remote sensing

MDPI

Article

Investigating Very-High-Resolution Land Cover Mapping in the Pearl River Delta with Remote Sensing Foundation Models and Multi-Source Data Bayesian Fusion

Junshen Luo †, Yikai Zhao †, Mingyang Xuan, Jizhou Zheng, Yan Zhou *,‡ and Xiaoping Liu

School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China; luojsh7@mail2.sysu.edu.cn (J.L.); zhaoyk25@mail2.sysu.edu.cn (Y.Z.); xuanmy@mail2.sysu.edu.cn (M.X.); zhengjzh7@mail2.sysu.edu.cn (J.Z.); liuxp3@mail.sysu.edu.cn (X.L.)

* Correspondence: zhouy825@mail.sysu.edu.cn

† These authors contributed equally to this work.

‡ current affiliation: School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China.

Highlights

What are the main findings?

- A comprehensive benchmark dataset including 262,436 unlabeled VHR images, 33,342 annotated samples, and 15,000 sample points is constructed for land cover mapping in the Pearl River Delta.
- A remote sensing foundation model pretraining method (SDMAE) with a value-aware masking strategy and Edge-Enhanced Loss for VHR land cover mapping is proposed, effectively preserving critical features in high-value objects and boundary regions.
- A scene-based semantic segmentation network (SBFNet) is developed, specifically designed to capture key scene-level features for VHR land cover mapping of complex heterogeneous landscapes.

What is the implication of the main finding?

- A decision-level Bayesian fusion framework combining statistic and spatial consistency tests with Bayesian probabilistic modeling effectively integrates multi-source remote sensing data, achieving robust land cover mapping.

Abstract

Very-high-resolution (VHR) land cover mapping in highly heterogeneous regions faces critical challenges including strong annotation dependence, significant image heterogeneity, and insufficient spectral information. To address these challenges, this study proposes a novel framework integrating remote sensing foundation models with multi-source data Bayesian fusion for VHR land cover mapping in the Pearl River Delta (PRD), which is one of the most complex and heterogeneous landscapes in China. To implement this framework, we first construct three datasets including PRD262K containing 262,436 unlabeled VHR images for pretraining, PRDLC-PRO with 33,342 annotated samples for semantic segmentation, and a 15,000-point sample set for medium-resolution (MR) classification. A Segmentation-Driven Masked AutoEncoder (SDMAE) is developed to learn robust feature representations from large-scale unlabeled VHR imagery, which is subsequently integrated with a Scene-Based Feature Network (SBFNet) to capture multi-scale semantic features for accurate land cover segmentation. Finally, a decision-level Bayesian fusion method is proposed to effectively integrate the fine spatial details of VHR imagery with the spectral stability of MR data. Experiments demonstrate that the proposed framework outperforms

Check for updates

Academic Editor: Xiaodong Li

Received: 22 January 2026
Revised: 2 March 2026
Accepted: 10 March 2026
Published: 15 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

existing methods across multiple datasets, achieving an overall accuracy of 87.98% and mIoU of 66.61% on PRDLC-PRO. The subsequent decision-level Bayesian fusion further enhances spatial consistency and robustness, providing an effective solution for large-scale VHR land cover mapping in highly heterogeneous regions with limited annotations.

Keywords: remote sensing foundation model; VHR imagery; land cover mapping; multi-source remote sensing data; multi-source data fusion

1. Introduction

Land cover describes the spatial distribution of biological and physical cover types on the Earth's surface, serving as fundamental information for human–environment interactions and surface process research [1]. Rapid urbanization continuously drives land cover pattern changes. Therefore, accurate and rapid mapping is essential for remote sensing applications in climate change assessment, ecological conservation, and territorial spatial planning [2–6].

Since the launch of the Landsat satellite in the 1970s, the continuous acquisition of medium-resolution (MR) imagery has contributed to large-scale land cover mapping [7]. Researchers have combined multi-temporal and multi-resolution imagery with machine learning methods, including decision trees, support vector machines, random forests, and artificial neural networks, to conduct land cover classification, enabling effective regional and even global mapping [8–13]. However, these methods primarily rely on single-pixel spectral or shallow artificial features, with limited capability to characterize complex spatial structures and superior semantic information, and are still facing limitations in highly heterogeneous surface-type regions.

As land cover mapping requirements for refined expression and high spatial consistency continue to increase, in recent years, the rapid development of very-high-resolution (VHR) remote sensing imagery has enabled the extraction of fine surface information. Meanwhile, breakthrough advances in deep learning, particularly semantic segmentation models capable of automatically learning multi-scale spatial features and superior semantic representations, have opened new avenues for fine characterization of complex ground objects [14–16]. These models typically adopt encoder–decoder architectures that recover spatial details while preserving semantic information through multi-scale feature fusion [17–21]. While early encoders were primarily based on convolutional neural networks [22–24], recent research has shifted towards Transformer architectures like ViT [25] and Swin Transformer [26] or hybrid structures [27] to enhance long-range dependency modeling.

Early MR land cover products (such as GLC2000 [28] and CGLS-LC100 [29]) and subsequent 10 m level products (such as FROM-GLC10 [30] and Dynamic World [31]), developed using machine learning, exemplify continuous progress in the automation and accuracy of traditional methods. The application of deep learning methods, particularly semantic segmentation, has elevated mapping resolution to the meter level, as evidenced by representative products such as SinoLC-1 [32] and SCLC [33]. To further enhance feature representation, targeted optimizations such as dual-attention mechanisms [34], multi-scale context aggregation [35], and boundary-aware mechanisms [36] have been introduced. To fundamentally reduce reliance on labels in the target domain, research has shifted towards cross-domain knowledge transfer. Recent studies have explored pseudo-labeling [37] and knowledge distillation [38] to mitigate domain shift, aiming to transfer discriminative features learned from source domains to new geographic regions.

However, the effectiveness of such a transfer is often compromised in highly heterogeneous landscapes, and the fundamental dependence on massive, high-quality labels remains a critical bottleneck [39].

To alleviate this annotation dependence, self-supervised pretrained remote sensing foundation models have emerged as a promising solution. By learning universal feature representations from massive unlabeled remote sensing imagery, self-supervised training can capture multi-scale spatial features and semantic information in images without relying on the distribution of labeled data, thereby improving model transfer capability and mapping performance across different regions and tasks [40,41]. Initially, paradigms of remote sensing foundation models such as masked image reconstruction [42–44] and contrastive learning [45,46] established the basis for learning robust visual features. These have recently evolved into vision–language joint pretraining [47–50], while model scales have expanded from millions to billions of parameters to further enhance universal generalization [51–56]. These advancements aim to provide a more robust feature foundation for downstream mapping tasks. However, when applied to fine-grained mapping, reconstruction-based models often suffer from the loss of high-value object details and blurred boundaries. This limitation stems from the difficulty of recovering sharp textures in highly heterogeneous landscapes during the self-supervised process.

Moreover, large-scale deployment of such models on VHR imagery faces several intrinsic limitations. VHR imagery can capture rich texture and geometric details, which enables the precise mapping of object boundaries and the detailed analysis of internal structures. While VHR imagery captures rich texture and geometric details that enable precise boundary delineation, large-scale applications face multiple challenges (Figure 1). On one hand, limited imaging conditions, such as weather effects (clouds), illumination variations, and differences in observation angle (shadows), can easily cause significant image heterogeneity. On the other hand, VHR imagery typically contains only RGB bands, lacking key spectral information such as near-infrared, making it difficult to effectively distinguish ground object types with similar spectral characteristics [37,57]. Additionally, due to image mosaicking caused by low temporal resolution, inconsistencies arise between source domain and target domain and thus lower the accuracy of VHR land cover mapping [58]. Consequently, even powerful mapping models struggle to bridge the gap caused by these intrinsic spectral and imaging deficiencies.

In contrast, MR imagery possesses rich spectral information, mature interpretation methods and stable spatiotemporal characteristics, which are valuable for large-scale, consistent mapping. Previous research indicates that multi-source data fusion (e.g., optical and SAR) can effectively improve classification accuracy [59,60], suggesting that integrating MR and VHR data is a highly promising strategy for high-precision VHR mapping. To address the spatial mismatch between these sources, scholars have explored various cross-scale modeling and weak-supervised strategies [61,62]. However, achieving large-scale, high-precision mapping by integrating VHR and MR imagery remains a significant challenge. While recent studies have found that Bayesian modeling can effectively fuse prior knowledge to improve stability [63,64], its potential for collaborative VHR-MR cross-scale fusion in land cover mapping has yet to be fully explored. Therefore, how to integrate the fine spatial information of VHR imagery with the robustness of MR imagery to achieve VHR land cover mapping via a unified probabilistic framework has become a critical problem.

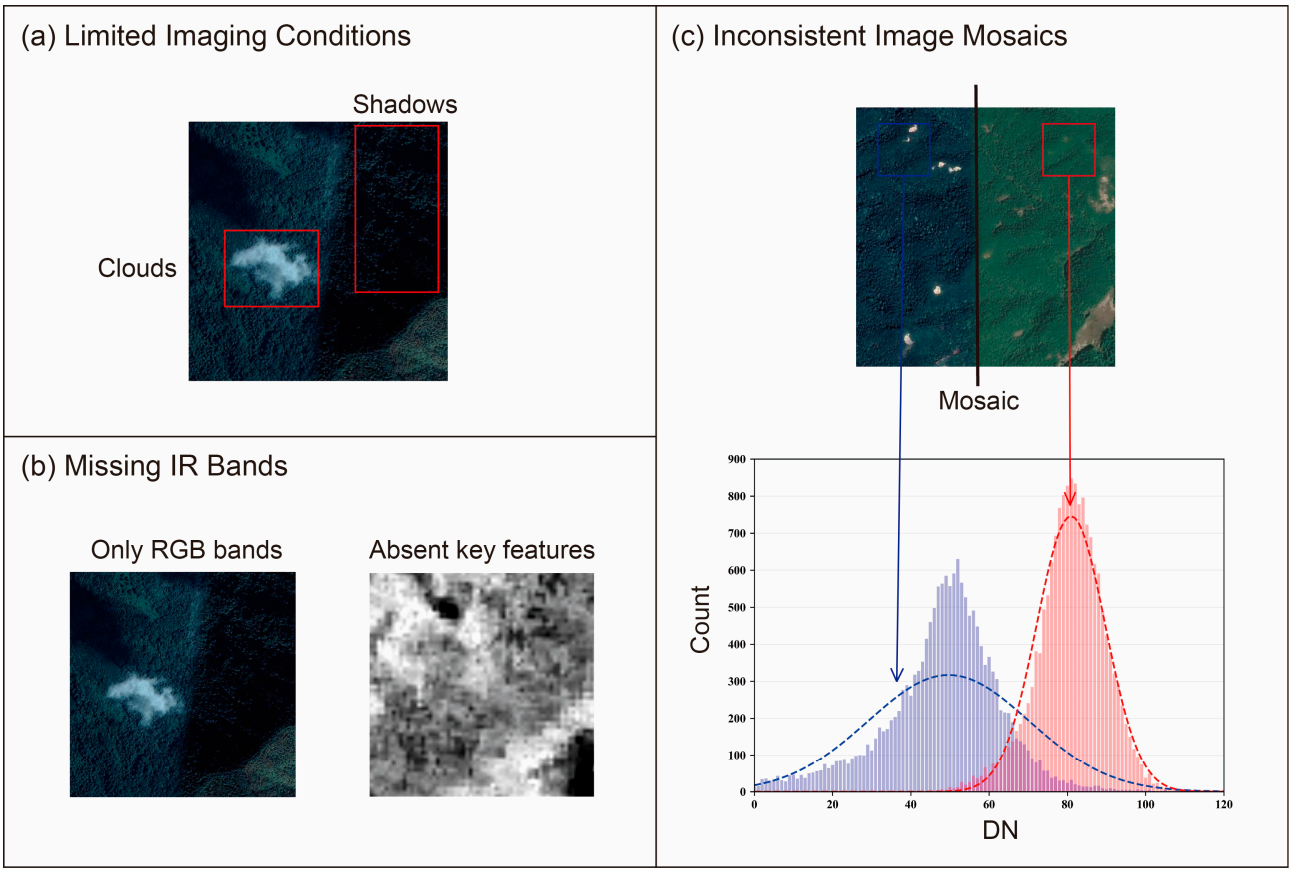


Figure 1. Main challenges in large-scale VHR land cover mapping. (**a**) Limited imaging conditions including clouds and shadows that cause image heterogeneity; (**b**) missing infrared bands with only RGB bands available, resulting in absent key spectral features; (**c**) inconsistent image mosaics due to low temporal resolution, leading to domain shift between source and target domains. In panel (**c**), the blue and red shaded histograms represent the DN value distributions of the source and target domain images, respectively; dashed curves indicate the fitted distributions.

Therefore, this study aims to construct a novel framework integrating a remote sensing foundation model with multi-source remote sensing data to address the above challenges. The technical roadmap of this research is illustrated in Figure 2. Taking the Pearl River Delta (PRD), a region characterized by extreme spatial heterogeneity and complex landscapes, as an example, we focus on VHR land cover mapping, with the main contributions including:

(1) Construction of a VHR land cover mapping benchmark dataset for the complex environment of the PRD, including a training point dataset for MR classification, a large-scale unlabeled image dataset (PRD262K) for pretraining, and a high-quality pixel-level annotated dataset (PRDLC-PRO) for training and validation.

(2) Proposal of a remote sensing foundation model pretraining method using Segmentation-Driven MAE (SDMAE) and a Scene-Based Feature Network (SBFNet) for VHR land cover mapping. An SDMAE addresses the loss of value details and edge features by introducing a value-aware masking strategy (VMask Strategy) and Edge-Enhanced Loss during pretraining. The SBFNet implements multi-scale feature extraction based on the "pixel–object–scene" semantic framework, enhancing the accuracy of VHR land cover mapping.

(3) Design of a novel decision-level Bayesian fusion framework that jointly utilizes the fine spatial information of VHR imagery and the spectral stability of MR imagery. By combining statistic and spatial consistency tests and Bayesian probabilistic modeling, the output probabilities of the VHR semantic segmentation model and MR random forest

model are collaboratively optimized. This approach prioritizes the rectification of systemic mapping errors, effectively improving the consistency and robustness of VHR mapping results in complex regions.

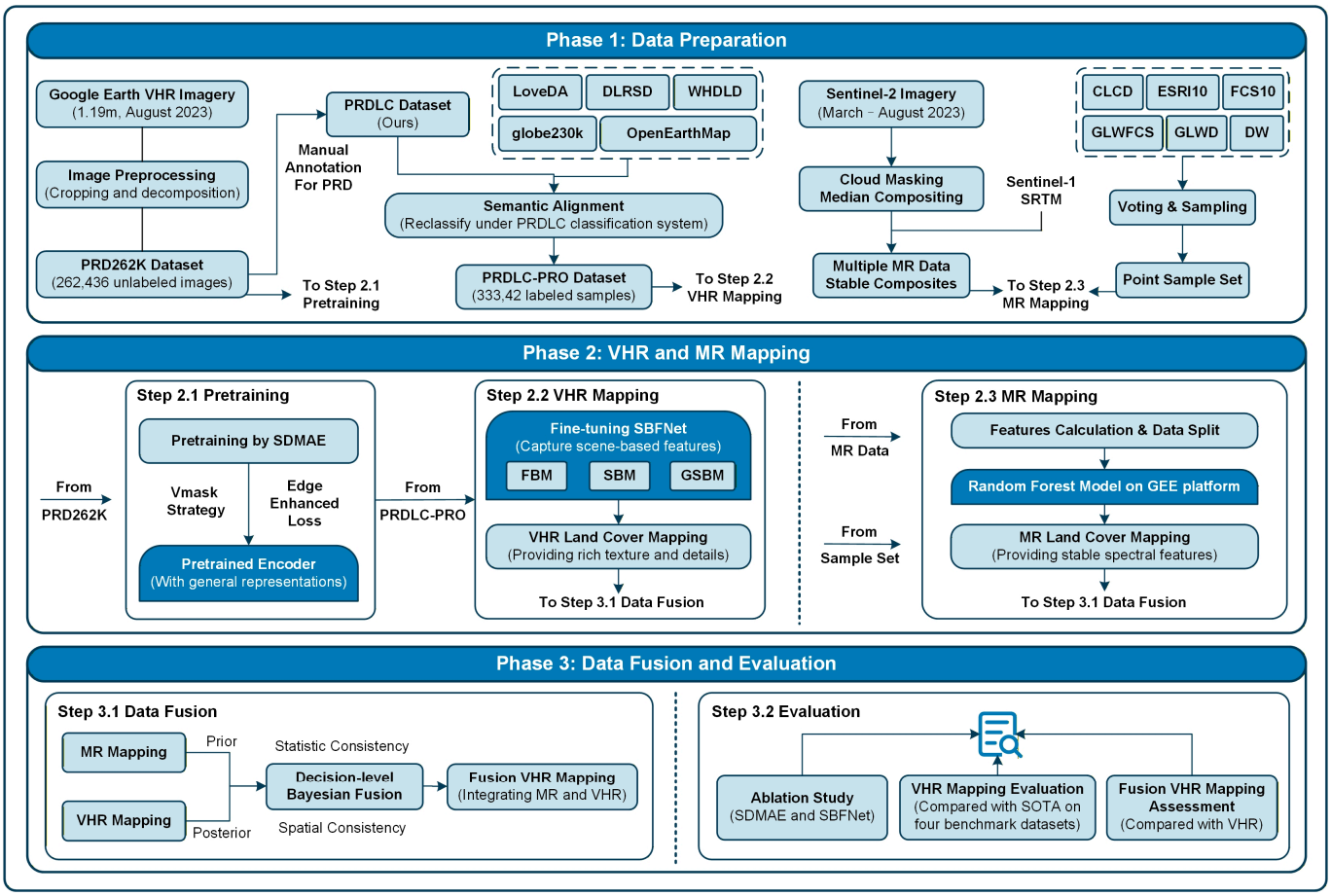


Figure 2. Technical roadmap of the study, illustrating the integrated pipeline from multi-source data preparation (Phase 1) to hierarchical mapping (Phase 2) using the SDMAE and SBFNet for VHR data alongside random forest for MR data and finally to decision-level Bayesian fusion and systematic performance evaluation (Phase 3).

