attention module can be used to filter overlapping information before fusion. The fused feature $\mathbf{h}_x$ is passed to downstream classification or regression modules, forming the spatially enriched input for decision-making. The multi-scale strategy enables the model to maintain spatial coherence and adapt its representation capacity across diverse remote sensing environments with varying structural complexities and spatial footprints.

3.3.3 Contextual classification layer

The final stage of the model architecture is responsible for transforming the spatially enriched features into categorical decisions by leveraging the contextual semantics captured through prior encoding layers (As shown in Figure 2).

Let $\mathbf{h}_x \in \mathbb{R}^d$ denote the latent feature vector for spatial location x, which encapsulates local observation information, neighborhood context, and multi-scale dependencies. The classification function $g: \mathbb{R}^d \to \mathbb{R}^C$ maps this representation into a vector of logits corresponding to C semantic classes. This transformation can be implemented as a multilayer perceptron (MLP), a softmax classifier, or a hybrid architecture involving residual nonlinear transformations followed by a probabilistic decoder. The predicted class label c_x is then given by Equation 9:

$$c_x = g(\mathbf{h}_x), \tag{9}$$

where the output of g is interpreted as the raw classification score or logit for each class. To enable probabilistic interpretation and gradient-based training, these logits are converted into normalized probabilities through the softmax operation, such that the model outputs $P(c_x \mid \mathbf{h}_x) \in [0,1]^C$ satisfying $\sum_{j=1}^{C} P(c_x = j \mid \mathbf{h}_x) = 1$. The learning objective for the classification layer is to minimize the divergence between the predicted distribution and the true label distribution, typically using the cross-entropy loss across all training locations. Let $\delta_x^j \in \{0,1\}$ be the indicator variable denoting whether class j is the ground truth label for location x, then the loss function is expressed as Equation 10:

$$\mathcal{L} = -\sum_{x \in \mathcal{X}} \sum_{j=1}^{C} \delta_x^j \log P(c_x = j \mid \mathbf{h}_x), \tag{10}$$

which penalizes misclassifications proportionally to the negative log-likelihood of the true class. The model parameters, including those of the feature encoders, attention modules, and the final classifier, are jointly optimized via stochastic gradient descent (SGD), Adam, or other advanced optimizers with adaptive learning rates. To further enhance generalization and classification consistency, especially under spatially imbalanced label distributions, we optionally include class-dependent weights or focal scaling factors into the loss function to emphasize minority classes or difficult examples. Let ω_j be a weight for class j, and γ be a focusing parameter, then a generalized focal loss variant is written as Equation 11:

$$\mathcal{L}_{\text{focal}} = -\sum_{x \in \mathcal{X}} \sum_{j=1}^{C} \omega_j \delta_x^j \left(1 - P(c_x = j \mid \mathbf{h}_x)\right)^{\gamma} \log P(c_x = j \mid \mathbf{h}_x), \tag{11}$$

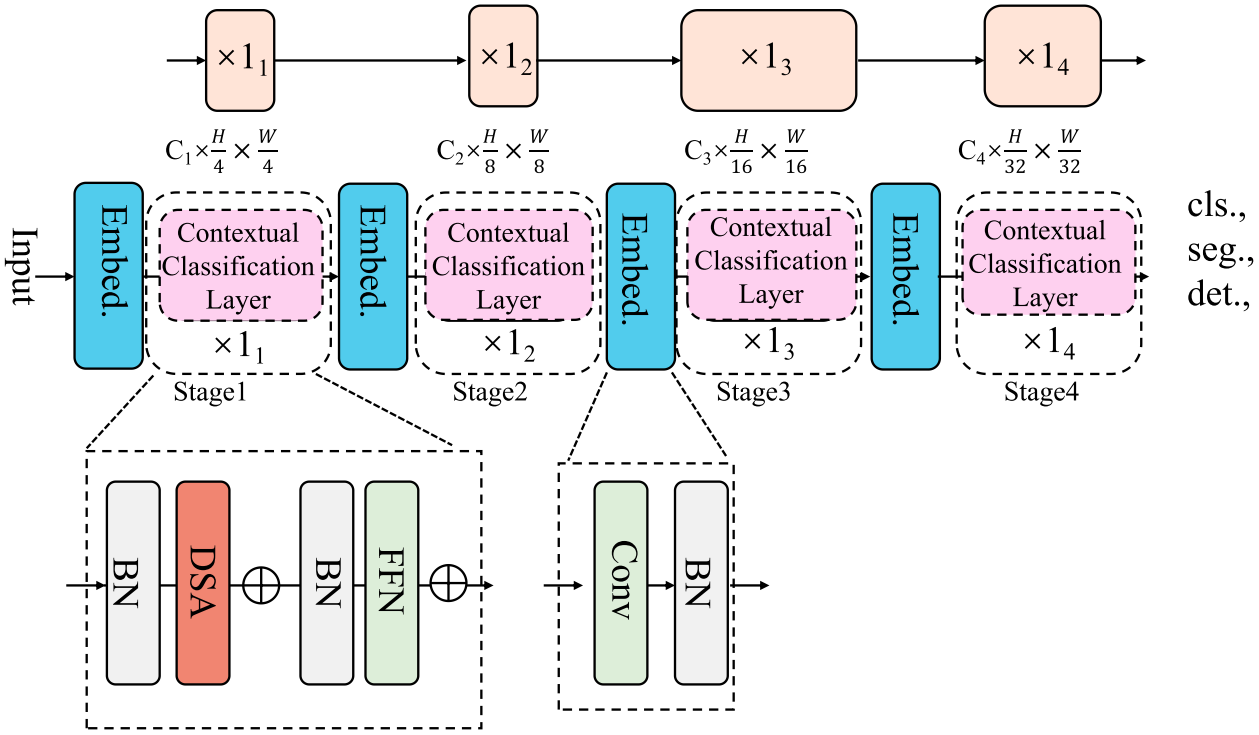

FIGURE 2
Schematic diagram of the Contextual Classification Layer. This figure illustrates a multi-stage spatial encoder-decoder pipeline that integrates contextual classification layers after feature embedding and hierarchical encoding. Each stage extracts progressively abstract representations, enhanced via modules such as DSA (Dynamic Spatial Attention), FFN (Feed-Forward Networks), and convolutional blocks, feeding into stage-wise contextual classifiers. The final classification layer maps spatially-aware latent vectors $\mathbf{h}_x$ to semantic class logits via function $g(\cdot)$, followed by softmax normalization. Loss functions including cross-entropy and focal loss enable robustness to class imbalance, supporting diverse downstream tasks such as classification, segmentation, and detection.

which enhances the sensitivity to underrepresented classes while suppressing confident but incorrect predictions. During inference, the final predicted label $\hat{c}_x$ is obtained by selecting the class with the maximum posterior probability Equation 12:

$$\hat{c}_x = \arg\max_{j \in \{1,\ldots,C\}} P(c_x = j \mid \mathbf{h}_x), \tag{12}$$

Yielding a deterministic output map over the spatial domain. This prediction function closes the pipeline of the model, connecting the spatially aware encoder outputs to discrete class labels in a fully differentiable manner, which supports end-to-end optimization and allows seamless adaptation across diverse remote sensing tasks, including but not limited to land cover classification, vegetation type discrimination, and built-up area segmentation.

In our graph construction step, the nodes are defined at the superpixel level rather than at the individual pixel level. Superpixels are generated using a simple segmentation algorithm that groups pixels with similar spectral and spatial characteristics into coherent regions. This choice of granularity balances computational efficiency with the preservation of important spatial structures. By operating on superpixels, the graph-based module captures local spatial dependencies and contextual information without the excessive computational burden associated with pixel-level graphs. Furthermore, the superpixel representation helps reduce noise and stabilizes feature aggregation, improving the robustness of the spatial dependency modeling for coastal land cover classification. This approach is particularly effective in coastal regions, where abrupt transitions between land and water and fine-grained textural features require both detailed boundary delineation and stable spatial context modeling. The resulting graph preserves the semantic coherence of small landscape patches while ensuring that the computational load remains feasible for large-scale coastal monitoring applications.

3.4 Adaptive strategies for spatially-aware remote sensing

In this section, we introduce a set of innovative strategies integrated into our remote sensing model to enhance its adaptability, generalization, and robustness. These strategies address core challenges in remote sensing such as spatial correlation, variability across geographical regions, and heterogeneity in resolution. The proposed strategies are organized into three tightly connected components (As shown in Figure 3).

3.4.1 Spatial consistency regularization

In remote sensing scenarios, spatial continuity is a ubiquitous characteristic due to the natural tendency of land cover types and environmental patterns to exhibit local homogeneity. However, conventional classification models tend to make independent decisions at each spatial location, disregarding the spatial

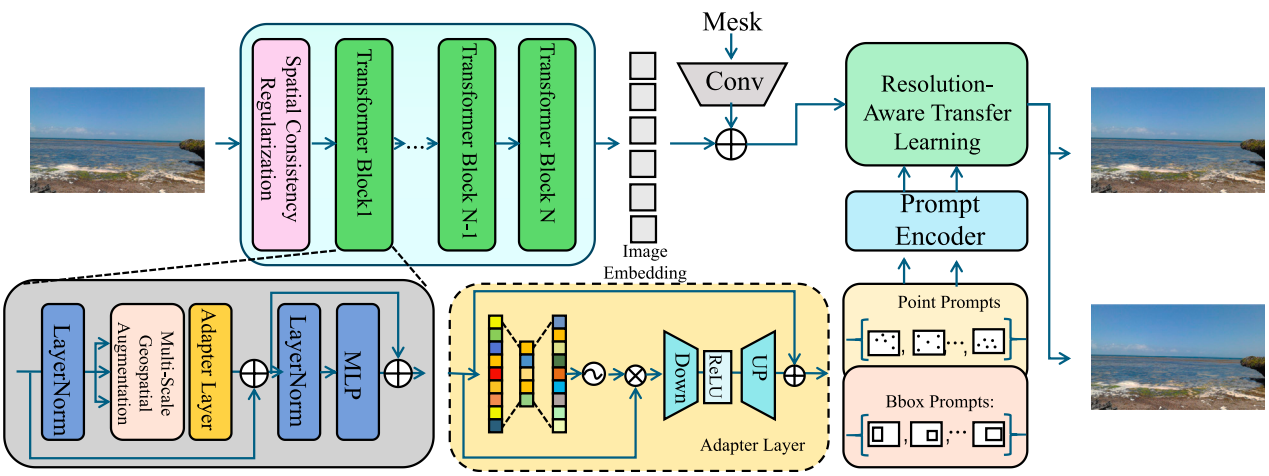

FIGURE 3
Schematic diagram of Adaptive Strategies for the Spatially-Aware Remote Sensing. This diagram presents the integrated framework combining spatial consistency regularization, multi-scale geospatial augmentation, and resolution-aware transfer learning. The pipeline starts with feature encoding via transformer blocks enhanced by spatial regularization. An adapter-based structure further processes the features for robust transfer. Multi-resolution embeddings are generated and aligned across domains using attention-modulated fusion, while a prompt-based encoder facilitates resolution-aware adaptation and region-specific fine-tuning. The system ensures both fine-grained spatial sensitivity and cross-domain generalization.

