CoMiX: Cross-Modal Fusion with Deformable Convolutions for HSI-X Semantic Segmentation

Xuming Zhang, *Student Member, IEEE,* Naoto Yokoya, *Member, IEEE,*
Xingfa Gu, Qingjiu Tian, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Improving hyperspectral image (HSI) semantic segmentation by exploiting complementary information from supplementary modalities (termed X-modality) is promising but challenging due to significant differences in imaging sensors, image content, and resolution. Existing methods often underutilize the unique spatial–spectral features of HSIs by processing them uniformly with X-modality data. In addition, current cross-modality fusion strategies often suffer from limited intermodal interaction or significantly increased model complexity. To address these limitations, we propose CoMiX, an asymmetric encoder-decoder architecture with deformable convolutions (DCNs) for HSI-X semantic segmentation. CoMiX includes an encoder with two parallel, interacting backbones and a lightweight all-multilayer perceptron (ALL-MLP) decoder. The encoder consists of four stages, each incorporating 2D DCN blocks for the X-modality to accommodate geometric variations and 3D DCN blocks for HSIs to adaptively capture spatial-spectral features. Each stage also incorporates a Cross-Modality Feature enhancement and eXchange (CMFeX) module and a feature fusion module (FFM). CMFeX exploits spatial-spectral correlations across modalities to recalibrate and enhance modality-specific and modality-shared features, while adaptively exchanging complementary information. Its outputs are subsequently fused in the FFM and propagated to the next stage for further learning. Finally, the ALL-MLP decoder aggregates the fused features from all stages to produce the final predictions. Extensive experiments demonstrate that CoMiX achieves state-of-the-art performance and generalizes well to various multimodal datasets. The CoMiX code will be released soon.**

***Index Terms*—Deep learning, hyperspectral image, classification, multimodel semantic segmentation, cross-modality fusion, deformable convolution**

This work was funded in part by the Guangzhou Municipal Science and Technology Bureau under Grant 2025A03J3171, National Natural Science Foundation of China under Grants L2424330 and 42101321, the National Key Research and Development Program of China under Grant 2023YFF1303903, and the Open Fund of State Key Laboratory of Urban and Regional Ecology under Grant SKLURE2023-2-6. (Corresponding authors: Xingfa Gu and Qingjiu Tian).

Xuming Zhang is with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266404, China (This is the first unit. e-mail: xumingzhang2018@163.com).

Qingjiu Tian is with the International Institute for Earth System Science, Nanjing University, Nanjing 210023, China (e-mail: tianqj@nju.edu.cn).

Naoto Yokoya is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8561, Japan, and with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

Xingfa Gu is with the Institute of Aerospace Remote Sensing Innovations, Guangzhou University, Guangzhou 510006, China, School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: guxf@aircas.ac.cn).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

I. Introduction

Hyperspectral images (HSIs), with continuous spectral information over a wide range of wavelengths, enable detailed analysis and discrimination of different materials [1], [2]. HSI semantic segmentation is a fundamental task in scene understanding and supports many applications, including land-cover analysis [3] and urban development monitoring [4]. It is closely related to HSI classification, as both are per-pixel classification [5]. However, single-modality HSI classification has encountered performance bottlenecks, especially under challenging conditions like low illumination and atmospheric interference. Multimodal fusion mitigates these limitations by integrating complementary information from additional sensors (X-modality), such as digital surface models (DSM), light detection and ranging (LiDAR), and synthetic aperture radar (SAR), thereby enhancing performance.

Numerous multimodal fusion approaches [6], [7] have been developed and are typically classified into three main categories: pixel-level fusion [8], feature-level fusion [9], and decision-level fusion [10]. Among these, feature-level fusion is generally recognized as effective and scalable. For example, Rasti *et al.* [9] adopted extinction profiles [11] to extract spatial and elevation information from HSI and LiDAR data, respectively, which was then fused to produce classification maps. Subsequent studies [12], [13] further refined strategies to improve the integration of multimodal information. However, these traditional techniques suffer from inherent drawbacks, including reliance on expert feature engineering and limitations in feature representation and generalization.

Deep learning (DL), an automatic feature learning technique, has demonstrated remarkable success in vision tasks [14], [15]. In particular, convolutional neural networks (CNNs) are widely used due to their scalability, flexibility, and ease of optimization [16]. Numerous CNN-based approaches [17], [18], [19] have been developed to enhance feature extraction or information fusion in multi-source data. Nonetheless, CNNs process all inputs indiscriminately, which is suboptimal because different sources contribute unequally to accurate classification [20]. To mitigate this limitation, attention mechanisms [21], [22] have been introduced to guide models toward the most informative features. Building on this idea, the vision transformer (ViT) [23] employs self-attention for adaptive spatial aggregation and long-range dependency modeling, marking a significant advance. This foundation has enabled transformers to greatly improve multimodal data processing by enhancing feature extraction and integration [24], [25], [26].

Although transformers have achieved impressive results in multimodal segmentation, their low sensitivity to local features and the lack of the inductive biases inherent in CNNs limit their performance. Therefore, hybrid networks [27], [28] were developed to combine the strengths of local convolutions and global transformers. Alternatively, some studies [29], [30] have attempted to introduce long-range dependencies into CNNs by using large dense kernels (e.g., 31×31).

Recently, Mamba [31], which linearly encodes sequential features, was introduced as an efficient alternative to transformers. Mamba-based cross-modality fusion networks (e.g., Sigma [32], MambaDFuse [33], SegMamba [34]) have demonstrated strong effectiveness in modeling long-range inter-modal dependencies while reducing computational overhead. Due to Mamba's one-dimensional selective scanning mechanism, these models exhibit limited scalability when applied to high-dimensional data like HSI, where capturing both spatial and spectral dependencies is essential.

Deformable convolutional networks (DCNs) [35], [36], [37] offer a lightweight yet effective alternative for capturing spatial-spectral information in high-dimensional data. Unlike standard convolutions with fixed receptive fields, DCNs introduce learnable offsets that adapt sampling locations based on inputs, enabling more flexible and context-aware feature aggregation. This adaptability is especially advantageous in HSI-X segmentation, where geometric distortions and scale discrepancies between modalities (e.g., HSI vs. LiDAR/SAR) frequently compromise alignment and feature consistency. Moreover, DCNs maintain high computational efficiency by employing small kernel sizes (e.g., 3×3), striking a favorable balance between performance and efficiency. These characteristics make DCNs well-suited for HSI-X feature extraction.

Despite these advances, existing approaches for HSI-X segmentation still face two key challenges:

1) Many methods [38], [39] are tailored to specific modality combinations (e.g., HSI and LiDAR), restricting their applicability to other data types. Although some architectures [24], [40] were proposed for HSI-X segmentation, they process HSIs using the same strategies as X-modality data, overlooking the unique 3D spatial-spectral structure of HSIs. This leads to underutilization of the rich spectral information in HSI. Therefore, there is an urgent need for HSI-X segmentation architectures that adapt to diverse X modalities while effectively exploiting the rich discriminative information of HSIs.

2) Early fusion strategies, including feature concatenation [18], [41], feature addition [28], [38], [42], and cross-attention fusion [27], [43], [44], often suffer from limited inter-modal interaction, resulting in suboptimal performance. Although recent multi-stage fusion approaches [40], [14] enhance cross-modality integration, they generally increase model complexity and reduce robustness, as multi-stage attention mechanisms are computationally intensive and harder to optimize.

To overcome these challenges, we propose CoMiX, a universal semantic segmentation framework designed for HSI-X segmentation, where X includes DSM, LiDAR, SAR, or other complementary modalities. CoMiX integrates deformable convolutions with transformer-inspired block-level designs, facilitating the development of 3D DCN and 2D DCN blocks for adaptive modality-specific feature extraction from HSI and X-modality data, respectively. As illustrated in Fig. 1, CoMiX consists of an encoder with two parallel and interactive backbones, along with a lightweight all-multilayer perceptron (ALL-MLP) decoder. Within the encoder, a Cross-Modality Feature enhancement and eXchange (CMFeX) module is introduced at each stage to recalibrate modality-specific features and explicitly facilitate cross-modality interactions across spatial and spectral dimensions. These recalibrated features are then fused via a Feature Fusion Module (FFM) and passed to the next stage for progressive learning. Finally, the All-MLP decoder aggregates the multi-stage fused features to produce the final segmentation map.

The main contributions of this study are as follows.

1) Proposing CoMiX, a universal framework for HSI-X semantic segmentation that extracts modality-specific, modality-shared, and complementary features from HSI and X-modality data, while enhancing cross-modal interaction and fusion. Specifically, modality-specific features denote information unique to a single modality (e.g., structural details from LiDAR); modality-shared features represent common patterns (e.g., semantic or spatial structures); and complementary features are distinct yet mutually reinforcing cues that, when fused, yield more complete and robust semantic representations.

2) Developing parallel and interactive backbones with deformable convolutions, where 3D DCN blocks capture spectral-spatial information from HSIs and 2D DCN blocks learn features from X-modality data, thereby enabling comprehensive cross-modal representation learning.

3) Designing the CMFeX module to recalibrate and enhance modality-specific and modality-shared features across spatial and spectral dimensions, while facilitating more effective cross-modality information exchange.

The remainder of this paper is organized as follows. Section II reviews related work, followed by a description of the proposed CoMiX framework in Section III. Section IV presents the experimental setup and results, while Section V provides further discussion. Finally, Section VI concludes this paper with remarks and directions for future research.

II. Related Work

A. CNN-based cross-modality segmentation

CNNs have been widely used for cross-modality semantic segmentation due to their efficient feature extraction and ease of optimization [16]. Chen *et al.* [17] proposed a two-branch CNN to extract features from multispectral images (MSIs)/HSIs and LiDAR data, which are subsequently fused through fully connected (FC) layers. EndNet [18], an FC-dominated encoder-decoder network, integrates multimodal features by forcing the fused features to reconstruct the inputs. The unified multimodal fusion framework [19] comprises extraction and fusion subnetworks, with the latter employing a novel cross-fusion strategy that outperforms concatenation-based fusion. Unlike these patch-based learning models [17], [18], [19], Fusion-FCN [41], the winner of the 2018 IEEE Data Fusion Contest (DFC), was built with a fully convolutional

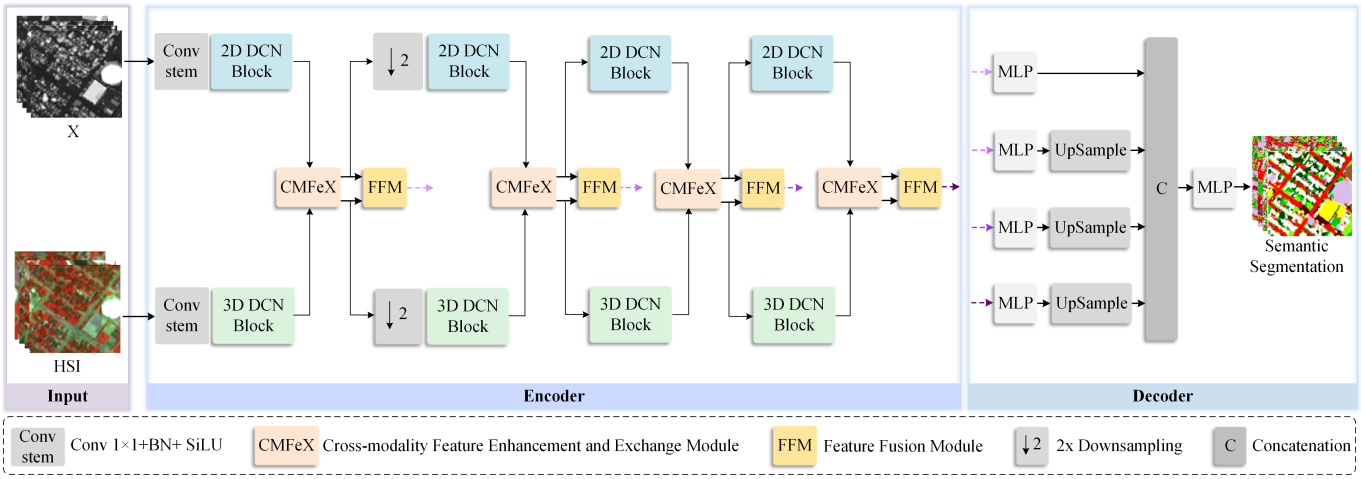


Fig. 1. Overview of the proposed CoMiX framework for HSI-X semantic segmentation.

network (FCN), demonstrating the significant potential of FCNs in cross-modality segmentation.