2. Study Area and Data

2.1. Study Area

This study focuses on the Pearl River Delta (PRD) region in Guangdong Province, China, which is geographically located between 21.57°N–24.40°N and 111.35°E–115.41°E, as illustrated in Figure 3. Located south of the Tropic of Cancer, the PRD belongs to the tropical monsoon climate zone, characterized by abundant heat, rich precipitation, and synchronous rainfall and temperature patterns. Influenced by the high-temperature and high-humidity environment, the PRD experiences persistent heavy cloud cover during the summer and autumn seasons. Combined with complex and variable illumination conditions, optical remote sensing imagery, particularly VHR imagery, exhibits significant radiometric inconsistency and image heterogeneity across space. Such issues are typically difficult to completely eliminate. Furthermore, multispectral remote sensing imagery in this region is severely affected by clouds and rain, making it challenging to acquire stable and continuous time-series observation data, thereby limiting the effective extraction of temporal characteristics of land cover types.

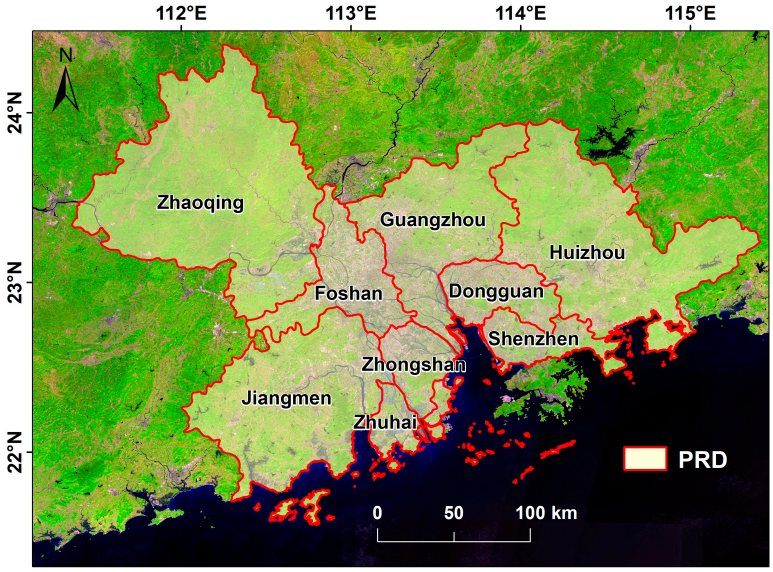


Figure 3. Geographic location of the study area in the PRD, Guangdong Province, China.

The complex climatic conditions and highly heterogeneous land cover types in the PRD make it difficult for single-source remote sensing data to comprehensively and stably capture land surface features. Therefore, it is essential to leverage the advantages of multi-source remote sensing data and develop novel land cover mapping methodologies through multi-source data fusion strategies, thereby achieving more accurate and robust VHR land cover mapping.

2.2. VHR Data

We used Google Earth VHR optical imagery (1.19 m resolution) acquired in August 2023 as the primary data source for feature extraction and fine-grained land cover mapping. The study area, covering approximately 87,800 km^2, is divided into 170 map sheets according to the 1:50,000 cartographic standard. The original sheet images were fully decomposed into 512 $\times$ 512 sub-images to accommodate the requirements of both pretraining and downstream semantic segmentation tasks. Ultimately, 262,436 complete VHR images were generated, forming the remote sensing foundation models' pretraining dataset PRD262K.

Furthermore, we established a land cover classification system based on the climatic and topographic characteristics of the PRD, comprising eight classes: cropland, forest, grass, shrubland, wetland, water, impervious and bare (Appendix A). Based on this classification system, we constructed a VHR land cover mapping dataset for the PRD. We manually annotated 7370 high-quality semantic segmentation samples. This set of annotations, which we refer to as the PRDLC dataset, is visualized with example images in Figure 4. Moreover, we integrated several existing land cover mapping datasets, including LoveDA [65], DLRSD [66], WHDLD [67], Globe230k [68], and OpenEarthMap [69], to complement our manually annotated PRDLC dataset. While PRDLC provided 7370 region-specific samples for the PRD, these external sources contributed an additional 25,972 samples from similar climatic conditions. By reclassifying them under a unified taxonomy, we constructed the PRDLC-PRO dataset (totaling 33,342 samples), as detailed in Table A2.



Figure 4. Examples of VHR imagery and corresponding land cover annotations in the PRDLC dataset. For each batch, (left to right) there is the original image and annotation label, respectively.

2.3. MR Data

This study employed Sentinel-1 C-band SAR imagery and Sentinel-2 multispectral imagery as MR data sources. Sentinel-1 enables all-weather observations, complementing optical imagery with land surface geometry and texture. Sentinel-2 spans the visible-to-shortwave infrared spectrum, offering rich spectral characteristics. All high-quality Sentinel-1 VV/VH polarization data acquired in 2023 were retrieved via the Google Earth Engine platform. Sentinel-2 surface reflectance products from March to August 2023 were selected to capture the peak phenological characteristics of vegetation and ensure temporal consistency with the VHR imagery for accurate multi-source data fusion. During this period, images with high cloud cover were excluded, and median compositing was applied to generate temporally stable multispectral composites. Furthermore, a digital elevation model (DEM) and slope data from NASADEM were included to improve land cover mapping performance.

The workflow to construct a high-quality land cover point sample set is illustrated in Figure 5. We integrated multiple land cover datasets (Table A3), including general land cover products (the China Land Cover Dataset (CLCD) [70], Dynamic World (DW) [31], Esri 10 m Land Cover (ESRI10) [71], and GLC_FCS10 (FCS10) [72]) and wetland-specific products (Global Lakes and Wetlands Database v2 (GLWD) [73] and GWL_FCS30 (GWLFCS) [74]). All datasets were clipped to the PRD region and reclassified according to the PRDLC classification system, then resampled to 10 m resolution to ensure spatial consistency.

A weighted voting strategy based on multi-source agreement was applied to select high-confidence samples. For easily distinguishable classes with abundant samples (forest, water and impervious), a strict agreement threshold of 3/4 was set. For classes prone to confusion (cropland, grass, shrubland, and bare), a loose threshold of 2/4 was adopted. For wetland, a weighted fusion approach assigned higher weights (1.5) to wetland-specific products and lower weights (0.6) to general products, identifying samples exceeding the weighted threshold as valid wetland instances. These thresholds and weights were empirically optimized through visual inspection to ensure label purity, while remaining adjustable as hyperparameters to adapt to the consistency of multi-source products in different geographic contexts. Finally, manual verification was performed using Google Earth VHR imagery, and stratified sampling was applied to balance class distribution. This process yielded a high-quality land cover sample set containing 15,000 points, with the composition shown in Table 1.

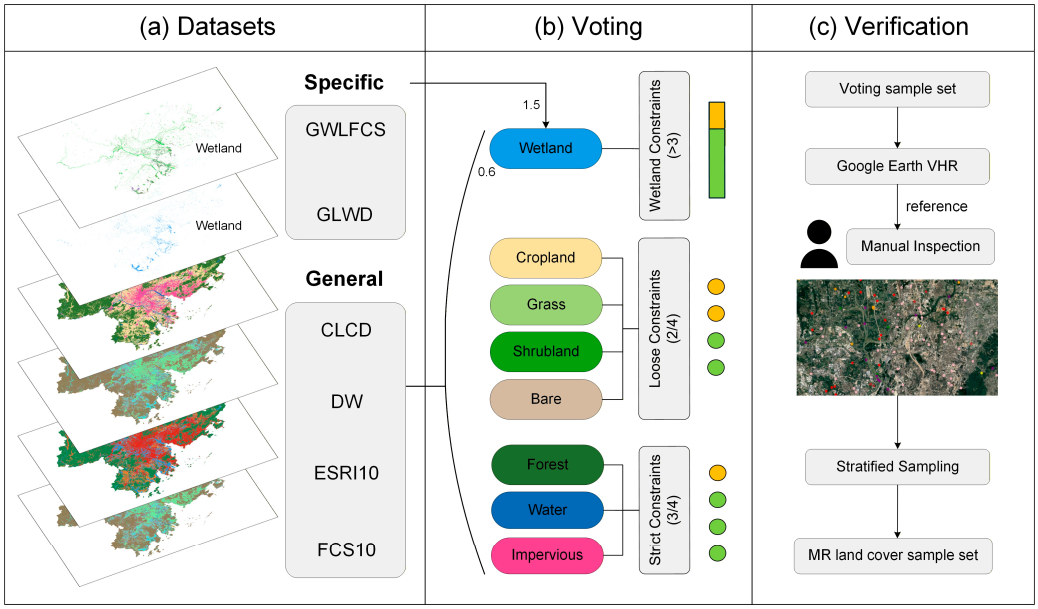


Figure 5. Workflow for constructing the high-quality MR land cover sample set. (**a**) Multiple MR land cover datasets including general products and wetland-specific datasets were integrated; (**b**) a weighted voting strategy was applied with different constraints for various land cover classes; (**c**) manual verification was conducted using Google Earth VHR imagery as a reference, followed by stratified sampling to generate the final sample set.

Table 1. Composition of the land cover sample set used for MR mapping.

Class	Samples per Class
Cropland	1898
Forest	3329
Grass	1710
Shrubland	2369
Wetland	1910
Water	943
Impervious	1426
Bare	1415

3. Methods

3.1. Segmentation-Driven Mask AutoEncoder (SDMAE)

We adopt the masked autoencoder (MAE) framework as our foundation due to its superior performance in learning fine-grained spatial features compared to contrastive learning methods [42]. However, its standard random masking strategy is suboptimal for VHR imagery, as it often leads to the loss of critical high-value details and edge features. To address this limitation, we propose a remote sensing foundation model pretraining method called a Segmentation-Driven Masked AutoEncoder (SDMAE), which incorporates a value-aware masking strategy (VMask Strategy) and Edge-Enhanced Loss (Figure 6). Similarly to previous studies, the SDMAE is a monomodal self-supervised pretraining method that reconstructs complete images from partial observations using an autoencoder. The training process of the SDMAE consists of four main steps: image masking, encoding, decoding, and image reconstruction.

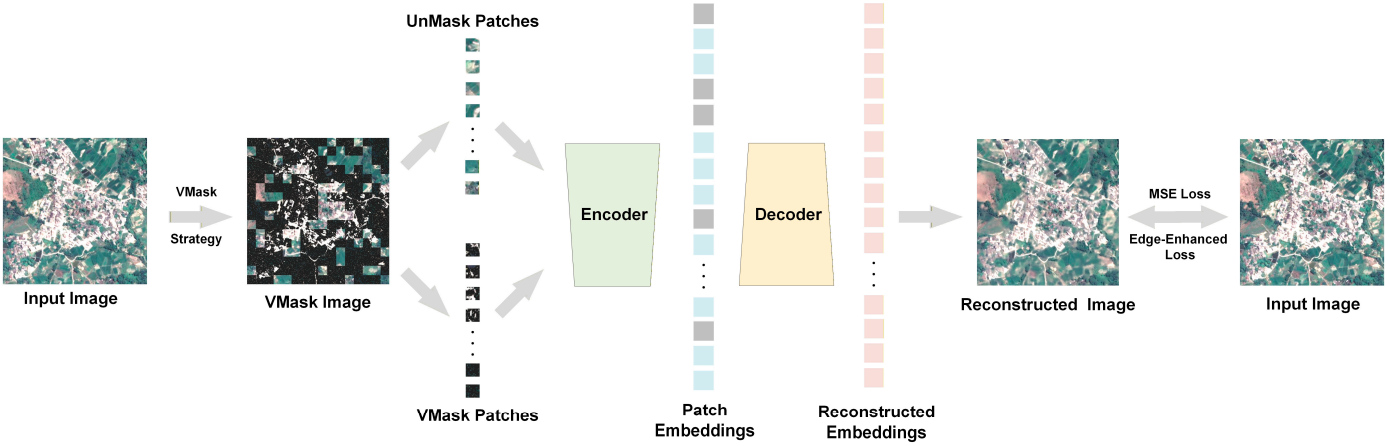


Figure 6. Overall architecture of the proposed SDMAE pretraining framework. The input image is processed through the VMask Strategy to generate a value-masked image, from which unmasked patches and VMask patches are extracted. These patches are encoded by an encoder into patch embeddings, then decoded to produce reconstructed embeddings. The model is optimized using a combined loss function consisting of MSE loss and Edge-Enhanced Loss to reconstruct the original image.

(1) Image Masking: The SDMAE divides the input remote sensing image into regular and non-overlapping patches. Most existing methods randomly sample and retain a certain proportion of patches following a uniform distribution, then completely mask the remaining patches [42], or randomly retain some pixels within masked patches [43]. The former, while effective for natural images, tends to overlook small objects in remote sensing imagery. Although the latter alleviates this issue to some extent, its mitigation remains insufficient. This is because the overlooked features in VHR land cover mapping, typically high-value pixels such as impervious and bare, often correspond to small yet critical objects that are easily missed by random masking, leading to feature loss. To address this problem, we propose the VMask Strategy, with the main workflow described as follows. Given the patch masking ratio $r \in (0,1)$ and value masking coefficient $r_v \in (0,1)$, for an input image $X \in \mathbb{R}^{H\times W\times C}$, we compute its value map $V \in \mathbb{R}^{H\times W}$ and calculate the percentile value threshold based on

$$v_h = Percentile(V, 75r_v + 25) \tag{1}$$

$$v_l = Percentile(V, 25 - 25r_v) \tag{2}$$

where v_h and v_l represent the high-value and low-value percentile thresholds, respectively. These thresholds define the medium-value region $(v_l \le V \le v_h)$, and pixels within this region are randomly masked with probability $p = 0.5r_v + 0.5$ to generate the value mask M_v, which is then applied to the original image X to obtain the value-masked image (VMask image) X_V:

$$M_v = I(v_l \le V \le v_h) \cap B(0.5r_v + 0.5) \tag{3}$$

$$X_V = (1 - M_v) \odot X \tag{4}$$

where $I(\cdot)$ is the indicator function, $B(\cdot)$ denotes the Bernoulli distribution, and $\odot$ represents element-wise multiplication. Both the input image X and the VMask image X_V undergo a Patchify operation, which transforms them into patch features via convolution and positional encoding:

$$P = Patchify(X) \in \mathbb{R}^{N\times D} \tag{5}$$

$$P_V = Patchify(X_V) \in \mathbb{R}^{N\times D} \tag{6}$$

where P and P_V denote the patch features of the input image X and the VMask image X_V, respectively. Finally, P and P_V are synchronously shuffled using an identical random index (idx) and then concatenated to obtain the complete feature sequence P_{final}:

$$idx \sim RandomPermutation(1, 2, \ldots, N) \tag{7}$$

$$P_{final} = concat\left(P\left[idx_{1:(1-r)N}\right], P_V\left[idx_{(1-r)N+1:N}\right]\right) \tag{8}$$

The first $(1-r)\%$ proportion of features corresponds to the unmasked patches from the original image X (i.e., $P\left[idx_{1:(1-r)N}\right]$), which is consistent with the MAE. The remaining $r\%$ proportion consists of features from the value-guided masked patches (i.e., $P_V\left[idx_{(1-r)N+1:N}\right]$), which preserves information from extreme-value regions. This strategy protects high-value object features while incorporating retained low-value features and random masking, thereby preventing the model from learning fixed-value patterns.

