Cross Hyperspectral and LiDAR Attention Transformer: An Extended Self-Attention for Land Use and Land Cover Classification

Swalpa Kumar Roy, *Senior Member, IEEE,* Atri Sukul, Ali Jamali,
Juan M. Haut, *Senior Member, IEEE,* Pedram Ghamisi, *Senior Member, IEEE*

***Abstract*—The successes of attention-driven deep models like the Vision Transformer (ViT) have sparked interest in cross-domain exploration. However, current transformer-based techniques in remote sensing primarily focus on single-modal data, limiting their potential to exploit the growing array of multimodal Earth observation data fully. Enhancing these models for multimodal integration is crucial for comprehensive remote sensing applications. To achieve this, we extend the traditional self-attention mechanism by introducing Cross Hyperspectral and LiDAR (Cross-HL) attention. We present a novel multimodal deep learning framework that effectively fuses remote sensing (RS) data, intending to improve land use and land cover (LULC) recognition. To enhance the accurate exchange of information across different modalities, we fuse their patch projections using the Cross-HL self-attention module. In this process, LiDAR patch tokens serve as queries (Q), while keys (K) and values (V) are derived from HS patch tokens. To demonstrate the superiority of Cross-HL in the proposed multimodal deep learning framework, we conducted extensive experiments on three multimodal RS benchmark datasets: Houston, Trento, and MUUFL. These datasets contain hyperspectral and light detection and ranging (LiDAR) data. The source code for Cross-HL will be made available publicly at https://github.com/AtriSukul1508/Cross-HL.**

***Index Terms*—Deep Learning, Convolutional Neural Networks, Vision Transformers, Hyperspectral Image Classification.**

I. Introduction

The significance and influence of remote sensing (RS) applications have been steadily growing for several decades Hyperspectral (HS) data has proven to be a valuable category of remotely sensed signals in various disciplines [1], including soil research [2], geology [3], water resource management [4], and vegetation monitoring [5]. HS data possesses unique high dimensionality, a significant relationship between neighboring bands, and an extensively nonlinear structure of data, enabling them to accurately represent the precise spectra emitted by different substances in an area of interest [1]. This facilitates improved recognition and distinction of underlying objects, particularly those sharing similar characteristics in single- and multiple-band remote sensing imagery (such as panchromatic, RGB, and MS) [6]. Consequently, the abundance of spectral and spatial data present in HS data has significantly enhanced the perception capability for Earth observation, making HS remote sensing technology vital in fields like precision agriculture [7], pollution monitoring [8], space exploration [9], and military applications [10]. HS data, however, may encounter challenges when attempting to effectively differentiate between complex land covers that yield similar spectral responses [11].

The extraction of multimodal characteristics from multiple sources and efficiently combining these heterogeneous characteristics for integrated land-use and land-cover (LULC) mapping is found to be helpful for accurate classification and semantic segmentation of complex areas [11]. Various fusion techniques have been introduced to integrate complementary information from diverse modalities, resulting in improved classification accuracy and better discrimination of complex land covers [12], particularly in distinguishing intricate objects with high spectral similarity such as wetlands [13]. Specifically, the integration of HS Light Detection and Ranging (LiDAR) and HS data can significantly enhance classification performance. This is because LiDAR data, aided by photoelectric detection systems, contains accurate information on elevation [14], [15].

Deep learning approaches have emerged and been extensively deployed in various vision tasks due to their inherent and automated feature engineering capabilities [16]–[18]. Numerous vision projects have demonstrated that DL techniques offer distinct advantages over conventional methods, as they can learn meaningful features directly from the raw data [19], [20]. Specifically, due to their distinct receptive field architecture and parameter connection of the convolution kernel, convolutional neural networks (CNNs) have risen as highly reliable tools for analyzing satellite imagery [21].

LiDAR data can offer valuable elevation information that

This research was funded by the Institute of Advanced Research in Artificial Intelligence (IARAI) and Consejería de Economía, Ciencia y Agenda Digital of the Junta de Extremadura and the European Regional Development Fund (ERDF) of the European Union, under Grant GR21040, and also Science and Engineering Research Board under Grant SRG/2022/001390. (*Corresponding author: Pedram Ghamisi*)

S. K. Roy is with the Department of Computer Science and Engineering, Alipurduar Government Engineering and Management College, West Bengal 736206, India (e-mail: swalpa@agemc.ac.in).

A. Sukul is with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, West Bengal 735102, India (e-mail: as2511@cse.jgec.ac.in).

A. Jamali is with the Department of Geography, Simon Fraser University, 8888 University Dr, Burnaby, BC V5A 1S6, Canada (e-mail: alij@sfu.ca).

J. M. Haut is with Departamento de Tecnología de Computadores y Comunicaciones, Escuela Politécnica, Universidad de Extremadura, Avenida de la Universidad s/n, 10001-Cáceres, España (e-mail: juanmariohaut@unex.es).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and is also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

aids in distinguishing among urban land covers with similar spectral responses in the HS data but varying elevations. For instance, LiDAR can aid HS data in differentiating between buildings and roads, even when they exhibit comparable spectral properties. The integration of LiDAR and HS imagery has undergone extensive study in the literature concerning land-cover classification tasks [22], [23]. The development of effective fusion techniques for HS and LiDAR data represents an intriguing and challenging research area. These techniques harness the spatial and spectral information from HS, along with the height information from LiDAR, to enhance classification performance and address the limitations inherent in each data source. For instance, Fan *et al.* [24] devised a three-module fusion algorithm called MSLAENet, which capitalizes on self-attention, cross-attention, and self-calibrated convolutions. Furthermore, Li *et al.* [25] introduced a fusion model founded on 1-D and 2-D CNNs for spectral and spatial feature extraction, as well as Cascade Net for extracting LiDAR features. Roy *et al.* [22] introduced feature extraction employing a joint CNN and morphological network for effective LULC classification. In addition, Fan *et al.* [26] implemented a graph-based fusion algorithm integrating dimension reduction techniques and geometry and spectral information fusion of the HS and LiDAR data.

Transformers, in contrast to CNNs and RNNs, stand out as some of the most advanced foundational models because of their utilization of self-attention mechanisms, proving highly effective in time series data analyzing and processing [27]. The rising popularity of transformers in computer vision has pushed to the invention of various innovative transformer models in recent years [28], [29]. While transformers excel at aggregating information embedded in spectral signatures, they struggle to define local semantic aspects and underutilize local spatial information. To unlock the full capability of transformers in modeling extensive dependencies alongside CNNs—a task characterized by locally linked features in the combined classification of LiDAR and HS imagery —Ding *et al.* introduced the GLT-Net to interact both global and local feature using transformer network [30]. For the purpose of integrating heterogeneous information from multiple sensors and enhancing classification performance through combined features, Zhao *et al.* introduced a distinctive dual-branch network that integrates a transformer network with hierarchical convolutional layers [31]. Yu *et al.* implemented the capsule vision transformer to leverage extensive cross-context feature dependencies across various contextual scales using LiDAR and multispectral images for land cover mapping [32]. Addressing the challenges of LULC classification, where both HS and LiDAR data play a role in capturing the spatial, spectral, and elevation relationship while fusing features from distinct modalities to avert redundancy, Zhang *et al.* developed the local information interaction transformer network [33]. Meeting the demand for the fusion and classification of HS features with LiDAR data, the multiple attention hierarchically fused network was developed as detailed in [34].

To harness modality-specific data concurrently from spatial to channel domains, as well as from local to global, and considering the intermodality relationship between features and positional embeddings, Yao *et al.* introduced separable convolutional stems, followed by subspace projections and connected layers of position embedding to enhance LULC classification performance [35]. Xue *et al.* utilizes a spatial hierarchical transformer structure to extract hierarchical spatial features from both HS and LiDAR data for effective classification [36]. The incorporation of a cross-attention (CA) feature fusion allows adaptive and dynamic fusion of heterogeneous features from multi-modality data, enhancing contextual awareness which improves the collaborative classification performance. Zhang *et al.* explored the multimodal transformer network (MTNet), an exciting advancement of transformer in the fusion of HS and LiDAR data, offering a promising approach for improving the accuracy and informativeness of land cover analysis [37]. Notably, Roy *et al.* have recently introduced a multimodal fusion transformer (MFT) that extracts features from HS data and integrates them with a CLS token obtained from LiDAR data, yielding substantial improvements in joint classification performance [38].

The consideration of dependency relationships across the spectral dimension and the extraction of hierarchical characteristics from the spatial domain are two areas in which current classification techniques fall short. These heterogeneous characteristics are also crucial for analyzing multi-modality data, especially for higher-dimensional datasets. Due to the potential impact of the Hughes phenomenon and the challenge of imbalances among various features [39]–[41], there is no suitable-unified architecture available for effectively combining these heterogeneous characteristics to achieve interconnected classification across different domains. These aforementioned factors, to some extent, constrain the precision of combined HS and LiDAR data classification.

To address the issues mentioned above, this paper aims to fully leverage multi-sensor RS imagery and enhance LULC accuracy by making better use of the complementary elevation data provided by the LiDAR sensor. To achieve this, we extend the conventional self-attention mechanism, introducing the Cross HS and LiDAR (Cross-HL) attention. Furthermore, we present a novel multimodal deep learning architecture designed to efficiently fuse HS and LiDAR data. The primary contributions can be outlined as follows:

- We introduce a novel Cross-HS and LiDAR (Cross-HL) attention, an extended self-attention module for RS data fusion.
- In the Cross-HL attention module, the query value Q_L is computed by extracting complementary information from LiDAR data, while K and V are calculated from the HS patch token.
- The newly developed ViT-based transformer architecture utilizes the Cross-HL self-attention mechanism, which efficiently fuses information taken from HS patch tokens and LiDAR patch tokens to learn a new CLS token that combines multimodal features.
- On three available HS datasets—Houston, Trento, and the University of Southern Mississippi Gulfpark (MUUFL)—respectively, we implement comprehensive experiments. These tests exhibit the efficacy of the proposed strategy. We carry out ablation tests utilizing HS