To overcome the limitations of standard convolutional kernels that treat all features equally in CNNs, attention mechanisms have been incorporated to enhance feature representation. Wang *et al.* [45] and Zhang *et al.* [42] applied attention modules to improve feature learning from optical and LiDAR data, respectively. The coupled adversarial learning-based classification (CALC) framework [38] introduced a spatial attention module to capture high-level semantic features from HSI and LiDAR data.

Moreover, various studies [27], [46], [47] explored cross-modality attention mechanisms to enhance cross-modality information fusion. Xue *et al.* [48] employed self-attention to enhance both feature learning and fusion between HSI and LiDAR. Roy *et al.* [49] developed a cross-attention mechanism where LiDAR patch tokens serve as queries, while HSI patch tokens act as keys and values, facilitating more effective integration of both modalities.

B. Transformer-Based cross-modality segmentation

Transformers [23] have revolutionized cross-modality segmentation by effectively capturing long-range dependencies and facilitating interactions between heterogeneous modalities. [23]. For instance, multimodal fusion transformers (MFT) [39] and deep hierarchical vision transformers [50] extract complementary features from HSI and LiDAR data, which are then fused via cross-attention modules. Similarly, the local information interaction transformer network [43] employs dual-branch transformers to model sequential dependencies in each modality, followed by fusion through a convolutional transformer for classification. However, these approaches typically adopt relatively simple cross-modal interaction strategies, limiting their generalization to diverse modality combinations.

In computer vision, various transformer-based cross-modality segmentation frameworks have been developed, particularly for RGB-X tasks [51], [52], [53], [54]. DFormer [55] introduces a pretraining framework tailored for RGB-Depth segmentation, improving the transferability of learned representations. GeminiFusion [51] proposes a pixel-wise fusion mechanism based on spatially aligned features across modalities. DPLNet [52] leverages a dual-prompt learning paradigm that adapts a frozen RGB backbone to multimodal inputs, significantly reducing training overhead. Other frameworks, such as CMNeXt [53] and CMX [54], adopt dual-branch architectures for RGB-X segmentation.

Flex-MCFNet [40] extends CMX by integrating a flexible mixup data augmentation strategy to improve performance in HSI-X segmentation. Given the high computational burden of transformers on large images, the local-to-global cross-modal attention-aware fusion (LoGoCAF) framework [24] integrates a local-to-global encoder—using CNNs in shallow layers and transformers in deeper layers—combined with cross-modality modules for effective HSI-X fusion.

Despite these advances, most HSI-X segmentation methods are adapted from RGB-X frameworks designed for natural images. In contrast, HSI-X segmentation in remote sensing poses unique challenges—including high spectral dimensionality, spatial heterogeneity, and modality disparities—that require specialized architectures. Moreover, models such as CMX, Flex-MCFNet, and LoGoCAF employ the same feature extraction structure across modalities, limiting their ability to exploit modality-specific features.

To overcome this limitation, CoMiX introduces 3D DCN blocks for HSI and 2D DCN blocks for the X-modality, enabling modality-aware feature extraction. In addition, unlike the cross-modal feature rectification module in CMX, CoMiX incorporates CMFeX for feature calibration and interaction, improving both performance and efficiency. Finally, CMX relies on a two-stage FFM, while CoMiX retains only the lightweight second stage, significantly reducing computational cost while maintaining effective cross-modality fusion. Collectively, these distinctions indicate that CoMiX is not merely an extension of CMX but a specialized framework tailored to the unique characteristics of HSI-X segmentation.

III. Methodology

A. Methodological Overview

Fig. 1 presents an overview of the proposed CoMiX framework. The encoder employs two parallel and interacting back-

bones to adaptively aggregate local- and long-range dependencies, as well as fine-grained and coarse-level features from HSI and X-modality. The All-MLP [5] decoder is used to integrate multi-level fused features and produce the final semantic segmentation maps. The encoder begins with a convolutional stem composed of a 1×1 convolution, batch normalization, and a SiLU activation function, projecting HSI and X-modality inputs into a unified feature space with equal channel dimensions. Following the stem, four sequential stages are deployed. In each stage, the proposed 3D DCN and 2D DCN blocks are used for modality-specific feature extraction from HSI and X-modality data, respectively. To balance segmentation accuracy and computational efficiency, spatial resolution is downsampled only once—by a factor of two—before Stage 2. At each stage, we further introduce a CMFeX module to capture spatial and spectral correlations across modalities. It recalibrates and enhances modality-specific and modality-shared features while facilitating the exchange of complementary information. The recalibrated features from CMFeX are passed to FFM for cross-modality fusion while being forwarded to the next stage for further learning. The decoder then aggregates the outputs from all FFMs to generate pixel-wise segmentation predictions.

B. 2D DCN Block

Geometric variations induced by scale, viewpoint, etc., pose significant challenges in land-cover recognition. Currently, DCNs are the state-of-the-art method to address this problem [36]. The DCN series extends traditional convolutional operations by introducing learnable offsets that dynamically adjust sampling locations. This allows the network to perform adaptive spatial aggregation and effectively adapt to geometric variations. DCNv1 [56] is the pioneering work that developed the deformable convolution. Subsequently, DCNv2 [36], DCNv3 [35], and DCNv4 [37] introduced tailored modifications to further enhance information representation and bridge the gap between CNNs and ViTs. Inspired by [35], we refine it to develop a faster and more effective feature extraction operator for the X-modality.

Given a convolution kernel with K sampling locations, let p_k denote the pre-specified offset for the k-th location. For example, a regular 3×3 convolution kernel is defined as $K = 9$ and $p_k \in$ (-1, -1), (-1, 0), . . ., (1, 1). Let $x\,(p)$ and $y\,(p)$ denote the features at location p of the input $\mathbf{x}$ and output $\mathbf{y}$, respectively. Following the multi-group designed in group convolution, our 2D DCN splits the spatial aggregation process into multi-groups to learn richer information. Then, 2D DCN at location p can be expressed as:

$$y\,(p) = \sum_{g=1}^{G} \sum_{k=1}^{K} w_g m_{gk} x_g\,(p + p_k + \Delta p_{gk})\,, \tag{1}$$

where G represents the number of spatial aggregation groups. For the g-th group, $w_g \in \mathbb{R}^{C \times C^{'}}$ is the location-irrelevant projection weight, $m_{gk} \in \mathbb{R}$ is the modulation scalar for the k-th sampling point, and $x_g \in \mathbb{R}^{C^{'} \times H \times W}$ denotes the input feature maps, where $C^{'} = C/G$ represents the group dimension. Δp_{gk} signifies the sampling offset corresponding to the pre-specified offset p_k within the g-th group. In practice, the modulation scalar m_{gk} and the sampling offset Δp_{gk} are generated from the input $\mathbf{x}$ by a 3×3 depth-wise convolution (DWConv) followed by a linear projection.

Equation (1) illustrates that the sampling offset Δp_{gk} is flexible and can capture both short- and long-range dependencies. Additionally, the sampling offset Δp_{gk} and the modulation scalar m_{gk} are input-dependent and learnable, allowing the network to effectively model geometric transformations. Meanwhile, the 2D DCN retains the inductive bias of convolution, leading to improved efficiency with reduced training time and data requirements. It is also easier to optimize and more computationally and memory-efficient than other techniques, such as self-attention [23], large kernel convolution [29], and Mamba [31], [57].

As shown in Fig. 2(a), by combining the aforementioned the 2D DCN with advanced transformer block designs, the 2D DCN block can be expressed as:

$$\mathbf{X}_1 = \mathrm{LN}(\text{2D DCN}(\mathbf{X}_\mathrm{x})) + \mathbf{X}_\mathrm{x}, \tag{2}$$

$$\mathbf{X}_{\text{out1}} = \mathrm{LN}(\mathrm{FFN}(\mathbf{X}_1)) + \mathbf{X}_1, \tag{3}$$

$$\mathrm{FFN}(\mathbf{X_1}) = \mathrm{Conv}_{1\times1}(\mathrm{GELU}(\mathrm{Conv}_{1\times1}(\mathbf{X_1}))), \tag{4}$$

where $\mathbf{X}_\mathrm{x}$, $\mathbf{X}_1$ and $\mathbf{X}_{\text{out1}}$ represent the input X-modality data, intermediate features, and output features of the 2D DCN block, respectively. $\mathrm{Conv}_{1\times1}$ refers to a 1×1 convolution, GELU is a activation function, and LN denotes layer normalization [58]. We stack 2D DCN blocks in each stage for X-modality feature learning.

Unlike previous DCN variants that use 3×3 sparse windows throughout the network, we adopt 3×3 windows in the first two stages and 7×7 windows in the last two. This design enables the network to selectively focus on salient regions and capture finer details, thereby enhancing feature representation. This capability is particularly crucial for remote sensing image semantic segmentation, where scenes exhibit complex structures and small objects require context-aware feature extraction and precise localization.

C. 3D DCN Block

Given that HSI data are 3D cubes, it is intuitive to extend the 2D DCN block to 3D for feature extraction. However, this extension significantly increases the number of parameters and computational cost as the number of channels scales along the third dimension [59]. To address this, we adopt the efficient paired-attention (EPA) block [60] for HSI processing. The EPA block employs a shared keys-queries scheme between the spatial and spectral attention modules, allowing mutual information exchange while improving efficiency. Specifically, the $\mathbf{K}$ and $\mathbf{Q}$ matrices are shared to enable consistent attention across spatial and spectral dimensions, thereby capturing modality-shared dependencies. In contrast, the $\mathbf{V}$ matrices remain separate, allowing each attention to retain modality-specific information. This design effectively balances cross-dimensional feature alignment with representational diversity.

As illustrated in Fig. 2(b), EPA consists of a spatial attention module and a spectral attention module, both of which use

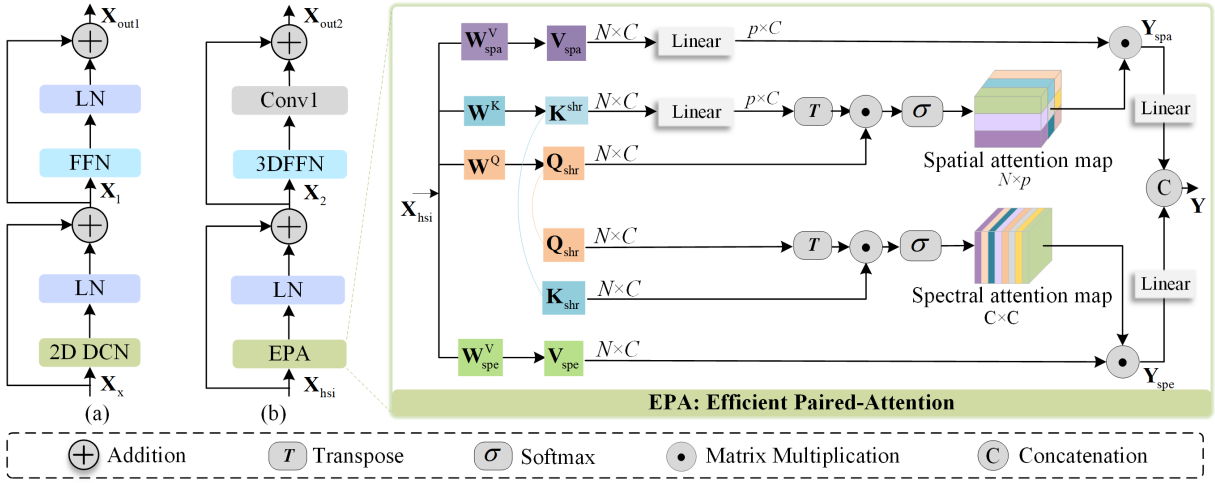


Fig. 2. Illustration of the proposed deformable convolution (DCN) blocks: (a) The 2D DCN block processes X-modality data using 2D DCN, followed by layer normalization (LN) and a feed-forward network (FFN); (b) The 3D DCN block captures spectral–spatial features via the efficient paired-attention (EPA) module, which shares keys and queries between the two attention modules to enrich representations while reducing complexity. Outputs are refined with a 3D FFN and a $1 \times 1 \times 1$ convolution (Conv_1)

.

shared keys and shared queries to minimize computational overhead. Given an input $\mathbf{X}_{\text{hsi}} \in \mathbb{R}^{HW \times C}$, where H, W, and C denote the height, width, and number of spectral channels of $\mathbf{X}_{\text{hsi}}$, respectively. We linearly transform it into spatial values $\mathbf{V}_{\text{spa}} \in \mathbb{R}^{N \times C}$, shared queries $\mathbf{Q}_{\text{shr}} \in \mathbb{R}^{N \times C}$, shared keys $\mathbf{K}_{\text{shr}} \in \mathbb{R}^{N \times C}$ and spectral values $\mathbf{V}_{\text{spe}} \in \mathbb{R}^{N \times C}$ via $\mathbf{V}_{\text{spa}} = \mathbf{W}^{\text{V}}_{\text{spa}}\mathbf{X}_{\text{hsi}}$, $\mathbf{Q}_{\text{shr}} = \mathbf{W}^{\text{Q}}\mathbf{X}_{\text{hsi}}$, $\mathbf{K}_{\text{shr}} = \mathbf{W}^{\text{K}}\mathbf{X}_{\text{hsi}}$ and $\mathbf{V}_{\text{spe}} = \mathbf{W}^{\text{V}}_{\text{spe}}\mathbf{X}_{\text{hsi}}$, where $N = HW$, $\mathbf{W}^{\text{V}}_{\text{spa}}$, $\mathbf{W}^{\text{Q}}$, $\mathbf{W}^{\text{K}}$, and $\mathbf{W}^{\text{V}}_{\text{spe}}$ are the corresponding projection weights. For simplicity, we omit the transformation from $H \times W \times C$ to $N \times C$. $\mathbf{V}_{\text{spa}}$ and $\mathbf{V}_{\text{spe}}$ are specific to the spatial and spectral attention modules, respectively, while $\mathbf{Q}_{\text{shr}}$ and $\mathbf{K}_{\text{shr}}$ are shared between them. The computational complexity of regular self-attention is $O\left(N^2\right)$, indicating that the computational complexity and memory requirements increase quadratically with the input resolution. Consequently, it quickly becomes a computational bottleneck for high-resolution inputs. To mitigate this, $\mathbf{K}_{\text{shr}}$ and $\mathbf{V}_{\text{spa}}$ are projected onto lower dimensions before the attention operation.