(2) Encoder: The encoder employed in this study is a Vision Transformer (ViT). We use ViT-Base to process the features of both the unmasked patches and the value-guided masked patches obtained previously, thereby learning a universal feature representation. Consistent with the ViT, the encoder first projects the input patches linearly and adds positional embeddings to obtain the feature embeddings for all patches. Subsequently, these feature embeddings are processed by a series of Transformer blocks.

(3) Decoder: The decoder consists of 8 Transformer layers, each with 512 channels, followed by a linear projection layer. The decoder takes the deep feature representations output by the encoder as input. During pretraining, the decoder first concatenates the encoded unmasked patch features with the value-guided masked patch features. This concatenated sequence is then reordered according to the original patch positions, and positional embeddings are added. After processing by multiple Transformer layers, the decoder produces final feature representations for reconstructing the input image.

(4) Image Reconstruction: The SDMAE reconstructs the original image by predicting the pixel values of the masked regions. Specifically, the final layer of the decoder is a linear projection head that maps the feature vector of each patch to the predicted values for all pixels within that patch. The output channels of the projection head correspond to the total number of pixel values in a single patch. By applying a reshape operation to the output of the decoder, the complete reconstructed image $\hat{X}$ is obtained. During pretraining, we design a loss function that consists of mean squared error loss and Edge-Enhanced Loss to simultaneously optimize pixel-level reconstruction accuracy and the preservation of edge texture. The loss function consists of two components:

Mean squared error (MSE) loss (L_{MSE}): This is the fundamental loss for pretraining, aiming to minimize the pixel-wise difference between the reconstruction image $\hat{X}$ and the original input image X within the masked regions (indicated by the total mask matrix M). It is calculated as follows:

$$L_{MSE} = \frac{1}{\sum M}\sum\left(\left(\hat{X} - X\right)^2 \odot M\right) \tag{9}$$

where $\odot$ denotes element-wise multiplication. This loss forces the model to focus on inferring occluded content from the visible context.

Edge-Enhanced Loss (L_{Edge}): To mitigate the common issue of lost edge features in land cover mapping, we introduce an additional edge constraint. A binary edge map E is extracted from the original image X using a Laplacian operator combined with Otsu's

thresholding method. This loss reinforces reconstruction accuracy specifically for edge areas within the masked regions and is defined as

$$L_{Edge} = \frac{1}{\sum(M \odot E)} \sum \left((\hat{X} - X)^2 \odot M \odot E \right) \tag{10}$$

Edge-Enhanced Loss significantly increases the weight of reconstruction errors at image edges, guiding the model to better recover the contours and boundary details of ground objects during reconstruction. The final pretraining loss function is the weighted sum of these two terms:

$$L_{Total} = L_{MSE} + \lambda \cdot L_{Edge} \tag{11}$$

where λ is a weighting factor (set to 1 in this study). The VMask Strategy prioritizes masking non-extreme-value regions (which often correspond to textured, homogeneous and non-edge areas), while the Edge-Enhanced Loss heavily penalizes reconstruction errors in edge regions. The synergistic effect of these two components enables the model to learn more balanced and refined feature representations from homogeneous regions to sharp boundaries in remote sensing imagery, thereby establishing a superior foundation for downstream land cover mapping tasks.

3.2. Scene-Based Feature Network (SBFNet)

The SBFNet is a semantic segmentation model designed for VHR land cover mapping in the PRD. Its architecture comprises a pretrained encoder, a scene-based decoder, and a scene-based feature fusion module (Figure 7). The SBFNet is grounded in the "pixel–object–scene" semantic framework and leverages it to decode and fuse features across three tiers. Pixel-level features capture fine details, and object-level features delineate object boundaries and overall forms, while scene-level features encapsulate the global image context, such as the high-value objects of urban areas. Existing decoders are often constrained by limited receptive fields and struggle to effectively capture scene-level semantics. To address this, the SBFNet incorporates a Scene-Based Module (SBM), a Feature-Based Module (FBM), and a Global Scene-Based Module (GSBM) within its decoder. These modules explicitly extract and fuse scene-level global features with local detailed features, thereby enhancing the accuracy of land cover mapping.

(1) Encoder: The encoder transfers the pretrained ViT-Base encoder from the SDMAE. After processing the original input image through this encoder, a feature embedding F_0 that conforms to the dataset distribution is obtained.

(2) Scene-Based Decoder: The decoder operates along two pathways. The first step directly uses the GSBM to extract the global scene. That is, the feature embedding F_0 is fed into the GSBM, sequentially passing through a 3×3 convolutional layer and a ReLU layer, and is then directly upsampled to obtain the global feature embedding F_G:

$$F_G = GSBM(F_0) \tag{12}$$

The second step involves the original feature embedding F_0 sequentially passing through several layers of scene-based feature extraction modules to obtain detailed decoding feature F_D. Specifically, the original feature embedding F_0 enters the SBM to produce a scene feature map. At the same time, F_0 enters the FBM to produce a detail feature map. Finally, these two maps are summed to obtain the overall feature F_1 at the corresponding scale. Repeating this process yields the feature map F_5 which matches the original image size. This process can be expressed by the following formula:

$$F_{i+1} = SBM(F_i) + FBM(F_i), i = 0, 1, 2, 3, 4 \tag{13}$$

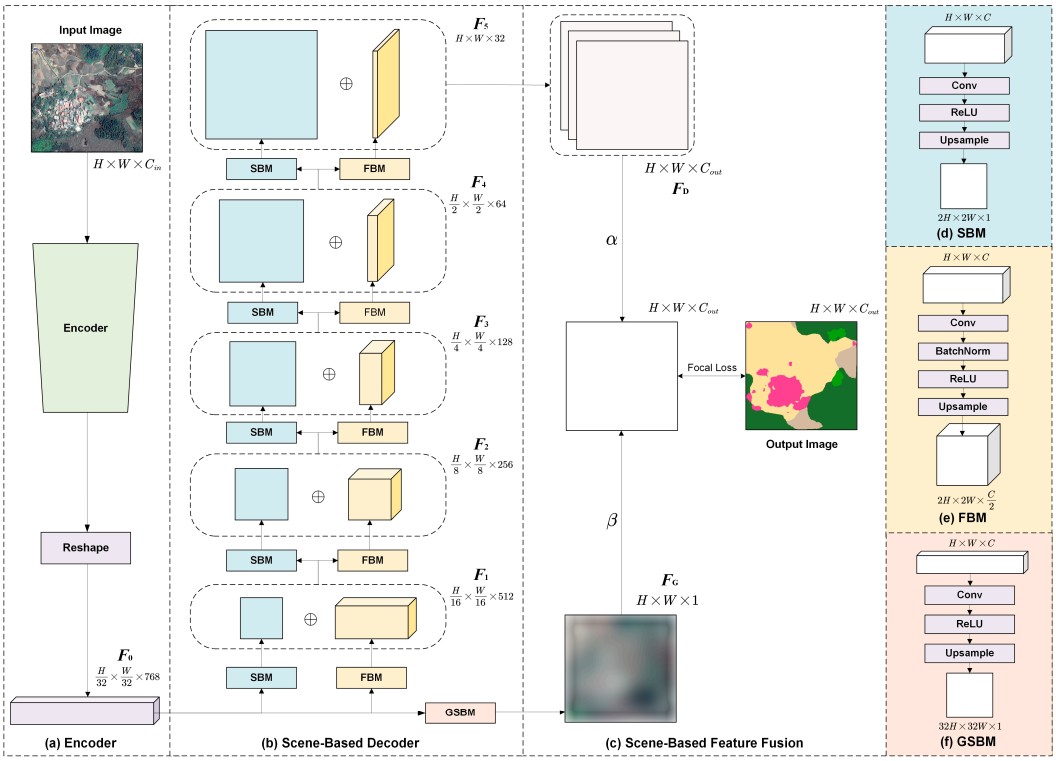


Figure 7. Overall architecture of SBFNet. (**a**) Encoder with pretrained ViT-Base; (**b**) scene-based decoder with SBMs, FBMs, and GSBMs for hierarchical feature extraction; (**c**) scene-based feature fusion combining detail and global features; (**d**) detailed structure of SBM; (**e**) detailed structure of FBM; (**f**) detailed structure of GSBM.

Finally, the detail decoding feature F_D is obtained by passing the feature map F_5 (same size as the input image) through a 3×3 convolutional layer, producing an output with the same channels as the target land cover classes.

(3) Scene-Based Feature Fusion: The final network output feature map F is obtained by a weighted summation of the detail decoding feature F_D and the global feature embedding F_G:

$$F = \alpha \cdot F_D + \beta \cdot F_G \tag{14}$$

where α and β are learnable vectors representing the weights for the detail decoding feature F_D and the global feature embedding F_G, respectively. Finally, the network output F and the ground truth labels are used to compute Focal Loss as the loss function, which addresses class imbalance and difficult sample issues during training:

$$Focal\ Loss = \theta \cdot (1 - p_{ij})^{\gamma} \cdot \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(p_{ij}) \tag{15}$$

where N is the number of samples, C is the number of classes, y_{ij} is the indicator function for the j-th class in the ground truth label of sample i, p_{ij} is the predicted probability for the j-th class in sample i, and θ and γ are tunable factors for the Focal Loss.

3.3. Random Forest Model for MR Mapping

This study employs the random forest model to classify MR remote sensing imagery, aiming to generate spectrally stable prior probability knowledge. As an ensemble learning method, random forest enhances model robustness and generalization capability by constructing multiple decision trees and aggregating their predictions, which also allows it to handle high-dimensional features effectively.

The implementation is carried out entirely on the Google Earth Engine platform. First, multi-source MR data is processed for feature extraction and fusion. This includes computing quarterly mean and variance from Sentinel-1 SAR data to suppress speckle noise and characterize temporal stability, performing cloud removal and median compositing on Sentinel-2 multispectral imagery along with deriving various spectral indices to improve separability of land cover types and generating slope features from NASADEM data to incorporate topographic information (see Table A4 for a detailed list of all features). Subsequently, the MR point sample set constructed in Section 2.3 is used, randomly split into training and testing subsets at a ratio of 7:3.

To mitigate the impact of random sample partitioning and systematically evaluate model performance, 10 independent repeated experiments are designed. In each experiment, a different random seed is used to reshuffle the samples, and the random forest API on GEE is invoked for training and validation, with the number of decision trees set to 100 and the output mode configured as MULTIPROBABILITY. Based on the evaluation results of these 10 experiments, the optimal model parameters are selected. Finally, using the optimized model parameters, classification inference is performed on the MR data to produce 10 m resolution land cover probability maps. These maps serve as key spectral prior knowledge and are utilized in the subsequent cross-resolution decision-level Bayesian fusion.

3.4. Decision-Level Bayesian Fusion

To leverage the advantages of multi-source data in VHR land cover mapping, we propose a decision-level Bayesian fusion framework (Figure 8). This framework processes distribution inference between the VHR and MR mapping results through statistic and spatial consistency tests and subsequently applies different strategies.

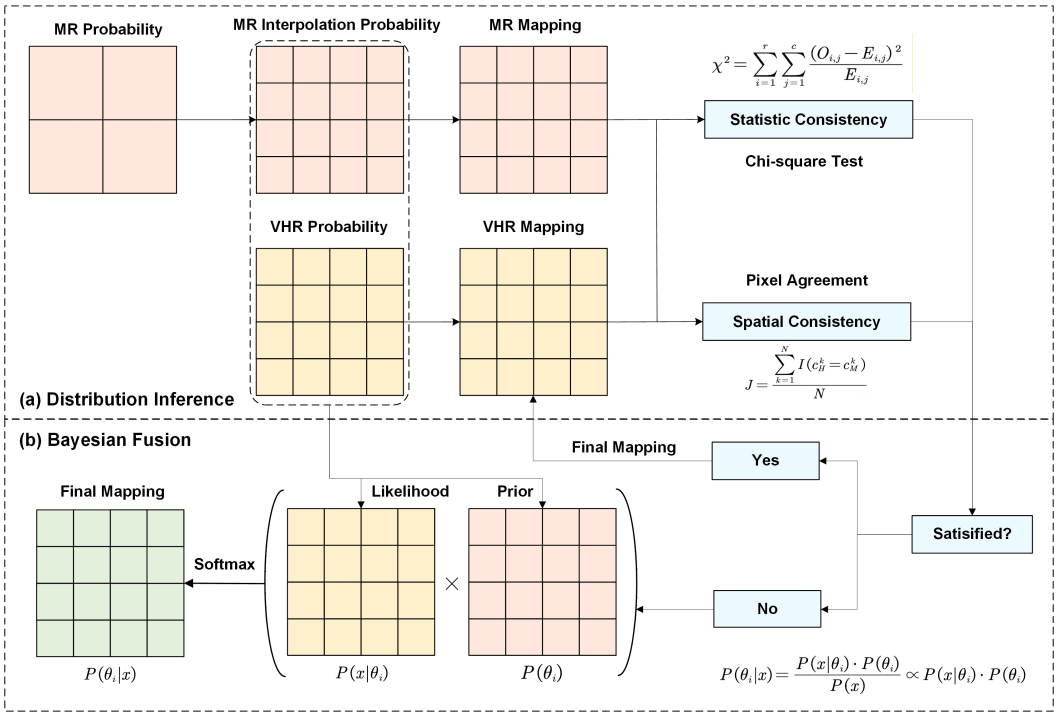


Figure 8. Decision-level Bayesian fusion framework for VHR and MR land cover mapping integration. (**a**) Distribution inference through statistic (chi-square test) and spatial (pixel agreement) consistency test; (**b**) adaptive fusion strategy with direct adoption of VHR results when consistency is high, or Bayesian fusion combining MR prior and VHR likelihood when inconsistency is detected.

First, the MR classification result (including probability outputs) is resampled to match the spatial resolution of the VHR result using bilinear interpolation. For each 512 × 512 image patch, distribution inference is performed by assessing statistical consistency via a chi-square test and spatial consistency via pixel agreement. The chi-square statistic χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{16}$$

where $O_{i,j}$ is the observed joint frequency, $E_{i,j}$ is the expected joint frequency under independence, and n is the number of land cover classes. Pixel agreement quantifies the spatial consistency between the two classification maps:

$$J = \frac{\sum_{k=1}^{N} I\left(c_H^k = c_M^k\right)}{N} \tag{17}$$

where c_H^k and c_M^k are the class labels of the k-th pixel in the VHR and MR maps, respectively, N is the total number of pixels in the patch, and $I(\cdot)$ is the indicator function. Given the significance level p, if the p-value of the chi-square test is greater than p and the pixel agreement J is greater than $1 - p$, the two maps are considered highly consistent in both statistical distribution and spatial layout. In this case, the VHR map is adopted as the final output to maximally preserve its detailed spatial features. Conversely, if the consistency criteria are not met, indicating notable discrepancies, a decision-level Bayesian fusion method is utilized to reconcile the two results.

The Bayesian framework is inherently suited for multi-source data fusion as it provides a principled probabilistic mechanism to integrate stable but coarse MR priors with detailed but uncertain VHR observations. Within the Bayesian framework, the MR classification probability is treated as the prior probability $P(\theta_i)$ since it provides reliable spectral information, and the VHR classification probability serves as the likelihood probability $P(x|\theta_i)$. The posterior probability for class θ_i, given pixel x is computed as

$$P(\theta_i|x) = \frac{P(x|\theta_i)\cdot P(\theta_i)}{P(x)} \propto P(x|\theta_i)\cdot P(\theta_i) \tag{18}$$

where $P(\theta_i|x)$ represents the posterior probability. Since the marginal probability $P(x)$ is constant across all classes for a given pixel, the final normalized posterior probability is obtained via the softmax function:

$$P(\theta_i|x) = softmax\{P(x|\theta_i)\cdot P(\theta_i)\} \tag{19}$$

For each pixel, the class with the highest posterior probability is selected as the final label. This framework adaptively chooses the fusion strategy based on consistent distribution inference. It preserves the fine spatial details of the VHR result when consistency is high and leverages the Bayesian mechanism to integrate the spectrally stable prior from MR data with the VHR observational evidence when inconsistency occurs. This approach effectively reduces systematic errors in VHR land cover mapping attributable to image heterogeneity, domain shift, and limited spectral information.

