International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: www.tandfonline.com/journals/tjde20

Object based Markov random field model for hierarchical semantic segmentation of remote sensing imagery

Jun Wang, Chen Zheng, Haoyu Fu, Yili Zhao, Qinling Dai, Xin Huang, Junfeng Xie & Leiguang Wang

To cite this article: Jun Wang, Chen Zheng, Haoyu Fu, Yili Zhao, Qinling Dai, Xin Huang, Junfeng Xie & Leiguang Wang (2025) Object based Markov random field model for hierarchical semantic segmentation of remote sensing imagery, International Journal of Digital Earth, 18:1, 2521795, DOI: 10.1080/17538947.2025.2521795

To link to this article: https://doi.org/10.1080/17538947.2025.2521795

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2025, VOL. 18, NO. 1, 2521795
https://doi.org/10.1080/17538947.2025.2521795

Object based Markov random field model for hierarchical semantic segmentation of remote sensing imagery

Jun Wang [a,b], Chen Zheng[c], Haoyu Fu[a], Yili Zhao[d], Qinling Dai[e], Xin Huang[f], Junfeng Xie[g] and Leiguang Wang [a]

[a]College of Landscape Architecture and Horticulture, Southwest Forestry University, Kunming, People's Republic of China; [b]Yunnan Earthquake Agency, Kunming, People's Republic of China; [c]School of Mathematics and Statistics, Henan University, Kaifeng, People's Republic of China; [d]College of Big Data and Intelligence Engineering, Southwest Forestry University, Kunming, People's Republic of China; [e]College of Art and Design, Southwest Forestry University, Kunming, People's Republic of China; [f]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, People's Republic of China; [g]Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing, People's Republic of China

ABSTRACT

Capturing hierarchical relationships among land-cover classes is crucial for accurate semantic segmentation of remote sensing images. Traditional object-based methods face inherent limitations in modeling these complex relationships. To overcome these limitations, we proposed a novel object-based Markov random field (OMRF) model for hierarchical semantic segmentation. The objective of our model is to address two key challenges: (i) the representation of hierarchical semantic features, and (ii) the edge preservation of segmentation results. To address the first challenge, we developed hierarchical semantic representations of images for two distinct land-cover class sets and incorporated a transition probability matrix into OMRF to capture the interaction between these two semantic layers. For the second challenge, we devised an innovative spatial energy function that effectively enforces hierarchical predictions and dynamically regulates boundary smoothness by evaluating spectral dissimilarities among neighboring objects. Furthermore, a generative cross-layer inference strategy was introduced to iteratively exchange and update information across semantic layers for improved prediction. Experimental results on 11 remote sensing images demonstrate the robustness and accuracy of the proposed method, achieving an average Kappa coefficient exceeding 0.96. In comparison to 15 state-of-the-art methods, our model achieved optimal performance in 9 instances and suboptimal performance in 2 instances.

ARTICLE HISTORY
Received 16 August 2024
Accepted 10 June 2025

KEYWORDS
Remote sensing image; Markov random field (MRF); hierarchical semantic segmentation; spectral dissimilarity

CONTACT Leiguang Wang wlgbain@126.com College of Landscape Architecture and Horticulture, Southwest Forestry University, Kunming 650224, People's Republic of China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17538947.2025.2521795.

1. Introduction

Remote sensing images are essential sources of information required for the development of current human society (Hossain and Chen 2019). Extracting meaningful information from remote sensing imagery remains a pressing challenge (Dronova et al. 2012). In particular, semantic segmentation of remote sensing images plays a vital role in many applications, such as environmental conservation (He et al. 2022), precision agriculture (Duro, Franklin, and Dubé 2012), urban planning (Grinias, Panagiotakis, and Tziritas 2016) and land resource management (Zhang et al. 2019). The gradual improvement in the spatial resolution of remote sensing images in recent years has revealed richer object detail information in high-resolution remote sensing (HRRS) images. Meanwhile, different imaging conditions lead to distortion of images acquired by a sensor, which tends to aggravate the uncertainty and diversity in the semantic segmentation (Feng et al. 2021). These effects lead to diverse land-cover classes often having similar spectral responses within specific regions, but the same land-cover classes have different spectral responses within different regions (X. Zhang et al. 2017). Typically, buildings located in urban and agricultural areas have different spectral responses. However, some farmland and water have similar spectral responses. The semantic segmentation of remote sensing images has been a formidable task to date (Zhao et al. 2017).

To obtain segmentation results closer to the ground truth, a large collection of state-of-the-art methods utilize not only the spectral features but also the spatial features of images (Ghamisi et al. 2018). Classically, L. Zhang et al. (2006) proposed a pixel-based spatial feature, the pixel shape index. This method combines the spectral features of individual pixels with the shape characteristics of their surrounding local regions, thereby achieving better performance than approaches relying solely on spectral features. However, pixel-based segmentation methods have a limited ability to capture macroscopic features of an image, and pixel-based methods usually introduce salt-and-pepper noise into the segmentation map (Chan, Ho, and Nikolova 2005). Therefore, geographic object-based image analysis (GEOBIA) was proposed, which transforms the basic unit of the image from pixels to over-segmented regions (Blaschke 2010; Hay and Castilla 2008). GEOBIA enables richer spatial feature extraction and effectively suppresses the noise typically introduced by pixel-based segmentation (Blaschke et al. 2014). In addition, each over-segmented region contains basic semantic information (Zheng, Zhang, and Wang 2016). For example, if the over-segmented region consists of multiple pixels labeled as crops, it may contain the basic semantic information of farmland. Multiresolution and multiscale techniques were introduced for semantic segmentation to improve the accuracy of the segmentation (Pont-Tuset et al. 2016). Multiscale segmentation expands the model's receptive field by incorporating features at different scales, enabling it to capture more informative spatial context (Bouman and Shapiro 1994; Cheng and Bouman 2001). Here, the method of constructing an image-based pyramid structure is widely applied to semantic segmentation because of its regular structure and complete theory (Z. Li et al. 2018). In addition, the wavelet transform, a method of obtaining feature vectors by stacking multiscale wavelet coefficients, can also improve the segmentation accuracy (Noda, Shirazi, and Kawaguchi 2002). The wavelet-transform-based method provides features of the image at different resolutions, which allows the images to be interpreted from multiple perspectives. Given the rich hierarchical information in HRRS

images, combining pixel-level and object-level features has become a common strategy to enhance semantic interpretation across different scales. Establishing interactions between different granularities allows for a more comprehensive interpretation of land-cover structures from both fine and coarse spatial perspectives. Multigranularity approaches combine the advantages of both granularities, which can capture more image spatial context information (Zheng et al. 2021). As the number of remote sensing data sources continues to grow, the practice of integrating multi-sensor systems to fuse diverse information sources is becoming increasingly prevalent. It has been demonstrated that data obtained through multi-source information fusion (MSIF) techniques exhibit higher reliability than data obtained from a single sensor (X. Li, Dunkin, and Dezert 2023).

In recent years, deep learning methods have shone a light on remote sensing image analysis thanks to their robust feature extraction and learning capabilities (LeCun, Bengio, and Hinton 2015; Ma et al. 2019). C. Zhu et al. (2024) employed a convolutional autoencoder (CAE) model based on deep learning and trained on a simulated time-series InSAR interferogram dataset. The trained model effectively mitigated atmospheric noise in coseismic deformation extraction, resulting in improved accuracy and robustness in displacement detection. Some classical networks, such as fully convolutional neural networks (FCNs) (Long, Shelhamer, and Darrell 2015), U-Net (Ronneberger, Fischer, and Brox 2015), feature pyramid networks (FPNs) (T. Y. Lin et al. 2017), and DeepLab (Chen et al. 2018), have received extensive attention. Representative techniques include atrous convolution and atrous spatial pyramid pooling (ASPP). Among these methods, atrous convolution can expand the receptive field of the model to capture more spatial features, and the ASPP technique effectively extracts multi-scale information from images (Chen et al. 2018). Aside from CNNs, graph neural networks (GNNs) have demonstrated considerable potential in several other fields. Nevertheless, GNNs display a higher propensity for instability than conventional machine learning techniques. In this regard, Z. Huang et al. (2024) addressed the issue of prediction variability in GNNs by proposing the Graph Relearn Network (GRN), which refines predictions of unstable nodes to reduce prediction variance and enhance accuracy. Although neural network-based remote sensing image analysis techniques offer convenient solutions, these techniques are susceptible to adversarial attacks. Z. Zhang et al. (2024) introduced a novel dual-branch sparse self-learning framework with instance binding augmentation and sparse depth wise separable convolution for adversarial detection in remote sensing images to counteract the performance degradation of deep neural network models under adversarial attacks.

Furthermore, statistical segmentation methods are another reliable approach in the field of semantic segmentation of remote sensing images (Masson and Pieczynski 1993). Specifically, land-cover classes usually present a regular distribution in remote sensing images. Interestingly, objects in the images are correlated with each other. The spatial correlation between adjacent objects is stronger than that between non-adjacent ones, and it decreases with increasing spatial distance (Tobler 1970). Statistical segmentation methods can distinguish the different land-cover classes more logically by capturing this statistical regularity. In particular, the Markov random field (MRF) model provides an ideal theoretical framework to integrate the spectral and spatial features of images (Cross and Jain 1983). The MRF framework has been widely employed in semantic segmentation of remote sensing imagery (S. Z. Li, 2009).

Although many semantic segmentation methods have been proposed, they still face unresolved problems. Specifically, GEOBIA solves the problem of salt and pepper noise that appears in pixel-based segmentation methods, but the ensuing consequences may be the loss of detailed information on the image or the over-smoothing of boundaries (Zheng et al. 2021). In addition, the multiresolution and multiscale methods can build multiscale representations on the pixel or object granularity, but they are still linear transformation methods. Unfortunately, these methods cannot effectively capture the hierarchical semantic information contained in HRRS images to guide semantic segmentation. The land-cover classes with high-level semantic information are usually composed of different subclasses. However, the spectral features of these subclasses are highly variable. For example, towns, as a land-cover class with high-level semantic information, comprise a collection of land-cover subclasses with low-level semantic information, such as buildings, roads, and trees. Without corresponding hierarchical semantic information to assist in semantic segmentation, it is difficult to integrate these subclasses into one class with high-level semantic information by spectral features alone. Therefore, the introduction of hierarchical semantic features can further optimize semantic segmentation. For deep learning, many models are implemented using supervised learning, including the aforementioned neural networks U-Net, FPNs, and DeepLabV3+. To obtain a reliable deep learning model, it is usually necessary to provide a large amount of training data and long training times. If the number of training samples is small, it may lead to the overfitting phenomenon (Rice, Wong, and Kolter 2020). To address similar challenges, S. Zhu et al. (2023). presented an innovative model for the remote sensing change detection of forests. The model employs data augmentation utilizing forest fragments generated by deep convolutional generative adversarial networks (DCGANs). However, in most instances, it is difficult to obtain a large number of training samples for model training. While some transferable deep models (Tong et al. 2020) can be adopted, achieving good segmentation results remains challenging due to the high variability among diverse datasets. That greatly restricts the availability of deep learning for semantic segmentation with small sample datasets. In addition, deep learning models are generally regarded as black-box models, which achieve good segmentation results but have low interpretability (Guidotti et al. 2018).

MRF methods can be divided into two groups based on the basic units of modeling: pixel-based MRF and object-based MRF (OMRF) methods. Of these methods, the OMRF methods have attracted widespread attention because they consider a larger-scope spatial context than the pixel-based MRF methods. Most OMRF approaches use a uniform smoothness scheme to model spatial interactions between objects (Pan et al. 2020). While these approaches have improved the accuracy of segmentation results, they do not always provide ideal results at image boundaries. To preserve image detail information, the OMRF model should decrease the smoothing effect in regions with sudden changes in spectral feature levels, like image boundaries. This requires extracting additional features from the image to model the complex spatial interactions between objects.

Based on previous literature, many MRF approaches use complementary information to improve image segmentation beyond spectral features, including image granularity and semantic and spectral dissimilarities. The existing MRF methods can flexibly define an MRF model based on any kind of auxiliary information. Although these methods solve the corresponding problems in semantic segmentation, there is still a

lack of a comprehensive MRF framework that integrates different kinds of auxiliary information. This work presents a unique OMRF approach that employs multi-layer semantic and spectral penalty information (MRF-MSSP) to extract hierarchical semantic features of remote sensing images and preserve boundary information among diverse classes. Over-segmented regions were utilized as the basic unit of the MRF-MSSP model. Based on the rich semantic information contained in HRRS images, we expanded the semantic layers described by the classic OMRF from one to two. To do this, images were first divided into two class layers representing high-level and low-level semantic information, respectively. Then, the transition probability matrix was used to extract and learn the semantic context information of observed images. The iterative update of the transition probability matrix enabled the capture of object interactions across different semantic layers. Furthermore, an object-based spectral dissimilarity function was developed to prevent over-smoothing of the segmentation result and to adaptively regulate the smoothing effect of the model in different regions. The MRF-MSSP model has a relatively low smoothing effect against areas with highly variable spectral features, especially in the boundary areas. It is anticipated that the discrepancies between the delineated boundaries of the various labels in the segmentation map and those established as the reference boundaries in the ground truth will be addressed through the implementation of a dissimilarity function. In the segmentation process of the MRF-MSSP model, the semantic context information was first provided by relying on the transition probability matrix. Subsequently, the spectral dissimilarity between adjacent objects helped us roughly determine the boundary information. Finally, combining both pieces of information to guide the semantic segmentation ultimately led to the segmentation results.

The principal contributions of this work are as follows:

- The Exploitation of HRRS Images: Our method further exploits the spectral, spatial, and hierarchical semantic information contained in HRRS images, thereby improving the accuracy of the segmentation process.
- Hierarchical Semantic Modeling: This work proposes a multilayer label field scheme for modeling hierarchical semantics in HRRS images, with the objective of capturing interactions among high – and low-level land-cover classes.
- The Edge-preserving Mechanism: The incorporation of a spectral dissimilarity-based spatial energy function enables the model to achieve boundary awareness and to regulate the smoothing effect across different regions adaptively, thereby effectively preserving boundary information among diverse classes.
- Cross-Layer Inference: A generative cross-layer inference approach is implemented, which provides for iterative information interchange and updates across hierarchical semantic layers.

The remainder of the article is structured as follows: Section 2 provides a concise overview of pertinent prior research. The specifics and structure of the proposed methodology are outlined in Section 3. Section 4 presents the results of the experimental investigation and a sensitivity analysis of the parameters. Section 5 assesses the efficacy of the MRF-MSSP model, delineating its advantages and constraints. Finally, Section 6 presents the conclusions of our work and suggests avenues for future inquiry.

2. Related work

Previous studies have proven the effectiveness of MRF-based methods on the semantic segmentation of remote sensing images. In this section, we briefly review the background of semantic segmentation in remote sensing and summarize the existing MRF-based methods as follows.

2.1. Image segmentation

Image segmentation is widely applied in many fields, such as computer vision (Felzenszwalb and Huttenlocher 2004), medical image analysis (D. Shen, Wu, and Suk 2017), and remote sensing image interpretation. For remote sensing image interpretation, image segmentation is mainly divided into two groups: basic segmentation and semantic segmentation approaches. Both approaches divide the remote sensing image into several homogeneous regions, each region consisting of many pixels. Mathematically, they are defined as follows (Blaschke et al. 2014). A remote sensing image $\boldsymbol{I}$ is segmented into n regions based on features such as the texture, color, spectrum, and shape.

The basic segmentation methods segment the image into many small regions with high spectral homogeneity. These regions usually contain low-level semantic information and are labeled as fine land-cover classes, such as industrial land, paddy fields, and trees. As an alternative, the basic segmentation approaches include the mean shift (MS) (Comaniciu and Meer 2002; X. Huang and Zhang 2008), watershed (Vincent and Soille 1991), and normalized cuts (Shi and Malik 2000) methods. The homogeneous regions generated by the basic segmentation method are referred to as over-segmented regions in this article. The size and shape of the over-segmented region directly impact the accuracy of the subsequent model for object feature extraction and segmentation. Therefore, the generation of over-segmented regions is the key step in the segmentation process of the MRF-MSSP model.

By contrast, the semantic segmentation approaches segment the image into several large-size homogeneous regions with high amounts of semantic information. These are generally composed of some over-segmented regions with lower amounts of semantic information. Consequently, semantic segmentation approaches usually generate homogeneous regions with large intra-class variations. For remote sensing images, it is a great challenge to tag objects with significant spectral variations into the same class. This is also the main issue addressed in our work.

2.2. Markov random field (MRF) model

The MRF model is a probabilistic graphical model that uses graphs to represent the correlations between variables (Mitchell 1997). Specifically, the MRF model is a type of undirected graph model. A vertex in a probability graph represents a group or one random variable. The potential functions, real functions defined on a subset of variables, are present in the MRF model. Furthermore, potential functions are mainly utilized to define the probability distribution functions and provide a quantitative depiction of the correlation between the vertices of the probability graph.

2.2.1. Classic MRF model

For remote sensing images, the MRF model typically considers the image's spectral and spatial characteristics. The MRF model is composed of two sub-models, the feature model and the label model (Wang et al. 2017). The feature model employs the likelihood function to compute pixel-class conditional probabilities. It is sometimes referred to as a spectral model, as the primary features considered by this model are the spectral characteristics of the image. The label model aims to represent the spatial context relationship among land cover classes in remote sensing images, with potential functions describing this relationship. Figure 1 illustrates the connection between the label and feature models.

