MCAFNet: Multiscale Cross-Attention Fusion Network for HSI and LiDAR Data Joint Classification

Junhui Cai, Xiaohui Huang, Xiaofei Yang, Jiangtao Peng, *Senior Member, IEEE*, Yifang Ban, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—Hyperspectral images (HSIs) and light detection and ranging (LiDAR) data provide complementary spatial–spectral and elevation information, respectively, which can significantly enhance land cover classification. However, existing methods are often limited by shallow cross-modal interactions and static architectures that struggle with multiscale complex scenes. This article proposes a multiscale cross-attention fusion network (MCAFNet) for joint classification of HSI and LiDAR data. To address the issues of insufficient feature interaction and weak scale adaptability in existing multimodal methods, MCAFNet adopts a three-branch architecture: a LiDAR branch for extracting spatial elevation features, an HSI branch for capturing global spectral features, and a multimodal branch that fuses local features through a multiscale feature enhancement (MSFE) module. The MSFE module extracts spatial–spectral features using depthwise separable and multiscale dilated convolutions, and enhances them via adaptive channel weighting, enabling dynamic multiscale feature perception. To achieve deep feature interaction, a multihead bidirectional cross-attention (MBCA) module is proposed, which realizes deep bidirectional interaction through feature reorganization and cross-branch attention, and introduces triple residual connections to preserve critical information. Extensive experiments conducted on multiple datasets demonstrate that MCAFNet outperforms existing state-of-the-art methods across all evaluation metrics. The source code will be available at https://github.com/2265526/MCAFNet**

***Index Terms*—Hyperspectral image (HSI), joint classification, light detection and ranging (LiDAR), multimodal fusion, multiscale.**

Received 14 February 2026; revised 12 March 2026 and 13 April 2026; accepted 6 May 2026. Date of publication 12 May 2026; date of current version 15 May 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62462031 and Grant 62301174 and in part by Jiangxi Provincial Natural Science Foundation under Grant 20242BAB26023 and Grant 20252BAC220012. *(Corresponding author: Xiaohui Huang.)*

Junhui Cai and Xiaohui Huang are with the School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, China (e-mail: 2024218085405024@ecjtu.edu.cn; hxh016@hotmail.com).

Xiaofei Yang is with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou 511370, China (e-mail: xiaofeiyang@gzhu.edu.cn).

Jiangtao Peng is with the Faculty of Mathematics and Statistics, Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China (e-mail: pengjt1982@hubu.edu.cn).

Yifang Ban is with the Division of Geoinformatics, School of Architecture and the Built Environment, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden (e-mail: yifang@kth.se).

Yicong Zhou is with the Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Digital Object Identifier 10.1109/TGRS.2026.3691799

I. Introduction

REMOTE sensing technology plays a vital role in Earth observation (EO), generating increasingly diverse multisource data through collaborative observations from multiple platforms and sensors. These data find widespread applications in land cover classification [1], resource exploration [2], urban planning [3], and environmental monitoring [4]. With continuous advancements in EO technology, the fusion and analysis of multisource remote sensing data have become a research focus [5], [6].

Hyperspectral images (HSIs) and light detection and ranging (LiDAR) data represent two significant remote sensing data sources [7]. HSI captures hundreds of contiguous narrow spectral bands, enabling fine discrimination of material spectral characteristics, but suffers from low spatial resolution and sensitivity to environmental conditions [8]. In contrast, LiDAR data acquires 3-D elevation information through laser ranging technology, effectively compensating for the limitations of HSI in spatial structure description [9], [10]. Integrating these two data types leverages their complementary advantages, significantly improving land cover classification accuracy [11], [12]. However, fusing HSI and LiDAR data is challenging. Key challenges include the high dimensionality of HSI, the lack of spectral information in LiDAR, and significant variations in object scale within complex scenes [13]. Therefore, effectively integrating their strengths and constructing robust feature representations has become a core research focus in remote sensing [14].

Early research primarily concentrated on convolutional neural network (CNN)-based methods. These approaches significantly improved classification accuracy by designing complex network architectures [15] and incorporating attention mechanisms [16], [17] to extract and fuse multimodal features. To overcome the limitations of CNNs, Transformer models have recently been introduced to remote sensing. By leveraging the powerful global modeling capability of self-attention mechanisms, Transformers have been applied to both HSI classification [18], [19] and HSI–LiDAR fusion [20]. These methods explore intermodal interactions through cross-modal attention [21] or mutual information constraints [22], demonstrating advantages in capturing long-range dependencies.

Recently, the emergence of state-space models (SSMs), represented by Mamba [23], offers a new paradigm for addressing Transformer computational efficiency bottlenecks. Mamba's

core innovation lies in its selective scanning mechanism and optimized structured state-space sequence modeling (S4). This mechanism dynamically propagates or forgets state information based on input content, enabling effective long-range dependence capture with linear time complexity [$O(N)$]. This linear scalability presents significant advantages when processing long-sequence data, substantially reducing computational and memory overhead, making it a strong alternative to CNNs and Transformers in efficiency-critical remote sensing applications. Mamba-based models [24], [25] have been applied to HSI–LiDAR fusion, preliminarily validating their efficiency in processing multimodal remote sensing data.

In summary, although CNNs, Transformers, and Mamba each possess distinct advantages in HSI–LiDAR fusion classification, the current HSI–LiDAR fusion classification methods based on these architectures not only exhibit respective limitations of the base models but also face prominent common challenges in practical implementation

1) CNNs struggle to model long-distance dependencies, while Transformers and Mamba lack flexibility in multiscale feature fusion: CNNs are constrained by local receptive fields [26]; Transformers capture long-range relationships but suffer from high computational complexity, and their fixed tokenization strategy has multiscale adaptability limitations [27]; and Mamba-type models are efficient, but their selective scanning mechanism may be insufficient for handling complex spatial dependencies [28].
2) Single-modality feature extraction or shallow interaction hinders full utilization of HSI and LiDAR complementarity. Many methods often perform simple concatenation or addition at high-level features, failing to achieve deep-level cross-modal information interaction and collaborative learning [29].
3) Object scale variations in complex scenes limit model adaptability. Existing fusion methods require improved dynamic integration capability for multiscale features, challenging the effective simultaneous representation of fine structures and macro scenes.

Overall, static architectures with fixed feature extraction schemes struggle to capture structural variations of objects at different scales in complex scenes and fail to provide scale-adaptive representations for cross-modal interaction. Meanwhile, shallow interactions prevent the spectral advantage of HSI and the spatial advantage of LiDAR from complementing each other effectively across scales, ultimately leading to inaccurate recognition of fine-grained objects and inadequate characterization of macroscopic scenes.

A promising direction is to explore hybrid architectures. Such architectures should combine the local perception strengths of CNNs with the long-range dependence modeling of Transformers or Mamba. They must also support deep cross-modal interaction and adapt to multiscale objects. Developing these architectures is key to advancing HSI–LiDAR joint classification [30].

To address these challenges, we propose a novel triple-branch architecture. Unlike traditional dual-branch networks that often suffer from either premature feature confusion in early fusion or insufficient interaction in late fusion, our tripartite design strategically decouples the feature learning process. This architecture enables dedicated modeling of modality-specific characteristics while providing a structured framework for deep cross-modal interaction. To achieve this goal, specifically, the LiDAR branch focuses exclusively on spatial-elevation patterns, the HSI branch captures global spectral–spatial dependencies, and the multimodal branch integrates both sources for comprehensive representation learning. This tripartite structure naturally aligns with the complementary nature of HSI and LiDAR data, where spectral information, elevation structures, and their interactions each play distinct yet interrelated roles in land cover discrimination.

Based on this analysis, this article proposes a multiscale cross-attention fusion network (MCAFNet), aiming to integrate CNN local perception advantages with Transformer/Mamba long-range dependence modeling capabilities, systematically exploring multiscale, cross-modal complementary characteristics of HSI and LiDAR data to enhance joint classification accuracy. The main contributions are summarized as follows.

1) We propose the MCAFNet framework. To decouple modality-specific features and avoid early fusion confusion in traditional dual-branch networks, this framework adopts three parallel encoding paths for elevation, spectral, and fused features. This structure preserves modality-specific information while enabling deep feature interaction.
2) We design the MSFE module. To handle significant object scale variations in complex remote sensing scenes without increasing computational cost, this module employs depthwise separable convolutions and multiscale dilated convolutions, followed by channel attention for adaptive weighting. This allows the simultaneous characterization of both fine-grained and macrostructures, highly improving adaptability to complex scene scales.
3) We design the MBCA module. To achieve deep bidirectional interaction among the HSI, LiDAR, and multimodal branches, this module performs feature reorganization, cross-attention computation, and triple residual connections within a shared latent space. This design facilitates semantic alignment and bidirectional information exchange, effectively enhancing the utilization of complementary information across modalities.
4) We construct an end-to-end efficient training framework and conduct systematic validation on four public datasets. Results demonstrate that MCAFNet achieves effective and highly competitive performance while maintaining linear complexity. Ablation studies further confirm complementary gains of MSFE and MBCA modules.

The rest of this article is structured as follows. Section II introduces the related work in HSI and LiDAR data joint classification. Section III details the proposed MCAFNet, followed by a more detailed description of the MSFE and MBCA modules. In Section IV, we report and analyze the

experimental results. Finally, Section V presents a summary of our work.

II. Related Work

The joint classification of HSI and LiDAR data has evolved through various multimodal learning architectures, which can be categorized into two main paradigms: single-architecture and hybrid-architecture approaches.

A. Multimodal Learning Based on a Single Architecture

This paradigm employs homogeneous core modules—such as pure CNN, Transformer, or SSM—as a unified feature-extraction backbone. CNNs are widely used for their strong local feature extraction ability. For instance, 3-D-CNN [31] introduced 3-D convolution to HSI classification. Lu et al. [22] proposed a global–local transformation network for joint classification, and Ye et al. [32] used multiscale deep convolutions to alleviate catastrophic forgetting. However, CNNs are limited in capturing long-range dependencies due to their local receptive fields. To overcome the limitations of CNNs, researchers have explored Transformer models, centered on the self-attention mechanism, as an alternative unified backbone. Roy et al. [33] introduced a multimodal fusion transformer using external classification tokens, and Hoffmann et al. [34] designed a synchronous token fusion architecture for cross-modal interaction. While Transformers excel at global context modeling, their quadratic computational complexity becomes a bottleneck for high-resolution images.

Recently, SSMs such as Mamba have gained attention for their linear complexity and effective long-sequence modeling. Liao et al. [35] proposed HLMamba, a Mamba-based network for the joint classification of HSI and LiDAR data, which effectively models long-range dependencies in multimodal sequences and better explores intramodal features and intermodal relationships. However, SSMs may still struggle with complex 2-D spatial structures and risk information loss during selective scanning. In summary, single-architecture approaches face a fundamental tradeoff among local perception, global modeling, and computational efficiency.