3.5. Evaluation Metrics and Implementation Details

For VHR-only mapping, overall accuracy (OA) and mean intersection over union (mIoU) are used to quantify overall model performance. Additionally, computational cost is quantitatively analyzed using floating point operations (FLOPs) and model parameters (Params). For the final fused mapping results, besides OA and mIoU, intersection over

union (IoU) is also used to evaluate the classification accuracy for each class. For MR mapping, user accuracy (UA) and producer accuracy (PA) are adopted to evaluate per-class accuracy, in conjunction with OA and the Kappa coefficient (Kappa) for comprehensive performance evaluation.

Assuming a classification system consisting of n land cover classes, the terms of the confusion matrix are defined as follows: p_{ii} denotes the number of pixels of class i that are correctly classified, while p_{ij} represents the number of pixels of class j that are misclassified as class i. When class i is considered as the positive class and all remaining classes are treated as negative, the basic accuracy components are defined under a binary classification framework as: True Positive (TP), where both the prediction and the reference label are class i; True Negative (TN), where both the prediction and the reference label are not class i; False Positive (FP), where pixels belonging to any other class are incorrectly classified as class i; and False Negative (FN), where pixels of class i are misclassified as other classes. These components are then used to compute standard metrics (OA, mIoU, UA, PA, and Kappa) as well as metrics for each class in the decision-level Bayesian fusion results (Precision, Recall, and F1). Where TP_i, FP_i, and FN_i correspond to the TP, FP, and FN for class i, the abovementioned metrics are computed as follows:

$$\mathrm{IoU}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i + \mathrm{FP}_i} = \frac{p_{ii}}{\sum_{j=1}^{n} p_{ij} + \sum_{i=1}^{n} p_{ij} - p_{jj}} \tag{20}$$

$$\mathrm{mIoU} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IoU_i} \tag{21}$$

$$\mathrm{OA} = \frac{\sum_{i=1}^{n} p_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}} \tag{22}$$

$$\mathrm{UA} = \frac{p_{ii}}{\sum_{j=1}^{n} p_{ij}} \tag{23}$$

$$\mathrm{PA} = \frac{p_{ii}}{\sum_{i=1}^{n} p_{ij}} \tag{24}$$

$$\mathrm{P_o} = \mathrm{OA} = \frac{\sum_{i=1}^{n} p_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}} \tag{25}$$

$$\mathrm{P_e} = \frac{\sum_{i=1}^{n}\left(\sum_{i=1}^{n} p_{ij} \times \sum_{j=1}^{n} p_{ij}\right)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\right)^2} \tag{26}$$

$$\mathrm{Kappa} = \frac{\mathrm{P_o} - \mathrm{P_e}}{1 - \mathrm{P_e}} \tag{27}$$

$$\mathrm{Precision}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + FP_i} \tag{28}$$

$$\mathrm{Recall}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{29}$$

$$\mathrm{F1}_i = 2 \cdot \frac{\mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{30}$$

IoU characterizes the spatial overlap precision of specific classes in semantic segmentation by calculating the ratio of the intersection to the union of the "model prediction" and the "ground truth" sets. mIoU, defined as the mean of the IoU across all classes, assesses the model's overall segmentation performance for multi-class features. OA is defined as the proportion of correctly classified pixels relative to the total number of pixels, serving as a fundamental metric for intuitively measuring the global classification accuracy of the model. UA represents the proportion of pixels correctly classified among those belonging to a certain class in the ground truth, reflecting the probability that a user observing a class in the mapping result encounters that feature in reality. PA denotes the proportion of pixels predicted as a certain class that match the ground truth, reflecting the effectiveness of omission error control from the perspective of the data producer. Kappa measures the agreement between classification results and ground truth that exceeds random chance, calculated by comparing observed agreement with expected random agreement.

The experiments were conducted on a server equipped with four NVIDIA GeForce RTX 4090 D GPUs. The operating system was Ubuntu 22.04.1 LTS with kernel version 6.8.0-85-generic. The software stack was built upon CUDA 12.4, Python 3.11.13, and the PyTorch 1.11.0 deep learning framework. For all models, results are based on three independent runs with different random seeds, with mean values reported.

During pretraining, we employed the AdamW optimizer with an initial learning rate of 5×10^{-4}, betas of 0.9 and 0.95, weight decay of 0.05, and a batch size of 4. The learning rate followed a LambdaLR schedule with a warm-up phase, defined as

$$lr_{function}(epoch) = \min\left\{\frac{epoch+1}{warmup_epoch+1\times 10^{-8}}, \frac{1}{2}\left(cos\left(\frac{\pi \cdot epoch}{total_epoch}\right)+1\right)\right\} \quad (31)$$

where the *total_epoch* was 200 and the *warmup_epoch* was 20. Input images (512 × 512) were partitioned into 32 × 32 patches.

For the fine-tuning of SBFNet and all comparative models, we used the Adam optimizer with betas set to 0.9 and 0.999 and weight decay set to 0. The learning rate was scheduled using CosineAnnealingLR, with an initial learning rate of 1×10^{-4} and a minimum learning rate of 1×10^{-6}, and *T_max* was set to 10. Training ran for 100 epochs with a batch size of 8. All fine-tuning experiments were conducted using PyTorch's Distributed Data Parallel mode with mixed-precision training. The dataset was split into training, validation, and test sets in a 7:1:2 ratio. The final model for accuracy assessment was selected based on the lowest validation loss. The threshold p for Bayesian fusion was set to 0.05, which is widely adopted as the standard significance level in statistical hypothesis testing and remote sensing uncertainty analysis.

4. Results

4.1. Pretraining Performance of SDMAE

We first evaluated the reconstruction performance of the SDMAE on PRD262K. Regarding the computational efficiency, pretraining the MAE on the PRD262K dataset required approximately 3.34 h per epoch on our server, while the proposed SDMAE required approximately 3.37 h per epoch. The results indicate that there is no significant difference in training time between the two, demonstrating that our proposed enhancements do not introduce substantial computational overhead.

To further address the visual discriminability of the reconstruction, we compared the results of the SDMAE (Figure 9) with those of the standard MAE (Figure A1). While both models accurately recover global structures and spectral characteristics, the SDMAE exhibits superior fidelity in restoring high-frequency details compared to the smoother results in Figure A1. As shown in Figure 9, the SDMAE maintains high visual consistency

with the original images, especially in areas with continuous spatial structures and stable textures, such as forests and water bodies. The boundaries between farmlands and impervious surfaces are also well-recovered. This indicates that the encoder effectively captures the representative and universal features inherent in the PRD.

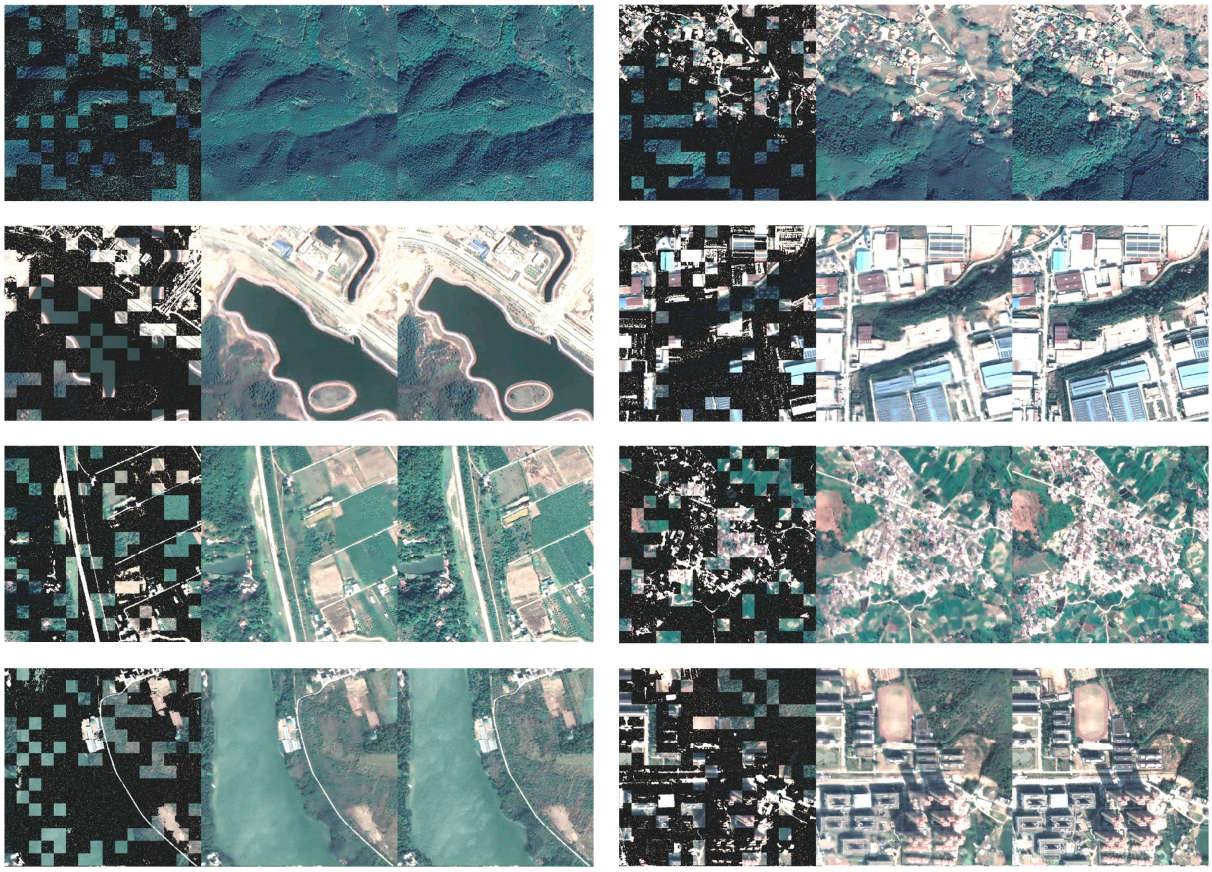


Figure 9. Examples of visualized results for our SDMAE. For each batch, (left to right), there is a masked image, reconstructed image and original image, respectively.

Furthermore, the model demonstrates robust reconstruction capabilities in regions with complex land cover types and strong spatial heterogeneity. The VMask Strategy and Edge-Enhanced Loss improve reconstruction performance. Transition zones between forests and farmlands are accurately preserved. Similarly, small-scale urban targets, such as buildings and fine-grained roads, remain distinct in the results. This shows the proficiency of the SDMAE in modeling critical structural information and fine-grained features.

4.2. Evaluation of VHR Mapping

To verify the effectiveness of our proposal in VHR land cover mapping tasks, several mainstream models including UNet [17], PSPNet [19], DeepLabV3+ [18], SETR [20], and SegFormer [21] were selected as benchmarks. The performance of these models, along with various configurations of our SBFNet series (including different pretraining strategies like the MAE [42] and SDMAE), is evaluated in this section. Specifically, the distinction between the MAE and SDMAE when pretraining is omitted lies in the decoder's incorporation of value-guided masked patches.

Comparative experiments were conducted on four datasets, including LoveDA [65], DLRSD [66], WHDLD [67] and our proposed PRDLC-PRO. The quantitative comparison results for each model across the LoveDA, DLRSD, WHDLD, and PRDLC-PRO datasets are summarized in Tables 2–5, respectively. The corresponding visual comparisons of the prediction results are presented in Figures 10–13.

Table 2. Quantitative comparison of different models on the LoveDA dataset.

Model	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
UNet	68.94	50.35	124.58	**13.40**
PSPNet	70.15	52.97	59.22	46.71
DeepLabV3+	69.68	52.88	56.58	40.52
SETR	66.56	45.86	41.97	89.77
SegFormer	66.40	46.01	153.89	136.66
MAE-SBFNet-0P-20E	57.29	32.78	**29.18**	92.71
MAE-SBFNet-0P-100E	66.66	46.17	**29.18**	92.71
SDMAE-SBFNet-0P-20E	60.54	38.24	29.85	93.30
SDMAE-SBFNet-0P-100E	68.79	49.31	29.85	93.30
MAE-SBFNet-50P-20E	66.94	46.89	**29.18**	92.71
MAE-SBFNet-50P-100E	69.35	50.00	**29.18**	92.71
MAE-SBFNet-100P-20E	67.81	47.99	**29.18**	92.71
MAE-SBFNet-100P-100E	69.34	50.34	**29.18**	92.71
SDMAE-SBFNet-50P-20E	66.78	46.33	29.85	93.30
SDMAE-SBFNet-50P-100E	68.85	50.54	29.85	93.30
SDMAE-SBFNet-100P-20E	69.85	50.77	29.85	93.30
SDMAE-SBFNet-100P-100E	**71.53**	**53.21**	29.85	93.30

Note: "XP-YE" indicates X epochs of pretraining and Y epochs of fine-tuning (e.g., 0P means no pretraining, 50P means 50 epochs of pretraining, and 100E means 100 epochs of fine-tuning). Bold indicates the best result among all methods.

Table 3. Quantitative comparison of different models on the DLRSD dataset.

Model	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
UNet	70.25	50.06	124.58	13.40
PSPNet	70.43	**50.12**	59.22	46.71
DeepLabV3+	69.36	47.07	56.58	40.52
SETR	66.80	44.65	41.97	89.77
SegFormer	68.88	49.14	153.89	136.66
MAE-SBFNet-0P-20E	60.46	30.48	**29.18**	92.71
MAE-SBFNet-0P-100E	69.72	44.24	**29.18**	92.71
SDMAE-SBFNet-0P-20E	50.72	24.47	29.85	93.30
SDMAE-SBFNet-0P-100E	70.49	46.30	29.85	93.30
MAE-SBFNet-50P-20E	59.46	30.10	**29.18**	92.71
MAE-SBFNet-50P-100E	69.36	46.90	**29.18**	92.71
MAE-SBFNet-100P-20E	48.37	21.81	**29.18**	92.71
MAE-SBFNet-100P-100E	71.09	43.76	**29.18**	92.71
SDMAE-SBFNet-50P-20E	65.83	39.00	29.85	93.30
SDMAE-SBFNet-50P-100E	69.53	46.32	29.85	93.30
SDMAE-SBFNet-100P-20E	67.74	44.10	29.85	93.30
SDMAE-SBFNet-100P-100E	**72.22**	46.86	29.85	93.30

Note: "XP-YE" indicates X epochs of pretraining and Y epochs of fine-tuning (e.g., 0P means no pretraining, 50P means 50 epochs of pretraining, and 100E means 100 epochs of fine-tuning). Bold indicates the best result among all methods.

Table 4. Quantitative comparison of different models on the WHDLD dataset.

Model	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
UNet	78.41	50.32	124.58	13.40
PSPNet	78.13	51.53	59.22	46.71
DeepLabV3+	78.01	51.67	56.58	40.52
SETR	76.39	48.53	41.97	89.77
SegFormer	77.15	49.22	153.89	136.66
SBFNet-0P-20E	74.55	44.89	**29.18**	92.71
SBFNet-0P-100E	76.15	48.37	**29.18**	92.71
SDSBFNet-0P-20E	75.35	46.36	29.85	93.30
SDSBFNet-0P-100E	75.84	47.24	29.85	93.30
MAE-SBFNet-50P-20E	77.84	50.30	**29.18**	92.71
MAE-SBFNet-50P-100E	77.84	50.30	**29.18**	92.71
MAE-SBFNet-100P-20E	76.64	48.33	**29.18**	92.71
MAE-SBFNet-100P-100E	78.21	50.55	**29.18**	92.71
SDMAE-SBFNet-50P-20E	78.84	51.11	29.85	93.30
SDMAE-SBFNet-50P-100E	79.22	**52.28**	29.85	93.30
SDMAE-SBFNet-100P-20E	76.22	48.69	29.85	93.30
SDMAE-SBFNet-100P-100E	**79.49**	52.26	29.85	93.30

Note: "XP-YE" indicates X epochs of pretraining and Y epochs of fine-tuning (e.g., 0P means no pretraining, 50P means 50 epochs of pretraining, and 100E means 100 epochs of fine-tuning). Bold indicates the best result among all methods.

Table 5. Quantitative comparison of different models on the PRDLC-PRO dataset.