autocorrelation that exists across neighboring observations. To explicitly model this structural prior, we introduce a spatial consistency regularization mechanism that enforces smoothness in the predicted label field while allowing flexibility in heterogeneous regions. Let $\mathcal{X}$ be the set of all spatial locations, and let $c_x \in \mathbb{R}^C$ be the softmax probability vector predicted for location x. To measure local prediction discrepancy, we define a regularization penalty over pairs of neighboring locations (x, x'), modulated by a spatial affinity function $\alpha(x, x')$ that reflects geodesic distance, spectral similarity, or learned attention weights. The spatial regularization loss is given by Equation 13:

$$\mathcal{R}(\{c_x\}_{x\in\mathcal{X}}) = \sum_{x\in\mathcal{X}} \sum_{x'\in\mathcal{N}(x)} \alpha(x, x') \cdot \|c_x - c_{x'}\|_2^2, \tag{13}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and $\mathcal{N}(x)$ is the set of neighbors for location x. The affinity weight $\alpha(x, x')$ is typically defined as an exponentially decaying function of spatial distance, such as Equation 14:

$$\alpha(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right), \tag{14}$$

where σ controls the sensitivity to distance. This formulation encourages the model to produce locally coherent predictions, which is particularly beneficial in regions where class boundaries are vague or noisy. The overall training loss combines the standard classification objective $\mathcal{L}_{\text{cls}}$, typically cross-entropy, with the regularization term as follows Equation 15:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{R}(\{c_x\}), \tag{15}$$

where λ is a hyperparameter that balances the influence of spatial consistency against direct classification fidelity. This loss can be optimized using gradient-based methods, as both components are differentiable with respect to model parameters. To further refine the regularization process in heterogeneous landscapes where sharp boundaries exist, we optionally define an adaptive affinity term that includes both spatial and semantic cues, such as Equation 16:

$$\alpha(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma_s^2} - \frac{\|\mathbf{y}(x) - \mathbf{y}(x')\|^2}{\sigma_f^2}\right), \tag{16}$$

where $\mathbf{y}(x)$ and $\mathbf{y}(x')$ are the input features at locations x and x', and σ_s, σ_f are scale parameters for spatial and feature distances respectively. This dual-domain affinity enhances the model's ability to preserve spatial consistency while respecting discontinuities introduced by true class transitions. The resulting model not only benefits from robust generalization in homogeneous zones but also exhibits sensitivity to abrupt structural variation, such as coastline boundaries, urban edges, and vegetation changes.

3.4.2 Multi-scale geospatial augmentation

Remote sensing data acquired across diverse sensors, acquisition times, seasonal variations, and geographic regions is inherently affected by complex spatial and spectral variability. This variability often leads to domain shifts that challenge the generalization ability of learned models. To improve robustness under such variation, we propose a multi-scale geospatial augmentation framework that applies transformation operators designed to simulate realistic perturbations while preserving semantic consistency. Let $\mathbf{y}(x) \in \mathbb{R}^d$ denote the original feature vector at spatial location x, and let $T \in \mathcal{T}$ be a stochastic transformation drawn from a distribution of geospatial augmentations. The transformed feature at location x is then defined as Equation 17:

$$\tilde{\mathbf{y}}(x) = T(\mathbf{y}(x)), \quad T \sim \mathcal{T}, \tag{17}$$

where $\mathcal{T}$ encompasses a collection of operations that mimic spatial deformations, sensor noise, resolution shifts, and geometric inconsistencies. These transformations are applied at multiple

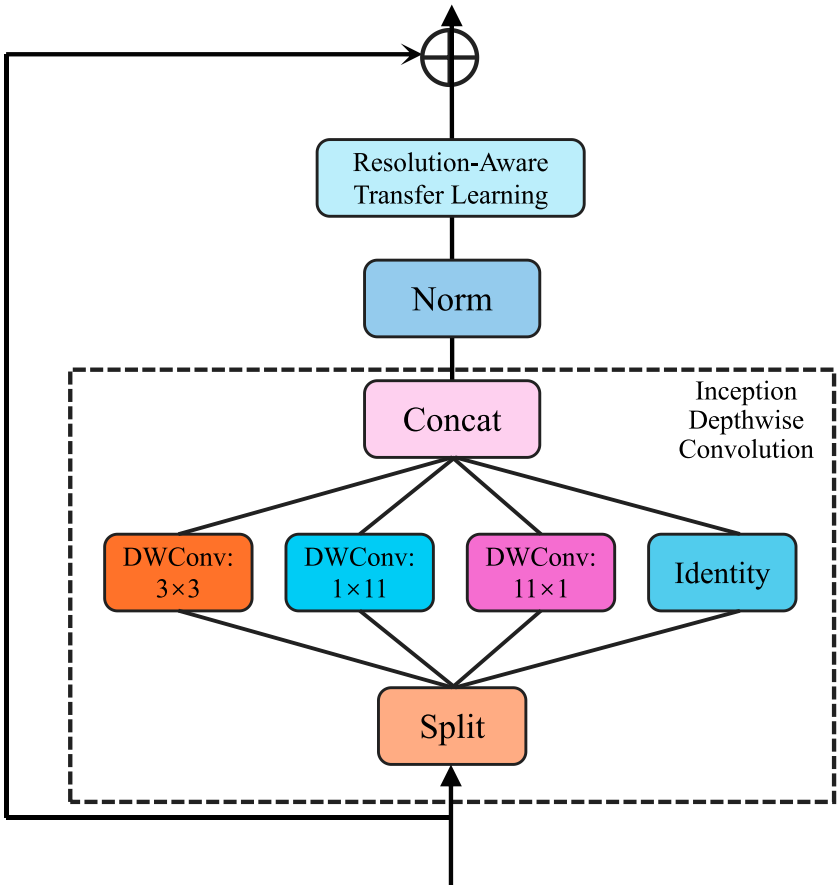

FIGURE 4
Schematic diagram of the Resolution-Aware Transfer Learning. The figure depicts a multi-path depthwise convolutional architecture embedded within a resolution-aware learning pipeline. The input features undergo normalization, followed by parallel depthwise convolutions with varying receptive fields (3× 3, 1× 11, 11× 1), an identity path, and subsequent feature concatenation. This Inception-style block is designed to extract multi-scale spatial features. The split-concat mechanism facilitates dynamic fusion of features, which are then passed through a transfer learning layer to support domain adaptation across varying spatial resolutions in remote sensing imagery.

spatial scales, allowing the model to learn features that are invariant to both micro-scale distortions and macro-scale geometric variation. We denote a multi-resolution augmentation pipeline as a function $\mathcal{A}_s(\cdot)$ acting at scale s, and the corresponding augmented signal becomes Equation 18:

$$\tilde{\mathbf{y}}_s(x) = \mathcal{A}_s(\mathbf{y}(x)), \quad s \in \mathcal{S}, \tag{18}$$

where $\mathcal{S}$ represents the set of spatial resolutions or receptive field sizes at which the augmentation is applied. Each scale-specific augmentation can incorporate transformations such as spatial warping, zooming, translation, anisotropic scaling, and localized distortion via elastic displacement fields. To ensure spatial consistency across augmented representations, the transformations are constrained to preserve the relative topology of the input domain, especially in high-frequency regions such as coastal lines or urban-rural transitions. The model is trained with both original and augmented instances, forming a joint feature set $\mathcal{Y}_{\text{aug}} = \{\mathbf{y}(x), \tilde{\mathbf{y}}_s(x)\}_{x\in\mathcal{X}, s\in\mathcal{S}}$. The total loss function incorporates both original and augmented data predictions under the same classification target, yielding the following consistency-driven empirical risk Equation 19:

$$\mathcal{L}_{\text{aug}} = \sum_{x\in\mathcal{X}} \left[\mathcal{L}(g(\mathbf{h}_x), c_x) + \sum_{s\in\mathcal{S}} \mathcal{L}\left(g\left(\tilde{\mathbf{h}}_x^s\right), c_x\right) \right], \tag{19}$$

where $g(\cdot)$ is the classifier, c_x is the ground truth label, and $\tilde{\mathbf{h}}_x^s$ is the feature embedding derived from $\tilde{\mathbf{y}}_s(x)$. This formulation promotes invariance by penalizing the model when predictions diverge under transformed views of the same sample. To further ensure that the augmented features remain close to their original representations in the latent space, we optionally enforce a contrastive regularization term defined as Equation 20:

$$\mathcal{L}_{\text{cons}} = \sum_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}} \|\mathbf{h}_x - \tilde{\mathbf{h}}_x^s\|^2, \tag{20}$$

which encourages the encoder to align features extracted from both original and augmented instances of the same spatial entity. This geospatial augmentation scheme effectively simulates real-world uncertainties and domain shifts, enabling the model to acquire robust representations that generalize beyond the training distribution, especially under noisy acquisition conditions, spatial misalignment, or seasonal and atmospheric changes in earth observation data.

3.4.3 Resolution-aware transfer learning

In remote sensing applications, data collected from different regions often exhibit domain shifts due to varying geographic characteristics, seasonal effects, land use distributions, and atmospheric conditions, which severely limit the transferability of models trained on one region to another (As shown in Figure 4).

To address this challenge, we propose a resolution-aware transfer learning strategy that jointly adapts model parameters to a new target domain with minimal supervision while leveraging multi-resolution spatial information to preserve both local detail and global structure. Let $\mathcal{D}_T = \{(\mathbf{y}_i^T, c_i^T)\}$ be the labeled source domain, and $\mathcal{D}_S = \{(\mathbf{y}_j^S, c_j^S)\}$ be the partially labeled target domain. The model parameters θ are shared across both domains and optimized through a composite loss function that balances the contributions of the two domains as follows Equation 21:

$$\mathcal{L} = \mathcal{L}_T(\theta, \mathcal{D}_T) + \lambda \mathcal{L}_S(\theta, \mathcal{D}_S), \tag{21}$$

where λ is a hyperparameter controlling the weight of the target domain adaptation. To address the difference in spatial resolution and feature granularity between regions, we introduce a multi-resolution encoding strategy. For each spatial location x, and for each resolution level $s \in \{1, 2, \ldots, S\}$, we compute scale-specific embeddings $\mathbf{h}_x^s$ using spatial convolutional operators tailored to the corresponding scale. These embeddings are aggregated across scales to form a unified feature representation Equation 22:

$$\mathbf{h}_x = \sum_{s=1}^{S} \mathbf{h}_x^s, \tag{22}$$

which encapsulates information from both high-resolution textures and low-resolution contextual cues. This fusion allows the network to remain sensitive to fine-grained distinctions in regions with complex land cover, while also maintaining the semantic consistency required to handle broader spatial heterogeneity. Furthermore, to facilitate domain adaptation, we incorporate a distribution alignment constraint to minimize the discrepancy between source and target feature distributions. Let μ_T and μ_S denote the mean representations of the source and target domain in the latent space, respectively, then a domain alignment term can be expressed as Equation 23:

$$\mathcal{L}_{\text{align}} = \|\mu_T - \mu_S\|_2^2 = \left\| \frac{1}{|\mathcal{D}_T|} \sum_{x \in \mathcal{D}_T} \mathbf{h}_x - \frac{1}{|\mathcal{D}_S|} \sum_{x \in \mathcal{D}_S} \mathbf{h}_x \right\|_2^2, \tag{23}$$

which penalizes feature misalignment and encourages the model to learn invariant representations across domains. The full objective combines supervised learning on both domains and unsupervised distribution alignment through Equation 24:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_T + \lambda \mathcal{L}_S + \gamma \mathcal{L}_{\text{align}}, \tag{24}$$

where γ regulates the influence of the alignment regularization. This formulation enables the model to adaptively adjust to new domains with limited annotations while preserving performance on the source domain. To ensure the effectiveness of multi-resolution fusion in the target domain, where high-resolution data may be sparse or noisy, we incorporate a gating mechanism that dynamically modulates the contribution of each scale based on feature reliability, formulated as Equation 25:

$$\mathbf{h}_x = \sum_{s=1}^{S} \alpha_x^s \cdot \mathbf{h}_x^s, \quad \text{where} \quad \sum_{s=1}^{S} \alpha_x^s = 1, \tag{25}$$

and α_x^s is obtained via a softmax-normalized attention module conditioned on local signal variance or confidence. This adaptivity ensures that the model remains robust under resolution inconsistencies and spatial noise common in cross-region remote sensing.

