Land Cover Classification for Remote Sensing Images Based on MCM-Net

Peilong SHI, Shuxin YIN*

College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

Abstract A novel CNN-Mamba hybrid architecture was proposed to address intra-class variance and inter-class similarity in remote sensing imagery. The framework integrates: (1) parallel CNN and visual state space (VSS) encoders, (2) multi-scale cross-attention feature fusion, and (3) a boundary-constrained decoder. This design overcomes CNN's limited receptive fields and ViT's quadratic complexity while efficiently capturing both local features and global dependencies. Evaluations on LoveDA and ISPRS Vaihingen datasets demonstrate superior segmentation accuracy and boundary preservation compared to existing approaches, with the dual-branch structure maintaining computational efficiency throughout the process.

Key words Semantic segmentation; Remote sensing images; CNN; Mamba

DOI:10.19759/j.cnki.2164-4993.2025.05.008

Semantic segmentation of high-resolution remote sensing imagery is challenged by intra-class variance and inter-class similarity[1-3]. Conventional CNNs capture local patterns efficiently but have a limited receptive field, while Transformer-style models capture global context at the cost of higher quadratic complexity[4-5]. To address these issues, we proposed MCM-Net, which integrates a CNN branch (local detail) and a VSS/Mamba branch (global context) and fuses them using MSCA. A boundary-aware decoder further sharpens object contours. Experiments on LoveDA and Vaihingen demonstrate state-of-the-art or competitive performance with balanced accuracy across categories. Contributions: (1) A CNN-VSS dual-branch network simultaneously extracts local details and global semantic features; (2) A Multi-Scale Spatial-Channel Attention (MSCA) fusion mechanism optimizes cross-branch feature integration through dual spatial-channel attention; and (3) Progressively decoding the output information from each branch to restore resolution and performing weighted summation (WS) for feature fusion. Notably, the MSCA mechanism significantly enhances multi-scale feature representation via parallel processing of spatial relationships and channel dependencies, enabling precise segmentation of complex geographical objects. Experimental results demonstrate the framework's superior performance across diverse remote sensing scenarios.

Materials and Methods

A dual-branch encoder integrates ResNet18 (local features) and VSS (long-range dependencies) for remote sensing segmentation. The VSS branch employs patch embedding and a 2D Selective Scan (SS2D) module that: (1) scans features along four spatial directions, (2) dynamically weights multi-directional information via Selective Scan Mechanism (S6), and (3) reconstructs 2D feature maps preserving both local-global contexts.

To address feature discrepancies between CNN-VSS branches in remote sensing imagery, a Multi-Scale Spatial-Channel Attention (MSCA) module is proposed (Fig. 4). The module: (1) concatenates ResNet (F1-F4) and VSS (G1-G4) features, (2) performs channel compression and multi-scale fusion using parallel convolutions ($1\times1/3\times3/5\times5$ kernels), and (3) generates spatial-channel attention maps through dual branches (3×3 conv + Sigmoid for spatial; 1×1 conv + Sigmoid for channel) to optimize feature fusion.

In deep learning networks, the loss function plays a pivotal role in precisely quantifying the discrepancy between predicted values and ground truth, which directly reflects the robustness and accuracy of the network model. To this end, this study employed an integrated loss function that combines Binary Cross-Entropy (BCE) loss, Dice loss, and boundary loss[6]. The boundary loss can be expressed as:

$$L_{Edge}=1-2\sum_{i=1}^{N}\hat{y}_i y_i+smooth/\sum_{i=1}^{N}(\hat{y}_i^p+y_i^p)+smooth,$$

where: N represents the total number of samples in the batch; $\hat{y}$ denotes the predicted segmentation probability map; y indicates the ground truth binary mask; and p is the exponential factor (default $p=2$). A smoothing term (default value = 1) is added to prevent division by zero. The final model output is optimized through a weighted combination of BCE loss and Dice loss:

$$L_{Bce}=-\frac{1}{N}\sum_{1}^{N}[y_i\log(\hat{y})+(1-y_i)\log(1-\hat{y}_i)],$$

$$L_{Dice}=1-\frac{2}{N}\sum_{n=1}^{N}\sum_{i=1}^{N}y\frac{\hat{y}_i y_i}{(\hat{y}_i^p+y_i^p)},$$

$$L_{BD}=L_{Bce}+L_{Dice},$$

$$L=L_{BD}+\lambda L_{Edge}.$$

Results and Analysis

Datasets

To validate the model's generalization across diverse scenarios, two benchmark remote sensing datasets were utilized: the LoveDA

Received: July 19, 2025 Accepted: September 23, 2025

Peilong SHI (2000-), male, P. R. China, master, devoted to research about processing of remote sensing image data.