Model	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
UNet	89.30	63.58	124.58	13.40
PSPNet	87.91	62.04	59.22	46.71
DeepLabV3+	88.41	64.13	56.58	40.52
SETR	83.71	56.82	41.97	89.77
SegFormer	85.14	57.98	153.89	136.66
SBFNet-0P-20E	73.33	35.96	**29.18**	92.71
SBFNet-0P-100E	85.18	57.58	**29.18**	92.71
SDSBFNet-0P-20E	76.47	42.50	29.85	93.30
SDSBFNet-0P-100E	85.68	58.83	29.85	93.30
MAE-SBFNet-50P-20E	84.74	55.17	**29.18**	92.71
MAE-SBFNet-50P-100E	88.69	64.77	**29.18**	92.71
MAE-SBFNet-100P-20E	83.16	54.57	**29.18**	92.71
MAE-SBFNet-100P-100E	88.42	63.44	**29.18**	92.71
SDMAE-SBFNet-50P-20E	83.99	51.56	29.85	93.30
SDMAE-SBFNet-50P-100E	88.28	66.57	29.85	93.30
SDMAE-SBFNet-100P-20E	81.99	49.88	29.85	93.30
SDMAE-SBFNet-100P-100E	87.98	**66.61**	29.85	93.30

Note: "XP-YE" indicates X epochs of pretraining and Y epochs of fine-tuning (e.g., 0P means no pretraining, 50P means 50 epochs of pretraining, and 100E means 100 epochs of fine-tuning). Bold indicates the best result among all methods.

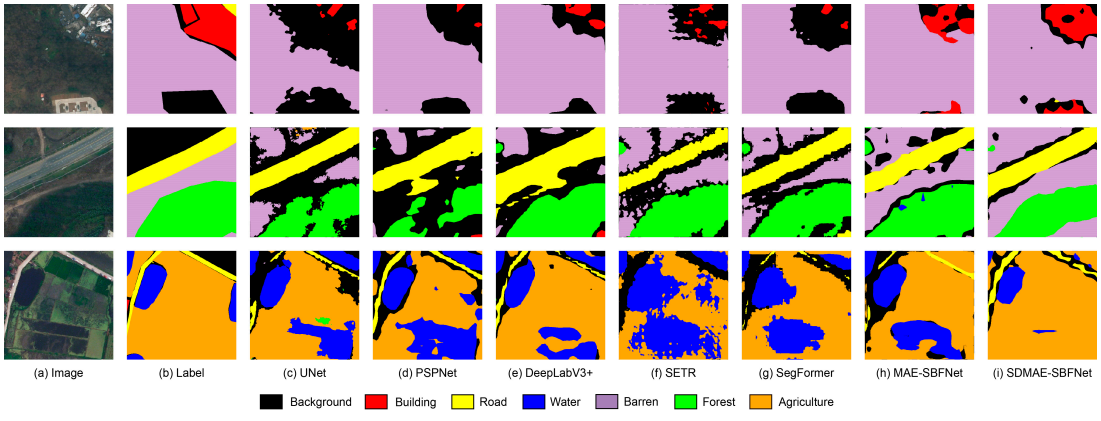


Figure 10. Visual comparison of land cover mapping results on the LoveDA dataset. (**a**) Original VHR image; (**b**) ground truth label; (**c**–**i**) prediction results from UNet, PSPNet, DeepLabV3+, SETR, SegFormer, MAE-SBFNet-100P-100E, and SDMAE-SBFNet-100P-100E, respectively.

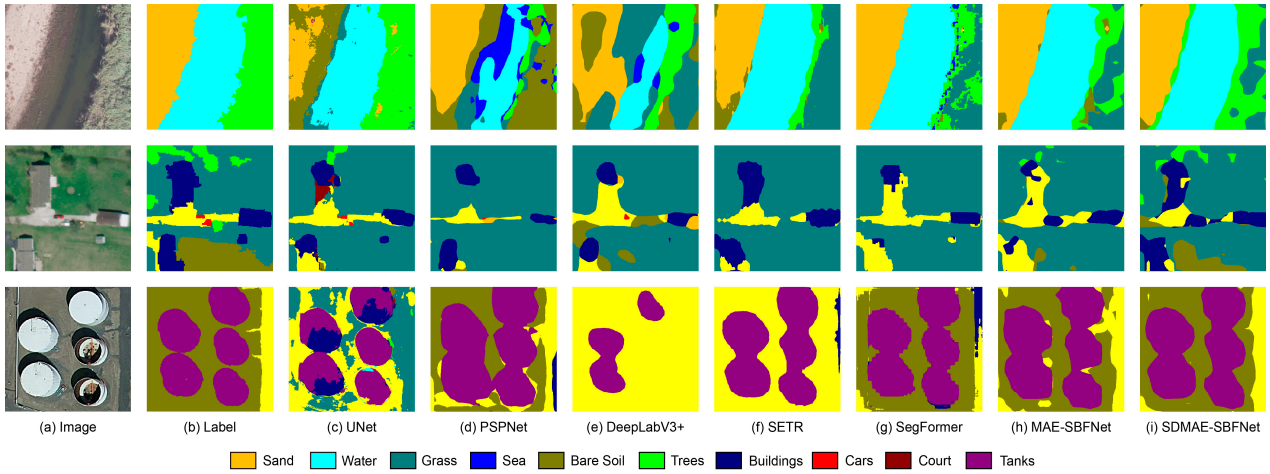


Figure 11. Visual comparison of land cover mapping results on the DLRSD dataset. (**a**) Original VHR image; (**b**) ground truth label; (**c**–**i**) prediction results from UNet, PSPNet, DeepLabV3+, SETR, SegFormer, MAE-SBFNet-100P-100E, and SDMAE-SBFNet-100P-100E, respectively.

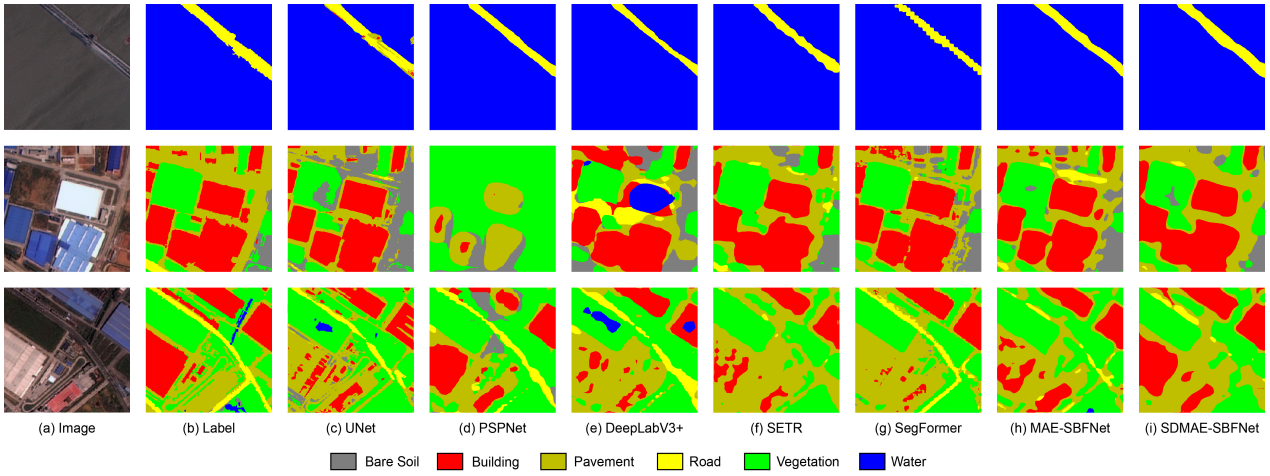


Figure 12. Visual comparison of land cover mapping results on the WHDLD dataset. (**a**) Original VHR image; (**b**) ground truth label; (**c**–**i**) prediction results from UNet, PSPNet, DeepLabV3+, SETR, SegFormer, MAE-SBFNet-100P-100E, and SDMAE-SBFNet-100P-100E, respectively.

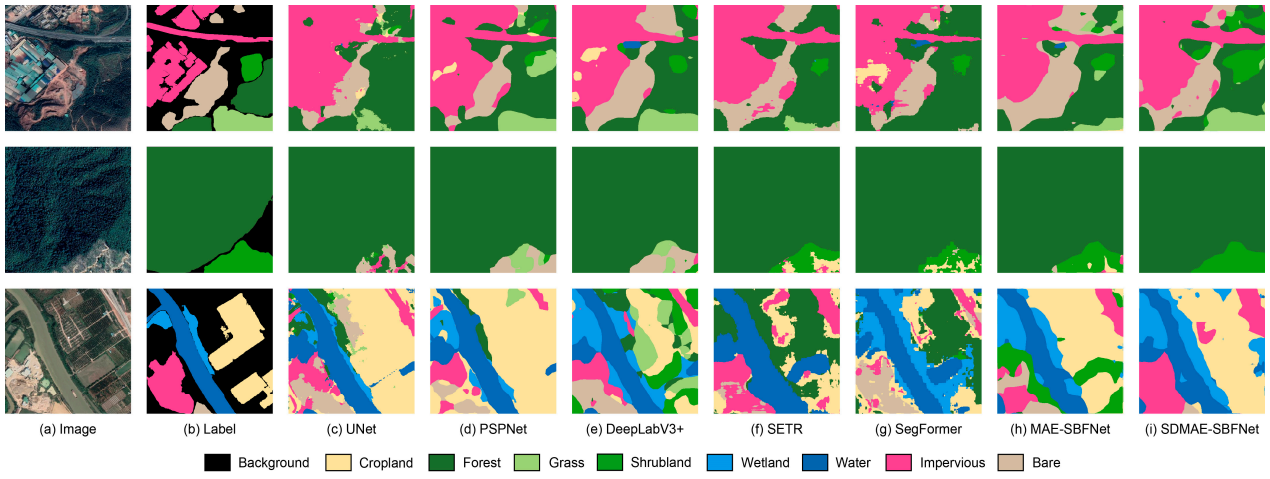


Figure 13. Visual comparison of land cover mapping results on the PRDLC-PRO dataset. (**a**) Original VHR image; (**b**) ground truth label; (**c**–**i**) prediction results from UNet, PSPNet, DeepLabV3+, SETR, SegFormer, MAE-SBFNet-100P-100E, and SDMAE-SBFNet-100P-100E, respectively.

Across all four datasets, SDMAE-SBFNet demonstrated consistent outperformance over baseline methods, with improvements ranging from 1.38% to 3.00% in OA and 0.24% to 3.00% in mIoU. These gains highlight its strong discriminative power and cross-scene adaptability for VHR land cover mapping, which are attributed to the emphasis on high-value and edge areas, as well as its robust scene-based feature extraction capability.

On LoveDA, which features complex urban–rural transition scenes, SDMAE-SBFNet-100P-100E reached an OA of 71.53% and an mIoU of 53.21%, surpassing PSPNet by 1.38% and 0.24% and MAE-SBFNet by 2.19% and 2.87%, respectively. As can be seen in Figure 10, our proposal produces the fewest missed detections for scattered buildings and small water bodies, maintains superior road continuity and smoother boundaries, and achieves higher internal consistency in forest and cropland regions.

In the dense small-object scenarios of DLRSD, SDMAE-SBFNet-100P-100E achieved an OA of 72.22% and an mIoU of 46.86%, while MAE-SBFNet-100P-20E dropped to 48.37% in OA. SDMAE-SBFNet-100P-20E maintained an OA of 67.74%, demonstrating robustness under limited fine-tuning. Figure 11 highlights the accurate boundary preservation and geometric structure maintenance of our method in coastal, rural road, and oil facility scenes.

For the relatively homogeneous urban scenes in WHDLD, SDMAE-SBFNet-100P-100E reached an OA of 79.46% and an mIoU of 52.26%, exceeding UNet by 1.94% in mIoU and MAE-SBFNet by 1.71%. Our SDMAE-SBFNet achieves the fewest misclassifications for the building and pavement, while also producing smooth boundaries and high spatial consistency in these areas (Figure 12).

In the heterogeneous landscapes of PRDLC-PRO, our SDMAE-SBFNet-100P-100E achieved the highest mIoU of 66.61%, surpassing UNet and DeepLabV3+ by 3.00% and 2.44%, with OA remaining stable at 87.98%. Figure 13 demonstrates the robustness of the SDMAE and SBFNet under illumination variations and shadow effects, with fewer misclassifications and closer alignment with the ground truth, indicating adaptability to complex imaging conditions and heterogeneous landscapes typical of the PRD.

Overall, SDMAE-SBFNet consistently surpasses MAE-SBFNet, reflecting the advantages of pretraining with the VMask Strategy and Edge-Enhanced Loss, which enhance feature discrimination and cross-scene generalization. Compared to Transformer-based models such as SETR and SegFormer, which still exhibit edge fragmentation, and convolutional models like UNet, PSPNet, and DeepLabV3+, which are prone to class confusion or boundary blurring in complex scenes, SDMAE-SBFNet achieves more balanced segmentation, effectively reducing misclassification while preserving fine edge details. Pretraining further elevates the performance ceiling and accelerates fine-tuning convergence, demonstrating its critical role in achieving robust VHR land cover mapping.

4.3. Ablation Study

We conducted ablation studies to validate the effectiveness of our proposed SDMAE, and the results are presented in Table 6. Generally, the introduction of either module improves model performance to varying degrees. When the VMask Strategy and Edge-Enhanced Loss are utilized simultaneously, the model achieves optimal results in both OA and mIoU, indicating a robust complementarity between the two modules in feature modeling. This synergy likely arises because the VMask Strategy preserves high-value object features critical for capturing dominant structures, while the Edge-Enhanced Loss maintains edge features, enabling the model to learn both dominant contextual features and fine features effectively. On the LoveDA dataset, the OA of the full model increases to 71.53% and the mIoU to 53.21%, demonstrating that both modules significantly promote the discriminative capability for multi-class geographical features in complex urban scenes. The DLRSD dataset results show that the Edge-Enhanced Loss yields a more pronounced

improvement in OA and mIoU, reflecting the importance of edge information for VHR aerial image classification. On the WHDLD dataset, the combination of both modules further enhances the mIoU, indicating that the model remains robust on datasets with balanced class distributions. On the PRDLC-PRO dataset, after introducing the VMask Strategy, the mIoU improves from 63.44% to 66.47%, proving that this module remains effective for fine-grained class discrimination within complex backgrounds. Moreover, the performance gains brought by both modules are accompanied by only a marginal increase in FLOPs and Params, embodying a favorable balance between performance and computational efficiency.

Table 6. Ablation study of SDMAE on four datasets.

Datasets	VMask Strategy	Edge-Enhanced Loss	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
LoveDA	×	×	69.34	50.34	**29.18**	**92.71**
LoveDA	×	✓	69.35	50.76	**29.18**	**92.71**
LoveDA	✓	×	70.86	52.07	29.85	93.30
LoveDA	✓	✓	**71.53**	**53.21**	29.85	93.30
DLRSD	×	×	71.09	43.76	**29.18**	**92.71**
DLRSD	×	✓	**74.50**	45.94	**29.18**	**92.71**
DLRSD	✓	×	71.47	42.77	29.85	93.30
DLRSD	✓	✓	72.22	**46.86**	29.85	93.30
WHDLD	×	×	78.21	50.55	**29.18**	**92.71**
WHDLD	×	✓	77.04	48.84	**29.18**	**92.71**
WHDLD	✓	×	78.93	52.29	29.85	93.30
WHDLD	✓	✓	**79.49**	**52.56**	29.85	93.30
PRDLC-PRO	×	×	**88.42**	63.44	**29.18**	**92.71**
PRDLC-PRO	×	✓	86.73	63.09	**29.18**	**92.71**
PRDLC-PRO	✓	×	87.82	66.47	29.85	93.30
PRDLC-PRO	✓	✓	87.98	**66.61**	29.85	93.30

Note: "×" indicates the component is ablated and "✓" indicates the component is included. Bold indicates the best result of the ablation study.

Furthermore, we performed ablation experiments to assess the individual contributions of the SBMs and GSBMs within the SBFNet, with detailed results provided in Table 7. As the FBM serves as the backbone feature extraction component and cannot be omitted, we evaluated the effects of adding the SBM and GSBM incrementally based on the FBM baseline. Experimental results indicate that integrating either module consistently enhances model performance, and the combination of all three modules yields the best results across all four datasets. This improvement can be attributed to the complementary roles of the modules, where the SBM strengthens the model's ability to capture scene-level contextual cues essential for interpreting complex land cover structures, while the GSBM supplies global spatial context, facilitating the fusion of local details with broader semantic information. For instance, on the LoveDA dataset, the full model attains 71.53% OA and 53.21% mIoU, significantly outperforming the FBM-only baseline. On the PRDLC-PRO dataset, the inclusion of the SBM alone leads to a notable mIoU gain from 62.89% to 65.01%, underscoring its utility in distinguishing fine-grained categories in heterogeneous backgrounds. Importantly, these performance improvements are achieved with only a slight increase in computational cost, demonstrating an effective trade-off between accuracy and efficiency.