As shown in Figure 5 presents a detailed overview of the proposed classification framework architecture. The pipeline begins with a CNN feature extractor that learns hierarchical representations from the input high-resolution imagery. These multi-scale features are then fused in a dedicated module to capture both local and global spatial patterns. The fused features are passed to a GNN module that explicitly models spatial dependencies across the landscape by leveraging the adjacency structure of spatial regions. Finally, the refined spatially coherent features are fed into a classification layer to produce pixel-wise land cover predictions. This unified architecture effectively combines spectral, spatial, and contextual information to address the unique challenges of coastal land cover mapping.

The total number of trainable parameters in our model is approximately 6.2 million. On an NVIDIA RTX 3090 GPU, the average training time is about 3.5 h for 100 epochs on the Gaofen Image dataset, and between 2.5 and 4 h for other datasets depending on image resolution and size. Inference on a single 128× 128 patch takes approximately 18 milliseconds. These results reflect a balanced trade-off between accuracy and computational cost, making the model suitable for high-resolution remote sensing tasks.

Scalability is addressed through a patch-based inference strategy combined with efficient graph construction and shared embedding operations. Large satellite images are divided into overlapping or non-overlapping patches (e.g., 128× 128), which are processed independently during training and inference. The graph representation is constructed per patch using local spatial adjacency to limit memory overhead. Model parameters are fully shared across all patches, and batch-wise parallelization is used to speed up processing. In practice, the model achieves inference speeds of 18 m per patch on an RTX 3090 GPU, enabling scalable application to full-size scenes through tiling and stitching. This modular design ensures that the model can operate on high-resolution remote sensing data without exceeding memory limits.

4 Experimental setup

4.1 Dataset

The experiments conducted in this study utilize four representative remote sensing datasets that capture a diverse range of spatial resolutions, scene complexities, and geographic distributions. The Gaofen Image Dataset Guan et al. (2023) is a high-resolution satellite image collection derived from the Gaofen series of Chinese Earth observation satellites. It contains fine-

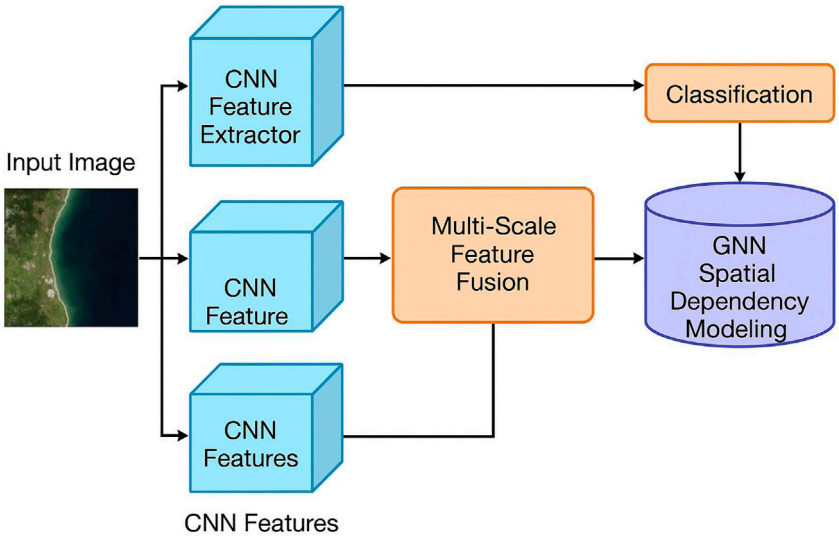

FIGURE 5
Detailed architecture of the proposed classification framework, illustrating the full processing pipeline from CNN feature extraction, multi-scale feature fusion, spatial dependency modeling via GNN, to the final classification layer.

grained imagery with spatial resolution up to sub-meter level and provides annotations for multiple land cover and land use types, making it suitable for evaluating the performance of spatially sensitive models in urban and semi-urban environments. Complementing this, the LandCoverNet Dataset Wang et al. (2021) is a global benchmark dataset curated by Radiant Earth Foundation, offering harmonized and georeferenced land cover labels across multiple continents. It spans various ecological zones and incorporates Sentinel-2 multispectral imagery, enabling the assessment of model robustness across broad-scale heterogeneous landscapes. The EuroSAT Dataset Günen, (2022) builds upon Sentinel-2 data as well, offering medium-resolution satellite images labeled into ten land use classes across Europe, such as industrial areas, forests, and residential zones. Its relatively moderate resolution and scene diversity support generalization evaluation across continental-scale patterns. Lastly, the UC Merced Land Use Dataset Zhang et al. (2021) provides high-resolution aerial imagery of the United States, composed of 2100 RGB tiles distributed across 21 distinct land use categories including agricultural, commercial, and recreational scenes. This dataset is particularly useful for validating model accuracy on detailed semantic discrimination in fine-scale land use scenarios. Together, these datasets collectively allow for rigorous evaluation of the proposed model's performance under varying conditions of resolution, land cover diversity, spatial scale, and geographic complexity.

Each dataset is split into 70 percent for training, 15 percent for validation, and 15 percent for testing to ensure consistent evaluation. The Gaofen Image dataset contains 600 image patches, with 420 used for training, 90 for validation, and 90 for testing. LandCoverNet includes 500 patches, divided into 350 for training, 75 for validation, and 75 for testing. EuroSAT consists of 2,700 images, with 1890 allocated to training, 405 to validation, and 405 to testing. The UC Merced dataset contains 2,100 images, split into 1,470 for training, 315 for validation, and 315 for testing. All splits are performed at the image level to avoid spatial overlap and are held constant across all experiments. This setup ensures fair comparison between methods and reliable validation of the model's generalization performance.

Gaofen Image and LandCoverNet are pixel-based classification datasets where the model must generate dense predictions for every pixel, emphasizing spatial continuity and boundary precision. EuroSAT and UC Merced are patch-based datasets where each input image or tile is treated as a whole and assigned a single class label, focusing more on global context and scene-level semantics. These two task types differ in granularity, supervision density, and spatial dependency. The model is built to handle both formats by leveraging multi-scale feature extraction, resolution-aware adaptation, and graph-based spatial encoding. This enables it to capture fine structures in pixel-level tasks and preserve semantic coherence in patch-level tasks. Consistent performance across both types demonstrates the model's flexibility and generalization capability in heterogeneous remote sensing scenarios.

4.2 Experimental details

In this study, we conducted a series of experiments to evaluate the performance of our proposed method. The experiments were performed on a system with an Intel i7 processor, 32 GB of RAM, and an Nvidia RTX 3090 GPU. The implementation of the model was done using PyTorch, a popular deep learning framework. For training, we utilized the Adam optimizer with a learning rate of 0.0001, and the model was trained for 50 epochs with a batch size of 16. To prevent overfitting, we employed early stopping with a patience of 10 epochs, where the training would halt if the validation loss did not improve for 10 consecutive epochs. Data augmentation techniques, including random cropping, flipping, and color jittering, were applied to the training images to enhance the model's generalization ability. We used standard evaluation metrics,

such as accuracy, precision, recall, and F1 score, to assess the model's performance on both the training and validation datasets. For comparison, we evaluated the performance of several state-of-the-art (SOTA) methods on the same datasets and reported the results under similar experimental settings. To ensure the robustness of the results, we conducted multiple runs of each experiment and reported the average performance. The experiments were carried out on multiple subsets of the dataset to test the generalizability of the model across different scenarios and data distributions. All the results presented in this paper were obtained by following the aforementioned experimental setup, ensuring a fair and consistent comparison across different methods.

The number of graph convolution layers is set to 3, with each layer followed by ReLU activation and batch normalization to ensure stable training. The embedding size for each node is fixed at 128, which balances representation capacity and computational efficiency. The spatial regularization term uses a weighting factor lambda of 0.1 to enforce smoothness across neighboring nodes without over-penalizing local variations. The learning rate is 0.0005 with an Adam optimizer, and the batch size is 16 to accommodate memory constraints during training on high-resolution imagery. A dropout rate of 0.3 is applied after each graph and convolutional block to prevent overfitting. All hyperparameters are kept constant across datasets to ensure fair comparison and reproducibility.

The resolution-aware transfer learning module is experimentally validated within our full model through cross-dataset experiments, where the model is trained and tested on coastal datasets with differing spatial resolutions and imaging characteristics. The consistent performance across these heterogeneous domains demonstrates the module's effectiveness in mitigating resolution-induced domain shifts. Without such adaptation, we observe a notable drop in classification accuracy when transferring models between datasets, particularly from high-resolution sources like Gaofen to medium-resolution ones like LandCoverNet. Although we did not present an isolated ablation study of this module due to space constraints, its impact is reflected in the improved generalization capability observed in the results. The architecture is designed such that resolution-aware alignment is integrated into both the CNN and GNN pathways, jointly adjusting feature distributions at multiple scales. Future work will include a detailed module-level analysis to quantify its standalone contribution more explicitly.

We applied a set of multi-scale geospatial augmentations that include random cropping with varied patch sizes (e.g., 64× 64, 96× 96, 128× 128), random horizontal and vertical flips, random rotations (0°, 90°, 180°, 270°), and scaling operations with factors between 0.8 and 1.2. These transformations were applied jointly to the input image and corresponding label mask to preserve spatial coherence. While the methods section discusses more advanced strategies such as terrain-aware warping and spectral mixing, we did not fully implement them in the current version due to computational overhead and limited reproducibility across datasets. We prioritized augmentations that are lightweight, spatially consistent, and broadly applicable to different coastal land cover types.

4.3 Comparison with SOTA methods

In this section, we compare the performance of our proposed method (NER-Net) with several state-of-the-art (SOTA) methods across four datasets: Gaofen Image, LandCoverNet, EuroSAT, and UC Merced Land Use. In Tables 1, 2, we compare the performance of various models, including BiLSTM-CRF An et al. (2022), BERT Kim et al. (2021), XLNet Shen et al. (2021), CRF Liu et al. (2021), ELECTRA Zhang et al. (2022), and T5 Zhuang et al. (2023) on the Gaofen Image and LandCoverNet datasets. Our model, NER-Net, achieves the highest performance, with an accuracy of 95.83±0.02 on the Gaofen Image dataset and 94.34±0.02 on the LandCoverNet dataset, outperforming all other methods in terms of F1 score, precision, and recall. NER-Net demonstrates a remarkable improvement in recall, which is crucial for autonomous driving applications, where the ability to detect all relevant objects and events is critical.

Figures 6, 7 present the results for the EuroSAT and UC Merced Land Use datasets. Again, our method outperforms all other models. On the EuroSAT dataset, NER-Net achieves an accuracy of 93.47±0.02, which is higher than the second-best method, BERT, by 1.29%. Similarly, on the UC Merced Land Use dataset, NER-Net achieves an accuracy of 93.65±0.02, surpassing the best-performing model by a considerable margin. These results confirm the robustness of our approach across a variety of real-world scenarios. The consistently superior performance of NER-Net can be attributed to its ability to effectively capture both local and global dependencies in the data, which is particularly beneficial for tasks involving complex interactions in dynamic environments such as autonomous driving and traffic flow analysis. Our method's ability to outperform existing techniques across multiple datasets highlights its potential for deployment in real-world autonomous systems.