Originally, the MRF model was established on a lattice of pixels with a regular spatial context, known as the classic MRF model (Besag 1986). Here, the vertices of the probability graph denote the pixels. For the classic MRF model, the feature model considers solely the spectral features of a single pixel, whereas the label model only captures a small-

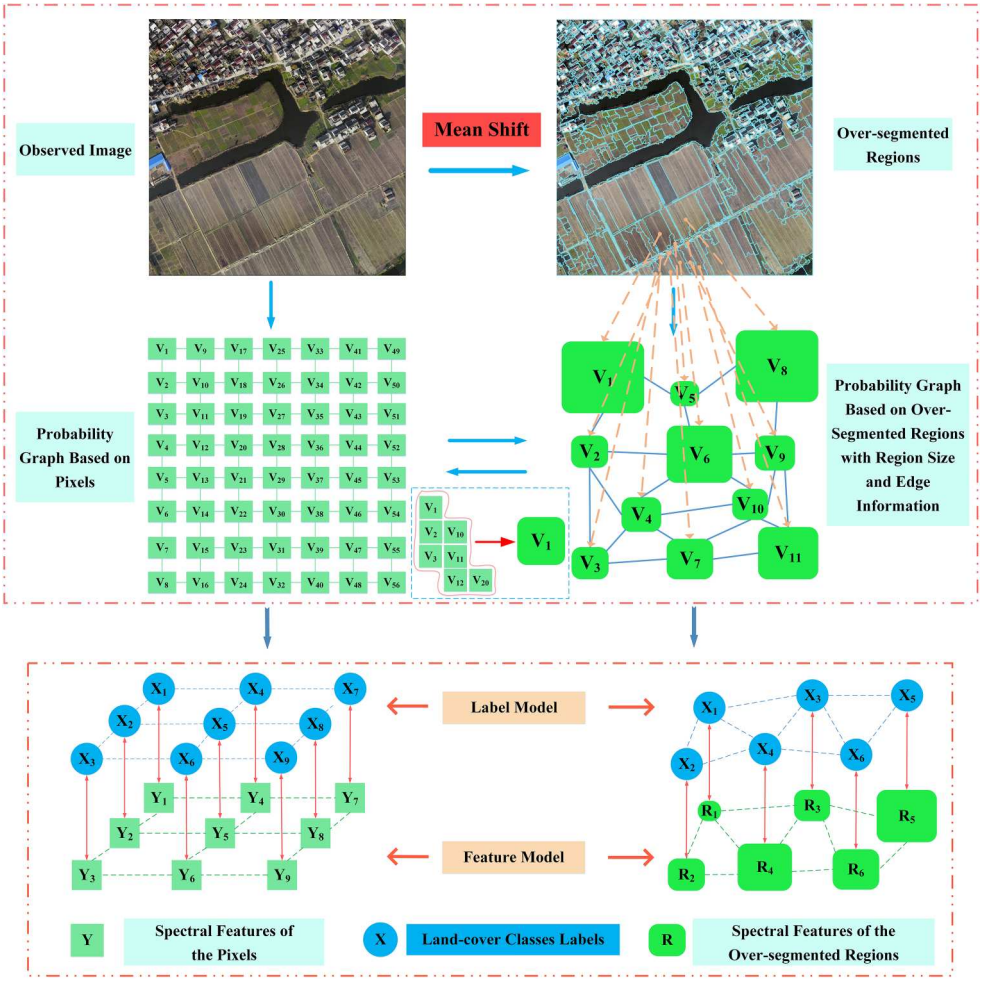

Figure 1. Feature model and label models for object Markov random field (OMRF) and classical pixel-based Markov random field (MRF) methods.

scope spatial context, such as a 4-connected or 8-connected neighborhood of pixels. To expand the receptive field of the model, the multi-scale technique combined with the MRF model (MRMRF) was proposed. For instance, Noda, Shirazi, and Kawaguchi (2002) decomposed the original image into images with a pyramidal structure using the wavelet transform method. Subsequently, the MRF models were defined on the decomposed images. The MRMRF method can capture a more expansive range of spatial contexts than the classical MRF method. However, as the spatial resolution of remote sensing images improves, the numbers of vertices in pixel-based MRF models increase, which adds to the computational burden of the models.

2.2.2. The improved MRF models

The Object-based MRF Model: In the field of remote sensing image analysis, GEOBIA has achieved great success (Blaschke et al. 2014). Thus, the OMRF model was proposed (Xia, He, and Sun 2006). The OMRF model starts by segmenting the image into over-segmented regions using basic segmentation methods. Then, a region adjacency graph (RAG) is constructed according to the spatial distribution of the over-segmented regions. Correspondingly, the vertices of the probability graph represent over-segmented regions. The OMRF methods reduce the computational burden by decreasing the number of vertices in the probabilistic graph (Zheng and Wang 2015). In addition, they can capture more features of the image shape, semantics, and spatial context than the classical MRF method. However, the modeling guidelines of the OMRF and pixel-based MRF methods differ because the over-segmented regions do not have a fixed size, and the shared boundary lengths between neighboring regions vary. Therefore, constructing and finding a solution for the OMRF model is more complex than the pixel-based MRF model. One of the solution approaches for the OMRF model is region growing (Adams and Bischof 1994). The solution to the model is determined by the continuous merging of neighboring vertices in the probability graph that satisfy the requisite growth criteria. For instance, Kuo and Sun merged adjacent regions with similar statistical properties in images generated by the watershed transform (Kuo and Sun 2010). Notably, this error is not rectified if regions are improperly merged during the following solution procedure. Consequently, another approach, generative probabilistic inference, is widely employed to solve OMRF models. For example, Xia, He, and Sun (2006) utilized the multilevel logistic (MLL) structure to characterize the interactions between adjacent vertices and used the maximization of the posterior marginal (MPM) criterion to find the model solution.

(Hierarchy-Agnostic) OMRF Model: To reflect the interactions between adjacent regions more accurately, Zheng and Wang (2015) developed the weighted RAG (WRAG), which takes into account the area of the region and the length information of the shared boundary between adjacent regions, and the OMRF model with regional penalties (OMRF-RP) was defined based on the WRAG. The OMRF-RP model solution was obtained using the MAP criterion. In addition to methods that use models based on pixel or object granularity, some methods attempt to combine the multi-granularity features. The hybrid MRF method with multigranularity features (HMRF-MG) was proposed for the semantic segmentation of remote sensing images (Zheng et al. 2021). The multi-layer structured probability graph was utilized by the HMRF-MG model to characterize the multi-granularity information of the image. In recent years, the performance of methods combining the OMRF and multiscale techniques has also been

impressive. Dai et al. (2020) constructed the regional multi-scale representation of the image by extracting the high-frequency and low-frequency features of the over-segmented image. The results at the final scale combined the feature information on all scales. Although the MRF-based methods mentioned above improved the segmentation accuracy, they considered only one level of semantic information. The interactions between the hierarchical semantics were not further explored.

Hierarchical Semantic Segmentation: Semantic information is one of the most essential features of HRRS images (Tong et al. 2020). In remote sensing images, land-cover classes with high semantic features generally comprise several land-cover classes with low-level semantic features. Moreover, subclasses belonging to the same class are spatially close typically, and these subclasses usually show distinct spectral responses, as illustrated in Figure 2. Solely considering spectral features poses challenges in classifying these over-segmented regions with diverse spectral responses into the same class. Therefore, in order to enhance the accuracy of image interpretation, it is imperative to have semantic information on the HRRS image. In particular, semantic context information between different semantic layers could weaken the adverse impact of intra-class heterogeneity on semantic segmentation. To investigate the semantic context information of images, the OMRF with auxiliary label fields method (OMRF-A) was proposed (Zheng, Zhang, and Wang 2017). OMRF-A utilizes two auxiliary label fields to explore the interactions between different semantic layers of the image. It can interpret remote sensing images

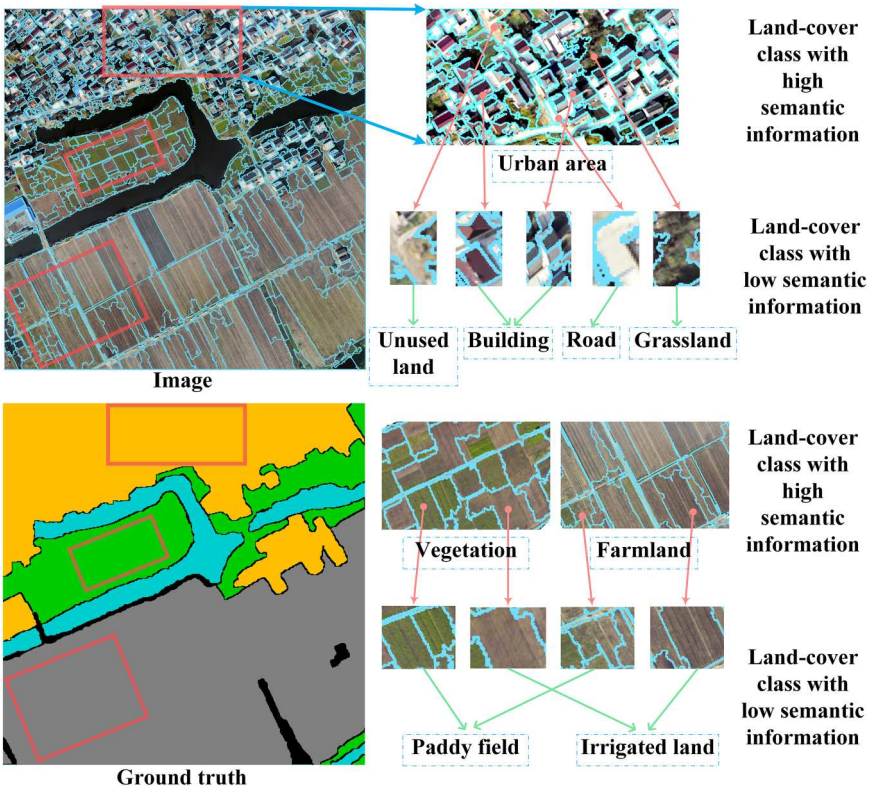

Figure 2. Hierarchical semantic representation of remote sensing image.

from different perspectives by setting a different number of auxiliary classes. Zheng, Zhang, and Wang (2016) developed an OMRF model with two semantic layers (HOMRF) that utilized the transition probability matrix to model the semantic context of images. For deep learning, D. Lin et al. (2018) proposed a multi-scale context intertwining (MSCI) scheme to aggregate features from different scales. The features of the images at multiple scales were connected using long short-term memory (LSTM) chains in a bidirectional cyclic manner. Zhou et al. (2019) designed a multi-scale deep context convolutional network for semantic segmentation that integrated the features of images from different levels of the network. L. Li et al. (2022) presented the Hierarchical Semantic Segmentation Network (HSSN) that is capable of structured scene parsing using taxonomic semantic relations to enable hierarchical semantic segmentation. Besides remote sensing images, medical images also employ analogous methodologies. In a recent publication, Kang et al. (2024) proposed a multi-modality based multi-resolution CNN for 3D-MRI super-resolution reconstruction, fusing features at two scales for enhanced detail recovery in HR T2w images. While existing hierarchical semantic segmentation methods are impressive, they seldom consider the influence of local spectral heterogeneity on the segmentation outcomes. This study aims to develop a framework for integrating hierarchical semantic segmentation with a spectral dissimilarity measure that more accurately reflects the complex variations observed in HRRS images.

Local Spectral Dissimilarity Measure: In capturing the spatial context, the aforementioned MRF models fail to take into account the variability induced by different spectral features between neighboring regions. Consequently, the vertices in these MRF-based approaches are generally treated equally, which does not accurately describe the distinctions across diverse neighborhoods. In this circumstance, segmentation results often fail to reflect the ground truth accurately. To describe the differences in the properties between neighboring objects more accurately, certain approaches that take the spectral heterogeneity into account have been proposed. Y. Shen et al. (2019) designed a local spectral heterogeneity measure for objects that integrates both inter and intra-heterogeneity. In multiscale segmentation algorithms, this method is able to select appropriate scales for different tested images adaptively. Wang et al. (2017) developed a novel MRF model (NED-MRF) to combat the issue of over-smoothing in the segmentation maps generated by the MRF method. The NED-MRF model introduces a spatial adaptive interactive parameter into the potential function to preserve object boundaries when defining quantitative relationships between neighboring pixel pairs. Similarly, X. Zhang et al. (2024) proposed a novel infrared maritime small target detection method, named LDMGGC. The method employs the Wasserstein distance for local dissimilarity measurements to detect suspicious targets. However, neither method utilizes the hierarchical semantic information in remote sensing images to assist in segmentation. In summary, there are advantages and disadvantages to the aforementioned methods. Motivated by the above-mentioned literature, we developed a new MRF-MSSP model to extract hierarchical semantic information and measure spectral dissimilarity between neighboring objects.

3. Proposed method

This section is organized as follows. First, the notation and problem formulation of the classical MRF model are reviewed. Next, the framework of the MRF-MSSP model is

discussed in detail. Finally, the entire algorithm is presented and the hyperparameters of the model are discussed.

3.1. Notation and problem formulation of the MRF model

Given an observed remote sensing image $\boldsymbol{Y}$, let $\boldsymbol{S} = \{s_1, s_2, s_3, \ldots, s_n\}$ be defined as the set of nodes of image $\boldsymbol{Y}$, and the observed image is $\boldsymbol{Y} = \{y_i \mid i \in S\}$. Here, n is the number of nodes of the observed image $\boldsymbol{Y}$, and each node represents a pixel or an over-segmented region. The probability graph $\boldsymbol{G} = (V, E)$ of image $\boldsymbol{Y}$ is defined on the set of nodes $\boldsymbol{S}$, where $V = \{v_i \mid i \in S\}$ and $E = \{e_{i,j} \mid i, j \in S,\ i \neq j\}$. v_i denotes the vertex and $e_{i,j}$ denotes the relationship between neighboring nodes v_i and v_j. If the probability graph $\boldsymbol{G}$ is defined at the pixel granularity, $e_{i,j}$ represents the quantitative relationship between neighboring pixels. For example, in the 8-neighborhood spatial context of the classical MRF model, $e_{i,j}$ denotes the quantitative relationship between the central pixel and the neighboring eight pixels. If the probability graph $\boldsymbol{G}$ is defined at the object granularity, $e_{i,j}$ denotes the length of the boundary shared between two adjacent regions i.e. the number of pixels. The label model $\boldsymbol{X} = \{X_i \mid i \in S\}$ is defined on the probability graph $\boldsymbol{G}$. Each vertex v_i has a land-cover class label X_i, and a random variable X_i of $\boldsymbol{X}$ takes values from the set of land-cover classes $\omega = \{1, 2, 3, \ldots, k\}$, where k is the number of classes. If $\boldsymbol{x} = \{x_i \mid i \in S\}$ denotes a realization of $\boldsymbol{X}$, the MRF model transforms the basic task of semantic segmentation of remote sensing images into finding the best class label $\hat{x}$ of v_i, i.e.

$$\begin{aligned} \hat{x} &= \underset{x \in \aleph}{\operatorname{argmax}} P(\boldsymbol{X} = x \mid \boldsymbol{Y}) \\ &= \underset{x \in \aleph}{\operatorname{argmax}} P(\boldsymbol{Y} \mid \boldsymbol{X} = x) \cdot P(\boldsymbol{X} = x). \end{aligned} \tag{1}$$

Here, $\aleph = \{x\}$ is the set of realizations. According to the Bayesian formula (Mitchell 1997), $P(\boldsymbol{Y})$ has no impact on the final value of $\hat{x}$, so we can obtain the third line of equation (1). In equation (1), the value of the optimal image label $\hat{x}$ is impacted by two parts, the feature model and the label model, as mentioned in Section 2.

The likelihood function $P(\boldsymbol{Y} \mid \boldsymbol{X} = x)$ of the feature model is used to estimate the conditional probability that y_i belongs to class x_i when given the features at position s_i. Each node s_i within the observed image $\boldsymbol{Y}$ is usually assumed to be independent of other nodes in the feature model. Therefore, the likelihood function $P(\boldsymbol{Y} \mid \boldsymbol{X} = x)$ is assumed to obey the naive Bayesian assumption. That is,

$$P(\boldsymbol{Y} \mid \boldsymbol{X} = x) = \prod_{i \in S} P(Y_i \mid X_i = x_i). \tag{2}$$

The objective function $P(\boldsymbol{X} = x)$ of the label model is the joint probability distribution. It is used to capture the spatial relationship between the vertex v_i and the neighborhood v_j in the probability graph $\boldsymbol{G}$. According to Hammersley–Clifford theorem (S. Z. Li 2012), $P(\boldsymbol{X} = x)$ follows a Gibbs distribution. That is,

$$P(\boldsymbol{X} = x) = \frac{1}{Z} \exp(-U(x)), \tag{3}$$

where $Z = \sum_x \exp(-U(x))$ is the normalization constant that makes $P(\boldsymbol{X} = x)$ range from 0 to 1. $U(x) = \sum_{c \in \boldsymbol{C}} \varphi_c(x)$ is the energy function that is the sum of all of the pairwise clique potentials $\varphi_c(x)$ in the label model. In particular, the pairwise clique potentials $\varphi_c(x)$ are employed to capture the interaction between adjacent vertices constituting the pairwise cliques $\boldsymbol{C}$ with the Markov property. Based on the Markov property, we have

$$P(\boldsymbol{X} = x_i | x_j) = P(\boldsymbol{X} = x_i | x_j, \quad x_j \in N_i), \tag{4}$$

where N_i denotes the set of vertices adjacent to vertex v_i in space. The Markov property indicates that the label x_i of vertex v_i in the probability graph is only affected by the label x_j of the vertex v_j adjacent to it, whereas the non-adjacent vertices are independent of each other. According to equations (3) and (4), $P(\boldsymbol{X} = x)$ can be rewritten as

$$P(\boldsymbol{X} = x) = \frac{1}{Z} \exp \sum_i \sum_{j \in N_i} (-\varphi(x_i, \ x_j)). \tag{5}$$

In the classical MRF model, the pairwise clique potentials $\varphi(x_i, \ x_j)$ are equal to

$$\varphi(x_i, \ x_j) \begin{cases} -\beta & \text{if } x_i = x_j \\ \beta & \text{otherwise} \end{cases}, \tag{6}$$

where β is regarded as a smoothing parameter. To facilitate the solution of the two model objective functions, the negative logarithmic is applied to $P(\boldsymbol{Y} \mid \boldsymbol{X})$ and $P(\boldsymbol{X})$. Equation (1) becomes

$$\begin{aligned} \hat{x} &= \underset{x \in \aleph}{\operatorname{argmin}} \ -\ln P(\boldsymbol{Y} \mid \boldsymbol{X} = x) - \ln P(\boldsymbol{X} = x) \\ &= \underset{x_i \epsilon \omega, \ i \in S}{\operatorname{argmin}} \sum_i \left(-P(Y_i \mid X_i = x_i) + \sum_{j \in N_i} -\varphi(x_i, \ x_j) \right). \end{aligned} \tag{7}$$

3.2. Hierarchically semantic layers of MRF-MSSP model

In HRRS images, the MRF model with the pixel as the basic unit has difficulty modeling the complex spatial interactions. As a result, we use the OMRF as the fundamental structure of the MRF-MSSP model. In this article, the mean shift (MS) algorithm is used to provide over-segmented regions. Principally, the over-segmented regions generated by the MS algorithm have more homogeneity and more precise boundaries. In addition, the most crucial aspect is that the MS controls the size of the generated minimum region by setting the parameter. Numerous studies have proven the effectiveness of the MS algorithm (Dai et al. 2020; X. Huang and Zhang 2008). Effective semantic information contributes to semantic segmentation, as mentioned in Section 2. In the proposed MRF-MSSP model framework, we investigate the interactions between two semantic layers of the image. Specifically, the MS algorithm is utilized to generate the set of over-segmented regions $\boldsymbol{R} = \{R_1, R_2, R_3, \ \ldots, R_n\}$ of the image $\boldsymbol{Y}$. Then, the RAG is built based on $\boldsymbol{R}$. Let $\boldsymbol{N} = \{1, 2, 3, \ \ldots, n\}$ denote the set of the number of regions in the RAG. The MRF-MSSP model is defined on the RAG, where each vertex $v_l | l \in \boldsymbol{N}$ denotes an over-segmented region R_l. Since the vertices have only one class label $\boldsymbol{X} = \{x_l\}$ in the label model of the classical OMRF, it has a limited capacity to extract

semantic information from images. Hence, we extend the labels of the vertices in the label model to two sets $\boldsymbol{X} = \{X_l\}$ and $\boldsymbol{X}^1 = \{X_l^1\}$. Here, $\boldsymbol{X}$ is the taken value from the class set $\omega = \{1, 2, 3, \ldots, k\}$, and $\boldsymbol{X}^1$ is the taken value from the class set $\omega^1 = \{1, 2, 3, \ldots, k_1\}$. Correspondingly, each vertex v_l in the label model will obtain two land-cover class labels. The two class labels $\boldsymbol{X}$ and $\boldsymbol{X}^1$ constitute the hierarchical semantic representation of the image, as illustrated in Figure 3.