Table 7. Ablation study of SBFNet on four datasets.

Datasets	FBM	SBM	GSBM	OA (%)	mIoU (%)	FLOPs (G)	Params (M)
LoveDA	✓	×	×	68.86	49.86	**29.81**	**93.29**
LoveDA	✓	✓	×	70.38	51.95	29.82	**93.29**
LoveDA	✓	×	✓	71.38	52.45	29.84	93.30
LoveDA	✓	✓	✓	**71.53**	**53.21**	29.85	93.30
DLRSD	✓	×	×	68.47	44.39	**29.81**	**93.29**
DLRSD	✓	✓	×	69.92	46.51	29.82	**93.29**
DLRSD	✓	×	✓	68.51	45.92	29.84	93.30
DLRSD	✓	✓	✓	**72.22**	**46.86**	29.85	93.30
WHDLD	✓	×	×	77.24	50.55	**29.81**	**93.29**
WHDLD	✓	✓	×	78.25	48.84	29.82	**93.29**
WHDLD	✓	×	✓	77.94	52.29	29.84	93.30
WHDLD	✓	✓	✓	**79.49**	**52.56**	29.85	93.30
PRDLC-PRO	✓	×	×	87.26	62.89	**29.81**	**93.29**
PRDLC-PRO	✓	✓	×	87.86	65.01	29.82	**93.29**
PRDLC-PRO	✓	×	✓	87.76	64.47	29.84	93.30
PRDLC-PRO	✓	✓	✓	**87.98**	**66.61**	29.85	93.30

Note: "×" indicates the component is ablated and "✓" indicates the component is included. Bold indicates the best result of the ablation study.

4.4. Evaluation of MR Mapping

The classification accuracy of the MR land cover mapping results, based on multi-source remote sensing MR data and the random forest model, is presented in Table 8. Overall, the model achieved an OA of 76.54% and a Kappa of 0.684 in the PRD. These results indicate that the MR mapping results possess sufficient reliability, serving as a stable source of prior information for subsequent fusion analysis.

Table 8. Accuracy assessment of MR land cover mapping using random forest model.

Class	UA (%)	PA (%)
Cropland	61.06	52.33
Forest	92.83	94.21
Grass	67.09	59.49
Shrubland	62.33	64.96
Wetland	67.37	76.33
Water	96.29	97.52
Impervious	85.25	84.69
Bare	82.29	85.70
OA (%)	76.54	
Kappa	0.684	

From the perspective of accuracy at the class level, the forest and water classes exhibited the most prominent performance, with both UA and PA exceeding 92%. This is primarily attributed to the high discriminability of these features within the multispectral, SAR, and topographic feature spaces, combined with their strong spatial distribution continuity. The classification accuracies for impervious and bare also remained at a high level, reflecting the complementary advantages of multi-source MR remote sensing data in identifying artificial surfaces and bare land.

In contrast, natural land surface types such as cropland, grass, and shrubland yielded relatively lower accuracies, mainly constrained by the similarities in seasonal variations,

management practices, and spectral characteristics among these classes. This phenomenon is particularly evident in the discrepancies between PA and UA, indicating that MR mapping still encounters certain confusion within transitional land cover types. Nevertheless, these results remain at a reasonable level for regional-scale land cover mapping and provide stable, credible prior information for VHR mapping.

4.5. Performance of Decision-Level Bayesian Fusion

To evaluate the enhancement of mapping accuracy achieved by the proposed decision-level Bayesian fusion, a quantitative comparative analysis was conducted between the results based on the VHR model output (denoted as "VHR-only") and the results integrated with MR prior information (denoted as "Fusion"). The results are summarized in Table 9. All results were obtained based on the PRDLC dataset.

Table 9. Quantitative comparison of land cover mapping performance before and after decision-level Bayesian fusion on the PRDLC dataset.

VHR-Only	Cropland	Forest	Grass	Shrubland	Wetland	Water	Impervious	Bare
Precision	**0.945**	0.979	0.921	**0.805**	0.684	0.925	0.929	**0.907**
Recall	0.920	0.985	0.805	0.842	0.869	0.908	**0.951**	0.860
F1	**0.932**	0.982	0.859	**0.823**	0.765	0.916	**0.940**	**0.883**
IoU(%)	**87.27**	96.44	75.22	**69.93**	61.98	84.56	**88.63**	**79.04**
OA(%)				95.37				
mIoU(%)				80.38				
Fusion	**Cropland**	**Forest**	**Grass**	**Shrubland**	**Wetland**	**Water**	**Impervious**	**Bare**
Precision	0.892	**0.981**	**0.922**	0.802	**0.721**	**0.969**	**0.941**	0.898
Recall	**0.927**	**0.988**	**0.816**	**0.845**	**0.861**	**0.933**	0.939	**0.866**
F1	0.909	**0.984**	**0.866**	**0.823**	**0.785**	**0.95**	**0.940**	0.882
IoU(%)	83.34	**96.93**	**76.34**	**69.93**	**64.56**	**90.54**	**88.63**	78.81
OA(%)				**95.99**				
mIoU(%)				**81.13**				

Note: Bold indicates the better result between VHR-Only and Fusion for each metric.

Generally, after the integration of MR prior information, the OA of the fusion results increased from 95.37% to 95.99% and the mIoU improved from 80.38% to 81.13%. This indicates that the proposed fusion strategy enhances overall classification accuracy while simultaneously improving the balance across various classes. The IoU for most land cover classes exhibited an upward trend, with wetland and water showing the most significant improvements. This suggests that the stable MR priors can effectively suppress misclassifications by the VHR model in areas with noise or low discriminability, thereby enhancing the classification performance.

For classes with distinct spatial structures, such as impervious and forest, the IoU remained at high levels both before and after fusion. This demonstrates that the fusion process does not adversely affect high-confidence classes, preserving the inherent advantages of the VHR model. Notably, a marginal decline in metrics was observed for cropland, which can be attributed to the interplay between the resolution constraints of the MR prior and the inherent fragmentation of cropland landscapes. The coarser spatial resolution of MR data often introduces mixed-pixel interference in small or geometrically complex patches, slightly limiting the precision of the fusion in these transition zones. Nevertheless, this slight trade-off is acceptable given the overall gains in consistency and error rectification.

Figure 14 presents a visual comparison of the mapping results in typical regions before and after fusion. From the visual comparison of typical regions before and after fusion,

it can be observed that for areas characterized by image mosaicking heterogeneity, the fusion results effectively smooth the discontinuities at land cover boundaries and rectify localized classification disorder. In regions with complex boundaries, such as grass and shrubland, the fusion strategy ensures higher continuity and consistency of land cover classes, resulting in clearer boundary delineations. Furthermore, in areas affected by clouds and shadows, the fusion results perform significantly better than the standalone VHR model in terms of mitigating noise and recovering information of obscured geographical features. This further validates the effectiveness of the fusion strategy in handling localized image interference and enhancing mapping stability. The visual evidence in Figure 14 confirms that the framework's primary strength lies in its ability to flip localized systemic errors into correct predictions, which is crucial for the practical confidence of large-scale mapping.

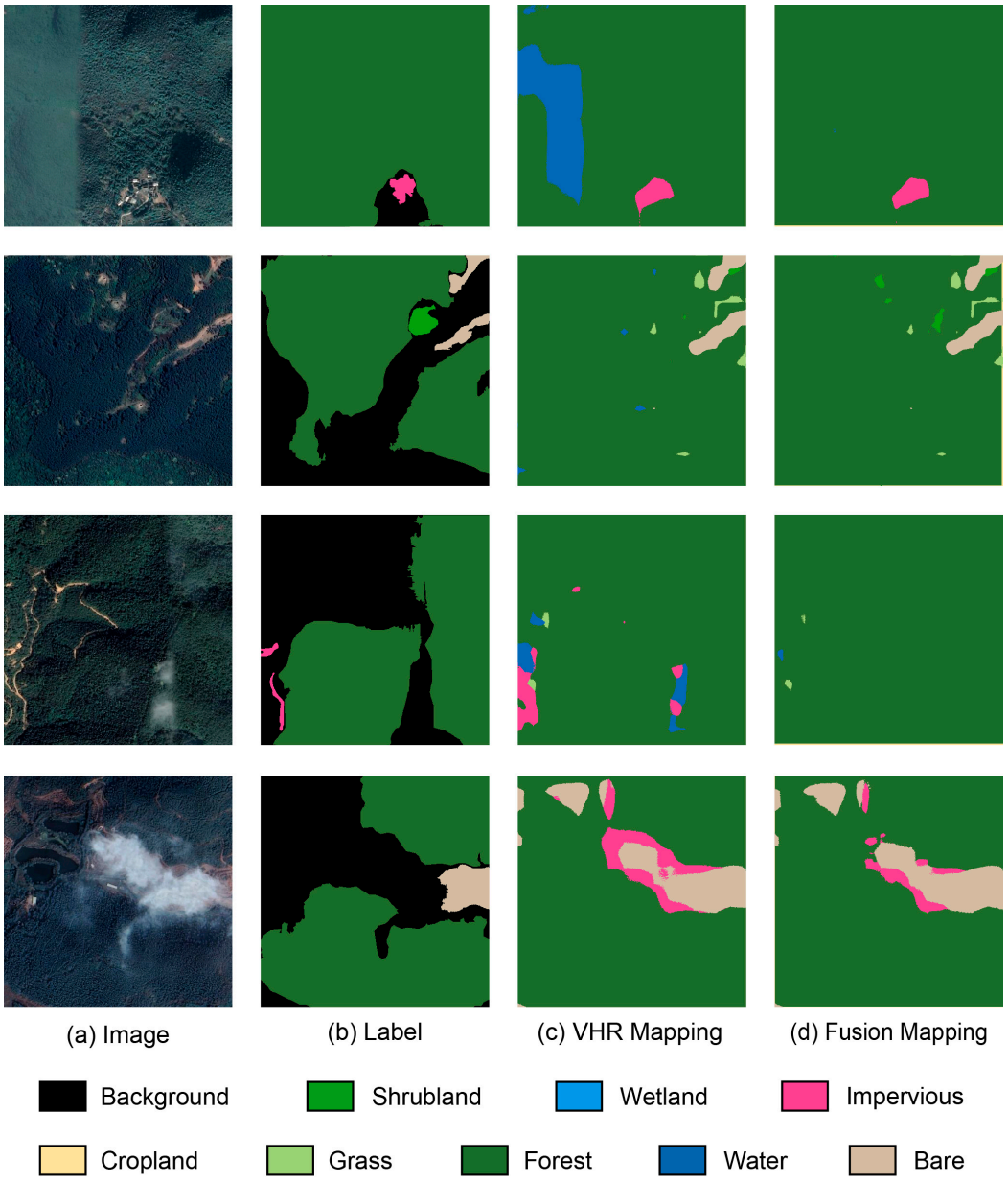


Figure 14. Visual comparison of land cover mapping results before and after decision-level Bayesian fusion in typical regions. (**a**) Original VHR image; (**b**) ground truth label; (**c**) VHR-only mapping result; (**d**) fusion mapping result integrating MR data.

In summary, our proposed decision-level Bayesian fusion based on non-parametric statistics and the Bayesian prior framework successfully integrates the spatial detail capture capability of a VHR model with the stable features of MR mapping. While maintaining the fine-grained representation of geographical features, it significantly improves the consistency and reliability of the overall classification results.

5. Discussion

5.1. Implications of the Decision-Level Bayesian Fusion

The decision-level Bayesian fusion we proposed integrates non-parametric statistical testing with Bayesian priors, achieving a synergistic "macro-prior from MR and micro-posterior from VHR" framework for multi-source remote sensing data. This aligns with the paradigm of using coarse-resolution data as a context to guide fine-scale mapping [75,76], while advancing it through a formal probabilistic fusion mechanism. This section elaborates on the implications of this mechanism using the first example in Figure 14.

The upper part of Figure 15 displays (a) the original image, (b) the ground truth label, (c) the VHR mapping result, and (d) the fusion mapping result. Due to clouds and shadows in the VHR imagery, a well-documented source of error in optical remote sensing classification [77,78], the VHR mapping result misclassified forest as water and impervious. However, after incorporating the MR result as a stable prior, these errors were completely corrected. The lower part of Figure 15 shows (e) the probability distributions for three classes (forest, water, and impervious) in the three results. In the VHR mapping result, the region marked by the left black box exhibits abnormally high probability for water and correspondingly low probability for forest (the correct class) due to imaging inconsistencies, leading to its misclassification as water. In contrast, the MR probability for this erroneous class is significantly lower in this region, and its distribution is smoother and more stable. After processing by the decision-level Bayesian fusion, the fusion result shows an obvious increase in forest probability in the correct areas, while the probability for the misclassified classes in the error-prone regions is effectively suppressed, thereby achieving error correction.

The mechanism of the decision-level Bayesian fusion lies in the reconstruction and harmonization of probability distributions, which is grounded in the Bayesian theory for optimal information fusion under uncertainty [79]. The VHR semantic segmentation model can capture fine spatial structures, but its predictions are susceptible to local illumination conditions, shadow effects, and textural variations. This sensitivity often yields a probability distribution with a marked contrast between areas of high and low probability. In comparison, MR remote sensing mapping, based on temporally stable spectral features and SAR data, exhibits stronger spectral consistency and robustness against disturbances at the regional scale, leading to a more balanced and stable probability distribution. Our proposal introduces the MR mapping result as a prior constraint at the probability level. This compresses the VHR posterior probability into a more stable range, specifically by moderately adjusting extreme probability values from the VHR result, leading to more cautious class determinations. This mechanism can, to a certain extent, correct misclassifications caused by VHR image heterogeneity, enhancing overall spatial consistency while preserving the ability to express spatial detail.

One limitation of the decision-level Bayesian fusion is that it can only correct errors within certain limits. Nevertheless, based on our test results, the decision-level Bayesian fusion effectively mitigates systematic biases in VHR mapping, such as those caused by clouds, shadows, anomalous illumination and image mosaicking heterogeneity. Furthermore, it delivers more stable classification outcomes, particularly in complex transitional zones and along boundaries. This proves the practical advantage of the adaptive fusion

strategy selection mechanism, which is driven by statistic and spatial consistency tests. It preserves the fine-grained features of VHR results when the two maps are consistent and utilizes decision-level Bayesian fusion to integrate the stable prior knowledge from MR data with VHR data when significant inconsistency exists, thereby offering a novel perspective for land cover mapping with multi-source remote sensing data.

