M³amba: CLIP-driven Mamba Model for Multi-modal Remote Sensing Classification

Mingxiang Cao, Weiying Xie, *Senior Member, IEEE*, Xin Zhang, Jiaqing Zhang, Kai Jiang, Jie Lei, *Member, IEEE*, Yunsong Li, *Member, IEEE*

Abstract—Multi-modal fusion holds great promise for integrating information from different modalities. However, due to a lack of consideration for modal consistency, existing multi-modal fusion methods in the field of remote sensing still face challenges of incomplete semantic information and low computational efficiency in their fusion designs. Inspired by the observation that the visual language pre-training model CLIP can effectively extract strong semantic information from visual features, we propose M³amba, a novel end-to-end CLIP-driven Mamba model for multi-modal fusion to address these challenges. Specifically, we introduce CLIP-driven modality-specific adapters in the fusion architecture to avoid the bias of understanding specific domains caused by direct inference, making the original CLIP encoder modality-specific perception. This unified framework enables minimal training to achieve a comprehensive semantic understanding of different modalities, thereby guiding cross-modal feature fusion. To further enhance the consistent association between modality mappings, a multi-modal Mamba fusion architecture with linear complexity and a cross-attention module Cross-SS2D are designed, which fully considers effective and efficient information interaction to achieve complete fusion. Extensive experiments have shown that M³amba has an average performance improvement of at least 5.98% compared with the state-of-the-art methods in multi-modal hyperspectral image classification tasks in the remote sensing field, while also demonstrating excellent training efficiency, achieving a double improvement in accuracy and efficiency. The code is released at https://github.com/kaka-Cao/M3amba.

Index Terms—Deep learning, Remote Sensing, Multi-modal, Feature Fusion, CLIP model, Mamba.

I. INTRODUCTION

Multi-modal fusion tasks involve integrating features from various data modalities, such as visible light, infrared, and LiDAR, to achieve more complete feature representations [1]–[3]. This is crucial for applications like remote sensing scene understanding [4]. Unfortunately, due to hardware limitations, a single sensor cannot capture all the complex details of an image [5], but the semantic information in multi-modal images is
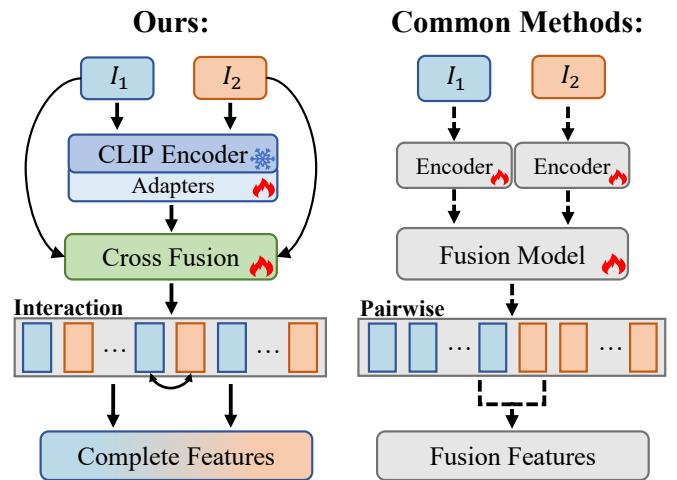
Fig. 1. **Left:** Proposed CLIP-driven modality-specific adapters guide the fusion process to produce complete features through semantic information interactions. **Right:** Common methods perform pairwise fusion by training encoders and fusion networks, which lacks consideration of semantic consistency and leads to incomplete representation. I_1 and I_2 represent inputs from different modalities.

always consistent, so it is particularly important to effectively leverage comprehensive semantic information across different modalities to guide fusion [6]. Existing methods focus on training different models for semantic feature extraction, but these models have limitations. CNN-based approaches cannot capture comprehensive semantic information due to their limited receptive fields, while Transformer-based methods have global modeling capabilities but incur high training costs and lack transferability. The advent of Contrastive Language-Image Pre-training (CLIP) [7] offers a solution to these issues. CLIP achieves significant success in various visual tasks by pre-training with a large corpus of image-text pairs obtained from the web. Given its robustness and transferability, CLIP is often used to directly infer effective image features. For example, Lin *et al.* [8] used frozen CLIP output features to capture semantic information between adjacent frames. Yang *et al.* [9] used CLIP image embedding for global understanding, allowing the model to learn from high-level semantic information. This confirms CLIP's powerful ability to learn rich semantic visual representations and its ability to transfer to other visual tasks. However, direct inference often results in weak perception capabilities of downstream tasks [10], and the potential of this approach in the field of multi-modal images remains underexplored.

It is noteworthy that multi-modal image data captured and processed by different sensors often exhibit characteristics of consistency and complementarity [11]. These traits can significantly enhance the capabilities of deep learning models, facilitating more detailed and nuanced task interpretation. Thus, obtaining complete multi-modal fusion features is crucial for optimizing tasks. Compared to traditional linear concatenation methods in CNN architectures, approaches based on attention mechanisms have shown satisfactory results. Ma *et al.* [12] introduced a novel universal image fusion framework based on cross-domain distance learning and Swin Transformer, achieving comprehensive integration and global interaction of multi-modal complementary information. Although the emergence of self-attention mechanisms has facilitated the interaction of information flow between modal features, they bring challenges of quadratic computational requirements and lack the rich semantic representation capabilities of multi-modal fusion features.

Thus, a natural question arises: Can we obtain comprehensive visual semantic information from different modalities with minimal computational overhead to guide a fusion network for complete feature fusion? To overcome the shortcomings of existing methods, we propose the first CLIP-driven **M**amba model for **Multi-M**odal fusion, named **M^{3}amba**. This unified framework benefits from the linear time complexity provided by Mamba [13] and demonstrates improved transferability through CLIP image encoder and modality-specific adapters, as shown in the application transfer from the natural image domain to the remote sensing domain. Specifically, as shown in Fig. 1, we implant CLIP-driven modality-specific adapters into the encoder branch to introduce modality-specific semantic understanding with minimal training, thereby learning comprehensive visual semantic representations. This semantically guided cross-modal fusion enhances the understanding of global and local associative sequences, ensuring more robust and coherent feature representations. To achieve complete fusion, we introduce a three-branch state space model (SSM) that effectively interacts with CLIP semantic features through the proposed Cross-SS2D module to extract consistent and complementary information. This fusion process models the feature space of sequential data without the need for quadratic computational complexity, thus capturing inter-modal dependencies at minimal computational cost. While common methods perform pairwise fusion by training the encoder and fusion network from scratch, incomprehensive semantic information leads to incomplete fusion features and often redundant representations. Extensive experiments show that M^{3}amba outperforms current mainstream CNN, Transformer, and Mamba architectures in both performance and training efficiency. This reflects the great potential of the CLIP and Mamba dual-driven unified framework in multi-modal fusion tasks. To summarize, our contributions are three-fold:

- We propose M^{3}amba, a novel end-to-end CLIP-driven Mamba model for effective and efficient multi-modal fusion. To our best knowledge, M^{3}amba is the first model that synergistically optimizes the powerful visual capabilities of CLIP and the efficient computational performance of Mamba.

- We design a multi-modal Mamba fusion architecture embedded with the proposed Cross-SS2D module, aiming to capture the comprehensive representation with linear complexity by enhancing the interactivity between modality mappings.

- Extensive experiments in the field of multi-modal remote sensing show that M^{3}amba surpasses SOTA methods in terms of effectiveness and training efficiency, demonstrating the potential of the CLIP and Mamba dual-driven multi-modal fusion framework.

II. RELATED WORKS

A. Large-scale Image Representation Learning

With the emergence of weakly labeled data at the scale of entire websites, we have witnessed a surge in new models for general visual representation learning. Concurrently, the scale of image models built through regular supervised learning has also been rapidly increasing [14], [15]. To further enhance visual representation capabilities, researchers have begun focusing on contrastive learning and self-supervised learning with large datasets and large models. The success of BERT [16] has sparked an emerging direction of constructing large-scale visual models using masked visual modeling [17], [18]. Vision-language models (VLMs), such as CLIP [7] and ALIGN [19], are trained on billion-scale, often noisy image-text datasets. These models consist of modality-specific encoders (image and text) that generate embeddings for each modality, therefore CLIP has been widely used as the vision backbone for its semantic representative feature and promising scalability even in the age of Multi-modal Large Language Models (MLLMs). Additionally, several methods apply the CLIP image encoder to 2D image variations to achieve better quality or controllability [20], [21]. Some methods use the CLIP image encoder to extract features for rendering images [22]–[24], while others adopt CLIP to extract semantically rich video representations [25]–[27]. Approaches like LiT [28] and BLIP-2 [29] reduce the training costs of CLIP-like models by deploying pre-trained unimodal models.