B. Multimodal Learning Based on a Hybrid Architecture

Hybrid architectures integrate heterogeneous modules to leverage their complementary strengths, offering more flexible solutions to address the limitations of single-architecture approaches. One common strategy combines convolution and attention mechanisms, where CNNs capture local patterns while attention modules enhance global context awareness. Zhu et al. [36] proposed a ConvGRU-based network with contourlet transform for texture extraction, and Ma et al. [37] designed intrasource and intersource interaction learning for feature fusion. Despite improvements in feature representation, global context modeling in such frameworks often remains constrained by the predominant CNN backbone. To more effectively leverage both local features and global context simultaneously, CNN–Transformer hybrid models have been increasingly explored. For example, Xu et al. [38] proposed a dual-stream Transformer with differential attention and gated fusion. Meanwhile, the interactive enhanced network by Gao et al. [39] leverages graph convolution and multi-head self-attention to jointly model spatial relationships and global spectral dependencies for HSI and LiDAR data fusion. However, many existing methods tend to simply cascade the two types of modules rather than designing truly integrated architectures, often failing to establish deep, bidirectional interaction mechanisms while introducing significant computational overhead due to the Transformer component [40].

With the recent emergence of SSMs like Mamba, the design space for hybrid architectures has further expanded, enabling new combinations, such as CNN–Mamba or Transformer–Mamba. For instance, Xie et al. [41] proposed FusionMamba for multimodal medical image fusion, demonstrating the efficiency of SSMs in cross-modal tasks; Dong et al. [42] introduced an improved Mamba architecture with gating mechanisms for finer fusion within hidden state spaces.

In contrast to prevailing hybrid paradigms that typically concatenate or parallelize CNN and Transformer (or Mamba) modules within a single-path or dual-branch framework, the proposed MCAFNet adopts a distinct strategy in its architectural design. Rather than directly stacking heterogeneous modules, MCAFNet introduces a triple-branch parallel encoding structure, which provides dedicated extraction and refinement paths for elevation features, spectral–spatial features, and fusion features, respectively. This decoupled design aims to avoid the blending of modality-specific information at its source and to establish a clearer, more stable feature foundation for subsequent cross-modal interaction. Regarding the fusion mechanism, MCAFNet employs an MBCA module, rather than simple attention weighting or feature concatenation, to facilitate systematic, bidirectional information exchange among the three feature groups. Furthermore, to address multiscale object variations, MCAFNet incorporates a dedicated MSFE module within the multimodal branch. This module explicitly captures multiscale contextual information through parallel dilated convolutions, rather than relying solely on the inherent sequence modeling capabilities of Transformers or Mamba, which may be less sensitive to spatial scale variations. Collectively, these design choices enable MCAFNet to more effectively tackle core challenges in HSI and LiDAR joint classification, including feature decoupling, deep interaction, and scale adaptation.

III. Proposed Method

This article proposes an MCAFNet for joint classification of HSI and LiDAR data. The overall framework, illustrated in Fig. 1, adopts a triple-branch encoder architecture. It separately extracts global spectral–spatial features from HSI data and elevation structural features from LiDAR data, followed by shallow fusion and multiscale feature enhancement (MSFE) through a multimodal branch. Finally, deep bidirectional feature interaction among the three branches is achieved via the designed MBCA module, fully exploiting intermodal complementary information. This section details the design and principles of each core component.

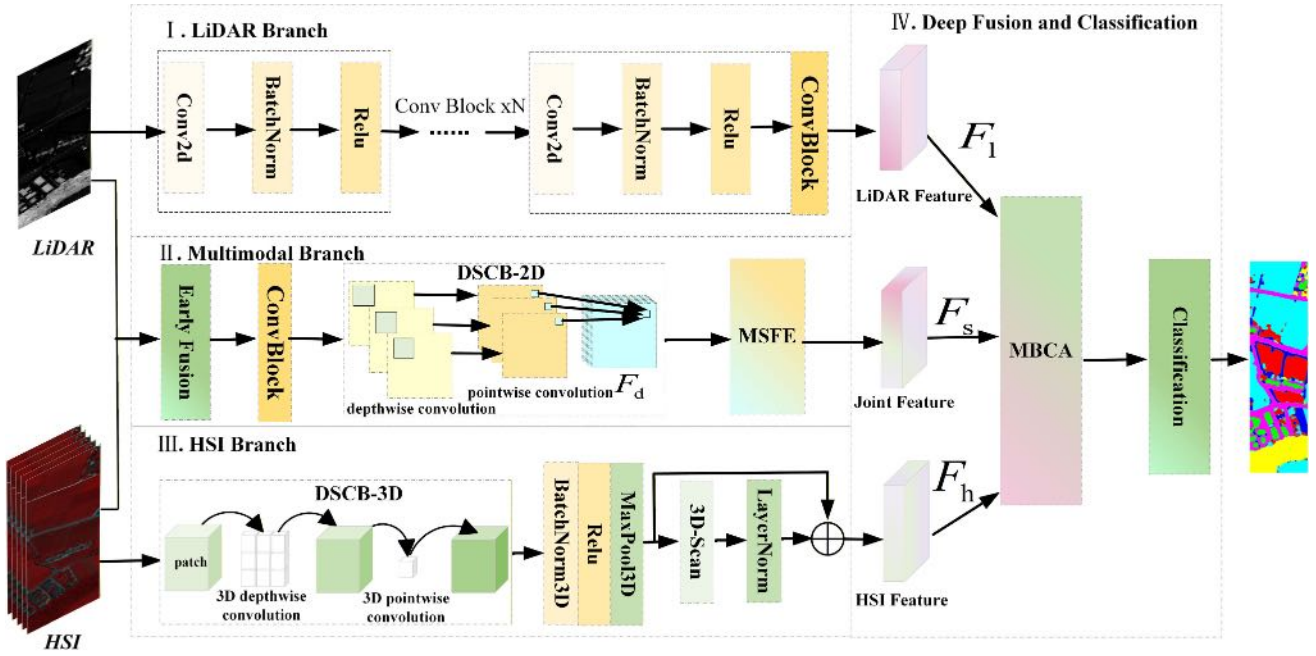


Fig. 1. Overall architecture of the proposed MCAFNet for HSI and LiDAR data joint classification. The framework consists of three dedicated branches: a LiDAR branch for spatial-elevation feature extraction, an HSI branch for global spectral representation learning, and a multimodal fusion branch. The fusion branch first integrates the features from the other two branches. It then employs the MSFE module to capture rich spatial–spectral contexts at various scales. Subsequently, the multihead bidirectional cross-attention (MBCA) module performs deep, interactive fusion between the LiDAR branch features (F_l), HSI branch features (F_h), and the multimodal branch features (F_s) through a cross-attention mechanism. The final fused representation is fed into a classifier to yield the prediction results.

A. Spatial–Spectral Feature Extraction

Given an input HSI patch $\mathbf{X}_h \in \mathbb{R}^{H\times W\times C_h}$ and a LiDAR elevation data patch $\mathbf{X}_l \in \mathbb{R}^{H\times W\times C_l}$, the model performs feature extraction through three parallel branches. Each branch outputs feature maps of identical dimensions, laying the foundation for subsequent deep fusion.

1) HSI Branch: Global Spectral–Spatial Feature Extraction: The HSI branch aims to effectively capture hyperspectral data spectral–spatial dependencies through 3-D convolution and a selective state-space scanning mechanism. This branch first expands the depth dimension of the input hyperspectral data $\mathbf{X}_h$ and converts it into a 3-D feature cube $\mathbf{F}_{3D}$. Subsequently, efficient feature extraction is performed using a 3-D depthwise separable convolutional block, primarily implemented through cascaded depthwise and pointwise 3-D convolutions

$$\mathbf{F}_{dw} = \zeta\left(\mathrm{BN}_{3D}\left(\mathrm{DWConv}_{3\times3\times3}\left(\mathbf{F}_{3D}\right)\right)\right) \tag{1}$$

$$\mathbf{F}_{pw} = \zeta\left(\mathrm{BN}_{3D}\left(\mathrm{Conv}_{1\times1\times1}\left(\mathbf{F}_{dw}\right)\right)\right) \tag{2}$$

where $\zeta(\cdot)$ denotes the GELU activation function. To enhance feature representation and reduce computational complexity, a maxpooling operation is introduced

$$\mathbf{F}_{pool} = \mathrm{MaxPool}_{3D}\left(\xi\left(\mathrm{BN}_{3D}\left(\mathrm{F}_{pw}\right)\right)\right) \tag{3}$$

where $\xi(\cdot)$ denotes the RELU activation function.

The core component of this branch is a scanning mechanism. Inspired by the selective SSM of Mamba, this mechanism has been appropriately simplified for our specific task. The spectral dimension naturally follows a sequential order of wavelengths, making it well-suited for capturing long-range dependencies. Spatial features have already been extracted by the preceding 3-D convolutions. Therefore, we tailor the scanning process specifically to the spectral dimension for adaptation to hyperspectral data. We also simplify the state-space update logic into a bidirectional linear transformation with shared parameters, which avoids the complex structured state-space computation in the original Mamba.

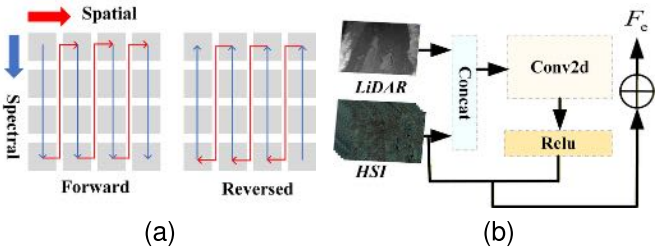


Fig. 2. Schematic of (a) spectral scanning paths and (b) early fusion module.

The 3-D-scan module models long-range dependencies through a bidirectional scanning strategy. The processing flow is as follows. First, input features are serialized along the spectral dimension to generate a sequence representation $\mathbf{X}_{seq} \in \mathbb{R}^{(B\times H\times W)\times D\times C}$.

A bidirectional scanning process is then executed: the forward scan processes the sequence front-to-back along the spectral dimension, integrating the current state with the previous hidden state; and the reverse scan processes the sequence in the opposite direction. Specific scanning paths are illustrated in Fig. 2(a). Both scanning processes share parameter matrices

A and **B** for processing the current input and the previous state, respectively. Scan outputs are controlled by a Sigmoid gating mechanism, and bidirectional scan results are ultimately averaged to form an enhanced feature representation.

A simplified state-space scanning process is then applied along the spectral dimension. This process yields the enhanced feature representation. Finally, output features are obtained through layer normalization and a residual connection to ensure training stability

$$\mathbf{F}_h = \text{LayerNorm}\left(Scan_{spectral}\left(\mathbf{F}_{pool}\right)\right) + \mathbf{F}_{pool}. \tag{4}$$

This branch outputs $\mathbf{F}_h \in \mathbb{R}^{C_f \times H \times W}$, effectively integrating HSI data spectral–spatial structural information and enhancing long-range dependence modeling through the selective scanning mechanism, where the feature dimension C_f is defined as a tunable hyperparameter and its empirical determination method will be detailed in Section IV.

2) LiDAR Branch: Elevation Feature Extraction: The LiDAR branch focuses on extracting multilevel elevation structural features from digital surface model (DSM) data. Given that LiDAR data typically contains rich spatial information but with relatively simple spectral characteristics compared to HSI, this branch employs a streamlined convolutional architecture to capture hierarchical elevation patterns.