* Corresponding author.

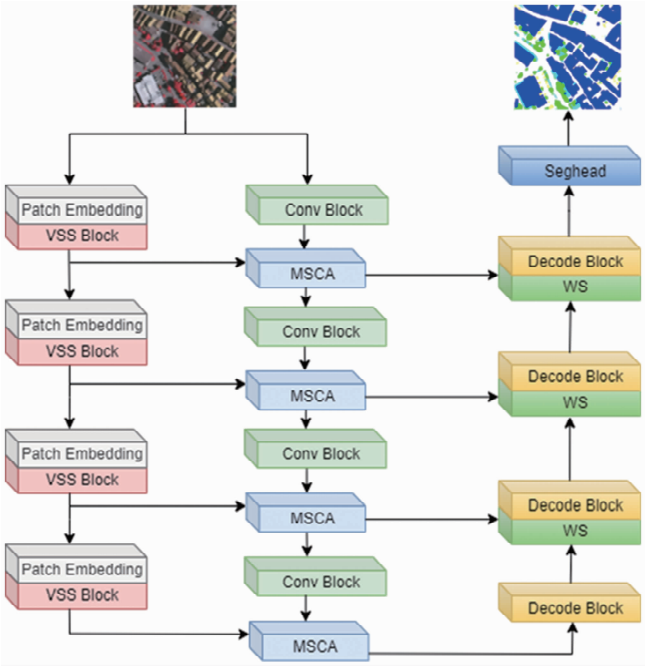

Fig. 1 The overall architecture of MCM-Net

dataset[7], covering urban-rural transitions in Chinese cities (Nanjing, Changzhou, Wuhan) with 0.3 m resolution imagery and seven land-cover classes (*e.g.*, buildings, roads, agricultural areas), split into 2 522 training, 1 669 validation, and 1 796 test images, and the ISPRS Vaihingen dataset[8], featuring ultra-high 9 cm resolution imagery from Germany with six classes (*e.g.*, buildings, cars, trees), comprising 16 images divided into a standard 3∶1 training-test split. These datasets collectively enable comprehensive evaluation under varying resolutions and geographic contexts.

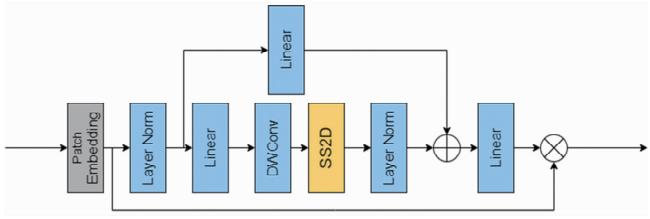

Fig. 2 The detailed architecture of VSS block

Experimental evaluation metrics

ResNet18 served as the backbone, with experiments conducted on an RTX 4090 GPU (24GB) using Python 3.10 and PyTorch 2.1.2. For fair comparison, MCM-Net was evaluated against state-of-the-art models (ABCNet, DeepLabV3+, CMTFNet, UNetformer, RS3Mamba, CM-Net). On LoveDA, AdamW (lr = 0.01, cosine annealing) and data augmentation (random scaling/flipping/rotation) were applied (batch = 4, epochs = 50). On Vaihingen, SGD (lr = 0.01, momentum = 0.9, weight decay = 5e-4, batch = 16, epochs = 50) was used. Performance was measured via mIoU and mF1.

Comparative experiments

Table 1 shows that our method achieves best mIoU on LoveDA, outperforming UNetformer by 2.93% and RS3Mamba by 1.13%. While slightly weaker on barren/forest/farmland classes, it excels in background/buildings/water, achieving overall SOTA performance. Fig. 5's visual comparison (red boxes) demonstrates our model's superior segmentation in complex areas, particularly at urban-rural boundaries. The results validate our architecture's enhanced feature extraction capability while maintaining balanced performance across all categories.