Specifically, for spatial attention, $\mathbf{K}_{\text{shr}}$ and $\mathbf{V}_{\text{spa}}$ are projected from $N \times C$ to $p \times C$ ($p \ll N$):

$$\mathbf{V}^{\text{proj}}_{\text{spa}}, \mathbf{K}^{\text{proj}}_{\text{shr}} = f\left(\mathbf{V}_{\text{spa}}\right), f\left(\mathbf{K}_{\text{shr}}\right), \tag{5}$$

where $f(\cdot)$ denotes a linear transformation, and $\mathbf{V}^{\text{proj}}_{\text{spa}}, \mathbf{K}^{\text{proj}}_{\text{shr}} \in \mathbb{R}^{p \times C}$ denote the projected spatial values and projected shared keys, respectively. The spatial self-attention is then computed as:

$$\mathbf{Y}_{\text{spa}} = \text{Softmax}\left(\frac{1}{\sqrt{d}}\mathbf{Q}_{\text{shr}}\mathbf{K}^{\text{proj}^{\text{T}}}_{\text{shr}}\right)\mathbf{V}^{\text{proj}}_{\text{spa}}, \tag{6}$$

where d is a normalization factor and $\mathbf{Y}_{\text{spa}} \in \mathbb{R}^{N \times C}$ is the output of the spatial attention module.

Similarly, the spectral attention module captures the interdependencies between channels. Using the same $\mathbf{Q}_{\text{shr}}$ and $\mathbf{K}_{\text{shr}}$ as in the spatial attention module, the spectral attention is computed as:

$$\mathbf{Y}_{\text{spe}} = \mathbf{V}_{\text{spe}} \cdot \text{Softmax}\left(\frac{1}{\sqrt{d}}\mathbf{Q}^{\text{T}}_{\text{shr}}\mathbf{K}_{\text{shr}}\right), \tag{7}$$

where $\mathbf{Y}_{\text{spe}} \in \mathbb{R}^{N \times C}$ represents the output of the spectral attention module.

Finally, a linear projection is employed to transform the outputs of the two attention modules, halving the spectral dimension and obtaining enriched feature representations. Then a concatenation operation is performed to fuse them. Therefore, the final output $\mathbf{Y} \in \mathbb{R}^{N \times C}$ of the EPA module is obtained by:

$$\mathbf{Y} = \left[f\left(\mathbf{Y}_{\text{spa}}\right), f\left(\mathbf{Y}_{\text{spe}}\right)\right], \tag{8}$$

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

With the aforementioned EPA module, the 3D DCN block can be formulated as:

$$\mathbf{X}_2 = \text{LN}(\text{EPA}(\mathbf{X}_{\text{hsi}})) + \mathbf{X}_{\text{hsi}}, \tag{9}$$

$$\mathbf{X}_{\text{out2}} = \text{Conv}_1(\text{3DFFN}(\mathbf{X}_2)) + \mathbf{X}_2, \tag{10}$$

$$\text{3DFFN} = \text{Conv}_3\left(\text{ReLU}\left(\text{Conv}_3\left(\mathbf{X}_2\right)\right)\right), \tag{11}$$

where $\mathbf{X}_2$ and $\mathbf{X}_{\text{out2}}$ denote the intermediate features and outputs of the 3D DCN block, respectively. Conv_1 and Conv_2 represent $1 \times 1 \times 1$ and $3 \times 3 \times 3$ convolutions, respectively.

D. Cross-modality Feature Enhancement and Exchange Module

Integrating information from different modalities can effectively enhance the quality and richness of information while mitigating undesired information. Thus, multimodal data fusion is essential for extracting comprehensive and reliable information from complex scenes. Consequently, the CMFeX module is developed to recalibrate and enhance modality-specific features while promoting cross-modality interaction.

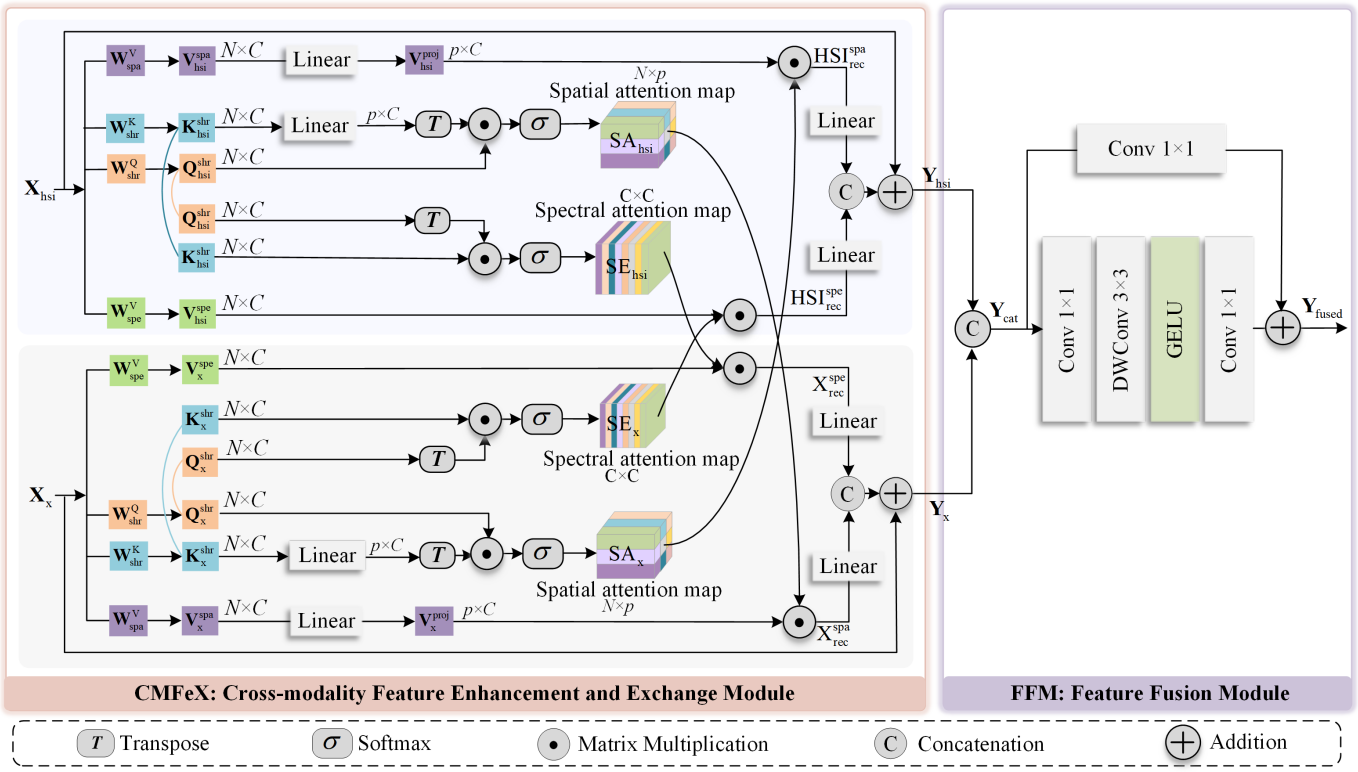


Fig. 3. Illustration of the proposed Cross-Modality Feature Extraction (CMFeX) and Feature Fusion Module (FFM). In CMFeX, features from each modality undergo linear projections, transposition, and attention computations to obtain spatial and spectral attention maps. Cross-modality matrix multiplications, linear projections, and concatenation are then applied to generate rectified HSI and X features. These features are then fused in FFM.

As shown in Fig. 3, CMFeX processes HSI and X modalities in spatial and spectral dimensions for feature calibration and interaction. Given an HSI input $\mathbf{X}_{\text{hsi}} \in \mathbb{R}^{HW\times C}$, it is linearly transformed into spatial values $\mathbf{V}_{\text{hsi}}^{\text{spa}} \in \mathbb{R}^{N\times C}$, shared queries $\mathbf{Q}_{\text{hsi}}^{\text{shr}} \in \mathbb{R}^{N\times C}$, shared keys $\mathbf{K}_{\text{hsi}}^{\text{shr}} \in \mathbb{R}^{N\times C}$, and spectral values $\mathbf{V}_{\text{hsi}}^{\text{spe}} \in \mathbb{R}^{N\times C}$. Similar to EPA, $\mathbf{K}_{\text{hsi}}^{\text{shr}}$ and $\mathbf{V}_{\text{hsi}}^{\text{spa}}$ are projected to $p \times C$ matrices through a linear layer before the attention operation:

$$\mathbf{K}_{\text{hsi}}^{\text{proj}},\ \mathbf{V}_{\text{hsi}}^{\text{proj}} = f\left(\mathbf{K}_{\text{hsi}}^{\text{shr}}\right),\ f\left(\mathbf{V}_{\text{hsi}}^{\text{spa}}\right), \tag{12}$$

where $\mathbf{K}_{\text{hsi}}^{\text{proj}},\ \mathbf{V}_{\text{hsi}}^{\text{proj}} \in \mathbb{R}^{p\times C}$ denotes the projected shared keys and projected spatial values of HSIs. Then, the spatial and spectral attention maps of HSI can be obtained by:

$$\text{SA}_{\text{hsi}} = \text{Softmax}\left(\frac{1}{\sqrt{d}}\mathbf{Q}_{\text{hsi}}^{\text{shr}}{\mathbf{K}_{\text{hsi}}^{\text{proj}}}^{\text{T}}\right), \tag{13}$$

$$\text{SE}_{\text{hsi}} = \text{Softmax}\left(\frac{1}{\sqrt{d}}{\mathbf{Q}_{\text{hsi}}^{\text{shr}}}^{\text{T}}\mathbf{K}_{\text{hsi}}^{\text{shr}}\right), \tag{14}$$

where $\text{SA}_{\text{hsi}} \in \mathbb{R}^{N\times p}$ and $\text{SE}_{\text{hsi}} \in \mathbb{R}^{C\times C}$ represent the spatial and spectral attention maps of HSIs, respectively.