Inspired by these developments, we utilize a pre-trained CLIP image encoder to extract semantic information from multi-modal images. However, its perception of a specific modality often suffers from understanding deviations due to domain gaps. Therefore, we introduce modality-specific adapters to further enhance intra-modal semantic understanding with minimal training overhead.

It is worth noting that in our implementation, we choose CLIP instead of ALIGN mainly based on the following considerations: 1) CLIP has stronger migration ability and more stable performance; 2) CLIP provides a more complete pre-trained model and interface; 3) Research in the field of remote sensing shows that CLIP is more suitable for processing the semantic features of remote sensing images [30].

B. Multi-modal Fusion in Remote Sensing

Faced with the increasingly complex and diverse needs of scene representation applications, deep learning has made significant contributions to overcoming the technical bottleneck
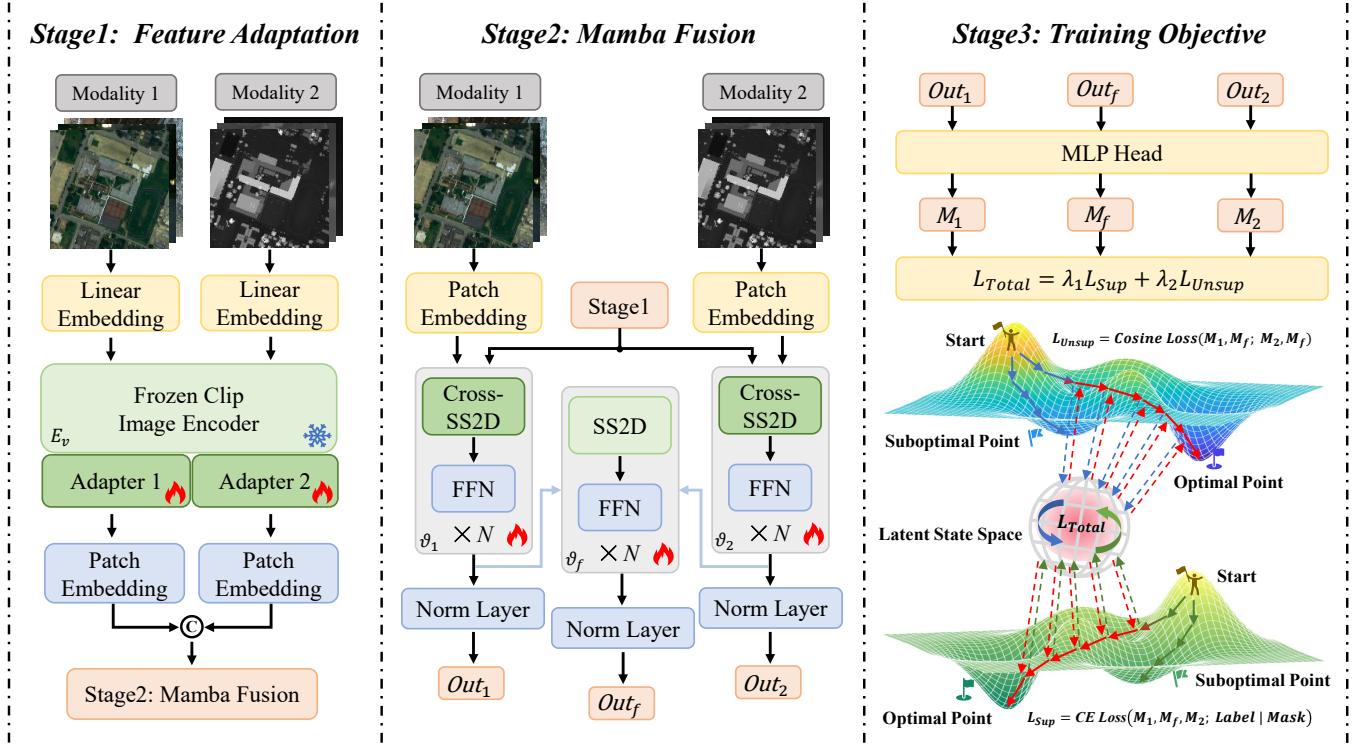
Fig. 2. Overview of the M³amba framework. For clarity, we split the end-to-end process of training into three stages: Feature Adaptation, Mamba Fusion with Cross-SS2D, and Training Objective. By utilizing the M³amba framework to perform a feature-level fusion of the two modalities, we can apply the complete fusion features to different downstream tasks. MLP Head consists of several convolutional and linear layers, and CE stands for Cross Entropy.

of multi-modal feature fusion [31], [32], and researchers have been advancing the study of multi-modal classification tasks in remote sensing, especially through CNN and Transformers [33]–[41]. Although these methods integrate complementary information between different modalities, they are still limited in terms of computational efficiency and fusion completeness. Recently, Mamba has emerged as a promising candidate for the next generation of base model backbones as it exhibits better scaling than Transformers while maintaining linear time complexity and has been developed in the field of multi-modal remote sensing. SpectralMamba [42] achieved improved performance by dynamically simplifying but fully modeling hyperspectral data in both spatial-spectral space and hidden state space. S²Mamba [43] mined spatial-spectral context features to achieve more efficient and accurate land cover analysis. MiM [44] introduced an innovative deployment architecture for hyperspectral image classification, improving the model's efficiency through a novel centralized Mamba Cross Scan (MCS) mechanism and T-Mamba encoder design.

Compared to these methods, M³amba uses comprehensive semantic features adapted by CLIP to guide the fusion of multi-modal features with linear complexity, while Cross-SS2D ensures feature consistency and uses complementary information as gain.

III. METHODOLOGY

A. Problem Formulation

In recent years, multi-modal data fusion has emerged as a crucial technique for enhancing the performance of various tasks, particularly in the field of remote sensing. In this paper, we focus on the application of multi-modal fusion to the pixel-level classification problem. This can be defined as accurately assigning each pixel in images from two modalities to the corresponding class. Given an image X consisting of q pixels, we aim to perform a pixel-level classification task using data from different modalities $X_1, X_2 \in \mathbb{R}^{h \times w \times c}$. Both modalities learn features of the same scene with label information $L \in \mathbb{R}^{h \times w \times m}$, and the label for the p-th pixel can be represented as a one-hot vector $L^q = \{0^{m-1}, 1\}$, where m denotes the number of classes. The objective of multi-modal pixel-level classification is to develop and train a model $\psi(X_1, X_2)$ that effectively maps images from different modalities to the classification distribution $C_{max}(X_1, X_2; L)$, indicating the probability that each pixel is associated with different classes under the guidance of the labels. A binary prediction map can be obtained through hard classification by a threshold τ on the maximum probability for different classes.

$$\psi(X_1, X_2) = \begin{cases} 0, & \text{if } C_{\max}(X_1, X_2; L) < \tau, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

Based on the above objectives, the logits can be expressed as $C_{max}(X_1, X_2; L)$, specifically formulated as $C_{max}(X_1, X_2; L) = \vartheta(X_1, X_2 \mid \beta_1, \beta_2)$. Here, we map the image feature space to the classification space using a nonlinear target model $\vartheta(\cdot)$, where β_1 and β_2 represent the parameters of the two modality branches.

B. Preliminaries

State Space Models State Space Models (SSMs) represent a class of sequence-to-sequence modeling systems characterized

by constant dynamics over time, often used to represent linear time-invariant systems. Due to their linear complexity, SSMs can efficiently capture the inherent dynamics of a system through implicit mapping to latent states. Mathematically, an SSM is typically represented as a linear ordinary differential equation (ODE):

$$\frac{dh(t)}{dt} = Ah(t) + Bx(t), y(t) = Ch(t) + Dx(t), \quad (2)$$

where $x(t) \in \mathbb{R}$, $h(t) \in \mathbb{R}^n$, and $y(t) \in \mathbb{R}$ represent the input, hidden state, and output, respectively, with n being the state size. The remaining parameters include the state transition matrix $A \in \mathbb{R}^{n \times n}$, projection parameters $B \in \mathbb{R}^{n \times 1}$, $C \in \mathbb{R}^{1 \times n}$, and skip connection $D \in \mathbb{R}$.

To handle discrete sequences such as images and text, the ODE needs to be converted into a discrete function by introducing a predefined time scale parameter $\Delta \in \mathbb{R}^D$. The discretization process of the above equations is as follows:

$$\overline{A} = \exp(\Delta A), \overline{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B, \quad (3)$$

$$h^t = \overline{A}h^{t-1} + \overline{B}x^t, y^t = Ch^t + Dx^t, \quad (4)$$

all matrices maintain the same dimensions across iterations of the operation. Furthermore, following Mamba, the matrix $\overline{B}$ can be approximated by a first-order Taylor series:

$$\overline{B} = (\exp(A) - I)A^{-1}B \approx (\Delta A)(\Delta A)^{-1}\Delta B = \Delta B. \quad (5)$$

2D-Selective-Scan Mechanism Due to the one-dimensional and causal properties of the Mamba selection mechanism, directly applying Mamba to visual tasks is unsuitable. For instance, while two-dimensional spatial information plays a crucial role in vision-related tasks, it is secondary in one-dimensional sequence modeling. Additionally, causal processing of input data prevents Mamba from absorbing information from parts of the data not yet scanned. These differences result in a limited receptive field, unable to capture potential correlations with unexplored patches. VMamba [45] introduces a Two-Dimensional Selective Scanning (SS2D) mechanism to solve this problem. SS2D transforms the input image into patch sequences along the horizontal and vertical axes, scanning in four directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right, generating independent sequences. This four-way scanning ensures that each element in the feature map contains information from all other positions in various directions. Consequently, it establishes a comprehensive global receptive field without linearly increasing computational complexity. The generated four sequences are then individually processed using selective SSM. Finally, all feature sequences are transformed back to their original 2D layout and merged to reconstruct the 2D feature map. SS2D is a core component of the Visual State Space (VSS) module, as shown in Figure 2, on which we build the hidden state space for cross-modal feature fusion.

C. Network Architecture

Overview The overall structure of M³amba, illustrated in Figure 2, is a multi-modal Mamba model built on top of a fixed CLIP image encoder, with images from different modalities

---

Algorithm 1 Algorithm of Cross-SS2D.

Input: Different modalities of data X_1, X_2.
Parameter: Weight matrices $weight_1$, $bias_1$, $weight_2$, $bias_2$, A_1, B_1, A_2, B_2.
Output: Fusion feature H_f.

1: **Initial** A_1, B_1, A_2, B_2.
2: **for** each block N = 1, 2, ..., 9 **do**
3:     **1. Compute System Matrices**
4:       $As_1 \leftarrow -exp(A_1) \quad As_2 \leftarrow -exp(A_2)$
5:       $Bs_1 \leftarrow Split(einsum(X_1, weight_1) + bias_1)$
6:       $Bs_2 \leftarrow Split(einsum(X_2, weight_2) + bias_2)$
7:     **2. Cross Assignment**
8:       $As_f \leftarrow (As_1 + As_2)/2$
9:       $Bs_1 \leftrightarrow Bs_2$
10:     **3. Extract Fusion Features**
11:       $H_f \leftarrow selective\_scan(As_f, Bs_1, Bs_2)$
12: **end for**
13: **return** H_f to the model.

---

used as inputs. Features corresponding to the same class from different modalities contain both consistent information and complementary insights. For instance, LiDAR provides enhanced spatial resolution, while hyperspectral data offers superior spectral resolution; infrared sensors excel at capturing thermal radiation, and visible light sensors are proficient at detailed texture data. However, a single sensor cannot capture all the complex details of an image, but the semantic information of images under the same scene is always consistent. Therefore, our goal is to **maximize the description of consistent information while leveraging complementary insights as benefits under comprehensive semantic feature guidance**.

Inspired by this, our proposed method aims to balance the informational discrepancies between the two modalities while utilizing their consistency and complementarity. For clarity, our network structure is conceptually divided into two parts, but it operates as a single end-to-end training process. Initially, modality-specific adapters are embedded in the pre-trained CLIP image encoder to capture comprehensive semantic information from different modalities. The output features, along with the input images, are then processed by Mamba's three-branch feed-forward network: ϑ_1, ϑ_2, and ϑ_f. Guided by an unsupervised consistency loss, the supervised fusion network with linear complexity is optimized to fully capture the features across modalities. Specifically, the fusion network consists of three main branches: two consistency branches ϑ_1 and ϑ_2, ϑ_1 processes the feature sequence of the first modality (HSI), and ϑ_2 processes the feature sequence of the second modality (LiDAR). They share the same network architecture, but use independent parameters to maintain the uniqueness of each modality feature, and map the unimodal and comprehensive semantic features to the hidden state space for interaction, so that complementary features and their consistency information can be thoroughly extracted. Additionally, the fusion branch ϑ_f employs a selective scanning mechanism for global modeling, further integrating the complete fusion

features. The entire fusion network undergoes N iterative updates, and the classification outputs of the three branches are finally used for loss function computation.

Feature Adaptation By learning from a large number of image-text pairs, CLIP can match images with their corresponding natural language descriptions. During training, the encoders are optimized by extracting features from input samples and aligning them in the embedding space using contrastive loss. This enables zero-shot classification and highlights CLIP's inherent advantage in extracting deep semantic visual feature representations. Leveraging this, we use the CLIP image encoder E_v to infer multi-modal input images X_1 and X_2, mapping them to the same feature space to reduce the domain gap between modalities and learn visual representations with rich semantics. In addition, in order to endow the original CLIP encoder with modality-specific perception capabilities for guiding more comprehensive feature fusion, we introduce modality-specific adapters. The resulting multi-modal features Y_1 and Y_2 have comprehensive modal scene semantic understanding and ensure more robust and coherent feature representations. This can be expressed as the equation:

$$Y_1, Y_2 = E_v^{Adapter_1}(X_1), E_v^{Adapter_2}(X_2), \quad (6)$$

where $E_v^{Adapter_1}$ and $E_v^{Adapter_2}$ represent the CLIP image encoder with modality-specific adapters added. The learnable adapters consist of two linear layers with residual connections, inserted after each attention block. We discard the last layer of the encoder because it is dedicated to classification. The resulting output features Y_1 and Y_2 are concatenated into Z after feature size transformation to provide guidance for the fusion network.

Mamba Fusion As analyzed in the related work, previous multi-modal fusion methods excessively focused on the differing relationships between modalities, often neglecting the consistency across the data. Our fusion objective is to obtain complete fusion features between modalities, involving capturing local differences and overall invariants. By leveraging the obtained CLIP semantic feature Z, we facilitate the fusion process to learn complete feature representations. Therefore, the entire encoding process can be expressed as:

$$Out_1 = \vartheta_1(X_1, Z|\beta_1), \quad (7)$$

$$Out_2 = \vartheta_2(X_2, Z|\beta_2), \quad (8)$$

$$Out_f = \vartheta_f(Out_1 + Out_2|\beta_f), \quad (9)$$

where Out_1, Out_2, and Out_f are the outputs of the three branches, and β_1, β_2, and β_f are their corresponding parameters. The fusion of features from both modalities significantly impacts the performance of downstream tasks. To that effect, designing a cross-modal complete attention mechanism becomes crucial.

Cross-SS2D As shown in Algorithm 1, the Cross-SS2D takes two features as input and generates a fused output while preserving the original shape of the features. The two inputs come from the comprehensive semantic features of CLIP and the unimodal image features, fully extracting independent features from each modality to complement the fused features and leveraging complementary insights to encompass

comprehensive modal information. To achieve the model's context-aware capability, linear projection layers are employed to produce the system matrices B, C, and Δ from the inputs. As specified in Equation 4, matrices $\overline{A}$ and $\overline{B}$ encode the previous hidden state h^{t-1} and the input x^t to compute the current state h^t.

Inspired by the cross-attention mechanisms [46] widely used in multi-modal tasks, we aim to facilitate the interaction of information between multiple modalities within the 2D selective scanning module. To achieve this goal, we use the $\overline{B}$ matrix generated by the complementary modality in the selective scanning operation, enabling the SSM to provide complementary information for the current modality guided by the features from another modality. Additionally, the $\overline{A}$ matrices corresponding to the two modalities are averaged to provide consistent information for the multi-modal input, ultimately resulting in the fusion feature H_f. Specifically, this process can be expressed as:

$$\overline{A}_1 = \exp(\Delta_1 A_1), \overline{A}_2 = \exp(\Delta_2 A_2), \quad (10)$$

$$\overline{B}_1 = \Delta_1 B_1, \overline{B}_2 = \Delta_2 B_2, \quad (11)$$

$$h_1^t = (\overline{A}_1 + \overline{A}_2)h_1^{t-1}/2 + \overline{B}_2 x_1^t, \quad (12)$$

$$H_f = y_1^t = C_1 h_1^t + D_1 x_1^t, \quad (13)$$

where Δ is a predefined time scale parameter, x^t represents the input at time t, and y^t denotes the selective scan output. Subscripts 1 and 2 represent unimodal features and CLIP features, respectively. To prevent redundant fusion, we choose y_1^t as the fusion feature. $\overline{A}_1$ and $\overline{A}_2$ are used as the averaged state transition matrices providing multi-modal consistency, $\overline{B}_1$ and $\overline{B}_2$ are used as the cross-modal projection matrices providing complementary information. By implementing this Cross-SS2D fusion mechanism, we achieve a balanced and comprehensive representation of multi-modal features, leveraging both the complementary and consistent aspects of the input data. In addition, the A matrix enhances the feature consistency between modalities, making the fusion process smoother. The B matrix guides feature selection and avoids the transmission of redundant information. Through the interaction of the A matrix and the B matrix, Cross-SS2D can establish an efficient information transmission mechanism between modalities, improving the efficiency of feature selection and fusion between different modalities.

D. Training Objective

Given two modality inputs X_1^i and X_2^i, our training strategy involves enhancing the supervised fusion network with linear complexity under the direction of an unsupervised loss to fully capture consistent and complementary features between modalities, where X_1^i represents the i-th sample of the first modality. For the supervised loss, we create an object mask map based on the ground truth, aiming to promote the consistency of the output probability distribution with the label L under the strong supervision constraint of the mask. The supervised loss calculation is as follows:

TABLE I
ABLATION STUDIES ON THE HOUSTON2013, THE AUGSBURG, AND THE MUUFL DATASET. THE BEST RESULT IS **HIGHLIGHTED**

Scheme	Houston2013			Augsburg			MUUFL		
	OA(%)	AA(%)	$\kappa(\times100)$	OA(%)	AA(%)	$\kappa(\times100)$	OA(%)	AA(%)	$\kappa(\times100)$
(A)	95.02	95.06	94.66	95.65	84.52	93.72	96.25	91.03	95.00
(B)	95.84	95.84	95.55	96.27	86.68	94.66	97.16	94.70	96.22
(C)	96.85	96.86	96.62	97.47	88.88	96.36	97.40	95.02	96.54
(D)	**97.31**	**97.32**	**97.11**	**98.19**	**93.32**	**97.39**	**97.84**	**96.10**	**97.13**

TABLE II
COMPARISON RESULTS ON THE HOUSTON2013, THE AUGSBURG, AND THE MUUFL DATASET. THE BEST RESULT IS **HIGHLIGHTED**

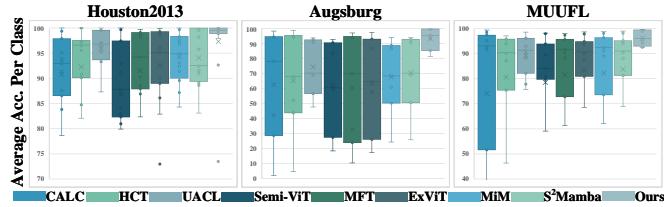| Method	Model	Houston2013			Augsburg			MUUFL		
		OA(%)	AA(%)	$\kappa(\times100)$	OA(%)	AA(%)	$\kappa(\times100)$	OA(%)	AA(%)	$\kappa(\times100)$
CALC[23]	CNN	88.97	90.78	88.06	91.73	62.62	87.98	93.94	74.09	92.00
HCT[23]	CNN&ViT	91.15	92.28	90.40	90.90	65.24	86.83	92.95	80.50	90.69
UACL[24]	CNN	95.37	95.99	95.00	89.19	74.30	84.80	88.29	89.11	84.78
Semi-ViT[22]	ViT	85.46	86.71	86.71	86.64	64.67	84.05	92.46	79.63	89.49
MFT[23]	ViT	89.80	91.51	88.89	90.49	60.36	86.26	94.34	81.48	92.51
ExViT[23]	ViT	91.40	92.60	90.66	87.91	63.20	82.82	94.37	83.37	92.54
MiM[24]	Mamba	92.89	94.21	92.28	88.63	67.57	86.56	92.65	82.22	91.59
S^2Mamba[24]	Mamba	93.36	94.09	92.79	89.34	70.41	89.45	94.19	83.72	91.98
M^3amba(Ours)	Mamba	**97.31**	**97.32**	**97.11**	**98.19**	**93.32**	**97.39**	**97.84**	**96.10**	**97.13**

Fig. 3. Comparison of box plots with other methods on three datasets.

$$\mathcal{L}_{Sup} = \sum_{i=1}^{X^i} \mathcal{L}_{CE}[(L, f_1(Out_1)); (L, f_2(Out_2));$$
$$(L, f_f(Out_f)) \mid Mask], \quad (14)$$

where $f_1(\cdot)$, $f_2(\cdot)$, and $f_f(\cdot)$ represent the MLP Head. To minimize the differences between the unique and complete feature predictions, we calculate the sum of cross-modal cosine similarity losses of three different features for unsupervised learning. The proposed unsupervised consistency loss is calculated as follows:

$$\mathcal{L}_{Unsup} = \sum_{i=1}^{X^i}[2 - \cos(f_1(Out_1), f_f(Out_f)) -$$
$$\cos(f_2(Out_2), f_f(Out_f))]. \quad (15)$$

The complete loss function consists of supervised loss and unsupervised consistency loss. However, there is a consensus that strong supervised learning can provide more accurate guidance for model optimization. Therefore, we utilize supervised loss as the primary update and the unsupervised consistency loss as auxiliary optimization, as shown in the following equation: $\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{Sup} + \lambda_2 \mathcal{L}_{Unsup}$. In particular, λ_1 is usually greater than λ_2.

Fig. 4. t-SNE for ablation on three datasets. The results from top to bottom correspond to Houston2013, Augsburg , and MUUFL datasets respectively.

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Settings

1) Datasets: To validate the effectiveness of the proposed method in analyzing multi-modal remote sensing data, we utilize three hyperspectral multi-modal datasets: the Houston2013 dataset [47], the Augsburg dataset [48], and the MUUFL dataset [49] for remote sensing classification tasks.
The Houston2013 Dataset The Houston2013 HSI-LiDAR Dataset is a widely used resource for remote sensing research, specifically designed for land cover classification tasks. It

TABLE III
A LIST OF THE NUMBER OF TRAINING AND TESTING SAMPLES FOR
EACH CLASS IN HOUSTON2013 DATASET

Class No.	Training	Testing	Class No.	Training	Testing
1	198	1053	9	193	1059
2	190	1064	10	191	1036
3	192	505	11	181	1054
4	188	1056	12	192	1041
5	186	1056	13	184	285
6	182	143	14	181	247
7	196	1072	15	187	473
8	191	1053			

TABLE IV
A LIST OF THE NUMBER OF TRAINING AND TESTING SAMPLES FOR
EACH CLASS IN AUGSBURG DATASET

Class No.	Training	Testing	Class No.	Training	Testing
1	146	13361	5	248	26609
2	7	1638	6	52	523
3	264	30065	7	23	1507
4	21	3830			

TABLE V
A LIST OF THE NUMBER OF TRAINING AND TESTING SAMPLES FOR
EACH CLASS IN MUUFL DATASET

Class No.	Training	Testing	Class No.	Training	Testing
1	1162	22084	7	112	2121
2	214	4056	8	312	5928
3	344	6538	9	69	1316
4	91	1735	10	9	174
5	334	6353	11	13	256
6	23	443			

information for terrain analysis. The scene covers 11 distinct land cover categories, with labeled pixels available for training and testing. The details are outlined in the Table V.

TABLE VI
AVERAGE TRAINING TIME ANALYSIS ON THREE DATASETS. THE BEST
RESULT IS **HIGHLIGHTED**

	Ours	MiM	ExViT	MFT	Semi-ViT	HCT	CALC
Train. (min)	**17.67**	18.71	44.03	42.62	50.92	24.28	29.15
AA (%)	**96.10**	82.22	83.37	81.48	79.63	80.50	74.09

TABLE VII
ABLATION STUDY OF FEATURE ADAPTATION WITH AVERAGE RESULTS
ON THREE DATASETS

	OA(%)	AA(%)	$\kappa(\times 100)$	Train. (min)
Infer	95.23	90.19	94.17	15.82
Fine-tune	98.10	95.36	97.40	46.08
Ours	97.78	95.58	97.21	17.67

features hyperspectral imagery (HSI) captured by the ITRES CASI-1500 sensor over the University of Houston and nearby rural areas in Texas, USA. This dataset was initially collected in June 2012 and later made available for the IEEE GRSS Data Fusion Contest in 2013. The HSI comprises 144 spectral bands, spanning from 380 nm to 1050 nm, with a spatial resolution of 10 nm, while the LiDAR data provide a single-channel elevation map. Both data modalities share a spatial resolution of 2.5 meters, and the entire dataset consists of 349×1905 pixels. Predefined training and testing sets are available for classification tasks, with details in Table III.

The Augsburg Dataset The Augsburg HS-SAR-LiDAR Dataset originates from a multi-modal data collection effort in Augsburg, Germany, and includes hyperspectral, synthetic aperture radar (SAR), and digital surface model (DSM) images. The hyperspectral imagery was collected using the HySpex sensor, delivering data across 180 spectral bands between 400 nm and 2500 nm. The selected study area includes 332,485 pixels with a ground sampling distance (GSD) of 30 meters. Additionally, the dataset provides four SAR bands and DSM elevation data, with the ground truth derived from manual labeling. This dataset is ideal for urban area classification, covering various land cover types. The details are outlined in the Table IV.

The MUUFL Dataset The MUUFL Gulfport Scene Dataset was gathered in November 2010 at the Gulf Park Campus of the University of Southern Mississippi in Long Beach, Mississippi. The dataset consists of hyperspectral images, captured by the CASI-1500 sensor, with 64 available spectral bands ranging from 375 nm to 1050 nm. The spatial resolution is approximately 0.54 by 1.0 meters. The dataset also includes LiDAR elevation data from two grates, providing further

2) Evaluation metrics: To evaluate the performance in classification, we employ three metrics: Overall Accuracy (OA), Average Accuracy (AA), and the kappa (κ). OA measures the ratio of correctly classified samples to the total number of samples. AA represents the average accuracy across all classes. The κ coefficient is a statistical metric assessing the agreement between the classification map generated by the model and the ground truth.

3) Implement details: All experiments are conducted on a server equipped with an NVIDIA A100 Tensor Core GPU. For hyperspectral remote sensing images, due to their dense objects and high resolution, the data samples are usually cropped to a size of 32 × 32 and use patch size 1. For all datasets, we use ViT-B/16 as the backbone of CLIP, and the training process employs the AdamW optimizer with an initial learning rate set to 1e-4, a weight decay of 1e-3, and a batch size of 8, spanning 200 epochs.

B. Ablation Study

1) Ablation analysis of different components: To evaluate the individual contributions of different components in our

TABLE VIII
OA, AA AND KAPPA COEFFICIENT ON THE HOUSTON2013 DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED**.
H AND L RESPECTIVELY INDICATE THAT OUR METHOD IS TRAINED USING ONLY HSI OR LiDAR DATA.

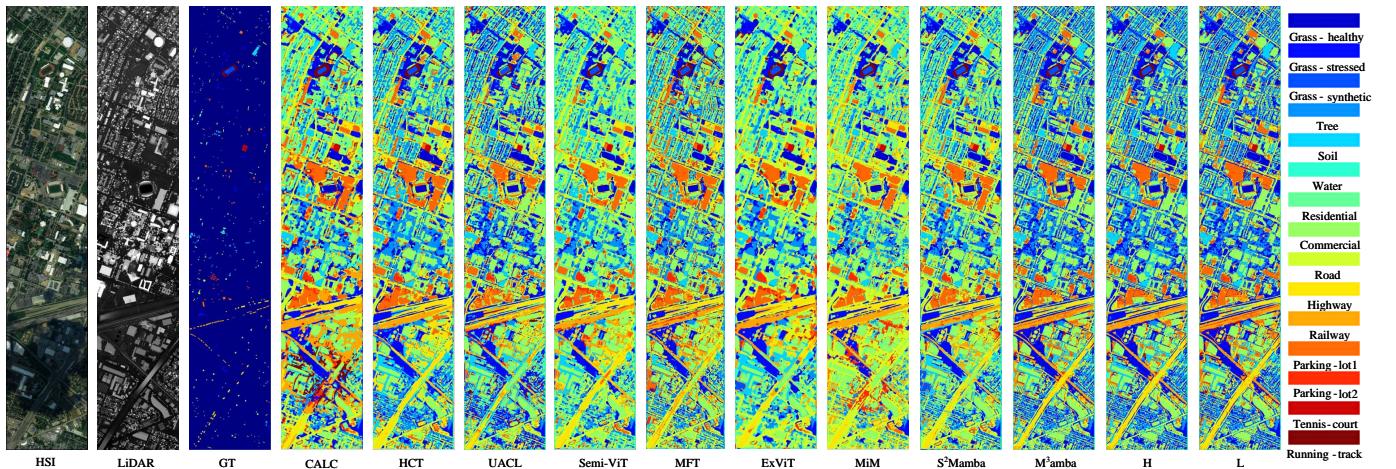|	Class Name	CALC$_{23}$	HCT$_{23}$	UACL$_{24}$	Semi-ViT$_{22}$	MFT$_{23}$	ExViT$_{23}$	MiM$_{24}$	S^{2}Mamba$_{24}$	M^{3}amba	H	L
1	Healthy grass	78.63	97.34	93.68	82.59	82.34	82.91	94.89	83.10	**100.00**	95.19	89.73
2	Stressed grass	83.83	96.62	97.10	82.33	88.78	98.68	98.27	**100.00**	**100.00**	98.13	84.66
3	Synthetic grass	93.86	84.75	99.84	97.43	98.15	99.60	**100.00**	99.60	**100.00**	92.12	91.67
4	Tree	86.55	96.78	96.82	92.93	94.35	99.15	96.87	98.20	**100.00**	79.58	81.32
5	Soil	99.72	**100.00**	99.55	99.84	99.12	99.91	99.86	**100.00**	99.90	97.13	88.29
6	Water	97.90	96.50	96.82	84.15	99.30	99.30	**99.65**	95.80	92.65	80.45	73.59
7	Residential	91.42	82.09	93.29	87.84	88.56	96.08	91.42	89.37	**99.81**	81.94	83.57
8	Commercial	92.88	95.54	87.31	79.93	86.89	90.03	87.17	88.60	**97.93**	88.39	87.48
9	Road	87.54	90.84	93.66	82.94	87.91	86.12	84.31	92.45	**98.87**	81.47	75.84
10	Highway	68.53	58.88	95.26	52.93	64.70	72.97	92.38	92.57	**99.48**	89.39	86.53
11	Railway	93.36	97.53	96.39	80.99	98.64	88.99	90.03	91.56	**99.20**	87.47	83.07
12	Park lot 1	**95.10**	90.11	93.52	91.07	94.24	90.39	89.58	90.97	73.50	70.18	52.95
13	Park lot 2	92.98	97.19	97.13	87.84	90.29	90.18	93.08	89.12	**99.25**	86.63	67.32
14	Tennis court	**100.00**	**100.00**	**100.00**	**100.00**	99.73	99.60	97.44	**100.00**	99.69	91.19	84.25
15	Running track	99.37	**100.00**	99.56	99.65	99.58	95.14	98.33	**100.00**	99.57	96.25	84.61
	OA(%)	88.97	91.15	95.37	85.46	89.80	91.40	92.89	93.36	**97.31**	86.59	79.97
	AA(%)	90.78	92.28	95.99	86.71	91.51	92.60	94.21	94.09	**97.32**	87.70	80.99
	$\kappa(\times 100)$	88.06	90.40	95.00	86.71	88.89	90.66	92.28	92.79	**97.11**	86.33	80.71

Fig. 5. Visualization of false-color HSI and LiDAR images using different comparison methods based on the Houston2013 dataset. H and L respectively indicate that our method is trained using only HSI or LiDAR data.

proposed method, we conduct ablation studies on three remote sensing datasets. Specifically, we design four schemes: (A) directly concatenating multi-modal images to replace the output of the feature adaptation stage; (B) not using our Cross-SS2D module and using the original SS2D module to process both inputs and concatenate the output to the fusion branch; (C) removing unsupervised consistency loss; (D) the complete M^{3}amba. The results of the ablation study are shown in Table I. We observe that M^{3}amba achieves optimal performance across the three datasets. In scheme (A), directly using the original image information to guide the fusion leads to an average OA reduction of 2.14% on the three datasets, because the modality adapter reduces the inter-domain differences between different modalities, making each modality more

accurate in specific tasks. When this module is removed, the fusion process of the modalities is affected, resulting in insufficient information interaction, which affects the overall classification effect. The effectiveness of our cross-attention fusion module Cross-SS2D is demonstrated in (B). Without the Cross-SS2D fusion module, OA, AA, and κ decreased by 1.36%, 3.17%, and 1.73% on average, respectively. This is because Cross-SS2D improves the synchronization and complementarity between modal features through the interaction of selective scanning and global information. After removing this module, the model cannot effectively fuse the information of the two modalities, resulting in a decrease in classification accuracy. The removal of the unsupervised consistency loss in scheme (C) also led to a significant decrease in performance.

TABLE IX
OA, AA AND KAPPA COEFFICIENT ON THE AUGSBURG DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED**. H AND L RESPECTIVELY INDICATE THAT OUR METHOD IS TRAINED USING ONLY HSI OR LiDAR DATA.

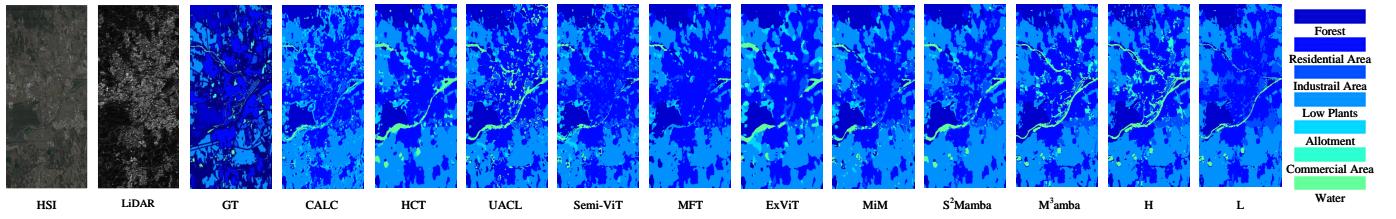|	Class Name	CALC$_{23}$	HCT$_{23}$	UACL$_{24}$	Semi-ViT$_{22}$	MFT$_{23}$	ExViT$_{23}$	MiM$_{24}$	S^{2}Mamba$_{24}$	M^{3}amba	H	L
1	Forest	94.34	94.23	93.66	90.41	94.65	92.89	93.91	93.90	**99.63**	96.13	96.88
2	Commercial Area	98.24	98.54	90.70	92.64	96.90	97.28	88.61	90.55	**99.24**	91.59	64.10
3	Residential Area	78.07	43.79	68.91	60.41	69.80	64.44	60.93	69.09	**85.19**	78.32	82.71
4	Industrial Area	94.57	95.33	92.38	83.40	93.98	86.63	86.55	92.70	**99.60**	97.63	55.17
5	Low Plants	28.68	67.88	56.58	59.41	32.70	57.74	68.29	70.30	**95.39**	79.59	61.35
6	Allotment	2.20	4.82	47.63	18.44	10.52	17.28	24.46	25.84	**81.29**	70.23	29.65
7	Water	42.27	52.09	70.26	27.51	23.98	26.14	50.24	50.49	**92.89**	85.48	63.76
	OA(%)	91.73	90.90	89.19	86.64	90.49	87.91	88.63	89.34	**98.19**	89.36	77.11
	AA(%)	62.62	65.24	74.30	64.67	60.36	63.20	67.57	70.41	**93.32**	85.57	64.80
	$\kappa(\times 100)$	87.98	86.83	84.80	84.05	86.26	82.82	86.56	89.45	**97.39**	88.10	78.79

Fig. 6. Visualization of false-color HSI and LiDAR images using different comparison methods based on the Augsburg dataset. H and L respectively indicate that our method is trained using only HSI or LiDAR data.

This is because the unsupervised consistency loss makes the features between the two modalities more matched by promoting the consistency of cross-modal features. After removing this loss function, the complementary information between the modalities cannot be effectively synchronized, resulting in inconsistency in information between the modalities, which affects the classification accuracy after fusion. In summary, our proposed CLIP feature adaptation and Mamba fusion method can perform joint optimization and effectively improve the performance of multi-modal learning. Additionally, to more intuitively demonstrate the impact of different components on performance, we use t-SNE to visualize the final classification results, as shown in Figure 4. The visualization results of different schemes have the same trend as the results in the table. In general, our M^{3}amba achieves complete feature fusion under the guidance of rich semantics. Comprehensive ablation experiments confirm that M^{3}amba achieves optimal results in both accuracy and visualization.

2) Ablation analysis of feature adaptation: To further explore the effectiveness of semantically guided fusion through feature adaptation, we set up three schemes, as shown in Table VII. We use ViT-B/16 as the backbone of CLIP. Infer refers to directly using the pre-trained CLIP image encoder to infer multi-modal images, and Fine-tune refers to fine-tuning the CLIP image encoder. We introduce modality-specific adapters to train only a very small number of parameters. It can be observed that direct inference can bring a small improvement in training efficiency, but the lack of modality-specific scene understanding leads to a significant decrease in accuracy. This problem can be alleviated by fine-tuning, but the time cost is

high. Our method aims to balance training efficiency and the performance of downstream tasks, achieving the best balance.

C. Comparisons with Previous Methods

To better demonstrate the superiority of our method, we conduct a comprehensive performance comparison with recent advanced multi-modal fusion methods, including CNN-based structures CALC [34], HCT [50], UACL [35]; Transformer-based structures Semi-ViT [51], MFT [33], ExViT [52]; and Mamba-based structures MiM [44], S^{2}Mamba [43]. The subscript of each method in the comparative experiment table represents the publication time of this method, for example, the subscript 24 means it was published in 2024. The average results are summarized in Table II, with our model achieving significant improvements in overall accuracy (OA), average accuracy (AA), and the kappa (κ) coefficient across all datasets. In the Houston2013 dataset, as shown in Table VIII, M^{3}amba achieves the highest overall performance, which is a significant improvement over the closest competitor, UACL. The superior performance of M^{3}amba can be attributed to its CLIP-driven modality-specific adapters, which allow for better cross-modal semantic understanding compared to CNN-based methods. This is especially important in this dataset, where high spectral resolution and spatial detail are critical for accurate classification. For example, M^{3}amba achieves perfect classification accuracy for similar categories like Stressed grass and Synthetic grass, which often pose a challenge in terms of distinguishing subtle spectral differences. The CLIP-driven modality-specific adapters ensure that the model can better differentiate these two categories by

TABLE X
OA, AA AND KAPPA COEFFICIENT ON THE MUUFL DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED**. H AND L RESPECTIVELY INDICATE THAT OUR METHOD IS TRAINED USING ONLY HSI OR LiDAR DATA.

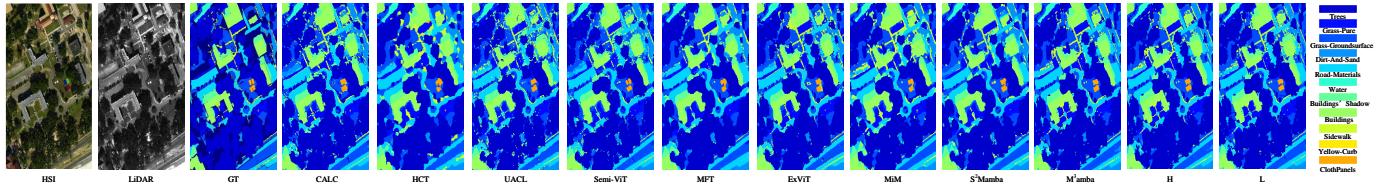|	Class Name	CALC$_{23}$	HCT$_{23}$	UACL$_{24}$	Semi-ViT$_{22}$	MFT$_{23}$	ExViT$_{23}$	MiM$_{24}$	S^{2}Mamba$_{24}$	M^{3}amba	H	L
1	Trees	97.31	97.03	91.29	97.89	97.90	98.58	98.64	98.93	**99.57**	91.13	96.79
2	Grass-Pure	**93.00**	90.29	82.06	79.71	92.11	87.70	92.85	88.05	92.50	87.04	86.07
3	Grass-Groundsurface	91.57	90.07	77.79	84.67	91.80	90.96	92.54	91.31	**96.18**	90.60	89.35
4	Dirt-And-Sand	95.10	94.18	90.13	89.40	91.59	90.61	92.33	90.96	**95.95**	93.79	81.41
5	Road-Materials	95.91	93.86	88.16	93.81	95.60	94.73	96.34	95.08	**98.88**	94.78	85.97
6	Water	99.32	95.71	98.50	81.04	88.19	93.68	88.93	94.03	**99.43**	92.75	80.19
7	Buildings'Shadow	**92.69**	87.09	91.45	83.92	90.27	90.05	91.01	90.40	92.32	80.10	72.91
8	Buildings	98.45	96.61	92.91	98.11	97.26	97.76	98.00	98.11	**99.44**	97.40	99.32
9	Sidewalk	51.60	46.35	75.71	59.12	61.35	68.54	62.09	68.89	**92.92**	85.37	51.32
10	Yellow-Curb	0.00	18.97	**96.07**	14.37	17.43	23.56	18.17	23.91	94.30	84.65	11.49
11	ClothPanels	0.00	75.39	**96.18**	80.86	72.79	80.86	73.53	81.21	95.61	90.19	78.93
	OA(%)	93.94	92.95	88.29	92.46	94.34	94.37	92.65	94.19	**97.84**	88.91	84.16
	AA(%)	74.09	80.50	89.11	79.63	81.48	83.37	82.22	83.72	**96.10**	89.80	75.80
	$\kappa(\times 100)$	92.00	90.69	84.78	89.49	92.51	92.54	91.59	91.98	**97.13**	85.75	83.78

Fig. 7. Visualization of false-color HSI and LiDAR images using different comparison methods based on the MUUFL Gulfport scene dataset. H and L respectively indicate that our method is trained using only HSI or LiDAR data.

refining the feature representations of both HSI and LiDAR modalities. This highlights how M^{3}amba's ability to capture complementary information from different modalities—thanks to its Cross-SS2D module—improves classification accuracy for fine-grained classes. Additionally, M^{3}amba demonstrates robust performance across nearly all land cover classes, but its significant improvement is particularly noticeable in categories like Highway and Road, where the fusion of spectral and spatial information provided by the model enhances feature extraction and reduces misclassification.

The Augsburg dataset presents a more challenging scenario, but M^{3}amba again surpasses all competing methods with an OA of 98.19%, an AA of 93.32%, and a kappa coefficient of 97.39%, as shown in Table IX. One key reason for this outstanding performance is the ability of the Cross-SS2D module to maintain feature integrity across different data modalities. For example, in the Forest category, M^{3}amba outperforms competitors by leveraging both the spectral details from HSI and the structural information from LiDAR, which is crucial for distinguishing different types of vegetation. Similarly, in Industrial areas, where spatial complexity and high variability in land cover types are prevalent, M^{3}amba shows consistent improvements by modeling the interactions between spectral and spatial features more effectively than the CALC model. The MUUFL dataset contains a mix of urban and natural land cover categories, as shown in Table X, M^{3}amba outperforms

all other methods, achieving an OA of 97.84%, an AA of 96.10%, and a kappa coefficient of 97.13%. Compared to the MiM model, which attains an OA of 92.65%, M^{3}amba offers a 5.19% improvement. In particular, M^{3}amba performs well in the sidewalk class, achieving an accuracy of 92.92%, far exceeding other comparison methods. These results highlight the effectiveness of the Cross-SS2D module in maintaining the integrity of fine-grained features across different data modalities and its ability to model long sequences, which is a key advantage over other Mamba-based methods such as MiM and S^{2}Mamba. In the Buildings category, the model's performance is also significantly enhanced. The fusion of HSI and LiDAR data provides a richer representation of building structure and surface features, which is often difficult to achieve with single-modal approaches. The model's ability to effectively leverage both spatial and spectral features allows it to better capture complex building shapes and variations in urban environments.

To investigate the specific contribution of the complementarity of different modal data to the fusion performance, we train M^{3}amba using only HSI or LiDAR data and compare it with the results of multi-modal data training. The results in Tables VIII, IX, X show that when combining the two data modalities, the spectral advantages of HSI and the spatial structure advantages of LiDAR can be simultaneously exploited to improve classification accuracy in complex scenes.

For example, although the spectral features of objects such as buildings may be similar in HSI, the elevation information of LiDAR helps to accurately identify these objects. Objects with obvious spectral differences, such as water and soil, have significantly improved accuracy in the fused classification results. We can also find that the complementarity of the two modalities allows some categories to reach 100% accuracy after fusion, which does not happen when using only single modal data.

Furthermore, as shown in Figure 3, the box plot of our method is taller and more compact than the competing methods, indicating that our method achieves robust detection performance for each class. This effectively mitigates the issue of low detection accuracy for individual classes observed in the competing methods.

In addition to the accuracy metrics, our method also demonstrates a notable reduction in training time compared to other models, as shown in Table VI. Our method achieves the lowest average training time across the three datasets, outperforming CNN-based, Transformer-based, and Mamba-based architectures and far exceeds all other methods in terms of average accuracy. This is attributed to the design of our efficient fusion architecture, in stark contrast to the quadratic complexity in traditional Transformer models. The combination of efficient feature fusion and comprehensive semantic guidance makes M^3amba both an accurate and computationally efficient model for multi-modal remote sensing tasks.

D. Analysis of Failure Cases

Although M^3amba achieves the highest accuracy on multiple categories in multiple datasets, there are still cases of poor classification, especially in complex scenes, where some categories perform poorly, partly due to the ambiguity of features or spectral overlap. Through the analysis of failed cases, we find the following two points. Texture fuzzy categories: such as Water (in Table VIII), because its texture features are similar to the surrounding environment in hyperspectral images, lead to classification errors. Especially in LiDAR data, these objects lack significant elevation differences, which affects the judgment of the model. Complex scene problems: In densely built-up areas in cities, objects such as parking lots (in Table VIII) lack clear spatial structure and have similar spectral features to the surrounding environment (such as roads and buildings), resulting in reduced classification accuracy.

E. Result Visualization

We visualize the results by assigning a unique color to each class. Figure 5, Figure 6, and Figure 7 demonstrate the excellent performance of our method on the full remote sensing image dataset, and our visualization is still better than most of the comparison methods even when trained with only single-modal data. M^3amba extracts fundamental properties of multi-modal images by establishing comprehensive fused features and invariant representations. It utilizes CLIP to generate comprehensive semantic features to guide the efficient fusion of multi-modal features. This enhances the model's generalization ability and preserves complete information between modalities, resulting in richer and more detailed representations in the classification maps. Overall, M^3amba is highly suitable for generating more robust and intricate detailed multi-modal classification maps.

V. CONCLUSION

In this paper, we introduce M^3amba, a novel end-to-end CLIP-driven Mamba unified framework for multi-modal fusion, combining CLIP's powerful multi-modal visual semantic representation capabilities with the efficiency of SSM. By introducing CLIP-driven modality-specific adapters and the Mamba fusion architecture embedded with the cross-attention module Cross-SS2D, M^3amba addresses the primary challenges of existing methods, including high computational complexity, limited generalization, and incomplete feature fusion, achieving optimal performance in both accuracy and efficiency. Extensive evaluations on multiple remote sensing multi-modal datasets demonstrate the generalizability and interpretability of M^3amba. This work showcases the tremendous potential for future research in efficient and effective multi-modal learning, highlighting the potential of combining pre-trained multi-modal models with state space modeling techniques to advance the field of multi-modal fusion. In addition, due to the efficient training and excellent performance of M^3amba, in future work, we will focus on verifying the effectiveness of M^3amba in specific application scenarios, especially in multiple practical scenarios such as agricultural monitoring, urban planning, and ecological protection, as well as exploring new directions such as dynamic target detection and real-time multi-modal data processing, to further demonstrate its wide applicability and strong potential in practical remote sensing tasks.

REFERENCES

[1] Y. Zhao, Q. Zheng, P. Zhu, X. Zhang, and W. Ma, "Tufusion: A transformer-based universal fusion algorithm for multimodal images," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 34, pp. 1712–1725, 2023.

[2] 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 Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2024.

[3] M. Ma, W. Ma, L. Jiao, X. Liu, F. Liu, L. Li, and S. Yang, "Mbsi-net: Multimodal balanced self-learning interaction network for image classification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 34, pp. 3819–3833, 2023.

[4] D. Li, W. Xie, J. Zhang, and Y. Li, "Mdfl: Multi-domain diffusion-driven feature learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 8, 2024, pp. 8653–8660.

[5] X. Xie, Y. Cui, C.-I. Ieong, T. Tan, X. Zhang, X. Zheng, and Z. Yu, "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba," *arXiv preprint arXiv:2404.09498*, 2024.

[6] X. Zhang, A. Liu, G. Yang, Y. Liu, and X. Chen, "Simfusion: A semantic information-guided modality-specific fusion network for mr images," *Information Fusion (IF)*, p. 102560, 2024.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.

[8] Z. Lin, S. Geng, R. Zhang, P. Gao, G. De Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 388–404.

[9] Z. Yang, H. Jiang, W. Hong, J. Teng, W. Zheng, Y. Dong, M. Ding, and J. Tang, "Inf-dit: Upsampling any-resolution image with memory-efficient diffusion transformer," *arXiv preprint arXiv:2405.04312*, 2024.

[10] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision (IJCV)*, vol. 132, no. 2, pp. 581–595, 2024.

[11] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 5, pp. 2402–2415, 2020.

[12] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica (IJAS)*, vol. 9, no. 7, pp. 1200–1217, 2022.

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

[14] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 104–12 113.

[15] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8583–8595, 2021.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009.

[18] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9653–9663.

[19] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 4904–4916.

[20] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[22] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa, "Clip-mesh: Generating textured meshes from text using pretrained image-text models," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–8.

[23] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "Clip-forge: Towards zero-shot text-to-shape generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 603–18 613.

[24] J. Xu, X. Wang, W. Cheng, Y.-P. Cao, Y. Shan, X. Qie, and S. Gao, "Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20 908–20 918.

[25] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[26] D. Luo, J. Huang, S. Gong, H. Jin, and Y. Liu, "Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 045–23 055.

[27] D. Luo, J. Huang, S. Gong, H. Jin, and Y. Liu, "Zero-shot video moment retrieval from frozen vision-language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 5464–5473.

[28] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 123–18 133.

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 19 730–19 742.

[30] S. Dong, L. Wang, B. Du, and X. Meng, "Changeclip: Remote sensing change detection with multimodal vision-language representation learning," *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, vol. 208, pp. 53–69, 2024.

[31] S. Mei, G. Zhang, N. Wang, B. Wu, M. Ma, Y. Zhang, and Y. Feng, "Lightweight multiresolution feature fusion network for spectral super-resolution," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 61, pp. 1–14, 2023.

[32] S. Mei, R. Jiang, M. Ma, and C. Song, "Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 61, pp. 1–13, 2023.

[33] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 61, pp. 1–20, 2023.

[34] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and lidar data," *Information Fusion (IF)*, vol. 93, pp. 118–131, 2023.

[35] K. Ding, T. Lu, and S. Li, "Uncertainty-aware contrastive learning for semi-supervised classification of multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2024.

[36] 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 Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2024.

[37] J. Wang and X. Tan, "Mutually beneficial transformer for multimodal data fusion," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 33, no. 12, pp. 7466–7479, 2023.

[38] 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 Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 33, no. 11, pp. 6804–6818, 2023.

[39] X. Wang, L. Song, Y. Feng, and J. Zhu, "S3f2net: Spatial-spectral-structural feature fusion network for hyperspectral image and lidar data classification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025.

[40] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 60, pp. 1–12, 2021.

[41] S. Mei, Y. Geng, J. Hou, and Q. Du, "Learning hyperspectral images from rgb images via a coarse-to-fine cnn," *Science China Information Sciences (SCIS)*, vol. 65, pp. 1–14, 2022.

[42] J. Yao, D. Hong, C. Li, and J. Chanussot, "Spectralmamba: Efficient mamba for hyperspectral image classification," *arXiv preprint arXiv:2404.08489*, 2024.

[43] G. Wang, X. Zhang, Z. Peng, T. Zhang, X. Jia, and L. Jiao, "S^{2}mamba: A spatial-spectral state space model for hyperspectral image classification," *arXiv preprint arXiv:2404.18213*, 2024.

[44] W. Zhou, S.-I. Kamata, H. Wang, M.-S. Wong *et al.*, "Mamba-in-mamba: Centralized mamba-cross-scan in tokenized mamba model for hyperspectral image classification," *arXiv preprint arXiv:2405.12003*, 2024.

[45] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[46] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 357–366.

[47] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 7, no. 6, pp. 2405–2418, 2014.

[48] 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 Journal of Photogrammetry and Remote Sensing (ISPRS)*, vol. 178, pp. 68–80, 2021.

[49] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "Muufl gulfport hyperspectral and lidar airborne data set," *Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570*, 2013.

[50] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer,"

IEEE Transactions on Geoscience and Remote Sensing (TGRS), vol. 61, pp. 1–16, 2022.

[51] Z. Cai, A. Ravichandran, P. Favaro, M. Wang, D. Modolo, R. Bhotika, Z. Tu, and S. Soatto, "Semi-supervised vision transformers at scale," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 25 697–25 710, 2022.

[52] 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 Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 61, pp. 1–15, 2023.

Kai Jiang received the B.E. degree in information engineering and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2019 and 2024, respectively. His research interests include image processing and deep learning.

Mingxiang Cao received the B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China in 2023. He is currently pursuing the M.S. degree with the Image Coding and Processing Center at State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China. His research interests include multimodal image processing, remote sensing classification, and object detection.

Jie Lei received his M.S. degree in Telecommunications and Information Systems and his Ph.D. degree in Signal and Information Processing from Xidian University, China, in 2006 and 2010, respectively. He was a Visiting Scholar at the Department of Computer Science at the University of California, Los Angeles, USA, from 2014 to 2015. He served as a Professor at the School of Telecommunications Engineering, Xidian University, until 2023. Currently, he is a Research Fellow at the School of Electrical and Data Engineering at the University of Technology Sydney. His research interests include wireless communication, remote sensing image processing, machine learning, and customized computing for big data applications.

Weiying Xie (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2017. Currently, she is a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. She has published over 50 articles in refereed journals and proceedings, including IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, IEEE Transaction on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, and Conference on IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) and Association for the Advancement of Artificial Intelligence (AAAI). Her research interests include neural networks, machine learning, hyperspectral image processing, and high-performance computing.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, China, in 1999 and 2002, respectively. He joined the School of Telecommunications Engineering, Xidian University in 1999 where he is currently a Professor. Prof. Li is the director of the image coding and processing center at the State Key Laboratory of Integrated Service Networks. His research interests focus on image and video processing and high-performance computing.

Xin Zhang received the B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China in 2019. She is currently pursuing the Ph.D. degree with the Image Coding and Processing Center at State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China. Her research interests span efficient deep learning, machine learning, and computer vision. Specifically, she is interested in model compression for computer vision models (CNN, ViT), knowledge distillation to both models and datasets, and general CV tasks (foundational model training and downstream applications).

Jiaqing Zhang received the B.E. degree in Telecommunications Engineering from Ningbo University, Zhejiang, China in 2019. She is currently pursuing the Ph.D. degree with the Image Coding and Processing Center at State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China. Her research interests include multimodal image processing, remote sensing object detection, and network compression.