Sains Malaysiana 55(2)(2026): 209-219
http://doi.org/10.17576/jsm-2026-5502-03

A Review of CNN-Based Typical Urban Land Cover Segmentation Techniques in Multispectral Remote Sensing Imagery

(Suatu Ulasan Teknik Segmentasi Litupan Tanah Bandar Tipikal Berasaskan CNN dalam Imej Penderiaan Jauh Multispektral)

ZHAO HAIMENG[1,2], RAIHANI MOHAMED[1,*] & NG SENG BENG[1]

[1]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*College of Artificial Intelligence, Guilin University of Aerospace Technology, Guilin, Guangxi, 541004, China*

Received: 20 June 2025/Accepted: 28 January 2026

ABSTRACT

Compared with visible-light remote sensing, multispectral remote sensing provides multi-band land surface information and enhances spectral separability through data fusion, thereby enabling more accurate surface representation. However, spectral redundancy, resolution discrepancies, and highly complex urban environments impose greater challenges on existing methods. Deep learning approaches based on convolutional neural network (CNN) offer superior capabilities in extracting and integrating multispectral features, enabling more accurate urban land cover segmentation. This review focuses on pixel-level urban land cover segmentation and systematically summarizes recent advances in deep learning for multispectral remote sensing. First, we emphasize that the rich spectral information and spatial complementarity of multispectral data effectively enhance segmentation performance and alleviate ambiguities caused by the 'same spectrum-different objects' and 'same object-different spectra'. Second, we review 19 publicly available multispectral datasets, highlighting differences in spectral bands, spatial resolution, and application scenarios, and summarize a standardized preprocessing pipeline including radiometric calibration, geometric correction, band normalization, and spectral dimensionality reduction to support reproducibility. Third, we discuss representative spectral-spatial feature extraction and cross-scale context modeling strategies, covering dilated convolution, 3D-2D hybrid structures, dual-branch architectures, and multi-scale enhancement modules. Extensive comparative experiments on ISPRS Potsdam and GID datasets further demonstrate the applicability and performance differences of representative models. Finally, future research trends and directions are discussed, encompassing multi-temporal and multi-scale temporal learning, cross-modal fusion, and the lightweight design of complex models.

Keywords: Convolutional neural network (CNN); multispectral features; remote sensing data; semantic segmentation; surface feature extraction

ABSTRAK

Dibandingkan dengan penderiaan jauh cahaya tampak, penderiaan jauh multispektral menyediakan maklumat permukaan tanah pelbagai jalur dan meningkatkan kebolehbezaan spektral melalui penggabungan data, seterusnya membolehkan perwakilan permukaan yang lebih tepat. Walau bagaimanapun, pertindihan spektral, perbezaan resolusi dan persekitaran bandar yang sangat kompleks menimbulkan cabaran lebih besar terhadap kaedah sedia ada. Pendekatan pembelajaran mendalam berasaskan rangkaian neural konvolusi (CNN) menawarkan keupayaan unggul dalam mengekstrak dan mengintegrasikan ciri multispektral, membolehkan pengasingan litupan tanah bandar yang lebih tepat. Ulasan ini memberi tumpuan pada pengasingan liputan tanah bandar per tahap piksel dan secara sistematik merumuskan kemajuan terkini dalam pembelajaran mendalam untuk penderiaan jauh multispektral. Pertama, kami menekankan bahawa maklumat spektral yang kaya dan pelengkap reruang data multispektral berkesan meningkatkan prestasi pengasingan dan mengurangkan kekeliruan akibat 'spektrum sama-objek berbeza' dan 'objek sama-spektrum berbeza'. Kedua, kami mengkaji 19 set data multispektral yang tersedia secara awam, menyoroti perbezaan dalam jalur spektral, resolusi spasial dan senario aplikasi, serta merumuskan saluran prapemprosesan piawai termasuk kalibrasi radiometrik, pembetulan geometri, normalisasi jalur dan pengurangan dimensi spektral untuk menyokong kebolehulangan. Ketiga, kami membincangkan strategi pengekstrakan ciri spektral-reruang dan permodelan konteks silang-skala, merangkumi konvolusi dilasi, struktur hibrid 3D-2D, seni bina dwi-cabang dan modul peningkatan multi-skala. Uji kaji perbandingan luas pada dataset ISPRS Potsdam dan GID seterusnya menunjukkan keberkesanan dan perbezaan prestasi model wakil. Akhirnya, tren dan arah

penyelidikan masa depan dibincangkan, termasuk pembelajaran temporal berbilang-skala dan berbilang-masa, penggabungan lintas-mod serta reka bentuk ringan bagi model kompleks.

Kata kunci: Ciri multispektral; data penderiaan jauh; pengekstrakan ciri permukaan; pengelasan semantik; rangkaian neural konvolusi (CNN)

INTRODUCTION

Compared with visible-light imagery, multispectral remote sensing integrates multidimensional spectral information with relatively high temporal availability, enabling a more comprehensive characterization of spectral differences and spatial heterogeneity among diverse urban land cover types. This capability is of great significance for improving the precision of urban land cover classification and dynamic monitoring, thereby providing essential data support for urban planning, land-use change analysis, and refined urban management. However, complex urban environments, multi-scale objects, and pronounced spectral heterogeneity pose substantial challenges to traditional methods that rely on handcrafted features (such as support vector machines and random forests), leading to limited robustness and generalization capability. In recent years, deep learning approaches, particularly those based on convolutional neural networks (CNNs), have demonstrated remarkable advantages owing to their powerful hierarchical feature learning ability. CNNs can effectively extract high-level semantic representations and multi-scale spatial structural features, thus, better handling complex urban scenes, nonlinear spectral characteristics, and scale variations. Consequently, they have achieved significant progress in fine-grained semantic segmentation of multispectral remote sensing imagery (Ding et al., 2025; Li et al., 2024; Ramos & Sappa 2024).