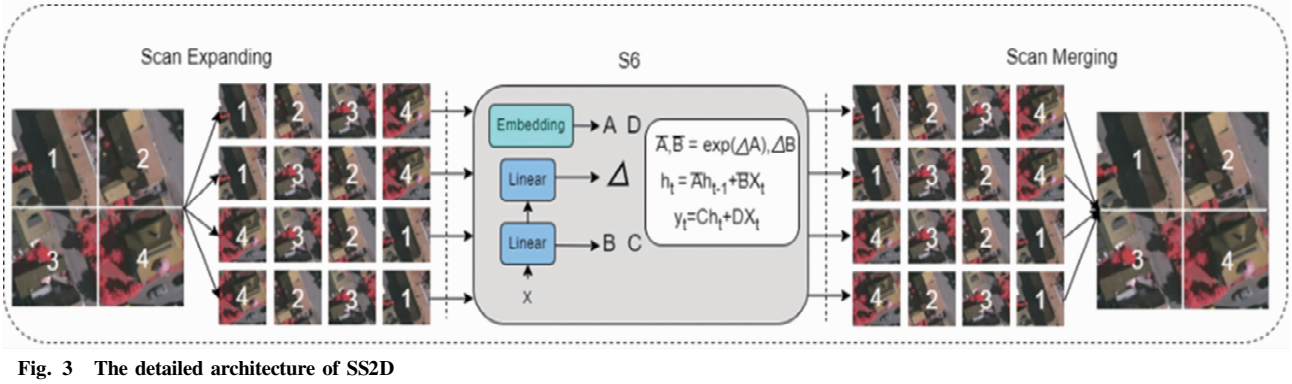

Fig. 3 The detailed architecture of SS2D

Table 1 Experimental results on the LoveDA urban dataset

Model	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU
ABCNet	41.82	56.65	52.42	46.73	28.51	37.54	45.23	45.47
DeepLabV3+	43.04	50.90	52.00	54.40	20.91	34.31	48.51	47.70
UNetformer	37.65	58.80	54.46	63.47	20.21	35.88	46.40	49.11
RS3Mamba	39.72	58.84	57.92	61.00	37.20	39.19	39.73	50.91
CM-Net	44.62	58.69	55.51	64.08	29.62	42.82	50.42	51.78
MCM-Net	45.23	59.93	53.22	66.21	27.53	37.67	41.40	52.04

