Land Cover Classification of Optical–SAR Imagery via Cross-Modal Interaction and Feature Alignment

Junqi Zhao, Min Chen*, Wei Guo,Jinbo Zhang, Zelan Fu,Xuming Ge, Han Hu, Bo Xu, Qing Zhu,

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, 611756, China

Keywords: Land Cover, Multi-Modal Fusion, Feature Interaction, Feature Alignment

Abstract

Land cover classification (LCC) plays a crucial role in geoscientific research and resource monitoring applications. Compared with traditional single-modal classification methods, multimodal fusion models can more effectively leverage the complementary information of optical and synthetic aperture radar (SAR) imagery, thereby improving classification performance in complex scenarios. However, due to the significant differences in the imaging mechanisms of the two sensors, inconsistencies in radiometric properties and spatial structures arise between optical and SAR images, posing challenges for cross-modal feature interaction and fusion. To address this issue, we propose a multimodal optical–SAR fusion network (MOSFNet) for high-precision LCC, which incorporates two core modules: the Feature Interaction Module (FIM) and the Feature Fusion Module (FFM). The FIM achieves complementary feature interaction between optical and SAR images through channel splitting and cross concatenation, while incorporating a coordinate attention mechanism to enhance the responsiveness of key land cover regions. The FFM leverages a 2D selective scan (SS2D) mechanism to implement bidirectional cross-modal feature alignment and gated fusion in the hidden state space, enabling deep correlation and adaptive integration of optical and SAR features. Experiments on the WHU-OPT-SAR dataset demonstrate that MOSFNet significantly outperforms existing methods in terms of classification accuracy and model generalization, providing an efficient and robust solution for high-precision land cover mapping with multi-source remote sensing imagery.

1. Introduction

Land Cover Classification (LCC) is widely applied in ecological monitoring, agricultural assessment, and urban planning (Li et al., 2020a, Li et al., 2014), with its accuracy directly impacting the reliability of surface process analysis and resource management. In recent years, with the rapid acquisition of multisource remote sensing imagery, the joint analysis of optical and synthetic aperture radar (SAR) data has become an active research topic (Gao et al., 2024, Liu et al., 2025, Liu et al., 2024b). Optical imagery provides rich spectral and textural information, while SAR imagery offers strong penetration capability and robustness against environmental interference. The complementary nature of these two sensing modalities provides new opportunities for achieving high-precision LCC in complex surface environments (Li et al., 2022a).

However, significant heterogeneity exists between optical and SAR imagery due to differences in imaging mechanisms, noise characteristics, and spatial representations. Optical imagery primarily captures the spectral and textural features of land surfaces, whereas SAR imagery characterizes structural and scattering properties. These disparities make cross-modal feature interaction and fusion highly challenging. Under complex conditions, single-modal approaches struggle to maintain both accuracy and robustness. For example, optical imagery tends to be unstable under cloud cover, shadowing, or illumination variations, whereas SAR imagery, despite its all-weather imaging capability, is susceptible to speckle noise and geometric distortions, leading to the loss of texture details and blurred object boundaries.

Early LCC approaches relied on shallow features and conventional machine learning models, such as random forests (Subedi et al., 2023) and maximum likelihood classifiers (Jassoom and Abdoon, 2024), which exhibit limitations in representing multiscale and complex land-cover types (Liu et al., 2024a). With the rapid development of deep learning, convolutional neural networks have been widely applied to remote sensing image interpretation, significantly improving classification accuracy and robustness. Representative architectures include UNet (Ronneberger et al., 2015) and DeepLab (Chen et al., 2018), which enhance spatial continuity and boundary precision through encoder–decoder structures and multiscale feature fusion. In addition, ResUNet, incorporating residual connections and pyramid pooling, has demonstrated excellent performance in high-resolution image segmentation (Diakogiannis et al., 2020).

More recently, the introduction of Transformer-based architectures has further strengthened the modeling of global contextual information. Methods such as Swin-Transformer are capable of capturing long-range dependencies, enabling global feature representation of optical imagery (Zhang et al., 2022). However, the high computational complexity of Transformers limits their practicality for high-resolution remote sensing images. To address this issue, the Mamba architecture, based on State Space Modeling , was proposed to model long-range dependencies with linear complexity (Liu et al., 2024c, Gu and Dao, 2024), demonstrating both high efficiency and expressive power in capturing global semantic relationships (Ma et al., 2024b).