This paper targets pixel-level land cover segmentation, a core task in intelligent remote sensing, aiming to enhance the effective utilization of multispectral information within a deep learning framework. To address spectral confusion (i.e., 'same spectrum, different objects' and 'different spectrum, same object'), this paper treats multi-band spectral-spatial fusion as a key technique, focusing on the extraction effect of typical urban features such as roads, buildings, and vegetation.

From the perspective of learning paradigms, remote sensing image segmentation has developed along three major technical routes: fully supervised, semi-supervised, and self-supervised learning. The fully supervised approach relies on large-scale labeled datasets, offering the highest accuracy but incurring significant annotation costs (Jia et al. 2023; Shen et al. 2022; Wang et al. 2024). The semi-supervised approach alleviates the annotation burden by employing pseudo-label generation and consistency regularization (Xue et al. 2025; Yu et al. 2023; Zhu et al. 2022). The self-supervised approach leverages proxy tasks such as contrastive learning and masked image modeling to achieve competitive segmentation performance using minimal manual annotations (He et al. 2022; Muhtar, Zhang & Xiao 2022; Zhang & Wang 2023).

Dilated convolution, depthwise separable convolution, and attention mechanisms are commonly employed to expand the receptive field while reducing computational complexity (Chen et al. 2017; Thisanke et al. 2023; Wang et al. 2022). Given the properties of multispectral imagery, models vary in their emphasis on spectral and spatial feature extraction. Spatial-Spectral 3D-2D CNN (SS-CNN) (Saralioglu & Gungor 2022) employs a 3D-2D hybrid convolution to jointly model spectral and spatial features, though its complexity results in high training costs. Dual-branch structures like Pseudo-Siamese Network (PSNet) and Multispectral Semantic Segmentation Network (MSNet) (Tao et al. 2022; Zheng et al. 2025) extract distinct bands or feature dimensions in parallel, mitigating spectral confusion while imposing higher demands on network design and parameter tuning. The use of feature fusion structure combined with multi-scale mechanism or dual encoding design can help improve the model's ability to express details and semantic information, but it will increase the computational burden. Typical representatives include Spatio-Temporal Attention Agricultural Network (STA-AgriNet) (Lin et al. 2024) and Dual-Encoder Deeplab V3+ Network (DEDNet) (Anandakrishnan, Sundaram & Paneer 2025). The combination of multi-scale structure and context fusion can build a more efficient lightweight segmentation model, especially when processing high-resolution images. According to the feature requirements of specific application scenarios, a reasonable trade-off is made between modeling depth, structural complexity and computational efficiency, as shown in Context-Aware Nested U-Net (ContextNested U-Net) (Ulku 2024).

Compared with traditional red, green, and blue (RGB) remote sensing images, the visible light band (400-700 nm) has obvious limitations in spectral recognition, while multispectral images include extended bands such as near-infrared (700-1300 nm) and short-wave infrared (1300-2500 nm), which significantly improves the ability to distinguish different ground objects such as roads, vegetation, and impermeable surfaces (Gui et al. 2022; Hong et al. 2023). In multispectral remote sensing, the visible light band mainly provides the color and texture characteristics of ground objects, the near-infrared band distinguishes vegetation from artificial surfaces through reflectivity differences, and the short-wave infrared band is more sensitive to the reflectance characteristics of water bodies and moist soil (Han et al. 2025; Wu et al.

2024). Multispectral collaborative observation, leveraging complementary spectral information, facilitates the extraction of more discriminative segmentation features (Yan et al. 2025). For instance, the Normalized Difference Vegetation Index (NDVI) highlights vegetation-covered areas by combining reflectance from the red and near-infrared bands, while the Normalized Difference Water Index (NDWI), derived from green and near-infrared bands, effectively identifies wet or shadowed regions. Leveraging multi-band features and spectral indices, the model more accurately distinguishes vegetation, bare soil, water bodies, building rooftops, and other objects in road and background environments, while maintaining high segmentation accuracy in complex scenes (Nagaraj & Kumar 2024; Tao et al. 2022).

Additionally, multispectral data exhibits strong resistance to occlusion. The near-infrared band can partially penetrate sparse foliage, enabling ground object edges occluded by tree canopies to remain visible in infrared imagery, thereby assisting the model in inferring ground object continuity. For regions fully covered by dense vegetation, multispectral imaging helps identify the occlusion itself, reducing missed or false detections of obscured targets (Gui et al. 2022; Hong et al. 2023). In shadowed areas with uneven illumination, near-infrared and short-wave infrared bands - being less influenced by visible light - provide discriminative cues independent of brightness, enabling robust target identification under shadow conditions.

Nevertheless, the high dimensionality of multi-band data, cross-sensor resolution inconsistencies, and registration errors in multi-source data present significant challenges for the efficient utilization of multispectral information. Effectively integrating multi-scale features and long-range dependencies while maintaining computational efficiency remains a critical challenge (Li et al. 2021; Wang et al. 2021). High-quality datasets serve as the foundation for algorithm evaluation and benchmarking. Currently, public datasets such as GID and EuroSAT include multi-platform and multi-temporal remote sensing imagery ranging from satellite to aerial sources; however, limitations remain in terms of geographical coverage, cross-modal consistency, and representation of extreme scenarios. Establishing a systematic and comprehensive preprocessing pipeline - including radiometric and geometric correction, noise reduction, band normalization, and data augmentation - can substantially enhance the generalization ability of segmentation models (Ramos & Sappa 2024).