The processing begins with the LiDAR data undergoing feature extraction through N consecutive convolutional blocks. The output from the ith convolutional block is given by the following mathematical formulation:

$$\mathbf{F}_l^i = \xi\left(\text{BN}\left(\text{Conv}_{3\times3}\left(\mathbf{F}_l^{i-1}\right)\right)\right) \tag{5}$$

where, for the initial condition $i = 0$, $\mathbf{F}_l^0 = \mathbf{X}_l$.

This branch ultimately outputs features $\mathbf{F}_l \in \mathbb{R}^{C_f \times H \times W}$. We align the dimension C_f with that of the HSI branch features $\mathbf{F}_h$ to ensure seamless integration in later fusion modules. The extracted multilevel elevation structural features provide crucial geometric information for the core MBCA module.

3) Multimodal Branch: Multiscale Fusion Feature Extraction: The multimodal branch is one of the cores of our network. Its purpose is to fuse HSI and LiDAR data at an early stage and capture target features through multiscale processing.

In the early fusion module, as illustrated in Fig. 2(b), the HSI and LiDAR input data are first concatenated along the channel dimension to create a unified representation. The combined features are then projected and nonlinearly transformed via a 1×1 2-D convolutional layer. Finally, the early fused feature $\mathbf{F}_e$ is obtained via a residual connection with the original HSI input to preserve critical spectral information. This early fusion process can be mathematically formulated as

$$\mathbf{F}_e = \xi\left(\text{Conv}_{1\times1}\left[\mathbf{X}_h, \mathbf{X}_l\right]\right) + \mathbf{X}_h \tag{6}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the channel dimension.

The output feature $\mathbf{F}_e$ from the early fusion module, while integrating multimodal information, remains a relatively shallow representation. Therefore, it is first refined by a convolutional block, which employs a 3×3 convolution followed by batch normalization and RELU activation to enhance its representational capacity. The refined features $\mathbf{F}_{conv}$ are then fed into a 2-D depthwise separable convolution block (DSCB-2-D) for further efficient feature extraction, implemented with depthwise and pointwise convolutions as follows:

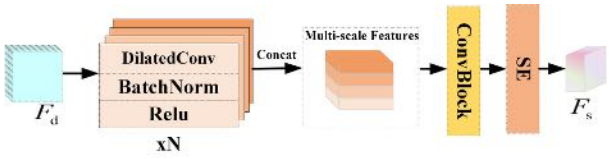


Fig. 3. Schematic of the MSFE.

$$\mathbf{F}_{dw} = \zeta\left(\text{BN}\left(\text{DWConv}_{3\times3}\left(\mathbf{F}_{conv}\right)\right)\right) \tag{7}$$

$$\mathbf{F}_d = \zeta\left(\text{BN}\left(\text{Conv}_{1\times1}\left(\mathbf{F}_{dw}\right)\right)\right). \tag{8}$$

However, remote sensing scenes exhibit significant scale variations in ground objects, ranging from extensive homogeneous regions to small isolated objects coexisting within the same imagery. To enhance the adaptability of the model to such multiscale characteristics, we design an MSFE module. As shown in Fig. 3, the MSFE module captures multiscale contextual information through N parallel dilated convolutions. Let $\mathbf{F}_{scale}^k$ denote the output of the kth scale branch; this process can be mathematically formulated as

$$\mathbf{F}_{scale}^k = \xi\left(\text{BN}\left(\text{Conv}_{3\times3,\text{dilation}=k}\left(\mathbf{F}_d\right)\right)\right) \tag{9}$$

$$\mathbf{F}_{fused} = \left[\mathbf{F}_{scale}^1, \mathbf{F}_{scale}^2, \ldots, \mathbf{F}_{scale}^N\right]. \tag{10}$$

The number of parallel branches N, which corresponds to the index k in $\mathbf{F}_{scale}^k$, is set to 3 in our implementation, with dilation rates ranging from 1 to 3. This configuration effectively captures spatial contexts at multiple scales. This design is motivated by the need to balance computational efficiency with the ability to represent objects at different scales commonly found in remote sensing scenes.

To address the information redundancy in fusing multiscale features and the varying contributions of different channels to classification, we first enhance feature representation via convolutional blocks, yielding the feature map F_{cf}. Subsequently, a channel attention mechanism is introduced. This mechanism adaptively recalibrates feature responses and highlights the most discriminative channels. The corresponding mathematical formulation is provided as follows:

$$\mathbf{F}_{attn} = \sigma\left(\text{Conv}_{1\times1}\left(\zeta\left(\text{Conv}_{1\times1}\left(\text{GAP}\left(\mathbf{F}_{cf}\right)\right)\right)\right)\right) \tag{11}$$

$$\mathbf{F}_s = \mathbf{F}_{cf} \otimes \mathbf{F}_{attn} \tag{12}$$

where GAP denotes global average pooling, σ is the Sigmoid activation function, and $\otimes$ represents elementwise multiplication along the channel dimension. This branch outputs multimodal fusion features $\mathbf{F}_s \in \mathbb{R}^{C_f \times H \times W}$.

B. Deep Fusion and Classification

1) Multihead Bidirectional Cross-Attention: To establish deep bidirectional information interaction among the three feature branches (HSI, LiDAR, and multimodal), an MBCA module is designed. This module facilitates semantic alignment and bidirectional information exchange between different

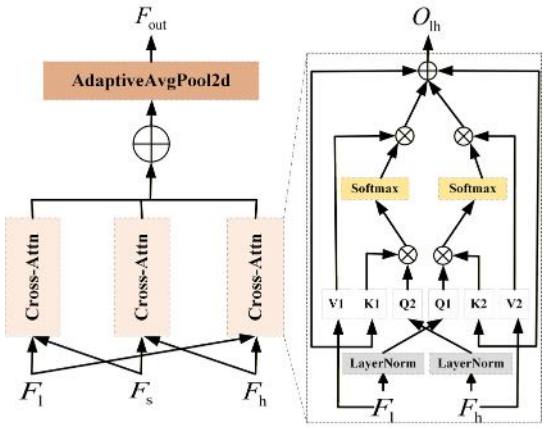


Fig. 4. Schematic of the MBCA, symbols F_{l}, F_{h}, and F_{s} denote the LiDAR branch features, HSI branch features, and multimodal branch features, respectively.

branch features, fully leveraging intermodal complementary information to enhance joint classification performance. As depicted in Fig. 4, take features $\mathbf{F}_h$ and $\mathbf{F}_l$ as example.

First, the two input features undergo layer normalization to stabilize training and improve feature consistency. Then, each feature is projected into query (Q), key (K), and value (V) vectors through linear projection layers, preparing for subsequent attention computation. In bidirectional cross-attention computation, the Q vector of each branch interacts with the K and V vectors of the other branch, thereby retrieving and aggregating relevant information from the features of the other branch. Specifically, the Q of branch A performs attention computation with the K–V pairs of branch B to obtain information beneficial to A; similarly, branch B extracts useful information from branch A. Each attention computation output is added to its original Q via residual connections to preserve critical information. The final output after this bidirectional interaction is denoted as $\mathbf{O}_{lh}$.

Finally, the above bidirectional interaction is sequentially performed for the three feature pairs ($\mathbf{F}_h$–$\mathbf{F}_s$, $\mathbf{F}_h$–$\mathbf{F}_l$, and $\mathbf{F}_s$-$\mathbf{F}_l$), and all interaction results are summed. To facilitate subsequent classification head predictions, the summed features are processed through GAP, ultimately forming a rich multimodal fused representation $\mathbf{F}_{\mathrm{final}}$. The MBCA module process is represented by the following mathematical expression:

$$\mathbf{F}_{\mathrm{out}} = \mathrm{MBCA}\left(\mathbf{F}_h, \mathbf{F}_l, \mathbf{F}_s\right). \tag{13}$$

2) Classifier: The final fused feature $\mathbf{F}_{\mathrm{out}}$, now in the form of a 1-D vector, is fed into a classifier composed of a dropout layer and a linear layer to predict the category label of each pixel. This design helps prevent overfitting and improves model generalization capability. The model is trained to minimize the cross-entropy loss between predictions and ground-truth labels. Through end-to-end training, MCAFNet can automatically learn how to effectively fuse HSI and LiDAR data to achieve optimal classification performance.

IV. Experiments and Analysis

This section provides a comprehensive overview of the datasets, evaluation metrics, and experimental settings employed. Furthermore, comparative experiments and ablation studies are conducted to demonstrate the effectiveness and superiority of the proposed MCAFNet.

A. Dataset Description

To validate the effectiveness of the proposed method, experiments are conducted on four publicly available datasets of multisensor remote sensing image classification: Houston2013, Augsburg, Trento, and MUUFL. These datasets comprise both HSI and LiDAR data, covering diverse scenarios of land cover such as urban and agricultural areas, ensuring fair public comparability.

1) Houston2013 Dataset: The Houston2013 dataset was acquired over the University of Houston area using the CASI-1500 sensor. This dataset has a spatial size of 349×1505 pixels and provides 144 spectral bands covering the wavelength range of 0.38–1.05 μm at 2.5-m spatial resolution. This dataset also includes a LiDAR-derived DSM at the same resolution. The dataset consists of 15 classification categories.

2) Augsburg Dataset: The Augsburg dataset was collected in the urban area of Augsburg, Germany, with a spatial size of 332×485 pixels. This dataset consists of hyperspectral imagery containing 180 spectral bands spanning 0.44–2.5-μm wavelength range at 30-m spatial resolution, along with LiDAR elevation data. The dataset includes ground truths for seven classes.

3) Trento Dataset: Representing an agricultural region in Italy, the Trento dataset was captured by the AISA Eagle sensor. The dataset has a spatial size of 600×166 pixels and offers 63 spectral bands at 1-m spatial resolution. This dataset also contains a coregistered LiDAR DSM at 1-m resolution. The dataset consists of six different ground-feature categories.

4) MUUFL Dataset: The MUUFL dataset was acquired over the University of Southern Mississippi Gulfport campus. This dataset has a spatial size of 325×220 pixels and includes hyperspectral data with 64 spectral bands covering visible and near-infrared wavelengths at 0.54×1.0 m spatial resolution, accompanied by LiDAR data at 0.60×0.78 m resolution. The dataset contains 11 land cover classes.

B. Experimental Setup

All experiments were conducted on a computer equipped with an Intel Core i5-14600KF CPU, an NVIDIA GeForce RTX 4060Ti GPU, and Ubuntu 20.04. The implementation is based on PyTorch. The Adam optimizer was used with a 0.05 learning-rate decay factor. Based on empirical validation, the batch size was set to 64, initial learning rate to 0.0005, and the models were trained for 100 epochs. The cross-entropy loss function was employed. Model performance was evaluated using four metrics: overall accuracy (OA), average accuracy (AA), kappa coefficient (κ), and $F1$ score ($F1$). All reported results are optimal values from three independent runs, enabling comprehensive classification result comparison with ground-truth maps.