To address the limitations of single-modal approaches, multimodal fusion methods have recently attracted increasing attention, aiming to integrate the complementary information of optical and SAR imagery to improve classification accuracy and robustness. Existing studies primarily include dual-stream networks (DDHRNet) (Ren et al., 2022), collaborative attention gated networks (CHGFNet) (Li et al., 2020b), bilinear feature fusion networks (MBFNet) (Li et al., 2020c), and multimodal

* Corresponding author

interaction methods based on complementary gating modules (Geng et al., 2023). These approaches partially enhance cross-modal feature complementarity. However, significant heterogeneous and nonlinear radiometric differences exist between optical and SAR data, and direct feature fusion often disrupts underlying complementary relationships, reducing the effectiveness of fused features. Moreover, some methods overly emphasize feature complementarity while neglecting structural consistency constraints, resulting in redundant or incomplete semantic information and ultimately limiting classification performance.

To address these challenges, we propose a multimodal optical–SAR fusion network (MOSFNet), designed to achieve more robust and precise land cover classification (LCC). The network incorporates two key modules: the Feature Interaction Module (FIM) and the Feature Fusion Module (FFM). The FIM enables preliminary complementary interaction between optical and SAR features through channel splitting and cross concatenation, enhancing the collaborative representation across modalities. The FFM leverages a 2D selective scan (SS2D) mechanism to establish dynamic cross-modal dependencies in the hidden state space, and employs a bidirectional gating mechanism to achieve deep interaction and adaptive fusion of optical and SAR features while maintaining semantic consistency across modalities. The interacted features are then projected back to the original space via linear mapping and convolution, and fused with the original features through residual connections, yielding enriched multimodal feature representations.

The main contributions of this work can be summarized as follows:

(1)We propose MOSFNet, which leverages cross-modal feature interaction and state-space feature fusion to efficiently integrate optical and SAR information for fine-grained land cover classification.

(2)We design a Feature Interaction Module (FIM) and a Feature Fusion Module (FFM). FIM enables complementary cross-modal feature exchange through channel splitting and cross-concatenation, while FFM employs two-dimensional state-space modeling and a bidirectional gated mechanism to enhance feature consistency and discriminative capability.

(3)We conduct extensive experiments on the WHU-OPT-SAR dataset, which demonstrate that MOSFNet outperforms other methods in both classification accuracy and robustness.

2. Method

The MOSFNet proposed in this paper is designed to fully align the features of optical and Synthetic Aperture Radar (SAR) images and fuse their complementary information, thereby improving the accuracy of LCC in complex surface scenarios. This model adopts an encoder-decoder architecture: the encoder employs two independent ConvNeXtV2-tiny networks as the backbones for feature extraction (Woo et al., 2023), which achieves an improved design of depthwise convolution and layer normalization. While maintaining light weight, this backbone possesses both excellent capability of capturing local fine-grained features and strong performance in global semantic modeling, and it realizes the deep fusion of optical and SAR features via the FIM and FFM. For the decoder, UPerNet is adopted (Xiao et al., 2018), which extracts global contextual information through the pyramid pooling module and generates pixel-level classification results by combining layer-wise feature upsampling and fusion. This decoder can integrate multi-scale semantic information while preserving spatial details, thus enhancing the accuracy of land cover boundary recognition and the capability of category discrimination. The overall network architecture is illustrated in Fig.1.

2.1 Feature Interaction Module (FIM)

After feature extraction by their respective backbone networks, optical and SAR data exhibit strong modality-specific feature representations. However, due to differences in imaging mechanisms, significant feature space shifts remain, and direct fusion may introduce noise and semantic conflicts. To address this, MOSFNet incorporates the FIM immediately after the backbone outputs, serving as an intermediate layer for cross-modal semantic alignment and complementary information exchange. FIM performs channel-level splitting and cross-concatenation, enabling optical features to perceive structural information from SAR while guiding SAR features to focus on optical spectral responses. This facilitates shallow cross-modal feature alignment and dynamic interaction modeling, providing stable and consistent input features for the subsequent FFM. The structure of FIM is illustrated in Fig.2.