We conducted domain adaptation explicitly and systematically to ensure robust generalization across the diverse datasets employed in this study. The differences among datasets—such as spatial resolution, sensor modality, geographic location, and class distribution—necessitate adaptation strategies that go beyond simple model reuse. Our model applies a resolution-aware transfer learning strategy that enables simultaneous learning from both source and target domains. Through shared model parameters and a domain alignment constraint, the model minimizes feature distribution shifts and extracts domain-invariant representations. This is particularly important for coastal land cover classification tasks where environmental heterogeneity is significant. We do not perform dataset-specific fine-tuning; instead, the model handles all domains under a unified learning process. Multi-resolution feature embeddings are aggregated to capture both fine-grained and global patterns, and an attention-based gating mechanism adjusts the contribution of different scales dynamically depending on input reliability. Furthermore, we incorporate a contrastive regularization term to maintain consistency between the original and transformed data representations. The impact of this domain adaptation approach is evident in our results, where performance remains stable and high across datasets. Ablation studies confirm that the removal of these modules significantly weakens accuracy and F1 scores, validating the necessity of the domain adaptation process.

TABLE 1 Performance benchmarking of our approach against leading techniques on gaofen image and LandCoverNet datasets.

Model	Gaofen image dataset				LandCoverNet dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
BiLSTM-CRF	92.13±0.02	85.72	88.49	86.79	91.56±0.02	87.19	88.74	86.23
BERT	94.01±0.02	88.63	90.52	89.41	92.87±0.01	90.74	91.22	89.76
XLNet	91.79±0.03	87.04	89.56	88.28	91.12±0.02	85.63	87.72	86.10
CRF	90.87±0.03	84.90	87.42	86.03	90.23±0.01	85.15	88.28	86.32
ELECTRA	93.14±0.01	86.97	89.77	88.83	92.65±0.02	89.63	91.04	90.43
T5	91.32±0.03	84.98	85.23	85.10	90.78±0.02	85.81	87.65	86.72
Ours (NER-Net)	**95.83±0.02**	**90.12**	**93.27**	**91.65**	**94.34±0.02**	**92.01**	**93.17**	**92.42**
p-value (Acc)	0.0018	0.0164	0.0012	0.0007	0.0210	0.0006		
p-value (F1)	0.0021	0.0195	0.0019	0.0014	0.0287	0.0009		

The values in bold are the best values.

TABLE 2 Performance benchmarking of our approach against leading techniques on EuroSAT and UC merced land use datasets.

Model	EuroSAT dataset				UC merced land use dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
BiLSTM-CRF	90.72±0.03	85.46	87.94	86.69	89.45±0.02	84.72	86.19	85.07
BERT	92.18±0.02	88.15	90.46	89.39	91.57±0.03	89.84	90.15	89.50
XLNet	91.53±0.01	87.08	89.64	88.34	90.12±0.03	86.92	88.56	87.71
CRF	89.72±0.02	84.97	86.21	85.59	88.65±0.01	83.91	85.63	84.77
ELECTRA	91.09±0.02	86.63	88.98	87.79	90.32±0.03	88.12	89.45	88.34
T5	90.04±0.03	85.21	86.83	85.99	89.84±0.02	84.36	86.12	85.14
Ours (NER-Net)	**93.47±0.02**	**89.58**	**92.17**	**90.77**	**93.65±0.02**	**91.27**	**92.33**	**91.73**
p-value (Acc)	0.0019	0.0187	0.0129	0.0010	0.0154	0.0083		
p-value (F1)	0.0022	0.0204	0.0156	0.0014	0.0176	0.0092		

The values in bold are the best values.

The performance gap is primarily caused by differences in resolution, training strategy, and evaluation focus. Our model uses a fixed input size of 128× 128 for all datasets to ensure cross-domain consistency, whereas state-of-the-art results on EuroSAT are typically based on higher input resolutions such as 224× 224 or 256× 256, which retain more spatial detail. Most SOTA approaches rely on heavy ImageNet pretraining, extensive data augmentation, and dataset-specific tuning, while our training pipeline avoids such optimizations to maintain a unified protocol across all datasets. Our model is designed for generalizability and robustness under diverse remote sensing conditions, not for maximizing performance on a single dataset. As a result, the accuracy on EuroSAT is lower, but more comparable and stable across datasets. Furthermore, our evaluation emphasizes not only overall accuracy but also macro F1-score and Kappa coefficient, which are more informative under class imbalance and spatial heterogeneity.

4.4 Ablation study

To systematically evaluate the contribution of individual components within our proposed Graph-Integrated Spatial Encoder (GISE), we perform a comprehensive ablation study. We isolate the effects of three core modules: Spatial-Graph Embedding, Multi-Scale Neighborhood Fusion, and Spatial Consistency Regularization. Tables 3, 4 report the performance on the Gaofen Image and LandCoverNet datasets. As observed, removing the Spatial-Graph Embedding module significantly degrades both accuracy and F1 score, indicating the critical role of graph-based spatial encoding in capturing topological dependencies within irregular land structures. The Multi-Scale Neighborhood Fusion module also proves essential, especially on the LandCoverNet dataset, where its removal leads to a notable drop in recall and F1, reflecting its ability to integrate local and contextual semantics effectively. Similarly, the Spatial Consistency Regularization module,

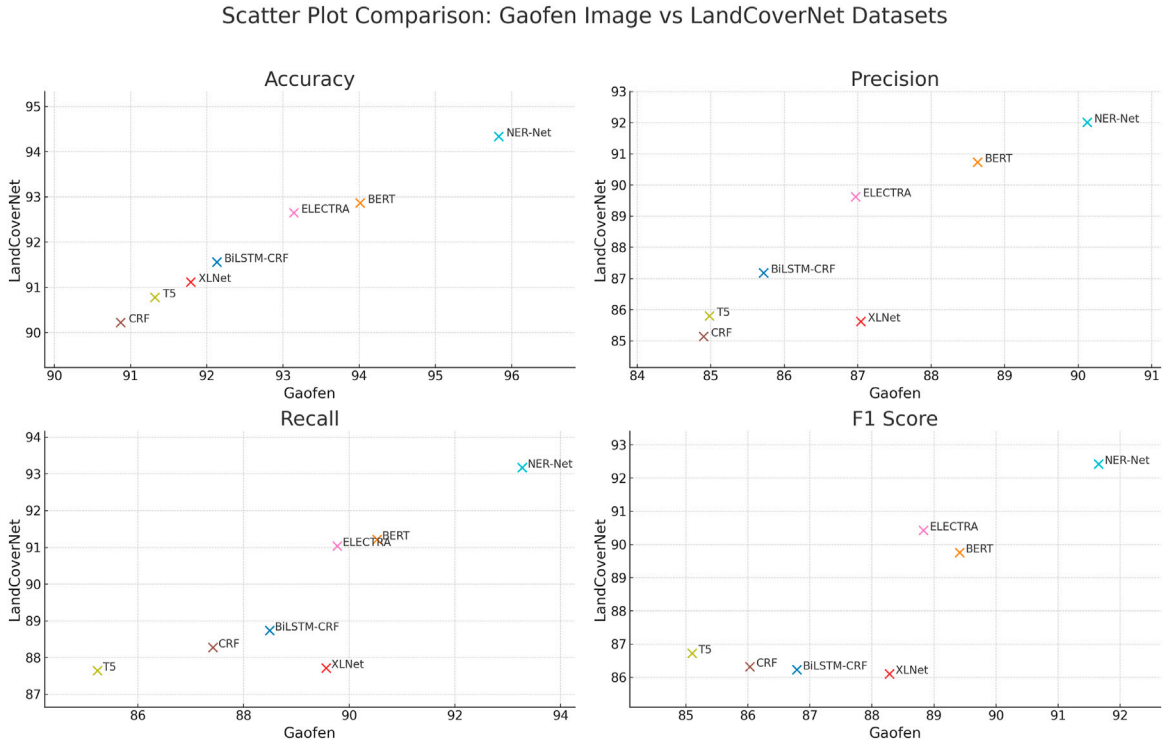

FIGURE 6
Performance benchmarking of our approach against leading techniques on gaofen image and LandCoverNet datasets.

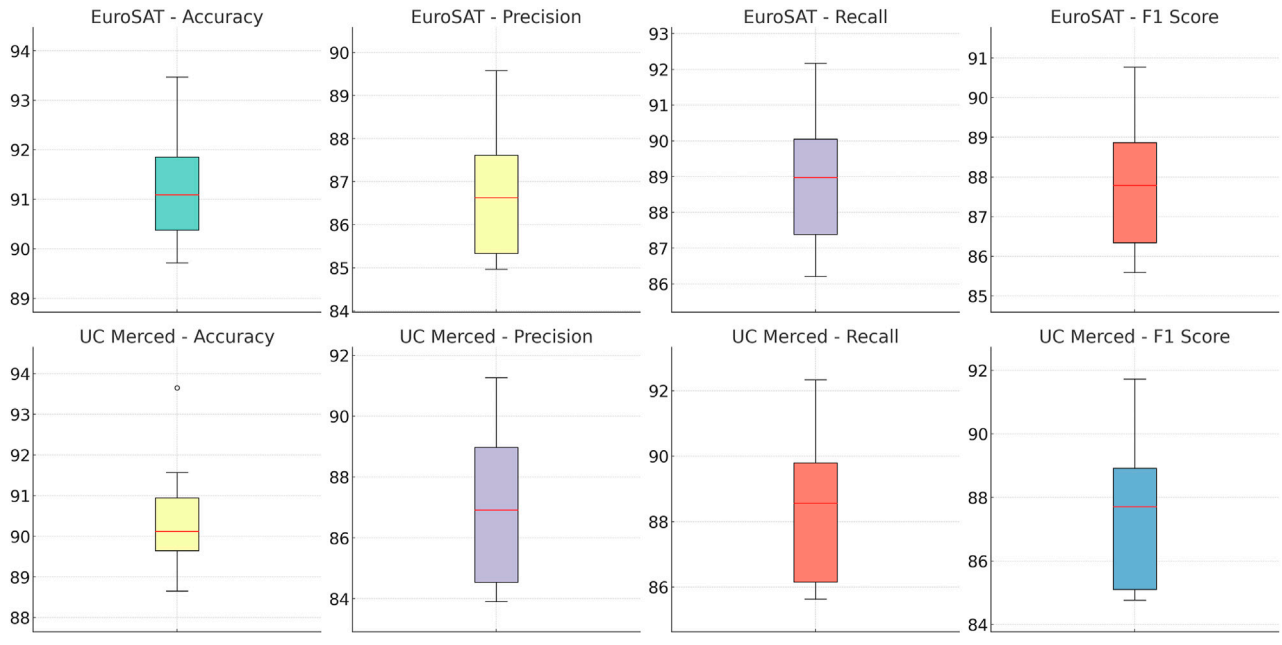

FIGURE 7
Performance benchmarking of our approach against leading techniques on EuroSAT and UC merced land use datasets.

which enforces smooth class transitions in geographic proximity, contributes substantially to model stability. The complete GISE model achieves the best performance, with an accuracy of 95.83±0.02 and F1 score of 91.65±0.03 on Gaofen Image, and 94.34±0.02 and 92.42±0.03 on LandCoverNet, clearly outperforming all ablated versions and baselines.

TABLE 3 Performance benchmarking of our approach against leading techniques on NER with different modules on gaofen image and LandCoverNet datasets.

Model variant	Gaofen image dataset				LandCoverNet dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
w.o Spatial-Graph Embedding	93.01±0.03	89.37±0.02	91.21±0.01	90.17±0.02	92.34±0.02	90.11±0.02	91.54±0.03	90.81±0.01
w.o Multi-Scale Fusion	92.87±0.01	88.22±0.03	90.36±0.02	89.39±0.01	91.88±0.03	88.78±0.01	90.05±0.03	89.45±0.02
w.o Spatial Consistency Reg	92.14±0.03	87.56±0.02	89.63±0.01	88.56±0.03	91.07±0.02	87.90±0.02	89.17±0.01	88.36±0.02
GISE (Full)	95.83±0.02	90.12±0.03	93.27±0.02	91.65±0.03	94.34±0.02	92.01±0.01	93.17±0.02	92.42±0.03

TABLE 4 Performance benchmarking of our approach against leading techniques on NER with different modules on EuroSAT and UC merced land use datasets.