C. Performance Comparison

To highlight the proposed method's performance, this section compared MCAFNet with various state-of-the-art methods

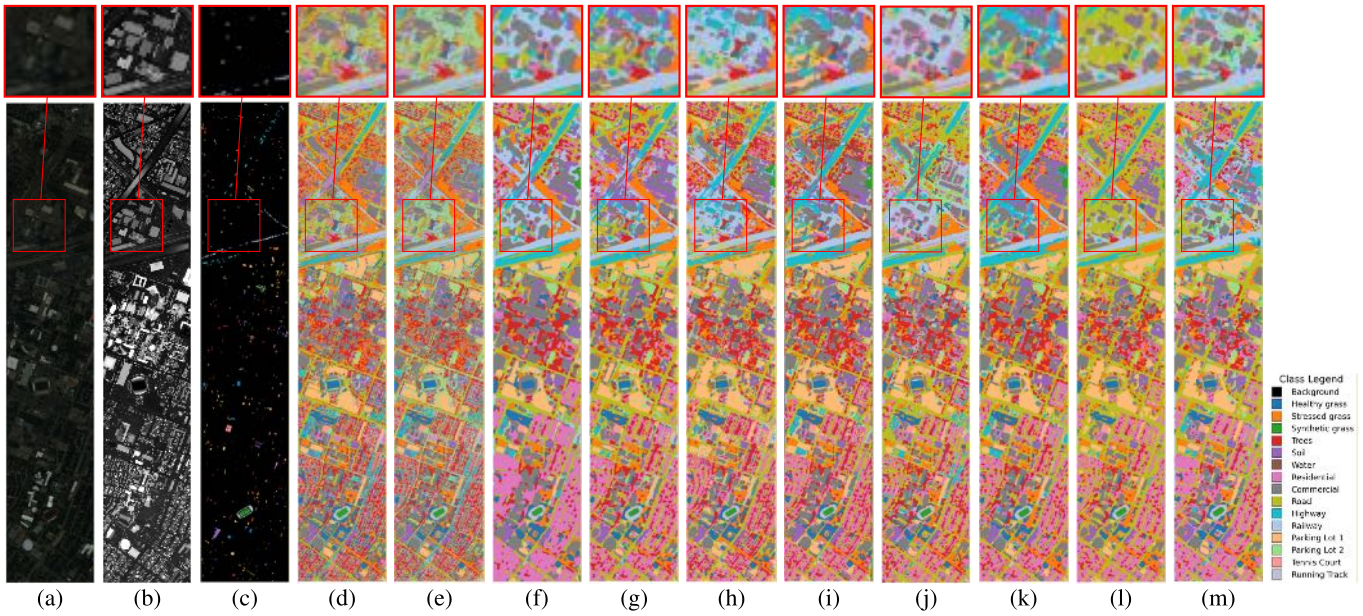


Fig. 5. Classification maps of different methods for the Houston2013 dataset. (a) False-color HIS, (b) gray image for LiDAR, (c) ground-truth map, (d) MDL-RS (89.22%), (e) EndNet (90.07%), (f) CALC (92.89%), (g) ExVit (92.22%), (h) Sal2RN (91.42%), (i) HLMamba (94.97%), (j) IDNet (92.10%), (k) MCFNet (94.57%), (l) S3F2Net (94.36%), and (m) MCAFNet (96.01%).

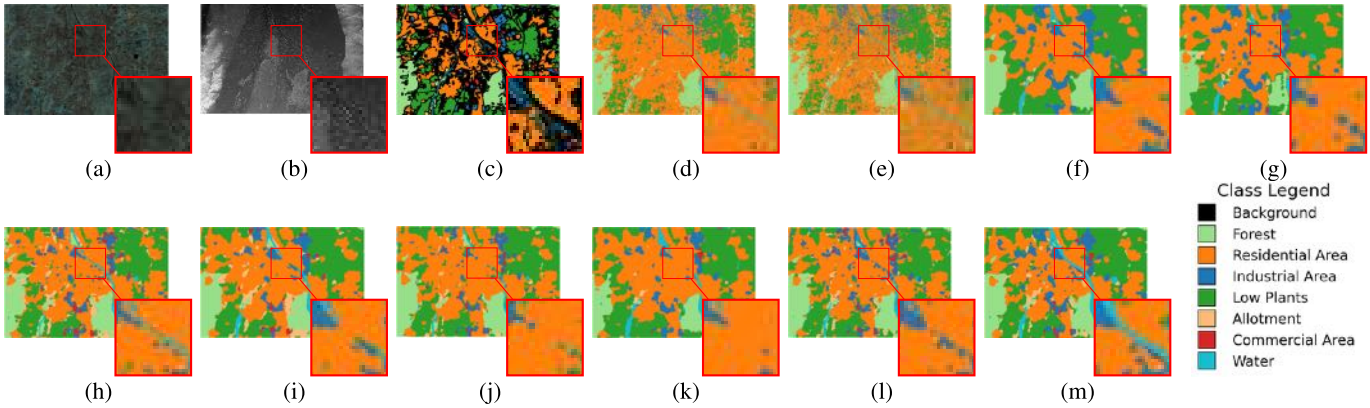


Fig. 6. Classification maps of different methods for the Augsburg dataset. (a) False-color HSI, (b) gray image for LiDAR, (c) ground-truth map, (d) MDL-RS (75.18%), (e) EndNet (80.55%), (f) CALC (91.77%), (g) ExVit (89.19%), (h) Sal2RN (90.26%), (i) HLMamba (89.71%), (j) IDNet (90.71%), (k) MCFNet (91.20%), (l) S3F2Net (92.23%), and (m) MCAFNet (92.39%).

from both visualization and quantitative perspectives, including: MDL-RS [28], EndNet [43], CALC [22], ExViT [21], Sal^{2}RN [16], HLMamba [35], IDNet [44], MCFNet [45], and S2F3Net [29]. For these comparative methods, we adhered to their optimal configurations as specified in the original publications; otherwise, we maintained consistent settings with our model for a fair comparison.

1) Comparison With Different Methods: Tables I–IV presented the classification results of different methods on the Houston2013, Augsburg, Trento, and MUUFL datasets, respectively. The best results of each category are highlighted in bold. Figs. 5–8 sequentially display the false-color HSI of each dataset, LiDAR, ground-truth map, and the classification results of different methods, with colors indicating the corresponding classes. It is worth noting that, to present the visual results clearly, we enlarged the local regions in the classification maps, which are indicated by the red rectangular boxes in the figures. Several conclusions could be drawn from the experimental results. CNN-based methods that employ multifeature fusion strategies perform significantly better. In contrast, methods that rely only on single deep feature extraction show comparatively weaker results. For instance, methods like Sal^{2}RN, which achieve deep integration of the spatial–spectral features from HSI and the elevation features from LiDAR, outperform fusion approaches that depend on single deep feature extraction (e.g., EndNet). Cross-modal interaction enhances classification performance. The cross-fusion mechanism designed in MDL-RS validates deep multimodal data interaction effectiveness, outperforming early simple concatenation or summation fusion strategies. Attention mechanisms enhance focus on key features: MCFNet, by introducing multiattention mechanisms, improves key object feature capture in complex scenes, especially for imbalanced sample distribution classes. The

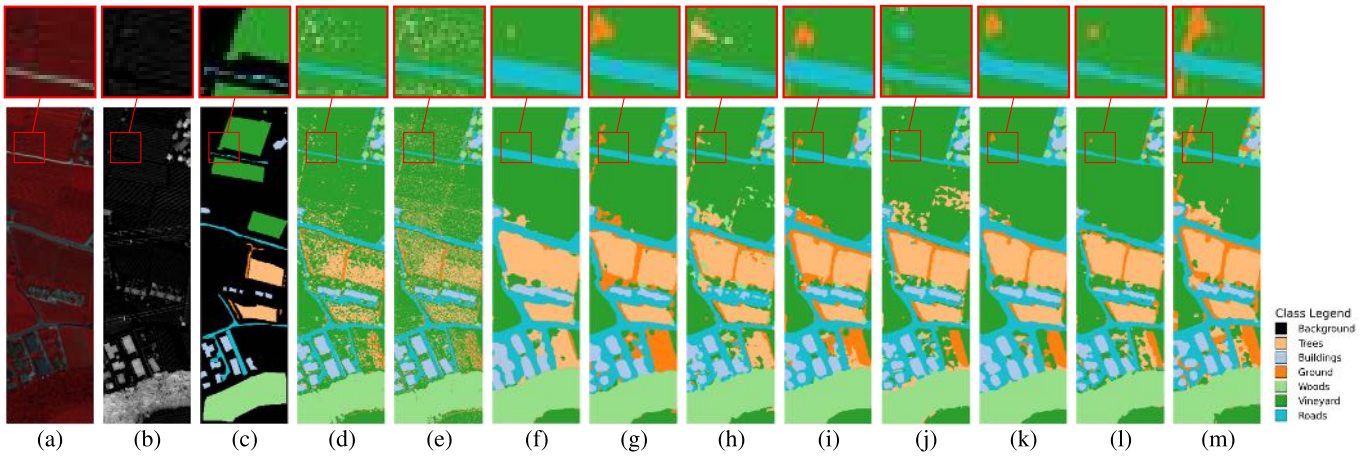


Fig. 7. Classification maps of different methods for the Trento dataset. (a) False-color HSI, (b) gray image for LiDAR, (c) ground-truth map, (d) MDL-RS (89.14%), (e) EndNet (90.82%), (f) CALC (99.25%), (g) ExVit (99.25%), (h) Sal2RN (98.26%), (i) HLMamba (99.25%), (j) IDNet (99.43%), (k) MCFNet (99.08%), (l) S3F2Net (99.06%), and (m) MCAFNet (99.51%).

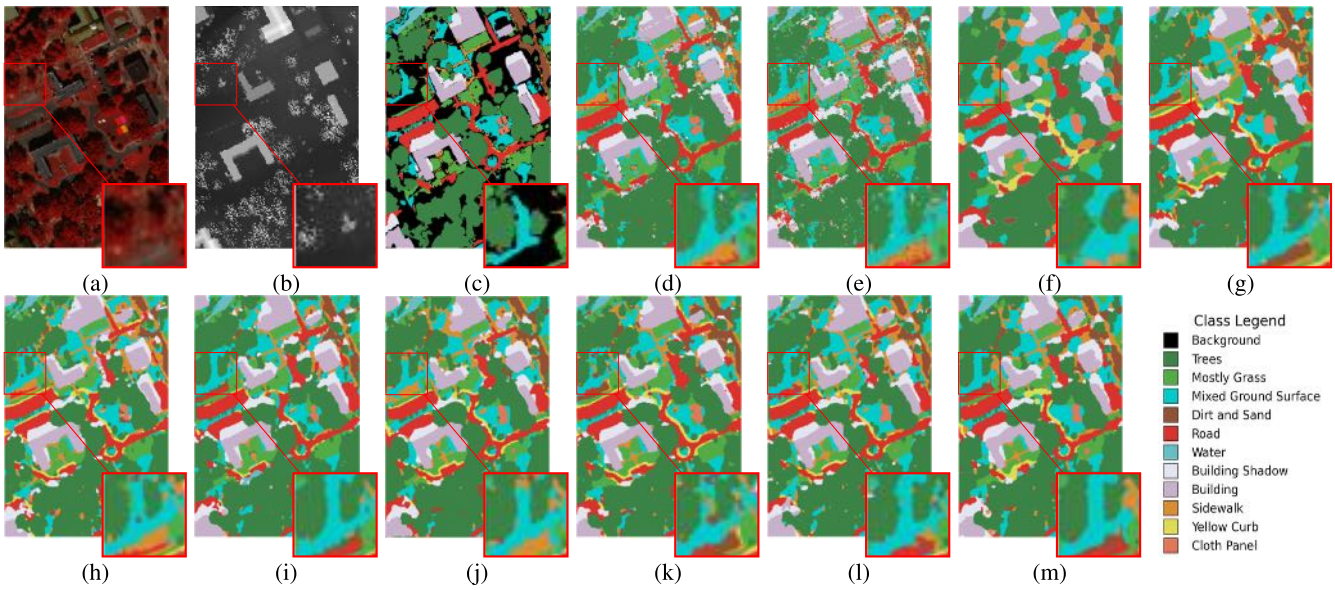


Fig. 8. Classification maps of different methods for the MUUFL dataset. (a) False-color HSI, (b) gray image for LiDAR, (c) ground-truth map, (d) MDL-RS (86.84%), (e) EndNet (86.61%), (f) CALC (77.63%), (g) ExVit (86.21%), (h) Sal2RN (86.60%), (i) HLMamba (88.78%), (j) IDNet (87.18%), (k) MCFNet (87.56%), (l) S3F2Net (86.80%), and (m) MCAFNet (90.30%).