Based on this background, this paper systematically reviews typical multispectral remote sensing image segmentation technologies. The main contributions are as follows: (1) Method Review: This paper focuses on fully supervised deep learning methods for multispectral remote sensing image segmentation. It systematically analyzes mainstream network architectures based on CNNs, with particular emphasis on the design principles and advantages of representative models such as SS-CNN and MSNet in extracting and utilizing multispectral features. (2) Data-Oriented Analysis: The physical significance of multispectral bands, along with preprocessing and fusion-based dimensionality reduction strategies, is summarized. Additionally, 19 mainstream public datasets are compared and analyzed in terms of data types, spatial resolution, coverage, and application scenarios. (3) Unified Experimental Framework: Six representative segmentation models (DeepLab V3+, UNet, MSNet) are trained and evaluated on two large-scale, high-resolution datasets - ISPRS Potsdam and GID. Their performance is quantitatively compared in terms of F1 Score, mean intersection-over-union (mIoU), and inference-time parameter count, establishing a reproducible baseline for future research. (4) Future Outlook: Focusing on hot topics such as multi-scale and multi-temporal, cross-modal fusion, and lightweight, several research directions are proposed.

MULTISPECTRAL REMOTE SENSING DATA RESOURCES AND PROCESSING STRATEGIES

This section examines spectral fusion and dimensionality reduction strategies within deep learning frameworks, emphasizing data preprocessing and feature modeling essential for high-precision segmentation, and highlighting the practical value of multispectral data in complex environments.

MULTISPECTRAL DATASETS

High-resolution, multi-dimensional, and semantically rich datasets help improve the training effect of segmentation algorithms. This paper collects and organizes multi-platform remote sensing image datasets, as shown in Table 1.

Tables 1 provide a comprehensive review of 19 public datasets in terms of data type, spatial scale, geographical coverage, and application scenarios, offering diverse image resources (RGB, multispectral, high-resolution, and multi-temporal) for land object segmentation tasks. The SpaceNet dataset includes subsets for land cover classification, building detection (covering approximately 685,000 buildings across six cities), and road extraction (comprising high-resolution images and road vector data from multiple urban areas). Tong et al. (2023) utilized this dataset for evaluating road segmentation performance. The Onera Satellite Change Detection dataset comprises 24 pairs of Sentinel-2 multispectral images (with spatial resolutions ranging from 10 to 60 m), covering regions including Brazil and parts of Europe. Mo et al. (2023) employed this dataset to validate the performance of dual neural networks in enhancing change detection accuracy. The Gaofen-2 (GF-2) China Road Dataset provides panchromatic and multispectral imagery of the

TABLE 1. Multispectral remote sensing image semantic segmentation datasets on aviation and satellite platforms

Dataset	Data Source	Image Quantity	Dataset	Data Source	Image Quantity
ISPRS Vaihingen	Aviation Platform	33 images	PASTIS	Satellite Platform	2433 images
ISPRS Potsdam	Aviation Platform	38 images	So2Sat LCZ42	Satellite Platform	1000 images
WeedNet	Aviation Platform	465 images	Satlas	Satellite Platform	10000 images
Weedmap	Aviation Platform	10196 images	Five-Billion-Pixels	Satellite Platform	150 images
RIT-18	Aviation Platform	18 images	CalCROP21	Satellite Platform	1839 images
UAV-HSI-Crop	Aviation Platform	443 images	Burned Area Delineation Dataset	Satellite Platform	73 images
SpaceNet	Satellite Platform	10 million images	EvLab-SS	Satellite Platform	60 images
Gaofen Image Dataset	Satellite Platform	150 images	38-Cloud dataset	Satellite Platform	38 images
EuroSAT	Satellite Platform	27000 images	xBD	Satellite Platform	22068 images
Onera Satellite Change Detection (OSCD)	Satellite Platform	24 pairs of images			

Chinese road network with a spatial resolution of 0.8 m. Du et al. (2023) utilized it to validate their MSI-guided segmentation network. The ISPRS dataset offers remote sensing data from two German cities - 9 cm resolution images of Vaihingen and 5 cm resolution aerial images of Potsdam - for semantic segmentation tasks. China's GF-2 and Gaofen-3 (GF-3) satellites provide both optical and SAR imagery with resolutions ranging from 1 to 4 m, supporting applications such as road extraction. IKONOS and GeoEye-1 satellites offer sub-meter resolution images for high-precision remote sensing analysis. Sentinel-2 satellites deliver multispectral data with spatial resolutions of 10 to 60 m for land cover classification.

SPECTRAL FUSION TECHNOLOGY

In the field of remote sensing image segmentation, commonly employed multispectral information fusion techniques include multi-band stacking, principal component analysis (PCA), feature-level fusion, and spectral dimensionality reduction with information preservation.

Multi-band superposition is a pixel-level spectral fusion technique that constructs multi-channel images by stacking multiple spectral bands. This approach preserves most of the original spectral information; however, the increased data dimensionality introduces redundancy, placing greater demands on the inference efficiency of segmentation algorithms. PCA transforms the correlated spectral bands into a set of uncorrelated principal components through linear transformation and ranks them based on their variance. A small number of principal components can approximate most of the original spectral information, significantly reducing dimensionality while minimizing information loss. PCA removes inter-band correlation redundancy and emphasizes the dominant spectral variations in the image, thereby enhancing segmentation efficiency and accuracy. Feature-level fusion uses a phased information processing mechanism to effectively alleviate the common information redundancy and noise interference problems in multi-source remote sensing data fusion. This method extracts and integrates features from each band of data separately, enhances the expressiveness of each band feature, and thus significantly improves the segmentation performance (Sun et al. 2024). Spectral dimension reduction and fidelity technology refers to the realization of dimensionality compression and noise control through band optimization and minimum noise fraction (MNF) transformation.