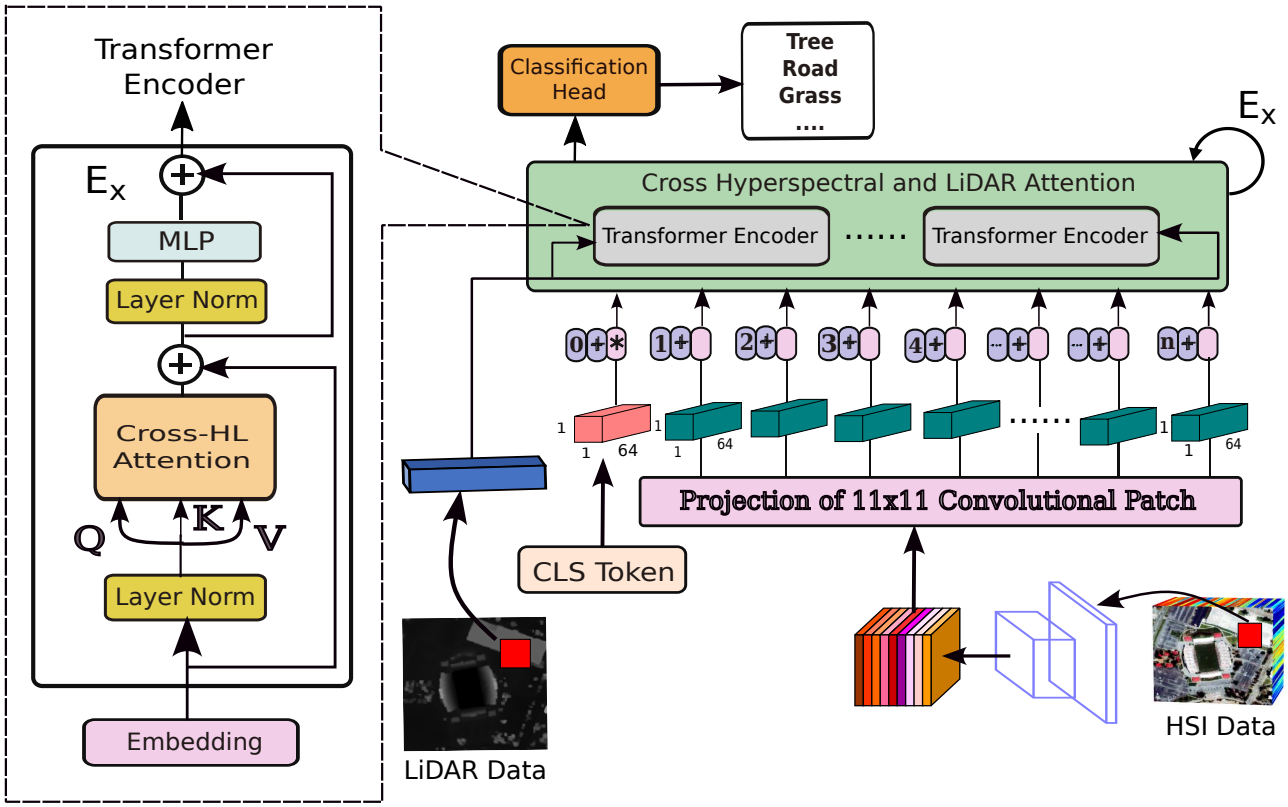


Fig. 1: Overview of the proposed multimodal deep learning framework for HS and LiDAR data fusion.

data and integrate it with additional sources of multimodal data, such as LiDAR data, to demonstrate the benefits of the proposed method.

The remainder of the paper is organized as follows. Section II outlines the motivation behind the proposed model. In Section III, we delve into the components of the proposed Cross-HL attention transformer model. Following this, we carry out extensive experiments, including hyperparameter sensitivity analysis and a discussion of the results obtained, as presented in Section IV. Section V concludes the paper with some concluding remarks.

II. Motivation

Earth Observation (EO) serves as a primary application of Remote Sensing (RS), primarily focusing on monitoring and analyzing changes in land cover. Alterations in land use and cover have significant implications for vital environmental factors, encompassing biodiversity loss, ecosystem degradation, and disruptive climate change. The accuracy of Land Use and Land Cover (LULC) classification is increasingly crucial for safeguarding the environment, advancing scientific research, managing land resources, and serving other critical purposes.

However, classifying LULC using Hyperspectral (HS) data presents various challenges. The substantial number of annotated training samples compels Convolutional Neural Networks (CNNs) to identify high-level representations of features. This approach is hindered by the limitation of high intra-class and low inter-class variability. The task becomes even more demanding for CNNs to derive effective representations in scenarios requiring long-range comprehension, which can be better addressed using the self-attention mechanism employed in Vision Transformers (ViTs). ViTs not only capture long-range structures and patterns but also incorporate spatial and positional information, enhancing accurate LULC mapping.

Moreover, LULC mapping through HS data encounters challenges stemming from sensors with narrow bandwidths. These challenges include data redundancy (i.e., similar spectral reflectance) and the presence of high noise levels in the acquired spectral reflectance data. This redundancy complicates the differentiation between various land cover classes, leading to potential inaccuracies in the mapping process. These challenges collectively underscore the need for refined techniques and methodologies to enhance the precision and reliability of LULC mapping utilizing HS data. For example, to solve these issues, multimodal fusion transformer (MFT) was introduced, which combines HS and LiDAR data using the CLS token obtained from LiDAR data. Although this was a significant breakthrough, the CLS token generated from LiDAR data primarily interacts with HS patch tokens through attention mechanisms, but its direct influence on how those tokens attend to each other is limited. Furthermore, The traditional attention used in transformer can be overly sensitive to noise within HS data, potentially leading to inaccurate feature extraction. Additionally, they may also struggle to fully capture the complex relationships between spectral and spatial

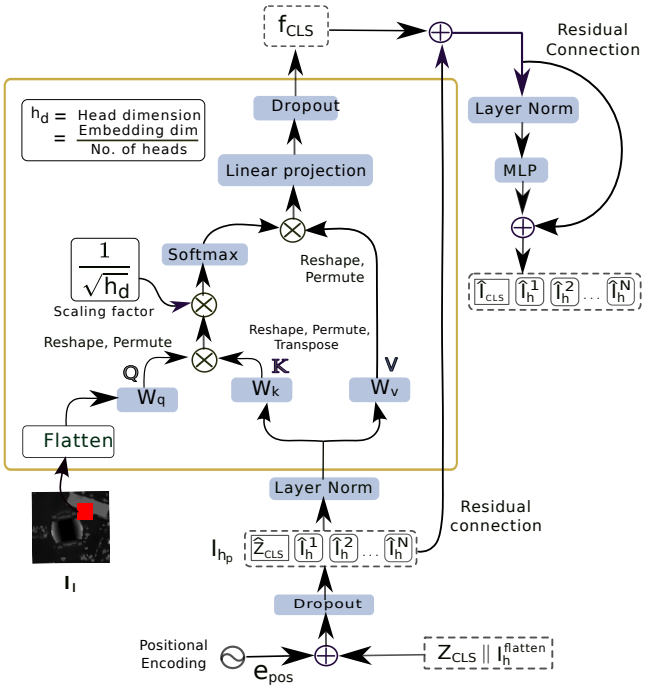


Fig. 2: Cross-Hyperspectral and LiDAR Attention Module. The LiDAR data serves as a query feature vector to interact with the key and value tokens of other HS patches through the attention process.

features. In this context, we introduce the Cross-HL attention module within a multi-head Vision Transformer framework. This novel extended self-attention mechanism addresses the complexities of LULC mapping using HS and LiDAR data.

To tackle the aforementioned issues related to HS data, our proposed attention methodology leverages the spatial, contextual, and structural insights gained from LiDAR data. This integration leads to notable improvements compared to traditional methods, particularly in terms of discrimination capability in forested and urban areas.

III. Proposed Methodology

Fig. 1 illustrates the multimodal deep learning framework benefiting the cross HS and LiDAR (Cross-HL) attention module for accurate LULC mapping by utilizing data from both HS and LiDAR sensors. The following subsections describe the proposed multimodal deep learning framework step by step.

A. HS and LiDAR Patch Extraction

HS data can be shown by $I_H \in \mathcal{R}^{M\times N\times D}$ and corresponding LiDAR rasterized data by $I_L \in \mathcal{R}^{M\times N}$, which represent two spatial dimensions, i.e., the width M and the height N, and the number of spectral channels D. All the pixels under the area of interest are categorized into c different land-cover classes defined by $Y = (y_1, y_2, \dots y_c)$. A set of spectral-spatial cubes and spatial patches are extracted from raw HS and LiDAR data. These cubes and patches contain both spectral and spatial information, which strengthens the feature learning network's ability to discriminate.

For HS data, spectral-spatial cubes $I_H^{(m,n)}$ are extracted from the original HS dataset. Each cube represents a small spatial region within the HS data and contains the spectral information for each pixel in that region. Similar to this, spatial patches $I_L^{(m,n)}$ of the same size are derived from LiDAR data. These patches represent small spatial regions within the LiDAR data and capture the spatial information in that region. The superscripts (m, n) denote the spatial coordinates of the patch, and the subscripts H and L indicate that they belong to the HS or LiDAR datasets, respectively.

Once the spectral-spatial cubes and spatial patches are extracted, they are stacked into matrices I_h and I_l, respectively. The matrix I_h contains all the spectral-spatial cubes extracted from the HS dataset, and the matrix I_l contains all the spatial patches extracted from the LiDAR dataset. Finally, the training and test instances of each class can be represented by selecting the relevant spectral-spatial cubes from I_h and spatial patches from I_l, respectively. This enables improved discriminating and classification performance by enabling the feature learning network to concurrently leverage the spectral and spatial information.

B. Feature Learning via Heterogeneous Convolution

Acknowledging the significance of excavating spectral-spatial information within the context of HS classification tasks, we employ a combination of heterogeneous convolutional layers to process the raw HS data. The feature extraction network is composed of successive layers of Conv3D and HetConv [42] operations. The raw HS data cubes denoted as I_h, start with dimensions of $S \times S \times D$ and undergo an unsqueezing operation to acquire dimensions of $1\times S\times S\times D$. Here, S represents the patch size, which has been set to 11. Following this, the reshaped data cube proceeds through a Conv3D layer, succeeded by batch normalization [43] and ReLU activation layers. This sequence generates intermediate feature maps (I_{int}) with dimensions of $8 \times S \times S \times (D-8)$, which are subsequently reshaped to a suitable configuration of $S \times S \times ((D-8) \times 8)$. These reshaped feature maps then undergo the HetConv layer, batch normalization, and ReLU activation layers are next. The outcome is a collection of feature maps (I_{out}) with dimensions of $S \times S \times 64$. This comprehensive process is designed to effectively harness the rich spectral-spatial information inherently present in HS data, resulting in improved classification outcomes:

$$\mathrm{I_{int}} = ReLU(\phi_{bn}\left(Conv3D(\phi_u\left(I_h\right)\right))) \tag{1a}$$

$$\mathrm{I_{out}} = ReLU(\phi_{bn}\left(HetConv(\phi_r\left(I_{int}\right)\right))) \tag{1b}$$

where $\phi_{bn}(.)$ refers to batch normalization operation, $\phi_u(.)$, $\phi_r(.)$ represent an unsqueeze and a reshape operations, respectively.

C. Cross Hyperspectral and LiDAR Attention module

Our objective is to effectively fuse information from HS cubes (I_h) and LiDAR patches (I_l) to enhance classification performance. Within our proposed multimodal deep learning framework, we have introduced the Cross-HL attention module as a replacement for the commonly used conventional

multi-head self-attention (MSHA), as depicted in Fig. 2. The prevalent approach in current state-of-the-art algorithms [22], [29], [38], [44] involves combining or concatenating a pixel's elevation information, along with spectral and spatial features extracted from diverse modalities, for use within attention modules. Essentially, the extracted LiDAR and HS information are combined and input into attention modules without specific priority or emphasis. However, employing such methods can make LULC mapping more challenging and potentially lead to decreased classification accuracy due to spectral redundancy with HS data. To address the issues posed by HS data, our Cross-HL attention module takes a different approach. Complementary information from LiDAR imagery is utilized to calculate the query Q_L, while spectral and spatial information from HS data is employed to determine the values V and keys K for the query Q_L. The elevation information, including object height and width details, for a specific pixel is significantly influenced by its position or location within the scene. For instance, a pixel situated within water features like rivers will exhibit lower elevation compared to another pixel near or within urban structures like buildings. By incorporating elevation (height) information derived from the LiDAR sensor as queries Q_L—reflecting the positional relevance—into the developed Cross-HL attention module, we can generate more precise keys and values for a given query. This approach allows for the identification of the most pertinent features, facilitating the establishment of long-range relationships between a pixel's elevation (height and positional information) and its spectral and positional data from the HS sensor.