First, the optical characteristic $F_i^{opt} \in \mathbb{R}^{C\times H\times W}$ and the SAR characteristic $F_i^{sar} \in \mathbb{R}^{C\times H\times W}$ on the i -th layer are mapped to a $2C$ channel space by convolution. They are then reorganized through channel-wise splitting and interleaved concatenation, where each subblock has a size of $\mathbb{R}^{C/2\times H\times W}$. The reorganized features $F_{opt}^{int} \in \mathbb{R}^{2C\times H\times W}$ and $F_{sar}^{int} \in \mathbb{R}^{2C\times H\times W}$ inject complementary information while maintaining modality-specific independence. This process is formulated in Eq.(1) as follows:

$$\begin{aligned} f_1^{opt}, f_2^{opt}, f_3^{opt}, f_4^{opt} &= Split(Conv(F_i^{opt})) \\ f_1^{sar}, f_2^{sar}, f_3^{sar}, f_4^{sar} &= Split(Conv(F_i^{sar})) \\ F_{opt}^{int} &= Concat(f_1^{opt}, f_2^{sar}, f_3^{opt}, f_4^{sar}) \\ F_{sar}^{int} &= Concat(f_1^{sar}, f_2^{opt}, f_3^{sar}, f_4^{opt}) \end{aligned} \tag{1}$$

where $Conv$ denotes convolution, $Split$ indicates splitting along the channel dimension, and $Concat$ represents concatenation along the channel dimension.

Finally, the interleaved features are further integrated via a convolution to enhance cross-modal correlations. A coordinate attention mechanism is then applied to the fused features to explicitly model horizontal and vertical dependencies, reinforcing geometrically consistent regions while suppressing noise and modality-specific differences, thereby improving spatial consistency. The resulting features are denoted as $\tilde{F}_{OPT_i} \in \mathbb{R}^{C\times H\times W}$ and $\tilde{F}_{SAR_i} \in \mathbb{R}^{C\times H\times W}$, as formulated in Eq.(2).

$$\begin{aligned} \tilde{F}_{OPT_i} &= CA\left(Conv\left(F_{opt}^{int}\right)\right) \\ \tilde{F}_{SAR_i} &= CA\left(Conv\left(F_{sar}^{int}\right)\right) \end{aligned} \tag{2}$$

where CA represents the coordinate attention mechanism.

The obtained interaction-enhanced optical and SAR features, denoted as $\tilde{F}_{OPT_i}$ and $\tilde{F}_{SAR_i}$, are subsequently introduced into the FFM to achieve cross-modal alignment and comprehensive semantic representation.

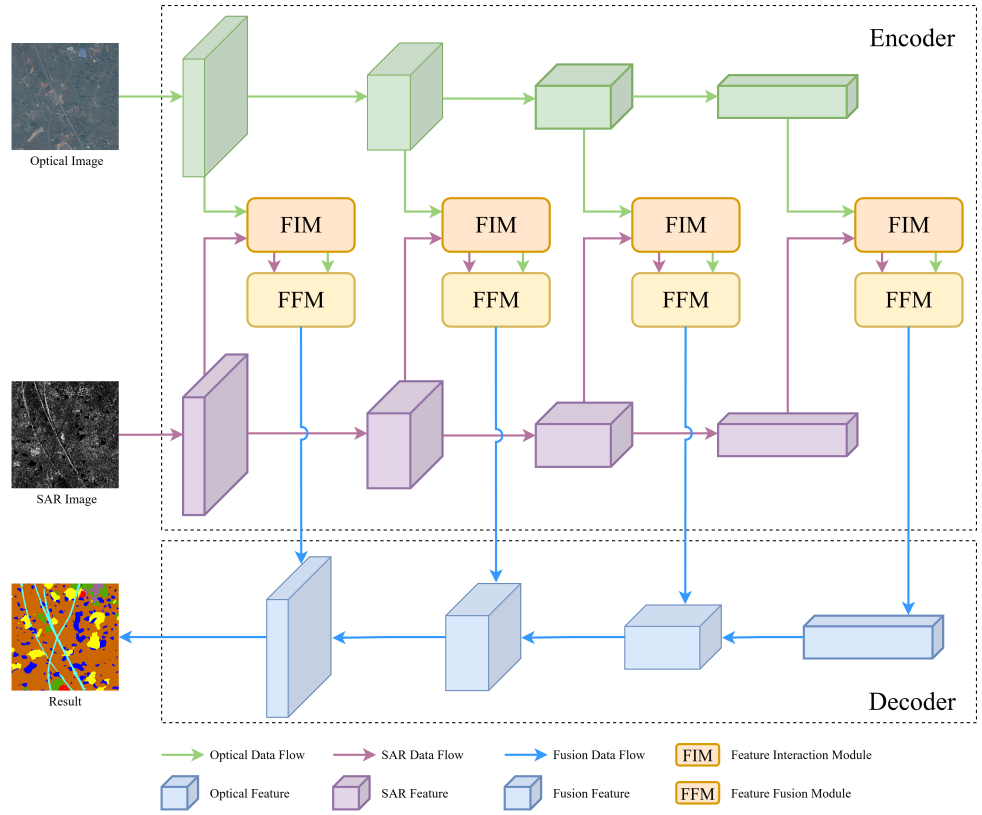

Figure 1. The overall framework of MOSFNet.

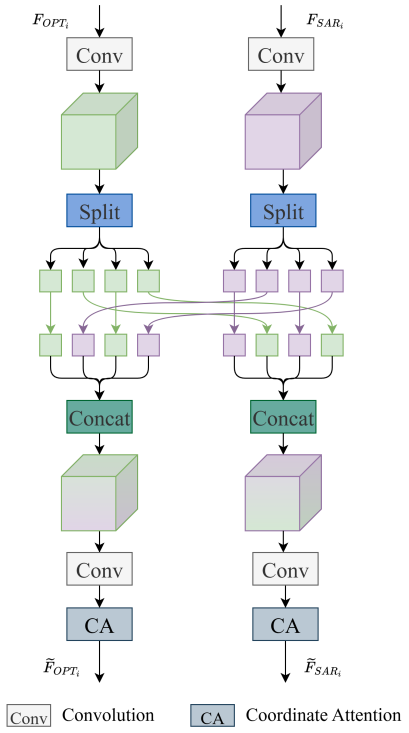

Figure 2. Diagram of the structure of the FIM.

2.2 Feature Fusion Module (FFM)

To fully exploit the complementary information between optical and SAR imagery while suppressing potential spurious features from single modalities, we design the Feature Fusion Module (FFM). The module is built upon the SS2D mechanism and employs a symmetric bidirectional structure to enable cross-modal feature interaction. In one branch, optical features modulate SAR features, while in the other branch, SAR features modulate optical features. A gating mechanism is applied in both branches during the interaction and fusion process to regulate the responses between modalities, enhancing the representation of complementary information while suppressing redundant or spurious features. The interacted hidden state features are projected back to the original space via linear mapping and convolution, and fused with the original features through residual connections. Finally, the outputs from both branches are directly summed to obtain enriched multimodal fusion features, providing the decoder with deeply complementary and adaptively optimized inputs. The structure of the FFM is shown in Fig.3.

Given the bidirectional symmetry of the FFM, we take the branch where SAR features modulate the optical features as an example. The features $\tilde{F}_{OPT_i}$ and SAR feature $\tilde{F}_{SAR_i}$ are fed into the hidden-state representation while being simultaneously projected. This process is formulated in Eq.(3).

$$\begin{aligned} Y_{OPT_i} &= P_i^1(Norm(\tilde{F}_{OPT_i})) \\ Y_{SAR_i} &= P_i^2(Norm(\tilde{F}_{SAR_i})) \\ Z_{OPT_i} &= L_i^1(Norm(\tilde{F}_{OPT_i})) \\ Z_{SAR_i} &= L_i^2(Norm(\tilde{F}_{SAR_i})) \end{aligned} \tag{3}$$

where P_i^1 and P_i^2 represent the operations of projecting features into the hidden-state space, which consist of linear layer, depthwise convolution (DW-Conv), the silu activation function, the SS2D operation, and Layer Normalization. L_i^1 and L_i^2 denote the feature mapping operations composed of linear layer

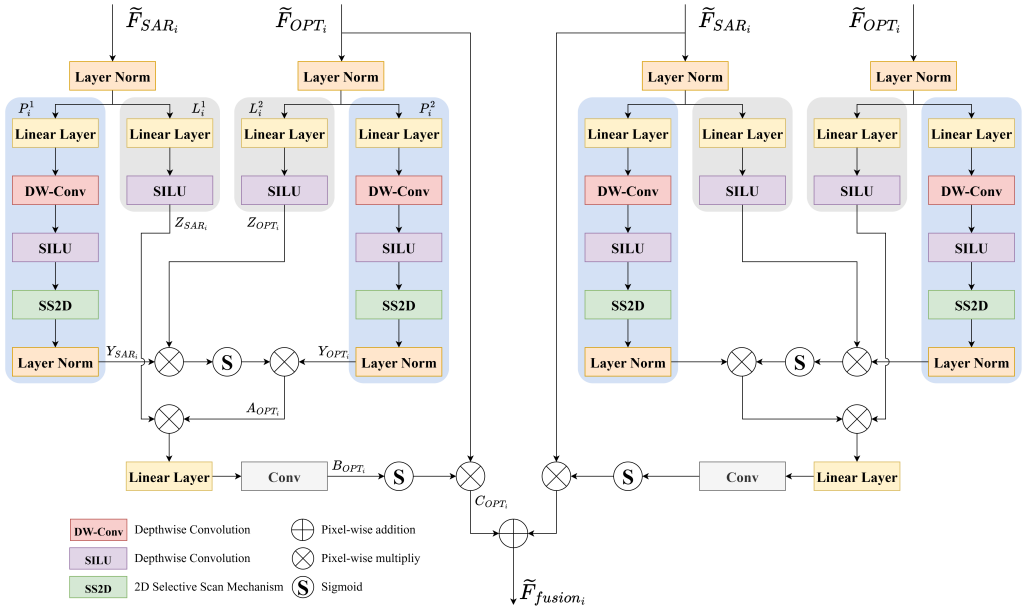

Figure 3. Diagram of the structure of the FFM.

and the silu activation function.

Subsequently, cross-modal feature interaction and fusion are performed, and a gating mechanism is introduced to modulate the response relationships between the two modalities, as formulated in Eq.(4).

$$\begin{aligned} A_{OPT_i} &= Y_{OPT_i} \otimes \sigma(Y_{SAR_i} \otimes Z_{OPT_i}) \\ B_{OPT_i} &= Conv(Linear(A_{OPT_i} \otimes Z_{SAR_i})) \\ C_{OPT_i} &= \tilde{F}_{OPT_i} \otimes \sigma(B_{OPT_i}) \end{aligned} \tag{4}$$

where $\otimes$ is the pixel-by-pixel multiplication, σ is the sigmoid function, and C_{OPT_i} represents the optical features after SAR feature adjustment.

Similarly, for the branch where optical features modulate SAR features, the output can be obtained as C_{SAR_i}. Finally, according to Eq.(5), the two branch outputs are pixel-wise added to generate the final fused feature $\tilde{F}_{fusion_i}$.

$$\tilde{F}_{fusion_i} = C_{OPT_i} \oplus C_{SAR_i} \tag{5}$$

where $\oplus$ denotes the pixel-wise addition operation.

3. Experiment

3.1 Dataset and Experimental Settings

Experiments are conducted on the WHU-OPT-SAR dataset (Li et al., 2022b) to validate the effectiveness and generalization capability of the proposed MOSFNet. We crop the dataset into 512×512 image tiles, of which 7040 tiles are used for training and 1760 for validation.

All experiments are implemented in PyTorch and executed on a single NVIDIA RTX 4090 GPU. We adopt AdamW as the optimizer with an initial learning rate of 1×10^{-4}, momentum parameter of 0.9, and a batch size of 4. Training is run for 100000 iterations. For model performance evaluation, three commonly used metrics are adopted: Mean Intersection over Union (mIoU), Overall Accuracy (OA), and the Kappa Coefficient. These metrics comprehensively assess the accuracy and robustness of the proposed model in multimodal LCC tasks.

3.2 Experimental results

The proposed MOSFNet is compared with several representative methods, including UNet (Ronneberger et al., 2015), MCANet (Li et al., 2022b), and ASMFNet (Ma et al., 2024a). Specifically, UNet (OPT) and UNet (SAR) represent single-modality baselines that take only optical or SAR images as input, respectively. UNet (OPT+SAR) denotes a simple multimodal baseline that concatenates the optical and SAR channels as the input. MCANet and ASMFNet serve as representative multimodal fusion approaches.

Table1 and Fig.4 present the quantitative evaluation and visual comparison of the proposed MOSFNet with other methods on the WHU-OPT-SAR dataset. As shown in Table1, MOSFNet achieves the best performance across all evaluation metrics. In terms of overall accuracy, MOSFNet attains a mIoU of 56.12%, outperforming all compared methods. Specifically, compared with the single-modality baselines U-Net (OPT) and U-Net (SAR), MOSFNet achieves improvements of 2.7% and 11.3%, respectively. It also surpasses U-Net (OPT+SAR)—which simply concatenates optical and SAR inputs without explicit cross-modal interaction—by 2.6%, indicating that input-level fusion alone is insufficient to fully exploit the complementary information between the two modalities. Compared with representative multimodal fusion methods, MCANet and ASMFNet, MOSFNet achieves mIoU gains of 4.4% and 6.2%, respectively. This improvement is mainly attributed to the fact that these two methods do not sufficiently achieve feature alignment and effective fusion between the two modalities. Moreover, MOSFNet attains the highest Overall Accuracy (84.52%) and Kappa coefficient (76.35%), both approximately 2 percentage points higher than those of other methods, further validating the superiority of the proposed model. From a class-wise perspective, MOSFNet achieves the highest IoU for major land-cover categories such as Farmland (70.20%), Water

Table 1. Quantitative evaluation of MOSFNet with other methods on the WHU-OPT-SAR dataset (%).

Model	Per Category IoU							mIoU	OA	Kappa
	Other	Farmland	City	Village	Water	Forest	Road			
Unet (OPT)	19.85	68.21	54.71	48.46	61.81	83.09	37.82	53.42	83.39	74.66
Unet (SAR)	13.79	61.02	52.11	31.67	52.89	80.46	22.09	44.86	79.32	68.00
Unet (OPT+SAR)	19.80	68.02	54.86	47.73	62.85	83.17	38.26	53.53	83.40	74.72
MCANet	17.15	67.05	53.16	46.92	60.45	82.62	34.65	51.71	82.75	73.55
ASMFNet	13.97	66.79	54.59	44.82	59.73	82.42	26.94	49.90	82.51	73.13
MOSFNet (Ours)	**25.05**	**70.20**	**57.21**	**49.48**	**65.48**	**83.85**	**41.54**	**56.12**	**84.52**	**76.35**

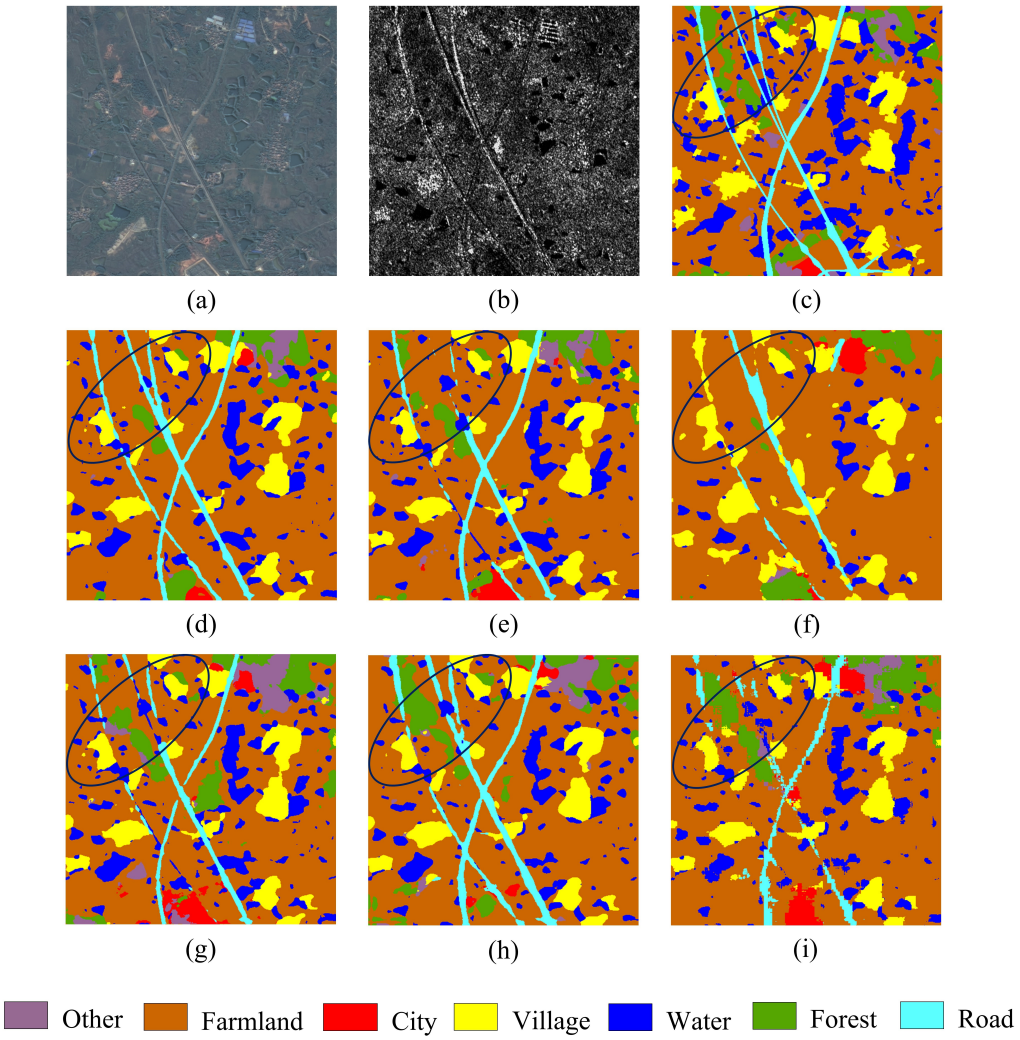

Figure 4. Visualization results of MOSFNet and other methods on the WHU-OPT-SAR dataset. (a) Optical image, (b) SAR image, (c) Ground truth, (d) MOSFNet, (e) Unet(OPT), (f) Unet(SAR), (g) Unet(OPT+SAR), (h) MCANet, (i) ASMFNet

(65.48%), and Forest (83.85%). These improvements can be attributed to the proposed FIM, which effectively models the semantic correlations between optical and SAR features, thereby enhancing the fusion of structural and spectral representations for more discriminative land-cover classification.

From the visualization results in Fig.4, it can be observed that the spatial distribution of MOSFNet's classification results is more consistent with the ground truth. Especially in the circled area of Fig.4, its performance in Road classification is outstanding, with complete road contours identified. In contrast, other comparative methods suffer from incomplete recognition or blurred boundaries in this category. MOSFNet can effectively preserve the detailed integrity of roads, which is mainly attributed to the bidirectional feature modulation mechanism of the FFM. This mechanism enables deep interaction and adaptive fusion of cross-modal features in the hidden state space, dynamically suppresses spurious target responses and enhances semantic correlations, thus significantly improving the classification accuracy and robustness in complex scenarios.

Although the MOSFNet proposed in this paper achieves significantly higher overall land cover classification accuracy than other comparative methods, clear limitations can still be observed from the visualization results in Fig.4. In the circled Road-Water intersection area in the figure, local discontinuit-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-609-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 2. Ablation study of FIM and FFM on the WHU-OPT-SAR dataset (%)

FIM	FFM	mIoU	OA	Kappa
×	×	55.27	84.38	76.10
✓	×	55.74	84.44	76.24
×	✓	55.94	84.44	76.23
✓	✓	**56.12**	**84.52**	**76.35**

ies occur in MOSFNet's classification results for water bodies, and some narrow and elongated water bodies are misclassified as farmland and village. This limitation mainly stems from the model's insufficient capture of fine-grained edge features of water bodies during bimodal feature fusion, as well as the incomplete adaptation to the differences in gray features of water bodies between SAR and optical images, which results in the insufficient robustness of feature representation for small-scale water bodies. Nevertheless, the overall classification performance of MOSFNet is still superior to all comparative models. In the future, the classification accuracy of such relevant areas can be further improved by optimizing the fine-grained feature fusion module.

3.3 Ablation Study

To evaluate the impact of FIM and FFM on the classification performance of MOSFNet, we conducted ablation experiments on the WHU-OPT-SAR dataset, with the results presented in Table 2. Here, "✓" denotes the incorporation of the corresponding module and "×" indicates its exclusion, and the MOSFNet without FIM and FFM—which only adopted pixel-wise addition for feature fusion—was used as the baseline model. The results show that compared with the baseline model, the model's mIoU increases by 0.47% when only FIM is added and by 0.67% when only FFM is incorporated. When both FIM and FFM are integrated simultaneously, the model achieves the maximum values of mIoU, OA and Kappa, which are 56.12%, 84.52% and 76.35%, respectively. The experimental results demonstrate that FIM can effectively facilitate cross-modal interaction between optical and SAR features and enhance the complementarity of multi-modal features. The symmetric bidirectional structure of FFM is able to adaptively regulate the cross-modal response relationship between optical and SAR features, which effectively suppresses the interference of spurious target information and captures more discriminative fused features. When the two modules act synergistically, the effects of cross-modal feature interaction and response modulation are superimposed, which further boosts the representational capability of the fused features and thus realizes the optimal improvement in the model's classification performance.

4. Conclusion

In this study, we propose MOSFNet, a multimodal fusion network designed to effectively integrate optical and SAR remote sensing imagery for accurate LCC. The framework incorporates two core modules, the FIM and the FFM, enabling deep cross-modal interaction and adaptive alignment of complementary information. The FIM enhances the discriminative and complementary characteristics of modality-specific features through channel reconstruction and attention modulation, while the FFM introduces SS2D-based dynamic fusion with a bidirectional gating mechanism to achieve robust feature interaction within the hidden state space. Experiments conducted on the WHU-OPT-SAR dataset demonstrate that MOSFNet significantly outperforms existing single-modal and multimodal methods in terms of mIoU, OA, and Kappa, confirming its robustness and generalization ability. Future work will focus on exploring adaptive modality weighting and dynamic feature modeling strategies to further improve the scalability and adaptability of the proposed network under varying imaging conditions and spatial resolutions.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U24A20589.

References

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Diakogiannis, F. I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 94–114.

Gao, G., Wang, M., Zhang, X., Li, G., 2024. DEN: A new method for SAR and optical image fusion and intelligent classification. *IEEE Transactions on Geoscience and Remote Sensing*.

Geng, X., Jiao, L., Li, L., Liu, F., Liu, X., Yang, S., Zhang, X., 2023. Multisource joint representation learning fusion classification for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–14.

Gu, A., Dao, T., 2024. Mamba: Linear-time sequence modeling with selective state spaces. *First Conference on Language Modeling*.

Jassoom, H. H., Abdoon, R. S., 2024. Monitoring LULC Changes in Babil Province for Sustainable Development Purposes Within the Period 2004–2023. *Photogrammetric Engineering & Remote Sensing*, 90(12), 745–753.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022a. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102926.

Li, W., Dong, R., Fu, H., Wang, J., Yu, L., Gong, P., 2020a. Integrating Google Earth imagery with Landsat data to improve 30-m resolution land cover mapping. *Remote Sensing of Environment*, 237, 111563.

Li, X., Lei, L., Sun, Y., Li, M., Kuang, G., 2020b. Collaborative attention-based heterogeneous gated fusion network for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 3829–3845.

Li, X., Lei, L., Sun, Y., Li, M., Kuang, G., 2020c. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1011–1026.

Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022b. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102638.

Li, Z., Jiao, Q., Liu, L., Tang, H., Liu, T., 2014. Monitoring geologic hazards and vegetation recovery in the Wenchuan earthquake region using aerial photography. *ISPRS International Journal of Geo-Information*, 3(1), 368–390.

Liu, C., Sun, Y., Xu, Y., Sun, Z., Zhang, X., Lei, L., Kuang, G., 2024a. A review of optical and SAR image deep feature fusion in semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Liu, C., Sun, Y., Zhang, X., Xu, Y., Lei, L., Kuang, G., 2025. OSHFNet: A heterogeneous dual-branch dynamic fusion network of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 141, 104609.

Liu, R., Ling, J., Zhang, H., 2024b. SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218, 277–293.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y., 2024c. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37, 103031–103063.

Ma, X., Xu, X., Zhang, X., Pun, M.-O., 2024a. Adjacent-scale multimodal fusion networks for semantic segmentation of remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Ma, X., Zhang, X., Pun, M.-O., 2024b. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5.

Ren, B., Ma, S., Hou, B., Hong, D., Chanussot, J., Wang, J., Jiao, L., 2022. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102896.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Subedi, M. R., Portillo-Quintero, C., Kahl, S. S., McIntyre, N. E., Cox, R. D., Perry, G., 2023. Leveraging NAIP imagery for accurate large-area land use/land cover mapping: A case study in central Texas. *Photogrammetric Engineering & Remote Sensing*, 89(9), 547–560.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16133–16142.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., Wang, C., 2022. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–20.