Similarly, the spatial values $\mathbf{V}_{\text{x}}^{\text{spa}} \in \mathbb{R}^{HW\times C}$, shared queries $\mathbf{Q}_{\text{x}}^{\text{shr}} \in \mathbb{R}^{HW\times C}$, shared keys $\mathbf{K}_{\text{x}}^{\text{shr}} \in \mathbb{R}^{HW\times C}$, and spectral values $\mathbf{V}_{\text{x}}^{\text{spe}} \in \mathbb{R}^{HW\times C}$ of the X-modality can be obtained by linearly transforming the input $\mathbf{X}_{\text{x}} \in \mathbb{R}^{H\times W\times C}$. The spatial and spectral attention maps for the X-modality are computed as:

$$\mathbf{K}_{\text{x}}^{\text{proj}},\ \mathbf{V}_{\text{x}}^{\text{proj}} = f\left(\mathbf{K}_{\text{x}}^{\text{shr}}\right),\ f\left(\mathbf{V}_{\text{x}}^{\text{spa}}\right), \tag{15}$$

$$\text{SA}_{\text{x}} = \text{Softmax}\left(\frac{1}{\sqrt{d}}\mathbf{Q}_{\text{x}}^{\text{shr}}{\mathbf{K}_{\text{x}}^{\text{proj}}}^{\text{T}}\right), \tag{16}$$

$$\text{SE}_{\text{x}} = \text{Softmax}\left(\frac{1}{\sqrt{d}}{\mathbf{Q}_{\text{x}}^{\text{shr}}}^{\text{T}}\mathbf{K}_{\text{x}}^{\text{shr}}\right), \tag{17}$$

where $\mathbf{K}_{\text{x}}^{\text{proj}},\ \mathbf{V}_{\text{x}}^{\text{proj}} \in \mathbb{R}^{p\times C}$ represents the projected shared keys and projected spatial values of the X-modality, $\text{SA}_{\text{x}} \in \mathbb{R}^{N\times p}$ and $\text{SE}_{\text{x}} \in \mathbb{R}^{C\times C}$ denote the spatial and spectral attention maps of the X-modality, respectively.

Spatial rectification is achieved through a spatial cross-attention process, where the spatial values are multiplied by the spatial attention map of the other modality. This process can be represented as:

$$\text{HSI}_{\text{rec}}^{\text{spa}} = \text{SA}_{\text{x}} \cdot \mathbf{V}_{\text{hsi}}^{\text{proj}}, \tag{18}$$

$$\text{X}_{\text{rec}}^{\text{spa}} = \text{SA}_{\text{hsi}} \cdot \mathbf{V}_{\text{x}}^{\text{proj}}, \tag{19}$$

where $\text{HSI}_{\text{rec}}^{\text{spa}},\ \text{X}_{\text{rec}}^{\text{spa}} \in \mathbb{R}^{N\times C}$ denote the recalibrated spatial information of HSI and X, respectively.

Similar to the spatial rectification, the spectral rectification is formulated as:

$$\text{HSI}_{\text{rec}}^{\text{spe}} = \mathbf{V}_{\text{hsi}}^{\text{spe}} \cdot \text{SE}_{\text{x}}, \tag{20}$$

$$\text{X}_{\text{rec}}^{\text{spe}} = \mathbf{V}_{\text{x}}^{\text{spe}} \cdot \text{SE}_{\text{hsi}}, \tag{21}$$

where $\text{HSI}_{\text{rec}}^{\text{spe}},\ \text{X}_{\text{rec}}^{\text{spe}} \in \mathbb{R}^{N\times C}$ are the recalibrated spectral information of HSI and X, respectively.

The final recalibrated representations, $\mathbf{X}_{\text{hsi}}^{\text{rec}}$ and $\mathbf{X}_{\text{x}}^{\text{rec}}$, are obtained by reducing the channel dimensions of the recalibrated

spatial and spectral features using a linear layer, followed by concatenation to integrate spatial and spectral information:

$$\mathbf{X}_{\text{hsi}}^{\text{rec}} = \left[f\left(\text{HSI}_{\text{rec}}^{\text{spa}}\right),\ f\left(\text{HSI}_{\text{rec}}^{\text{spe}}\right)\right], \tag{22}$$

$$\mathbf{X}_{\text{x}}^{\text{rec}} = \left[f\left(\text{X}_{\text{rec}}^{\text{spa}}\right),\ f\left(\text{X}_{\text{rec}}^{\text{spe}}\right)\right], \tag{23}$$

where $\mathbf{X}_{\text{hsi}}^{\text{rec}}, \mathbf{X}_{\text{x}}^{\text{rec}} \in \mathbb{R}^{N\times C}$ denote the rectified features for HSI and X modalities, respectively.

To facilitate gradient flow and mitigate the vanishing gradient problem, residual connections are established between the rectified features and their corresponding inputs:

$$\mathbf{Y}_{\text{hsi}} = \mathbf{X}_{\text{hsi}}^{\text{rec}} + \mathbf{X}_{\text{hsi}}, \tag{24}$$

$$\mathbf{Y}_{\text{x}} = \mathbf{X}_{\text{x}}^{\text{rec}} + \mathbf{X}_{\text{x}}, \tag{25}$$

where $\mathbf{Y}_{\text{hsi}}, \mathbf{Y}_{\text{x}} \in \mathbb{R}^{N\times C}$ are the final outputs of the CMFeX. These outputs are subsequently sent to the next stage for further information learning and simultaneously fed to FFM for cross-modality information fusion.

E. Feature Fusion Module

After obtaining the outputs $\mathbf{Y}_{\text{hsi}}$ and $\mathbf{Y}_{\text{x}}$ of CMFeX, they are reshaped to $H\times W\times C$ and concatenated along the channel dimension:

$$\mathbf{Y}_{\text{cat}} = \left[\mathbf{Y}_{\text{hsi}},\ \mathbf{Y}_{\text{x}}\right]. \tag{26}$$

The concatenated features $\mathbf{Y}_{\text{cat}}$ are then processed by a depthwise feedforward network (DWFFN) with a residual connection to produce the cross-modality fused features:

$$\mathbf{Y}_{\text{fused}} = \text{DWFFN}\left(\mathbf{Y}_{\text{cat}}\right) + \text{Conv}_{1\times1}\left(\mathbf{Y}_{\text{cat}}\right), \tag{27}$$

where the DWFFN is defined as two 1×1 convolution layers separated by a 3×3 depthwise convolution (DWConv) and a GELU activation function:

$$\text{DWFFN} = \text{Conv}_{1\times1}(\text{GELU}\left(\text{DWConv}\left(\text{Conv}_{1\times1}\left(\mathbf{Y}_{\text{cat}}\right)\right)\right). \tag{28}$$

Through this process, $\mathbf{Y}_{\text{cat}} \in \mathbb{R}^{H\times W\times 2C}$ is fused into $\mathbf{Y}_{\text{fused}} \in \mathbb{R}^{H\times W\times C}$, which is passed to the decoder for further feature learning.

F. Lightweight All-MLP Decoder

Given that the encoder allows for a larger effective receptive field, we adopt the lightweight ALL-MLP decoder [5] for efficient decoding and prediction. As illustrated in Fig. 1, the decoder contains only a few MLP layers. The multi-level fused features from each stage are first fed into an MLP layer to unify their channel dimension and then upsampled. The results are concatenated and then passed through another two successive MLP layers for fusion and prediction, respectively. By avoiding computationally intensive components, this streamlined design makes CoMiX simpler, more efficient, and easily adaptable.

IV. EXPERIMENTS AND RESULTS

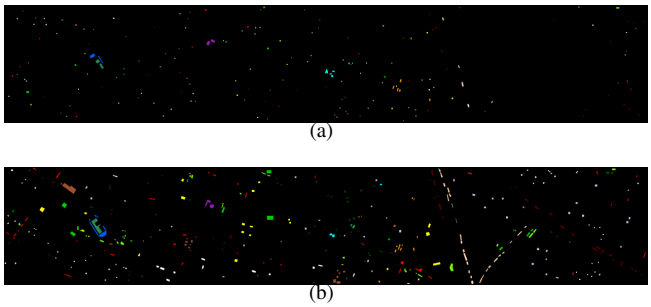
A. Description of Datasets

We conducted experiments on three public multi-modality benchmarks that combine HSI-DSM, HSI-SAR, and HSI-multispectral LiDAR (MS-LiDAR) data, respectively.

The Houston2013 dataset [61] covers the University of Houston campus and its adjacent urban region. This dataset comprises two data sources: an HSI and a DSM derived from LiDAR, both sharing identical spatial size (349×1905) and spatial resolution (2.5 m). The HSI contains 144 spectral bands, spanning the wavelength range from 380 nm to 1050 nm. This scene contains 15 classes, as illustrated in Fig. 4 and Table I.

TABLE I
THE LAND-COVER TYPES AND THE NUMBER OF TRAINING AND TEST SAMPLES ON THE DFC2013 DATASET

ID	Color	Land-cover Type	Training	Test
C1		Healthy grass	198	1053
C2		Stressed grass	190	1064
C3		Synthetic grass	192	505
C4		Trees	188	1056
C5		Soil	186	1056
C6		Water	182	143
C7		Residential	196	1072
C8		Commercial	191	1053
C9		Road	193	1059
C10		Highway	191	1036
C11		Railway	181	1054
C12		Parking Lot 1	192	1041
C13		Parking Lot 2	184	285
C14		Tennis Court	181	247
C15		Running Track	187	473



(a) (b)

Fig. 4. Houston2013 dataset: spatial distribution of (a) the training samples and (b) the test samples.

The Berlin dataset [62], obtained from the urban and surrounding rural areas of Berlin, is a comprehensive collection of HSI and SAR images. The HSI is a $797\times220\times224$ data cube with a spatial resolution of 30 m and covers a wavelength range of 0.4 to 2.5 μm. In contrast, the SAR image consists of 1723×476 pixels with a spatial resolution of 13.89 m. To match the spatial resolution of the SAR image, the HSI was interpolated using the nearest neighbor technique. Details on land-cover classes, including the number of samples and their spatial distribution, can be found in Fig. 5 and Table II.

The DFC2018 dataset [41] covers the University of Houston campus and its adjacent urban areas. The dataset adopted in this study includes HSI data with a spatial resolution of 1

TABLE II
THE LAND-COVER TYPES AND THE NUMBER OF TRAINING AND TEST SAMPLES ON THE BERLIN DATASET

ID	Color	Land-cover Type	Training	Test
C1		Forest	443	54511
C2		Residential Area	423	268219
C3		Industrial Area	499	19067
C4		Low Plants	376	58906
C5		Soil	331	17095
C6		Allotment	280	13025
C7		Commercial Area	298	24526
C8		Water	170	6502

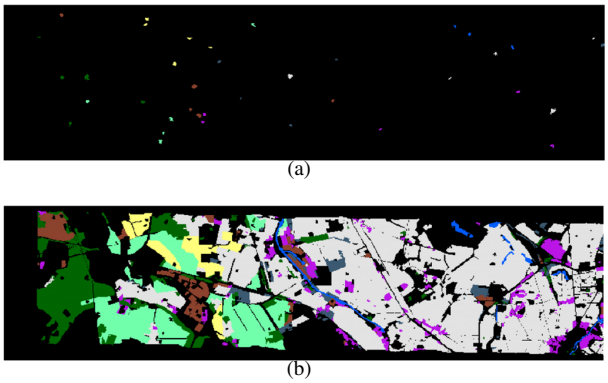


Fig. 5. Berlin dataset: spatial distribution of (a) the training samples and (b) the test samples.

TABLE III
THE LAND-COVER TYPES AND THE NUMBER OF TRAINING AND TEST SAMPLES ON THE DFC2018 DATASET

ID	Color	Land-cover Type	Training	Test
C1		Healthy grass	39196	20000
C2		Stressed grass	130008	20000
C3		Artificial turf	2736	20000
C4		Evergreen trees	54322	20000
C5		Deciduous trees	20172	20000
C6		Bare earth	18064	20000
C7		Water	1064	1628
C8		Residential buildings	158995	20000
C9		Non-residential buildings	894769	20000
C10		Roads	183283	20000
C11		Sidewalks	136035	20000
C12		Crosswalks	6059	5345
C13		Major thoroughfares	185438	20000
C14		Highways	39438	20000
C15		Railways	27748	11232
C16		Paved parking lots	45932	20000
C17		Unpaved parking lots	587	3524
C18		Cars	26289	20000
C19		Trains	21479	20000
C20		Stadium seats	27296	20000

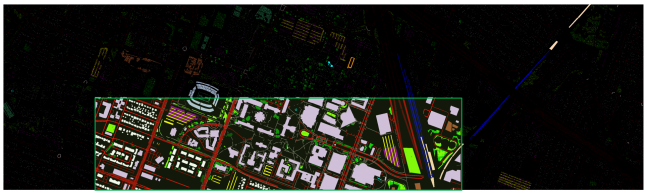


Fig. 6. DFC2018 dataset: spatial distribution of training (green box) and test (outside the green box) samples.

m and MS-LiDAR point cloud data with a spatial resolution of 0.5 m. The HSI dataset encompasses 4172×1202 pixels distributed across 48 bands, covering the spectral range from 380 to 1050 nm. Meanwhile, MS-LiDAR contains a DSM, a digital elevation model (DEM), and three intensity rasters at distinct wavelengths: 1550 nm (near-infrared), 1064 nm (mid-infrared), and 532 nm (green). To match the spatial resolution of MS-LiDAR images, the HSI was interpolated to a spatial resolution of 0.5 m using the nearest neighbor interpolation algorithm. To obtain the actual elevation of objects, the normalized DSM (NDSM) value was calculated via NDSM $=$ DSM$-$DEM. The dataset includes 20 land-cover classes, with more detailed information available in Table III and Fig. 6.