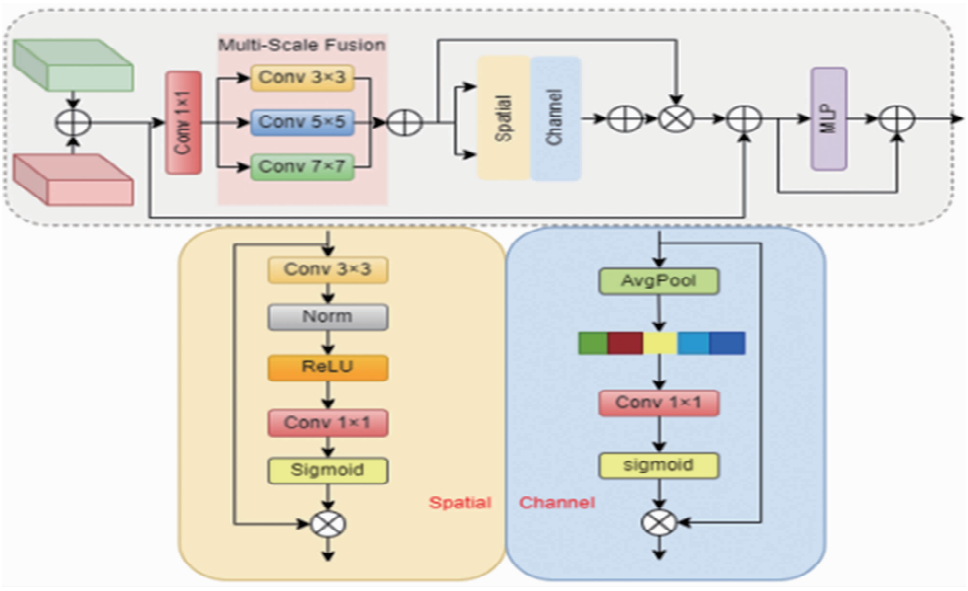

Fig. 4 The detailed architecture of MSCA

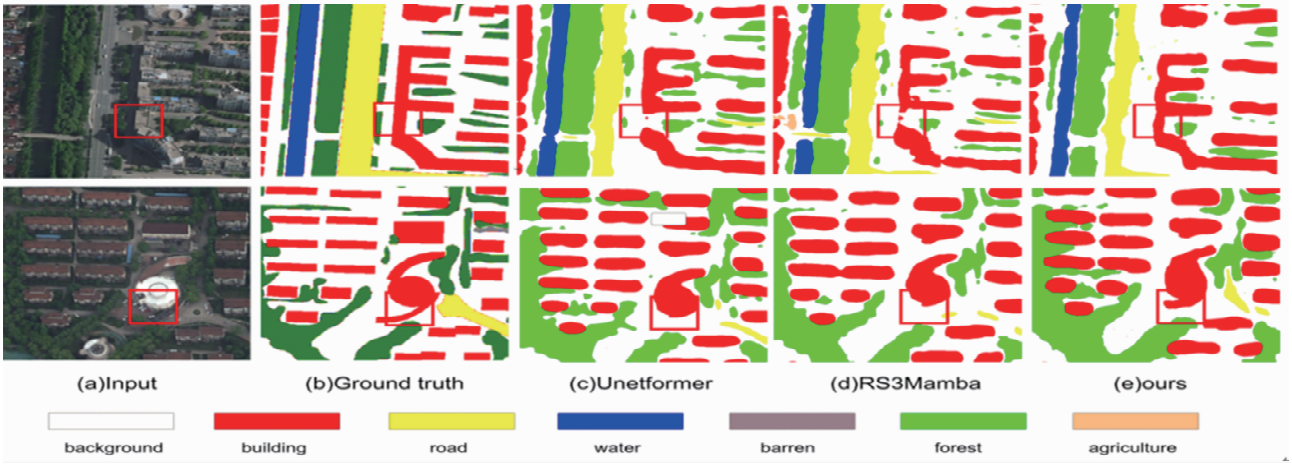

Fig. 5 Qualitative performance comparisons on the LoveDA Urban

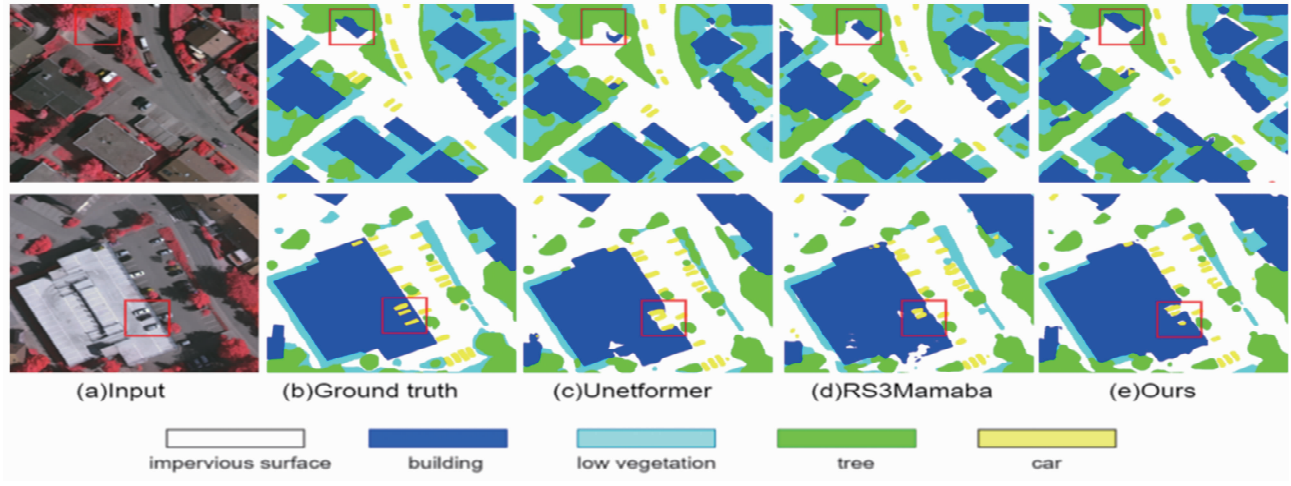

Fig. 6 Qualitative performance comparisons on the ISPRS Vaihaigen

Table 2 shows MCM-Net's performance on Vaihingen, achieving 90.73% mF1 and 83.65% mIoU, surpassing UNetformer by 0.85% (mF1) and 1.66% (mIoU), and RS3Mamba by 0.54% (mF1) and 1.12% (mIoU). Fig. 6's red-boxed comparisons

demonstrate MCM-Net's superior boundary smoothness and object integrity, particularly in complex urban scenes. These quantitative and qualitative results confirm the model's robust segmentation capability on high-resolution imagery.