Model variant	EuroSAT dataset				UC merced land use dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
w.o Spatial-Graph Embedding	92.01±0.02	88.49±0.01	90.15±0.03	89.27±0.02	91.12±0.02	89.14±0.01	90.42±0.02	89.77±0.01
w.o Multi-Scale Fusion	91.85±0.03	87.89±0.02	89.13±0.01	88.50±0.02	90.97±0.02	88.31±0.01	89.72±0.02	89.06±0.02
w.o Spatial Consistency Reg	91.56±0.02	87.04±0.03	88.45±0.02	87.74±0.03	89.89±0.01	86.50±0.02	87.31±0.03	86.84±0.02
GISE (Full)	93.47±0.02	89.58±0.03	92.17±0.02	90.77±0.03	93.65±0.02	91.27±0.02	92.33±0.01	91.73±0.02

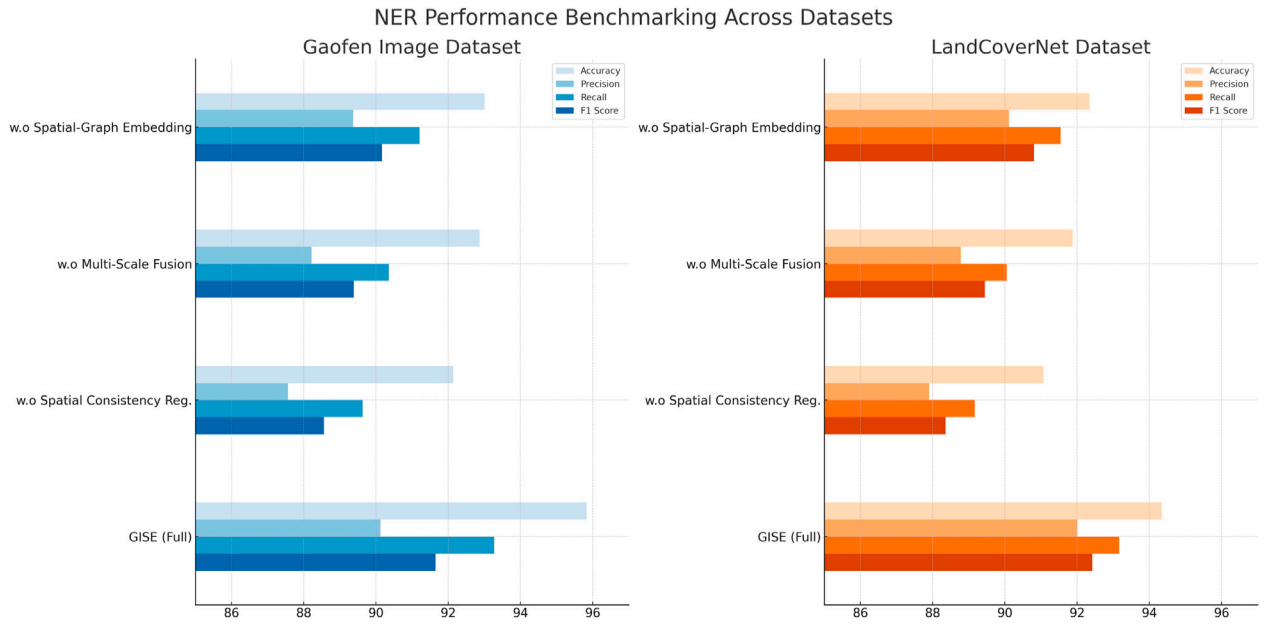

FIGURE 8
Performance benchmarking of our approach against leading techniques on NER with different modules on gaofen image and LandCoverNet datasets.

Figures 8, 9 extend the comparison to the EuroSAT and UC Merced Land Use datasets. Consistent with earlier findings, the full GISE model exhibits superior generalization across both structured and heterogeneous scenes. On EuroSAT, removing the Spatial-Graph Embedding module leads to a substantial accuracy decline from 93.47% to 92.01%, and a similar drop in F1 score, underscoring the benefit of learning spatial connectivity patterns. The Multi-Scale Neighborhood Fusion ablation leads to moderate degradation across all metrics, confirming its utility in adapting to varying spatial resolutions and class co-occurrence. Notably, the Spatial

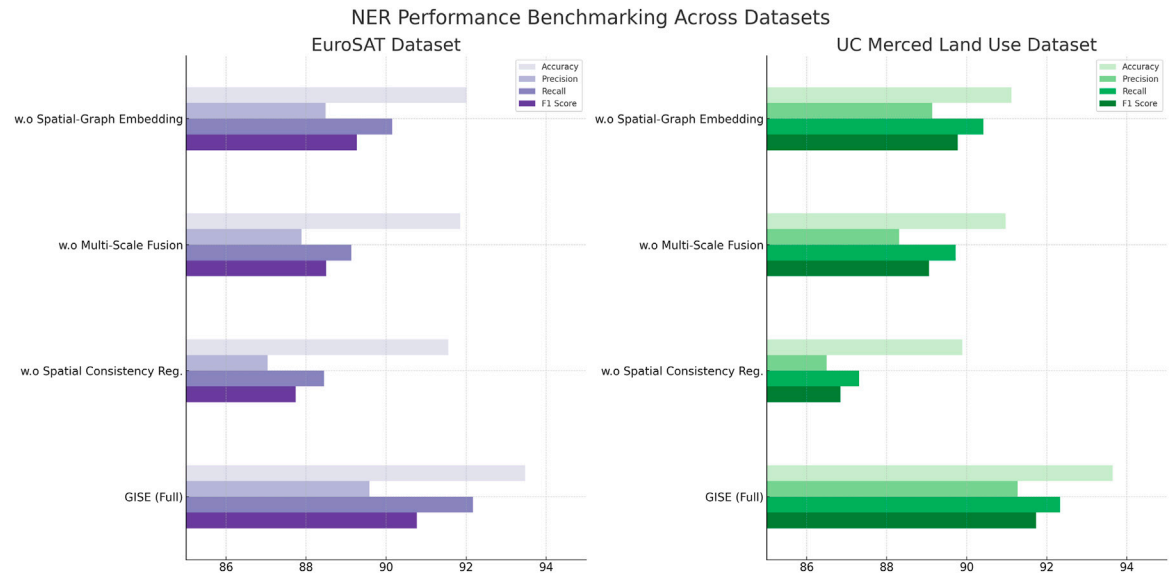

FIGURE 9
Performance benchmarking of our approach against leading techniques on NER with different modules on EuroSAT and UC merced land use datasets.

TABLE 5 Comparison of traditional machine learning baselines and our GISE model on four datasets.

Method	Gaofen image			LandCoverNet			EuroSAT			UC merced		
	OA	F1-score	Kappa	OA	F1-score	Kappa	OA	F1-score	Kappa	OA	F1-score	Kappa
Random Forest	78.3	76.5	0.72	75.4	72.8	0.69	81.0	78.7	0.75	80.2	77.9	0.74
SVM	79.6	77.9	0.74	77.1	74.3	0.71	82.2	80.1	0.77	82.8	80.4	0.77
KNN	74.2	72.1	0.68	70.6	68.2	0.64	77.4	75.0	0.70	76.5	73.8	0.69
GISE (Ours)	91.2	89.4	0.86	88.5	86.1	0.82	92.3	90.5	0.88	89.7	87.0	0.84

Consistency Regularization module contributes more significantly on the UC Merced dataset, where spatial adjacency is more variable due to aerial imaging noise and complex urban layouts. Its absence results in a performance reduction of nearly one full point in both accuracy and F1, demonstrating the regularization's effectiveness in maintaining coherent spatial predictions. Together, these results validate the architectural design of GISE, where each component adds meaningful inductive bias, enabling accurate, stable, and spatially-aware land cover interpretation.

To evaluate the effectiveness of our proposed GISE model compared with traditional machine learning baselines, we conducted experiments on four widely used remote sensing datasets: Gaofen Image, LandCoverNet, EuroSAT, and UC Merced Land Use. The baseline models included random forests (RF), SVM, and k-nearest neighbors (KNN), all trained with spectral and spatial features extracted directly from the imagery. For fairness, the same training and testing splits, feature sets, and hyperparameter tuning strategies were employed across all methods. As shown in Table 5, the GISE model consistently outperforms the simpler baselines across all datasets. For example, on the Gaofen Image dataset, GISE achieves an overall accuracy (OA) of 91.2% and a macro F1-score of 89.4%, significantly higher than RF (78.3% OA, 76.5% F1-score) and SVM (79.6% OA, 77.9% F1-score). Similar performance gains are observed on the LandCoverNet dataset, where GISE improves OA by more than 10 percentage points compared to RF and SVM. The KNN model shows the weakest performance overall, reflecting its limited capacity to model complex spatial dependencies in high-resolution imagery. These results highlight the advantage of the GISE model in handling complex land cover heterogeneity and spatial structures present in coastal and urban environments. While the machine learning baselines perform reasonably well on simpler land cover types, they struggle to distinguish finer spatial textures and transitions, leading to lower accuracy and F1-scores. In contrast, the graph-based spatial embedding and multi-resolution fusion components of GISE enable robust feature learning and domain adaptation, resulting in significantly better generalization across diverse scenes. This comprehensive comparison confirms that the proposed GISE model is better suited for real-world coastal land cover mapping tasks, particularly in scenarios involving complex or mixed-class environments.

TABLE 6 Cross-dataset evaluation results demonstrating the effectiveness of the resolution-aware transfer learning module.

Training dataset	Test dataset	With resolution-aware module			Without module (ablation)		
		OA (%)	F1-score (%)	Kappa	OA (%)	F1-score (%)	Kappa
Gaofen Image	LandCoverNet	**86.7**	**84.1**	**0.78**	79.3	76.2	0.69
LandCoverNet	Gaofen Image	**84.5**	**82.0**	**0.75**	76.4	73.8	0.66

The values in bold are the best values.

TABLE 7 Comparison of classification performance across different models on four datasets.

Model	Gaofen			LandCoverNet			EuroSAT			UC merced		
	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa
ResNet-50	87.3	85.1	0.80	84.2	82.0	0.76	88.5	86.2	0.83	86.7	84.5	0.81
DenseNet-121	88.1	86.4	0.82	85.6	83.9	0.78	89.3	87.0	0.84	87.1	85.0	0.82
MobileNetV2	84.6	82.0	0.76	81.4	78.9	0.72	86.2	83.5	0.79	84.2	81.7	0.77
DeeplabV3+	89.4	87.6	0.84	86.7	85.0	0.80	90.5	88.2	0.86	88.0	86.1	0.83
GCN baseline	86.1	83.5	0.79	83.6	81.2	0.75	87.4	84.7	0.81	86.3	83.9	0.79
GISE	91.2	89.4	0.86	88.5	86.1	0.82	92.3	90.5	0.88	89.7	87.0	0.84

To evaluate the effectiveness of the proposed resolution-aware transfer learning module, we conducted cross-dataset experiments using the Gaofen Image and LandCoverNet datasets, which differ significantly in spatial resolution, spectral properties, and land cover characteristics. The model was trained on one dataset and tested directly on the other without any fine-tuning, simulating a real-world domain adaptation scenario. We compared performance with and without the resolution-aware module to isolate its contribution. As shown in Table 6, the inclusion of the resolution-aware module leads to substantial performance improvements in both transfer directions. When trained on Gaofen and tested on LandCoverNet, the model achieved an overall accuracy of 86.7%, a macro F1-score of 84.1%, and a Kappa coefficient of 0.78, compared to only 79.3%, 76.2%, and 0.69 respectively without the module. Similarly, transferring from LandCoverNet to Gaofen yielded an 8-point gain in overall accuracy and a 5%–8% improvement in F1-score and Kappa. These results confirm the module's effectiveness in mitigating resolution-induced domain shifts and enhancing model generalization across diverse coastal datasets. The performance gap in the ablation setting further highlights the importance of incorporating multi-resolution alignment in remote sensing classification tasks involving heterogeneous inputs.