The query vector appears as a fundamental component in the context of self-attention, playing a significant role in learning important positionally-aware features. To get the most relevant features from an input image, a query vector that acts as an inquisitive question, which helps the attention mechanism to focus selectively on certain areas of the image. By comparing the query vector to the key vectors associated with each patch embedding, the mechanism can determine which patches are most similar to the query and should be given more weight in the final output. In this way, by considering complementary information of HS data extracted from LiDAR data to obtain the query Q_L values, the developed attention module can overcome or reduce the inherent issues associated with HS sensors (e.g., data redundancy and image noise). In other words, the spectral and spatial information obtained from the HS sensor is not utilized in the calculation of query Q_L of the developed Cross-HL attention module. Thus, we significantly emphasize the positional information (i.e., pixel height) of object pixels as compared to their spectral and spatial information during the attention map calculation.

In the proposed Cross-HL attention module, as illustrated in Fig. 2, the extracted HS feature maps (I_{out}) are first flattened and transposed, similarly, the LiDAR patch is also flattened as shown in Eq. (2b).

$$I_h^{flatten} = \phi_t(Flatten\,(I_{out})) \tag{2a}$$

$$I_{l_p} = Flatten\,(I_l) \tag{2b}$$

where $I_h^{flatten} \in \mathcal{R}^{N\times 64}$, $I_{l_p} = [I_l^1, I_l^2, \dots I_l^N] \in \mathcal{R}^{N\times 1}$, $N\left(=S^2\right)$ is the number of flatten patch embeddings (same for both modalities i.e. HS and LiDAR) and $\phi_t(.)$ is a transpose function.

Then, the patch embeddings are concatenated with an external learnable CLS (z_{CLS}) embedding and performed element-wise ($\oplus$) addition with position embedding (e_{pos}) to preserve positional information as shown in Eq. (3b).

$$I_{h_p} = \phi_{dp}\left([z_{CLS} \oplus e_{pos} || I_h^{flatten} \oplus e_{pos}]\right) \tag{3a}$$

$$= [\hat{z}_{CLS} || \hat{I}_h^{flatten}] \tag{3b}$$

where $z_{CLS} \in \mathcal{R}^{1\times 64}$, $I_{h_p} = [\hat{z}_{CLS}, \hat{I}_h^1, \hat{I}_h^2, \dots \hat{I}_h^N] \in \mathcal{R}^{(N+1)\times 64}$ and ϕ_{dp} represents a dropout layer with a dropout rate of 0.1. The learnable CLS (z_{CLS}) token utilizes the HS patch embeddings to exchange information among themselves, thereby deriving the comprehensive abstract representation of the entire HS patch.

The resultant vectors of HS patch embeddings (I_{h_p}) serve as values V and keys K for the Cross-HL attention module. Conversely, the flattened LiDAR image patches (I_{l_p}) extracted from the identical spatial region as the HS patches serve as queries Q_L. Mathematically, these values can be expressed as follows:

$$\begin{aligned} Q_L &= W_q I_{l_p}, \\ K &= W_k I_{h_p}, \\ V &= W_v I_{h_p} \end{aligned} \tag{4}$$

The entire process of the attention module takes place within the transformer encoder blocks. Each encoder block comprises two sub-blocks: a multi-head self-attention block and an MLP block. A residual connection is employed around each sub-block. The attention block consists of a layer normalization operation and a self-attention layer. Within the attention block, the output maps of the queries and keys are reshaped and multiplied. Subsequently, the resulting feature maps are divided by the head dimension h_d and fed into a $softmax$ layer. The resulting feature map is then multiplied with the reshaped values V. This resultant feature map (A) is passed through a linear layer followed by a dropout layer to obtain the feature map of the proposed Cross-HL attention module, as depicted in Fig. 2. Mathematically, the aforementioned process can be summarized by the following equations:

$$\mathtt{A}\,(Q_L, K, V) = softmax\left(\frac{Q_L K^T}{\sqrt{h_d}}\right) V \tag{5a}$$

$$CrossHL\left(I_{l_p}, \phi_{ln}\left(I_{h_p}\right)\right) = \phi_{dp}\left(WA\right) \tag{5b}$$

where $\phi_{ln}(.)$ refers to LN operation, $W \in \mathcal{R}^{64\times 64}$ is the weights of linear projection layer and h_d is head dimension (= embedding dimension/number of heads), which is chosen as 8.

The output of the Cross-HL attention module, denoted as f_{CLS}, for a given HS patch embedding $I_{h_p}^{(j-1)}$, generates f_j when combined with the corresponding HS patch embedding $I_{h_p}^{(j-1)}$, as demonstrated in equations (6a) and (6b). After applying layer normalization, f_j is passed through an MLP block. This block consists of two layers: a hidden layer and

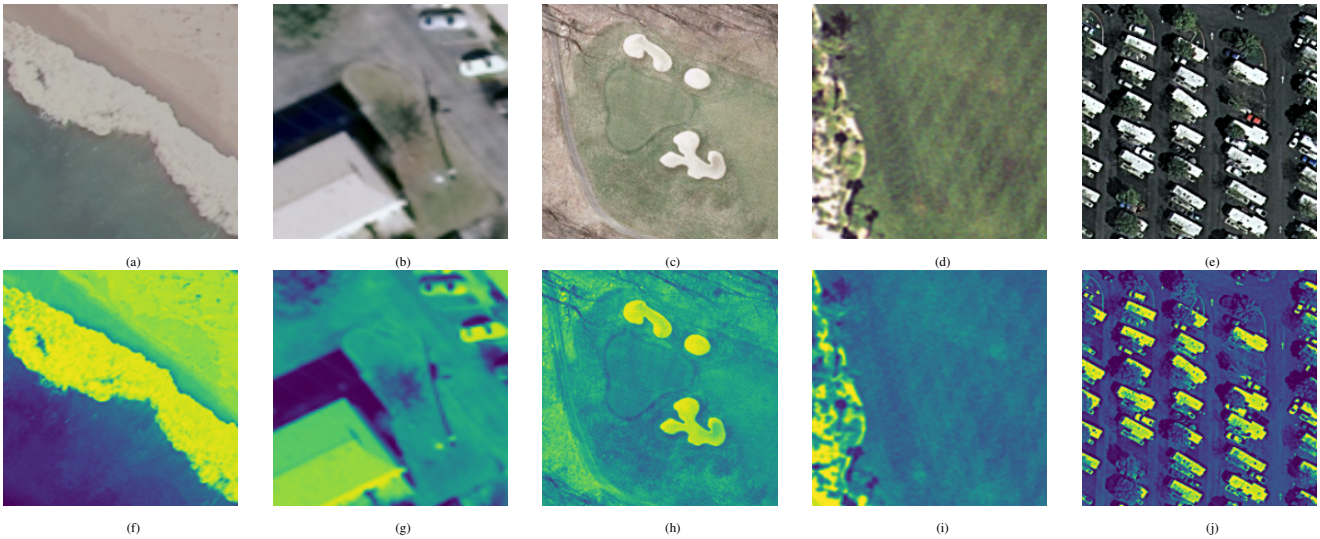


Fig. 3: Visualizing the Cross-HL attention module, as demonstrated in the bottom row (f)–(j), applied to randomly selected images from the UC Merced Land Use Dataset, presented in the top row (a)–(e).(http://weegee.vision.ucmerced.edu/datasets/landuse).

an output layer, with dimensions 512 and 64 respectively, both utilizing the GELU non-linearity. The output of this MLP block is then integrated with the input using a residual connection to produce $I_{h_p}^{j}$, as illustrated in Equation (6c).

$$f_{CLS} = CrossHL\Big(I_{l_p}, \phi_{ln}\left(I_{h_p}^{(j-1)}\right)\Big) \tag{6a}$$

$$f_j = f_{CLS} \oplus I_{h_p}^{(j-1)} \tag{6b}$$

$$I_{h_p}^{j} = f_j \oplus MLP\left(\phi_{ln}\left(f_j\right)\right) \tag{6c}$$

The proposed framework is stacked with E transformer encoder blocks and j is the jth transformer encoder ($1 \leq j \leq E$).

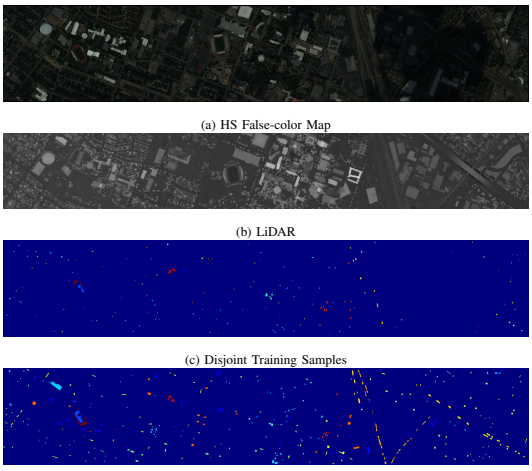
The first element of the final encoder block output $I_{h_p}^{E}$ (i.e. $I_{h_p}^{E}[0]$) is sent to the classification head to obtain the classification score.

$$y_{out} = I_{h_p}^{E}[0] W_{ca} \tag{7}$$

where $y_{out} \in \mathcal{R}^{1\times c}$, $W_{ca} \in \mathcal{R}^{64\times c}$ is the weights of the linear layer used for the implementation of the classification head block, and the number of transformer encoder blocks E is set to 2 in the proposed model. The Cross-HL attention is visually demonstrated in Fig. 3 over some randomly selected example images. This illustration showcases its capacity to emphasize spatial locations of objects while filtering out irrelevant feature information.

IV. Experimental Results and Discussion

In order to assess the performance of the proposed transformer network, we have employed three distinct and challenging datasets. We then compared the experimental outcomes with those of state-of-the-art methods to establish the superiority of our proposed approach. The datasets used in our experiments encompass the University of Houston (UH), Trento, and MUUFL Gulfport (MUUFL) scenes. Below, you will find comprehensive details concerning these experimental datasets.



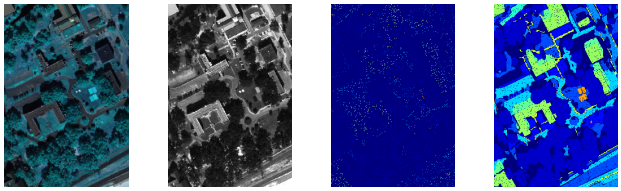
Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	662013	652648		Grass-healthy	198	1053
	Grass-stressed	190	1064		Grass-synthetic	192	505
	Tree	188	1056		Soil	186	1056
	Water	182	143		Residential	196	1072
	Commercial	191	1053		Road	193	1059
	Highway	191	1036		Railway	181	1054
	Parking-lot1	192	1041		Parking-lot2	184	285
	Tennis-court	181	247		Running-track	187	473

Fig. 4: Pictorial view of the University of Houston (UH) scenario: (a) A false-color representation of the HS data for bands 64, 43, and 22, respectively. (b) An 8-bit representation of the LiDAR data, (c) the annotation of the disjoint training samples, and (d) the disjoint test samples. The table displays the number of disjoint training and test instances, as well as land-cover types relevant to each class.