IDNet model, which focuses on learning high-frequency detail features through intensity-constrained Transformer blocks and Laplacian pyramid decomposition, shows competitive performance, particularly in capturing fine spatial–spectral details. Multiscale and decision-level fusion optimize model adaptability: HLMamba, through its efficient long-range dependence modeling combined with multiscale feature extraction, further optimizes classification consistency, indicating the importance of integrating global context with multiscale structures for enhancing model robustness. In contrast, the proposed MCAFNet surpassed all comparative methods on all four datasets in OA, AA, Kappa, and $F1$ scores. MCAFNet's advantage stems from multiple complementary design choices. Fundamentally, its unique three-branch architecture synthesizes the strengths of multiple frameworks. The MSFE and MBCA modules further enhance multiscale adaptability and deep cross-modal interaction, thus exhibiting stronger adaptability and classification accuracy in complex and varying remote sensing scenarios.

2) Quantitative Analysis: Tables I–IV present the comprehensive quantitative evaluation results across the four benchmark datasets. Our proposed MCAFNet method achieved the best results across all four datasets.

Table I indicates that MCAFNet achieves the highest OA among all comparative methods. In contrast, methods such as MDL-RS and EndNet exhibit relatively lower OA values of 89.22% and 90.07%, respectively. This performance gap may be attributed to their insufficient exploration of cross-modal complementarity, where MDL-RS primarily relies on a decision-level fusion strategy that fails to achieve deep feature interaction, while EndNet adopts a relatively simple fusion approach that may lead to feature confusion between

TABLE I
COMPARATIVE EXPERIMENTAL RESULTS ON THE HOUSTON2013 DATASET

No.	Class(Train/Test)	MDL-RS	EndNet	CALC	ExVit	Sal^{2}RN	HLMamba	IDNet	MCFNet	S2F3Net	MCAFNet
1	Health grass(198/1053)	84.29	**98.12**	82.33	81.96	82.84	83.09	84.42	82.53	86.93	83.10
2	Stressed grass(190/1064)	96.13	**100**	96.40	99.72	**100**	99.90	96.61	**100**	99.51	98.78
3	Synthetic grass(192/505)	94.19	97.63	94.03	98.81	96.61	95.84	**100**	95.25	96.64	**100**
4	Trees(188/1056)	98.62	98.48	95.92	98.39	96.53	**99.71**	98.95	99.62	99.13	96.21
5	Soil(186/1056)	98.25	99.71	99.43	**100**	96.62	**100**	**100**	**100**	99.60	**100**
6	Water(182/143)	95.70	80.88	95.80	**100**	**100**	95.80	95.80	95.80	95.43	**100**
7	Residential(196/1072)	80.17	87.56	92.91	**96.92**	91.92	88.89	88.71	93.75	93.28	90.76
8	Commercial(191/1053)	87.80	88.95	84.49	91.26	85.68	92.40	95.82	94.78	92.79	**95.92**
9	Road(193/1059)	80.39	80.21	84.89	86.97	83.18	**97.63**	92.44	89.24	89.23	97.54
10	Highway(191/1036)	62.87	92.98	88.99	80.02	83.81	**97.20**	65.05	97.10	96.14	95.85
11	Railway(181/1054)	91.06	86.26	**100**	88.33	95.70	95.92	96.39	97.44	98.94	99.43
12	Parking lot1(192/1041)	77.01	70.88	86.61	88.76	88.43	92.60	93.37	87.99	91.95	**98.27**
13	Parking lot2(184/285)	84.02	**99.60**	89.82	91.93	86.27	92.28	88.77	91.23	92.26	92.63
14	Tennis court(181/247)	**100**	98.31	**100**	**100**	94.45	97.16	**100**	**100**	99.20	**100**
15	Running track(187/473)	98.45	83.48	99.36	**100**	**100**	99.57	99.78	99.79	97.07	**100**
	OA%(2832/12197)	89.22	90.07	92.89	92.22	91.42	94.97	92.10	94.57	94.36	**96.01**
	AA%	90.85	90.47	93.53	93.54	92.46	95.20	93.08	94.97	95.24	**96.57**
	κ%	88.32	89.22	92.28	91.55	92.16	94.55	91.42	94.11	94.37	**95.67**
	F1%	88.01	89.86	93.54	93.43	92.31	94.28	92.34	94.69	94.84	**96.15**

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS ON THE AUGSBURG DATASET

No.	Class(Train/Test)	MDL-RS	EndNet	CALC	ExVit	Sal^{2}RN	HLMamba	IDNet	MCFNet	S2F3Net	MCAFNet
1	Forest(146/13361)	91.86	90.32	94.83	92.54	96.49	93.97	**98.34**	87.86	96.16	95.40
2	Residential-Area(264/30065)	78.84	87.14	95.21	95.98	97.04	98.15	**99.09**	98.43	98.38	98.19
3	Industrial-Area(21/3830)	26.68	21.57	**89.74**	62.77	61.72	60.96	51.46	67.78	68.68	80.00
4	Low-Plants(248/36609)	75.46	74.89	96.04	92.10	90.36	90.77	91.88	**97.47**	95.49	93.87
5	Allotment(52/523)	44.93	31.83	28.46	39.77	54.38	50.47	51.62	36.33	58.55	**69.41**
6	Commercial-Area(7/1638)	10.74	8.81	1.23	9.65	**17.54**	14.22	5.92	4.52	5.99	8.30
7	Water(23/1507)	53.22	34.39	39.30	43.53	29.77	33.44	40.41	38.55	30.77	**54.41**
	OA%(761/77533)	75.18	80.55	91.77	89.19	90.26	89.71	90.71	91.20	92.23	**92.39**
	AA%	54.53	49.85	64.11	62.33	63.90	63.14	62.68	61.56	64.99	**71.37**
	κ%	64.62	71.47	88.11	84.50	86.02	85.21	86.52	87.16	88.94	**89.11**
	F1%	52.02	51.13	64.46	60.91	63.54	61.59	63.19	64.20	64.91	**67.15**

TABLE III
COMPARATIVE EXPERIMENTAL RESULTS ON THE TRENTO DATASET

No.	Class(Train/Test)	MDL-RS	EndNet	CALC	ExVit	Sal^{2}RN	HLMamba	IDNet	MCFNet	S2F3Net	MCAFNet
1	Trees(40/3994)	98.92	98.85	99.87	99.85	99.37	99.22	99.49	99.57	**100**	99.18
2	Buildings(29/2874)	68.57	67.51	**98.92**	97.83	93.01	98.46	98.57	**98.92**	97.79	97.61
3	Ground(5/474)	98.83	98.77	87.34	98.96	83.54	98.10	98.73	83.54	85.63	**100**
4	Woods(91/9032)	90.45	92.89	**100**	**100**	99.79	**100**	**100**	**100**	**100**	**100**
5	Vineyard(105/10396)	90.45	90.48	99.98	99.97	99.92	**100**	**100**	**100**	**100**	**100**
6	Roads(32/3142)	58.21	64.70	95.98	94.99	94.02	95.51	96.72	95.29	98.04	**98.65**
	OA%(302/29912)	89.14	90.82	99.25	99.25	98.26	99.25	99.43	99.08	99.06	**99.51**
	AA%	84.24	85.53	97.01	98.60	94.94	98.55	98.92	96.22	96.91	**99.23**
	κ%	85.39	87.65	99.00	98.76	97.68	99.00	99.24	98.78	98.75	**99.35**
	F1%	86.18	87.60	97.70	98.62	95.92	98.25	98.73	97.02	97.59	**98.86**

modalities. In contrast, the triple-branch architecture of the MCAFNet method effectively overcomes these limitations by establishing dedicated pathways for modality-specific feature extraction while providing a structured framework for deep cross-modal interaction, thereby achieving more comprehensive feature representation without information loss.

Table II illustrates the robustness of MCAFNet in handling class imbalance, achieving the highest OA of 92.39%. Methods, including IDNet and ExVit, show lower performance with OA values of 90.71% and 89.19%, respectively. These results may be explained by the limitations of each comparator. IDNet excels at enhancing high-frequency details but may lack the capacity to fully model long-range dependencies in complex scenes. Conversely, while ExVit employs attention mechanisms, it might not sufficiently exploit the complementary information between HSI and LiDAR modalities. This suggests that the MBCA module in MCAFNet facilitates more effective feature learning through deep cross-modal interaction.

Table III shows that MCAFNet achieves the highest OA of 99.51% among all methods. Other approaches display lower OA values. The slightly reduced performance of these methods could be due to their insufficient modeling of the complementary relationship between spectral and elevation

TABLE IV
COMPARATIVE EXPERIMENTAL RESULTS ON THE MUUFL DATASET

No.	Class(Train/Test)	MDL-RS	EndNet	CALC	ExVit	Sal^{2}RN	HLMamba	IDNet	MCFNet	S2F3Net	MCAFNet
1	Trees(1184/22062)	75.85	74.91	93.29	92.90	92.62	95.29	**95.77**	94.04	95.60	95.73
2	Mostly grass(216/4054)	74.79	75.19	60.73	78.42	80.76	81.27	74.83	**85.69**	79.38	84.85
3	Mixed Ground Surface(345/6537)	70.26	75.79	64.92	82.61	75.87	79.54	80.23	75.42	79.73	**87.17**
4	Dirt and Sand(91/1735)	86.04	86.02	64.14	**91.01**	78.03	71.35	82.07	87.78	79.03	82.77
5	Road(335/6352)	79.46	75.85	64.64	78.21	84.58	85.89	81.21	82.86	**86.79**	86.07
6	Water(23/443)	77.41	79.91	46.95	53.50	59.27	80.81	72.91	69.98	52.10	**84.20**
7	Building shadow(113/2120)	91.18	**91.83**	42.92	68.87	70.29	86.83	67.97	79.86	73.18	81.93
8	Building(313/5927)	74.98	61.74	87.54	89.67	91.20	**93.67**	91.56	91.24	88.57	89.46
9	Sidewalk(86/1299)	**91.20**	89.60	29.56	71.67	77.97	77.52	67.43	71.13	74.01	83.14
10	Yellow curb(58/125)	91.73	87.80	88.80	80.80	93.09	81.60	95.20	**99.20**	80.60	97.60
11	Cloth panels(15/254)	**94.48**	94.05	83.07	87.01	68.92	49.21	81.10	69.69	83.28	76.77
	OA%(2779/50908)	86.84	86.61	77.63	86.21	86.60	88.78	87.18	87.56	86.80	**90.30**
	AA%	82.49	81.15	66.05	79.51	79.28	80.28	80.94	82.44	79.47	**86.34**
	κ%	82.63	82.36	70.10	82.01	82.39	85.33	83.11	83.76	82.78	**87.27**
	F1%	**79.20**	77.83	61.89	72.64	72.72	76.14	74.94	75.37	74.96	78.64

features. In contrast, MCAFNet's bidirectional cross-attention mechanism effectively leverages LiDAR elevation information. This helps resolve spectral ambiguities, leading to more accurate classification.