DATA PREPROCESSING TECHNOLOGY

Data preprocessing technology provides a standardized process guarantee for improving the accuracy of multispectral remote sensing image segmentation (Bishoff et al. 2023). Radiometric correction effectively eliminates illumination changes and aerosol interference by compensating for pixel value errors, so that the radiation response of the ground object more realistically presents its spectral characteristics. Geometric correction relies on image registration and control point matching technology to correct the spatial distortion and projection error of the image, thereby ensuring the consistency of spatial geometric relationships. Band normalization unifies the scale of the values of each band of the multispectral spectrum to ensure that different bands have balanced weights in segmentation. Image enhancement uses contrast stretching and histogram equalization to highlight the edge and texture information of the ground object and improve the segmentation effect. Noise suppression combines spatial filtering and image smoothing to effectively reduce random errors and imaging noise.

APPLICATION OF DEEP LEARNING IN MULTISPECTRAL REMOTE SENSING

Multispectral image segmentation network architecture based on CNN

Multispectral images are increasingly employed in land cover segmentation, with deep learning methods progressively evolving from the initial processing of three-band RGB data to the integration of multispectral information. With the growing demand for land cover segmentation in environmental monitoring and resource management, traditional machine learning algorithms (K-means) exhibit limitations in processing complex geospatial data, whereas deep learning techniques offer a more promising alternative.

Saralioglu and Gungor (2022) proposed a segmentation model SS-CNN based on a hybrid 3D-2D architecture, which achieved the coordinated optimization of spectral and spatial features and effectively overcame the limitations of traditional 2D networks in spatial context modeling. Lin et al. (2024) adopted a DEDNet network with a parallel coding structure, and alleviated the common problems of feature information loss and semantic expression ambiguity in traditional methods through a cross-level feature fusion strategy. Zheng et al. (2025) proposed a dual-branch independent encoding PSNet model to address the problem of decreased segmentation accuracy caused by multi-band feature aliasing. Its core advantage is that it achieves more refined boundary segmentation while retaining the advantages of each band's characteristics. Ulku (2024) focused on model structure optimization and reasoning efficiency improvement, and designed a lightweight segmentation network ContextNested U-Net, which combines a nested context aggregation module with a lightweight attention mechanism, significantly reducing parameter scale and computational overhead while maintaining accuracy. The Multiscale Spatial-Spectral Fusion Network (MSSFNet) model proposed by Yu et al. (2024) pays special attention to the expressive power of high-dimensional spectral features. The MSSFNet model adds a spectral-feature enhancement mechanism and a multi-scale attention module to achieve feature enhancement, thereby significantly improving the overall segmentation performance.

The spectral-spatial feature collaborative optimization method of SS-CNN is highly representative among the above segmentation technologies. The model integrates the advantages of two-dimensional spatial feature extraction and three-dimensional spectral-spatial joint modeling by constructing a 2D-3D dual-branch structure. This hybrid modeling method effectively enhances the model's ability to segment different types of ground objects and is suitable for multispectral remote sensing image segmentation tasks in complex environments. In addition, SS-CNN introduces a parameterized ReLU (PReLU) activation function to replace the traditional ReLU, further enhancing the representation ability of complex feature patterns. PReLU alleviates the problem of neuron inactivation caused by standard ReLU when negative inputs return to zero by adding a learnable negative semi-axis slope parameter. Its mathematical form is shown in formula (1).

$$PReLU(x_i) = \begin{cases} x_i if x_i > 0 \\ a_i x_i if x_i \le 0 \end{cases} \quad (1)$$

where x_i is the input value and a_i is a learnable parameter that controls the slope of the negative semi-axis. With this mechanism, PReLU can enable the model to retain its learning ability in the negative region, thereby accelerating convergence and improving overall performance. To evaluate the model effect, this paper uses the classification cross entropy loss function, which is very suitable for multi-category classification tasks. Its mathematical expression is shown in formula (2).

$$L(X_i Y_i) = -\sum_{j=1}^{c} y_{ij} * log(p_{ij}) \quad (2)$$

Y_i classification vector $(y_{i1}, y_{i2}, \dots, y_{ic})$.

$$y_{ij} = \begin{cases} 1 \text{ if } i.elementis \in j - class \\ 0 \text{ if not} \end{cases} \quad (3)$$

The SS-CNN model was used to study the Sürmene area in Trabzon, Turkey, utilizing WorldView-2 (WV-2) multispectral satellite imagery, which provides a panchromatic resolution of 0.5 m and a multispectral resolution of 1.84 m. The dataset includes eight spectral bands (R, G, B, NIR1, NIR2, Coastal, Yellow, and Red Edge), covering various land cover types such as crops, roads, and bare soil. Additionally, multispectral satellite data from IKONOS, Pleiades, and Deimos-2 were incorporated as supplementary experimental datasets. To evaluate the segmentation performance of SS-CNN, it was compared with traditional segmentation methods, and the experimental results are presented in Table 2.

Experimental results indicate that, compared to traditional pixel-based methods such as Support Vector Machine (SVM) and Random Forest (RF), the proposed model achieves higher segmentation accuracy across various image sources and land cover types, including WV-2, IKONOS, Pleiades, and Deimos-2. In addition to enhancing the recognition accuracy of land cover categories, the method offers high computational efficiency and low hardware demands, enabling training on standard computing platforms.

Multispectral image segmentation network architecture based on FCN

Alhassan et al. (2020) introduced contextual enhancement modules and adversarial learning mechanisms into traditional fully convolutional networks (FCNs), significantly improving the model's ability to extract

multispectral features and segmentation accuracy. Zhu et al. (2022) integrated spectrally separable modules into FCNs from the perspective of independent modeling of spectral dimensions and constructed a DSSM model, which not only improved reasoning efficiency but also improved segmentation accuracy. Buttar and Sachan (2024) designed a 3D FCN model combined with a coordinate attention mechanism to enhance the model's ability to capture temporal coherence and multispectral features. Through joint modeling of temporal, spectral, and spatial dimensions, the model's semantic consistency and segmentation performance in complex scenes were improved. The STA-AgriNet model proposed by Anandakrishnan, Sundaram and Paneer (2025) integrates multi-scale 3D-2D convolution modules to specifically extract spatiotemporal and spectral features from multispectral images, and exhibits better segmentation stability and accuracy in a multi-source data fusion environment.