A. HS Datasets

• The **University of Houston** (UH) dataset, published by the IEEE Geoscience and Remote Sensing Society, was collected in June 2012 using the Compact Airborne Spectrographic Imager (CASI) [45]. This dataset comprises HS and LiDAR data. The HS and LiDAR data exhibit a spatial resolution of 340×1905 pixels and encompass 144 spectral bands. The spatial resolution is 2.5 meters per pixel (MPP), signifying

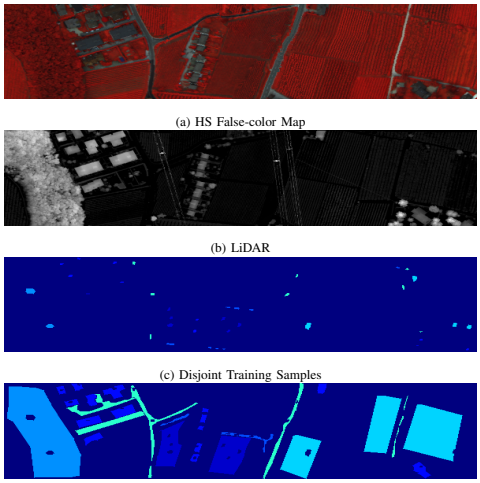
that each pixel represents an area of 2.5 meters by 2.5 meters on the ground. The wavelength range covered by the dataset extends from 0.38 to $1.05\mu\ m$. Further details concerning the number of instances per class, and the distribution of disjoint training and test samples as geographical maps, are provided in Fig. 4.



(a) HS False-color Map (b) LiDAR (c) Training Samples (d) Test Samples

Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	68816	20497		Trees	1162	22084
	Grass-Pure	214	4056		Grass-Groundsurface	344	6538
	Dirt-And-Sand	91	1735		Road-Materials	334	6353
	Water	23	443		Buildings'-Shadow	112	2121
	Buildings	312	5928		Sidewalk	69	1316
	Yellow-Curb	9	174		ClothPanels	14	255

Fig. 5: Pictorial view of the MUUFL scenario: (a) False-color image for the HS data across bands 40, 20, and 10, and (b) an 8-bit image for the LiDAR data. (c) - (d) Annotated training and test samples. The table shows the types of land cover that are distinctive to each class, as well as the percentages of training samples (5%) and test samples (95%), which were chosen at random.



(a) HS False-color Map

(b) LiDAR

(c) Disjoint Training Samples

(d) Disjoint Test Sample

Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	98781	70205		Apples	129	3905
	Buildings	125	2778		Ground	105	374
	Woods	154	8969		Vineyard	184	10317
	Roads	122	3052				

Fig. 6: Pictorial view of the Trento data: (a) False-color representation of the HS data by combining bands 40, 20, and 10, respectively. (b) An 8-bit representation of the LiDAR data. (c) Annotation derived from disjoint training samples. (d) test samples. The table shows the number of disjoint training and test samples and the types of land-cover that are unique to each class, respectively.

• The **MUUFL Gulfport scene**, captured by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor in November 2010 [46], [47], encompasses the University of Southern Mississippi Gulf Park campus in Long Beach, Mississippi. This dataset comprises 325 by 220 pixels with 72 spectral bands. Alongside spectral data, the dataset incorporates LiDAR information in the form of elevation measurements. The LiDAR component consists of two raster data layers. However, due to noise considerations, the first and last eight spectral bands are excluded, resulting in a total of 64 bands available for analysis. Ground truth information within this dataset encompasses 53,687 pixels, classifying into 11 distinct urban land-cover categories. For additional insights into class samples and the distribution of disjoint training and test samples as geographical maps, please refer to Fig. 5.

• The **Trento scene** was acquired utilizing the AISA Eagle sensor across rural regions in the southern part of Trento, Italy. Correspondingly, LiDAR data was collected via the Optech ALTM 3100EA sensor. HS data within the Trento scene comprises 63 spectral bands, spanning a wavelength range of 0.42 to 0.99 $\mu\ m$. The LiDAR dataset in Trento incorporates two raster layers, capturing elevation details. Dimensions of the Trento scene stand at 600×166 pixels, with a spatial resolution of 1 meter per pixel (MPP). The HS data's spectral resolution is 9.2 nm. As with the aforementioned datasets, samples in the Trento scene are segregated into disjoint training and test sets for the six vegetation land-cover classes. For additional specifics regarding class samples, disjoint training and test samples, and geographical distribution, please refer to Fig. 6.

B. Experimental Setting

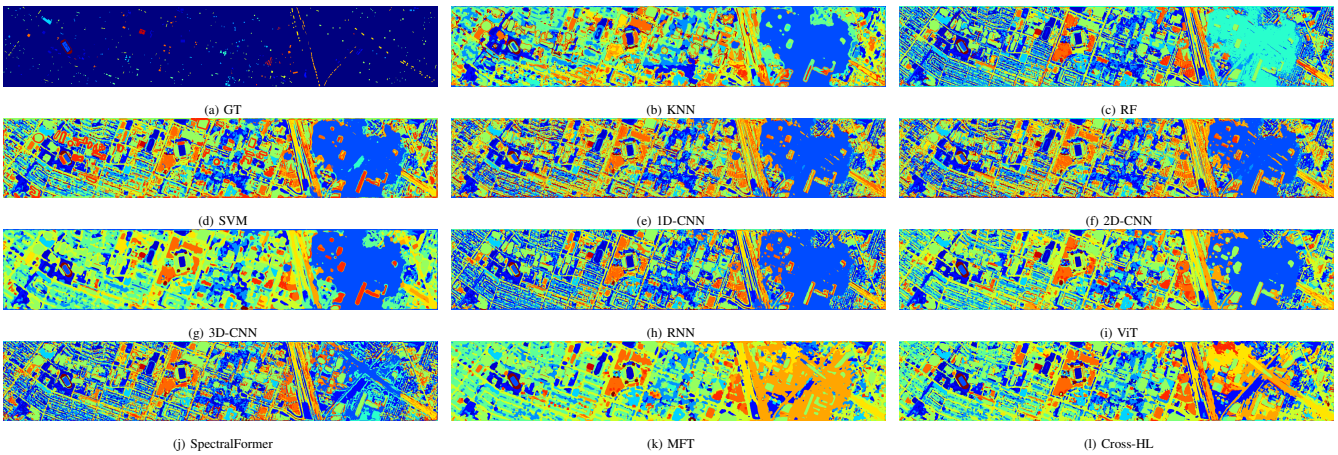
This section outlines the experimental settings employed to assess the proposed model, alongside the comparative methodologies. The accuracies presented in the preceding sections are derived from disjoint training and test samples, irrespective of other works documented in the literature. To facilitate result comparison, average accuracy (AA), overall accuracy (OA), Kappa accuracy (κ), and per-class accuracies are computed across all datasets and compared methods. Overall and average accuracies predominantly concern the percentage of correctly mapped samples. Kappa (κ) accuracy, on the other hand, is derived from statistical testing, providing insights into the performance of classification models relative to random selection. In essence, Kappa (κ) accuracy considers the dataset's class count and the likelihood of random label assignments to sample points. Consequently, it serves as a more robust accuracy measure compared to OA and AA, which might be misleading in cases of imbalanced datasets.

The proposed method is compared with three different categories of the models proposed in the literature and extensively utilized for comparative purposes. The comparative methods include conventional classification models, for instance, K-Nearest Neighbours (KNN) [48], Random Forest (RF) [49], and Support Vector Machine (SVM) [50], classical Convolutional Neural Networks, for instance, 1-D CNN [48], 2-D CNN [51], 3-D CNN [52], and Recurrent Neural Network (RNN) [53] models and finally with vision transformers (ViTs) networks, for instance, vanilla ViT [28], SpectralFormer [54] and Multimodal Fusion Transformer (MFT) [38].

The performance evaluation of the models involved training and testing with batch sizes of 64 and 100, respectively. During training and validation, patches of size $11 \times 11 \times D$ were employed for both HS and LiDAR data. Except for KNN, RF, SVM, and RNN, all models were trained using the Adam

TABLE I: Results in terms of OA, AA, and Kappa (in %) obtained on the University of Houston dataset using HS and LiDAR data.

Class No.	Traditional Learning			Convolutional Neural Networks				Transformer Models			
	KNN	RF	SVM	1D-CNN	2D-CNN	3D-CNN	RNN	ViT	SpectralFormer	MFT	Cross-HL
1	77.30	79.33 ± 00.38	79.96	81.32 ± 00.16	82.62 ± 00.08	82.30 ± 00.29	81.80 ± 00.62	82.59 ± 00.18	82.65 ± 00.24	81.96±0.66	**83.10 ± 00.00**
2	81.58	71.49 ± 01.69	82.89	81.92 ± 00.19	81.83 ± 00.66	78.70 ± 00.98	71.40 ± 05.21	82.33 ± 01.46	83.33 ± 00.29	91.26±5.93	**97.37 ± 04.44**
3	97.82	98.15 ± 00.25	60.79	57.49 ± 00.61	62.90 ± 00.37	96.96 ± 00.47	76.04 ± 14.49	97.43 ± 00.70	75.78 ± 12.48	99.03±0.89	**99.34 ± 00.35**
4	80.59	78.09 ± 00.47	86.65	86.74 ± 00.23	88.57 ± 00.04	80.49 ± 01.57	88.51 ± 02.14	92.93 ± 01.72	91.10 ± 01.29	94.61±3.19	**97.30 ± 02.57**
5	91.86	89.68 ± 00.08	95.36	95.93 ± 00.27	97.19 ± 00.19	98.11 ± 00.54	85.76 ± 04.87	99.84 ± 00.04	98.30 ± 01.12	99.80±0.21	**99.94 ± 00.09**
6	78.32	93.71 ± 00.57	69.23	58.04 ± 00.57	65.03 ± 01.98	73.89 ± 01.84	85.78 ± 02.87	84.15 ± 03.30	89.04 ± 02.70	92.45±5.95	**96.97 ± 01.36**
7	80.97	79.32 ± 01.03	85.26	77.95 ± 00.88	80.22 ± 01.42	81.09 ± 00.75	82.77 ± 01.71	**87.84 ± 01.49**	81.72 ± 01.19	83.47±5.02	80.15 ± 03.15
8	42.45	53.37 ± 02.26	49.38	54.42 ± 02.33	60.81 ± 04.32	44.63 ± 06.52	61.44 ± 07.10	79.93 ± 00.16	67.81 ± 09.00	89.31±3.87	**91.01 ± 02.72**
9	69.22	76.27 ± 00.46	78.28	66.13 ± 00.19	67.74 ± 01.08	74.76 ± 03.84	67.42 ± 07.89	82.94 ± 00.85	74.47 ± 01.89	88.58±4.68	**90.82 ± 01.58**
10	43.44	38.42 ± 01.07	50.48	47.30 ± 02.18	51.74 ± 01.33	37.52 ± 11.23	38.45 ± 03.12	52.93 ± 05.14	56.76 ± 06.31	**59.50±6.23**	58.86 ± 07.47
11	65.84	65.02 ± 00.47	49.34	44.40 ± 02.06	39.91 ± 03.52	40.80 ± 08.50	64.39 ± 06.38	80.99 ± 03.06	59.93 ± 07.67	96.24±1.62	**98.77 ± 00.77**
12	81.94	82.20 ± 00.56	72.05	63.66 ± 08.15	82.20 ± 03.85	66.38 ± 02.36	77.07 ± 05.79	91.07 ± 02.55	70.00 ± 02.98	**93.97±3.17**	91.55 ± 01.78
13	60.35	67.72 ± 00.50	77.19	50.41 ± 01.86	52.40 ± 01.29	68.77 ± 12.16	47.13 ± 03.73	87.84 ± 01.91	66.20 ± 00.44	81.37±1.97	**90.49 ± 02.00**
14	83.81	82.46 ± 00.83	75.30	86.50 ± 00.69	92.44 ± 02.89	92.85 ± 01.16	97.98 ± 00.87	**100.0 ± 00.00**	92.04 ± 04.50	99.84±0.20	99.96 ± 00.13
15	91.33	95.56 ± 00.30	55.18	40.10 ± 03.20	52.08 ± 04.14	96.90 ± 01.41	73.50 ± 09.82	**99.65 ± 00.50**	77.45 ± 13.31	98.75±1.69	98.64 ± 02.01
OA	73.50	73.83 ± 00.31	71.94	68.11 ± 00.29	71.88 ± 00.36	71.41 ± 00.35	72.31 ± 02.44	85.05 ± 00.47	76.87 ± 02.21	88.93±0.76	**89.66 ± 00.64**
AA	75.12	76.72 ± 00.28	71.16	66.15 ± 00.09	70.51 ± 00.27	74.28 ± 00.30	73.30 ± 02.50	86.83 ± 00.53	77.77 ± 02.78	90.01±0.74	**91.25 ± 00.59**
$\kappa(\times 100)$	71.41	71.78 ± 00.33	69.67	65.48 ± 00.31	69.56 ± 00.39	69.12 ± 00.36	70.14 ± 02.62	83.84 ± 00.51	75.03 ± 02.37	87.98±0.82	**88.78 ± 00.70**