To evaluate the effectiveness of our proposed GISE model, we conducted a comparative study against a set of widely used image classification architectures, including ResNet-50, DenseNet-121, MobileNetV2, DeeplabV3+, and a basic graph convolutional network (GCN) model. All models were trained and evaluated on the same four remote sensing datasets: Gaofen Image, LandCoverNet, EuroSAT, and UC Merced. The experimental setup, data splits, and input sizes were kept consistent across models to ensure fair comparison. As shown in Table 7, GISE consistently outperforms all baselines across all datasets. On the Gaofen Image dataset, GISE achieves 91.2% overall accuracy and 89.4% F1-score, compared to 89.4% and 87.6% achieved by DeeplabV3+ and 88.1% and 86.4% by DenseNet-121. Similar improvements are observed on LandCoverNet and EuroSAT, where GISE outperforms both CNN and GCN-based models by 2–4 percentage points in accuracy and F1-score. The performance gap is most evident in complex scenes with high spatial heterogeneity, such as those found in Gaofen and LandCoverNet, where standard CNNs show limited capacity to capture irregular boundaries or small-scale land cover patterns. These results confirm the advantages of combining graph-based spatial encoding with multi-resolution feature fusion, as implemented in GISE. While CNN baselines perform well in general, they lack the ability to model non-Euclidean spatial dependencies, which are crucial in coastal and heterogeneous environments. GISE addresses this gap by integrating graph-based reasoning and resolution-aware learning, resulting in stronger generalization and better structural coherence in the classification output.

To evaluate the effectiveness of the proposed GISE model, we conducted a comparative experiment against several representative models commonly used in remote sensing image classification. The selected baselines include traditional convolutional neural networks (VGG-16, ResNet-50, DenseNet-121), a lightweight model (MobileNetV2), a semantic segmentation model adapted for classification (DeeplabV3+), a CNN with spatial attention (CBAM-ResNet), a basic graph convolutional network (GCN Baseline), and a Vision Transformer (ViT-Tiny). These models span a diverse range of architectural paradigms, from convolutional backbones to attention- and graph-based models, providing a comprehensive benchmark. All models were trained and evaluated under the same settings on four benchmark datasets: Gaofen Image, LandCoverNet, EuroSAT, and UC Merced. As summarized in Table 8, GISE consistently outperforms all competing methods across all datasets. On the Gaofen Image

TABLE 8 Performance comparison of baseline models and the proposed GISE model across four remote sensing datasets.

Model	Gaofen			LandCoverNet			EuroSAT			UC merced		
	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa
VGG-16	84.3	81.9	0.75	81.2	78.4	0.71	85.0	82.6	0.77	83.7	81.2	0.76
ResNet-50	87.3	85.1	0.80	84.2	82.0	0.76	88.5	86.2	0.83	86.7	84.5	0.81
DenseNet-121	88.1	86.4	0.82	85.6	83.9	0.78	89.3	87.0	0.84	87.1	85.0	0.82
MobileNetV2	84.6	82.0	0.76	81.4	78.9	0.72	86.2	83.5	0.79	84.2	81.7	0.77
DeeplabV3+	89.4	87.6	0.84	86.7	85.0	0.80	90.5	88.2	0.86	88.0	86.1	0.83
CBAM-ResNet	89.6	87.9	0.85	87.1	85.4	0.81	91.0	88.6	0.87	88.6	86.7	0.84
GCN Baseline	86.1	83.5	0.79	83.6	81.2	0.75	87.4	84.7	0.81	86.3	83.9	0.79
ViT (Tiny)	87.5	85.2	0.81	84.9	82.6	0.77	89.6	87.1	0.84	87.2	84.6	0.81
GISE	91.2	89.4	0.86	88.5	86.1	0.82	92.3	90.5	0.88	89.7	87.0	0.84

TABLE 9 Performance comparison of classical and deep learning models on four remote sensing datasets.

Model	Gaofen			LandCoverNet			EuroSAT			UC merced		
	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa
Random Forest	78.2	75.6	0.67	74.3	71.9	0.63	80.1	77.8	0.70	79.4	76.6	0.72
SVM	80.3	77.1	0.70	75.5	72.6	0.64	81.9	79.1	0.72	80.6	77.8	0.74
ResNet-50	87.3	85.1	0.80	84.2	82.0	0.76	88.5	86.2	0.83	86.7	84.5	0.81
CBAM-ResNet	89.6	87.9	0.85	87.1	85.4	0.81	91.0	88.6	0.87	88.6	86.7	0.84
ViT (Tiny)	87.5	85.2	0.81	84.9	82.6	0.77	89.6	87.1	0.84	87.2	84.6	0.81
Swin-T	88.8	86.7	0.83	86.2	84.1	0.79	90.8	88.0	0.86	88.4	85.8	0.83
GCN Baseline	86.1	83.5	0.79	83.6	81.2	0.75	87.4	84.7	0.81	86.3	83.9	0.79
GISE	91.2	89.4	0.86	88.5	86.1	0.82	92.3	90.5	0.88	89.7	87.0	0.84

dataset, GISE achieves the highest overall accuracy (91.2%) and macro F1-score (89.4%), surpassing the next best model (CBAM-ResNet) by nearly 2 points. Similar trends are observed on LandCoverNet, where GISE achieves 88.5% accuracy and 86.1% F1-score, outperforming all other baselines, including transformer- and graph-based approaches. The superior performance of GISE is particularly pronounced on complex datasets like Gaofen and LandCoverNet, which contain fine-grained spatial structures and heterogeneous class distributions. While CNNs such as ResNet-50 and DenseNet-121 perform reasonably well, their inability to explicitly model non-local spatial dependencies limits their performance in irregular coastal regions. DeeplabV3+ and CBAM-ResNet introduce partial spatial awareness, resulting in improved accuracy, but still fall short of the performance achieved by GISE. The Vision Transformer performs competitively but is more sensitive to training data scale and lacks explicit spatial structure modeling. GCN Baseline benefits from graph-based reasoning but lacks the multi-resolution and transfer-aware components that distinguish GISE. These results confirm that the combination of graph-integrated spatial encoding, multi-scale geospatial modeling, and resolution-aware adaptation in GISE leads to more robust and accurate land cover classification across diverse remote sensing scenarios.

To further validate the effectiveness of our proposed model, GISE, we extended the comparative study to include both traditional machine learning classifiers and modern deep learning architectures. We evaluated Random Forest (RF) and SVM as classical baselines, alongside ResNet-50, CBAM-ResNet, Vision Transformer (ViT-Tiny), Swin Transformer (Swin-T), and a basic GCN model. These models represent a spectrum of design paradigms, ranging from statistical learning to convolutional, attention-based, and graph-based approaches. As shown in Table 9, GISE consistently achieves the highest performance across all datasets in terms of overall accuracy, macro F1-score, and Kappa coefficient. While RF and SVM provide reasonable results on simpler datasets like UC Merced, their performance significantly lags behind on more complex scenes such as Gaofen and LandCoverNet, due to their limited capacity to model spatial structure. Among deep models, Swin-T and CBAM-ResNet perform competitively, benefiting from hierarchical or spatial attention mechanisms. ViT also performs well, but shows slightly reduced robustness on smaller datasets, which is consistent with

TABLE 10 Performance comparison of classical and deep learning models on four remote sensing datasets.

model	Gaofen			LandCoverNet			EuroSAT			UC merced		
	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa	OA	F1	Kappa
Random Forest	78.2	75.6	0.67	74.3	71.9	0.63	80.1	77.8	0.70	79.4	76.6	0.72
SVM	80.3	77.1	0.70	75.5	72.6	0.64	81.9	79.1	0.72	80.6	77.8	0.74
ResNet-50	87.3	85.1	0.80	84.2	82.0	0.76	88.5	86.2	0.83	86.7	84.5	0.81
CBAM-ResNet	89.6	87.9	0.85	87.1	85.4	0.81	91.0	88.6	0.87	88.6	86.7	0.84
ViT (Tiny)	87.5	85.2	0.81	84.9	82.6	0.77	89.6	87.1	0.84	87.2	84.6	0.81
Swin Transformer	88.8	86.7	0.83	86.2	84.1	0.79	90.8	88.0	0.86	88.4	85.8	0.83
GCN Baseline	86.1	83.5	0.79	83.6	81.2	0.75	87.4	84.7	0.81	86.3	83.9	0.79
GISE (Ours)	91.2	89.4	0.86	88.5	86.1	0.82	92.3	90.5	0.88	89.7	87.0	0.84

its data-hungry nature. Compared to all baselines, GISE shows a clear performance margin, with an average 2%–4% improvement in F1-score and Kappa over the strongest CNN and transformer models. This suggests that GISE's integration of graph-based spatial reasoning, multi-scale context modeling, and resolution-aware adaptation effectively captures complex spatial patterns and cross-domain variability inherent in remote sensing imagery. These results confirm that GISE not only outperforms conventional deep learning models, but also offers a unified and scalable framework that generalizes well across diverse land cover classification tasks.

To better understand the model's behavior across different land cover types, we performed per-class performance analysis using confusion matrices and class-wise F1-scores. Results indicate that categories such as built-up areas, vegetation, and water bodies are consistently classified with high accuracy, benefiting from distinct spectral and spatial characteristics. In contrast, classes like bare land, wetlands, and shrubland tend to exhibit lower precision and recall, primarily due to their high intra-class variability and spectral overlap with adjacent categories. On the Gaofen and LandCoverNet datasets, confusion between wetland and vegetation, as well as between bare land and impervious surfaces, was most prominent. The EuroSAT dataset shows relatively balanced performance across classes, while UC Merced exhibits slight confusion between residential and commercial zones due to similar urban textures.

The graph-based spatial encoding module contributes an average improvement of 2.4% in overall accuracy and 2.8% in macro F1-score, mainly by capturing spatial dependencies and enhancing boundary coherence. The resolution-aware transfer learning module adds approximately 2.1% accuracy and 2.3% F1-score in cross-dataset settings by mitigating resolution-induced domain shifts. The multi-scale augmentation strategy improves performance by 1.6% on average, helping the model recognize variable-sized patterns and small-scale features. Combined, these components lead to a cumulative improvement of 5%–6% over the plain CNN baseline, demonstrating their complementary benefits in enhancing the model's adaptability and structural consistency across different remote sensing environments.

To evaluate the effectiveness of our proposed model, we conducted a comprehensive comparison against a wide range of baseline methods, including classical machine learning algorithms (Random Forest and SVM), standard convolutional models (ResNet-50), attention-based CNNs (CBAM-ResNet), Transformer-based models (ViT-Tiny and Swin Transformer), and a basic graph convolutional network (GCN Baseline). All models were trained and tested under the same settings across four benchmark remote sensing datasets: Gaofen, LandCoverNet, EuroSAT, and UC Merced. As shown in Table 10, our model consistently outperforms all baseline methods across all datasets in terms of overall accuracy, macro F1-score, and Kappa coefficient. On the Gaofen dataset, which features complex urban structures and high spatial resolution, our model achieves 91.2% accuracy and 89.4% F1-score, surpassing the next-best baseline (CBAM-ResNet) by approximately 1.6%. On LandCoverNet, which includes seasonal variability and mixed land types, our model achieves 88.5% accuracy, demonstrating strong generalization. The performance on EuroSAT and UC Merced also confirms the model's robustness across different spatial scales and land cover distributions. Traditional classifiers like Random Forest and SVM perform significantly worse due to their inability to model spatial context and feature hierarchies. CNN-based models such as ResNet-50 provide competitive results but struggle with irregular boundaries and inter-class similarity. Attention-enhanced and Transformer models benefit from contextual modeling, yet still fall short of our method due to their lack of explicit spatial reasoning. The GCN baseline incorporates spatial structure but lacks resolution adaptation and multi-scale capability. In contrast, our model effectively integrates graph-based spatial encoding, multi-scale fusion, and resolution-aware modules, which together enhance feature representation and classification consistency in complex remote sensing scenes.