Among these methods, Alhassan et al. (2020) proposed a representative deep learning framework that significantly improved the performance of traditional FCN in multispectral feature extraction. This framework is based on traditional FCN, integrating the context enhancement module and the adversarial learning mechanism, aiming to improve the model's ability to express multispectral features and segmentation accuracy. The context enhancement module expands the receptive field by dilating convolution, effectively capturing long-distance spatial dependencies, and helps the model understand the contextual connection between different ground objects, thereby improving the segmentation effect. The adversarial learning mechanism introduces a generator-discriminator structure to enhance the robustness of the model. The generator is responsible for semantic segmentation of perturbed samples, and the discriminator determines the difference between its output and the true label. The two networks are jointly optimized using a mixed loss function, which improves the model's generalization and robustness. This study used Landsat5/7 multispectral satellite images to verify the effects of the context module and adversarial network. Tables 3 shows the results of the FCN network when the context and adversarial network modules are not introduced and introduced, respectively.

Experimental results demonstrate that incorporating context modules and adversarial networks into the three FCN architectures significantly enhances classification accuracy. The best-performing model achieved an overall accuracy of 90.46% when both modules were integrated, surpassing the 88.25% accuracy of the baseline model without these extensions. The findings indicate that the context module effectively integrates multi-scale spatial information, while adversarial training enhances the model's ability to capture high-level semantic features. Their combination leads to a substantial improvement in classification performance.

PERFORMANCE COMPARISON OF MULTIPLE MODELS IN MULTISPECTRAL REMOTE SENSING IMAGE SEGMENTATION TASKS

This paper selects two representative publicly available datasets - ISPRS Potsdam and Gaofen Image Dataset (GID) - to conduct comparative evaluations of commonly used semantic segmentation models. The ISPRS Potsdam dataset, released by the International Society for Photogrammetry and Remote Sensing (ISPRS), is a

TABLE 2. Comparison results between SS-CNN and traditional methods on different datasets

Method	WV-2			IKONOS			Pleiades			Deimos-2		
	OA	AA	K	OA	AA	K	OA	AA	K	OA	AA	K
SVM	86.4	—	82.7	78.51	74.02	73.09	76.38	79.01	69.34	78.29	74.25	66.16
RF	89.2	—	86.4	78.12	72.31	71.72	73.86	73.21	65.44	81.58	83.44	71.49
SS-CNN	**95.6**	**94.8**	**95.3**	**87.50**	**86.01**	**84.57**	**87.93**	**89.31**	**83.99**	**94.08**	**95.69**	**91.15**

TABLE 3. Comparative experiment of FCN network with context and adversarial network modules

Method	Baseline			Baseline + Context + Adversarial		
	Global Accuracy	Mean Accuracy	Mean IoU	Global Accuracy	Mean Accuracy	Mean IoU
VGG-16	87.99	81.50	72.19	90.34	83.63	75.36
ResNet-101	88.25	81.92	73.53	90.46	84.14	75.66
GoogleNet	62.13	37.13	29.06	77.71	59.81	49.20

benchmark high-resolution aerial remote sensing dataset with a spatial resolution of 5 cm. It was acquired in 2013 and covers typical urban regions of Potsdam, Germany, with annotations for six major land-cover categories. The GID dataset was constructed by Wuhan University using Gaofen-2 satellite imagery with a spatial resolution of 4 m. It has been collected since 2014 and contains 15 urban land-cover categories under diverse environmental conditions and seasonal periods across different cities in China. These two datasets are selected because they are authoritative public benchmarks, provide diverse scene coverage, differ in spatial resolution, and offer high-quality pixel-level annotations. Moreover, they are widely recognized and extensively used in the academic community, making them well suited for fair, comprehensive, and robust evaluation of segmentation models under complex urban and land-cover scenarios. Sample images from the datasets are presented in Figure 1.

The Potsdam dataset is a widely used benchmark in the field of multispectral remote sensing. It comprises 38 aerial images with a resolution of 6000 × 6000 pixels, capturing typical urban features in Potsdam, Germany, such as dense buildings, roads, and green spaces. It provides panchromatic images with a spatial resolution of 5 cm and multispectral images (including RGB and NIR bands), and is widely employed in urban land cover semantic segmentation research.

The GID is constructed from multispectral data captured by China's domestically developed GF-2 satellite, encompassing typical urban areas in both southern and northern China. It includes 1 m resolution panchromatic and 4 m resolution multispectral images (including RGB and NIR bands), providing a valuable resource for large-scale global land cover segmentation research.

This section focuses on the analysis of DeepLab V3+, Feature Pyramid Network (FPN), Pyramid Scene Parsing Network (PSPNet), UNet, RGB-Thermal Fusion Network (RTFNet), and MSNet models. Based on the Potsdam and GID, the core training parameter settings of each model are summarized. The batch sizes are flexibly adjusted according to the structural characteristics of each model to optimize the utilization of GPU memory resources. The input data of each channel is normalized according to the samples of the multispectral remote sensing dataset before training. The number of training epochs is set to 100; the Adam optimizer is used with an initial learning rate of 0.00001 and a weight decay coefficient of 0.0001.