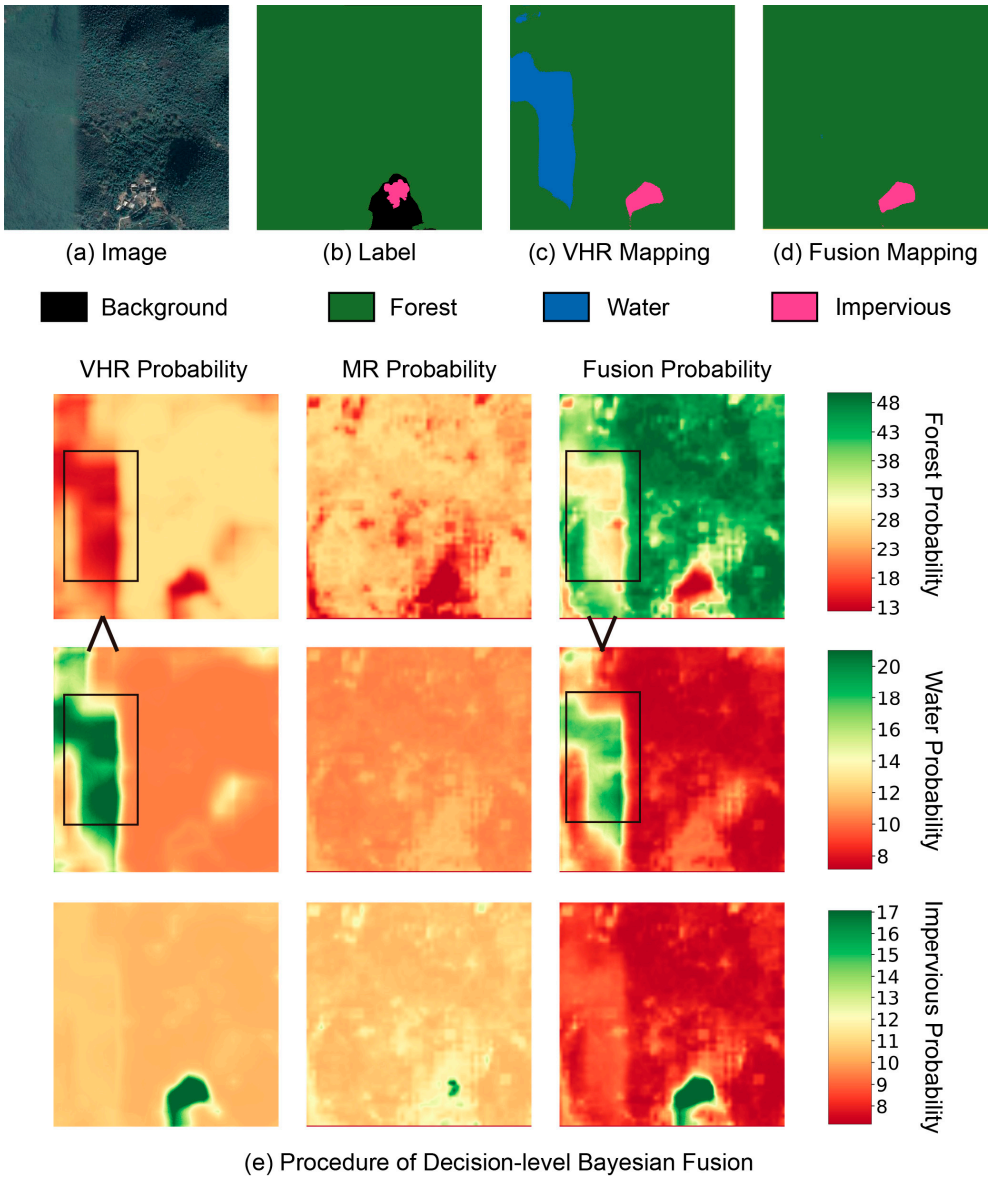


Figure 15. Detailed illustration of the decision-level Bayesian fusion. Upper panel: (**a**) original VHR image; (**b**) ground truth; (**c**) VHR-only mapping; (**d**) fusion mapping. Lower panel: (**e**) class probability distributions (forest, water, and impervious) for VHR probability, MR probability, and fusion probability. The black boxes highlight regions where VHR mapping errors are corrected through fusion with stable MR probability.

5.2. Impact, Limitation and Future Work

The mapping framework proposed in this study, which integrates remote sensing foundation models with multi-source data Bayesian fusion, achieved high OA and mIoU in the highly urbanized and complex landscape of the PRD. This underscores the framework's robustness in navigating the extreme spatial heterogeneity and fragmented land cover patches characteristic of the PRD. Compared to baseline semantic segmentation methods relying merely on VHR imagery, this approach demonstrates more stable mapping results

with superior spatial consistency. While maintaining high precision, its key contribution lies in significantly enhanced suppression of textural heterogeneity and systematic noise.

From a methodological perspective, this study makes three key contributions to VHR land cover mapping under limited annotation. First, the proposed SDMAE pretraining framework mitigates value heterogeneity and preserves boundary features through VMask Strategy and Edge-Enhanced Loss, offering a foundation model explicitly optimized for land cover mapping tasks. Second, the SBFNet explicitly models the inherent "pixel–object–scene" semantic hierarchy to enable effective multi-scale feature fusion. Third, the decision-level Bayesian fusion integrates multi-source data via adaptive, distribution-driven consistency tests, enhancing the robustness of land cover mapping.

The decision-level Bayesian fusion is not confined to multi-source remote sensing data fusion but represents a general probabilistic modeling paradigm, analogous to Bayesian approaches used in limited-sensor fusion [80] and spatiotemporal Bayesian modeling for analyzing spatially structured relationships [81]. Any data source that can be expressed probabilistically can be integrated. That is, existing knowledge or experience can serve as prior information and newly acquired observations as posterior evidence, and the model can be iteratively optimized through our proposal. This idea can be generalized to various scenarios, such as temporal data updating, multi-sensor collaborative observation, and cross-scale information fusion, providing a sustainable technique for uncertainty quantification and knowledge accumulation in remote sensing information extraction.

Beyond the case study of the PRD, the framework proposed in this study offers transferable insights for VHR land cover mapping in other complex regions facing similar challenges. The emphasis on leveraging MR data as stable priors provides a practical path for improving VHR mapping without a proportional increase in high-quality annotations at the target resolution. Consequently, the effectiveness of the method should be evaluated not only by the magnitude of accuracy gain but also by its capacity to maintain mapping stability under data-constrained and high-noise environments.

Despite the aforementioned progress, several limitations remain in this study. A critical dependency of our framework is the thematic accuracy of the MR prior, as the Bayesian fusion requires a reliable prior probabilistic baseline. The scale of the pretraining dataset and the annotated segmentation samples remains limited by data availability and annotation costs but could be expanded in future work. Additionally, the framework was only compared against the MAE foundation model, as benchmarking against other outstanding foundation models was not feasible due to the lack of publicly available implementations. The mapping pipeline involves multi-stage modeling and multi-source fusion, resulting in high computational complexity. Furthermore, the cross-regional transferability and robustness of the proposed framework should be systematically tested in more regions.

Future work can proceed in the following directions. The decision-level Bayesian fusion framework can be extended by incorporating more rigorous prior constraints and multi-temporal observations to reduce reliance on a single prior source. Furthermore, it is worth exploring how historical mapping results or time-series data can serve as dynamic priors to enable continuous land cover monitoring and adaptive updating. Additionally, the generalization capability of the method should be systematically evaluated across more regions, and integration with emerging vision–language foundation models can be explored to further reduce reliance on task-specific annotations.

6. Conclusions

This study conducted systematic research on VHR land cover mapping in the PRD. We constructed a series of region-specific datasets and proposed a VHR land cover map-

ping framework that combines remote sensing foundation models with multi-source data Bayesian fusion. The main conclusions are as follows:

(1) Using Google Earth VHR optical imagery, we built PRD262K, a large-scale pretraining dataset for remote sensing foundation models in the PRD. Furthermore, we established PRDLC-PRO, a semantic segmentation dataset containing 33,342 high-quality annotated samples. Together with a newly developed point sample set, these resources establish a multi-scale, complementary annotation system that underpins fine-scale land cover mapping in the PRD.

(2) We proposed a VHR mapping framework that combines remote sensing foundation models with scene-based feature decoding. By training a remote sensing foundation model using the SDMAE for the PRD and integrating it with the specifically designed semantic segmentation network SBFNet, the framework significantly improved the accuracy and consistency of VHR land cover mapping.

(3) Based on statistic and spatial consistency tests and Bayesian probability modeling, we introduced a decision-level Bayesian fusion framework. By incorporating the stability of MR remote sensing as prior knowledge to guide the optimization of VHR posterior results, this approach effectively mitigates classification errors in VHR imagery. It provides a novel methodological perspective for multi-source remote sensing data fusion in land cover mapping.

Author Contributions: Conceptualization, J.L.; methodology, J.L.; software, J.L. and Y.Z. (Yikai Zhao); validation, J.L. and Y.Z. (Yikai Zhao); formal analysis, J.L. and Y.Z. (Yikai Zhao); investigation, J.L. and Y.Z. (Yikai Zhao); resources, Y.Z. (Yan Zhou); data curation, J.L., Y.Z. (Yikai Zhao) and M.X.; writing—original draft, J.L. and Y.Z. (Yikai Zhao); writing—review and editing, J.L., Y.Z. (Yikai Zhao), M.X., J.Z. and Y.Z. (Yan Zhou); visualization, J.L. and Y.Z. (Yikai Zhao); supervision, Y.Z. (Yan Zhou) and X.L.; project administration, Y.Z. (Yan Zhou) and X.L.; funding acquisition, Y.Z. (Yan Zhou) and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant No. 42571408, the National Natural Science Foundation of China under Grant No. 42501550 and the Natural Science Foundation of Guangdong Province under Grant No. 2025A1515010093.

Data Availability Statement: Our code is available at https://github.com/JeasunLok/SDMAE-SBFNet (accessed on 21 January 2026). Our datasets (PRD262K and PRDLC-PRO) and land cover point sample set are available at https://doi.org/10.5281/zenodo.18301135.

Acknowledgments: Thanks to the National Natural Science Foundation of China under Grant No. 42571408, the National Natural Science Foundation of China under Grant No. 42501550 and the Natural Science Foundation of Guangdong Province under Grant No. 2025A1515010093 for funding this work. Special thanks to Youcheng Xu for improving the logic of the original draft.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Detailed Land Cover Classification System of PRDLC

To ensure the objectivity and reproducibility of the image annotation process, we established a rigorous interpretation protocol for the eight-class system used in this study. Our classification criteria and interpretation symbols are strictly aligned with the Globe230k dataset [73] and the national standard of China (GB/T 21010-2017 [82]), which were then adapted to the 1.19 m VHR imagery. The identification of each category relies on a combination of spectral and spatial keys, including color, shape, texture, and site association. For instance, forest and shrubland are distinguished not only by their spectral signatures but also by their distinctive "popcorn-like" grainy textures and shadow heights in high-resolution imagery. Table A1 provides the formal definition and specific visual keys for each class to facilitate the replication of our methodology.

Table A1. Definition and interpretation symbols for land cover classification system of PRDLC.

Class	Definition	Interpretation Symbols
Cropland	Areas for crop production	Color: Bright green to yellow-brown; Shape: Regular patches with clear boundaries; Texture: Striated or uniform row patterns.
Forest	Land with >10% tree canopy cover	Color: Dark green; Shape: Irregular clusters; Texture: Coarse, "popcorn-like" grainy appearance caused by tree crowns.
Grass	Herbaceous vegetation, mainly grass	Color: Light or yellowish-green; Shape: Large patches or strips; Texture: Fine, smooth, and uniform and lacks the grainy texture of trees.
Shrubland	Perennial woody plants with low canopy closure.	Color: Dull green; Shape: Scattered or clumped; Texture: Rougher than grass but smoother than forest and lower height shadows.
Wetland	Areas seasonally or permanently flooded	Color: Dark green to brownish-black; Site: Adjacent to water or low-lying; Texture: Mottled, mixed with water and vegetation.
Water	Natural or artificial water bodies (rivers, ponds)	Color: Dark blue to black (clear) or cyan (turbid); Shape: Linear or smooth polygons; Texture: Very smooth.
Impervious	Man-made surfaces (buildings, roads, runways)	Color: Gray, white, blue or terracotta; Shape: Geometric, sharp edges, distinct shadows; Texture: Smooth but structured.
Bare	Land with <10% vegetation (soil, sand, rock)	Color: Brown, light gray, or white; Shape: Irregular; Texture: Earthy or grainy and often shows signs of human activity (e.g., construction).

Appendix B.

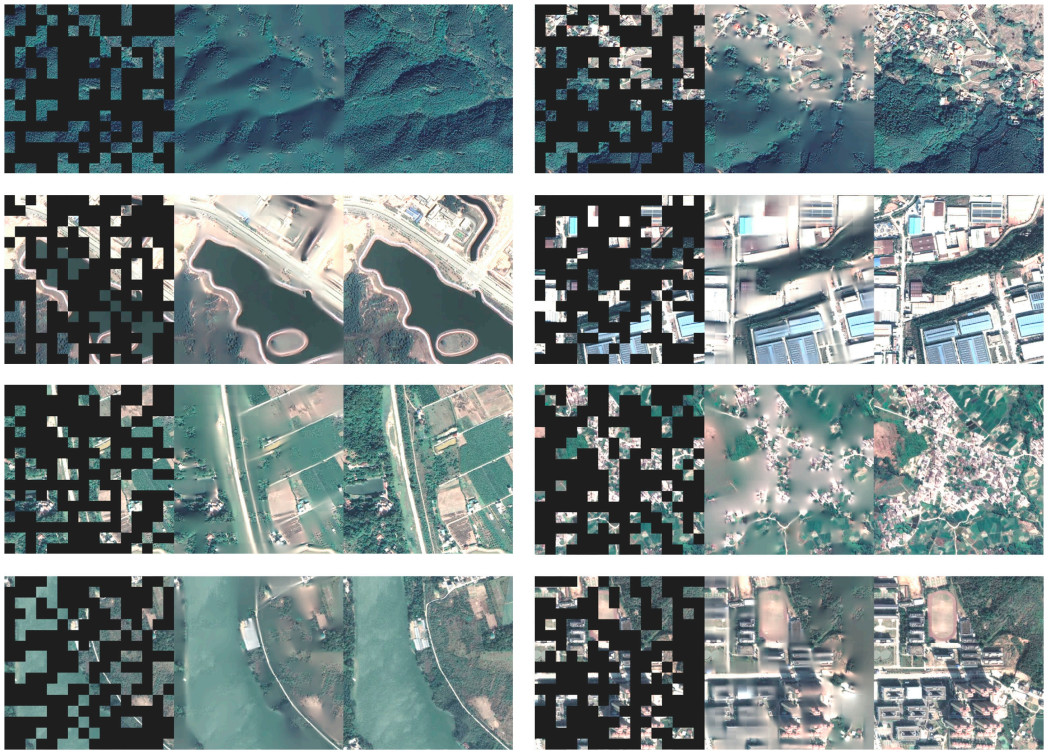


Figure A1. Examples of visualized results for the MAE. For each batch, (left to right), there is the masked image, reconstructed image and original image, respectively.

Table A2. Rules for constructing PRDLC-PRO including dataset contributions and label conversion.

PRDLC		Globe230k			OneEarthMap			LoveDA			WHDLD			DLRSD		
Class	**Label**	**Class**	**Original Label**	**Converted Label**	**Class**	**Original Label**	**Converted Label**	**Class**	**Original Label**	**Converted Label**	**Class**	**Original Label**	**Converted Label**	**Class**	**Original Label**	**Converted Label**
Background	0	Background	0	0	Background	0	0	Background	1	0	Building	1	7	Airplane	1	0
Cropland	1	Cropland	1	1	Bareland	1	8	Building	2	7	Road	2	7	Bare soil	2	8
Forest	2	Forest	2	2	Rangeland	2	3	Road	3	7	Pavement	3	7	Buildings	3	7
Grassland	3	Grass	3	3	Developed space	3	7	Water	4	6	Vegetation	4	0	Cars	4	0
Shrubland	4	Shrub	4	4	Road	4	7	Barren	5	8	Bare soil	5	8	Chaparral	5	4
Wetland	5	Wetland	5	5	Tree	5	2	Forest	6	2	Water	6	6	Court	6	7
Water	6	Water	6	6	Water	6	6	Agriculture	7	1				Dock	7	7
Impervious	7	Tundra	7	0	Agriculture land	7	1							Field	8	1
Bare	8	Impervious surface	8	7	Building	8	7							Grass	9	3
		Bareland	9	8										Mobile home	10	0
		Ice/snow	10	0										Pavement	11	7
														Sand	12	8
														Sea	13	6
														Ship	14	0
														Tanks	15	0
														Trees	16	2
														Water	17	6
Sample contribution (number of images)																
1–7370		7370–8810			8811–9538			9539–26,302			26,303–31,242			31,243–33,342		

Table A3. Composition and label conversion details of the MR sample set.

PRDLC-PRO		SinoLC-1		CLCD		Dynamic World		Esri 10		GLC_FCS10		GLWD V2		GWL_FCS30	
Class	**ID**	**Class**	**ID**	**Class**	**ID**	**Class**	**ID**	**Class**	**ID**	**Class**	**ID**	**Class**	**ID**	**Class**	**ID**
Cropland	1	Traffic route	1	Cropland	1	Water	1	Water areas	1	Cropland	11–20	Open water bodies	1–7	Permanent water	180
Forest	2	Tree cover	2	Forest	2	Trees	2	Trees	2	Forest	51–92	Lacustrine wetlands	8–9	Swamp	181
Grassland	3	Shrubland	3	Shrub	3	Grass	3	Flooded vegeta-tion areas	4	Shrubland	121–122	Riverine wetlands	10–19	Marsh	182
Shrubland	4	Grassland	4	Grassland	4	Flooded vegeta-tion	4	Crops	5	Grassland	130	Peatlands	22–27	Flooded flat	183
Wetland	5	Cropland	5	Water	5	Crops	5	Built areas	7	Tumdra	140	Coastal wetlands	28–31	Saline	184
Water	6	Building	6	Snow/Ice	6	Shrub&scrub	6	Bare ground areas	8	Wetland	181–187	Special wetlands	32–33	Mangrove forest	185
Imprevious	7	Barren and sparse vegetation	7	Barren	7	Built area	7	Snow/ice	9	Imprevious surfaces	191–192			Salt marsh	186
Bare	8	Snow and ice	8	Impervious	8	Bare ground	8	Clouds	10	Bare areas	150–200			Tidal flat	187
		Water	9	Wetland	9	Snow/ice	9	Rangeland areas	11	Water	210				
		Wetand	10							Premanent ice and snow	220				
		Moss and lichen	12												

Table A4. Multi-source MR remote sensing features for the random forest model.