Before the experiments, these three datasets were normalized to [0, 1] to standardize the magnitude of the data, thereby enhancing network convergence during training.

B. Experimental Settings

1) Comparison Methods: We compared our CoMiX with the support vector machine (SVM) [63], and several DL-based models, including FusAtNet [27], CALC [38], Fusion_HCT [44], Fusion-FCN [41], Flex-MCFNet [40], and LoGoCAF [24]. SVM is a pixel-wise classifier that works at the pixel level. FusAtNet, CALC, and Fusion_HCT are patch-based classification networks designed for HSI and LiDAR data. Specifically, FusAtNet generates HSI-derived and LiDAR-derived attention maps that enhance spectral and spatial information of HSIs, respectively. CACL develops a coupled adversarial feature learning subnetwork to extract semantic features from HSI and LiDAR data, followed by a multi-level feature fusion classification subnetwork. Fusion_HCT sequentially employs CNN and transformer blocks for feature extraction, followed by a cross-token attention fusion module. Unlike the aforementioned networks, Fusion-FCN is an FCN-based architecture in which feature maps retain the same spatial resolution as the input images at all levels. Flex-MCFNet integrates a flexible mixup data augmentation strategy to enable comprehensive multimodal fusion of HSI and X-modality data (e.g., MSI, SAR, LiDAR) within a patch-based classification framework. For a fair comparison, we implemented the model without the mixup strategy, termed MCFNet, which retains the same architecture as CMX. Finally, LoGoCAF is a HSI-X multimodal segmentation framework that optimally balances accuracy, efficiency, and versatility.

2) Implementation Details: We implemented our CoMiX framework on the PyTorch platform using the Adam optimizer and the cross-entropy loss function. CoMiX was trained for 500 epochs with a weight decay of 0.01. The initial learning rate was set to 6×10^{-5} for the Houston2013 and DFC2018 datasets, and 1×10^{-3} for the Berlin dataset. The learning rate was updated during training using the poly-learning rate schedule. The input size and batch size were set to 128×128 and 4, respectively, due to GPU memory constraints. The entire network is trained in an end-to-end manner.

For the comparison methods, publicly available source codes and literature-reported hyperparameters were used, with

TABLE IV
CLASSIFICATION ACCURACY ON THE HOUSTON2013 DATASET. THE NUMBERS AFTER ± ARE THE STANDARD DEVIATIONS OF THE CORRESPONDING METRICS. THE BEST VALUES ARE MARKED IN **BOLD**, AND THE SECOND-BEST VALUES ARE UNDERLINED

Method	SVM-X [63]	SVM-HSI [63]	SVM [63]	FusAtNet [27]	CACL [38]	Fusion_HCT [44]	Fusion-FCN [41]	MCFNet [40]	LoGoCAF [24]	CoMiX (Ours)
OA (%)	17.16±0	55.51±0	64.68±0	89.02±1.03	86.24±1.56	88.95±1.48	83.84±1.67	91.76±1.31	92.11±1.18	**95.75**±0.98
AA (%)	16.66±0	59.83±0	67.14±0	92.28±1.18	88.10±2.18	90.48±1.91	85.42±2.16	93.17±1.45	93.33±1.06	**96.23**±0.86
$\kappa \times 100$	10.76±0	52.12±0	61.82±0	88.09±1.14	85.11±1.92	88.00±1.62	82.50±1.61	91.11±1.56	91.44±1.36	**95.39**±0.79
C1	55.56	81.96	82.43	82.96	80.63	82.48	83.00	85.77	83.10	**90.60**
C2	0.00	75.00	80.83	**97.70**	80.83	83.93	84.02	87.40	85.06	85.34
C3	**100.00**	99.01	99.01	**100.00**	85.55	98.22	**100.00**	99.86	**100.00**	**100.00**
C4	15.63	89.21	88.35	95.74	91.00	92.36	90.72	94.37	**100.00**	**100.00**
C5	0.00	85.23	90.06	98.96	99.34	99.91	99.34	99.92	**100.00**	**100.00**
C6	0.00	78.32	78.32	**100.00**	93.01	91.72	99.30	**100.00**	**100.00**	98.60
C7	47.30	28.64	69.50	92.19	83.12	87.35	82.46	93.22	86.29	**93.75**
C8	31.43	12.92	54.42	82.81	85.28	91.19	90.98	93.81	83.95	**93.92**
C9	0.00	81.30	83.00	85.35	86.03	79.55	75.54	95.53	80.55	**96.03**
C10	0.00	1.83	2.90	66.12	60.91	67.31	54.83	71.39	94.40	**99.81**
C11	0.00	55.88	64.52	85.06	94.88	96.40	92.60	83.81	**99.24**	94.97
C12	0.00	0.10	0.00	89.63	88.38	95.89	70.70	96.76	**100.00**	**99.23**
C13	0.00	11.93	15.09	86.32	92.98	90.88	59.30	**95.74**	87.72	91.23
C14	0.00	97.17	99.60	99.19	99.60	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
C15	0.00	98.94	99.15	99.58	**100.00**	**100.00**	98.52	**100.00**	99.58	**100.00**

TABLE V
CLASSIFICATION ACCURACY ON THE BERLIN DATASET. HE NUMBERS AFTER ± ARE THE STANDARD DEVIATIONS OF THE CORRESPONDING METRICS. THE BEST VALUES ARE MARKED IN **BOLD**, AND THE SECOND-BEST VALUES ARE UNDERLINED

Method	SVM-X [63]	SVM-HSI [63]	SVM [63]	FusAtNet [27]	CACL [38]	Fusion_HCT [44]	Fusion-FCN [41]	MCFNet [40]	LoGoCAF [24]	CoMiX (Ours)
OA (%)	27.54±0	61.34±0	62.98±0	73.86± 1.84	75.39±1.76	76.46±1.51	55.14±1.74	70.84±1.41	76.53±1.32	**76.81**±1.01
AA (%)	20.11±0	61.15±0	62.37±0	66.21±2.08	63.83±2.31	61.47±2.13	57.52±2.32	**68.57**±1.51	64.69±0.98	64.81±0.95
$\kappa \times 100$	10.25±0	47.59±0	49.53±0	60.72±1.97	61.89±2.08	63.38±1.92	40.23±1.74	58.51±1.59	64.07±1.27	**64.21**±1.08
C1	**88.27**	80.34	80.23	50.05	47.33	59.50	43.19	74.14	73.02	65.13
C2	26.18	57.41	58.34	81.09	**86.43**	85.89	54.44	70.92	88.75	84.81
C3	46.43	46.43	48.30	48.51	41.13	65.57	34.15	53.12	52.64	**67.20**
C4	0.00	76.36	84.88	85.58	80.30	85.57	69.40	83.14	68.45	**87.86**
C5	0.00	74.24	74.40	**96.38**	96.27	83.54	78.15	77.06	80.95	79.68
C6	0.00	60.85	59.75	59.74	60.14	55.03	64.16	**75.08**	61.65	50.39
C7	0.00	27.84	27.25	29.88	26.09	12.72	**46.77**	38.86	20.46	19.45
C8	0.00	65.73	65.80	**78.42**	72.95	43.92	69.92	76.28	71.58	63.93

identical preprocessing and training/test sets to ensure fair comparisons. Five independent runs were performed for each experiment, and the average metrics were reported to ensure result objectivity.

3) Metric: The performance of different approaches was evaluated using four metrics: overall accuracy (OA), average accuracy (AA), the kappa coefficient (κ), and the producer accuracy (PA) for each category.

C. Comparison with Other Methods

1) Quantitative Results and Analysis: The quantitative results are presented in Tables IV–VI, with the best and second-best results in each row highlighted in bold and underlined, respectively. Note that SVM classifies HSI data, X data, and combined HSI-X data, respectively, while the other approaches focus on HSI-X data.

The experimental results across the three datasets show a downward trend in accuracy for all comparison methods: over 80% on Houston2013, around 70% on Berlin, and approximately 60% on DFC2018. This pattern reflects increasing recognition difficulty, likely due to differences in data sources, category definitions, and the distribution of training and test samples [20]. Nevertheless, our CoMiX consistently demonstrated superior performance in terms of OA across all three datasets. For example, on the Houston dataset, CoMiX achieved the best OA of 95.75%, outperforming SVM, FusAtNet, CALC, Fusion_HCT, Fusion-FCN, MCFNet, and LoGoCAF by remarkable margins of 31.07%, 6.73%, 9.51%, 6.80%, 11.91%, 3.99%, and 3.64% respectively.

SVM performs significantly better on HSI data than on X-modality data, with margins of 38.35%, 33.80%, and 15.48% on Houston2013, Berlin, and DFC2018, respectively, indicating that HSIs contain more discriminative features. Combining HSI and X data further improves SVM accuracy across all datasets, highlighting the important of multimodal feature fusion. Nevertheless, SVM still underperforms compared to DL-based methods. As a conventional pixel-level classifier, it processes raw pixels directly and cannot fully exploit discriminative contextual information.

In contrast, DL-based approaches that incorporate spatial context consistently achieve improved performance across all three datasets. For instance, patch-based frameworks (i.e., FusAtNet, CALC, Fusion_HCT, and MCFNet) achieve higher

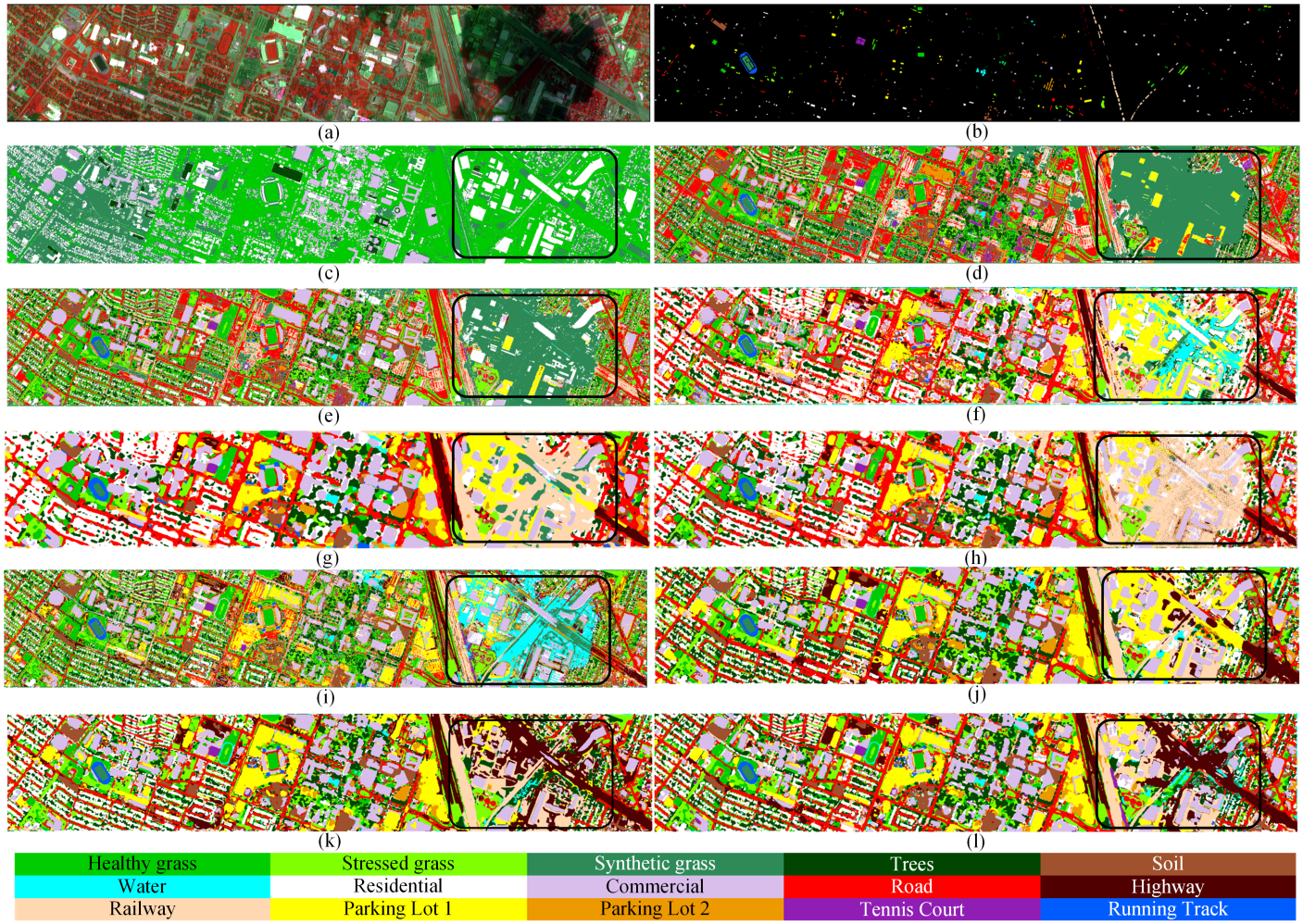


Fig. 7. Classification maps (with a black box on an area of interest) provided by different methods on the Houston2013 dataset. (a) False Color Image. (b) Ground Truth. (c) SVM-X. (d) SVM-HSI. (e) SVM. (f) FusAtNet. (g) CACL. (h) Fusion_HCT. (i) Fusion-FCN. (j) MCFNet. (k) LoGoCAF. (l) CoMiX.