In particular, the values of the two labels are interactive. If $x^1 = \{x_l^1\}$ is a realization of $\boldsymbol{X}^1$, the optimal label $\hat{x}$ is determined by x^1 and x. Since it is NP-hard to derive the MAP solution of the joint probability function for the two labels directly, this work adopts the strategy of iteratively updating the two labels to obtain the MAP solution of the objective function. One of the labels is optimized by utilizing another label throughout the iterative approach (refer to Algorithm I for specifics). When $\boldsymbol{X}^1 = \{x_l^1\}$ is provided, $\hat{x}$ of the label model becomes

$$\begin{aligned} \hat{x} &= \underset{x \in \aleph}{\operatorname{argmax}} P(\boldsymbol{X} = x | \boldsymbol{Y}, \boldsymbol{X}^1 = x^1) \\ &= \underset{x \in \aleph}{\operatorname{argmax}} P(\boldsymbol{Y} | \boldsymbol{X} = x) \cdot P(\boldsymbol{X} = x | \boldsymbol{X}^1 = x^1). \end{aligned} \tag{8}$$

Correspondingly, when $\boldsymbol{X} = \{x_l\}$ is provided, $\widehat{x^1}$ becomes

$$\begin{aligned} \widehat{x^1} &= \underset{x^1 \in \aleph^1}{\operatorname{argmax}} P(\boldsymbol{X}^1 = x^1 | \boldsymbol{Y}, \boldsymbol{X} = x) \\ &= \underset{x^1 \in \aleph^1}{\operatorname{argmax}} P(\boldsymbol{Y} | \boldsymbol{X}^1 = x^1) \cdot P(\boldsymbol{X}^1 = x^1 | \boldsymbol{X} = x). \end{aligned} \tag{9}$$

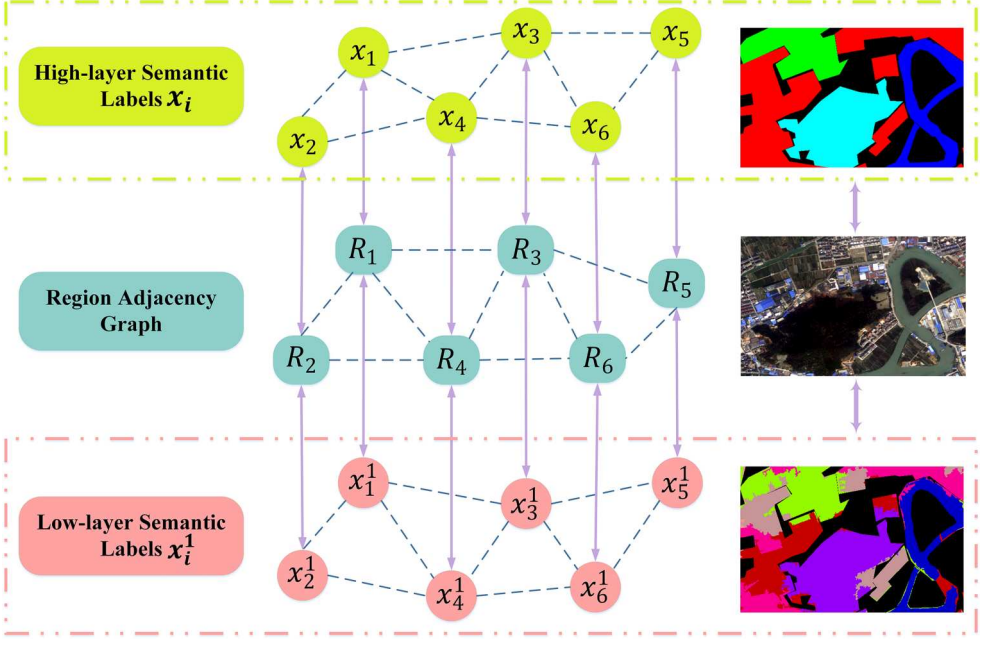

Figure 3. Demonstration of the two labels in the label model.

Here, $\aleph^1 = \{x^1\}$ is the set of realizations. In equations (8) and (9), $P(\boldsymbol{X} = x|\boldsymbol{X}^1 = x^1)$ and $P(\boldsymbol{X}^1 = x^1|\boldsymbol{X} = x)$ are the joint probability distributions of the label models for $\boldsymbol{X}$ and $\boldsymbol{X}^1$, respectively. Notably, the spatial context interaction types in the label models have been expanded from the previous intra-class interactions to inter-class and intra-class interactions. The Markov property of the MRF-MSSP model is assumed to be

$$\begin{aligned} P(X_l|X_j,\ X_j \in \boldsymbol{X}^1 \cup \boldsymbol{X}/\{X_l\}) &= P(X_l|X_j,\ j \in N_l \cup \{X_l^1\}) \\ &= \prod_{l \in N} P(x_l|x_j,\ j \in N_l \cup \{x_l^1\}). \end{aligned} \tag{10}$$

The alteration of the Markov property leads to a change in the way that the energy function $U(x)$ captures the interactions of pairwise clique potentials $\varphi(x_i,\ x_j)$ in the label model. The energy function $U(x)$ captures not only the intra-layer interactions of pairwise cliques but also the inter-layer interactions of pairwise cliques. $U(x)$ is updated to

$$\begin{aligned} U(x_l|x_j,\ j \in N_l \cup \{x_l^1\}) &= U_{\text{intra}}(x_l|x_j, j \in N_l) + U_{\text{inter}}(x_l|x_j, j \in \{x_l^1\}) \\ &= \sum_{j \in N_l} [\beta \cdot \varphi_1(x_l, x_j) + \varphi_2(x_l, x_l^1)], \end{aligned} \tag{11}$$

where $\varphi_1(x_l, x_j)$ represents the interaction between the label x_l of vertex v_l and the label x_j of vertex v_j at the intra-layer, $\varphi_2(x_l, x_l^1)$ represents the interactions between the label x_l and the bel x_l^1 of vertex v_l at the inter-layer, and β is a smoothing parameter. The larger the value of β, the stronger the effect of the label model on the segmentation results, and vice versa. Here, the intra-layer pairwise clique potential $\varphi_1(x_l, x_j)$. becomes

$$\varphi_1(x_l, x_j) = \begin{cases} e_{lj}, & \text{if } x_l = x_j \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

The e_{lj} denotes the length of the shared boundary between regions R_l and R_j. The value of e_{lj} reflects the correlation strength between the two regions. Based on the literature Zheng, Zhang, and Wang (2017), the inter-layer pairwise clique potential is analytically expressed as follows:

$$\varphi_2(x_l, x_l^1) = \sqrt{\frac{R \times C}{n}} \cdot \frac{\sum_{R_j \in R,\ x_j = x_l,\ x_j^1 = x_l^1} |R_j|}{\sum_{R_j \in R,\ x_j = x_l} |R_j|}. \tag{13}$$

In equation (13), the R and C denote the rows and columns of the observed image $\boldsymbol{Y}$, respectively, and n is the number of over-segmented regions. $\sqrt{((R \times C)/n)}$ is a constant to balance the strength of $\varphi_2(x_l, x_l^1)$. $|R_j|$ is the number of pixels composing the over-segmented region Rj. The negative logarithm was applied to the objective function for the label model. Consequently, according to equations (10) and (11), $P(\boldsymbol{X} = x|\boldsymbol{X}^1 = x^1)$ is rewritten as

$$\begin{aligned} &P(\boldsymbol{X} = x|\boldsymbol{X}^1 = x^1) = -\ln P(\boldsymbol{X} = x|\boldsymbol{X}^1 = x^1) \\ &= \sum_{l \in N} \left[\sum_{j \in N_l} [\beta \cdot (-\varphi_1(x_l, x_j)) + (-\varphi_2(x_l, x_l^1))] \right]. \end{aligned} \tag{14}$$

Similarly, $P(\boldsymbol{X}^1 = x^1|\boldsymbol{X} = x)$ becomes

$$\begin{aligned} P(\boldsymbol{X}^1 = x^1|\boldsymbol{X} = x) &= -\ln P(\boldsymbol{X}^1 = x^1|\boldsymbol{X} = x) \\ &= \sum_{l\in N}\left[\sum_{j\in N_l}[\beta\cdot(-\varphi_1(x_l^1, x_j^1)) + (-\varphi_2(x_l^1, x_l))]\right]. \end{aligned} \tag{15}$$

This paper presents two approaches for solving the likelihood function of the feature model. The likelihood function $P(\boldsymbol{Y}|\boldsymbol{X} = x)$ is determined by obtaining the class probabilities of each region through a probabilistic support vector machine (SVM). That is,

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X} = x) &= \prod_{l\in N}\mathrm{SVM}_{R_l,x_l} \\ &= -\sum_{l\in N}\ln(\mathrm{SVM}_{R_l,x_l}). \end{aligned} \tag{16}$$

In equation (16), the second equation was found by applying the negative logarithm of the first. SVM_{R_l,x_l} denotes the probability that region R_l belongs to a given label class x_l. For the likelihood function $P(\boldsymbol{Y}|\boldsymbol{X}^1 = x_l^1)$, it is assumed that each vertex follows a Gaussian distribution based on the conditional independence assumption. $P(\boldsymbol{Y}|\boldsymbol{X}^1 = x_l^1)$ becomes

$$\begin{aligned} P(Y_{R_l}|x_l^1) = \prod_{o_l\in\boldsymbol{Y}} P(o_l|x_l^1 = q) = \prod_{o_l\in\boldsymbol{Y}} &\frac{1}{(2\pi)^{D/2}\cdot\det|\Sigma_q|^{1/2}} \\ &\times\exp\left(-\frac{1}{2}(o_l-\mu_q)^{\mathrm{T}}\Sigma_q^{-1}(o_l-\mu_q)\right). \end{aligned} \tag{17}$$

The negative logarithmic operation is applied to equation (17), and it is rewritten as

$$\begin{aligned} P(Y_{R_l}|x_l^1) = \sum_{o_l\in\boldsymbol{Y}} -\ln[P(o_l|x_l^1 = q)] &= -\frac{n\cdot D}{2}\ln 2\pi \\ &+\sum_{o_l\in Y} -\frac{1}{2}\ln(\det|\Sigma_q|) - \frac{1}{2}(o_l-\mu_q)^{\mathrm{T}}\Sigma_q^{-1}(o_l-\mu_q), \end{aligned} \tag{18}$$

where o_l denotes the average spectrum value of region R_l, D is the spectral dimension of image $\boldsymbol{Y}$, and Σ_q and μ_q are the mean and variance of the Gaussian distribution, respectively, which can be automatically estimated using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977).

3.3. Spectral dissimilarity representation of MRF-MSSP model

Most MRF methods use the same criterion to determine the potential energy $\varphi_1(x_l, x_j)$ between adjacent vertices. They do not account for the variability between vertices when capturing interactions between vertices. Equation (12) indicates that the model solely considers the border information e_{lj} between vertices. Modeling the relationships between vertices more accurately requires the inclusion of additional information to describe the intensity of the interactions. Consequently, a method for the adaptive

adjustment of the smoothness is presented with the goal of preserving boundary information and enhancing the separability between classes. Specifically, we calculate the spectral dissimilarity ψ_{diff} between adjacent regions and then integrate this information into the label model, as illustrated in Figure 4.

Correspondingly, $\varphi_1(x_l, x_j)$ is updated as follows:

$$\varphi_1(x_l, x_j) = \begin{cases} e_{lj} \cdot \text{diff}(S_{R_l,R_j}), & \text{if } x_l = x_j \\ 0, & \text{otherwise} \end{cases}, \tag{19}$$

The $\text{diff}(S_{R_l,R_j})$ is a function representing the spectral dissimilarity between R_l and R_j. It is monotonically decreasing with respect to S_{R_l,R_j}. The definition of $\text{diff}(S_{R_l,R_j})$ is as follows:

$$\text{diff}(S_{R_l,R_j}) = \exp(-|S_{R_l,R_j}|), \tag{20}$$

the S_{R_l,R_j} denotes the value of the spectral dissimilarity. In contrast to pixels, the spectral features of the regions are highly complicated and diverse, particularly in urban areas. It is difficult to quantify the spectral dissimilarity between regions. In this work, a novel approach was created to measure the spectral dissimilarity between regions. S_{R_l,R_j} is defined as follows:

$$S_{R_l,R_j} = \frac{\sum_{d=1}^{D} \left(\left| \frac{a_{ld} - a_{jd}}{a_{ld} + a_{jd}} \right| \right)}{D} \quad \text{with } l \in N,\ j \in N_l, \tag{21}$$

where S_{R_l,R_j} ranges from 0 to 1. The value of S_{R_l,R_j} (dependent variable) is positively correlated with the spectral variability (independent variable) of R_l and R_j. In the abovementioned function, a_{ld} and a_{jd} are the spectral average values of the over-segmented regions R_l and R_j in the d-band, respectively. Notably, R_j is the neighborhood of R_l. a_{ld} and a_{jd}

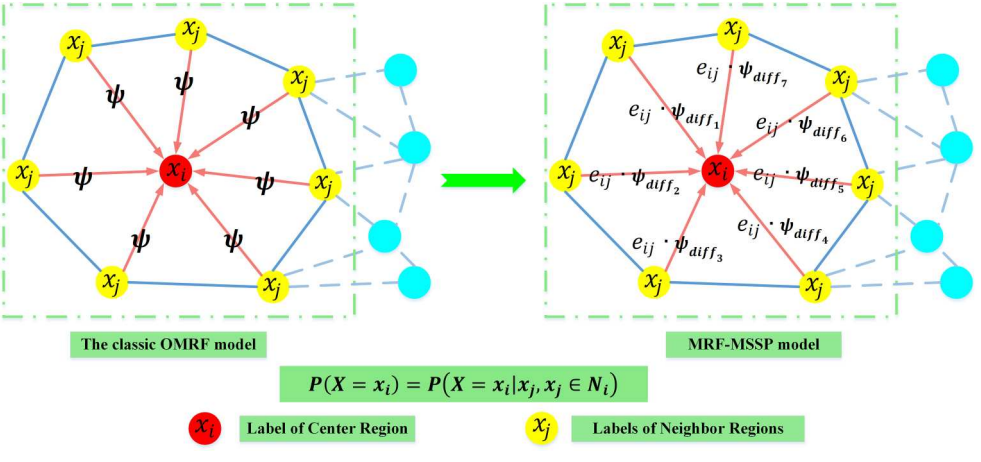

Figure 4. Demonstrate the distinctions between the classic OMRF model and the proposed MRF-MSSP model in reflecting the interactions between various regions.

can be obtained as follows:

$$a_{ld} = \frac{\sum_{z=1}^{Q_l} g_{zd}}{Q_l} \quad \text{and} \quad a_{jd} = \frac{\sum_{z=1}^{Q_j} h_{zd}}{Q_j}. \tag{22}$$

The Q_l and Q_j are the number of pixels composing the regions R_l and R_j, respectively, and g_{zd} and h_{zd} denote the spectral values of the pixels in the regions R_l and R_j in the d-band, respectively.

$$\begin{aligned}\hat{x} &= \underset{x\in\aleph}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}=x)P(\boldsymbol{X}=x|\boldsymbol{X}^1=x^1) \\ &= \underset{x\in\aleph}{\operatorname{argmin}} \{-\ln P(\boldsymbol{Y}|\boldsymbol{X}=x) - \ln P(\boldsymbol{X}=x|\boldsymbol{X}^1=x^1)\} \\ &= \underset{x\in\aleph}{\operatorname{argmin}} \left\{ \begin{array}{c} -\sum\limits_{l\in N} \ln(\mathrm{SVM}_{R_l,x_l}) \\ +\sum\limits_{l\in N}\left[\sum\limits_{j\in N_l}[\beta\cdot(-\varphi_1(x_l,x_j)) + (-\varphi_2(x_l,x_l^1))]\right] \end{array} \right\},\end{aligned} \tag{23}$$

In summary, according to the abovementioned equation, equation (8) is rewritten as equations (23) and (9) is rewritten as

$$\begin{aligned}\widehat{x^1} &= \underset{x^1\in\aleph^1}{\operatorname{argmax}} P(Y|\boldsymbol{X}^1=x^1)P(\boldsymbol{X}^1=x^1|\boldsymbol{X}=x) \\ &= \underset{x\in\aleph^1}{\operatorname{argmin}} \{-\ln P(\boldsymbol{Y}|\boldsymbol{X}^1=x^1) - \ln P(\boldsymbol{X}=x|\boldsymbol{X}^1=x^1)\} \\ &= \underset{x^1\in\aleph^1}{\operatorname{argmin}} \left\{ \begin{array}{c} =\sum\limits_{o_l\in Y} -\frac{1}{2}\ln(|\Sigma_q|) - \frac{1}{2}(o_l-\mu_q)^{\mathrm{T}}\Sigma_q^{-1}(o_l-\mu_q) \\ \sum\limits_{l\in N}\left[\sum\limits_{j\in N_l}[\beta\cdot(-\varphi_1(x_l^1,x_j^1)) + (-\varphi_2(x_l^1,x_l))]\right] \end{array} \right\}.\end{aligned} \tag{24}$$

3.4. Parameters of MRF-MSSP model

Before obtaining the final segmentation results, certain parameters of the MRF-MSSP model need to be set. The means and variances of the Gaussian distributions in the likelihood function (17) can be estimated automatically, respectively, as follows:

$$\mu_q = \frac{\sum_{R_l\in\boldsymbol{R},x_l^1=q}|R_l|\cdot o_l}{\sum_{R_l\in\boldsymbol{R},x_l^1=q}|R_l|}, \tag{25}$$

$$\Sigma_q = \frac{\sum_{R_l\in\boldsymbol{R},x_l^1=q}\sum_{f\in R_l}(o_f-\mu_q)^{\mathrm{T}}(o_f-\mu_q)}{\sum_{R_l\in\boldsymbol{R},x_l^1=q}|R_l|}. \tag{26}$$

The $|R_l|$ denotes the number of pixels in region R_l, and o_l is average spectral value of region R_l.

In addition to the parameters that can be estimated automatically, the remaining three parameters in the MRF-MSSP model need to be set empirically. These are k_1, β, and MRA.