(a) GT (b) KNN (c) RF (d) SVM (e) 1D-CNN (f) 2D-CNN (g) 3D-CNN (h) RNN (i) ViT (j) SpectralFormer (k) MFT (l) Cross-HL

Fig. 7: The predicted land cover maps were created for the UH data set.

optimizer [55] with a learning rate of 5×10^{-4} and weight decay of 5×10^{-3}. The RNN used a higher learning rate of 1×10^{-3} without weight decay. The models also utilized a step scheduler with a step size of 50 and gamma of 0.9, trained for 200 epochs. Each experiment was repeated 10 times, and the reported results are averaged with standard deviations. The proposed model was implemented using PyTorch 2.0.1 and Python 3.10.11, with the server configuration of an Intel i9-9940X processor and 128 GB of DDR4 RAM, along with an NVIDIA Titan RTX with 24 GB of DDR4 RAM.

C. Performance Analysis

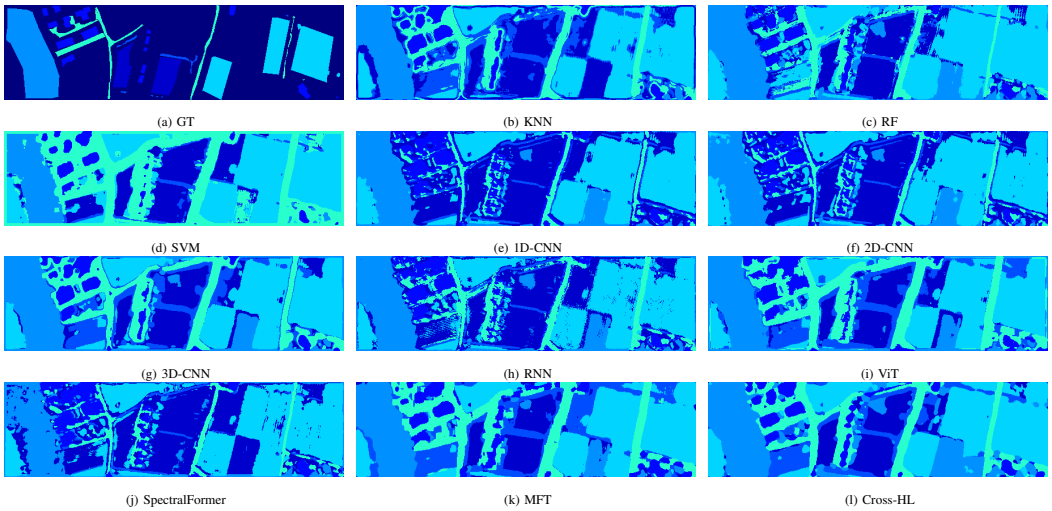
The proposed method and the other comparative methods are evaluated on three widely used datasets: Houston, Trento, and MUUFL. The experimental results are presented in Tables I, II, and III. All the comparative methods are compared with the proposed Cross-HL attention transformer network. Based on the obtained results reported in Tables I, II, and III, the proposed method outperforms all the comparative methods due to the ability of the proposed network to represent discriminative features.

All the comparative methods, along with the Cross-HL attention transformer, are evaluated with a fixed number of disjoint training and test samples—specifically, 5% for training and the remaining 95% for test samples—to evaluate the performance of the model. It is important to note that the samples used for training and test remain the same for all the compared and proposed methods to ensure the validity and soundness of the claims. In other words, the allocation of training and test samples is conducted only once, and all the competing methods are executed concurrently with the proposed method. Changing these samples for each method could potentially introduce bias in the results, as some samples might appear easier to classify, especially those near the decision boundary, which cannot be guaranteed to be consistently selected for every method due to the random nature of sample selection.

Table I presents the classification performance of the proposed method and other compared methods on the Houston dataset, considering both HS images and LiDAR data. The table showcases quantitative metrics, including Overall Accuracies (OAs), Average Accuracies (AAs), Cohen's Kappa (κ), and class-wise accuracy. The proposed Cross-HL transformer network surpasses other methods in 12 out of 15 class-wise classification accuracies over the Houston dataset. Among conventional classifiers, Random Forest (RF) performs the best, yielding mean OAs, AAs, and κ values of 73.83%, 76.72%, and 71.78%, respectively, with corresponding standard deviations of 0.31%, 0.28%, and 0.33%. Despite RNN displaying improved results compared to other classical convolutional networks, RF maintains its lead over RNN. However, transformer-based methods outperform RF. The proposed Cross-HL attention transformer excels over all compared mod-

TABLE II: Results in terms of OA, AA, and Kappa (in %) obtained on the Trento dataset using HS and LiDAR data.

Class No.	Traditional Learning			Convolutional Neural Networks				Transformer Models			
	KNN	RF	SVM	1D-CNN	2D-CNN	3D-CNN	RNN	ViT	SpectralFormer	MFT	Cross-HL
1	87.94	83.73 ± 00.06	97.44	97.00 ± 00.50	96.98 ± 00.21	92.95 ± 00.10	91.75 ± 04.30	90.87 ± 00.77	96.76 ± 01.71	93.70±1.19	**99.32 ± 00.31**
2	95.79	96.30 ± 00.06	98.12	96.51 ± 01.70	97.56 ± 00.14	98.09 ± 00.23	**99.47 ± 00.37**	99.32 ± 00.77	97.25 ± 00.66	93.63±2.01	95.32 ± 01.42
3	81.28	70.94 ± 01.55	56.15	42.34 ± 06.33	55.35 ± 00.00	93.85 ± 01.09	79.23 ± 16.47	92.69 ± 01.53	58.47 ± 11.54	89.07±4.62	**97.62 ± 01.17**
4	96.25	99.73 ± 00.07	97.53	99.77 ± 00.05	99.66 ± 00.03	99.32 ± 00.05	99.58 ± 00.42	**100.0 ± 00.00**	99.24 ± 00.21	99.53±0.42	99.89 ± 00.10
5	95.29	95.35 ± 00.25	98.13	99.27 ± 00.09	99.56 ± 00.07	98.74 ± 00.04	98.39 ± 00.65	97.77 ± 00.86	93.52 ± 01.75	99.71±0.26	**99.97 ± 00.04**
6	83.85	72.63 ± 00.90	78.96	76.91 ± 03.62	76.91 ± 00.15	88.15 ± 00.20	85.86 ± 02.89	86.72 ± 02.02	73.39 ± 06.78	88.75±2.56	**94.17 ± 01.36**
OA	93.29	92.57 ± 00.07	95.33	95.81 ± 00.13	96.14 ± 00.03	96.93 ± 00.03	96.43 ± 00.79	96.47 ± 00.49	93.51 ± 01.27	97.01±0.35	**98.69 ± 00.19**
AA	90.07	86.45 ± 00.32	87.72	85.30 ± 00.72	87.67 ± 00.04	95.18 ± 00.18	92.38 ± 03.50	94.56 ± 00.57	86.44 ± 02.96	94.06±0.94	**97.47 ± 00.30**
$\kappa(\times 100)$	91.11	90.11 ± 00.09	93.76	94.39 ± 00.17	94.83 ± 00.04	95.89 ± 00.04	95.21 ± 01.06	95.28 ± 00.65	91.36 ± 01.67	95.99±0.47	**98.25 ± 00.26**



(a) GT (b) KNN (c) RF (d) SVM (e) 1D-CNN (f) 2D-CNN (g) 3D-CNN (h) RNN (i) ViT (j) SpectralFormer (k) MFT (l) Cross-HL

Fig. 8: The predicted land cover maps were created for the Trento data set.

TABLE III: Results in terms of OA, AA and Kappa (in %) values obtained on the MUUFL dataset using HS and LiDAR data.

Class No.	Traditional Learning			Convolutional Neural Networks				Transformer Models			
	KNN	RF	SVM	1D-CNN	2D-CNN	3D-CNN	RNN	ViT	SpectralFormer	MFT	Cross-HL
1	92.12	95.42 ± 00.09	96.63	95.05 ± 00.22	95.79 ± 00.11	95.10 ± 00.19	95.84 ± 00.14	97.85 ± 00.29	97.30 ± 00.83	97.40±0.73	**98.33 ± 00.21**
2	51.85	74.03 ± 00.11	59.25	70.35 ± 03.30	72.76 ± 00.58	63.72 ± 03.18	81.93 ± 02.11	76.06 ± 02.40	69.35 ± 05.16	**90.21±2.45**	89.65 ± 01.18
3	69.35	75.81 ± 00.38	81.46	75.80 ± 02.09	78.92 ± 00.52	69.94 ± 03.42	80.47 ± 02.13	87.58 ± 03.46	78.48 ± 03.41	89.15±2.28	**89.86 ± 01.37**
4	57.00	68.59 ± 00.77	73.54	78.60 ± 02.77	83.59 ± 00.99	63.90 ± 01.84	87.01 ± 01.46	92.05 ± 02.31	82.63 ± 03.68	92.86±3.39	**96.28 ± 01.01**
5	83.87	88.17 ± 00.18	83.79	78.31 ± 01.48	78.29 ± 01.12	79.48 ± 01.43	90.65 ± 00.65	94.73 ± 00.60	87.91 ± 02.97	94.49±0.69	**95.64 ± 00.65**
6	19.19	77.28 ± 00.93	15.35	46.35 ± 02.49	50.34 ± 02.13	02.86 ± 03.00	54.25 ± 03.14	82.02 ± 01.13	58.77 ± 02.76	82.44±5.28	**86.55 ± 02.97**
7	44.60	64.83 ± 00.97	77.04	78.31 ± 00.20	79.70 ± 00.26	47.96 ± 01.00	81.24 ± 01.32	87.11 ± 01.54	85.87 ± 00.62	91.32±1.53	**92.73 ± 01.39**
8	76.97	93.29 ± 00.27	86.94	66.72 ± 01.17	71.95 ± 01.10	70.47 ± 01.25	88.39 ± 01.50	97.60 ± 00.16	95.60 ± 01.26	96.56±0.68	**98.02 ± 00.42**
9	09.95	19.15 ± 01.37	21.28	40.15 ± 02.96	43.92 ± 01.24	06.28 ± 04.91	60.54 ± 04.40	57.83 ± 04.45	53.52 ± 04.32	53.89±6.22	**68.81 ± 02.92**
10	00.00	04.41 ± 00.72	00.00	09.20 ± 01.24	12.45 ± 00.27	00.00 ± 00.00	26.44 ± 02.82	**31.99 ± 08.86**	08.43 ± 02.22	1.26±1.78	26.12 ± 08.56
11	64.45	71.88 ± 00.84	62.89	25.65 ± 02.89	26.82 ± 02.60	66.93 ± 01.76	**87.50 ± 02.92**	58.72 ± 03.85	35.29 ± 06.00	72.30±5.35	71.53 ± 03.22
OA	76.83	85.32 ± 00.09	84.24	81.50 ± 00.03	83.40 ± 00.04	77.99 ± 00.06	88.79 ± 00.45	92.15 ± 00.19	88.25 ± 00.56	93.20±0.17	**94.49 ± 00.12**
AA	51.76	66.62 ± 00.16	59.83	60.41 ± 00.48	63.14 ± 00.21	51.51 ± 00.40	75.84 ± 00.62	78.50 ± 01.28	68.47 ± 01.44	78.35±1.05	**82.81 ± 01.01**
$\kappa(\times 100)$	68.92	80.39 ± 00.12	78.80	75.43 ± 00.07	77.94 ± 00.06	70.31 ± 00.03	85.18 ± 00.60	89.56 ± 00.27	84.40 ± 00.77	91.00±0.22	**92.71 ± 00.16**

TABLE IV: OA, AA and Kappa values for different patch sizes.