OA, exceeding 80% on Houston2013 and 70% on Berlin. However, their limited patch size restricts the ability to capture long-range spatial dependencies, resulting in the loss of critical details. Overcoming this limitation, Fusion-FCN and LoGo-CAF process larger input images for pixel-wise classification. Despite this, Fusion-FCN performs comparably to or worse than patch-based networks, as its effective receptive field is still constrained by stacked 3×3 convolutions, limiting its capacity to model long-range dependencies. LoGoCAF overcomes this by integrating CNNs and transformers for local-to-global information modeling and employing two cross-modality modules to enhance cross-modality information enhancement, interaction, and fusion.

Compared with the aforementioned approaches, CoMiX achieves the highest OA and κ values. CoMiX not only processes larger input images for pixel-wise classification but also effectively models local- and long-range dependencies. Moreover, the 2D DCN and 3D DCN blocks for adaptive modality-specific feature extraction, together with CMFeX and FFM for cross-modality interaction and fusion, enables a robust and comprehensive approach to HSI-X semantic segmentation. Although CoMiX's AA is slightly lower than that of MCFNet and FusAtNet on the Berlin dataset, it still achieves the best OA and κ across all three datasets. The slightly reduced AA is primarily due to CoMiX's sensitivity to categories with fewer samples. This is a common challenge in class-imbalanced scenarios where the network may prioritize dominant classes over underrepresented ones.

2) Qualitative Results and Analysis: The classification maps, along with the corresponding false color and ground truth (GT) images, are shown in Figs. 7–9.

As shown in Figs. 7(c), 8(c), and 9(c), SVM applied to X-modality data can only distinguish broad categories, such as vegetation, industrial, and residential areas. Its performance improves on HSI data, highlighting the importance of rich spectral information for accurate object identification. Combining HSI and X data further refines the classification maps, consistent with the quantitative results. Nevertheless, SVM-generated maps still exhibit significant salt-and-pepper noise due to the pixel-level fusion strategy, and the lack of spatial context hinders discrimination of land covers with similar spectral characteristics.

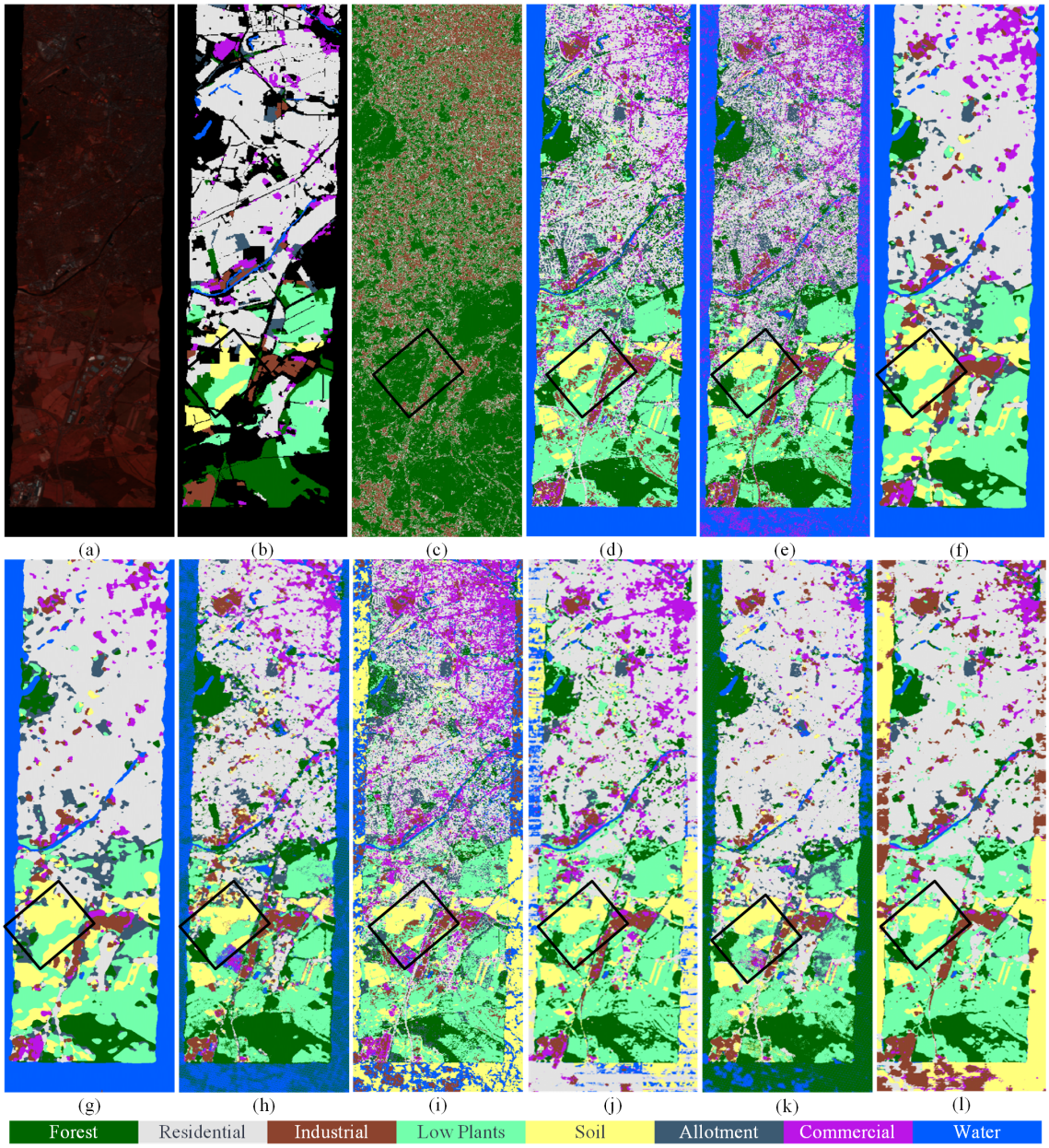


Fig. 8. Classification maps (with a black box on an area of interest) provided by different methods on the Berlin dataset. (a) False Color Image. (b) Ground Truth. (c) SVM-X. (d) SVM-HSI. (e) SVM. (f) FusAtNet. (g) CACL. (h) Fusion_HCT. (i) Fusion-FCN. (j) MCFNet. (k) LoGoCAF. (l) CoMiX.

Patch-based networks, such as FusAtNet, Fusion_HCT, and MCFNet, achieve improved visual performance by integrating spectral and spatial features and employing adaptive cross-modality feature fusion. However, some models, including CACL and Fusion_HCT, produce distortions at object boundaries because they assume that all pixel within a patch contributes equally, which is ineffective in heterogeneous regions. Similarly, Fusion-FCN yields blurred boundaries and numerous misclassifications, particularly in low-light or shaded areas.

In contrast, CoMiX consistently produces high-quality classification maps with sharper boundaries and smoother object representations. Benefiting from a single 2× downsampling and full contextual capture, CoMiX preserves high-resolution details and performs robustly under challenging conditions. For example, accurately delineating road, residential, and green vegetation in the Houston2013 and DFC2013 scenes, and the well-preserved low plants on the Berlin dataset. By enhancing modality-specific feature extraction and cross-modality fusion, CoMiX produces highly accurate semantic segmentation maps, particularly excelling in texture and edge details. To facilitate a more intuitive evaluation, regions of interest (ROIs; black boxes in Figs. 7–9) were selected, revealing that CoMiX produces maps with finer and more reliable details.

Overall, the qualitative analysis confirms CoMiX's effectiveness in multimodal data fusion and its ability for robust semantic scene understanding.

TABLE VI
CLASSIFICATION ACCURACY ON THE DFC2018 DATASET. THE NUMBERS AFTER ± ARE THE STANDARD DEVIATIONS OF THE CORRESPONDING METRICS. THE BEST VALUES ARE MARKED IN **BOLD**, AND THE SECOND-BEST VALUES ARE UNDERLINED

Method	SVM-X [63]	SVM-HSI [63]	SVM [63]	FusAtNet [27]	CACL [38]	Fusion_HCT [44]	Fusion-FCN [41]	MCFNet [40]	LoGoCAF [24]	CoMiX (Ours)
OA (%)	25.67±0	41.15±0	56.60±0	61.24±0.98	60.06±1.56	62.90±1.18	62.87±1.47	63.04±1.56	65.50±1.17	**68.26**±0.59
AA (%)	21.93±0	36.45±0	49.90±0	56.17±1.11	53.26±1.89	56.84±2.11	57.00±2.39	57.32±1.49	61.23±1.08	**63.57**±0.85
$\kappa\times100$	21.05±0	37.50±0	53.91±0	58.79±1.06	57.61±1.50	60.66±1.71	60.62±1.71	60.76±1.62	63.47±1.26	**66.31**±0.61
C1	8.11	95.82	96.18	95.27	94.60	96.19	96.53	95.54	**98.34**	97.26
C2	90.89	**91.86**	91.59	87.13	89.99	86.01	87.80	86.74	84.13	85.61
C3	0.00	40.50	54.81	15.72	60.15	53.30	48.79	40.42	14.08	**84.20**
C4	66.37	86.95	95.71	97.28	94.43	95.39	95.99	96.61	**98.23**	93.29
C5	0.00	28.02	55.32	57.97	48.75	53.00	**60.02**	52.86	58.87	54.95
C6	0.00	26.72	49.12	85.20	45.86	56.58	62.17	66.74	**88.34**	49.13
C7	0.00	28.13	33.35	81.23	31.70	56.05	59.37	72.85	**99.82**	86.79
C8	59.82	24.34	35.32	**79.32**	51.96	64.65	59.33	69.66	76.51	72.84
C9	87.02	85.53	88.34	92.15	90.72	91.04	89.46	92.10	91.95	**95.76**
C10	84.09	61.28	73.35	87.95	65.07	56.54	78.37	85.44	**89.04**	82.98
C11	39.57	35.86	53.98	64.73	56.84	59.94	60.62	55.15	66.73	**77.96**
C12	0.00	0.00	0.00	8.61	9.36	9.02	10.05	0	11.73	**14.61**
C13	2.53	44.05	**49.84**	33.41	49.47	44.46	45.72	41.96	31.12	49.13
C14	0.00	8.52	9.45	14.22	24.81	26.75	21.65	**41.39**	18.92	19.67
C15	0.00	0.05	0.47	9.43	6.94	8.76	9.04	5.10	11.39	**33.41**
C16	0.00	15.56	46.19	**85.34**	49.72	68.57	55.61	64.56	85.13	69.47
C17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C18	0.03	0.83	38.39	33.78	50.35	48.36	**52.53**	38.27	46.44	47.64
C19	0.16	6.74	59.47	74.12	75.12	88.12	82.19	84.65	90.47	**96.51**
C20	0.01	48.22	67.13	20.48	69.41	**74.00**	64.77	56.36	63.33	60.17

V. DISCUSSION

A. Efficiency Analysis

The structural complexity of CoMiX was assessed on the Houston2013 dataset by comparing its number of parameters (Params), floating-point operations (FLOPs), training time, and inference time with those of other methods, as summarized in Table VII.