The segmentation results depend heavily on how the class numbers k and k_1 of sets ω and ω^1 are set, respectively. The k value, which is the number of classes of set ω, is the same as number of ground truth classes. The value of k_1 is taken in two circumstances when the number of classes k is provided. If $k_1 > k$, $\boldsymbol{X}^1$ will attempt to collect detailed information of images from a low-level-semantic perspective. If $k_1 < k$, $\boldsymbol{X}^1$ will generate homogeneous regions with large sizes, aiming to acquire the macroscopic spatial context structure of the image. Notably, some datasets comprise two distinct semantic layers of class annotation. As an illustration, Tong et al. (2020) constructed a Gaofen Image Dataset (GID) employing the Gaofen-2 (GF-2) satellite images. The annotated images in the GID consist of two parts. One is a large-scale classification dataset containing 150 images in five categories. The other part consists of 10 images annotated into five main categories and 15 fine land-cover categories. For a dataset such as this one, it is not necessary to additionally set the k_1 value. Unfortunately, most of the tested images have only one ground truth label. Therefore, k_1 generally needs to be set empirically.

Many MRF-based methods require the setting of the β value. These methods are often relatively robust. The smoothing parameter β is used to regulate the energy function $U(x)$ contribution in the MRF model. A large β will prompt the model to generate homogeneous regions with large sizes. Conversely, a small β will provide more detailed information in the segmentation result.

In this work, the MS is used to provide the over-segmented region. As mentioned in Section 2, the size and quality of the over-segmented region have a significant impact on the model's ability to capture the spatial context of the image. Therefore, the parameter *MRA*, which regulates the generation of the minimum over-segmented region $R_{\min}$ area in the MS algorithm, is another critical parameter that influences the segmentation performance. An increase in the value of *MRA* will result in a reduction in the number of vertices present in the RAG, thereby alleviating the computational burden on the model. In general, a larger value of *MRA* will encompass a broader range of spatial context. Conversely, lower values of *MRA* will lead to the preservation of more accurate object boundaries. The effect of different k_1, β, and *MRA* values on the MRF-MSSP model will be tested through additional experiments. The experimental results and analysis will be presented in Section 4.

3.5. Overall algorithm

The framework of the MRF-MSSP model has been described in detail above. In this subsection, the entire process of the algorithm for obtaining the optimal solution to the model is presented. Figure 5 illustrates the workflow of the algorithm, and the complete process is outlined in the following algorithm.

Algorithm of MRF-MSSP model

Input: Observed image $\boldsymbol{Y}$, potential parameter β, k_1, and *MRA*.
Output: Multi-layer segmentation results of MRF-MSSP model.

1. Use the MS (Comaniciu and Meer 2002) provided by EDISON (http://www.wisdom.weizma-nn.ac.il/bagon/MATLAB.html) to obtain the over-segmented region set $\boldsymbol{R} = \{R_1, R_2, R_3, \ldots, R_n\}$, and define the RAG on $\boldsymbol{R}$.
2. Obtain the classification results with k classes and the initial class probability $\mathrm{SVM}^0_{R_i,x_l}$ using the SVM classifier. Then, obtain the classification results with k^1 classes by the classical MRF method.
3. Initialize the label models $\boldsymbol{X}^{(0)} = \{x_l^{(0)} | l \in \{1, 2, 3, \ldots, n\}, \ x_l^0 \in \omega\}$ and $\boldsymbol{X}^{1(0)} = \{x_l^{1(0)} | x_l^{1(0)} \in \omega^1\}$ according to the SVM classifier and classical MRF results.
4. Set $t = 0$.
5. Update $\boldsymbol{X}_l^{1(t+1)}$ based on $\boldsymbol{X}_l^{1(t)}$ and $\boldsymbol{X}_l^{(t)}$.

5.1 Estimate the mean $\mu_q^{(t)}$ and variance $\Sigma_q^{(t)}$ of the feature model for label $\boldsymbol{X}_l^{1(t)}$ using the EM algorithm. Then, obtain the likelihood function $P(Y_{R_l}|x_l^{1(t)}=q)$ based on equation (17).

5.2 For $\boldsymbol{X}_l^{1(t)}$ of the label model, calculate the energy function $U(x_l^{1(t)}|x_j^{1(t)},\ j\in N_l\cup\{x_l^{(t)}\})$ based on equations (11), (13), and (19).

5.3 Sequentially update $\boldsymbol{X}_l^{1(t)}=\{x_l^{1(t)}|x_l^{1(t)}\in\omega^1\}$ to $\boldsymbol{X}_l^{1(t+1)}=\{x_l^{1(t+1)}|x_l^{1(t+1)}\in\omega^1\}$ according to equation (24).

6. Update $\boldsymbol{X}_l^{(t+1)}$ based on $\boldsymbol{X}_l^{(t)}$ and $\boldsymbol{X}_l^{1(t+1)}$.

6.1 Utilize the class probability $\mathrm{SVM}_{R_l,x_l}^{(t)}$ provided by the SVM classifier to solve the likelihood function according to equation (16).

6.2 Update the energy function $U(x_l^{(t)}|x_j^{(t)},\ j\in N_l\cup\{x_l^{1(t+1)}\})$ in the label model $\boldsymbol{X}_l^{(t)}$ in accordance with $\boldsymbol{X}_l^{1(t+1)}$ obtained in step (5).

6.3 Sequentially update $\boldsymbol{X}_l^{(t)}=\{x_l^{1(t)}\}$ to $\boldsymbol{X}_l^{(t+1)}=\{x_l^{(t+1)}\}$ according to equation (23).

7. If $\boldsymbol{X}^{(t+1)}\neq\boldsymbol{X}^{(t)}$ or $\boldsymbol{X}^{1(t+1)}\neq\boldsymbol{X}^{1(t)}$, set $t=t+1$ and go to step 5; otherwise, output $\boldsymbol{X}^{(t+1)}$ and $\boldsymbol{X}^{1(t+1)}$.

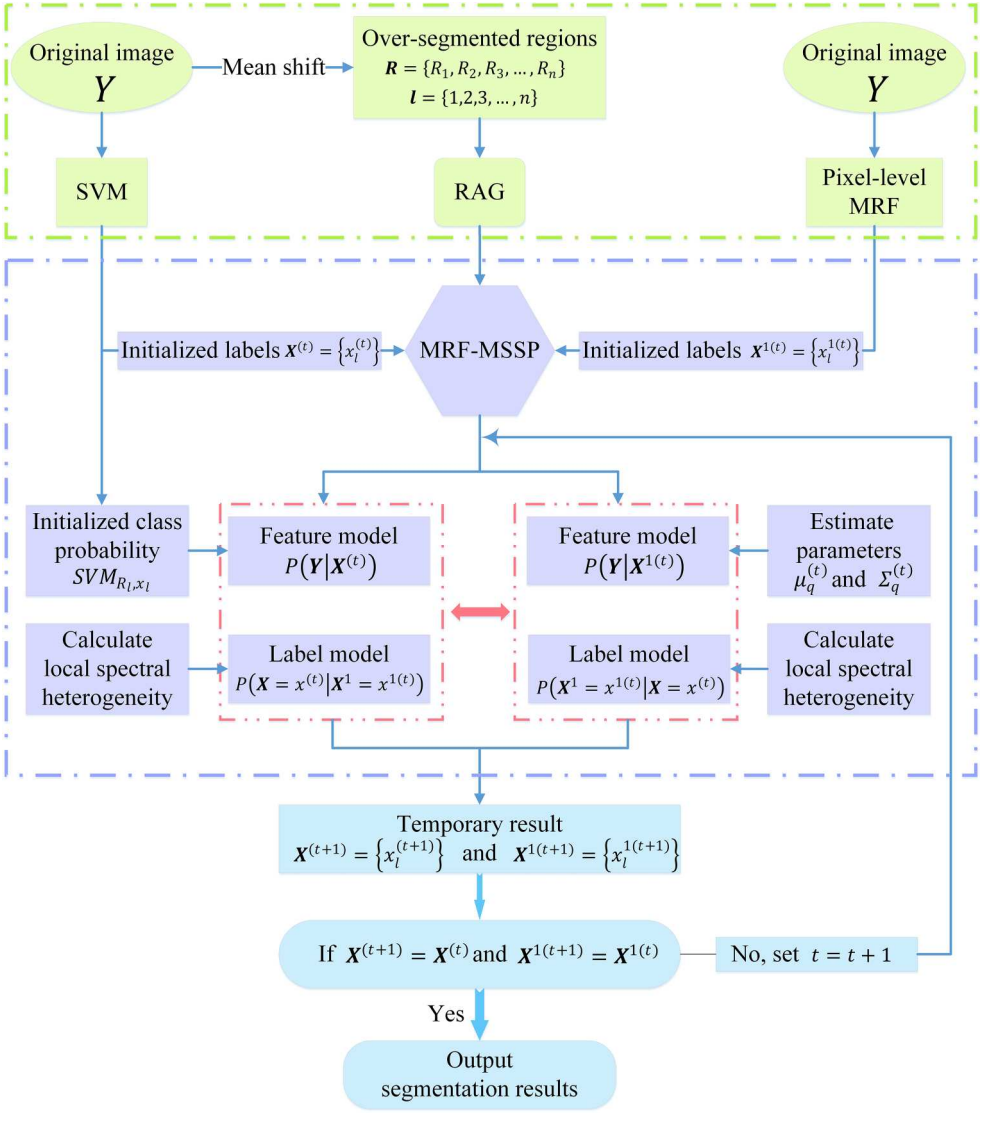

Figure 5. Workflow of the proposed method.

4. Experimental results

To demonstrate the reliability of the proposed MRF-MSSP model for semantic segmentation, experiments were conducted on six datasets, and the results were evaluated using kappa coefficients and overall accuracy (OA). In addition, 15 competing methodologies were utilized for comparison The data sets used for the segmentation experiments are presented in Section 4.1. Section 4.2 reports the comparison methods. The experimental results are presented and analyzed in Section 4.3. Section 4.4 tests the effect of the parameters on the MRF-MSSP model. The computational complexity of the model is analyzed in Section 4.5.

4.1. Data

In this work, six distinct types of tested data were employed to evaluate the dependability of the proposed method. Table 1 reports the properties of the tested data.

Texture images I and II were generated by the Prague texture segmentation data generator (Mikeš and Haindl 2022) to evaluate the effectiveness of the MRF-MSSP model against texture data. To examine the efficacy of the MRF-MSSP model on images with varying resolutions, we selected a 10-m SPOT5 image with a medium-high spatial resolution and a portion of a 1.65-m Washington DC mall image with a high spatial resolution. In addition, a subset of the GID images was selected for experimentation to validate the reliability of the MRF-MSSP model against images with large sizes and high spatial resolutions. The sixth tested image was an aerial image with a 0.4-m spatial resolution in Taizhou, China. The remote sensing images introduced above were composed of blue, green, and red spectral bands. To examine the feasibility of our approach comprehensively, the Salinas image, a hyperspectral remote sensing image, was used to test the robustness of the MRF-MSSP against the hyperspectral image. The Salinas image was obtained over an agricultural area by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS). After discarding 20 water absorption bands, a total of 204 bands were available for interpretation. The spectral range was from 0.4 to 2.5 μm. The Salinas images included 16 classes and 54,129 reference pixels. The Salinas images contain hyperspectral data and are widely employed by academics to evaluate the efficacy of segmentation algorithms.

4.2. Comparison methods

The proposed method was compared with other competing methods, which were broadly categorized into two groups: statistical learning methods and deep learning-based methods. Details of the MRF-based competing methods are provided below.

Table 1. Tested data sets.

Name	Sensor name	Bands	Spatial resolution	Size
Texture image I	/	/	/	512 × 512
Texture image II	/	/	/	512 × 512
Washington DC	GeoEye	3	1.65-m	730 × 770
SPOT5 image	HRG	3	10-m	428 × 428
Gaofen-2 image	Gaofen-2(MS)	3	4-m	2048 × 2048
Aerial image	/	3	0.4-m	1024 × 1024
Salinas	AVIRIS	224	3.7-m	512 × 217

- Iterated conditional mode (ICM) (Besag 1986): This is a classical pixel-based MRF model that employs the iterated conditional mode (ICM) to solve the objective function.
- Support vector machine (SVM) (Chang and Lin 2011): This is a supervised statistical learning method that is less influenced by the probability distribution of the data.
- Object-based MRF model (OMRF) (Xia, He, and Sun 2006): This is a classical object-based MRF model. It adopts a generated probabilistic inference approach to obtain segmentation results.
- Object-based MRF model with regional penalties (OMRF-RP) (Zheng and Wang 2015): This is an improved OMRF model that introduces the area relationship of adjacent regions as weights into the potential function.
- Hybrid object-based Markov random field model (HOMRF) (Zheng, Zhang, and Wang 2016): This is a method that integrates the semantic features of two label fields for hierarchical semantic segmentation.
- Hybrid MRF model with multigranularity information (HMRF-MG) (Zheng et al. 2021): The HMRF-MG method constructs hybrid probability graphs integrating two types of granularity information
- Fully connected conditional random fields (CRFs) with Gaussian edge potentials (Krähenbühl and Koltun 2011).
- Markov random field integrating spectral dissimilarity and class co-occurrence dependency (NED-MRF) (Wang et al. 2017): NED-MRF is a pixel-based method that introduces the local spectral dissimilarity into the spatial energy function.

Among the abovementioned eight comparison methods, ICM, OMRF, OMRF-RP, HOMRF, and HMRF-MG can obtain segmentation results without prior information. These models assume that tested data follow a Gaussian distribution. The class conditional probabilities of the tested images are solved by calculating the mean μ and variance Σ of the Gaussian distribution, similar to equation (17). However, initial labels and class conditional probabilities provided by SVM or other classifiers are necessary for the fully connected CRFs and NED-MRF.

To compare the performance of the MRF-MSSP model to those of the other models in a fair manner, all the object-based MRF models used over-segmented regions generated by the MS, and all of the methods that required prior information were started with the same initial SVM outputs. Finally, all the parameters of these models were set to optimal. In the experimental setup strategy, the differences between the single-layer and multilayer semantic models were discussed by comparing the results of the ICM, OMRF, OMRF-RP, HMRF-MG, and MRF-MSSP. The differences between the proposed MRF-MSSP and pixel-based methods can be seen by comparing the performances of the MRF-MSSP and NED-MRF with the same prior information.

To evaluate the performance of the MRF-MSSP model more comprehensively, we employed seven popular deep-learning methods beyond the MRF-based methods outlined above:

- U-net: convolutional networks for biomedical image segmentation (U-net) (Ronneberger, Fischer, and Brox 2015).
- Feature pyramid network (FPN) (T. Y. Lin et al. 2017).

- Rethinking Atrous Convolution for Semantic Image Segmentation (DeepLabV3+) (Chen et al. 2018).
- CGGLNet: Semantic Segmentation Network for Remote Sensing Images Based on Category-Guided Global-Local Feature Interaction (Ni et al. 2024).
- CMLFormer: CNN and Multi-scale Local-context Transformer network for remote sensing images semantic segmentation (Wu et al. 2024).
- CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation (Wu et al. 2023).
- SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation (Yang, Yuan, and Li 2024).

The proposed method was further validated by comparing it with the seven convolutional neural networks (CNNs) mentioned earlier. It should be noted that the seven CNN-based methods require substantial amounts of training data. Consequently, the performance of both the CNN and MRF-MSSP models was evaluated solely on the GID dataset.

4.3. Segmentation experiment

In this section, the experimental results are presented in two parts. First, a comparative analysis of the experimental outcomes of the MRF-MSSP model is conducted against those of other statistical learning methods. Then, the performance of MRF-MSSP is compared with deep-learning-based approaches on the GID dataset.

4.3.1. Comparison with statistical learning methods

Segmentation experiments were conducted on seven images using nine methods. Since HRRS images typically contain a wealth of texture information, segmentation experiments were first performed on two synthetic texture images.

The texture image I consisted of six different flower textures, each containing numerous flowers and some shadow areas. The different textures appeared visually similar, as shown in Figure 6(a). To facilitate quantitative analysis of the segmentation results, class codes were defined for each texture, as illustrated in Figure 6(a,b). From a visual interpretation perspective, Textures 2 and 4 appeared extremely similar. In contrast, Texture 3 exhibited greater intra-class heterogeneity, with significant variation in the appearance of green leaves and white flowers. Additionally, each texture contained shadow areas, which significantly impacted the models' ability to extract image features. Figure 6 shows the segmentation results of the nine methods. Among them, the ICM method performed the worst, as it only considered the spatial context interactions of 8-neighborhood pixels. From Figure 6, it is evident that all unsupervised methods (ICM, OMRF, OMRF-RP, HOMRF, and HMRF-MG) not only misclassified the shadow areas as a single class but also mistakenly classified most of the Texture 2 area as Texture 4. Although the SVM correctly recognized Textures 2 and 4, there was a lot of salt-and-pepper noise in the results because it was a pixel-based method and it did not consider spatial neighborhood relationships. Fully connected CRFs and NED-MRF took spatial context relationships into account, so they reduced the noise in the segmentation results. However, these two methods still failed to distinguish the different classes

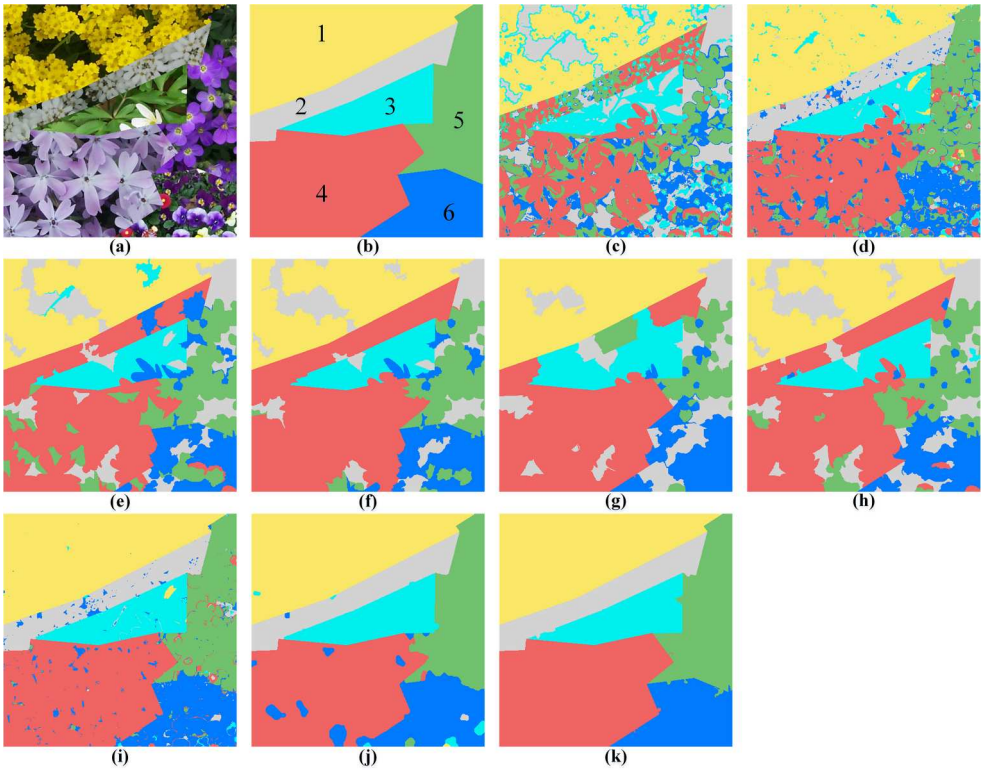

Figure 6. Segmentation results for the texture image I. (a) texture image I. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

of textures precisely, because the intra-class heterogeneity and inter-class homogeneity of the six textures were large, e.g. Texture 6 consisted of purple flowers, green leaves, shadows, red flowers, and white flowers. Compared with other methods, the hierarchical semantic representation constructed by the MRF-MSSP model could extract image features from different semantic scales to further distinguish similar texture areas, and it could adaptively adjust the smoothness to preserve the boundary information, as shown in Figure 6(k). Table 2 presents the quantitative indices OA and kappa for evaluating the segmentation results of texture image I, which demonstrated the promising properties of the MRF-MSSP model.