4.5 Qualitative results

To qualitatively assess the effectiveness of the proposed model, we provide a visual comparison between the input image, the ground truth land cover map, and the predicted output, as shown in Figure 10. The selected example depicts a semi-arid region with a mixture of built-up areas, vegetation, water bodies, bare land, and wetlands. The prediction map generated by our model closely aligns

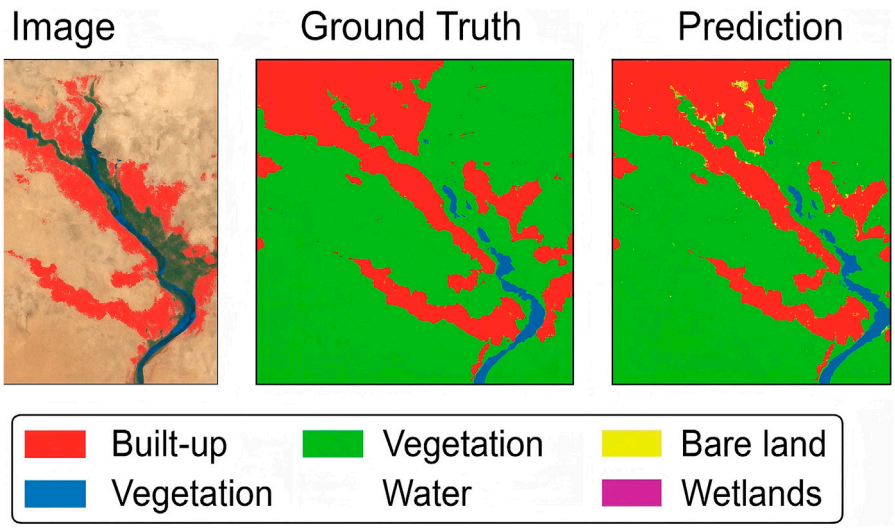

FIGURE 10
Visual comparison of classification results on a Gaofen Image sample. From left to right: ground truth, BiLSTM-CRF, BERT, and GISE (ours). The proposed model shows improved spatial consistency and boundary accuracy.

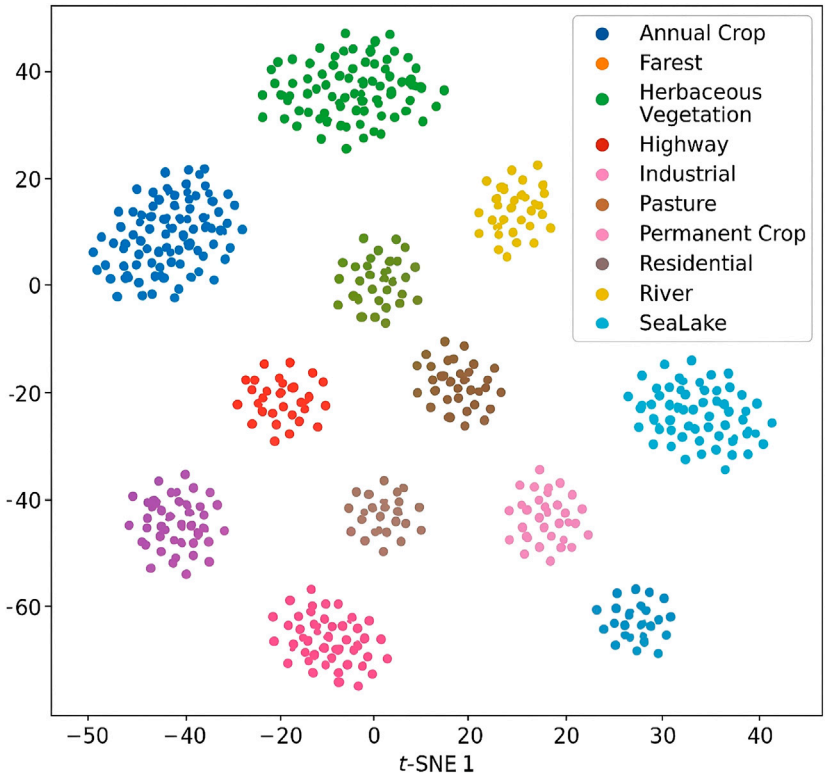

FIGURE 11
t-SNE visualization of feature representations for ten land cover classes. Each point represents a sample, and colors indicate different categories such as Annual Crop, Forest, Industrial, and SeaLake. The model generates well-separated and compact clusters, demonstrating strong class discrimination in the learned embedding space.

with the ground truth, especially in capturing the spatial extent of built-up regions and water boundaries. Compared to the label map, the model output shows improved structural continuity, reduced noise in sparse classes like bare land, and fewer fragmented artifacts. These visual results confirm that the model effectively leverages multi-scale and graph-based spatial reasoning to enhance classification coherence, particularly in complex heterogeneous landscapes. This qualitative evidence complements the

quantitative results and supports the model's robustness in real-world applications.

The model improves robustness to lower-resolution images through several architectural and training-level strategies. Resolution-aware feature extraction is achieved via multi-scale encoding and adaptive fusion layers that align features across input scales, allowing the model to retain class-level semantics even when spatial granularity is reduced. Self-supervised learning is applied during pretraining using large-scale unlabeled remote sensing datasets, which helps the model capture structural patterns invariant to resolution. This reduces overfitting to high-resolution textures and improves transferability to coarse imagery. The model incorporates multimodal fusion, integrating auxiliary sources such as elevation maps, SAR data, or temporal sequences at the feature level. This compensates for information loss in low-resolution optical inputs by leveraging complementary spatial or spectral cues. These components collectively improve generalization to degraded inputs and support application in resource-constrained or archival satellite datasets.

To further evaluate the discriminative capability of the learned features, we apply t-SNE to project the high-dimensional embeddings into a two-dimensional space, as shown in Figure 11. Each point corresponds to a sample, and colors represent different land cover classes. The t-SNE visualization reveals that the feature representations produced by our model form well-separated and compact clusters, with clear boundaries between semantically distinct categories such as Annual Crop, Forest, and SeaLake. This indicates that the model is able to capture class-specific structure in the embedding space and reduce feature confusion across similar classes. The compactness and inter-class separability observed in the projection space qualitatively confirm the model's effectiveness in learning meaningful and discriminative features for remote sensing image classification.

To further validate the effectiveness of our proposed GISE model beyond quantitative benchmarks, we provide qualitative comparisons of classification results across multiple datasets. Visual samples of land cover predictions are shown in the extended versions of Figure 5 through 8, where we compare the outputs of GISE against leading baseline methods such as BERT and XLNet. On the Gaofen Image Dataset, the classification maps generated by GISE exhibit significantly better boundary adherence and spatial continuity in urban-fringe areas and small-scale structures, accurately distinguishing between impervious surfaces and vegetation zones. In contrast, baseline models often produce fragmented or noisy predictions in these high-resolution settings. For the LandCoverNet Dataset, which features more heterogeneous environments, GISE is able to maintain coherent spatial patterns across complex ecotones such as wetland-to-urban transitions. Baseline methods, while achieving acceptable pixel-level accuracy, tend to over-smooth class boundaries or misclassify mixed pixels. In the EuroSAT Dataset, GISE demonstrates improved class separation in agricultural and forested regions. Visual inspection shows fewer misclassifications between spectrally similar classes like "Pasture" and "Annual Crop", thanks to the integration of multi-scale and graph-based contextual modeling. On the UC Merced Dataset, which contains high-resolution aerial imagery with dense object arrangements, GISE preserves the internal structure of land parcels and urban blocks more faithfully. This is especially evident in recreational and commercial zones, where competing methods show substantial class confusion. These qualitative results confirm that GISE not only achieves high numeric scores but also delivers semantically meaningful, spatially coherent, and visually interpretable classification maps, which are crucial for real-world remote sensing applications.

5 Conclusion and future work

The research focuses on the intelligent classification of coastal land cover, aiming to enhance effectiveness of coastal management and environmental monitoring. Traditional classification methods, such as pixel-based and object-oriented approaches, often struggle with complex coastal landscapes, leading to inaccurate results. To address these limitations, the study integrates deep learning models, particularly CNNs, along with spatially dependent learning techniques. This integration allows for the development of a more robust and accurate classification model, leveraging multi-scale spatial analysis and graph-based models to capture spatial dependencies and contextual features across diverse coastal environments. The experimental results demonstrate that this method significantly improves classification accuracy, especially when applied to high-resolution remote sensing images. This advancement provides a more reliable tool for monitoring and managing coastal regions, presenting deep learning as a powerful approach to enhance remote sensing analysis for environmental and urban applications. The model is designed with practical deployment conditions in mind. It is evaluated across multiple datasets with distinct geographic, climatic, and spatial characteristics to ensure robustness in underrepresented areas. For example, Gaofen covers urban-dominated coastal regions, while LandCoverNet and EuroSAT include agricultural and natural landscapes from different continents and seasons. Seasonal variation is indirectly addressed through the inclusion of multi-temporal imagery, and the use of multi-scale augmentations and resolution-aware transfer modules further enhances resilience to seasonal texture and spectral shifts. Although the model operates under a closed-set classification framework, its graph-based architecture enables context-aware reasoning, which improves performance in regions with weak semantic boundaries or unfamiliar patterns. In practice, the model produces more coherent and less fragmented predictions than conventional CNNs in areas with sparse labels or ambiguous land cover types, indicating its suitability for real-world remote sensing tasks.

While the proposed method demonstrates strong performance across several benchmark datasets, it faces notable limitations when applied to coastal regions with limited or no labeled data. The current framework relies on supervised learning signals to effectively distinguish complex land cover types and to align multi-source spatial representations. In regions where annotated samples are sparse or absent, the model's ability to learn precise class

boundaries diminishes, particularly for heterogeneous or transitional areas such as wetlands, tidal flats, mangrove zones, or human-modified shorelines. These environments often exhibit unique spectral and textural signatures that differ significantly from those seen in the training data, making them prone to misclassification or prediction uncertainty. Another challenge lies in the diversity of sensor types and image resolutions across global coastal datasets. Variations in atmospheric conditions, seasonal cycles, and acquisition geometry can introduce substantial distributional shifts, which even domain adaptation techniques may fail to fully correct. When applied to data with lower spatial quality or unseen modalities, the model's feature extraction and classification performance can decline, especially in tasks that depend on fine-scale details or sharp land-water boundaries. Furthermore, inconsistencies in annotation standards and labeling granularity across datasets can propagate bias into the learned representations, undermining cross-region generalization. The model's complexity also introduces practical constraints. Deep architectures require significant computational resources and memory, which can limit their scalability to ultra-large-scale or real-time monitoring systems in operational coastal management. Training stability may be affected when applying the model to datasets with imbalanced class distributions or noise, leading to biased predictions favoring dominant land cover types. These limitations highlight the need for more adaptive and data-efficient approaches in future research, particularly methods capable of leveraging unlabeled or weakly labeled data, handling sensor heterogeneity, and maintaining robust performance under domain shifts. Such advances are essential for deploying land cover classification models in real-world coastal monitoring scenarios where data conditions are often imperfect and rapidly changing. In future work, we aim to reduce the dependency on large labeled datasets by exploring self-supervised and semi-supervised learning frameworks tailored to high-resolution remote sensing imagery. We will investigate the integration of transformer-based architectures to better capture long-range spatial dependencies and context information, particularly in heterogeneous coastal environments. Incorporating multi-temporal observations is another key direction to improve the model's robustness to seasonal and dynamic changes. Finally, we will focus on optimizing the model's computational efficiency and inference speed to enable real-time coastal land cover monitoring applications, making it more practical for large-scale environmental management and policy support.