TABLE VII
NUMBER OF PARAMS AND FLOPS, TRAINING (TRAIN) AND INFERENCE (INFER) TIME OF DIFFERENT METHODS ON THE HOUSTON2013 DATASET

Method	Params (M)	FLOPs (G)	Train (s)	Infer (s)
SVM [63]	—	—	0.18	2.85
FusAtNet [27]	36.90	221.61	50583.30	57.15
CACL [38]	0.34	0.81	1696.08	54.21
Fusion_HCT [44]	0.43	0.51	2126.89	10.45
Fusion-FCN [41]	0.09	6.22	6720.78	9.60
MCFNet [40]	7.14	83.81	7859.25	64.21
LoGoCAF [24]	7.75	93.68	9444.23	22.63
CoMiX (Ours)	21.87	197.02	8282.43	7.72

SVM takes the shortest training and inference time among all methods. Among the DL-based approaches, CoMiX achieves the fastest inference speed despite its higher number of Params and FLOPs compared to others. As emphasized in previous studies [37], although Params and FLOPs are commonly used to evaluate model complexity, they do not necessarily correlate with efficiency. In practice, efficiency is influenced by many factors beyond just Params and FLOPs. For example, despite Fusion_HCT has more Params than CACL, Fusion-FCN, and MCFNet, it achieves shorter training and inference times due to its optimized design.

Although CoMiX has more Params and FLOPs than methods like Fusion-FCN and LoGoCAF, its performance significantly surpasses theirs. CoMiX strikes an excellent balance between accuracy and efficiency, with competitive or faster training and inference speed compared to networks with fewer parameters, such as FusAtNet and MCFNet. On the other hand, models like CACL and MCFNet have lower Params and FLOPs, but suffer from excessively long inference times, making them impractical for large-scale or real-time applications.

The superior performance of CoMiX is largely attributed to its innovative architecture, which integrates 3D DCN and 2D DCN blocks for HSI and X feature extraction, respectively, and incorporates CMFeX and FFM modules for effective cross-modal feature enhancement, interaction, and fusion. In contrast, simpler models like Fusion-FCN rely solely on convolutional and pooling layers with point-wise addition fusion, limiting their ability to capture complex dependencies and interactions between modalities.

In summary, CoMiX strikes an effective balance between accuracy and efficiency. It delivers superior cross-modality feature extraction and inference speed, making it well-suited for real-world applications despite its relatively elaborate architecture.

B. Ablation Studies

A series of ablation experiments were conducted on the Houston2013 dataset to evaluate the effectiveness of 2D DCN block, 3D DCN block, CMFeX and FFM. The corresponding results are presented in Table VIII. The ablation studies take a two-branch backbone with ViT blocks as the encoder and the lightweight All-MLP decoder as the baseline. If CMFeX is ablated, features are extracted independently in their respective branches. If FFM is removed, we simply average the two features for fusion.

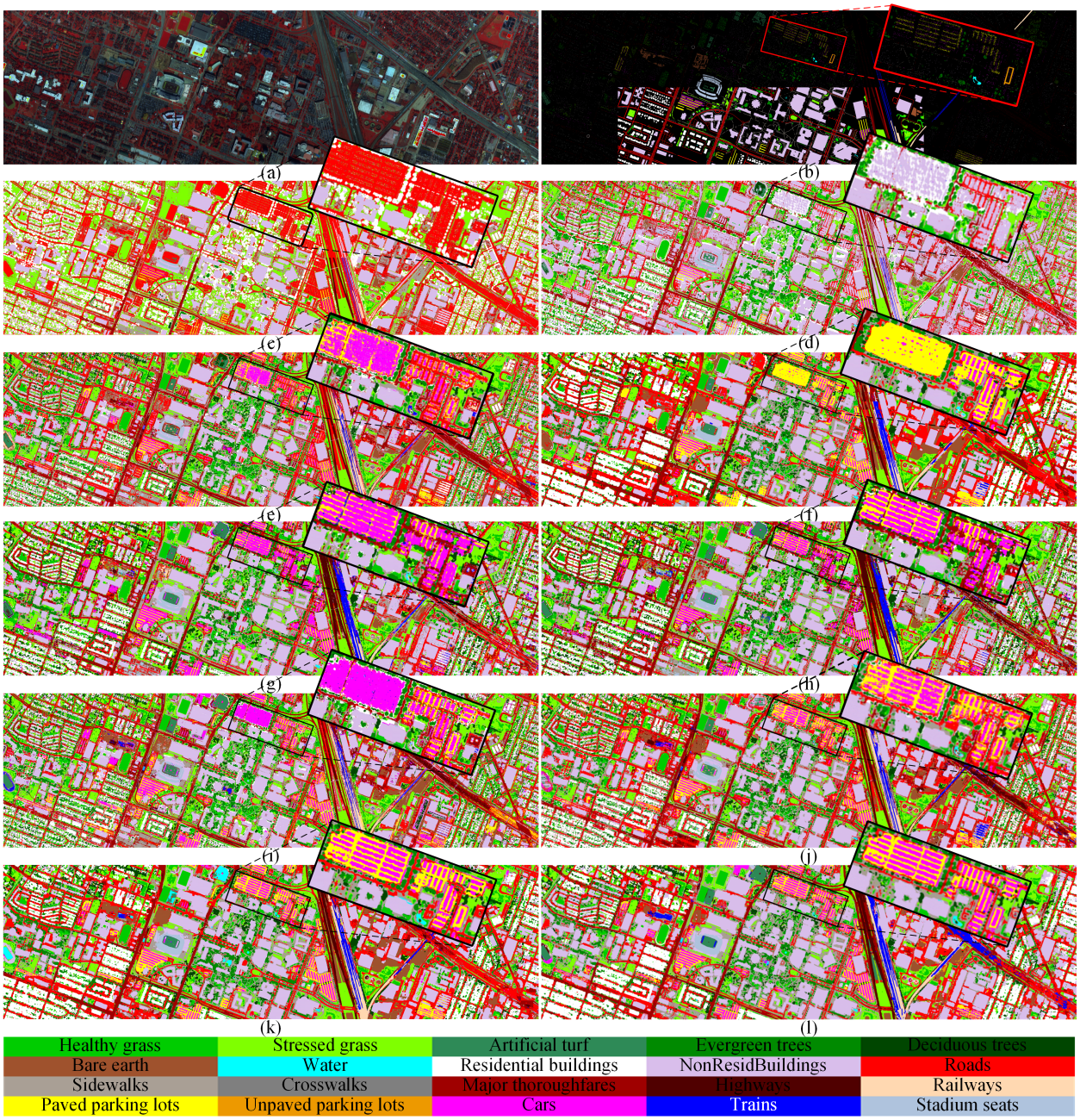


Fig. 9. Classification maps (with a zoom on an area of interest) provided by different methods on the DFC2018 dataset. (a) False Color Image. (b) Ground Truth. (c) SVM-X. (d) SVM-HSI. (e) SVM. (f) FusAtNet. (g) CACL. (h) Fusion_HCT. (i) Fusion-FCN. (j) MCFNet. (k) LoGoCAF. (l) CoMiX.

1) Effectiveness of the 2D DCN Block: We first replaced the ViT blocks in the two branches with 2D DCN blocks. The results in Table VIII(b) demonstrate that integrating the 2D DCN block improves the OA by 4.72%, 1.93%, and 2.31% on the Houston, Berlin and DFC datasets, respectively, compared with the baseline, highlighting its superior feature representation capabilities.

2) Effectiveness of the 3D DCN Block: Next, we replaced the 2D DCN block in the HSI branch with our 3D DCN block. As illustrated in Table VIII(c), this substitution improved the OA from 87.23%, 71.27%, and 64.01% to 89.08%, 72.67%, and 64.81 on the Houston, Berlin, and DFC datasets, respectively, demonstrating the superior performance of the 3D DCN block over the 2D DCN block for HSI feature extraction. In addition, this suggests that deploying 2D DCN blocks for the X-modality and 3D DCN blocks for the HSI modality is an effective design choice to fully explore the specific features of both modalities.

3) Effectiveness of CMFeX: We evaluated several variants of CMFeX, as summarized in Table VIII. Here, "spa" denotes spatial-only rectification, while "spe" represents spectral-only rectification. The spatial-only variant increases OA by 2.84%, 0.90%, and 1.87%, whereas the spectral-only variant yields gains of 2.19%, 0.61%, and 2.17% on the Houston, Berlin, and

TABLE VIII
ABLATION ANALYSIS OF THE PROPOSED COMIX

Module		(a) Baseline	(b) 2D DCN block	(c) 3D DCN block	(d) CMFeX (spa)	(e) CMFeX (spe)	(f) CMFeX	(g) FFM
(a) Baseline		✓	✓	✓	✓	✓	✓	✓
(b) 2D DCN block			✓	✓	✓	✓	✓	✓
(c) 3D DCN block				✓	✓	✓	✓	✓
(d) CMFeX (spa)					✓	✗	✓	✓
(e) CMFeX (spe)					✗	✓	✓	✓
(f) CMFeX							✓	✓
(g) FFM								✓
OA (%)	Houston	82.51	87.23	89.08	91.92	91.27	93.72	**95.75**
	Berlin	69.34	71.27	72.67	73.57	73.28	75.79	**76.81**
	DFC	61.70	63.01	63.81	65.68	65.98	67.68	**68.26**

DFC datasets, respectively. Replacing these variants with the full CMFeX module further enhances performance, demonstrating its effectiveness in calibrating and fusing cross-modal features across both spatial and spectral dimensions.

4) Effectiveness of FFM: As listed in Table VIII(g), using CMFeX alone increases OA to 93.72%, 75.79%, and 67.68% on the Houston, Berlin, and DFC datasets, respectively. Incorporating FFM further boosts OA to 95.75%, 76.81%, and 68.26%, respectively, demonstrating the importance of combining CMFeX and FFM for effective HSI-X information fusion.

After integrating the 2D DCN block, 3D DCN block, CMFeX and FFM, CoMiX achieves superior performance by effectively learning, calibrating, and fusing features from heterogeneous data sources, significantly enhancing its accuracy and robustness in segmentation tasks.

C. Cross-Modality Fusion Modules Analysis

To further validate the effectiveness of the proposed CMFeX and FFM modules, we performed a set of controlled experiments that compare them with several representative fusion strategies. Specifically, we replace CMFeX and FFM with the following alternatives:

- Feature addition: Element-wise addition of features from both modalities, followed by averaging to obtain the fused representation.
- Feature concatenation: Channel-wise concatenation of features from both modalities, followed by a linear projection layer to compress the concatenated feature map to the original dimensionality.
- Cross-attention: A basic cross-attention mechanism applied between HSI and X-modality features.
- Pag: A pixel-attention-guided (Pag) fusion module proposed in [64], used as a replacement for both CMFeX and FFM.
- Pag + FFM: Pag replaces CMFeX, and its original additive fusion is further replaced by our Feature Fusion Module (FFM).
- CM-FRM + FFM: The cross-modal feature rectification module (CM-FRM) and feature fusion module (FFM) introduced in CMX [14].

Table IX summarizes the results of all variants in terms of OA, AA, training time, and inference time. Although simple addition and concatenation offer fast training and inference, they fail to capture the complex interdependencies between HSI and X modalities, leading to limited segmentation accuracy. Cross-attention improves performance, but comes at the cost of increased training overhead.

TABLE IX
COMPARISON OF CROSS-MODALITY FUSION MODULES

Fusion Module	OA (%)	AA (%)	Train (s)	Infer (s)
Feature addition	89.08	89.56	**5209.32**	**5.75**
Feature concatenation	89.97	90.26	6109.31	5.89
Cross-attention	91.24	92.68	7860.28	7.16
Pag [64]	92.15	92.46	7569.00	6.35
Pag [64] + FFM (Ours)	93.54	94.62	7845.40	8.52
CM-FRM + FFM [14]	93.73	94.25	8365.77	9.71
CoMiX (Ours)	**95.75**	**96.23**	8282.43	7.72

The Pag module [64] alone offers a notable improvement over basic fusion strategies, achieving an OA of 92.15%, with moderate training and inference costs. When integrated with FFM (Pag + FFM), the performance is further boosted to 93.54% OA, demonstrating that FFM effectively enhances the quality of fused features. However, this gain comes with increased training and inference time.

On the other hand, CM-FRM + FFM [14] achieves comparable performance (93.73% OA), benefiting from the rectification mechanism in CM-FRM. Nevertheless, it incurs the highest training and inference time among these fusion modules, indicating lower efficiency than our design.

In contrast, CoMiX achieves the highest OA and AA while maintaining competitive training and inference efficiency, demonstrating the effectiveness of its structured design.

D. Impact of the Number of Training Samples

The performance of DL-based methods is highly dependent on the number of training samples, making it necessary to assess CoMiX's sensitivity to varying sample sizes. On the Houston2013 dataset, we varied the number of training samples per class from 40% to 100% in 20% intervals. As shown in Fig. 10, CoMiX consistently outperforms other methods across all sample ratios, demonstrating its outstanding performance and robustness.