For the MUUFL dataset in Table IV, MCAFNet demonstrates superior performance with an OA of 90.30%. Comparative methods, such as MCFNet and S3F2Net, achieve lower OA values of 87.56% and 86.80%, respectively. The performance difference may be attributed to the integrated approach of MCAFNet, combining multiscale processing in the MSFE module with deep feature interaction in the MBCA module, enabling more comprehensive capture of subtle spectral–spatial characteristics. In contrast, the fusion strategy of MCFNet might not fully exploit the multiscale characteristics of objects, while S3F2Net could face challenges in effectively integrating spatial and spectral information across different scales.

In summary, our proposed MCAFNet method consistently achieved the best across the majority of classification categories in all three datasets. The ability of the method to maintain high accuracy across varying scenarios—from complex urban environments to agricultural landscapes—suggests the effectiveness of its architectural design in addressing the challenges of multimodal remote sensing data classification.

3) Qualitative Analysis: As illustrated in the classification maps (see Figs. 5–8), MCAFNet has showcased superior performance across all four datasets, exhibiting minimal noise and delivering highly accurate classification results when viewed overall. The red rectangular boxes highlight local regions that are magnified for detailed comparison.

In the Houston dataset (see Fig. 5), methods such as MDL-RS and EndNet exhibit fragmented boundaries and salt-and-pepper artifacts within Commercial zones. This behavior stems from the intrinsic limitation of their fully connected architectures, which process each pixel independently without considering spatial context. Such pixelwise processing completely disregards the spatial relationships among adjacent urban structures, leading to incoherent boundary delineation. The MSFE module of MCAFNet mitigates this issue by adaptively aggregating contextual information across multiple scales, thereby supporting more coherent boundary delineation without introducing excessive smoothing.

For the Augsburg dataset (see Fig. 6), methods employing early fusion strategies, such as Sal2RN, display blurred transitions between water bodies and adjacent vegetation. Given the spectral similarity of these classes, the early fusion approach of Sal2RN, which inadequately preserves spatial details, may struggle to leverage complementary elevation cues. MCAFNet addresses this through bidirectional cross attention between the spectral and elevation streams, which helps emphasize subtle topographic discontinuities at water edges and improves boundary fidelity in spectrally ambiguous regions.

In the Trento dataset (see Fig. 7), attention-enhanced methods like MCFNet and ExViT show occasional confusion between Ground and Vineyard classes in areas with minimal elevation contrast. This may reflect an imbalance in modality utilization within these models, where spectral dominance overshadows structurally informative elevation features. By enabling iterative refinement between modalities, the cross-attention mechanism of MCAFNet promotes balanced feature integration, enhancing discrimination in scenarios where spectral signatures alone are insufficient.

Regarding the MUUFL dataset (see Fig. 8), inconsistent labeling of mixed ground surface—particularly near building shadows—appears in the results of methods, including IDNet and S2F3Net. Such inconsistencies may arise from the limited capacity of these models to capture fine-scale spatial variations within heterogeneous regions. The parallel multiscale convolutions of MCAFNet, combined with adaptive channel weighting, facilitate simultaneous representation of both local textures and broader contextual patterns, contributing to more stable predictions across complex surface types.

These observations indicate that recurring error patterns in comparative methods frequently arise from inherent architectural limitations. MDL-RS and EndNet are constrained by limited context modeling capabilities. MCFNet and ExViT encounter challenges due to imbalanced integration of spectral and elevation information. IDNet and S2F3Net exhibit reduced adaptability in capturing multiscale spatial variations. While MCAFNet does not eliminate all classification challenges, the design choices of MCAFNet appear to alleviate several

TABLE V
COMPLEXITY ANALYSIS RESULTS BETWEEN MCAFNET AND THE COMPARATIVE METHODS

Metric	MDL-RS	EndNet	CALC	ExVit	Sal^{2}RN	HLMamba	IDNet	MCFNet	S2F3Net	MCAFNet
Params(K)	120.56	91.29	284.14	229.10	940.79	330.69	111.79	549.02	275.20	403.50
FLOPS(M)	6.68	0.18	28.75	46.87	6.32	78.4	21.75	39.06	57.50	37.25
Train time(s)	26.90	24.34	579.61	254.80	88.98	185.05	245.69	131.20	692.47	360.40
Test time(s)	0.04	0.04	1.41	8.56	1.23	2.41	3.65	2.23	1.98	0.73
Memory usage(MB)	2.97	2.28	3591.91	879.71	2768.92	358.38	72.81	189.41	7639.61	188.69

TABLE VI
INFLUENCE OF MBCA AND MSFE ON CLASSIFICATION RESULTS OF MCAFNET

MBCA	MSFE	Houston2013				Augsburg				Trento				MUUFL			
		OA	AA	κ	F1	OA	AA	κ	F1	OA	AA	κ	F1	OA	AA	κ	F1
-	-	93.44	94.20	92.88	92.48	90.34	67.55	86.25	62.00	99.02	98.45	99.05	98.49	86.12	74.63	81.75	71.96
+	-	94.96	95.25	94.53	93.71	91.84	68.16	88.25	66.96	99.43	98.75	99.24	98.79	87.41	75.71	83.35	72.75
-	+	94.11	94.33	93.60	94.51	91.65	70.16	87.42	65.17	99.38	98.77	99.18	98.34	87.18	83.78	83.29	75.23
+	+	**96.01**	**96.57**	**95.67**	**96.15**	**92.39**	**71.37**	**89.11**	**67.15**	**99.51**	**99.23**	**99.35**	**98.86**	**90.30**	**86.34**	**87.27**	**78.64**

recurring issues observed in error-prone regions. Across all these cases, MCAFNet achieves consistency and accuracy.

D. Computational Complexity Analysis

We assessed various methods' model complexity on the Houston2013 dataset, including calculations of floating-point operations (FLOPs), Params, training time, testing time, and memory usage, as shown in Table V. The FLOPs and parameter count of MCAFNet fall within a moderate range among all compared methods. Lightweight models, such as EndNet and MDL-RS, have fewer parameters and lower computational costs, yet their overall accuracies are only 90.07% and 89.22%, respectively, which are substantially below the 96.01% achieved by MCAFNet. In contrast, nonlightweight models, like HLMamba and MCFNet, exhibit higher FLOPs or parameter counts than MCAFNet. Specifically, HLMamba requires more than twice the FLOPs of our model, and MCFNet has both larger parameter count and higher FLOPs. Nevertheless, neither model surpasses MCAFNet in accuracy, reaching only 94.97% and 94.57%, respectively. Despite the triple-branch architecture, the parameter increase remains controlled because we adopt depthwise separable convolutions, shared feature dimensions across branches, and efficient attention mechanisms, for example. Notably, the memory consumption of MCAFNet remains moderate compared to other high-performance methods, contributing to its practical deployability. In terms of testing time, the proposed method demonstrates efficient inference speed. Consequently, MCAFNet achieves a favorable tradeoff between classification accuracy and computational cost.

E. Ablation Study

1) Module Contribution Analysis: To validate the contribution of each proposed module, we conducted systematic ablation studies on the four datasets by sequentially removing the MBCA module and the MSFE module. Results are shown in Table VI.

The baseline model, which consists of the triple-branch encoder architecture without both the MBCA and MSFE modules, performs the worst across all datasets. For instance, on the Houston2013 and MUUFL datasets, it achieves OAs of only 94.49% and 89.01%, respectively. This result underscores the limitations of a simple architecture that lacks dedicated mechanisms for deep cross-modal interaction and multiscale feature extraction.

Adding only the MSFE module to the baseline brings stable improvements. A notable gain is observed on the Augsburg dataset, where the AA increases by 2.61%. This demonstrates the critical role of the MSFE module in enhancing the adaptability of the model to scale variations in complex scenes. By employing parallel dilated convolutions and adaptive channel weighting, the MSFE module empowers the multimodal branch to capture rich spatial–spectral contexts at various scales.

Introducing only the MBCA module to the baseline leads to more significant performance leaps. The OA increases by 1.52% on the Houston2013 dataset and 1.29% on the MUUFL dataset. These substantial improvements confirm the efficacy of the deep bidirectional interaction mechanism facilitated by the MBCA module. It effectively enables semantic alignment and information exchange among the HSI branch, the LiDAR branch, and the multimodal branch, thereby fully leveraging intermodal complementary information.

Finally, the full MCAFNet model, which integrates both the MSFE and MBCA modules, achieves the best performance on all datasets across all metrics. The consistent and comprehensive gains over the variants with single modules indicate that the MSFE and MBCA modules are highly complementary. In conclusion, each component within our design exhibits promising results and plays a crucial role in the feature extraction process.

2) Branch Combination Analysis: To further validate the effectiveness of the triple-branch architecture, we conducted experiments to evaluate the contribution of each branch and their combinations. Table VII presents the OA results of different branch combinations on all four datasets.

The experimental results clearly demonstrate that each branch contributes uniquely to the classification performance.

TABLE VII
OA (%) RESULTS FOR MCFNET METHOD VARIED BRANCH

Branch	Houston2013	Augsburg	Trento	MUUFL
LiDAR Branch	58.39	81.86	95.34	75.99
HSI Branch	93.74	91.07	98.11	84.51
Multimodal Branch	94.86	91.68	98.62	87.89
LiDAR+HSI Branch	94.40	91.92	99.21	86.54
LiDAR+HSI+Multimodal Branch	**96.01**	**92.39**	**99.51**	**90.30**

TABLE VIII
INFLUENCE OF VARIOUS SCANNING MECHANISMS ON CLASSIFICATION RESULTS OF MCAFNET

Scanning Mechanism	Houston2013	Augsburg	Trento	MUUFL
SS1D	95.24	91.69	99.47	87.59
SS2D	95.27	91.62	99.52	87.75
SS3D	95.92	92.18	**99.56**	89.63
3D-Scan	**96.01**	**92.39**	99.51	**90.30**

The individual LiDAR branch achieves the lowest accuracy across all datasets, which is expected given its limited spectral information. Conversely, the HSI branch consistently outperforms the LiDAR branch, highlighting the importance of spectral features for land cover discrimination. The multimodal branch alone yields competitive performance, surpassing both single-modality branches on most datasets. When combining only the LiDAR and HSI branches, the performance is slightly lower than the multimodal branch alone on Houston2013 and MUUFL datasets.