Patch Size	Metric	HOUSTON	TRENTO	MUUFL
3×3	OA	89.59±0.2694	97.47±0.3094	90.66±0.3005
	AA	91.31±0.2583	96.69±0.6207	80.01±1.2892
	Kappa(x100)	88.70±0.2874	96.61±0.4149	87.62±0.4098
5×5	OA	89.59±0.8897	98.40±0.153	92.44±0.1683
	AA	91.33±0.819	97.80±0.2945	81.54±1.2105
	Kappa(x100)	88.71±0.9623	97.85±0.205	89.99±0.2322
7×7	OA	89.33±0.6132	98.66±0.1021	93.50±0.1877
	AA	91.14±0.4624	97.91±0.2363	82.03±0.8805
	Kappa(x100)	88.43±0.6587	98.20±0.1371	91.40±0.2501
9×9	OA	89.01±0.4104	98.69±0.0742	94.25±0.1418
	AA	90.89±0.3743	97.79±0.1923	82.39±0.7148
	Kappa(x100)	88.08±0.4437	98.25±0.0983	92.38±0.1877
11×11	OA	89.66±0.6428	98.69±0.1907	94.49±0.1229
	AA	91.25±0.5857	97.47±0.2967	82.81±1.0142
	Kappa(x100)	88.78±0.6967	98.25±0.2559	92.71±0.1638
13×13	OA	88.70±0.3437	98.64±0.1101	—
	AA	90.62±0.2716	97.31±0.4876	
	Kappa(x100)	87.74±0.3758	98.18±0.1464	

TABLE V: OA, AA, and Kappa values using HS and with HS+LiDAR data

	HSI			HSI + LiDAR		
	HOUSTON	TRENTO	MUUFL	HOUSTON	TRENTO	MUUFL
OA	83.23 ± 00.47	94.62± 00.21	91.99 ± 00.35	89.66 ± 00.64	98.69 ± 00.19	94.49 ± 00.12
AA	85.88 ± 00.33	91.33 ± 00.22	79.54 ± 01.93	91.25 ± 00.59	97.47 ± 00.30	82.81 ± 01.01
$\kappa(\times 100)$	81.88 ± 00.51	92.81 ± 00.28	89.37 ± 00.46	88.78 ± 00.70	98.25 ± 00.26	92.71 ± 00.16

TABLE VI: OA, AA and Kappa (in %) values using LiDAR as query vector and HSI as query vector in the proposed Cross-HL attention.

Class	HOUSTON		TRENTO		MUUFL	
	HSI (Query)	LIDAR (Query)	HSI	LIDAR	HSI	LIDAR
1	80.29±1.81	**83.10±0.00**	98.34±0.93	**99.32±0.31**	96.10±0.58	**98.33±0.21**
2	82.61±1.18	**97.37±4.44**	**95.58±0.78**	95.32±1.42	64.76±5.43	**89.65±1.18**
3	**99.45±0.34**	99.34±0.35	58.40±6.51	**97.62±1.17**	75.22±3.92	**89.86±1.37**
4	82.62±2.28	**97.30±2.57**	98.86±0.80	**99.89±0.10**	77.73±5.68	**96.28±1.01**
5	99.15±0.46	**99.94±0.09**	98.70±0.75	**99.97±0.04**	94.08±1.15	**95.64±0.65**
6	88.39±5.49	**96.97±1.36**	92.96±1.72	**94.17±1.36**	71.94±5.83	**86.55±2.97**
7	**85.70±4.01**	80.15±3.15	-	-	72.62±3.92	**92.73±1.39**
8	87.11±6.32	**91.01±2.72**	-	-	95.64±0.67	**98.02±0.42**
9	83.57±2.27	**90.82±1.58**	-	-	41.37±8.39	**68.81±2.91**
10	**62.79±4.79**	58.86±7.47	-	-	4.20±4.85	**26.12±8.56**
11	89.49±5.39	**98.77±0.77**	-	-	10.43±16.83	**71.53±3.22**
12	91.22±4.54	**91.55±1.78**	-	-	-	-
13	60.67±4.04	**90.49±2.00**	-	-	-	-
14	96.60±2.42	**99.96±0.13**	-	-	-	-
15	**99.47±0.56**	98.64±2.01	-	-	-	-
OA	85.42±0.70	**89.66±0.64**	97.30±0.51	**98.69±0.19**	86.66±0.26	**94.49±0.12**
AA	85.94±0.56	**91.25±0.59**	90.47±1.45	**97.47±0.30**	64.01±1.75	**82.81±1.01**
Kappa(x100)	84.20±0.75	**88.78±0.70**	96.38±0.68	**98.25±0.26**	82.25±0.36	**92.71±0.16**

els, including both transformer-based models and classical methods. It achieves OAs, AAs, and κ values of 89.01%, 90.89%, and 88.08%, respectively, with standard deviations of 0.41%, 0.37%, and 0.44%. The proposed `Cross-HL` model

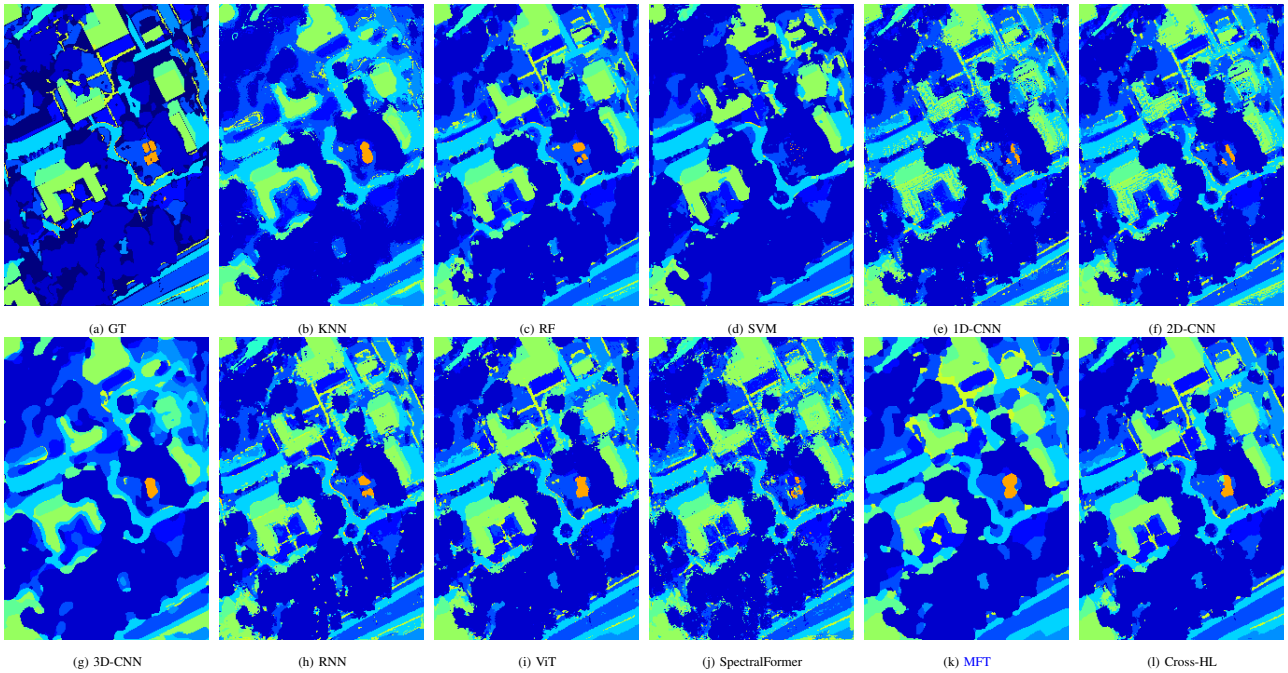


Fig. 9: The predicted land cover maps were created for the MUUFL data set.

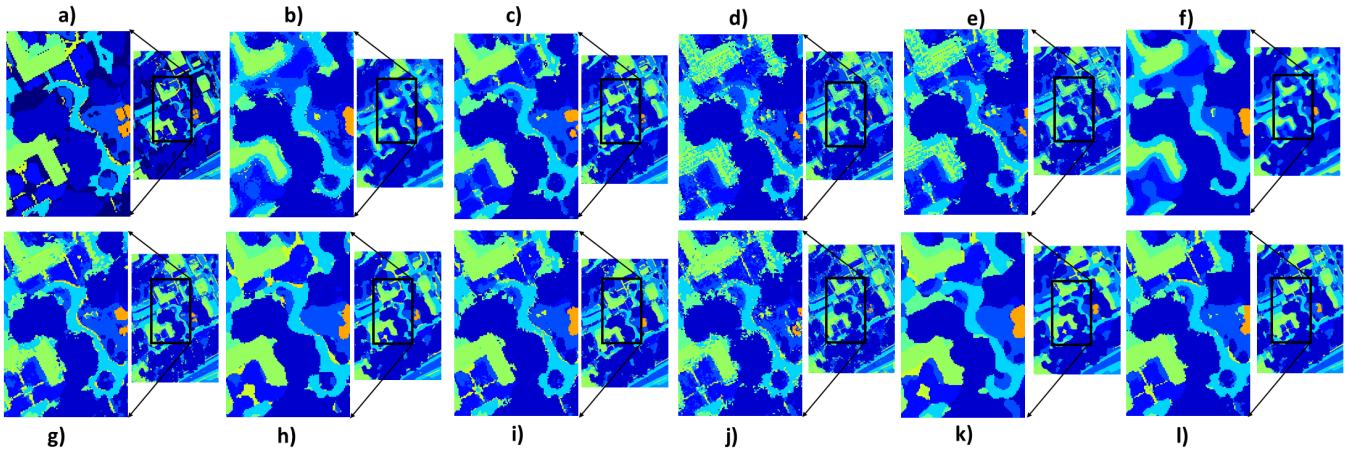


Fig. 10: The predicted land cover maps were created for the MUUFL dataset, magnifying the region of interest for the following methods: (a) Annotated map, (b) KNN, (c) RF, (d) SVM, (e) 1D-CNN, (f) 2D-CNN, (g) 3D-CNN, (h) RNN, (i) ViT, (j) SpectralFormer, (k) MFT, and (l) Cross-HL.

demonstrates superior performance compared to the second-best model, MFT, across the majority of class-wise accuracy metrics and all classification metrics. The proposed model consistently outperforms on the Houston dataset using both HS and LiDAR data, evident from its lower standard deviation in all three accuracy metrics in comparison to other transformer-based networks.