This paper conducts a comprehensive evaluation of model performance in terms of semantic segmentation accuracy and inference efficiency. Evaluation metrics include Precision, Recall, F1 Score, intersection-over-union (IoU), mIoU, and mean class-wise Precision (mPre). Among them, Precision and Recall rate, respectively, evaluate the accuracy and completeness of each category segmentation; F1 Score is used to evaluate the harmonic mean; IoU is used to evaluate the overlap between the predicted results and the true annotations; mIoU evaluates the overall balanced performance of each category; mPre evaluates the average accuracy of segmentation results of different categories. The formulas for these evaluation metrics are defined as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP} \tag{7}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

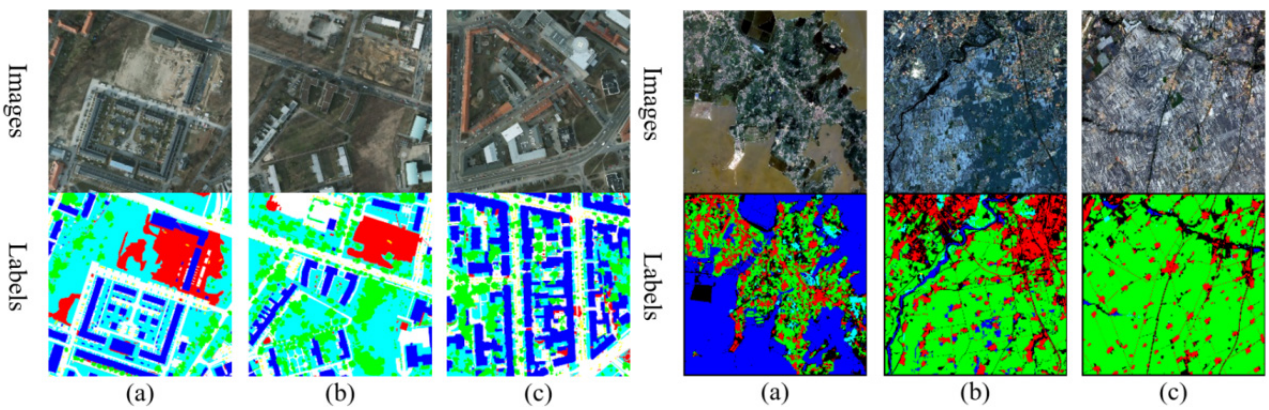

FIGURE 1. Samples from ISPRS Potsdam and GID

$$FWIoU = \frac{TP + FN}{TP + FN + FP + FN} \times \frac{TP}{TP + FP + FN} \quad (9)$$

$$MIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \quad (10)$$

$$MPre = \frac{1}{N}\sum_{i=1}^{N} Precision_i \quad (11)$$

This section comprehensively analyzes the performance indicators of each model on the Potsdam and GID. All models have completed training and testing in the RGB and NIR bands. The core of the evaluation is segmentation performance and computational efficiency. Precision and IoU are used to evaluate the segmentation performance of different ground objects, and F1 Score, mIoU and mPre are used to evaluate the overall performance, so as to comprehensively evaluate the performance of different models in different ground object segmentation categories.

Tables 4 and 5 present the comparative experimental results on the Potsdam and GID. As shown in Table 4, MSNet attained the highest scores of 93.6%, 88.0%, and 93.1% in F1 Score, mIoU, and mPre, respectively. U2-Net achieved the highest recall, scoring 89.7%. Table 5 summarizes the segmentation performance of each model on the GID test set. MSNet obtained the best scores in F1 Score, mPre, and Recall, with values of 91.5%, 90.1%, and 92.9%, respectively. CM-UNet achieved the highest mIoU score of 86.2%.

In summary, different multispectral remote sensing image segmentation models exhibit distinct advantages and limitations with respect to network architecture design and data processing strategies. Lightweight models (UNet and FPN) feature simple architectures and high inference efficiency, making them suitable for resource-constrained or real-time applications. However, they exhibit limitations in multispectral feature extraction and segmentation accuracy. Models based on deep architectures (DeepLab V3+ and PSPNet) excel in semantic representation and context modeling, enabling higher segmentation accuracy. However, their substantial parameter sizes significantly reduce inference speed, limiting applicability in large-scale or real-time scenarios. Models incorporating attention mechanisms or multi-branch structures (RTFNet and MSNet) enhance the spatiotemporal feature modeling of multispectral data while preserving moderate computational efficiency, particularly across multiple evaluation metrics. Some models improve adaptability to multi-source data through remote sensing-specific spectral indices, band grouping, and dual-branch processing strategies, demonstrating

TABLE 4. Evaluation result on the Potsdam Dataset

Model	Low_veg		Tree		Imp_surf		Building		Car		Clutter		F1	mIoU	mPre	Recall	Param
	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU					
DeepLab v3+	80.3	63.3	61.2	48.4	85.3	72.2	89.8	80.9	75.8	62.6	49.4	38.6	75.8	61.0	73.6	78.1	38.4
FPN	92.7	85.7	90.5	83.5	94.7	88.9	97.1	94.5	90.6	82.9	81.4	73.6	91.8	84.9	91.1	92.5	7.1
PSPNet	91.8	84.0	89.8	82.6	93.5	87.2	97.2	94.2	82.0	69.1	79.7	71.9	89.8	81.5	89.0	90.6	32.8
UNet	92.3	85.2	90.0	82.7	94.7	89.0	97.3	94.5	89.8	83.1	80.9	72.9	91.6	84.6	90.8	92.4	6.4
RTFNet	93.3	86.8	91.9	85.6	95.1	90.2	96.9	94.7	90.7	81.8	83.2	72.3	92.0	85.3	91.8	92.2	18.2
MSNet	94.6	88.7	92.3	86.9	95.8	91.5	98.0	95.8	92.8	86.4	85.3	78.6	93.6	88.0	93.1	94.1	9.9