Most importantly, the full triple-branch combination achieves the highest OA across all datasets. Compared to dual-branch combination (LiDAR+HSI), the full model improves OA by 1.61%, 0.47%, 0.30%, and 3.76% on Houston2013, Augsburg, Trento, and MUUFL, respectively. These consistent gains confirm that the multimodal branch, together with the modality-specific branches, provides a more comprehensive feature representation, capturing both source-specific characteristics and their deep complementary relationships.

3) Analysis of Scanning Mechanisms: To evaluate the impact of different scanning mechanisms on spectral feature extraction, we conducted ablation experiments. We compared four scanning strategies: SS1D, SS2D, SS3D, and 3-D-Scan. Our spectral scanning mechanism is designed based on the inherent characteristics of hyperspectral data. The spectral dimension naturally follows a sequential order of wavelengths, making it well-suited for state-space modeling to capture long-range spectral dependencies. Meanwhile, spatial structural information has already been effectively extracted by the preceding 3-D convolutional layers. Thus, the scanning operation can focus exclusively on the spectral dimension, thereby preserving the locality of spatial features. Table VIII presents the OA results of these scanning mechanisms across the four datasets.

The experimental results show clear patterns. The 3-D-Scan mechanism consistently achieves the best or near-best performance across all datasets. It obtains the highest OA on Houston2013, Augsburg, and MUUFL. This demonstrates the effectiveness of the proposed 3-D scanning approach.

A clear performance trend emerges as scanning mechanism complexity increases. The SS1D mechanism scans only along a single dimension. It yields the lowest performance on most datasets. SS2D shows marginal improvements over SS1D. SS3D provides more substantial gains. On the MUUFL dataset, SS3D improves OA by 2.04% over SS1D.

On the Trento dataset, SS3D achieves the highest OA (99.56%). This slightly outperforms 3-D-Scan (99.51%). This suggests that the optimal scanning strategy may be dataset-dependent. It may be influenced by factors like spectral resolution, spatial complexity, and land cover characteristics. However, the 3-D-scan mechanism shows the most consistent overall performance. These findings validate the design choice of the 3-D-Scan mechanism in the HSI branch. It confirms its effectiveness in modeling long-range spectral dependencies. It also maintains robustness across diverse remote sensing scenarios.

TABLE IX
CLASSIFICATION RESULTS BETWEEN MBCA AND VARIOUS FUSION METHODS

Fusion Methods	Houston2013	Augsburg	Trento	MUUFL
Feature Concatenation	94.11	91.65	99.38	87.18
Standard Cross-Attention	95.41	91.62	99.45	88.78
MBCA	**96.01**	**92.39**	**99.51**	**90.30**

4) Comparison With Alternative Interaction Mechanisms: To further validate the necessity of the proposed MBCA module, we compare it with two mainstream interaction mechanisms: standard cross-attention and feature concatenation. We replace the MBCA module with these alternative mechanisms while keeping the rest of the network architecture unchanged. The OA results are presented in Table IX. The feature concatenation method yields the lowest performance among the three interaction mechanisms. The standard cross-attention mechanism shows moderate improvements over feature concatenation. However, its performance still lags behind our proposed MBCA module. For instance, on the Houston2013 dataset, MBCA outperforms standard cross attention by 0.6% in OA. The performance gap is even more pronounced on the MUUFL dataset, where MCAFNet achieves 90.30% OA, surpassing standard cross attention by 1.52%. These results confirm that the proposed MBCA module is a more effective interaction mechanism for HSI and LiDAR data joint classification.

F. Parameter Discussion

1) Different Dimensions of C_f: Based on the quantitative results presented in Fig. 9, we conducted a systematic analysis to investigate the impact of the feature dimension C_f on the final classification performance. Experiments were configured with C_f values of 32, 64, and 128 to evaluate the sensitivity of the model to this key hyperparameter across the four datasets.

The results indicate that the optimal choice of C_f is contingent upon the specific characteristics of the dataset. For the Houston2013 dataset, a clear trend is observed where a larger feature dimension, specifically $C_f = 128$, yields superior performance across all metrics. This suggests that the complex urban scenes and diverse land cover categories

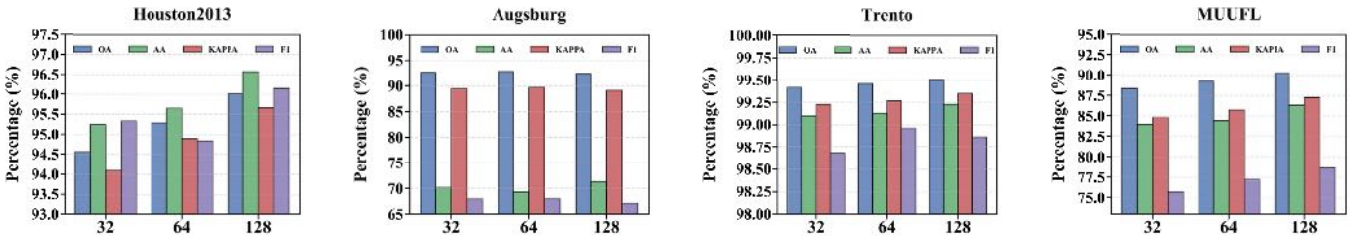


Fig. 9. Classification results of the proposed MCAFNet with different feature dimensions C_f on the Houston2013, Augsburg, Trento, and MUUFL datasets.

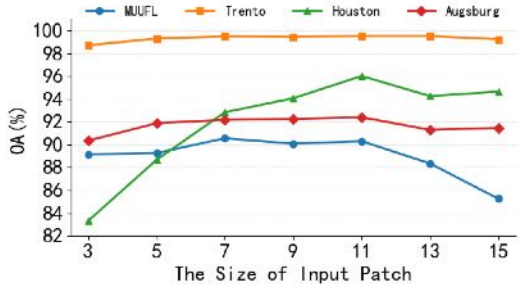


Fig. 10. OA (%) results for MCAFNet method across varied patch sizes.

of this dataset benefit from a higher-dimensional feature representation, which provides greater capacity for the encoding of discriminative information.

A similar preference for a larger feature dimension is noted on the MUUFL dataset, where $C_f = 128$ again leads to the best results. This implies that adequately modeling the spectral–spatial details of this campus environment requires a sufficiently large feature space. In contrast, the performance of the model on the Augsburg dataset remains relatively stable across the different values of C_f, showing only minor fluctuations. Notably, for the Trento dataset, the classification accuracy demonstrates remarkable robustness, with negligible variation across the tested dimensions. This insensitivity suggests that the model can effectively capture the essential features for the classification of the agricultural landscape of Trento, even with a more compact representation.

Considering the overall performance, a feature dimension of $C_f = 128$ achieves the best or highly competitive results on the majority of datasets without introducing significant computational overhead.

2) Different Patch Sizes: We further investigated the impact of the input patch size on the classification performance of the model. A comprehensive analysis was conducted by varying the patch size from 3×3 to 15×15 in increments of 2. The OA trends across the four datasets are summarized in Fig. 10.

The experimental results indicate that the optimal patch size is dataset-dependent, which can be attributed to the varying spatial characteristics, scene complexity, and land cover distributions. For the MUUFL dataset, the highest classification accuracy was achieved with a 7×7 patch size, as this compact size effectively captures the local spatial details prevalent in this scene. The performance of the model on the Trento dataset demonstrated notable robustness, with minimal fluctuations in accuracy for patch sizes ranging from 7×7 to 13×13. This suggests that the model is less sensitive to the spatial context scale in this particular agricultural environment. In contrast, the Houston2013 and Augsburg datasets attained their peak performance at a patch size of 11×11. This moderate size provides an optimal balance for these scenarios, offering sufficient spatial context for the discrimination of complex land cover categories without introducing excessive extraneous information that could act as noise. While the ideal patch size exhibits some dataset-specific variation, a size of 11×11 consistently delivers strong and stable performance across all datasets.

3) Influence of Dilation Rate Combinations in MSFE: We investigate the impact of different dilation rate combinations on classification performance. As shown in Table X, we evaluate configurations with varying numbers of parallel branches (N = 2, 3, 4) and different dilation rate sets.

For N = 2, the combination of dilation rates [2, 4] generally outperforms [1, 2] across most datasets, particularly on Augsburg and MUUFL. This suggests that capturing larger receptive fields is beneficial for distinguishing objects in complex scenes. When increasing to N = 3, the combination [1, 2, 3] achieves the best overall performance, obtaining the highest OA on Houston2013, Augsburg, and Trento. This balanced configuration captures both fine-grained details (dilation = 1) and broader contexts (dilation = 3). Further increasing to N = 4 with dilation rates [1, 2, 3, 4] does not yield additional improvements and even shows slight degradation on some datasets. This observation suggests that excessive parallel branches may introduce redundancy or dilute the discriminative power of features.

The consistent superiority of the [1, 2, 3] configuration across diverse datasets demonstrates its robustness and generalizability. This combination effectively balances the tradeoff between capturing local details and global contexts, which is crucial for handling the scale variations present in HSI and LiDAR data.

G. Robustness Evaluation

To assess the robustness of MCAFNet against data corruption, we conducted experiments with different types and intensities of noise added to the input data. The noise types include Gaussian noise, salt-and-pepper noise, and a mixture of both (70% Gaussian + 30% salt-and-pepper). The classification performance under varying noise levels is presented in Fig. 11.

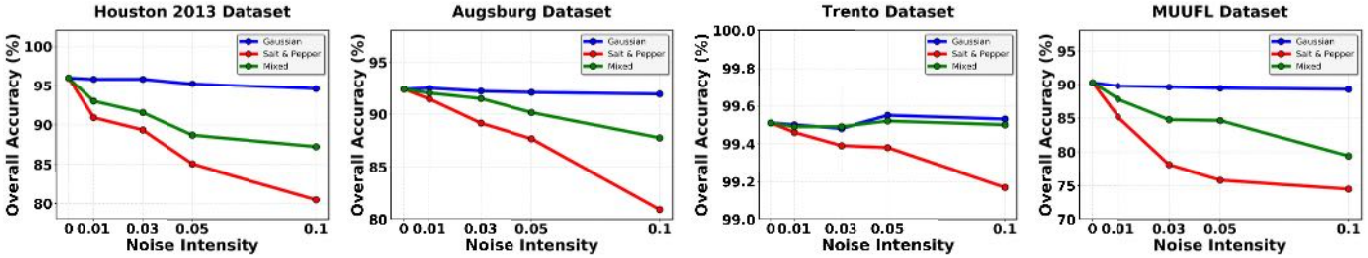


Fig. 11. Robustness evaluation of the proposed MCAFNet against varying noise intensities on the Houston 2013, Augsburg, Trento, and MUUFL datasets.

TABLE X
INFLUENCE OF DIFFERENT DILATION RATE COMBINATIONS ON CLASSIFICATION RESULTS OF MCAFNET

N	Dilation Rates	Houston2013				Augsburg				Trento				MUUFL			
		OA	AA	κ	F1	OA	AA	κ	F1	OA	AA	κ	F1	OA	AA	κ	F1
N=2	1, 2	94.97	95.44	94.53	95.60	91.17	69.07	87.44	65.43	99.46	**99.24**	99.28	98.72	89.35	85.30	86.04	77.90
	2, 4	95.12	95.62	94.70	95.48	92.12	68.26	88.68	65.77	99.46	99.13	99.27	98.93	**90.48**	84.15	**87.45**	78.05
N=3	1, 2, 3	**96.01**	**96.57**	**95.67**	**96.15**	**92.39**	**71.37**	**89.11**	**67.15**	**99.51**	99.23	**99.35**	98.86	90.30	86.34	87.27	78.64
	1, 3, 5	95.71	96.16	95.35	94.20	91.93	68.19	88.40	66.61	99.49	**99.24**	99.32	98.73	90.02	84.70	86.87	76.87
N=4	1, 2, 3, 4	95.24	95.71	94.84	95.88	92.16	69.35	88.76	66.69	99.40	98.92	99.20	**98.97**	89.86	**86.38**	86.69	**78.99**

The results demonstrate that MCAFNet maintains relatively stable performance under Gaussian noise across all intensity levels, which can be attributed to the multiscale feature extraction and adaptive channel weighting in the MSFE module. For salt-and-pepper noise, the performance degradation is more pronounced, particularly at higher noise intensities, as this type of noise creates extreme pixel value outliers that disrupt local spatial patterns. However, even under severe salt-and-pepper noise, MCAFNet maintains reasonable classification accuracy due to the combination of multiscale processing and cross-modal feature reinforcement.

The mixed noise scenario shows intermediate performance between pure Gaussian and salt-and-pepper noise. Notably, the Trento dataset exhibits exceptional robustness across all noise conditions, likely due to the distinct spectral and elevation characteristics of its agricultural land cover. In contrast, the more complex urban scenes in Houston2013 and MUUFL show greater sensitivity to noise, particularly salt-and-pepper noise that disrupts fine spatial structures. Overall, these robustness evaluations confirm that MCAFNet's architectural components contribute to resilience against various types of data corruption, making it suitable for real-world remote sensing applications where data quality may vary.

V. CONCLUSION

This article proposed a novel MCAFNet for joint classification of HSI and LiDAR data. The network employs a three-branch architecture to extract LiDAR elevation features, HSI global spectral–spatial features, and multimodal fusion features, effectively addressing the limitations of existing methods in cross-modal interaction depth and multiscale adaptability. In addition, we design an MSFE module that leverages depthwise separable and multiscale dilated convolutions to capture rich spatial–spectral information across varying object scales. We also design an MBCA module that enables deep bidirectional interaction among HSI, LiDAR, and multimodal branches through feature reorganization and triple residual connections, effectively enhancing semantic alignment and information retention. Furthermore, extensive experiments conducted on four datasets validate the effectiveness and generalization capability of the proposed method. Future work will focus on two main directions to enhance the model's practical utility. First, lightweight designs will be explored to reduce the parameter count and FLOPs, addressing the moderate computational overhead observed in Table V while aiming to preserve the performance gains demonstrated across all four datasets. Second, dynamic inference strategies will be investigated, inspired by the dataset-specific sensitivities to patch size and feature dimension shown in Figs. 9 and 10. The goal is to develop an adaptive mechanism that can adjust the model's receptive field and representation capacity based on input scene complexity, thereby optimizing the tradeoff between accuracy and efficiency for diverse real-world scenarios.

REFERENCES

[1] J. Yue, L. Fang, and M. He, "Spectral–spatial latent reconstruction for open-set hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 5227–5241, 2022.

[2] H. Shirmard, E. Farahbakhsh, R. D. Müller, and R. Chandra, "A review of machine learning in processing remote sensing data for mineral exploration," *Remote Sens. Environ.*, vol. 268, Jan. 2022, Art. no. 112750.

[3] C. Wang, L. Zhang, W. Wei, and Y. Zhang, "Dynamic super-pixel normalization for robust hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505713.

[4] N. Jiang, H.-B. Li, C.-J. Li, H.-X. Xiao, and J.-W. Zhou, "A fusion method using terrestrial laser scanning and unmanned aerial vehicle photogrammetry for landslide deformation monitoring under complex terrain conditions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707214.

[5] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 68–80, Aug. 2021.

[6] J. Zhang, J. Lei, W. Xie, G. Yang, D. Li, and Y. Li, "Multimodal informative ViT: Information aggregation and distribution for hyperspectral and LiDAR classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7643–7656, Aug. 2024.

[7] W. Dong, T. Yang, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Joint contextual representation model-informed interpretable network with dictionary aligning for hyperspectral and LiDAR classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6804–6818, Nov. 2023.

[8] Y. Kong, S. Yu, Y. Cheng, C. L. Philip Chen, and X. Wang, "Joint classification of hyperspectral images and LiDAR data based on candidate pseudo labels pruning and dual mixture of experts," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5505912.

[9] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 82, pp. 1–18, Jun. 2022.

[10] J. Lin, F. Gao, L. Qi, J. Dong, Q. Du, and X. Gao, "Dynamic cross-modal feature interaction network for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5508916.

[11] P. Ghamisi et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[12] K. Liu, T. Sun, H. Zeng, Y. Zhang, C.-M. Pun, and C.-M. Vong, "Spatial-aware conformal prediction for trustworthy hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 9, pp. 8754–8766, Sep. 2025.

[13] D. Li, W. Xie, Z. Wang, Y. Lu, Y. Li, and L. Fang, "FedDiff: Diffusion model driven federated learning for multi-modal and multi-clients," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10353–10367, Oct. 2024.

[14] S. Jia et al., "Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1437–1452, Feb. 2021.

[15] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[16] J. Li, Y. Liu, R. Song, Y. Li, K. Han, and Q. Du, "Sal^{2}RN: A spatial–spectral salient reinforcement network for hyperspectral and LiDAR data fusion classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500114.

[17] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.

[18] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[19] Y. Zhang, W. Li, W. Jia, M. Zhang, R. Tao, and S. Liang, "Cross-domain hyperspectral image classification based on bi-directional domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 12038–12051, Dec. 2025.

[20] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.

[21] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[22] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, 2023.

[23] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[24] Y. Li, D. Li, W. Xie, J. Ma, S. He, and L. Fang, "Semi-mamba: Mamba-driven semi-supervised multimodal remote sensing feature classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 10, pp. 9837–9849, Oct. 2025.

[25] 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.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[28] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[29] X. Wang, L. Song, Y. Feng, and J. Zhu, "S3F2Net: Spatial–spectral-structural feature fusion network for hyperspectral image and LiDAR data classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 4801–4815, May 2025.

[30] Y. Zhang et al., "A cross-modal feature aggregation and enhancement network for hyperspectral and LiDAR joint classification," *Expert Syst. Appl.*, vol. 258, Dec. 2024, Art. no. 125145.

[31] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

[32] Z. Ye, Y. Zhang, J. Zhang, W. Li, and L. Bai, "A multiscale incremental learning network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5606015.

[33] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.

[34] D. S. Hoffmann, K. N. Clasen, and B. Demir, "Transformer-based multi-modal learning for multi-label remote sensing image classification," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 4891–4894.

[35] D. Liao, Q. Wang, T. Lai, and H. Huang, "Joint classification of hyperspectral and LiDAR data based on mamba," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5530915.

[36] H. Zhu et al., "ConvGRU-based multiscale frequency fusion network for PAN-MS joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5406415.

[37] W. Ma et al., "Intra- and intersource interactive representation learning network for remote sensing images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5401515.

[38] L. Xu et al., "A dual-stream transformer with diff-attention for multi-spectral and panchromatic classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5409114.

[39] H. Gao et al., "Interactive enhanced network based on multihead self-attention and graph convolution for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5533716.

[40] S. Feng, H. Deng, Y. Hu, C. Zhao, W. Li, and R. Tao, "Fractional Fourier-enhanced fusion network based on Pareto optimization for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5513416.

[41] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "FusionMamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Vis. Intell.*, vol. 2, no. 1, pp. 1–18, Dec. 2024.

[42] W. Dong et al., "Fusion-mamba for cross-modality object detection," *IEEE Trans. Multimedia*, vol. 27, pp. 7392–7406, 2025.

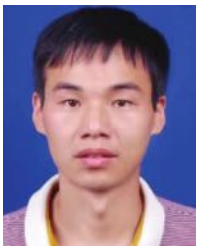
[43] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[44] C. Liang, Y. Zhao, Y. Song, and K. Ni, "IDNet: Intensity-constrained detail-enhanced network for hyperspectral and LiDAR collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5518515.

[45] Q. Song et al., "MCFNet: Multiscale cross-domain fusion network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4703912.

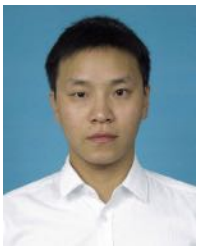
Junhui Cai received the B.S. degree from Yangtze College, East China Institute of Technology, Fuzhou, China, in 2024. He is currently pursuing the M.S. degree in computer technology with East China Jiaotong University, Nanchang, China.

He is currently a Researcher with the Institute for Data Science and Deep Learning, East China Jiaotong University. His research interests include deep learning and hyperspectral image classification.



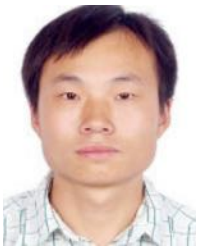
Xiaohui Huang received the B.Eng. and master's degrees from Jiangxi Normal University, Nanchang, China, in 2005 and 2008, respectively, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2014.

He is currently an Associate Professor of Computer Science with East China Jiaotong University, Nanchang. His research interests are in the areas of machine learning, deep learning, and clustering algorithms.



Xiaofei Yang received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2014 and 2019, respectively.

He was a Post-Doctoral Researcher with the Department of Computer and Information Sciences, University of Macau, Macau, China, from 2020 to 2023. He is currently with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou, China. His research interests are in the areas of semisupervised learning, deep learning, remote sensing, transfer learning, and graph mining.



Jiangtao Peng (Senior Member, IEEE) received the B.S. and M.S. degrees from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the Faculty of Mathematics and Statistics, Hubei University. His research interests include machine learning and hyperspectral image processing.

Yifang Ban (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Nanjing University, Nanjing, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996.

Before joining KTH Royal Institute of Technology (KTH), Stockholm, Sweden, in 2004, she was a Tenured Associate Professor at York University, Toronto, ON, Canada. She is currently the Chair Professor and the Director of the Geoinformatics Division, KTH Royal Institute of Technology, and an Associate Director of Digital Futures, Stockholm. Her research interests include Earth observation big data analytics, machine learning/deep learning, and their applications in mapping urban and land cover, monitoring urbanization, wildfires, and other environmental changes, and assessing environmental impact.

Yicong Zhou (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tufts University, Medford, MA, USA, in 2010.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE). He was recognized as one of the "Highly Cited Researchers" in 2020, 2021, 2023, and 2024. He serves as a Senior Area Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.