The classification performance over the Trento dataset, incorporating HS and LiDAR data, has been analyzed across various classification methods, with outcomes presented in Table II. The table encompasses conventional classification methods, classical convolutional networks, transformer-based methods, and the proposed Cross-HL attention transformer model. Transformer-based networks demonstrate superior performance compared to other methods. However, for the Trento dataset, Support Vector Machine (SVM) emerges as the top performer among conventional classifiers, achieving mean OAs, AAs, and κ values of 95.33%, 87.72%, and 93.76%, respectively. 3D-CNN surpasses other classical convolutional networks. Notably, 3D-CNN outperforms all methods, including transformer-based methods like ViT and SpectralFormer but MFT outperforms 3D-CNN in terms of OA and κ but exhibits inconsistent results in AA, with AA values lower than those of 3D-CNN and ViT. Nevertheless, the proposed Cross-HL attention transformer network outshines all compared methods in terms of OAs, AAs, and κ values. Additionally, the proposed Cross-HL attention transformer network displays an improvement of over 1.68%, 3.41%, and 2.26% for OA, AA, and κ, respectively, when compared to MFT.

Table III showcases the classification results obtained from methods based on conventional machine learning techniques, convolutional neural networks, vision transformers, and the

proposed Cross-HL attention transformer network over the MUUFL dataset. While RF and RNN demonstrate improved performance compared to other conventional methods, they fall short of transformer-based methods like ViT and MFT. In the case of MFT, a similar inconsistency in AA values is observable on the MUUFL dataset as well. However, the proposed Cross-HL attention transformer network outperforms all methods in terms of OAs, AAs, and κ values, demonstrating superiority across all three classification metrics.

Lastly, when HS data encounter complex land covers with high spectral similarity, distinguishing between these land covers becomes challenging. However, the incorporation of LiDAR data, providing elevation information, allows HS data to benefit from complementary data. While LiDAR alone might struggle to classify materials with similar elevation, integrating it into the transformer encoder in the form of query values (Q_L) enables effective fusion of HS and LiDAR data. This fusion facilitates the exploration of the long-range relationship between a pixel's elevation, spectral characteristics, and spatial features extracted from the HS sensor. As a result, the model's classification capabilities are enhanced.

D. Visual evaluation

Figs. 7, 8, and 9 demonstrate the classification maps obtained by the proposed method and other compared methods on UH, MUUFL, and Trento datasets, respectively in conjunction with their corresponding LiDAR modality. It can be observed that the predicted LULC maps are inconsistent with the classification results listed in Tables I, II, and III. We evaluated the quality of different classification methods by examining their resulting classification maps in terms of noise and edge boundary between the different land cover regionsWe analyzed the classification maps of various classification methods to assess their quality, specifically focusing on the noise and edge boundaries between different land cover regions. While traditional classifiers such as KNN, RF, and SVM produce detailed maps, they often have noises in the border area of LULC classes due to relying solely on spectral information from the data. However, deep learning models like 1D-CNN, 2D-CNN, and 3D-CNN due to their ability to capture non-linear relationships between the feature maps excel in exploiting the spatial-spectral information, resulting in smoother classification maps with distinct boundaries between different land-use and land-cover classes. Transformer-based models, such as SpectralFormer and MFT, are more effective for the HS classification task because of their capacity to derive highly abstract sequential representation. This is in contrast to the classical convolutional neural networks, which typically extract lower-level features. The ability to extract high-level features allows transformer-based methods to generate classification maps with enhanced visual quality. By using LiDAR modality as query Q_L, the Cross-HL attention module correlates the spectral and positional information of this pixel from the HS sensor with its elevation information obtained from the LiDAR sensor. Hence, the proposed model can effectively transmit positional information across layers, yielding extremely favorable classification maps, particularly with regards to texture and edge features, as compared to ViT, SpectralFormer, and MFT. Additionally, we chose a region of interest (ROI) from Fig. 9 and magnified it to emphasize the variations in classification maps across various models, as shown in Fig. 10.

E. Performance Over Varying Sizes of Training Samples

Furthermore, to assess the performance of the proposed model, varying sizes of training instances—namely 3%, 5%, 7%, and 9%—have been employed for training, while the remaining samples are designated for testing. The quantity of training samples plays a crucial role in determining a classification model's performance. The model may overfit or perform poorly on unobserved instances if there aren't enough training samples to allow it to fully understand the complexity of the data. Consequently, having an adequate number of training samples is essential to enable the model to generalize effectively, recognizing pertinent features and accurately classifying the data.

In this section, we analyze the proposed model's performance with respect to the size of the training samples while maintaining all other parameters constant, as outlined in Section IV-B. A random selection of 3%, 5%, 7%, and 9% from each dataset is utilized to train the model, while the remaining samples are used for evaluating the model's performance.

Figures 11(a)-(i) present a comparison of the classification performance of various transformer-based models—ViT, SpectralFormer (pixel-wise), SpectralFormer (patch-wise), MFT and the proposed Cross-HL attention transformer—across the three multimodal datasets: Houston (HS+LiDAR), Trento (HS+LiDAR), and MUUFL (HS+LiDAR), with varying training samples percentage. The performance of these models is assessed using three classification metrics: Overall Accuracy (OA), Average Accuracy (AA), and Cohen's Kappa (κ) values, displayed on the y-axis of the graphs. The horizontal axis represents the different training percentages. In Fig. 11, ViT, SpectralFormer (pixel-wise), SpectralFormer (patch-wise) and MFT are represented by distinct markers and orange, green, red and purple colors, respectively, while the proposed Cross-HL attention transformer is depicted in blue. Across all datasets, the evaluation metrics consistently demonstrate improvement or stability as the percentage of training samples increases. The graphs distinctly show that employing a 9% training sample consistently results in higher performance metric values compared to 3%, 5%, and 7% training samples.

The proposed Cross-HL attention transformer exhibits significant performance enhancement in terms of OA, AA, and κ accuracy over all comparative models for various training sample percentages in the MUUFL (HS+LiDAR) dataset. In the Trento (HS+LiDAR) and Houston (HS+LiDAR) datasets, the proposed Cross-HL attention transformer consistently outperforms all other models. However, the rate of improvement diminishes with larger training sample percentages. The graphs also indicate that the performance of transformer-based models like ViT, SpectralFormer (pixel-wise), SpectralFormer (patch-wise) and MFT is not consistently aligned with varying training sample sizes.

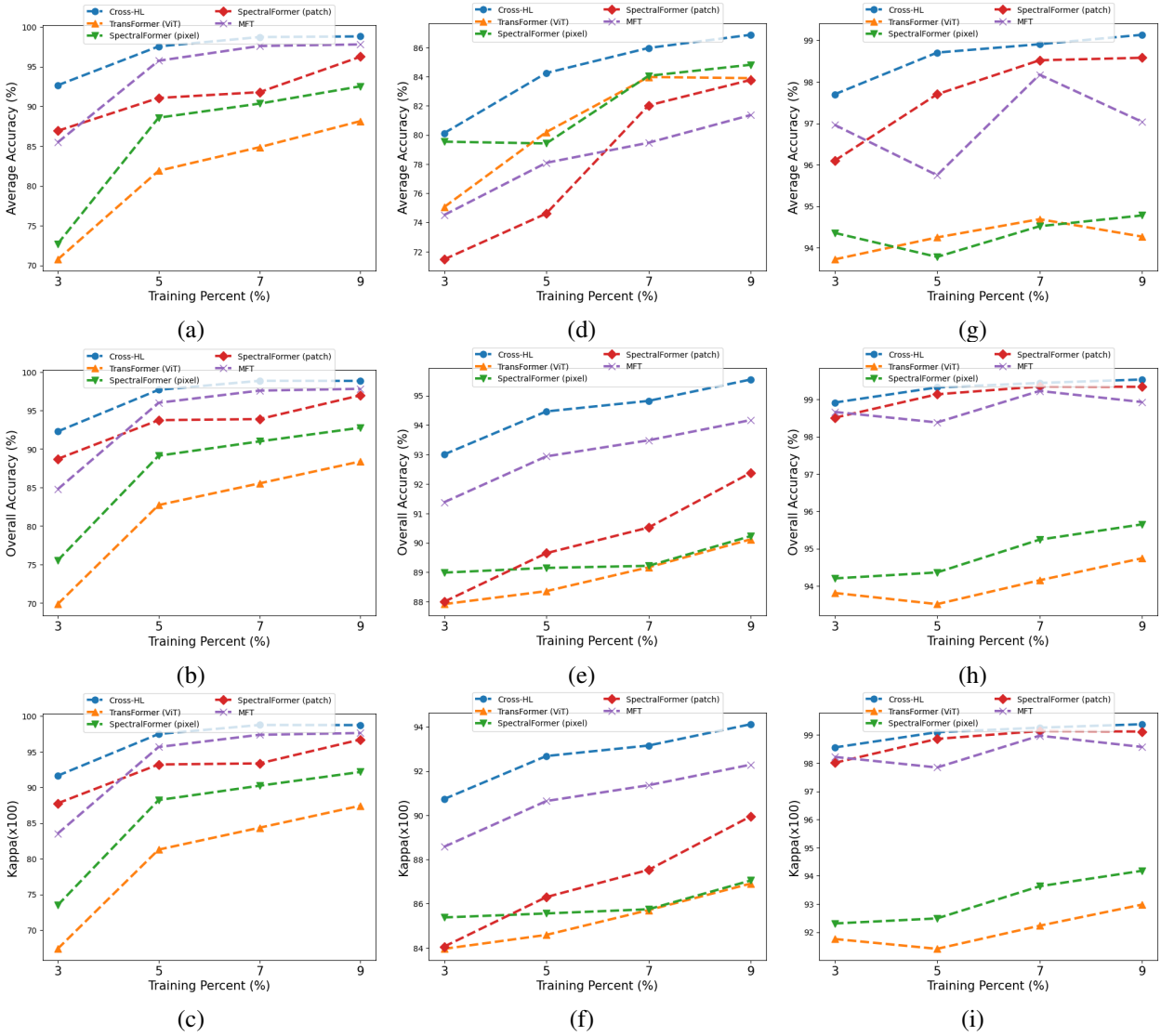


Fig. 11: The classification accuracies in terms of AA, OA, and kappa (κ) achieved by various techniques with different percentages of training samples randomly selected from: UH (left), MUUFL (center), and Trento (right) datasets are as follows.