Data Source	Features	Calculation Formula
Sentinel-1	VV, VH	-
	CrossVH	$\frac{VV}{VH}$
	RVI	$\frac{4\times VH}{VV}$
	PoL	$\frac{VH-VV}{VH+VV}$
Sentinel-2	B2–B5, B8–B9, B11–B12	-
	NDVI	$\frac{B8-B4}{B8+B4}$
	EVI	$\frac{2.5\times(B8-B4)}{B8+6\times B4-7.5\times B2+1}$
	SAVI	$\frac{(B8-B4)\times 1.5}{B8+B4+0.5}$
	NDWI	$\frac{B3-B8}{B3+B8}$
	MNDWI	$\frac{B3-B11}{B3+B11}$
	LSWI	$\frac{B8-B11}{B8+B11}$
	NDBI	$\frac{B11-B8}{B11+B8}$
	IBI	$\frac{2\times B11}{B8+B11}$
	NDBSI	$\frac{B4-B2}{B4+B2}$
	BSI	$\frac{(B6+B4)-(B8+B2)}{B6+B4+B8+B2}$
NASADEM	DEM	-
	Slope	-

References

1. Wang, Y.; Sun, Y.; Cao, X.; Wang, Y.; Zhang, W.; Cheng, X. A Review of Regional and Global Scale Land Use/Land Cover (LULC) Mapping Products Generated from Satellite Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 311–334. [CrossRef]
2. Venter, Z.S.; Barton, D.N.; Chakraborty, T.; Simensen, T.; Singh, G. Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. *Remote Sens.* **2022**, *14*, 4101. [CrossRef]
3. Thenkabail, P.S.; Schull, M.; Turral, H. Ganges and Indus River Basin Land Use/Land Cover (LULC) and Irrigated Area Mapping Using Continuous Streams of MODIS Data. *Remote Sens. Environ.* **2005**, *95*, 317–341. [CrossRef]
4. Luo, X.; Zhou, H.; Satriawan, T.W.; Tian, J.; Zhao, R.; Keenan, T.F.; Griffith, D.M.; Sitch, S.; Smith, N.G.; Still, C.J. Mapping the Global Distribution of C4 Vegetation Using Observations and Optimality Theory. *Nat. Commun.* **2024**, *15*, 1219. [CrossRef]
5. He, K.; Fan, C.; Zhong, M.; Cao, F.; Wang, G.; Cao, L. Evaluation of Habitat Suitability for Asian Elephants in Sipsongpanna under Climate Change by Coupling Multi-Source Remote Sensing Products with MaxEnt Model. *Remote Sens.* **2023**, *15*, 1047. [CrossRef]
6. Gunacti, M.C.; Gul, G.O.; Cetinkaya, C.P.; Gul, A.; Barbaros, F. Evaluating Impact of Land Use and Land Cover Change under Climate Change on a Lake System. *Water Resour Manag* **2023**, *37*, 2643–2656. [CrossRef]
7. Cihlar, J. Land Cover Mapping of Large Areas from Satellites: Status and Research Priorities. *Int. J. Remote Sens.* **2000**, *21*, 1093–1114. [CrossRef]
8. Grekousis, G.; Mountrakis, G.; Kavouras, M. An Overview of 21 Global and 43 Regional Land-Cover Mapping Products. *Int. J. Remote Sens.* **2015**, *36*, 5309–5335. [CrossRef]
9. Rogan, J.; Franklin, J.; Stow, D.; Miller, J.; Woodcock, C.; Roberts, D. Mapping Land-Cover Modifications over Large Areas: A Comparison of Machine Learning Algorithms. *Remote Sens. Environ.* **2008**, *112*, 2272–2283. [CrossRef]
10. Nemmour, H.; Chibani, Y. Multiple Support Vector Machines for Land Cover Change Detection: An Application for Mapping Urban Extensions. *ISPRS J. Photogramm. Remote Sens.* **2006**, *61*, 125–133. [CrossRef]
11. Foody, G.M. Land Cover Classification by an Artificial Neural Network with Ancillary Information. *Int. J. Geogr. Inf. Syst.* **1995**, *9*, 527–542. [CrossRef]
12. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]

13. Jin, S.; Yang, L.; Zhu, Z.; Homer, C. A Land Cover Change Detection and Classification Protocol for Updating Alaska NLCD 2001 to 2011. *Remote Sens. Environ.* **2017**, *195*, 44–55. [CrossRef]
14. Liu, L.; Zhang, X.; Gao, Y.; Chen, X.; Shuai, X.; Mi, J. Finer-Resolution Mapping of Global Land Cover: Recent Developments, Consistency Analysis, and Prospects. *J. Remote Sens.* **2021**, *2021*, 5289697. [CrossRef]
15. Vali, A.; Comai, S.; Matteucci, M. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
16. Šćepanović, S.; Antropov, O.; Laurila, P.; Rauste, Y.; Ignatenko, V.; Praks, J. Wide-Area Land Cover Mapping with Sentinel-1 Imagery Using Deep Learning Semantic Segmentation Models. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10357–10374. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
18. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 801–818.
19. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
20. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886.
21. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: https://arxiv.org/abs/1409.1556 (accessed on 21 January 2026).
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
27. Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; Hu, S. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
28. Bartholomé, E.; Belward, A.S. GLC2000: A New Approach to Global Land Cover Mapping from Earth Observation Data. *Int. J. Remote Sens.* **2005**, *26*, 1959–1977. [CrossRef]
29. Buchhorn, M.; Lesiv, M.; Tsendbazar, N.E.; Herold, M.; Bertels, L.; Smets, B. Copernicus Global Land Cover Layers-Collection 2. *Remote Sens.* **2020**, *12*, 1044. [CrossRef]
30. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y.; et al. Stable Classification with Limited Sample: Transferring a 30-m Resolution Sample Set Collected in 2015 to Mapping 10-m Resolution Global Land Cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef] [PubMed]
31. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, near Real-Time Global 10 m Land Use Land Cover Mapping. *Sci. Data* **2022**, *9*, 251. [CrossRef]
32. Li, Z.; He, W.; Cheng, M.; Hu, J.; Yang, G.; Zhang, H. SinoLC-1: The First 1 m Resolution National-Scale Land-Cover Map of China Created with a Deep Learning Framework and Open-Access Data. *Earth Syst. Sci. Data* **2023**, *15*, 4749–4780. [CrossRef]
33. Zhong, Y.; Su, Y.; Wu, S.; Zheng, Z.; Zhao, J.; Ma, A.; Zhu, Q.; Ye, R.; Li, X.; Pellikka, P.; et al. Open-Source Data-Driven Urban Land-Use Mapping Integrating Point-Line-Polygon Semantic Objects: A Case Study of Chinese Cities. *Remote Sens. Environ.* **2020**, *247*, 111838. [CrossRef]
34. Sang, Q.; Zhuang, Y.; Dong, S.; Wang, G.; Chen, H. FRF-Net: Land Cover Classification From Large-Scale VHR Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1057–1061. [CrossRef]

35. Zhang, W.; Tang, P.; Zhao, L. Fast and Accurate Land-Cover Classification on Medium-Resolution Remote-Sensing Images Using Segmentation Models. *Int. J. Remote Sens.* **2021**, *42*, 3277–3301. [CrossRef]
36. Chen, Y.; Huang, J.; Wang, J.; Zhou, Y.; Ge, Y. Deep Spatiotemporal Subpixel Mapping Network by Integrating a Prior Fine Land Cover Map with Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4412214. [CrossRef]
37. Tong, X.-Y.; Xia, G.-S.; Zhu, X.X. Enabling Country-Scale Land Cover Mapping with Meter-Resolution Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 178–196. [CrossRef]
38. Zhang, Y.; Chen, G.; Myint, S.W.; Zhou, Y.; Hay, G.J.; Vukomanovic, J.; Meentemeyer, R.K. UrbanWatch: A 1-Meter Resolution Land Cover and Land Use Database for 22 Major Cities in the United States. *Remote Sens. Environ.* **2022**, *278*, 113106. [CrossRef]
39. Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef]
40. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9726–9735.
41. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
42. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988.
43. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–22. [CrossRef]
44. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4065–4076.
45. Wu, K.; Zhang, Y.; Ru, L.; Dang, B.; Lao, J.; Yu, L.; Luo, J.; Zhu, Z.; Sun, Y.; Zhang, J.; et al. A Semantic-Enhanced Multi-Modal Remote Sensing Foundation Model for Earth Observation. *Nat. Mach. Intell.* **2025**, *7*, 1235–1249. [CrossRef]
46. Gong, Z.; Wei, Z.; Wang, D.; Hu, X.; Ma, X.; Chen, H.; Jia, Y.; Deng, Y.; Ji, Z.; Zhu, X.; et al. CrossEarth: Geospatial Vision Foundation Model for Domain Generalizable Remote Sensing Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**. [CrossRef]
47. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In Proceedings of the Tenth International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
48. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, online, 18–24 July 2021; pp. 8748–8763.
49. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 1–16. [CrossRef]
50. Liu, C.; Huang, W.; Zhu, X.X. LandSegmenter: Towards a Flexible Foundation Model for Land Use and Land Cover Mapping. *arXiv* **2025**, arXiv:2511.08156. [CrossRef]
51. Wang, Y.; Albrecht, C.M.; Zhu, X.X. Self-Supervised Vision Transformers for Joint SAR-Optical Representation Learning. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 139–142.
52. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
53. Wang, D.; Hu, M.; Jin, Y.; Miao, Y.; Yang, J.; Xu, Y.; Qin, X.; Ma, J.; Sun, L.; Li, C.; et al. HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 6427–6444. [CrossRef]
54. Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M.A.; Danish, M.S.; Fraccaro, P.; Watson, C.D.; Klein, L.J.; Khan, F.S.; et al. EarthDial: Turning Multi-Sensory Earth Observations to Interactive Dialogues. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 14303–14313.
55. Liu, C.; Chen, K.; Zhao, R.; Zou, Z.; Shi, Z. Text2Earth: Unlocking Text-Driven Remote Sensing Image Generation with a Global-Scale Dataset and a Foundation Model. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 238–259. [CrossRef]
56. Yao, K.; Xu, N.; Yang, R.; Xu, Y.; Gao, Z.; Kitrungrotsakul, T.; Ren, Y.; Zhang, P.; Wang, J.; Wei, N.; et al. Falcon: A Remote Sensing Vision-Language Foundation Model (Technical Report). *arXiv* **2025**, arXiv:2503.11070.
57. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]

58. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
59. Li, Q.; Qiu, C.; Ma, L.; Schmitt, M.; Zhu, X.X. Mapping the Land Cover of Africa at 10 m Resolution from Multi-Source Remote Sensing Data with Google Earth Engine. *Remote Sens.* **2020**, *12*, 602. [CrossRef]
60. Ienco, D.; Interdonato, R.; Gaetano, R.; Ho Tong Minh, D. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for Land Cover Mapping via a Multi-Source Deep Learning Architecture. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 11–22. [CrossRef]
61. Zhao, J.; Yuan, Z.; Mi, X.; Yang, J.; Chen, X.; Meng, X.; Zhu, H.; Meng, Y.; Jiang, Z.; Zhang, Z. A Cross-Spatiotemporal Weakly Supervised Framework for Land Cover Classification: Generating Temporally and Spatially Consistent Land Cover Maps. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 519–538. [CrossRef]
62. Yin, Z.; Li, X.; Wu, P.; Lu, J.; Ling, F. CSSF: Collaborative Spatial-Spectral Fusion for Generating Fine-Resolution Land Cover Maps from Coarse-Resolution Multi-Spectral Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *226*, 33–53. [CrossRef]
63. Shi, Q.; Pan, T.; Lu, D.; Li, H.; Chai, Z. BPUM: A Bayesian Probabilistic Updating Model Applied to Early Crop Identification. *J. Remote Sens.* **2025**, *5*, 0438. [CrossRef]
64. Luo, J.; Li, J.; Chu, X.; Yang, S.; Tao, L.; Shi, Q. BTCDNet: Bayesian Tile Attention Network for Hyperspectral Image Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 5504205. [CrossRef]
65. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Virtual Event, 6–14 December 2021.
66. Shao, Z.; Yang, K.; Zhou, W. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
67. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
68. Shi, Q.; He, D.; Liu, Z.; Liu, X.; Xue, J. Globe230k: A Benchmark Dense-Pixel Annotation Dataset for Global Land Cover Mapping. *J. Remote Sens.* **2023**, *3*, 0078. [CrossRef]
69. Xia, J.; Yokoya, N.; Adriano, B.; Broni-Bediako, C. OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023.
70. Yang, J.; Huang, X. The 30 m Annual Land Cover Dataset and Its Dynamics in China from 1990 to 2019. *Earth Syst. Sci. Data* **2021**, *13*, 3907–3925. [CrossRef]
71. Karra, K.; Kontgis, C.; Statman-Weil, Z.; Mazzariello, J.C.; Mathis, M.; Brumby, S.P. Global Land Use/Land Cover with Sentinel 2 and Deep Learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4704–4707.
72. Zhang, X.; Liu, L.; Zhao, T.; Zhang, W.; Guan, L.; Bai, M.; Chen, X. GLC_FCS10: A Global 10-m Land-Cover Dataset with a Fine Classification System from Sentinel-1 and Sentinel-2 Time-Series Data in Google Earth Engine. *Earth Syst. Sci. Data* **2025**, *17*, 4039–4062. [CrossRef]
73. Lehner, B.; Anand, M.; Fluet-Chouinard, E.; Tan, F.; Aires, F.; Allen, G.H.; Bousquet, P.; Canadell, J.G.; Davidson, N.; Ding, M.; et al. Mapping the World's Inland Surface Waters: An Upgrade to the Global Lakes and Wetlands Database (GLWD V2). *Earth Syst. Sci. Data* **2025**, *17*, 2277–2329. [CrossRef]
74. Zhang, X.; Liu, L.; Zhao, T.; Chen, X.; Lin, S.; Wang, J.; Mi, J.; Liu, W. GWL_FCS30: A Global 30 m Wetland Map with a Fine Classification System Using Multi-Sourced and Time-Series Remote Sensing Imagery in 2020. *Earth Syst. Sci. Data* **2023**, *15*, 265–293. [CrossRef]
75. Wang, C.; Sun, W. Semantic Guided Large Scale Factor Remote Sensing Image Super-Resolution with Generative Diffusion Prior. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 125–138. [CrossRef]
76. Hao, M.; Chen, S.; Lin, H.; Zhang, H.; Zheng, N. A Prior Knowledge Guided Deep Learning Method for Building Extraction from High-Resolution Remote Sensing Images. *Urban Inf.* **2024**, *3*, 6. [CrossRef]
77. Shao, C.; Li, H.; Shen, H. Generative Shadow Synthesis and Removal for Remote Sensing Images Through Embedding Illumination Models. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5620115. [CrossRef]
78. Liu, Y.; Li, W.; Guan, J.; Zhou, S.; Zhang, Y. Effective Cloud Removal for Remote Sensing Images by an Improved Mean-Reverting Denoising Model with Elucidated Design Space. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 17851–17861.
79. Zhang, R.; Aziz, I.; Houtz, D.A.; Zhao, Y.; Ford, T.W.; Watts, A.C.; Alipour, M. UAV-Based Remote Sensing of Soil Moisture across Diverse Land Covers: Validation and Bayesian Uncertainty Characterization. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–18. [CrossRef]

80. Fu, Y.; Wang, Z.; Maghareh, A.; Dyke, S.; Jahanshahi, M.; Shahriar, A.; Zhang, F. Effective Structural Impact Detection and Localization Using Convolutional Neural Network and Bayesian Information Fusion with Limited Sensors. *Mech. Syst. Signal Process.* **2025**, *224*, 112074. [CrossRef]
81. Rohleder, S.; Costa, D.; Bozorgmehr, P.K. Area-Level Socioeconomic Deprivation, Non-National Residency, and COVID-19 Incidence: A Longitudinal Spatiotemporal Analysis in Germany. *eClinicalMedicine* **2022**, *49*, 101485. [CrossRef]
82. *GB/T 21010-2017*; Land Use Classification Status. Standardization Administration of China: Beijing, China, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.