Texture image II contained 11 diverse textures, as shown in Figure 7(a,b). First, there was a large amount of misclassification in the ICM results. Since Textures 1, 3, 6, and 8 appeared similar, the SVM failed to distinguish these textures effectively by relying on appearance features alone, as shown in Figure 7(d). The OMRF method substantially reduced the salt-and-pepper noise in the segmentation result and improved the segmentation accuracy. The OMRF-RP model further optimized the segmentation results by calculating the area relationship between adjacent over-segmented regions, but it still failed to recognize diverse textures correctly. The HOMRF approach considered the semantic context of the image through the interactions between the two semantic layers, while the HMRF-MG method captured information at both granularities to assist in the

Table 2. Quantitative indexes of ICM (Besag 1986), SVM (Chang and Lin 2011), OMRF (Xia, He, and Sun 2006), OMRF-RP (Zheng and Wang 2015), HOMRF (Zheng, Zhang, and Wang 2016), HMRF-MG (Zheng et al. 2021), Fully connected CRFs (Krähenbühl and Koltun 2011), MRF-NED (Wang et al. 2017) and MRF-MSSP for experiments in two texture images of Figures 6 and 7.

Test image	Indexes (%)	ICM	SVM	OMRF	OMRF-RP	HOMRF	MRF-MG	FCRFs	NED-MRF	MRF-MSSP
Texture image I	Kappa	48.84	82.64	62.28	68.05	69.38	63.98	94.45	95.86	99.59
	OA	52.90	85.40	66.72	72.53	72.62	68.32	98.51	96.66	99.67
Texture image II	Kappa	49.58	67.12	60.65	73.96	72.62	74.69	96.42	98.94	99.75
	OA	52.59	69.76	63.14	76.21	75.09	77.03	96.81	99.06	99.78

segmentation. Interestingly, Textures 2 and 10 in the image were extremely similar in appearance, so all of the unsupervised methods failed to distinguish these two textures. The fully connected CRFs and NED-MRF methods correctly recognized both textures but still suffered from some misclassification. To summarize, the MRF-MSSP achieved the best segmentation results of all of the methods, as shown in Figure 7(k). Moreover, the quantitative indices OA and kappa were even close to 1 for both texture images, as shown in Table 2.

The performance of the MRF-MSSP model was subsequently tested on remote sensing data. The Washington DC image with a spatial resolution of 1.65 m and a size of 730 ×

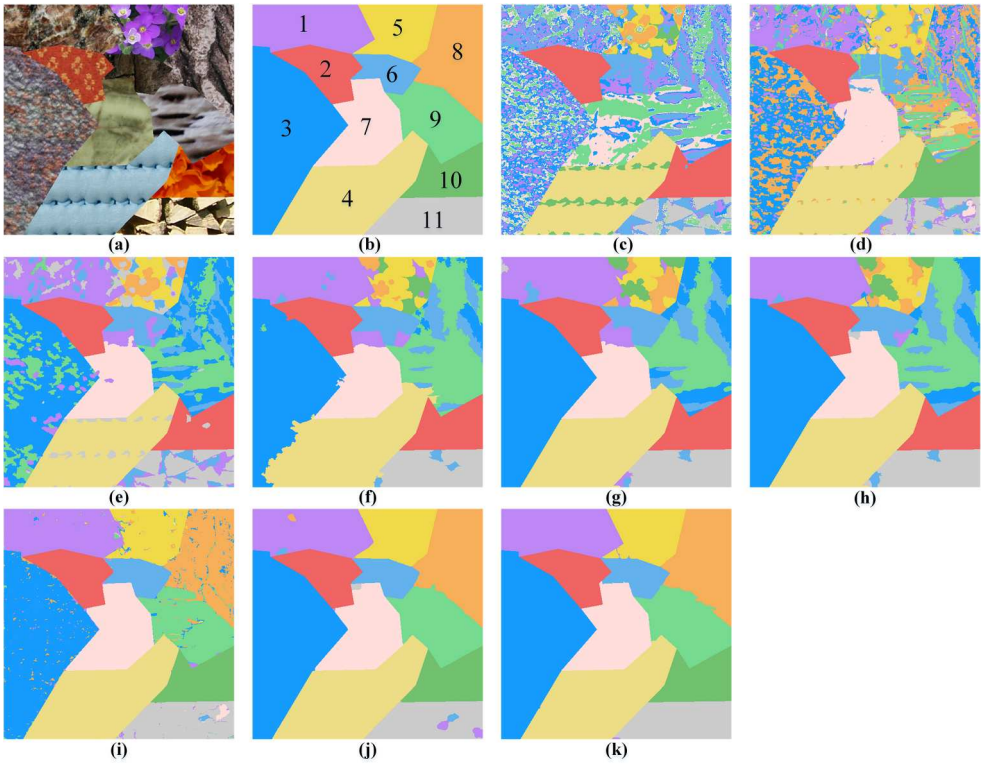

Figure 7. Segmentation results for the texture image II. (a) texture image II. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

770, obtained by the GeoEye sensor, was tested first. The Washington DC image contained four land-cover classes, i.e. buildings, grassland, trees, and roads. Among these classes, buildings and roads had similar spectral features, so the ICM and SVM methods incorrectly identified many roads as buildings, as shown in the bottom right corner of Figure 8(c,d). The object-based MRF approaches improved the segmentation accuracy. Unfortunately, the cost of the increased accuracy was the over-smoothing of image detail information. For example, the OMRF-RP misclassified many trees as grasslands, as shown in Figure 8(f). The analysis indicated that the OMRF-RP was weaker at distinguishing the classes with low spectral variability, as it mainly considered the area and boundary information between regions. The HMRF-MG method took the detailed information of the pixels and the macro-information of the objects into account, so the multigranularity-based methods could achieve better results than the HOMRF method on datasets where the hierarchical semantic features were not evident. The fully connected CRF method roughly distinguished between the four land-cover classes, but there was some salt-and-pepper noise around the boundaries. The MRF-MSSP further corrected the interactions between the two semantic scales by introducing prior information; therefore, the buildings within the black ellipse on the top left of Figure 8 could be correctly recognized. Table 3 reports the confusion matrix of the

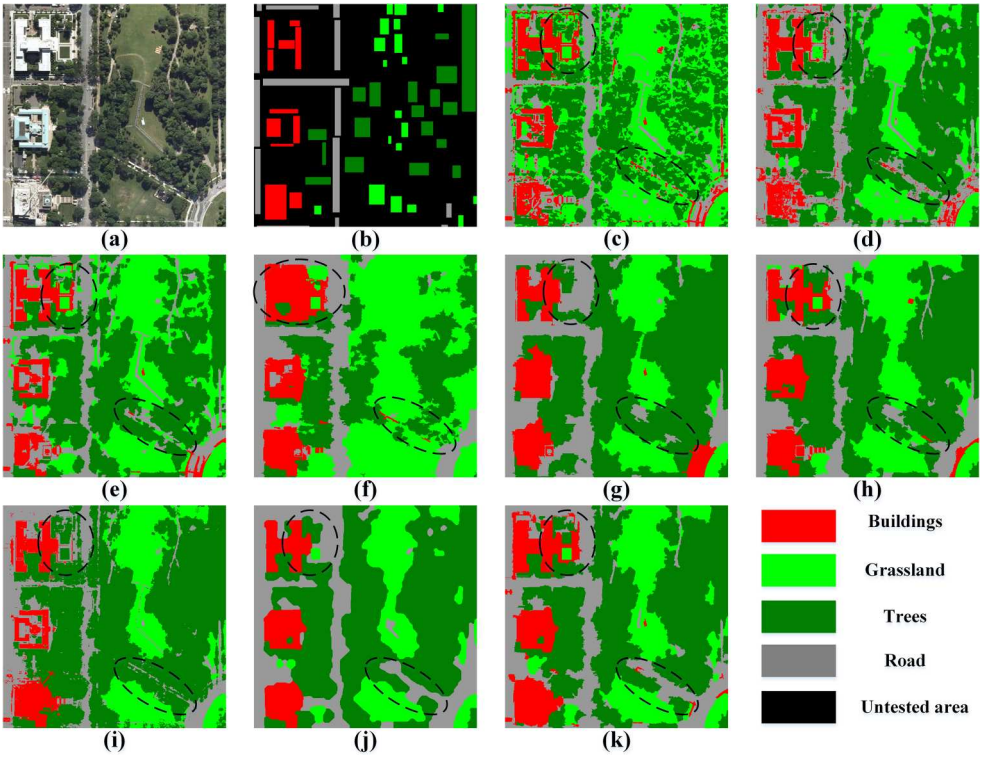

Figure 8. Segmentation results for the Washington DC image. (a) Washington DC image. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

Table 3. Quantitative indexes of ICM, SVM, OMRF, OMRF-RP, HOMRF, HMRF-MG, Fully connected CRFs, MRF-NED and MRF-MSSP for experiments in five remote sensing images of Figures 8–12.

Test image	Indexes (%)	ICM	SVM	OMRF	OMRF-RP	HOMRF	MRF-MG	FCRFs	NED-MRF	MRF-MSSP
Washington DC	Buildings	76.83	87.16	76.88	91.22	93.50	95.11	89.53	98.07	**99.92**
	Road	70.24	79.18	83.85	95.61	91.15	92.16	82.16	92.25	**96.68**
	Grassland	96.49	97.87	98.99	**100**	96.12	99.66	96.03	99.57	99.66
	Tress	72.72	95.38	96.33	77.74	98.40	**99.64**	99.84	99.59	99.50
	Kappa	71.88	86.55	86.32	85.33	93.42	95.68	89.93	96.27	**98.38**
	OA	76.38	89.71	89.56	88.45	95.14	96.85	92.46	97.28	**98.83**
SPOT5 image	Farmland	63.23	91.12	92.90	97.23	96.00	95.29	96.02	97.37	**97.91**
	Vegetation	77.80	79.43	82.37	80.58	86.68	83.66	81.97	80.69	**87.66**
	Urban area	49.78	83.16	84.56	86.62	91.27	88.24	87.22	**94.79**	93.17
	Kappa	56.44	80.18	82.65	84.87	88.39	85.67	85.00	87.93	**90.54**
	OA	64.07	85.33	87.35	89.16	91.82	89.75	89.25	91.49	**93.43**
Gaofen-2 image-I	Built-up	39.03	89.11	54.71	53.14	62.96	61.72	95.70	**97.00**	96.88
	Farmland	0.29	80.34	0.30	0.05	0.36	0.29	94.94	94.51	**95.87**
	Water	86.05	95.49	98.02	**99.97**	96.91	96.92	95.43	95.48	96.83
	Forest	97.72	96.32	96.83	97.56	96.59	96.91	98.29	98.67	**98.14**
	Kappa	29.31	80.86	36.08	35.72	39.16	38.75	92.81	93.83	**94.39**
	OA	33.77	87.27	43.21	42.34	47.94	47.23	95.67	96.33	**96.67**
Aerial image	Vegetation	79.93	76.59	90.00	68.90	**93.31**	92.23	98.29	86.62	90.33
	River	99.64	99.42	98.40	**99.97**	98.07	98.37	98.74	99.23	97.96
	Farmland	65.15	82.32	81.71	93.04	88.72	85.96	96.56	89.56	**99.36**
	Urban area	25.41	79.32	77.29	66.00	91.80	92.75	95.98	93.87	**98.11**
	Kappa	52.72	76.81	78.26	76.15	87.84	86.44	93.45	87.88	**96.29**
	OA	58.67	82.04	83.12	81.97	91.14	90.00	95.479	91.22	**97.49**
Salinas	Kappa	91.17	87.31	96.02	93.18	96.31	98.06	98.82	**99.33**	99.20
	OA	91.98	88.24	96.39	93.83	96.67	98.26	98.94	**99.40**	99.28

segmentation results, from which we can observe that the proposed method was the most reliable, especially for recognizing of buildings, where it achieved 99.92% accuracy.

To test the performance of the proposed method on remote sensing images with different spatial resolutions, the second tested remote sensing data was an image with a 10-m spatial resolution collected from the SPOT5 satellite. This image contained three categories: urban areas, vegetation, and farmland, as shown in Figure 9. As the spatial resolution decreased, the number of mixed pixels in the image increased, which exacerbated the uncertainty in the semantic segmentation. Although the SVM corrected most of the misclassified areas of the ICM, some vegetation was still misclassified as urban areas in the upper right corner of the image. The urban area in the image consisted of many subclasses, so the methods that utilize hierarchical semantic features, such as HOMRF and MRF-MSSP, provided better segmentation results and a higher segmentation accuracy, as shown in Figure 9(g,k). The fully connected CRF method performed poorly for images with many mixed pixels. The NED-MRF method exhibited positive properties in the boundaries between different classes. However, there were still misclassifications, such as the area within the red ellipse in the upper right corner of the image, as shown in Figure 9(j). The MRF-MSSP model has greater evaluation metrics than competing methods, and its visual presentation is superior.

An image of 2048 × 2048 pixels was selected from the GID to evaluate the model's performance on large-sized remote sensing images. It consisted of four land-cover categories: built-up areas, farmland, water, and forest, as shown in Figure 10. Interestingly, the four classes in this image had similar spectral responses. In particular,

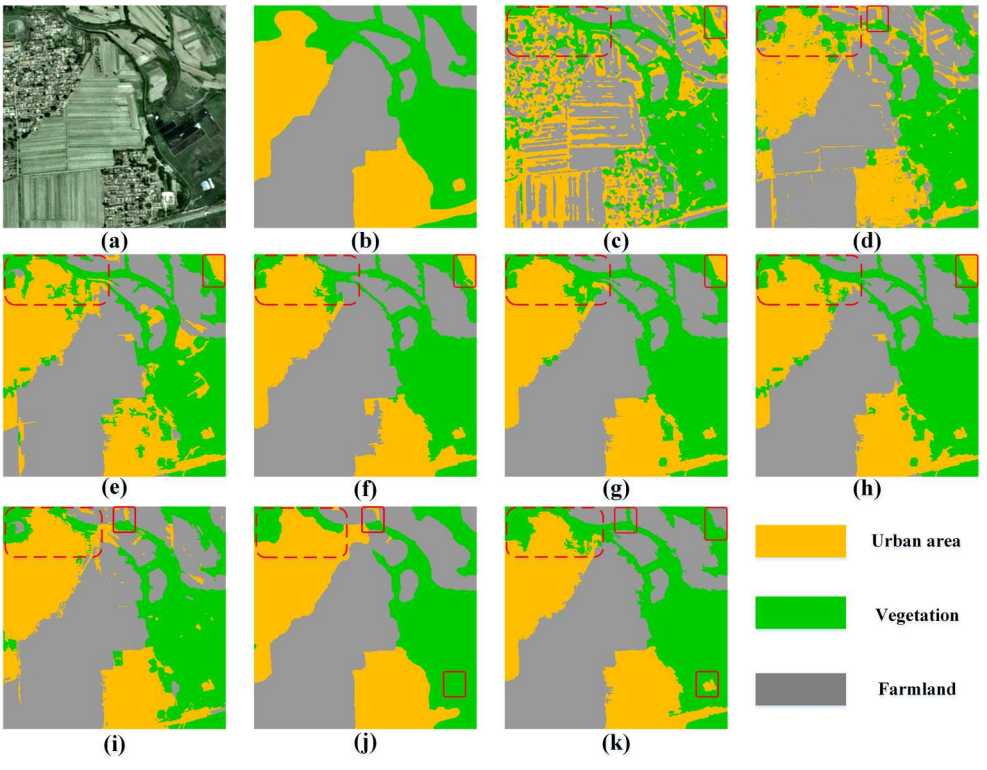

Figure 9. Segmentation results for the SPOT5 image. (a) SPOT5 image. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

the farmland, water, and forest at the top of the image appeared similarly. Moreover, there are many dark roofs in the built-up areas, and the difference between those areas and the farmland was slight from a visual perspective. Therefore, the five unsupervised methods failed to effectively distinguish the four classes, as shown in Figure 10(c,e–h). This error was corrected with the introduction of prior information. The MRF-MSPP smoothed the initial results provided by the SVM while preserving the local details of the image, such as it being able to distinguish between farmland and roads inside the farmland, as shown in Figure 10(k).