VI. CONCLUSION

This study proposes CoMiX, an encoder-decoder framework for HSI-X semantic segmentation. In the encoder, 3D and 2D

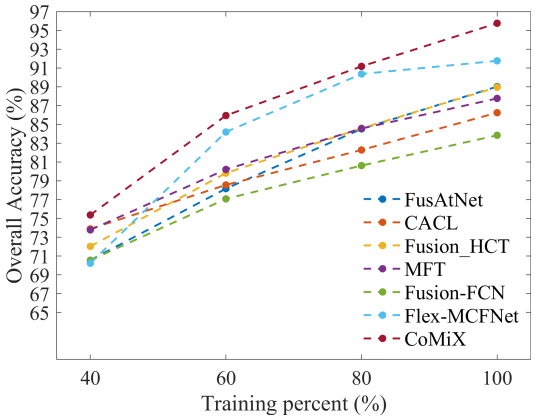


Fig. 10. Classification accuracy of different methods versus the percentage of training samples used on the Houston2013 dataset.

DCN blocks adaptively extract modality-specific features from HSI and X data, respectively. To enhance cross-modality learning, the CMFeX module is developed to recalibrate modality-specific and modality-shared information while dynamically exchanging complementary information across spatial and spectral dimensions. The refined features are then fused by the FFM. By integrating 2D DCN blocks, 3D DCN blocks, CMFeX, and FFM, CoMiX improves cross-modality feature extraction, interaction, and fusion, effectively mitigating challenges of insufficient HSI utilization and inefficient fusion mechanisms. Experimental results demonstrate that CoMiX achieves state-of-the-art performance with favorable efficiency, highlighting its robustness and adaptability across diverse multimodal datasets. Future work will extend CoMiX to arbitrary cross-modal fusion scenarios and incorporate three or more modalities, further enhancing its flexibility and broadening its applicability to a wider range of domains.

ACKNOWLEDGMENT

The authors would like to thank the IEEE GRSS IADF and Hyperspectral Image Analysis Lab, University of Houston for providing the Houston2013 and DFC2018 datasets.

REFERENCES

[1] Y. Su, L. Gao, M. Jiang, A. Plaza, X. Sun, and B. Zhang, "NSCKL: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification," *IEEE Trans. Cybern.*, pp. 1–14, 2022.

[2] X. Zhang, Y. Su, L. Gao, L. Bruzzone, X. Gu, and Q. Tian, "A lightweight transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[3] G. Sun *et al.*, "Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images," *Int.J. Appl. Earth Obs. Geoinf.*, vol. 91, p. 102157, Sep. 2020.

[4] Alamús *et al.*, "Ground-based hyperspectral analysis of the urban nightscape," *ISPRS J. Photogramm. Remote Sens.*, vol. 124, pp. 16–26, Feb. 2017.

[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Adv. Neural Inf. Process. Syst.*, Dec. 2021.

[6] W. Ma *et al.*, "A multi-scale progressive collaborative attention network for remote sensing fusion classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3897–3911, 2023.

[7] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for RGB-T semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2023.

[8] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, 2017.

[9] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, 2017.

[10] L. Lam and S. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Trans Syst Man Cybern A*, vol. 27, no. 5, pp. 553–568, 1997.

[11] P. Ghamisi, R. Souza, L. Rittner, J. A. Benediktsson, R. Lotufo, and X. X. Zhu, "Extinction profiles: a novel approach for the analysis of remote sensing data," in *IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 5122–5125, 2016.

[12] B. Chen, B. Huang, and B. Xu, "Multi-source remotely sensed data fusion for improving land cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 124, pp. 27–39, 2017.

[13] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, 2017.

[14] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, 2023.

[15] Y. Xu, B. Du, and L. Zhang, "Robust self-ensembling network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3780–3793, 2024.

[16] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv:2104.05704*, 2022.

[17] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, 2017.

[18] 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.

[19] 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, 2021.

[20] X. Zhang, J. Yan, J. Tian, W. Li, X. Gu, and Q. Tian, "Objective evaluation-based efficient learning framework for hyperspectral image classification," *GISci. Remote Sens.*, vol. 60, no. 1, p. 2225273, 2023.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, pp. 7132–7141, 2018.

[22] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE CVPR*, pp. 3141–3149, 2019.

[23] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021.

[24] X. Zhang, N. Yokoya, X. Gu, Q. Tian, and L. Bruzzone, "Local-to-global cross-modal attention-aware fusion for hsi-x semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[25] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 10012–10022, October 2021.

[26] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE CVPR*, pp. 14420–14430, 2023.

[27] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 416–425, 2020.

[28] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[29] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE CVPR*, pp. 11953–11965, 2022.

[30] X. Ding *et al.*, "UniRepLKNet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition," *arXiv*, vol. abs/2311.15599, 2023.

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

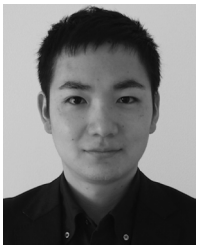
[32] Z. Wan *et al.*, "Sigma: Siamese mamba network for multi-modal semantic segmentation," *arXiv preprint arXiv:2404.04256*, 2024.

[33] Z. Li, H. Pan, K. Zhang, Y. Wang, and F. Yu, "MambaDFuse: A mamba-based dual-phase model for multi-modality image fusion," *arXiv preprint arXiv:2404.08406*, 2024.

[34] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, vol. LNCS 15008, Springer Nature Switzerland, October 2024.
[35] W. Wang *et al.*, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," *Proc. IEEE CVPR*, pp. 14408–14419, 2022.
[36] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE CVPR*, pp. 9300–9308, 2019.
[37] Y. Xiong *et al.*, "Efficient Deformable ConvNets: Rethinking dynamic and sparse operator for vision applications," *ArXiv*, vol. abs/2401.06197, 2024.
[38] 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.
[39] 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, pp. 1–20, 2023.
[40] J. Wang, M. Zhang, W. Li, and R. Tao, "A multistage information complementary fusion network based on flexible-mixup for hsi-x image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2023.
[41] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, 2019.
[42] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, 2023.
[43] Y. Zhang, Y. Peng, B. Tu, and Y. Liu, "Local information interaction transformer for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1130–1143, 2023.
[44] 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 Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[45] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, "Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 385–404, 2023.
[46] J. Li, Y. Ma, R. Song, B. Xi, D. Hong, and Q. Du, "A triplet semisupervised deep network for fusion classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[47] 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, 2022.
[48] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[49] S. K. Roy, A. Sukul, A. Jamali, J. M. Haut, and P. Ghamisi, "Cross hyperspectral and lidar attention transformer: An extended self-attention for land use and land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.
[50] 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.
[51] D. Jia *et al.*, "GeminiFusion: Efficient pixel-wise multimodal fusion for vision transformer," *arXiv preprint arXiv:2406.01210*, 2024.
[52] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan, "Efficient multimodal semantic segmentation via dual-prompt learning," *arXiv preprint arXiv:2312.00360*, 2023.
[53] J. Zhang *et al.*, "Delivering arbitrary-modal semantic segmentation," in *Proc. IEEE CVPR*, pp. 1136–1147, 2023.
[54] J. Guo *et al.*, "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE CVPR*, pp. 12175–12185, June 2022.
[55] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2024.
[56] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE ICCV*, pp. 764–773, 2017.
[57] Y. Liu *et al.*, "VMamba: Visual state space model," in *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 103031–103063, Curran Associates, Inc., 2024.
[58] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv*, vol. abs/1607.06450, Jul. 2016.
[59] R. Azad *et al.*, "Beyond self-attention: Deformable large kernel attention for medical image segmentation," *2024 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1276–1286, 2023.
[60] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "UNETR++: Delving into efficient and accurate 3D medical image segmentation," *IEEE Trans. Med. Imaging*, pp. 1–1, 2024.
[61] C. Debes *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, 2014.
[62] D. Hong, J. Hu, J. Yao, J. Chanussot, and 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, 2021.
[63] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
[64] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE CVPR*, pp. 19529–19539, June 2023.

Xuming Zhang received the B.S. and M.S. degrees from the China University of Petroleum (East China), Qingdao, China, in 2018 and 2021, respectively, and the Ph.D. degree from Nanjing University, Nanjing, China, in 2025. From 2023 to 2025, she was a Visiting Ph.D. Student with the Department of Information Engineering and Computer Science, University of Trento, Italy. She is currently a Lecturer with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China.

Her research interests include hyperspectral image processing, multisensor data fusion, high-resolution remote sensing processing, deep learning and its applications.



Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees from the Department of Aeronautics and Astronautics, the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He is currently an Associate Professor at the University of Tokyo and a Team Leader at the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, where he leads the Geoinformatics Team. He was an Assistant Professor at the University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow, working at the German Aerospace Center (DLR), Oberpfaffenhofen, Germany and Technical University of Munich (TUM), Munich, Germany. His research focuses on visual information processing of large-scale real-world scenes.

Dr. Yokoya won the First Place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was the Chair from 2019 to 2021, a Co-Chair of the IEEE GRSS IADF TC from 2017 to 2019, and also the Secretary of the IEEE GRSS All Japan Joint Chapter from 2018 to 2021. He was an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) from 2018 to 2021. He has been an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2021, as well as the ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING since 2024. He has been designated a Clarivate Highly Cited Researcher since 2022.

Xingfa Gu received the B.S. degree in Aerial Photogrammetry from Wuhan University, Wuhan, China, in 1982, and the M.S. and Ph.D. degrees in Remote Sensing Physics from Paris Diderot University – Paris VII, Paris, France, in 1988 and 1991, respectively.

He is a Professor at the Institute of Aerospace Remote Sensing Innovations, Guangzhou University. He is also a Professor and the director of National Engineering Research Center for Satellite Remote Sensing Applications (NECRSA), Aerospace Information Research Institute (AIR), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include quantitative remote sensing, data and information engineering science of aerospace remote sensing, and the design of remote sensing processing systems.

Dr. Gu is an Academician of the International Academy of Astronautics (IAA), an Academician of the International Eurasian Academy of Sciences (IEAS), and a Fellow of the International Society for Optical Engineering (SPIE). He serves as the Chief Designer of the Application System for China's Major National Science and Technology Project "High-Resolution Earth Observation System," the Head of the Requirements and Applications Group for the *National Medium- and Long-Term Development Plan for Space Infrastructure*, and the Chief Scientist of the National Basic Research Program (973 Program) project "Comprehensive Multi-Scale Aerosol Observation and Spatiotemporal Distribution." He is also the Deputy Leader of the Expert Group for the *Key Research and Development Program on Earth Observation and Navigation* of the Ministry of Science and Technology. In addition, he is a member of the Expert Group for the formulation of the GEO 2016–2025 Decadal Implementation Plan, Chairman of the China National Committee for Remote Sensing, Deputy Secretary-General of the Asian Association on Remote Sensing (AARS), and Co-Chair of the Asia–Oceania Group on Earth Observation (AOGEO).

Qingjiu Tian received the B.S. degree in infrared from Shandong University, Jinan, China, in 1987, the M.S. degree in cartography and remote sensing from the Chinese Academy of Sciences, Beijing, China, in 1996, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2003.

He is currently a Professor with the International Institute for Earth System Sciences, Nanjing University. His research interests include ground object detection and the retrieval of parameters by multi/hyperspectral remote sensing.

Lorenzo Bruzzone (Fellow, IEEE) received the M.S. degree in electronic engineering (summa cumlaude) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively. He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications.

Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory (https://rslab.disi.unitn.it/) in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning and pattern recognition.

He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is currently the Principal Investigator of the *Radar for icy Moon exploration (RIME)* instrument in the framework of the *JUpiter ICy moons Explorer (JUICE)* mission of the European Space Agency (ESA) and the Science Lead for the *High Resolution Land Cover* project in the framework of the Climate Change Initiative of ESA. He is the author (or coauthor) of more than 360 scientific publications in referred international journals, more than 350 papers in conference proceedings, and 22 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. His papers are highly cited, as proven from the total number of citations (more than 46000) and the value of the h-index (103) (source: Google Scholar). He was invited as keynote speaker in more than 40 international conferences and workshops. Since 2009 he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS), where since 2019 he is Vice-President for Professional Activities. Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since that he was recipient of many international and national honors and awards, including the recent *IEEE GRSS 2015 Outstanding Service Award*, the 2017 and 2018 *IEEE IGARSS Symposium Prize Paper Awards* and the 2019 *WHISPER Outstanding Paper Award*. Dr. Bruzzone was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE for which he has been Editor-in-Chief between 2013-2017.Currently he is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has been *Distinguished Speaker* of the IEEE Geoscience and Remote Sensing Society between 2012-2016.