TABLE 5. Evaluation result on the GID

Model	Clutter		Water		Farmland		Meadow		Forest		Built-up		F1	mIoU	mPre	Recall	Param
	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU	Pre	IOU					
DeepLab v3+	79.1	64.0	77.7	70.8	75.5	60.3	29.2	19.7	70.6	58.5	77.9	64.4	72.0	56.3	68.3	76.1	38.4
FPN	92.5	85.2	92.8	88.6	91.9	85.5	75.8	65.7	85.5	81.0	92.4	82.4	89.8	81.5	88.5	91.1	7.1
PSPNet	91.6	83.9	92.6	88.6	91.4	84.0	77.6	65.9	84.2	79.6	91.8	82.5	89.3	80.7	88.2	90.5	32.8
UNet	92.0	83.2	90.8	87.8	90.5	83.5	66.7	56.1	81.3	75.9	90.4	81.0	87.6	77.9	85.3	90.0	6.4
RTFNet	90.2	84.9	95.2	88.5	93.4	85.8	71.4	37.3	87.2	81.2	93.0	80.8	86.6	76.4	88.4	84.9	18.2
MSNet	93.6	87.7	96.4	93.4	93.7	88.2	75.2	66.9	87.4	83.5	94.2	86.1	91.5	84.3	90.1	92.9	9.9

stronger generalization across diverse ground object recognition tasks. Therefore, in practical applications, a trade-off must be made among segmentation accuracy, model complexity, and inference efficiency based on specific task requirements.

CONCLUSION

This paper provides a systematic review of the rapid advancements in deep learning technologies for multispectral remote sensing-based segmentation of typical urban land-cover categories in recent years. It explores key multispectral feature modeling techniques and focuses on analyzing network architectures employing diverse spectral-spatial fusion strategies. Through unified experimental evaluations on representative datasets such as ISPRS Potsdam and GID, this paper verifies the critical role of multispectral information fusion and network structural optimization in improving segmentation performance, while comprehensively analyzing the strengths and limitations of existing approaches. These models have been widely applied in practical tasks such as urban planning and land-use management. In urban planning, multispectral segmentation models enable precise extraction of impervious surfaces, road networks, and building footprints, providing reliable data support for urban spatial structure analysis, land development evaluation, and infrastructure planning. Compared with single visible-light imagery, multispectral data exhibit stronger capability in distinguishing spectrally similar urban materials such as concrete, asphalt, and roofing structures, while CNN-based models further enhance the ability to capture complex textures and boundary features, thereby significantly improving urban land-cover mapping accuracy. In land-use management, multispectral segmentation models facilitate the production of large-scale, high-precision land-cover classification maps for urban expansion monitoring and regional resource management. In particular, models incorporating attention mechanisms, multi-branch architectures, or spectral indices demonstrate enhanced robustness and generalization under heterogeneous environments and varying imaging conditions, supporting multi-scale, cross-regional, and multi-temporal mapping tasks.

When reviewing the relevant literature, this paper confronts challenges arising from multi-source data (such as multispectral images, panchromatic imagery, high-resolution optical data, and even LiDAR and SAR observations), particularly in terms of feature alignment and semantic complementarity across different data sources. Achieving collaborative representation and integrated modeling of multimodal and multi-temporal urban remote sensing imagery can effectively enhance the model's comprehensive understanding of urban spatial, spectral, structural, and temporal characteristics. In addition, data-related issues remain a critical factor constraining the development of urban multispectral segmentation. Significant differences in imaging conditions across urban regions, inconsistencies in spectral responses among sensors, insufficient availability of high-quality pixel-level annotated samples under complex urban environments, and domain shifts across cities and regions can all adversely affect model stability and generalization capability.

With the rapid growth of artificial intelligence and large-scale urban remote sensing data, multispectral semantic segmentation in urban environments still faces multiple challenges. First, modeling temporal consistency and change stability under multi-temporal scenarios remains an open problem in dynamic urban environments. Second, lightweight model design, network pruning, and knowledge distillation remain challenging, making it difficult to simultaneously satisfy high accuracy and real-time processing requirements. In addition, further breakthroughs are needed in software-hardware co-optimization to achieve more efficient computational performance. Meanwhile, with the increasing integration of heterogeneous remote sensing data sources, multimodal fusion introduces new challenges; effectively combining multisensor information to enhance segmentation accuracy and robustness remains a key research direction. Future work should therefore focus on improving multi-scene adaptability, computational efficiency, intelligence, temporal sensitivity, and multimodal integration to further enhance the precision and efficiency of urban remote sensing image analysis and promote broader real-world applications.

ACKNOWLEDGEMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant 62261016, and in part by the Guangxi Key Research and Development Program under Grant AB25069407.

REFERENCES

Alhassan, V., Henry, C., Ramanna, S. & Storie, C. 2020. A deep learning framework for land-use/land-cover mapping and analysis using multispectral satellite imagery. *Neural Computing and Applications* 32: 8529-8544.

Anandakrishnan, J., Sundaram, V.M. & Paneer, P. 2025. STA-AgriNet: A spatio-temporal attention framework for crop type mapping from fused multi-sensor multi-temporal SITS. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18: 1817-1826.

Bishoff, E., Godfrey, C., McKay, M. & Byler, E. 2023. Quantifying the robustness of deep multispectral segmentation models against natural perturbations and data poisoning. *In Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX*, SPIE, 12519: 200-213.

Buttar, P.K. & Sachan, M.K. 2024. Land cover segmentation using 3D FCN-based architecture with coordinate attention. *IEEE Geoscience and Remote Sensing Letters* 21: 2502905.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Ding, L., Hong, D., Zhao, M., Chen, H., Li, C., Deng, J., Yokoya, N., Bruzzone, L. & Chanussot, J. 2025. A survey of sample-efficient deep learning for change detection in remote sensing: Tasks, strategies, and challenges. *IEEE Geoscience and Remote Sensing Magazine* 13(3): 164-189.

Du, Y., Sheng, Q., Zhang, W., Zhu, C., Li, J. & Wang, B. 2023. From local context-aware to non-local: A road extraction network via guidance of multi-spectral image. *ISPRS Journal of Photogrammetry and Remote Sensing* 203: 230-245.

Gui, Y., Li, W., Xia, X.G., Tao, R. & Yue, A. 2022. Infrared attention network for woodland segmentation using multispectral satellite images. *IEEE Transactions on Geoscience and Remote Sensing* 60: 5627214.