The sixth tested image, illustrated in Figure 11(a), was an aerial image with a 0.4-m ultra-high spatial resolution. It comprised four land-cover classes: urban areas, rivers, vegetation, and farmland. There was rich hierarchical semantic information in this image, where the urban area consisted of shadows, roads, vegetation, farmland, and buildings of different colors. The high intra-class heterogeneity of this image led to many misclassifications in the ICM, SVM, OMRF, and OMRF-RP results, as shown in Figure 11(c–f). The HOMRF and HMRF-MG could perform better using one label or granularity layer than the MRF methods. The fully connected CRFs performed poorly on the SPOT5 image used in this study, but exhibited promising properties for aerial images. The NED-MRF optimized the results using anisotropic spatial energy functions. However, they were still pixel-based approaches and employed only one label layer to

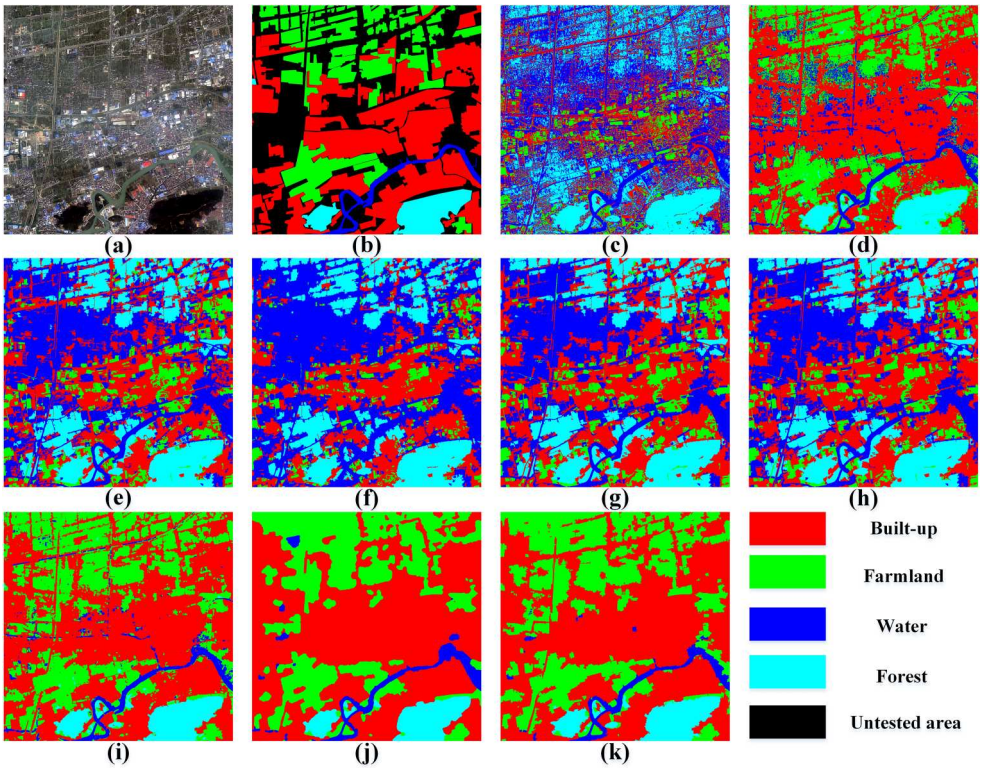

Figure 10. Segmentation results for the Gaofen-2 Image-I. (a) Gaofen-2 image-I. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

describe the image. Therefore, they failed to capture the large scope of the spatial interactions in the aerial images. As shown in Figure 11(j), the NED-MRF could not reliably distinguish farmland and vegetation in broad land-cover classes, and small misclassifications remained in the segmentation results of the fully connected CRFs, such as the red circled regions in Figure 11(i). The MRF-MSSP, introducing a spectral dissimilarity function, could generate the most complete homogeneous regions and preserve the boundaries between different classes. To further assess the contribution of equations (19–22), Figure 11(l) supplements the main results by displaying local segmentation details for enhanced visual evaluation. It can be observed that visual gaps between the different methods were evident and that the MRF-MSSP generated the closest boundaries to the ground truth.

Finally, the Salinas images were employed to evaluate the effectiveness of the MRF-MSSP model against hyperspectral remote sensing images. In this experiment, the SVM method provided the starting labels for all eight methods to achieve a fairer comparison. For the SVM method, 50 samples were randomly sampled from each ground truth class for training, and the remaining samples were used for testing. Except for the NED-MRF and fully connected CRF methods, the MRF-based methods employed an ICM to optimize the objective function in the model. If the image with 204 bands were segmented directly, it would cause the Gaussian distribution covariance matrices

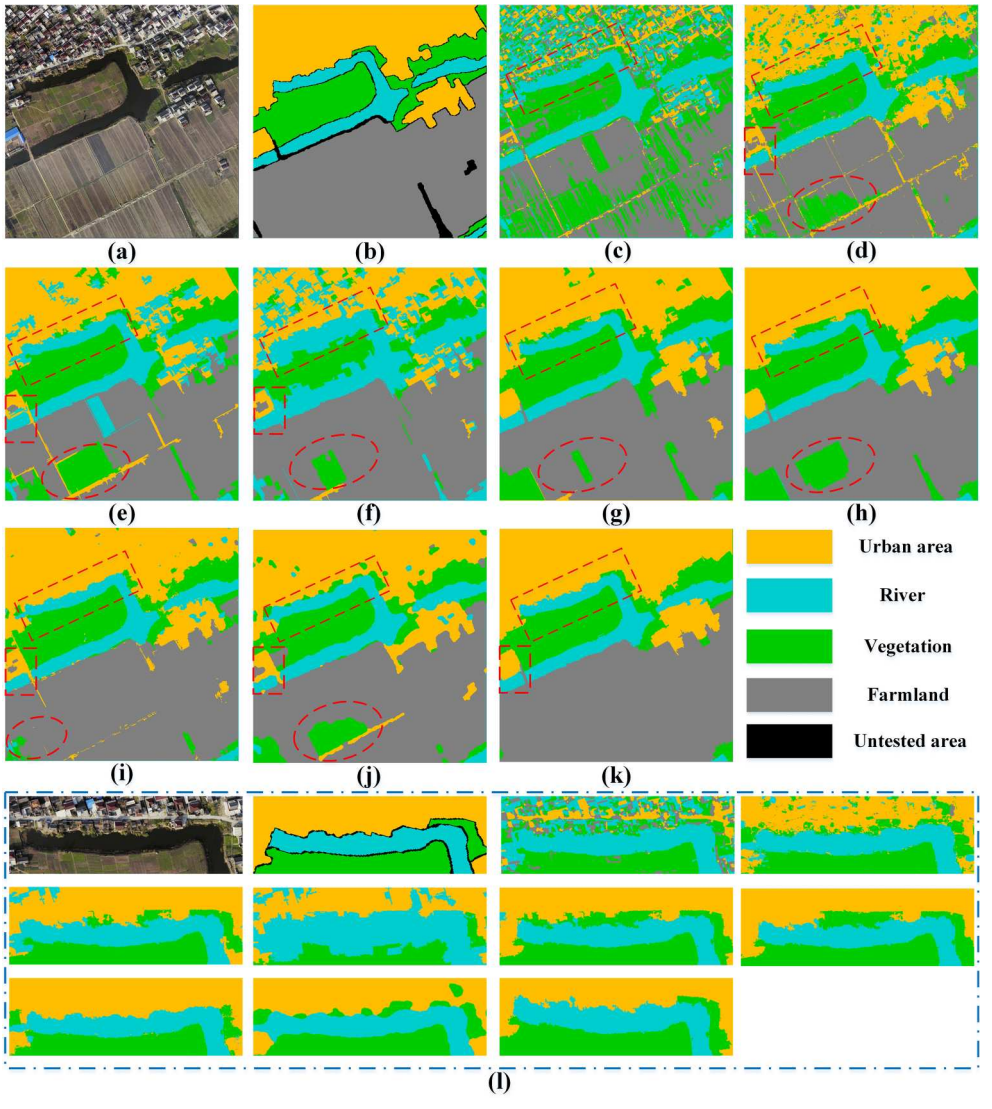

Figure 11. Segmentation results for the Aerial image. (a) Aerial image. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP. (l) Original local image patch and details of (a)-(k).

of these six methods to become singular or nearly singular, which could impact the final segmentation results. Therefore, the input data for these six methods were the three-band image of Salinas acquired by principal component analysis. The NED-MRF and fully connected CRF methods still input the original image with 204 bands. Figure 12 displays the segmentation results. The Salinas image had few hierarchical semantic features available, as its 16 classes are all crops. HOMRF failed to discriminate well across classes with similar spectral responses, particularly in the boundaries between classes. The MRF-MSSP and NED-MRF models incorporated the local spectral heterogeneity of the image to further optimize the segmentation results. Table 3 presents quantitative

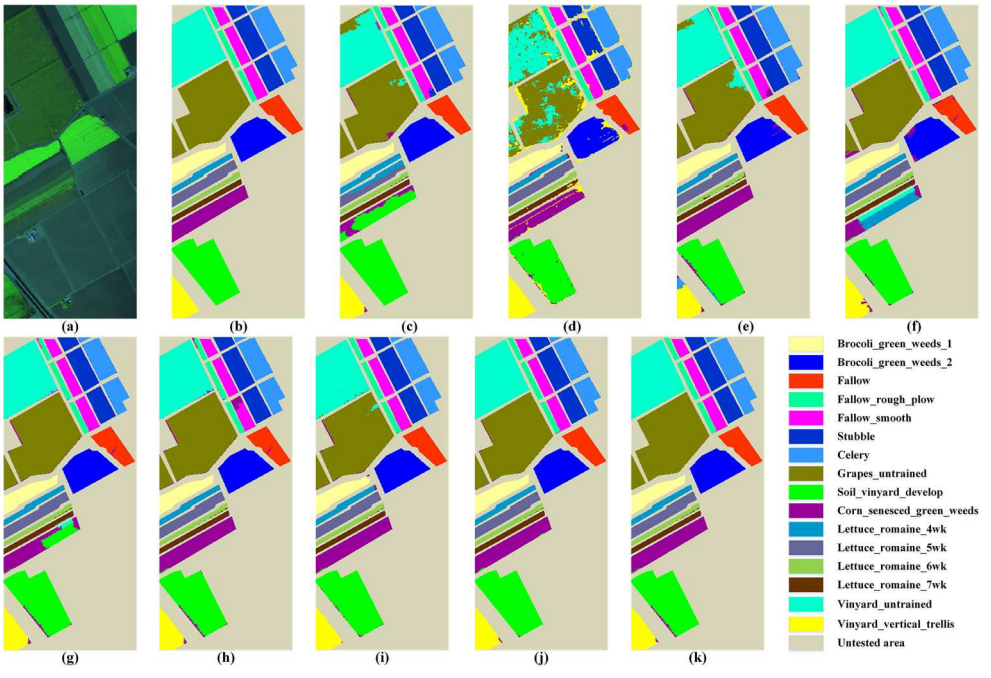

Figure 12. Segmentation results for the Salinas hyperspectral image. (a) Salinas hyperspectral image. (b) Ground truth. (c) Result of ICM. (d) Result of SVM. (e) Result of OMRF. (f) Result of OMRF-RP. (g) Result of HOMRF. (h) Result of HMRF-MG. (i) Result of Fully connected CRFs. (j) Result of MRF-NED. (k) Result of MRF-MSSP.

indices that also demonstrate the robustness of the MRF-MSSP model against hyperspectral remote sensing images. Figure 13 illustrates the confusion matrix of segmentation results on Salinas image using the MRF-MSSP model.

4.3.2. Comparison with deep-learning-based methods

For the deep-learning-based methods, four test images of size 2048×2048 pixels were selected from the GID dataset, while the remaining images were used for model training. Specifically, to examine the influences of the varied quantities of the training data on the performances of the CNN-based models, two patch sets were collected to provide training data for the seven CNN methods. First, patches with sizes of 256×256 were randomly sampled on each training image. Patch set I contained 14,600 patches, and patch set II contained 2920 patches. The four classes were randomly distributed in two patch sets. The FPN was trained on Patch set II with 2920 patches. Subsequently, the remaining six methods were trained on Patch set I with 14,600 patches.

The deep learning models employed in this study were implemented using the PyTorch framework. All experiments were conducted on a workstation equipped with an NVIDIA RTX A4000 GPU and 16GB of RAM, utilizing CUDA 11.2 to leverage efficient parallel computation capabilities. The operating system used was a 64-bit version of Microsoft Windows 10, with the development platform being Anaconda 5.2.0 and Python version 3.8.8.

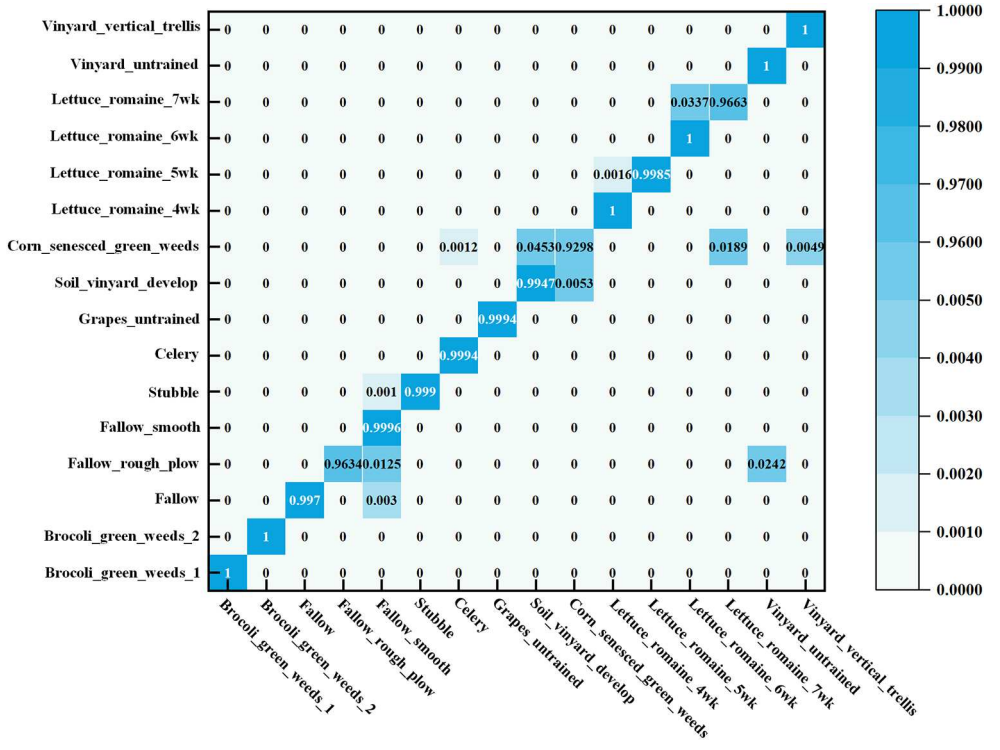

Figure 13. Confusion matrix of segmentation results on the 16 classes of Salinas image obtained using the MRF-MSSP model.

During model training, the SGD optimizer was used to ensure stability and efficiency in the optimization process. The batch size was set to 32, with a total of 50 epochs. To enhance the training performance, the initial learning rate was set to 0.01, and a cosine annealing schedule was applied to dynamically adjust the learning rate during training, gradually decreasing it to promote better convergence. A global weight decay parameter of 0.001 was also applied to improve generalization and reduce overfitting, ensuring the stability and reliability of the training process.

The experimental results are presented in Figure 14. It can be observed that the segmentation result of the MRF-MSSP model was the closest to the ground truth. Notably, the FPN had difficulty distinguishing between farmland and water. The FPN misclassified the large areas of water in Gaofen-2 image II as farmland and misclassified the farmland at the bottom of Gaofen-2 image IV as water, as shown in Figure 14(a3,c3), respectively. This may have been caused by the fact that Patch set II provided less available training data for the FPN and that the spectral responses of the two classes were similar. Patch set I had more training samples accessible to model than patch set II. As a consequence, the models such as U-net were able to generate satisfactory segmentation results. In particular, U-net obtained the highest kappa and OA, with values of 0.9707 and 0.9836, respectively, on the Gaofen-2 image-V. The seven CNN-based methods were more effective in recognizing the built-up areas but poorer in recognizing farmland. In the GID, the farmland comprised many heterogeneous regions with

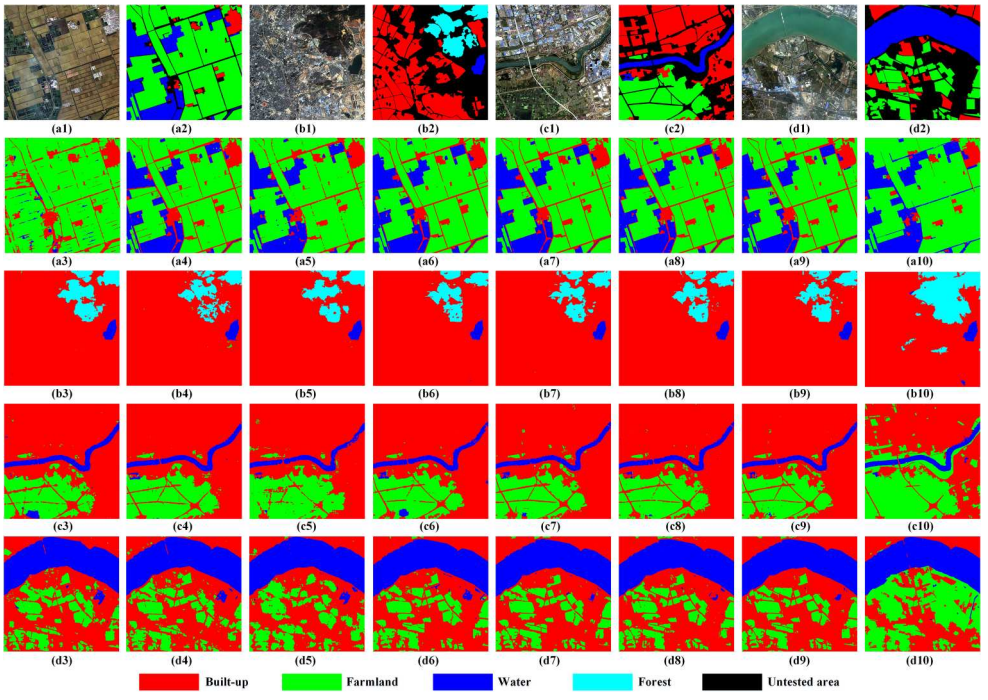

Figure 14. Segmentation results for the GID. (a1) Gaofen-2 image-II. (b1) Gaofen-2 image-III. (c1) Gaofen-2 image-IV. (d1) Gaofen-2 image-V. (a2–d2) Ground truth. (a3–d3) Result of FPN. (a4–d4) Result of U-net. (a5–d5) Result of DeepLabV3+. (a6–d6) Result of CGLNet. (a7–d7) Result of CMLFormer. (a8–d8) Result of CMTFNe. (a9–d9) Result of SFFNet. (a10–d10) Result of MRF-MSSP.

significant intra-class variations. The spectral responses of the crops varied. It can be observed from Figure 14(d1) that the appearances of built-up and farmland areas were similar. For these areas, the CNN-based methods generated more noise in the segmentation map. Compared to the CNN-based methods, MRF-MSSP proposed algorithm correctly recognized water, farmland, and built-up areas and minimized the salt-and-pepper noise in the segmentation results. The kappa, OA, and specific-class accuracies achieved by the eight methods are shown in Table 4. The abovementioned experimental findings further demonstrated the robustness and dependability of MRF-MSSP model for diverse HRRS images.

4.4. Analysis of MRF-MSSP model parameter

The abovementioned experimental results demonstrated the promising performance of the MRF-MSSP model. However, the parameters of the model will impact the segmentation results. This subsection analyzes and discusses the effects of parameters k_1, β, and *MRA* on the robustness of the MRF-MSSP model through additional experiments. To evaluate this effect more accurately, a strategy was adopted in which one parameter was tested at a time while the other two parameters were fixed. The effect of parameter k_1 on the model was first tested. Table 5 reports the optimal k_1 values for the seven

Table 4. Quantitative indexes of U-net (Ronneberger, Fischer, and Brox 2015), FPN (Lin et al. 2017), DeepLabV3+ (Chen et al. 2018), CGGLNet (Ni et al. 2024), CMLFormer (Wu et al. 2024), CMTFNet (Wu et al. 2023), SFFNet (Yang, Yuan, and Li 2024) and MRF-MSSP for experiments of eight Gaofen-2 images of Figure 14.