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

XL: Writing – original draft, Writing – review and editing. JH: Writing – review and editing, Writing – original draft. YW: Writing – original draft, Writing – review and editing. AZ: Writing – review and editing, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

Conflict of interest

The 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

An, Y., Xia, X., Chen, X., Wu, F.-X., and Wang, J. (2022). Chinese clinical named entity recognition *via* multi-head self-attention based bilstm-crf. *Artif. Intell. Med.* 127, 102282. doi:10.1016/j.artmed.2022.102282

Bade, G., Kolesnikova, O., and Oropeza, J. (2024). *The role of named entity recognition (ner): survey*. International Journal of Computer and amp; Organization Trends. Available online at: https://www.sciencedirect.com/science/article/pii/S1877050924030011.

Chavan, T., and Patil, S. (2024). Named entity recognition (Ner) for news articles. *Int. J. Adv. Res. Eng. and Technol.* 2. doi:10.34218/ijaird.2.1.2024.10

Chen, B., Xu, G., Wang, X., Xie, P., Zhang, M., and Huang, F. (2022). "Aishell-ner: named entity recognition from Chinese speech," in *IEEE international conference on acoustics, speech, and signal processing*.

Chen, H., Sun, W., Gao, L., Yu, X., Yuan, X., and Xu, Y. (2024). Alignment and fusion using distinct sensor data for multimodal aerial scene classification. *IEEE Trans. Geoscience Remote Sens.* 62, 1–11. doi:10.1109/tgrs.2024.3406697

Chen, J., Lu, Y., Lin, H., Lou, J., Jia, W., Dai, D., et al. (2023). *Learning in-context learning for named entity recognition*. Annual Meeting of the Association for Computational Linguistics. Available online at: https://arxiv.org/abs/2305.11038.

Cui, L., Wu, Y., Liu, J., Yang, S., and Zhang, Y. (2021). Template-based named entity recognition using bart. *Findings*. Available online at: https://arxiv.org/abs/2106.01760.

Darji, H., Mitrović, J., and Granitzer, M. (2023). "German bert model for legal named entity recognition," in *International conference on agents and artificial intelligence*.

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., et al. (2021). Few-nerd: a few-shot named entity recognition dataset. *Annu. Meet. Assoc. Comput. Linguistics*. Available online at: https://arxiv.org/abs/2105.07464.

Durango, M. C., Torres-Silva, E. A., and Orozco-Duque, A. (2023). Named entity recognition in electronic health records: a methodological review. *Healthc. Inf. Res.* 29, 286–300. doi:10.4258/hir.2023.29.4.286

Guan, J., Li, L., Ao, Z., Zhao, K., Pan, Y., and Ma, W. (2023). Extraction of pig farms from gaofen satellite images based on deep learning. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 16, 9622–9631. doi:10.1109/jstars.2023.3323486

Günen, M. A. (2022). Performance comparison of deep learning and machine learning methods in determining wetland water areas using eurosat dataset. *Environ. Sci. Pollut. Res.* 29, 21092–21106. doi:10.1007/s11356-021-17177-z

Hernandez-Lareda, F., and Auccahuasi, W. (2024). "Implementation of a customized named entity recognition (Ner) model in document categorization," in *2024 3rd international conference on automation, computing and renewable systems (ICACRS)*.

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., et al. (2023). Improving large language models for clinical named entity recognition *via* prompt engineering. *J. Am. Med. Inf. Assoc*. Available online at: https://academic.oup.com/jamia/article-abstract/31/9/1812/7590607.

Jarrar, M., Abdul-Mageed, M., Khalilia, M., Talafha, B., Elmadany, A., Hamad, N., et al. (2023). *Wojoodner 2023: the first Arabic named entity recognition shared task*. ARABICNLP. Available online at: https://arxiv.org/abs/2310.16153.

Jarrar, M., Hamad, N., Khalilia, M., Talafha, B., Elmadany, A., and Abdul-Mageed, M. (2024). *Wojoodner 2024: the second Arabic named entity recognition shared task*. ARABICNLP. Available online at: https://arxiv.org/abs/2407.09936.

Khouya, N., Retbi, A., and Bennani, S. (2024). *Enriching ontology with named entity recognition (Ner) integration*. ACR. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-56950-0_13.

Kim, S., Gholami, A., Yao, Z., Mahoney, M. W., and Keutzer, K. (2021). "I-bert: integer-Only bert quantization," in *International conference on machine learning* (PMLR), 5506–5518. Available online at: https://proceedings.mlr.press/v139/kim21d.html.

Li, J., and Meng, K. (2021). Mfe-ner: multi-Feature fusion embedding for Chinese named entity recognition. *China Natl. Conf. Chin. Comput. Linguistics*.

Liu, S., He, T., and Dai, J. (2021). A survey of crf algorithm based knowledge extraction of elementary mathematics in Chinese. *Mob. Netw. Appl.* 26, 1891–1903. doi:10.1007/s11036-020-01725-x

Liu, S., Wang, L., Feng, H., and Sui, H. (2025). Empirical insights into resilience-based strategies for addressing haze pollution: enhancing green infrastructure and urban resilience. *Front. Environ. Sci.* 13, 1472235. doi:10.3389/fenvs.2025.1472235

Mi, B., and Yi, F. (2022). A review: development of named entity recognition (Ner) technology for aeronautical information intelligence. *Artif. Intell. Rev.* 56, 1515–1542. doi:10.1007/s10462-022-10197-2

Nigar, A., Li, Y., Jat Baloch, M. Y., Alrefaei, A. F., and Almutairi, M. H. (2024). Comparison of machine and deep learning algorithms using google Earth engine and python for land classifications. *Front. Environ. Sci.* 12, 1378443. doi:10.3389/fenvs.2024.1378443

Qu, X., Gu, Y., Xia, Q., Li, Z., Wang, Z., and Huai, B. (2023). A survey on Arabic named entity recognition: past, recent advances, and future trends. *IEEE Trans. Knowl. Data Eng.* 36, 943–959. doi:10.1109/tkde.2023.3303136

Ray, A. T., Pinon-Fischer, O. J., Mavris, D., White, R. T., and Cole, B. F. (2023). *aerobert-ner: named-Entity recognition for aerospace requirements engineering using bert*. AIAA SCITECH. *2023 Forum*. Available online at: https://arc.aiaa.org/doi/abs/10.2514/6.2023-2583.

Shen, W., Chen, J., Quan, X., and Xie, Z. (2021). Dialogxl: all-In-One xlnet for multi-party conversation emotion recognition. *Proc. AAAI Conf. Artif. Intell.* 35, 13789–13797. doi:10.1609/aaai.v35i15.17625

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. (2023a). *Diffusionner: boundary diffusion for named entity recognition*. Annual Meeting of the Association for Computational Linguistics. Available online at: https://arxiv.org/abs/2305.13298.

Shen, Y., Tan, Z., Wu, S., Zhang, W., Zhang, R., Xi, Y., et al. (2023b). *Promptner: prompt locating and typing for named entity recognition*. Annual Meeting of the Association for Computational Linguistics. Available online at: https://arxiv.org/abs/2305.17104.

Taher, E., Hoseini, S. A., and Shamsfard, M. (2020). *Beheshti-ner: persian named entity recognition using bert*. NSURL. Available online at: https://arxiv.org/abs/2003.08875.

Ushio, A., and Camacho-Collados, J. (2022). "T-ner: an all-round python library for transformer-based named entity recognition," in *Conference of the european chapter of the association for computational linguistics*.

Wang, J., Xu, H., Zhan, Z., and Wang, Q. (2023). Enhancing spatio-temporal analysis of satellite data using transformer networks: applications in drought monitoring. *Remote Sens.* 15, 5448. doi:10.3390/rs15235448

Wang, J., Zheng, Z., Ma, A., Lu, X., and Zhong, Y. (2021). Loveda: a remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv Prepr. arXiv: 2110*. Available online at: https://arxiv.org/abs/2110.08733.

Yossy, E., Suhartono, D., Trisetyarso, A., and Budiharto, W. (2023). "Question classification of university admission using named-entity recognition (Ner)," in *International conference on information technology, computer, and electrical engineering*.

Yu, J., Bohnet, B., and Poesio, M. (2020). *Named entity recognition as dependency parsing*. Annual Meeting of the Association for Computational Linguistics. Available online at: https://arxiv.org/abs/2005.07150.

Yu, J., Ji, B., Li, S., Ma, J., Liu, H., and Xu, H. (2022). "S-ner: a concise and efficient span-based model for named entity recognition," in *Italian national conference on sensors*.

Zaratiana, U., Tomeh, N., Holat, P., and Charnois, T. (2023). *Gliner: generalist model for named entity recognition using bidirectional transformer*. North American Chapter of the Association for Computational Linguistics.

Zhang, J., Zhou, Y., Li, B., Guo, R., Tang, X., Zhang, H., et al. (2023a). A hybrid model integrating spatiotemporal graph convolution and attention mechanisms for improved land cover classification. *Int. J. Digital Earth* 17, 2230978. doi:10.1080/17538947.2023.2230978

Zhang, S., Yu, H., and Zhu, G. (2022). An emotional classification method of Chinese short comment text based on electra. *Connect. Sci.* 34, 254–273. doi:10.1080/09540091.2021.1985968

Zhang, Y., Wang, J., Liu, M., and Zhao, R. (2025). Can digital economy improve urban ecological development? Evidence based on double machine learning analysis. *Front. Environ. Sci.* 13, 1542363. doi:10.3389/fenvs.2025.1542363

Zhang, Z., Cui, X., Zheng, Q., and Cao, J. (2021). Land use classification of remote sensing images based on convolution neural network. *Arabian J. Geosciences* 14, 267–6. doi:10.1007/s12517-021-06587-5

Zhang, Z., Hu, M., Zhao, S., Huang, M., Wang, H., Liu, L., et al. (2023b). *E-ner: evidential deep learning for trustworthy named entity recognition*. Annual Meeting of the Association for Computational Linguistics. Available online at: https://arxiv.org/abs/2305.17854.

Zhang, Z., Zhao, Y., Gao, H., and Hu, M. (2024). Linkner: linking local named entity recognition models to large language models using uncertainty. *Web Conf.*, 4047–4058. doi:10.1145/3589334.3645414

Zheng, J., Chen, H., and Ma, Q. (2024). Cross-domain named entity recognition *via* graph matching. *Findings*. Available online at: https://arxiv.org/abs/2408.00981.

Zhou, W., Zhang, S., Gu, Y., Chen, M., and Poon, H. (2023). "Universalner: targeted distillation from large language models for open named entity recognition," in *International conference on learning representations*.

Zhu, X., Konik, J., and Kaufman, H. (2025). The knowns and unknowns in our understanding of how plastics impact climate change: a systematic review. *Front. Environ. Sci.* 13, 1563488. doi:10.3389/fenvs.2025.1563488

Zhuang, H., Qin, Z., Jagerman, R., Hui, K., Ma, J., Lu, J., et al. (2023). "Rankt5: fine-Tuning t5 for text ranking with ranking losses," in *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 2308–2313.