Han, Z., Tian, Q., Tian, J., Zhao, T., Xu, C. & Zhou, Q. 2025. Estimation of fractional cover based on NDVI-VISI response space using visible-near infrared satellite imagery. *International Journal of Applied Earth Observation and Geoinformation* 137: 104432.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. 2022. Masked autoencoders are scalable vision learners. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* https://doi.org/10.1109/CVPR52688.2022.01553

Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A. & Zhu, X.X. 2023. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment* 299: 113856.

Jia, J., Song, J., Kong, Q., Yang, H., Teng, Y. & Song, X. 2023. Multi-attention-based semantic segmentation network for land cover remote sensing images. *Electronics* 12(6): 1347.

Li, J., Cai, Y., Li, Q., Kou, M. & Zhang, T. 2024. A review of remote sensing image segmentation by deep learning methods. *International Journal of Digital Earth* 17: 2328827.

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L. & Atkinson, P.M. 2021. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 181: 84-98.

Lin, L., Liu, L., Liu, M., Zhang, Q., Feng, M., Khalil, Y.S. & Yin, F. 2024. DEDNet: Dual-Encoder DeeplabV3+ network for rock glacier recognition based on multispectral remote sensing image. *Remote Sensing* 16(14): 2603.

Mo, W., Tan, Y., Zhou, Y., Zhi, Y., Cai, Y. & Ma, W. 2023. Multispectral remote sensing image change detection based on twin neural networks. *Electronics* 12(18): 3766.

Muhtar, D., Zhang, X. & Xiao, P. 2022. Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* 60: 4411511.

Nagaraj, R. & Kumar, L.S. 2024. Extraction of surface water bodies using optical remote sensing images: A review. *Earth Science Informatics* 17(2): 893-956.

Ramos, L. & Sappa, A.D. 2024. Multispectral semantic segmentation for land cover classification: An overview. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17: 14295-14336.

Saralioglu, E. & Gungor, O. 2022. Semantic segmentation of land cover from high resolution multispectral satellite images by spectral-spatial convolutional neural network. *Geocarto International* 37: 657-677.

Shen, X., Weng, L., Xia, M. & Others. 2022. Multi-scale feature aggregation network for semantic segmentation of land cover. *Remote Sensing* 14: 6156.

Sun, J., Yin, M., Wang, Z., Xie, T. & Bei, S. 2024. Multispectral object detection based on multilevel feature fusion and dual feature modulation. *Electronics* 13(2): 443.

Tao, C., Meng, Y., Li, J., Yang, B., Hu, F., Li, Y., Cui, C. & Zhang, W. 2022. MSNet: Multispectral semantic segmentation network for remote sensing images. *GIScience & Remote Sensing* 59(1): 1177-1198.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R. & Herath, D. 2023. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence* 126(Part A): 106669.

Tong, Z., Li, Y., Zhang, J., He, L. & Gong, Y. 2023. MSFANet: Multiscale fusion attention network for road segmentation of multispectral remote sensing data. *Remote Sensing* 15(8): 1978.

Ulku, I. 2024. ContexNestedU-Net: Efficient context-aware semantic segmentation architecture for precision agriculture applications based on multispectral remote sensing imagery. *Traitement du Signal* 41(5): 2425-2436.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X. & Atkinson, P.M. 2022. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 190: 196-214.

Wang, L., Li, R., Wang, D., Duan, C., Wang, T. & Meng, X. 2021. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sensing* 13(16): 3065.

Wang, Q., Hu, C., Wang, H., Wang, R., Xie, Y. & Zhao, Y. 2024. Semantic segmentation of urban land classes using a multi-scale dataset. *International Journal of Remote Sensing* 45(2): 653-675.

Wu, X., Wang, P., Gong, Y., Zhang, Y., Wang, Q., Li, Y., Guo, J. & Han, S. 2024. Construction and application of dynamic threshold model for agricultural drought grades based on near-infrared and short-wave infrared bands for spring maize. *Remote Sensing* 16(17): 3260.

Xue, Z., Yang, G., Yu, X., Yu, A., Guo, Y., Liu, B. & Zhou, J. 2025. Multimodal self-supervised learning for remote sensing data land cover classification. *Pattern Recognition* 157: 110959.

Yan, Q., Zhang, S., Chen, X. & Zheng, Z. 2025. Multiscale superpixel depth feature extraction for hyperspectral image classification. *Scientific Reports* 15(1): 13529.

Yu, A., Quan, Y., Yu, R., Guo, W., Wang, X., Hong, D., Zhang, H., Chen, J., Hu, Q. & He, P. 2023. Deep learning methods for semantic segmentation in remote sensing with small data: A survey. *Remote Sensing* 15(20): 4987.

Yu, H., Hou, Y., Wang, F., Wang, J., Zhu, J. & Guo, J. 2024. MSSFNet: A multiscale spatial-spectral fusion network for extracting offshore floating raft aquaculture areas in multispectral remote sensing images. *Sensors* 24(16): 5220.

Zhang, W. & Wang, A. 2023. Research on semantic segmentation method of remote sensing image based on self-supervised learning. *International Journal of Advanced Computer Science and Applications* 14(8). https://doi.org/10.14569/IJACSA.2023.0140855.

Zheng, Y., Chen, Z., Zheng, T., Tian, C. & Dong, W. 2025. PSNet: A universal algorithm for multispectral remote sensing image segmentation. *Remote Sensing* 17(4): 563.

Zhu, H., Tan, R., Han, L., Fan, H., Wang, Z., Du, B., Liu, S. & Liu, Q. 2022. DSSM: A deep neural network with spectrum separable module for multi-spectral remote sensing image segmentation. *Remote Sensing* 14(4): 818.

*Corresponding author; email: raihanimohamed@upm.edu.my