Tested image	Indexes (%)	FPN	U-net	DeepLabV3 +	CGLNet	CML former	CMTFNe	SFFNet	MRF-MSSP
Gaofen-2 image-II	Water	3.3	95.23	94.21	97.36	97.24	97.07	97.88	**99.93**
	Farmland	97.08	98.64	98.38	99.13	98.92	98.88	**99.2**	98.32
	Built-up	98.33	99.47	**98.35**	88.01	87.03	87.36	89.28	97.81
	Kappa	52.50	95.12	94.06	94.38	93.79	93.75	95.04	**96.35**
	OA	77.80	97.98	97.52	97.16	96.87	96.85	97.5	**98.49**
Gaofen-2 image-III	Water	97.30	99.01	95.36	99.14	98.13	97.99	96.09	**99.41**
	Forest	81.75	72.56	76.43	75.54	79.59	80.75	78.73	**99.71**
	Built-up	**99.99**	99.97	**99.99**	99.38	99.09	99.16	99.36	99.85
	Kappa	90.13	85.53	87.28	85.04	86.29	87.27	86.69	**99.48**
	OA	96.46	94.74	95.41	96.66	96.76	96.95	96.89	**99.81**
Gaofen-2 image-IV	Water	91.24	**94.33**	83.46	93.19	94.23	91.65	93.34	91.09
	Farmland	85.89	96.86	94.03	92.29	92.28	94.5	93.52	**99.04**
	Built-up	**99.89**	**99.89**	99.74	98.11	97.57	98	98.2	97.75
	Kappa	90.25	97.07	93.59	92.26	91.11	92.69	**92.72**	95.82
	OA	94.31	**98.36**	96.36	96.65	96.17	96.84	96.86	97.62
Gaofen-2 image-V	Water	98.87	99.35	99.30	99.75	99.65	99.68	**99.76**	98.93
	Farmland	77.46	83.20	76.90	86.62	87.6	86.89	88.26	**91.24**
	Built-up	**99.37**	98.50	**99.08**	94.34	94.69	95.56	96.51	96.63
	Kappa	87.92	90.67	87.84	90.45	91.05	91.58	92.9	**93.73**
	OA	91.76	93.75	91.72	94.24	94.6	94.9	95.73	**95.88**

images, and Figure 15 illustrates the effects of different k_1 values on the segmentation accuracy.

As can be observed from Figure 15(a–c), the segmentation accuracy was improved with the increase in the k_1 value, and then it decreased. Given that Salinas was divided into 16 classes, if the strategy of $k_1 > k$ were still adopted, only a few available hierarchical semantic features could be extracted by the model. We set $k_1 < k$ with the goal of interpreting the images from a high-level-semantics perspective. As illustrated in Figure 15(d), when k_1 ranged from 4 to 10, the quantitative indices exhibited very minor fluctuations, and the kappa values always exceeded 0.99. The abovementioned analysis demonstrated that the MRF-MSSP model was highly robust to k_1.

When k and k_1 were specified, the parameters β and MRA impacted the performance of the model jointly. The segmentation accuracies of four images with different values β and MRA were compared. For Washington DC images, the β value was set from 1 to 15 with a step of 1, and the MRA value was set from 50 to 190 with a step of 10. The

Table 5. The optimal k_1 value of the MRF-MSSP model for seven images.

Image	k value	k_1 value	Kappa (%)	OA (%)
Texture image I	6	10	99.59	99.67
Texture image II	11	14	99.75	99.78
Washington DC	4	7	98.38	98.83
SPOT5 image	3	6	90.54	93.43
Gaofen-2 image-I	4	9	94.39	96.67
Aerial image	4	7	96.29	97.49
Salinas	16	6	99.20	99.28

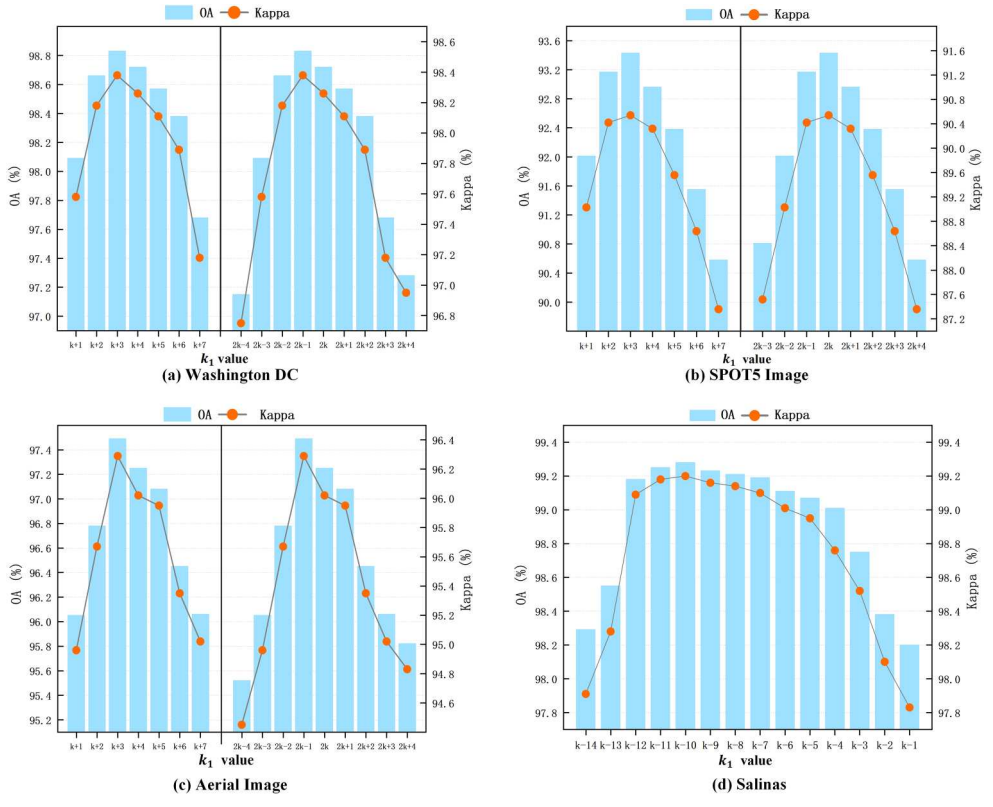

Figure 15. Quantitative indexes for results with different k_1 value.

parameter setting strategy for the other three images was similar to that used for the Washington DC image. As shown in Figure 16, the MRF-MSSP model was also quite robust against β and *MRA*. Based on the aforementioned discussion of the parameters, it can be concluded that the MRF-MSSP model exhibits a high degree of robustness with respect to parameter settings.

4.5. Computational cost

The computational complexity of the MRF-MSSP model was $O((k + k_1)nt)$, where n is the number of vertices, and t is the number of iterations. Compared with other competing methods, the computational complexity of the MRF-MSSP model was the same as that of the HOMRF model. In fact, the MRF-MSSP model additionally calculated the spectral dissimilarity between adjacent regions, and thus, its computation time was slightly higher than that of the HOMRF. For example, for the segmentation experiment of the Salinas image with a size of 512×217, the MRF-MSSP model required 8.7 s, the HOMRF model required 7.5 s, and the OMRF model required 6.3 s. Notably, all of the experiments (statistical learning methods) in this study were performed on a Windows 11 PC with an Intel i5-12400 CPU using 16 GB of memory. Table 6 reports the computing times of the MRF-MSSP model on the seven images. In general, the computational speed of the MRF-MSSP mode was acceptable.

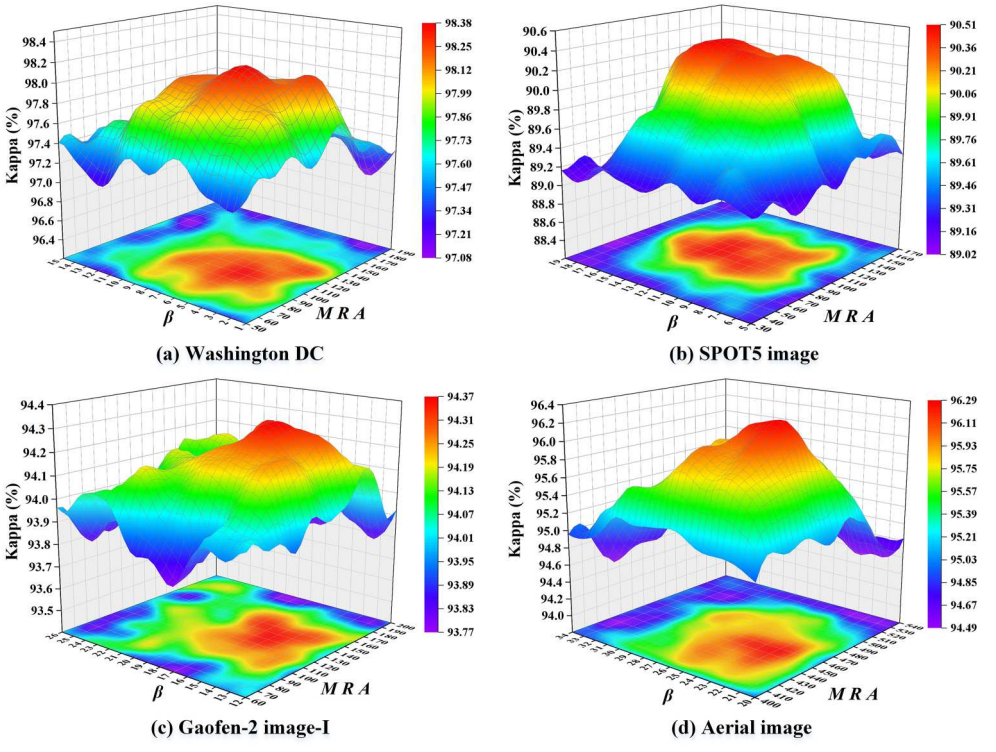

Figure 16. Kappa index of the MRF-MSSP model with different β values and *MRA* values.

5. Discussion

In this study, a theoretical and experimental investigation of the concept of multilayer semantic segmentation and local spectral dissimilarity of remote sensing images was conducted. To demonstrate the differences between the comparison methods and the MRF-MSSP method more intuitively, Figure 17 displays the kappa values obtained by nine methods, including the SVM, fully connected CRF, and MRF-based methods, on seven tested images. Figure 18 displays the kappa obtained by the deep-learning-based methods and the MRF-MSSP method on four remote sensing images from the GID. It can be observed from Figures 17 and 18 that the proposed MRF-MSSP model achieved the highest segmentation accuracy in most cases. Overall, the method proposed in this study was the most robust.

In recent years, with the improvement in the spatial resolution of remote sensing images, the images have been able to contain rich and complex spatial context

Table 6. The computational time of MRF-MSSP model in seven images (in seconds).

	Data						
Method	Texture image I	Texture image II	Washington DC	SPOT5 image	Gaofen-2 image I	Aerial image	Salinas
MRF-MSSP	6.0	5.1	7.3	6.9	102.2	14.2	8.7

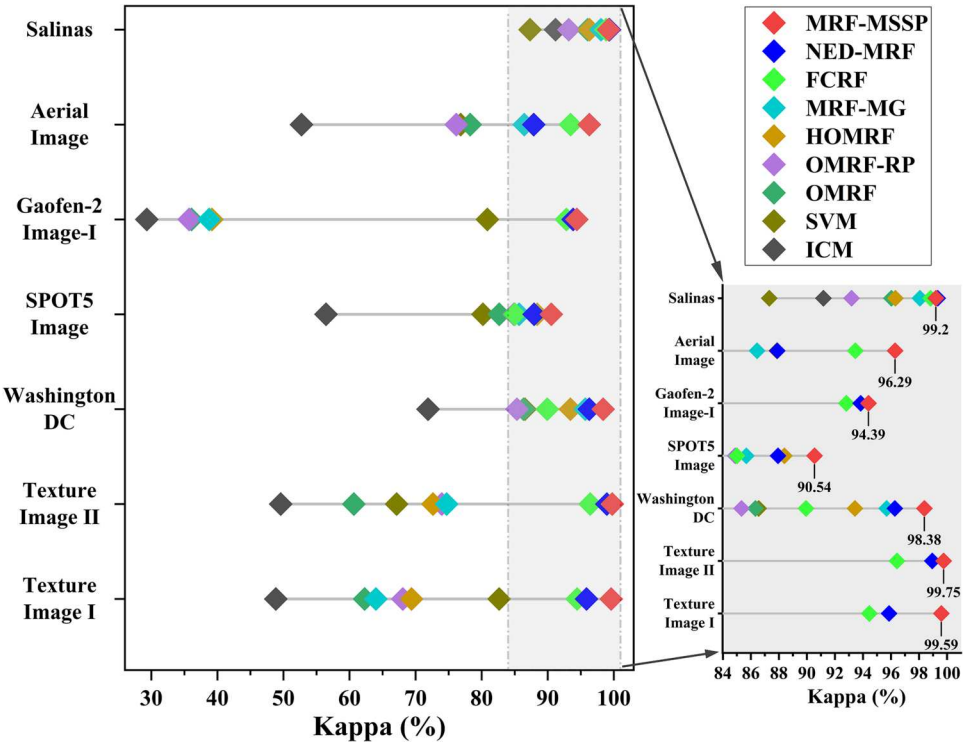

Figure 17. Kappa index for results with SVM, fully connected CRFs and six MRF-based methods.

information. Previous studies have demonstrated the use of multiclass-layer labels to effectively capture the spatial context information of images from different semantic perspectives (L. Li et al. 2022; Zheng, Zhang, and Wang 2016). A similar strategy was adopted in this work to interpret the images. For example, the MRF-MSSP divided urban areas of aerial images into four classes with low-level semantic features from a detailed perspective, and it divided farmland and vegetation into two classes with low-level semantic features, as shown in Figure 2. The model captured the interactions between low- and high-level semantic classes in the form of a transition probability matrix. However, for some complex remote sensing images, the unsupervised methods had difficulty capturing the semantic context without prior information, which led to final segmentation results not accurately reflecting the ground truth, as shown in Figure 10(c). As a result, the SVM was employed to initialize the label layers in the label model, instead of the classical pixel-based MRF method (ICM). It was demonstrated in Section 4 that incorporating prior information into semantic segmentation led to improved results. The next part will focus on explaining the introduction of regional spectral dissimilarity into the hierarchical semantic model.

MRF-based methods improve the semantic segmentation of remote sensing images. Unfortunately, they improve the segmentation accuracy at the cost of over-smoothing detailed image information. This is because most MRF models prefer to regard adjacent vertices in the probability graph G as the same class. Thus, they can obtain large-scale homogeneous regions. Conversely, this makes it possible to lose boundary information

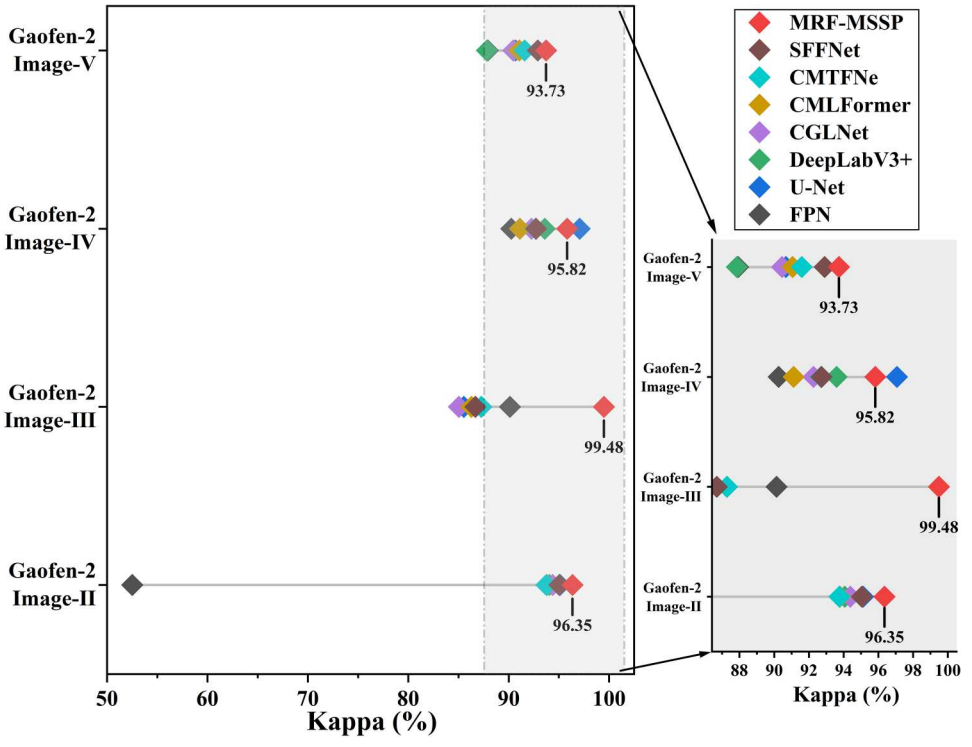

Figure 18. Kappa index for results with seven deep learning-based methods and MRF-MMSP method.

in the segmentation results. Expressly, the ICM, OMRF, HOMRF, and HMRF-MG methods assume the model is uniformly smooth over the whole image. That is, these methods assign the same weight when calculating the potential energy $\varphi(x_l, x_j)$ between adjacent nodes v_l and v_j in the probability graph G. However, this assumption is frequently violated in the boundary between different classes.

From Figures 8 and 12, it can be seen that these methods over-smoothed the boundaries of the images. To prevent the segmentation map from being over-smoothed, a new optimized scheme is proposed. Under the same sensor and illumination conditions, objects belonging to the same class should present similar spectral features. Based on this, the model should prefer to select vertices with similar spectra to those of the same class. For this purpose, the dissimilarity between adjacent regions was introduced into the potential function of the label model as adaptive weights (see equation (17)), which aims to preserve the dissimilarity of regions exhibiting discontinuities to prevent an image from being over-smoothed. A previous study (Wang et al. 2017) confirmed the effectiveness of this strategy, which was used as a competing method (NED-MRF) in this article. The NED-MRF model performed poorly when handling large-scope homogeneous regions but performed well in edge-preserving, as shown in Figures 10(j), 11(j), and 12(j). Different methods have their own strengths and weaknesses, and we can choose the appropriate solution based on the circumstances.

In the field of semantic segmentation of remote sensing images, the excellent performance of deep learning approaches has been proven (Ma et al. 2019). They can provide

satisfactory segmentation results with the support of massive training data. Consequently, they have been extensively studied and continuously improved. In this work, the MRF-MSSP algorithm was compared with seven CNN deep learning methods. In contrast to deep-learning-based methods, the MRF method is a statistical learning method. It has a robust theoretical basis and a flexible theoretical framework. Its segmentation process can be precisely characterized by mathematical formulations, which makes the results of the MRF-MSSP model easier to interpret than those of deep learning models. In addition, the MRF can capture the statistical regularity of the geographic distribution of each land-cover class from a probabilistic perspective, and this regularity is beneficial for the model to generate results that are closer to the ground truth. In particular, deep-learning-based methods are greatly impacted by the quality and quantity of the training samples. In the absence of sufficient training samples, the MRF models outperformed the CNN models. When the number of training samples is massive, deep-learning-based methods are regarded as the optimal option. However, since deep-learning-based methods require feature learning from the massive number of training samples, a substantial amount computational resources will be required to train the models. Several deep learning methods have forced researchers to reconsider using more efficient hardware architectures to accommodate the computational complexity and accelerate the training of models. By contrast, the MRF model requires fewer computational resources. For instance, in this study, the DeepLabV3+ model was trained using a dataset comprising 14,600 images, each with a size of 256 × 256 pixels. The training process took approximately 7 h and 23 min to complete, highlighting the significant computational cost associated with deep learning-based semantic segmentation approaches.

6. Conclusions

This paper presents a semantic segmentation algorithm applicable to both high-resolution remote sensing (HRRS) images and hyperspectral images. The proposed MRF-MSSP model has the following positive properties:

- Semantic-aware segmentation: The model automatically extracts semantic contextual information from the input image to guide the segmentation process. This capability enables improved performance, particularly in complex remote sensing scenarios.
- Adaptive boundary preservation: By measuring the spectral dissimilarity between adjacent objects, the model adaptively adjusts smoothness across different regions. As a result, it effectively preserves boundary details instead of over-smoothing them.
- Dual-layer semantic outputs: The method generates segmentation maps for two semantic levels – high-level and low-level. These multi-layer results provide more comprehensive scene interpretation. An example of these results for the Gaofen-2 images is presented in Appendix Figure A.1.

In addition, the proposed model significantly reduces salt-and-pepper noise commonly observed in segmentation maps, further enhancing the visual quality and accuracy of the results.

Different types of images were employed to evaluate the effectiveness of the MRF-MSSP model, and the effect of various parameter-setting strategies on the model's robustness was tested. The experimental results demonstrated that the MRF-MSSP model is capable of generating the closest ground truth results, and it is slightly influenced by the parameters. In addition, the segmentation quality indices indicate that our method provided the best segmentation accuracy on most of the tested images compared with other competing methods. In summary, the proposed MRF-MSSP model provides an optional solution for the semantic segmentation of remote sensing images.

Although the abovementioned experimental results showed an encouraging segmentation performance of the proposed method, it still has limitations. Two aspects are involved in the analysis of the deficiencies of the MRF-MSSP model. Firstly, it should be noted that not all of the images exhibited distinct hierarchical semantic features, as is the case with the Salinas image. In the case of images that lack distinct hierarchical semantic features, the MRF-MSSP model is unable to collect sufficient hierarchical semantic features to guide semantic segmentation effectively. The question of how to effectively extract semantic features in fine classification images is an intriguing challenge, and it will be the focus of our subsequent research efforts. Secondly, as illustrated in Figures 8–12, the segmentation outcomes obtained by the MRF-MSSP exhibited jagged boundaries between the different classes, a common occurrence in object-based methodologies. In the future, we intend to improve the quality of the over-segmented region generated by the basic segmentation techniques and integrate multi-granularity features of images further to refine the jagged boundaries of the segmentation map.

Acknowledgements

We deeply appreciate the anonymous reviewers for their invaluable feedback and meticulous efforts, which have played a pivotal role in refining the content and enhancing the rigor of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Natural Science Foundation of China [grant number 32160369], the Fundamental Research Project of Yunnan Province [grant number 202501AS070090], and the Ten Thousand Talents Program for Young Top-notch Talents of Yunnan Province [grant number YNWR-QNBJ-2019-026].

Data availability statement

The testing data and source code for the MRF-MSSP model have been made publicly available at the following GitHub repository: https://github.com/rexingxiaowang/MRF-MSSP_Model.

ORCID

Jun Wang http://orcid.org/0000-0003-0451-1433
Leiguang Wang http://orcid.org/0000-0003-2962-1508

References

Adams, R., and L. Bischof. 1994. "Seeded Region Growing." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16 (6): 641–647. https://doi.org/10.1109/34.295913.

Besag, J. 1986. "On the Statistical Analysis of Dirty Pictures." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 48 (3): 259–279. https://doi.org/10.1111/j.2517-6161.1986.tb01412.x.

Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Queiroz Feitosa, et al. 2014. "Geographic Object-Based Image Analysis – towards a New Paradigm." *ISPRS Journal of Photogrammetry and Remote Sensing* 87:180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014.

Bouman, C. A., and M. Shapiro. 1994. "A Multiscale Random Field Model for Bayesian Image Segmentation." *IEEE Transactions on Image Processing* 3 (2): 162–177. https://doi.org/10.1109/83.277898.

Chan, R. H., C.-W. Ho, and M. Nikolova. 2005. "Salt-and-Pepper Noise Removal by Median-Type Noise Detectors and Detail-Preserving Regularization." *IEEE Transactions on Image Processing* 14 (10): 1479–1485. https://doi.org/10.1109/TIP.2005.852196.

Chang, C.-C., and C.-J. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology* 2 (3): 1–27. https://doi.org/10.1145/1961189.1961199.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Cheng, H., and C. A. Bouman. 2001. "Multiscale Bayesian Segmentation Using a Trainable Context Model." *IEEE Transactions on Image Processing* 10 (4): 511–525. https://doi.org/10.1109/83.913586.

Comaniciu, D., and P. Meer. 2002. "Mean Shift: A Robust Approach toward Feature Space Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (5): 603–619. https://doi.org/10.1109/34.1000236.

Cross, G. R., and A. K. Jain. 1983. "Markov Random Field Texture Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5 (1): 25–39. https://doi.org/10.1109/TPAMI.1983.4767341.

Dai, Q., B. Luo, C. Zheng, and L. Wang. 2020. "Regional Multiscale Markov Random Field for Remote Sensing Image Classification." *J Remote Sens (Chinese)* 24 (03): 245–253.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39 (1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Dronova, I., P. Gong, N. E. Clinton, L. Wang, W. Fu, S. Qi, and Y. Liu. 2012. "Landscape Analysis of Wetland Plant Functional Types: The Effects of Image Segmentation Scale, Vegetation Classes and Classification Methods." *Remote Sensing of Environment* 127:357–369. https://doi.org/10.1016/j.rse.2012.09.018.

Duro, D. C., S. E. Franklin, and M. G. Dubé. 2012. "A Comparison of Pixel-Based and Object-Based Image Analysis with Selected Machine Learning Algorithms for the Classification of Agricultural Landscapes Using SPOT-5 HRG Imagery." *Remote Sensing of Environment* 118:259–272. https://doi.org/10.1016/j.rse.2011.11.020.

Felzenszwalb, P. F., and D. P. Huttenlocher. 2004. "Efficient Graph-Based Image Segmentation." *International Journal of Computer Vision* 59 (2): 167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77.

Feng, D., C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer. 2021. "Deep Multi-modal Object Detection and Semantic Segmentation for

Autonomous Driving: Datasets, Methods, and Challenges." *IEEE Transactions on Intelligent Transportation Systems* 22 (3): 1341–1360. https://doi.org/10.1109/TITS.2020.2972974.

Ghamisi, P., E. Maggiori, S. Li, R. Souza, Y. Tarablaka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, and M. Chi. 2018. "New Frontiers in Spectral-Spatial Hyperspectral Image Classification: The Latest Advances Based on Mathematical Morphology, Markov Random Fields, Segmentation, Sparse Representation, and Deep Learning." *IEEE Geoscience and Remote Sensing Magazine* 6 (3): 10–43. https://doi.org/10.1109/MGRS.2018.2854840.

Grinias, I., C. Panagiotakis, and G. Tziritas. 2016. "MRF-based Segmentation and Unsupervised Classification for Building and Road Detection in Peri-urban Areas of High-Resolution Satellite Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 122:145–166. https://doi.org/10.1016/j.isprsjprs.2016.10.010.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. "A Survey of Methods for Explaining Black box Models." *ACM Computing Surveys* 51 (5): 1–42. https://doi.org/10.1145/3236009.

Hay, G. J., and G. Castilla. 2008. "Geographic Object-Based Image Analysis (GEOBIA): A New Name for a New Discipline." In *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*, edited by T. Blaschke, S. Lang, and G. J. Hay, 75–89. Berlin: Springer. https://doi.org/10.1007/978-3-540-77058-9_4.

He, D., Q. Shi, X. Liu, Y. Zhong, and L. Zhang. 2022. "Generating 2 m Fine-Scale Urban Tree Cover Product over 34 Metropolises in China Based on Deep Context-Aware Sub-pixel Mapping Network." *International Journal of Applied Earth Observation and Geoinformation* 106:102667. https://doi.org/10.1016/j.jag.2021.102667.

Hossain, M. D., and D. Chen. 2019. "Segmentation for Object-Based Image Analysis (OBIA): A Review of Algorithms and Challenges from Remote Sensing Perspective." *ISPRS Journal of Photogrammetry and Remote Sensing* 150:115–134. https://doi.org/10.1016/j.isprsjprs.2019.02.009.

Huang, Z., K. Li, Y. Jiang, Z. Jia, L. Lv, and Y. Ma. 2024. "Graph Relearn Network: Reducing Performance Variance and Improving Prediction Accuracy of Graph Neural Networks." *Knowledge-Based Systems* 301:112311. https://doi.org/10.1016/j.knosys.2024.112311.

Huang, X., and L. Zhang. 2008. "An Adaptive Mean-Shift Analysis Approach for Object Extraction and Classification from Urban Hyperspectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 46 (12): 4173–4185. https://doi.org/10.1109/TGRS.2008.2002577.

Kang, L., B. Tang, J. Huang, and J. Li. 2024. "3D-MRI Super-resolution Reconstruction Using Multi-modality Based on Multi-resolution cnn." *Computer Methods and Programs in Biomedicine* 248:108110. https://doi.org/10.1016/j.cmpb.2024.108110.

Krähenbühl, P., and V. Koltun. 2011. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 109–117. https://doi.org/10.48550/arXiv.1210.5644.

Kuo, W.-F., and Y.-N. Sun. 2010. "Watershed Segmentation with Automatic Altitude Selection and Region Merging Based on the Markov Random Field Model." *International Journal of Pattern Recognition and Artificial Intelligence* 24 (01): 153–171. https://doi.org/10.1142/S021800141000783X.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Li, S. Z. 2009. *Markov Random Field Modeling in Image Analysis*. Berlin: Springer Science & Business Media.

Li, S. Z. 2012. *Markov Random Field Modeling in Computer Vision*. Berlin: Springer Science & Business Media.

Li, X., F. Dunkin, and J. Dezert. 2023. "Multi-source Information Fusion: Progress and Future." *Chinese Journal of Aeronautics* 37 (7): 24–58. https://doi.org/10.1016/j.cja.2023.12.009.

Li, S., Q. Hao, X. Kang, and J. A. Benediktsson. 2018. "Gaussian Pyramid Based Multiscale Feature Fusion for Hyperspectral Image Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (9): 3312–3324. https://doi.org/10.1109/JSTARS.2018.2856741.

Li, L., T. Zhou, W. Wang, J. Li, and Y. Yang. 2022. "Deep Hierarchical Semantic Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 18–24. New Orleans, LA: IEEE. https://doi.org/10.1109/CVPR52688.2022.00131.

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature Pyramid Networks for Object Detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125. Honolulu, HI: IEEE. https://doi.org/10.1109/CVPR.2017.106.

Lin, D., Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang. 2018. "Multi-scale Context Intertwining for Semantic Segmentation." In *Proceedings of the European Conference on Computer Vision (ECCV)*, edited by L. Leal-Taixé and S. Roth, 603–619. Munich: Springer. https://doi.org/10.1007/978-3-030-01219-9_37.

Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. Boston, MA: IEEE. 10.1109/TPAMI.2016.2572683.

Ma, L., Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. 2019. "Deep Learning in Remote Sensing Applications: A Meta-analysis and Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 152:166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Masson, P., and W. Pieczynski. 1993. "SEM Algorithm and Unsupervised Statistical Segmentation of Satellite Images." *IEEE Transactions on Geoscience and Remote Sensing* 31 (3): 618–633. https://doi.org/10.1109/36.225529.

Mikeš, S., and M. Haindl. 2022. "Texture Segmentation Benchmark." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (9): 5647–5663. https://doi.org/10.1109/TPAMI.2021.3075916.

Mitchell, T. M. 1997. *Machine Learning*. Vol. 1. New York: McGraw-hill New York.

Ni, Y., J. Liu, W. Chi, X. Wang, and D. Li. 2024. "CGGLNet: Semantic Segmentation Network for Remote Sensing Images Based on Category-Guided Global-Local Feature Interaction." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–17. https://doi.org/10.1109/TGRS.2024.3379398.

Noda, H., M. N. Shirazi, and E. Kawaguchi. 2002. "MRF-based Texture Segmentation Using Wavelet Decomposed Images." *Pattern Recognition* 35 (4): 771–782. https://doi.org/10.1016/S0031-3203(01)00077-2.

Pan, C., X. Jia, J. Li, and X. Gao. 2020. "Adaptive Edge Preserving Maps in Markov Random Fields for Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 59 (10): 8568–8583. https://doi.org/10.1109/TGRS.2020.3035642.

Pont-Tuset, J., P. Arbelaez, J. T. Barron, F. Marques, and J. Malik. 2016. "Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (1): 128–140. https://doi.org/10.1109/TPAMI.2016.2537320.

Rice, L., E. Wong, and Z. Kolter. 2020. "Overfitting in Adversarially Robust Deep Learning." In *International Conference on Machine Learning (ICML)*, edited by H. Daumé III and S. Aarti, 8093–8104. Virtual: PMLR. https://proceedings.mlr.press/v119/rice20a.html.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, 234–241. Cham: Springer. https://doi.org/10.1007/978-3-319-24574-4_28.

Shen, Y., J. Chen, L. Xiao, and D. Pan. 2019. "Optimizing Multiscale Segmentation with Local Spectral Heterogeneity Measure for High Resolution Remote Sensing Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 157:13–25. https://doi.org/10.1016/j.isprsjprs.2019.08.014.

Shen, D., G. Wu, and H.-I. Suk. 2017. "Deep Learning in Medical Image Analysis." *Annual Review of Biomedical Engineering* 19 (1): 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442.

Shi, J., and J. Malik. 2000. "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8): 888–905. https://doi.org/10.1109/34.868688.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (sup1): 234–240. https://doi.org/10.2307/143141.

Tong, X.-Y., G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang. 2020. "Land-cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models." *Remote Sensing of Environment* 237:111322. https://doi.org/10.1016/j.rse.2019.111322.

Vincent, L., and P. Soille. 1991. "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (6): 583–598. https://doi.org/10.1109/34.87344.

Wang, L., X. Huang, C. Zheng, and Y. Zhang. 2017. "A Markov Random Field Integrating Spectral Dissimilarity and Class co-occurrence Dependency for Remote Sensing Image Classification Optimization." *ISPRS Journal of Photogrammetry and Remote Sensing* 128:223–239. https://doi.org/10.1016/j.isprsjprs.2017.03.020.

Wu, H., P. Huang, M. Zhang, W. Tang, and X. Yu. 2023. "CMTFNet: Cnn and Multiscale Transformer Fusion Network for Remote Sensing Image Semantic Segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 61:1–12. https://doi.org/10.1109/TGRS.2023.3314641.

Wu, H., M. Zhang, P. Huang, and W. Tang. 2024. "CMLFormer: Cnn and Multi-scale Local-Context Transformer Network for Remote Sensing Images Semantic Segmentation." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17:7233–7241. https://doi.org/10.1109/JSTARS.2024.3375313.

Xia, G.-s., C. He, and H. Sun. 2006. "An Unsupervised Segmentation Method Using Markov Random Field on Region Adjacency Graph for SAR Images." In *CIE International Conference on Radar*, 1–4. Shanghai: IEEE. https://doi.org/10.1109/ICR.2006.343148.

Yang, Y., G. Yuan, and J. Li. 2024. "SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation." arXiv preprint arXiv:240501992. doi:10.48550/arXiv.2405.01992.

Zhang, X., Z. Gao, L. Jiao, and H. Zhou. 2017. "Multifeature Hyperspectral Image Classification with Local and Nonlocal Spatial Information via Markov Random Field in Semantic Space." *IEEE Transactions on Geoscience and Remote Sensing* 56 (3): 1409–1424. https://doi.org/10.1109/TGRS.2017.2762593.

Zhang, L., X. Huang, B. Huang, and P. Li. 2006. "A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 44 (10): 2950–2961. https://doi.org/10.1109/TGRS.2006.876704

Zhang, Z., X. Li, H. Li, F. Dunkin, B. Li, and Z. Li. 2024. "Dual-branch Sparse Self-learning with Instance Binding Augmentation for Adversarial Detection in Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–13. https://doi.org/10.1109/TGRS.2024.3436841.

Zhang, C., I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, and P. M. Atkinson. 2019. "Joint Deep Learning for Land Cover and Land use Classification." *Remote Sensing of Environment* 221:173–187. https://doi.org/10.1016/j.rse.2018.11.014.

Zhang, X., A. Wang, Y. Zheng, S. Mazhar, and Y. Chang. 2024. "A Detection Method with Anti-interference for Infrared Maritime Small Target." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17:3999–4014. https://doi.org/10.1109/JSTARS.2024.3357496.

Zhao, W., S. Du, Q. Wang, and W. J. Emery. 2017. "Contextually Guided Very-High-Resolution Imagery Classification with Semantic Segments." *ISPRS Journal of Photogrammetry and Remote Sensing* 132:48–60. https://doi.org/10.1016/j.isprsjprs.2017.08.011.

Zheng, C., Y. Chen, J. Shao, and L. Wang. 2021. "An MRF-Based Multigranularity Edge-Preservation Optimization for Semantic Segmentation of Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/LGRS.2021.3058939.

Zheng, C., and L. Wang. 2015. "Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model with Regional Penalties." *IEEE Journal of Selected Topics in*

Applied Earth Observations and Remote Sensing 8 (5): 1924–1935. https://doi.org/10.1109/JSTARS.2014.2361756.

Zheng, C., Y. Zhang, and L. Wang. 2016. "Multilayer Semantic Segmentation of Remote-Sensing Imagery Using a Hybrid Object-Based Markov Random Field Model." *International Journal of Remote Sensing* 37 (23): 5505–5532. https://doi.org/10.1080/01431161.2016.1244364.

Zheng, C., Y. Zhang, and L. Wang. 2017. "Semantic Segmentation of Remote Sensing Imagery Using an Object-Based Markov Random Field Model with Auxiliary Label Fields." *IEEE Transactions on Geoscience and Remote Sensing* 55 (5): 3015–3028. https://doi.org/10.1109/TGRS.2017.2658731.

Zhou, Q., W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki. 2019. "Multi-scale Deep Context Convolutional Neural Networks for Semantic Segmentation." *World Wide Web* 22 (2): 555–570. https://doi.org/10.1007/s11280-018-0556-3.

Zhu, S., W. Jing, P. Kang, M. Emam, and C. Li. 2023. "Data Augmentation and Few-Shot Change Detection in Forest Remote Sensing." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16:5919–5934. https://doi.org/10.1109/JSTARS.2023.3285389.

Zhu, C., X. Li, C. Wang, B. Zhang, and B. Li. 2024. "Deep Learning-Based Coseismic Deformation Estimation from InSAR Interferograms." *IEEE Transactions on Geoscience and Remote Sensing* 62:1–10. https://doi.org/10.1109/TGRS.2024.3357190.