Table 2 Experimental results on the ISPRS vaihingen dataset

Model	Background	Imp. Surf	Building	Low. Veg	Tree	Car	mF1	mIoU
ABCNet	ResNet18	84.84	91.29	65.42	82.14	72.51	87.39	78.15
DeepLabV3 +	ResNet50	85.21	91.98	66.31	82.31	76.91	88.31	79.51
Unetformer	ResNet18	85.41	92.11	65.46	82.55	80.21	89.88	81.99
RS3Mamba	ResNet18	85.98	93.20	67.11	82.89	81.46	90.19	82.93
CM-Net	ResNet18	86.24	92.48	66.52	82.71	82.21	90.11	83.01
MCM-Net	ResNet18	86.37	93.12	67.23	83.11	82.71	90.73	83.65

Conclusions and Discussion

To address the complex challenges in remote sensing image semantic segmentation, a dual-branch segmentation model based on visual state space architecture was proposed in this study. The MCM-Net framework employs two parallel branches to extract local features (detail preservation) and global contextual representations (long-range dependencies) respectively. Experimental results on LoveDA and Vaihingen benchmarks demonstrate that MCM-Net outperforms state-of-the-art methods. This work provides a new paradigm for fusing local processing and global reasoning in remote sensing segmentation, with potential extensibility to other dense prediction tasks.

References

[1] ZHANG D, WANG F, NING L, *et al.* Integrating SAM with feature interaction for remote sensing change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1 – 11.

[2] LI R, ZHENG S, DUAN C, *et al.* Land cover classification from remote sensing images based on multi-scale fully convolutional network [J]. Geo-spatial Information Science, 2022, 25(2): 278 – 294.

[3] ZHAO L, ZHANG Y, SHI C, *et al.* APNet: A novel antiperturbation network for robust hyperspectral image classification against adversarial attacks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1 – 14.

[4] HE D, SHI Q, LIU X, *et al.* Deep subpixel mapping based on semantic information modulated network for urban land use mapping[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(12): 10628 – 10646.

[5] CHEN,K.; ZOU, Z.; SHI, Z. Building extraction from remote sensing images with Sparse Token Transformers[J]. Remote Sens. 2021, 13: 4441.

[6] CAI J, TAO L, LI Y. CM-UNet + +: A multi-level information optimized network for urban water body extraction from high-resolution remote sensing imagery[J]. Remote Sensing, 2025, 17(6): 980.

[7] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal[M]. Mach. Intell. 2017.

[8] CHEN LC, ZHU Y, PAPANDREOU G, *et al.* Encoder-decoder with atrous separable convolution for semantic image segmentation[J]. in Proc IEEE Eur Conf Comput Vis, 2018: 801 – 818.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU

(Continued from page 37)

RCLA-PLE algorithm not only surpassed other multi-task models but also significantly outperformed all single-task models. Therefore, the RCLA-PLE model can effectively capture relationships within data while better utilizing potential information in different datasets to enhance predictive performance, establishing itself as an efficient multi-task learning algorithm.

References

[1] ORGIAZZI A, BALLABIO C, PANAGOS P, *et al.* LUCAS Soil, the largest expandable soil dataset for Europe: A review[J]. European Journal of Soil Science, 2018, 69(1): 140 – 153.

[2] BARNES RJ, DHANOA MS, LISTER SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra[J]. Applied spectroscopy, 1989, 43(5): 772 – 777.

[3] TANG H, LIU J, ZHAO M, *et al.* Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations[C]//Proceedings of the 14th ACM conference on recommender systems. 2020: 269 – 278.

[4] LECUN Y, BOTTOU L, BENGIO Y, *et al.* Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278 – 2324.

[5] LI Y, ZHU Z, KONG D, *et al.* EA-LSTM: Evolutionary attention-based LSTM for time series prediction[J]. Knowledge-Based Systems, 2019, 181: 104785.

[6] VANDENHENDE S, GEORGOULIS S, VAN GANSBEKE W, *et al.* Multi-task learning for dense prediction tasks: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(7): 3614 – 3633.

[7] PADARIAN J, MINASNY B, MCBRATNEY AB. Using deep learning to predict soil properties from regional spectral data[J]. Geoderma Regional, 2019, 16: e00198.

[8] SINGH S, KASANA SS. Estimation of soil properties from the EU spectral library using long short-term memory networks[J]. Geoderma Regional, 2019, 18: e00233.

[9] YANG J, WANG X, WANG R, *et al.* Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis – NIR spectroscopy[J]. Geoderma, 2020, 380: 114616.

[10] TSAKIRIDIS NL, KERAMARIS KD, THEOCHARIS JB, *et al.* Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network[J]. Geoderma, 2020, 367: 114208.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU