State Space Model-Based Fusion Modulation Network for Multimodal Semantic Segmentation

Yanli Shang, Fanghong Liu, Haowen Zhang, and Qiuze Yu

***Abstract*—Combining synthetic aperture radar (SAR) and optical images for geomorphic feature extraction offers a promising solution. However, the significant difference in data distribution between the two modalities poses a challenge for the fusion mechanism to achieve efficient segmentation. This letter presents a fusion modulation network based on state space model (SSM) (FMamba), designed for multimodal semantic segmentation tasks. The network can effectively fuse the complementary features of the two modalities to enhance the semantic features of the target regions. Specifically, FMamba integrates a dynamic fusion modulation module (DFMM) based on SSM at the encoding stage. This module dynamically fuses and modulates features from two different modalities to facilitate cross-modal feature extraction and semantic alignment. In the decoding phase, the final segmentation map is generated by progressively decoding and reconstructing the multiscale fusion features. Experimental results on the WHU-OPT-SAR and DFC2025 datasets demonstrate the superior semantic segmentation performance of the proposed method.**

***Index Terms*—Dynamic fusion modulation module (DFMM), semantic segmentation, state space model (SSM).**

I. Introduction

SEMANTIC segmentation is an image processing technique that aims to divide an image into regions, each of which represents a specific semantic category. This technique combines low-level visual features with high-level semantic understanding to generate pixel-level labeled segmentation maps. In remote sensing, the accurate segmentation of key categories, such as buildings, roads, water bodies, and vegetation, is essential for both image interpretability and reliable downstream analysis [1].

Optical and synthetic aperture radar (SAR) images are widely used in remote sensing, offering complementary advantages: optical images provide rich spectral details but are sensitive to weather, while SAR ensures all-weather and all-day stability via microwave imaging. The fusion of these modalities for collaborative segmentation is promising yet challenging, due to fundamental differences such as geometric distortions, contrast variations, noise, and scale inconsistencies. In recent years, deep learning has advanced multimodal remote sensing image fusion. Convolutional neural networks (CNNs) are widely used for feature extraction [2], attention mechanisms facilitate heterogeneous feature alignment [3], [4], and Transformers are increasingly adopted for global modeling [5]. Meanwhile, diverse innovative fusion methods continue to emerge. For example, nearest neighbor-based contrastive learning (NNCNet) [6] improves local semantic consistency using neighboring region contrasts. Prototype-based information compensation network (PICNet) [7] enhances intermodal complementarity via a prototype-guided mechanism.

Concurrently, Mamba, an emerging architecture represented by state space model (SSM), is rapidly growing in the vision domain. Unlike self-attention mechanisms, Mamba adopts a recurrent SSM approach that implicitly aggregates historical information through state compression and selective updates, effectively avoiding the global propagation of noise. Research built on the Mamba architecture has made remarkable progress across various domains. For example, Grouped Mamba [8] and 3-D Mamba [9] have been proposed for state modeling in remote sensing image classification, demonstrating significant performance gains. RSMamba [10] and MSFMamba [11] further enhance the global feature representation through dynamic multipath scanning and multiscale multipath strategies, respectively. In the field of medical image segmentation, VM-UNet [12], built upon VMamba [13], employs an asymmetric encoder–decoder architecture to improve the global context modeling. Moreover, in tasks such as remote sensing change detection and remote sensing image understanding, models like CDMamba [14] and DynamicVis [15] integrate both global and local information, enabling the effective capture of spatiotemporal dynamics and enhancing adaptability to complex scenes. These advances highlight the unique capabilities of the Mamba architecture, supporting its broad application and ongoing development in complex visual tasks.

Inspired by the above, we propose a multimodal feature fusion and segmentation method based on an encoder–decoder architecture (FMamba). This method effectively handles heterogeneous data and reconstructs segmentation by fusing hierarchical information. The encoder leverages a dual-branch, four-level CNN architecture to extract multiscale features from both SAR and optical images. After each stage, the dynamic fusion modulation module (DFMM) effectively mitigates the heterogeneous conflicts between SAR and optical images. Specifically, the DFMM begins by extracting shared features through a global modeling block (GMBlock). It then uses SAR and optical features as modulation conditions, applying a feature dynamic modulation block (FDMBlock) to selectively modulate and filter the shared features, ensuring that only the most relevant information is retained. The modulated features are subsequently fused and delivered as prior information,

Received 4 July 2025; revised 27 August 2025; accepted 1 September 2025. Date of publication 9 September 2025; date of current version 1 October 2025. *(Corresponding author: Qiuze Yu.)*

The authors are with the Laboratory of Advanced Detection and Intelligent Perception, School of Electronics and Information, Wuhan University, Wuhan, Hubei 430072, China (e-mail: aaysyl@126.com; 2018302080229@whu.edu.cn; 17361636069@163.com; yuhenry007@whu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2025.3605954

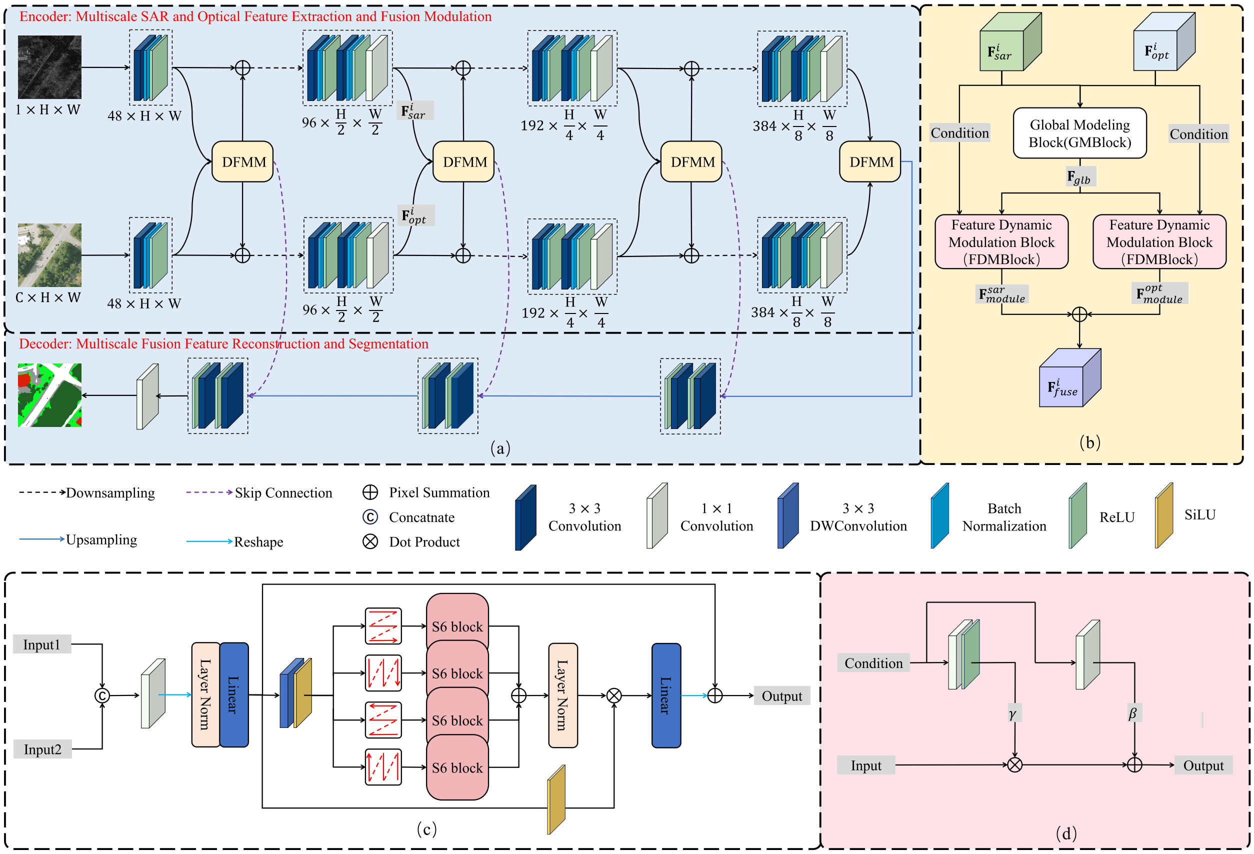

Fig. 1. Detailed structure of the proposed method. (a) Overall structure of the proposed method. (b) DFMM, including the GMBlock and the FDMBlock. (c) Detailed structure of GMBlock. (d) Detailed structure of the FDMBlock.

guiding and constraining the ongoing feature extraction process. Finally, at the decoder stage, the fused modulation results at each stage are reconstructed and segmented with a layered reconstruction method. In summary, our contributions include the following.

1) *FMamba Framework:* We propose the FMamba based on an encoder-decoder architecture, which effectively handles heterogeneous data from SAR and optical images through hierarchical information fusion, improving the segmentation accuracy and robustness.
2) *DFMM:* A design that resolves feature conflicts between SAR and optical images using a GMBlock and an FDMBlock, enhancing fusion feature representation.
3) *Guided Feature Extraction:* Feature expression consistency is improved by using modulation fusion features as a priori information to guide feature extraction in subsequent stages.

II. Methodology

A. Overview

1) Overall Framework: In Fig. 1(a), the proposed encoder–decoder network is designed for fusing and segmenting SAR and optical images. The encoder comprises a CNN-based multimodal feature extractor and a DFMM [see Fig. 1(b)], while the decoder uses a multilayer CNN to reconstruct fused features and generate segmentation results. Two independent multistage CNN branches are employed to extract multiscale features from SAR and optical images, respectively. At each stage, features from the two modalities are first extracted by CNN modules and subsequently fused and modulated through a DFMM, which consists of a GMBlock, [see Fig. 1(c)] and an FDMBlock, [see Fig. 1(d)]. The modulation outputs are propagated as prior information to the next stage, thereby guiding and constraining the feature extraction of both branches to ensure consistency in representation. For each stage $i \in \{0, 1, 2, 3\}$, the formulation is defined as follows:

$$\mathbf{F}^i_{\text{sar}} = \begin{cases} \text{Encoder\_}CNN^i_{\text{sar}}\left(\mathbf{I}_{\text{sar}}\right), & i = 0 \\ \text{Encoder\_}CNN^i_{\text{sar}}\left(\mathbf{F}^{i-1}_{\text{sar}} + \mathbf{F}^{i-1}_{\text{fuse}}\right), & i \in \{1, 2, 3\} \end{cases} \tag{1}$$

$$\mathbf{F}^i_{\text{opt}} = \begin{cases} \text{Encoder\_}CNN^i_{\text{opt}}\left(\mathbf{I}_{\text{opt}}\right), & i = 0 \\ \text{Encoder\_}CNN^i_{\text{opt}}\left(\mathbf{F}^{i-1}_{\text{opt}} + \mathbf{F}^{i-1}_{\text{fuse}}\right), & i \in \{1, 2, 3\} \end{cases} \tag{2}$$

where $\mathbf{I}_{\text{sar}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{I}_{\text{opt}} \in \mathbb{R}^{C\times H\times W}$ represent SAR and optical images, respectively. $\mathbf{F}^i_{\text{sar}}$ and $\mathbf{F}^i_{\text{opt}}$ are the feature maps extracted from the SAR and optical images, respectively. At each stage, these two features are fused, and their (SAR and optical) contributions are adjusted during training according to the importance of these two features in the network. The fusion operation is formalized as follows:

$$\mathbf{F}^i_{\text{fuse}} = DFMM\left(\mathbf{F}^i_{\text{sar}}, \mathbf{F}^i_{\text{opt}}\right), \quad i \in \{0, 1, 2, 3\} \tag{3}$$

where $\mathbf{F}^i_{\text{fuse}}$ denotes the fusion modulation result generated by DFMM.

Algorithm 1 Dynamic Feature Modulation Fusion Module

```
Input  : $\mathbf{F}_{sar}, \mathbf{F}_{opt} \in \mathbb{R}^{C\times H\times W}$
Output: $\mathbf{F}_{fuse} \in \mathbb{R}^{C\times H\times W}$
$\mathbf{F}_{concat} \leftarrow$ Conv(Concat($\mathbf{F}_{sar}, \mathbf{F}_{opt}$));//Feature concatenation
$\mathbf{F}_{x}, \mathbf{F}_{z} \leftarrow$ Linear(LayerNorm(Reshape($\mathbf{F}_{concat}$)));//Linear mapping
$\mathbf{F}_{dw} \leftarrow$ SiLU(DWConv(Reshape($\mathbf{F}_{x}$)));//Depthwise conv + activation
$\{\mathbf{S}^{i}_{dir}\}^{4}_{i=1} \leftarrow$ ViewSeq($\mathbf{F}_{dw}$);//Directional sequences
for $i = 1$to4 do
    $\mathbf{F}_{i} \leftarrow \text{S6Block}_{i}(\mathbf{S}^{i}_{dir})$
end
$\mathbf{F}_{merged} \leftarrow \text{Reshape}\left(\text{LayerNorm}\left(\sum_{i=1}^{4} \mathbf{F}_{i}\right)\right)$;//S6 merging
$\mathbf{F}'_{linear} \leftarrow$ Linear($\mathbf{F}_{merged}$) $\cdot$ SiLU($\mathbf{F}_{z}$);//Feature enhancement
$\mathbf{F}_{glb} \leftarrow$ Reshape($\mathbf{F}'_{linear}$) $+ \mathbf{F}_{concat}$;//Global feature construction
$\mathbf{F}^{sar}_{module} \leftarrow$ FDMBlock($\mathbf{F}_{glb}, \mathbf{F}_{sar}$)
$\mathbf{F}^{opt}_{module} \leftarrow$ FDMBlock($\mathbf{F}_{glb}, \mathbf{F}_{opt}$);//Dynamic modulation
$\mathbf{F}_{fuse} \leftarrow \mathbf{F}^{sar}_{module} + \mathbf{F}^{opt}_{module}$;//Final fusion
return $\mathbf{F}_{fuse}$
```

Finally, the decoder employs the CNN modules to progressively reconstruct the fused multiscale features and generate the final segmentation map $\mathbf{F}_{\text{seg}}$, as follows:

$$\mathbf{F}_{\text{seg}} = \text{Decoder\_}CNN\left(\mathbf{F}^{i}_{\text{fuse}}\right), \quad i \in \{0, 1, 2, 3\}. \tag{4}$$

2) Loss Function: The network is optimized using cross-entropy loss between the ground truth $\mathbf{F}_{\text{label}}$ and the predicted segmentation $\mathbf{F}_{\text{seg}}$ as follows:

$$\mathbf{L}_{\text{seg}} = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{n=1}^{\text{cls}}\left(\mathbf{F}_{\text{label}}(h, w, n)\log \mathbf{F}_{\text{seg}}(h, w, n)\right). \tag{5}$$

B. Dynamic Fusion Modulation Module

DFMM comprises two main components: GMBlock and FDMBlock. It first obtains fusion results from the two branches via a cascading operation and then applies GMBlock to extract shared features. FDMBlock further dynamically adjusts these features, using SAR and optical features as modulation conditions. The detailed process is outlined in Algorithm 1.

1) Global Modeling Block: In the encoder of FMamba, SAR and optical images are first extracted as initial features by CNN, respectively, and then spliced and fed into GMBlock built on VSSBlock [13] for global modeling. The GMBlock maps the fused features into four directional sequences that are input into the S6 module for multidirectional global feature extraction. The outputs are then fused through summation and linear mapping operations. In addition, a residual connection is incorporated to enhance the stability of information transmission.

2) Feature Dynamic Modulation Block: GMBlock first extracts global features, which are then modulated by FDMBlock to filter out irrelevant information. Specifically, FDMBlock uses SAR and optical features as conditional inputs to generate scaling parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, shifting parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$

$$\boldsymbol{\gamma}_1 = \text{ReLU}\left(\text{Conv}\left(\mathbf{F}_{\text{sar}}\right)\right), \quad \boldsymbol{\beta}_1 = \text{Conv}\left(\mathbf{F}_{\text{sar}}\right) \tag{6}$$

$$\boldsymbol{\gamma}_2 = \text{ReLU}(\text{Conv}(\mathbf{F}_{\text{opt}})), \quad \boldsymbol{\beta}_2 = \text{Conv}(\mathbf{F}_{\text{opt}}). \tag{7}$$

The scaling parameter enhances important channel features such as texture and structure, while the shifting parameter suppresses background noise and extraneous information. The pass target feature $\mathbf{F}_{\text{glb}}$ is linearly transformed using this parameter

$$\mathbf{F}^{\text{sar}}_{\text{module}} = \boldsymbol{\gamma}_1\mathbf{F}_{\text{glb}} + \boldsymbol{\beta}_1 \tag{8}$$

$$\mathbf{F}^{\text{opt}}_{\text{module}} = \boldsymbol{\gamma}_2\mathbf{F}_{\text{glb}} + \boldsymbol{\beta}_2 \tag{9}$$

$$\mathbf{F}_{\text{glb}} = \text{GMBlock}\left(\mathbf{F}^{i}_{\text{sar}}, \mathbf{F}^{i}_{\text{opt}}\right). \tag{10}$$

The two parameters adaptively learn to adjust the efficacy of each channel feature during training, thus filtering out redundant features in $\mathbf{F}_{\text{glb}}$. Finally, the SAR and optical modulation results are fused by a summing operation

$$\mathbf{F}^{i}_{\text{fuse}} = \mathbf{F}^{\text{sar}}_{\text{module}} + \mathbf{F}^{\text{opt}}_{\text{module}}. \tag{11}$$

III. Experiments

We validated our method on the WHU-OPT-SAR (WHU) [4] and DFC2025 datasets [16], evaluating segmentation using five metrics: class accuracy, overall accuracy (OA), mean intersection over union (mIoU), model parameters, and floating point operations (FLOPs). We compared against representative methods from CNN-based (DeepLabV3+ [2] and CMGFNet [17]), attention-based (CCNet [3] and MCANet [4]), transformer-based (CMLFormer [5]), and SSM-based (VM-UNet [12]) categories.

A. Datasets

The WHU dataset contains 100 optical–SAR image pairs from Hubei, China, with seven land use types (e.g., water, urban areas, and roads). These images were cropped into 29 400 nonoverlapping patches of size 256 × 256 pixels and subsequently divided into 80% training and 20% testing samples.

The DFC2025 dataset includes 4300 pairs covering eight land use types (e.g., bare land, trees, and buildings). For few-shot evaluation, 1000 image pairs were randomly selected and cropped into 4000 nonoverlapping patches of size 512 × 512 pixels, which were then partitioned into 20% training and 80% testing samples.

B. Network Training and Experimental Setup

The network was implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU workstation. The input SAR and optical images are normalized to $[0, 1]$. Training used the Adam optimizer with a learning rate of $1e^{-4}$, a batch size of 4 (WHU) and 2 (DFC2025), and 50 epochs for both datasets.

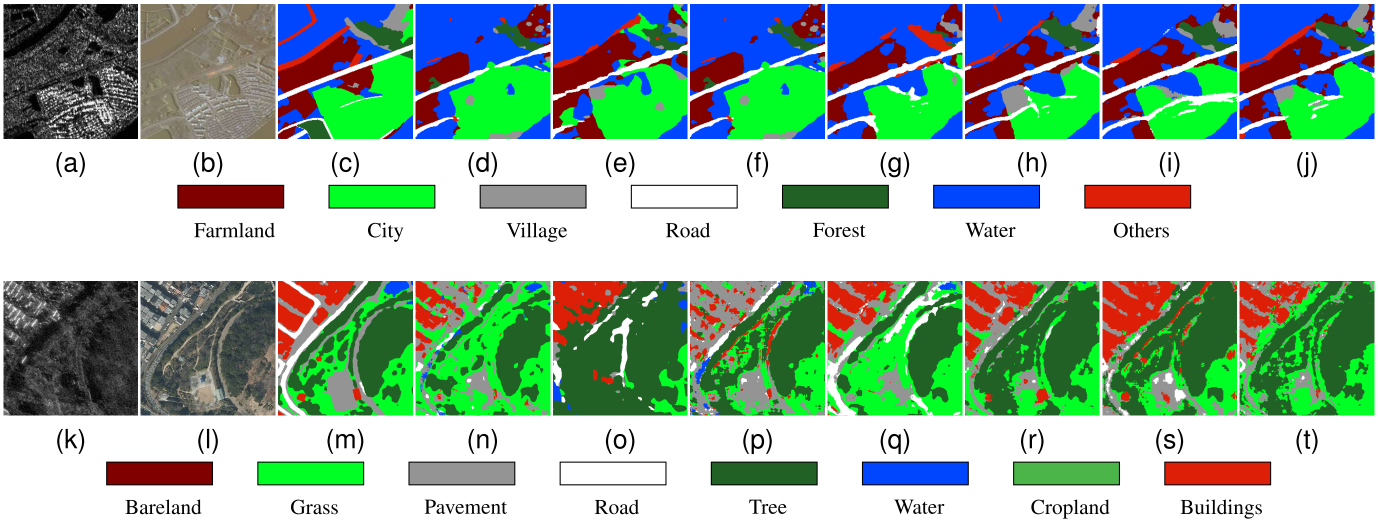

Fig. 2. Visualization results for (top) WHU and (bottom) DFC2025. (a) and (k) SAR. (b) and (l) Optical. (c) and (m) Label. (d) and (n) DeeplabV3+. (e) and (o) CCNet. (f) and (p) MCANet. (g) and (q) CMGFNet. (h) and (r) CMLFormer. (i) and (s) VM-UNet. (j) and (t) FMamba.

TABLE I

Comparison of Different Methods on the WHU Dataset Using an 8:2 Division Between Training and Testing Samples. The Highest Scores Are Highlighted in Bold

Methods	Class Accuracy							OA	mIoU	Params	FLOPs
	Farmland	City	Village	Water	Forest	Road	Others				
DeepLabV3+	0.811	0.718	0.591	0.763	0.887	0.467	0.171	0.798	0.566	26.927M	50.063G
CCNet	0.812	0.719	0.600	0.762	0.890	0.403	0.228	0.802	0.499	49.491M	54.696G
MCANet	0.791	0.720	0.617	0.813	0.901	0.482	0.282	0.811	0.565	55.791M	71.329G
CMGFNet	0.822	0.659	0.636	0.728	0.897	0.536	0.197	0.800	0.447	85.228M	39.048G
CMLFormer	0.823	0.734	**0.666**	**0.831**	0.889	0.535	**0.328**	0.820	0.492	11.736M	49.868G
VM-UNet	0.831	0.696	0.645	0.803	0.894	0.533	0.290	0.820	0.550	38.286M	**10.041G**
FMamba	**0.858**	**0.758**	0.656	0.798	**0.906**	**0.543**	0.306	**0.830**	**0.579**	**9.765M**	34.157G

TABLE II

Comparison of Different Methods on the DFC2025 Dataset Using a 2:8 Division Between Training and Testing Samples. The Highest Scores Are Highlighted in Bold

Methods	Class Accuracy								OA	mIoU
	Bareland	Grass	Pavement	Road	Tree	Water	Cropland	Buildings		
DeepLabV3+	0.001	0.781	0.681	0.180	0.634	0.295	0.326	0.537	0.568	0.277
CCNet	0.012	0.798	0.555	0.288	0.548	0.728	0.153	**0.660**	0.595	0.364
MCANet	0.002	0.754	0.525	0.093	0.688	0.551	0.035	0.201	0.500	0.227
GMCFNet	0.000	0.765	0.717	**0.470**	0.638	0.858	0.246	0.645	0.626	0.379
CMLFormer	0.000	**0.809**	0.668	0.374	**0.701**	0.805	0.259	0.522	0.644	0.352
VM-UNet	0.002	0.590	0.665	0.019	0.559	0.847	0.331	0.539	0.630	0.388
FMamba	**0.022**	0.725	**0.737**	0.266	0.665	**0.861**	**0.501**	0.576	**0.660**	**0.426**

C. Experimental Results

Table I reports the quantitative results on the WHU dataset. FMamba achieves the highest performance in major categories, including cropland (85.8%), urban (75.8%), and forest (90.6%). While CMLFormer outperforms in some classes, our method attains the best OA with an OA of 83.0% and mIoU of 57.9%. In terms of efficiency, our method has the lowest parameter count and ranks second in FLOPs, just above VM-UNet. As shown in the top row of Fig. 2, FMamba, alongside CMLFormer and VM-UNet, effectively segments roads, water, and forest regions. Notably, our method yields sharper boundaries compared with others.

The bottom row of Fig. 2 and Table II show results on the DFC2025 dataset under few-shot settings. Due to pseudo-labels and complex feature distributions, all methods exhibit degraded performance, with some classes near-zero accuracy. Despite this, our method achieves superior overall performance (OA = 0.660 and mIoU = 0.426), demonstrating the effectiveness of the GMBlock and FMBlock in mitigating SAR–optical modality conflict. Visual results confirm the overall challenge across methods, highlighting practical limitations in such settings.

D. Parameter Analysis

1) Number of Network Stages: A proper value of the stage will positively affect the results of FMamba. Table III shows the experimental results of setting different stages. From the results, increasing network stages from 3 to 4 significantly improves OA (62.4%→66.0%) with stable mIoU; further

TABLE III

Comparison of Different Numbers of Network Stages on the DFC2025 Dataset, With the Best Results Marked in Bold

stage	OA	mIoU	Params	FLOPs
3	0.624	**0.426**	**2.442M**	**23.641G**
4	0.660	**0.426**	9.868M	34.283G
5	**0.661**	0.425	39.392M	44.888G
6	0.645	0.381	157.128M	55.476G

TABLE IV

Comparison of Different Numbers of Scan Route Number in SS2D on the DFC2025 Dataset, With the Best Results Marked in Bold

K	OA	mIoU	Params	FLOPs
(1, 1, 1, 1)	0.631	0.372	**9.580M**	**33.299G**
(2, 2, 2, 2)	0.656	0.424	9.747M	33.779G
(4, 4, 2, 2)	0.646	0.401	9.765M	34.157G
(4, 4, 4, 4)	**0.660**	**0.426**	9.868M	34.308G

TABLE V

Ablation Study of Network Modules on the DFC2025 Dataset, With the Best Results Marked in Bold

GMBlock	FMBlock	OA	mIoU	Params	FLOPs
Concat	✓	0.585	0.345	9.028M	31.302G
Add	✓	0.612	0.347	**8.636M**	**30.119G**
✓	✗	**0.664**	0.369	9.082M	31.867G
ResBlock		0.581	0.278	11.376M	38.579G
✓	✓	0.660	**0.426**	9.868M	34.308G

stages bring little gain or degrade performance, making four stages the optimal balance.

2) Number of Scan Route Number in SS2D: Table IV shows the effect of different numbers of scan routes K in SS2D over the four stages of the network. The results show that setting $K = 4$ for each stage yields the best performance (OA 66.0% and mIoU 42.6%). This suggests that segmentation performance can be effectively enhanced by multidirectional modeling, carrying fewer parameters and increased computational overhead. Therefore, $K = 4$ is used as the default configuration for the SS2D module.

E. Ablation Study

Table V lists the results of the ablation experiments for the network modules. Replacing GMBlock and FMBlock with a standard ReBlock causes a sharp performance drop (OA: 58.1% and mIoU: 27.8%) and increased computation, showing their necessity. GMBlock alone significantly boosts accuracy, highlighting its role in semantic enhancement. While FMBlock alone has a limited effect, it complements GMBlock when combined, improving performance with minimal overhead by effectively modulating relevant features.

IV. Conclusion

This letter proposes a fusion modulation network based on SSM, named FMamba, for multimodal semantic segmentation. The method introduces a DFMM to effectively align and fuse multimodal features. In the encoding phase, GMBlock is first used to perform global modeling of the fused SAR and optical features. Then, FDMBlock dynamically adjusts and integrates the common information extracted from the GMBlock, further enhancing the feature representation and improving feature extraction performance. In the decoding phase, a CNN module is employed to progressively reconstruct and complete the segmentation task. Experimental results demonstrate that the proposed method significantly improves segmentation accuracy, validating its effectiveness and superiority. In future work, we will further explore strategies to enhance the model's generalization ability under data-limited scenarios, aiming to address the challenges posed by the scarcity of multimodal remote sensing data.

References

[1] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[3] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[4] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 106, Feb. 2022, Art. no. 102638.

[5] H. Wu, M. Zhang, P. Huang, and W. Tang, "CMLFormer: CNN and multiscale local-context transformer network for remote sensing images semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7233–7241, 2024.

[6] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.

[7] F. Gao et al., "Prototype-based information compensation network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5513615.

[8] Y. He, B. Tu, P. Jiang, B. Liu, J. Li, and A. Plaza, "IGroupSS-mamba: Interval group spatial–spectral mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5538817.

[9] Y. He, B. Tu, B. Liu, J. Li, and A. Plaza, "3DSS-mamba: 3D-spectral–spatial mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534216.

[10] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "RSMamba: Remote sensing image classification with state space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[11] F. Gao, X. Jin, X. Zhou, J. Dong, and Q. Du, "MSFMamba: Multiscale feature fusion state space model for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5504116.

[12] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision mamba UNet for medical image segmentation," 2024, *arXiv:2402.02491*.

[13] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 103031–103063.

[14] H. Zhang, K. Chen, C. Liu, H. Chen, Z. Zou, and Z. Shi, "CDMamba: Incorporating local clues into mamba for remote sensing image binary change detection," 2024, *arXiv:2406.04207*.

[15] K. Chen, C. Liu, B. Chen, W. Li, Z. Zou, and Z. Shi, "DynamicVis: An efficient and general visual foundation model for remote sensing image understanding," 2025, *arXiv:2503.16426*.

[16] J. Xia, H. Chen, C. Broni-Bediako, Y. Wei, J. Song, and N. Yokoya, "OpenEarthMap-SAR: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping," 2025, *arXiv:2501.10891*.

[17] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022.