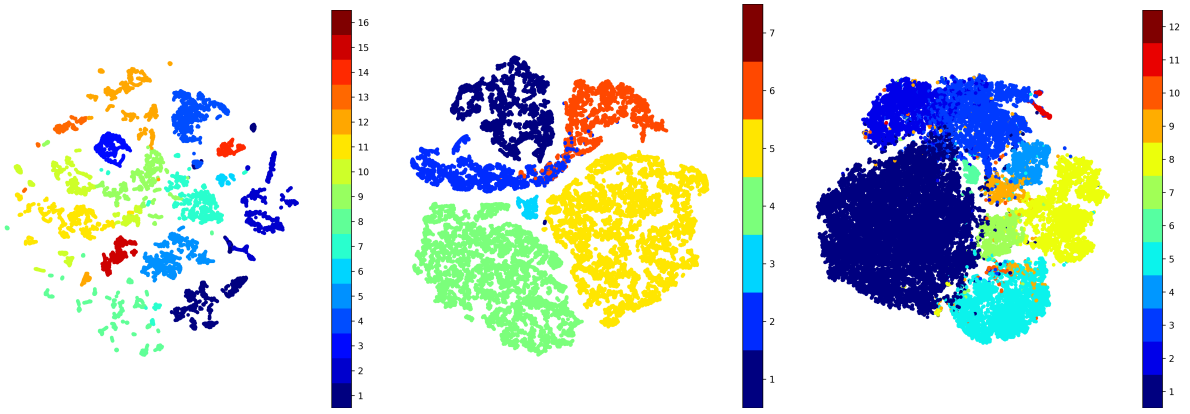


Fig. 12: The visualization displays the suggested spectral-spatial features for the test samples from various datasets using t-SNE. Each point represents the features of a test sample, and its class label is indicated by a distinct color: (a) HOUSTON, (b) TRENTO, and (c) MUUFL (optimal viewing with color).

For instance, over the MUUFL dataset, regarding AA, SpectralFormer (pixel-wise) is the second-best model following the proposed Cross-HL attention transformer when trained on 3% training samples. However, ViT takes the second spot when

trained on 5% of the training samples. For 7% and 9% training samples, SpectralFormer (pixel-wise) once again becomes the second-best model. Furthermore, the performance of MFT is on the rise with increasing percent of training samples but it exhibits suboptimal performance for 7% and 9%. In contexts of OA and κ for 3% training instances, SpectralFormer (pixel-wise) is the second-best model immediately after the proposed Cross-HL attention transformer, but from 5% training samples onward, SpectralFormer (patch-wise) emerges as the second-best model. Nonetheless, this inconsistency diminishes substantially in the proposed Cross-HL attention transformer model, exhibiting consistently superior performance compared to other models across all three datasets, regardless of the number of training samples. This underscores its superiority in HS data classification.

F. Feature Visualization using t-SNE

HS data consists of various spectral bands that capture extensive information across a wide range of electromagnetic wavelengths. Consequently, visualizing these high-dimensional features can pose a challenge. Nonetheless, t-Distributed Stochastic Neighbor Embedding (t-SNE) [56] has the potential to facilitate a more effective visualization of the intricate spectral-spatial features extracted by the proposed Cross-HL Attention Transformer within a 2D space. This visualization plays a crucial role in analyzing the representation capabilities of our proposed model, offering insights that might not be readily apparent through direct analysis of raw data. In Fig. 12, the spectral-spatial representation from the three datasets obtained by the proposed model are showcased. Each point corresponds to a feature, with its color denoting the assigned class label. The t-SNE plots in Fig. 12 (a), (b), and (c) originate from the Houston, Trento, and MUUFL datasets, respectively. The graphs clearly demonstrate that similar categories tend to cluster together, and the within-class variance is minimized across all three datasets. Consequently, it becomes evident from these plots that the proposed model effectively captures the semantic representation of spectral-spatial features.

G. Stability Analysis

• **Ablation Study:** To assess the efficacy of the proposed model, we performed ablation tests while maintaining all other experimental parameters constant. It is imperative to compare the performance of the proposed model under two scenarios: 1) utilizing solely HS data for generating query (Q), key (K), and value (V), and 2) integrating the LiDAR modality to provide complementary information to HS data by using LiDAR as query (Q_L), while employing HS for generating V and K values. Table V presents the performance metrics of the proposed model for these two scenarios across three widely used datasets: Houston, Trento, and MUUFL. Evidently, the table indicates that when LiDAR is used as the query (Q_L) in the Cross-HL attention module, the proposed model experiences a substantial enhancement in classification performance. For instance, within the Houston (HS with LiDAR) dataset, the proposed model exhibits a notable improvement of 6.43%, 5.37%, and 6.9%, respectively, in terms of OA, AA, and k.

• **Impact of Spatial Dimension:** Both spectral and spatial information hold crucial significance in HS data classification. The spatial dimension of the data patch can significantly influence the model's performance. A comparison of the model's performance across various window sizes is necessary to achieve the most stable results. The performance of the proposed Cross-HL Attention Transformer is evaluated on different patch sizes (3×3, 5×5, 7×7, 9×9, 11×11, and 13×13) across three datasets: Houston, Trento, and MUUFL. The outcomes are presented in Table IV, revealing that the optimal overall performance is achieved using a window size of 11×11 across all three datasets. Notably, altering the spatial dimensions of the patch has distinct effects on the model's performance for different datasets. For instance, the poorest result is obtained when using a patch size of 13×13 for the Houston dataset. Conversely, the patch size of 3×3 produces the lowest results for the Trento dataset. Hence, throughout the experimental evaluation, we adopted 11×11 HS and LiDAR patches.

• **Importance of LiDAR as query feature vector:** In the realm of developing attention mechanisms for HS data classification, choosing LiDAR as the query feature vector, interacting with key and value tokens of other HS patches, is more appropriate. LiDAR's precision in revealing spatial details, such as elevation, slope and topographical details, complements HS data, which struggles with spatial nuances, especially when classes share similar spectral signatures. Additionally, LiDAR's resilience to lighting and atmospheric distortions ensures reliable data, making it a robust choice for accurate classification in diverse scenarios. Moreover, using LiDAR for key or value vectors is less suitable as its primary strength lies in spatial analysis. Attempting to encode spectral information or provide context through LiDAR in these roles may not be optimal. key and value vectors play crucial roles in capturing and contextualizing information encapsulated in the query vector, and HS data, with its spectral richness, is better suited for these tasks. Therefore, LiDAR's use as the query feature vector enhances overall attention mechanism performance, particularly in image classification tasks. On the other hand, using HS data as a query feature vector to interact with LiDAR patches is less suitable. While HS excels in spectral analysis, it lacks the spatial resolution inherent in LiDAR, compromising the effectiveness of the attention mechanism. Table VI presents the classification results derived from employing LiDAR (the proposed model) and HS as the query feature vector. The results strongly substantiates the assertion that utilizing LiDAR as the query feature vector surpasses the efficacy of employing HS in the same role. The proposed Cross-HL model exhibits superior performance compared to the model utilizing HS as the query feature vector with LiDAR as key and value tokens across all three classification metrics. On the MUUFL dataset, the proposed model showcases a substantial margin of improvement: 7.83% in OA, 18.80% in AA, and 10.46% in k, respectively. Consistent trends in performance are observed across the other two datasets as well.

• **Hyperparameter Sensitivity Analysis:** In the field of

TABLE VII: Parameteres required (in K.) by different methods.

Convolutional Neural Networks				Transformer Models				
1D-CNN	2D-CNN	3D-CNN	RNN	ViT	Spectral Former (Pixel)	Spectral Former (Patch)	MFT	Cross-HL
100	318	290	138	90	217	226	313	457

HS data classification, it is crucial to consider the complexity and number of parameters in a model. Although our proposed model has a higher number of parameters compared to other models, this increase is not arbitrary. The trade-off is justified because our proposed model exhibits a significant improvement in classification performance, as evidenced by higher OA, AA, and k in Tables I, II and III. Moreover, the visual classification maps generated by our proposed model produce more appealing and distinguishable results as evident in the Figs. 7, 8, 9 and 10. This viewpoint highlights the idea that the trade-off in terms of increased model complexity(reflected in the higher number of parameters in Table VII) is aligned with a tangible and substantiated enhancement in the model's ability to accurately classify HS imagery.

V. Conclusions

In this paper, we propose the Cross-HL Attention Transformer, an extension of self-attention designed for land use and land cover classification. This approach leverages two distinct modalities: spectral information from HS data and elevation/structural information extracted from LiDAR data. These modalities are effectively fused using the attention mechanism to enhance the robustness of HS classification. Instead of employing conventional feature fusion techniques, we employ the LiDAR modality as the query value (Q_L), while the hyperspectral data serves as the key (K) and values (V) within the proposed Cross-HL attention block of the transformer encoder. This configuration facilitates the learning of long-range feature dependencies and captures relevant information from LiDAR, especially in scenarios where spectral information alone falls short in discerning subtle differences among similar land cover materials. Given its critical role in the classification framework, the CLS token contributes significantly to model generalization, enhancing classification accuracy by incorporating information complementary to that of the HS data. We evaluate the performance of the proposed model across three widely used benchmark datasets: Houston, Trento, and MUUFL. The outcomes of our proposed model demonstrate its superior performance across these benchmark datasets compared to state-of-the-art models.

References

[1] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.

[2] L. G. T. Crusiol, M. R. Nanni, R. H. Furlanetto, R. N. R. Sibaldelli, L. Sun, S. L. Gonçalves, J. S. S. Foloni, L. M. Mertz-Henning, A. L. Nepomuceno, N. Neumaier, and J. R. B. Farias, "Assessing the sensitive spectral bands for soybean water status monitoring and soil moisture prediction using leaf-based hyperspectral reflectance," *Agricultural Water Management*, vol. 277, p. 108089, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377422006369

[3] P. Ghamisi, K. R. Shahi, P. Duan, B. Rasti, S. Lorenz, R. Booysen, S. Thiele, I. C. Contreras, M. Kirsch, and R. Gloaguen, "The potential of machine learning for a more responsible sourcing of critical raw materials," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8971–8988, 2021.

[4] X. Cai, L. Wu, Y. Li, S. Lei, J. Xu, H. Lyu, J. Li, H. Wang, X. Dong, Y. Zhu, and G. Wang, "Remote sensing identification of urban water pollution source types using hyperspectral data," *Journal of Hazardous Materials*, vol. 459, p. 132080, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304389423013638

[5] B. F. R. Davies, P. Gernez, A. Geraud, S. Oiry, P. Rosa, M. L. Zoffoli, and L. Barillé, "Multi- and hyperspectral classification of soft-bottom intertidal vegetation using a spectral library for coastal biodiversity remote sensing," *Remote Sensing of Environment*, vol. 290, p. 113554, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723001050

[6] M. Wang, D. Hong, Z. Han, J. Li, J. Yao, L. Gao, B. Zhang, and J. Chanussot, "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 26–72, 2023.

[7] G. Avola, A. Matese, and E. Riggi, "An overview of the special issue on "precision agriculture using hyperspectral images"," *Remote Sensing*, vol. 15, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/7/1917

[8] A. Saha, B. Sen Gupta, S. Patidar, and N. Martínez-Villegas, "Identification of soil arsenic contamination in rice paddy field based on hyperspectral reflectance approach," *Soil Systems*, vol. 6, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2571-8789/6/1/30

[9] T. U. Rehman, L. Zhang, D. Ma, and J. Jin, "Common latent space exploration for calibration transfer across hyperspectral imaging-based phenotyping systems," *Remote Sensing*, vol. 14, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/2/319

[10] H. Qin, W. Xie, Y. Li, K. Jiang, J. Lei, and Q. Du, "Weakly supervised adversarial learning via latent space for hyperspectral target detection," *Pattern Recognition*, vol. 135, p. 109125, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322006057

[11] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[12] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and lidar data using ip-cnn," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[13] A. Jamali, S. K. Roy, and P. Ghamisi, "Wetmapformer: A unified deep cnn and vision transformer for complex wetland mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103333, 2023.

[14] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and lidar remote sensing data for classification of complex forest areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1416–1427, 2008.

[15] R. Hänsch and O. Hellwich, "Fusion of multispectral lidar, hyperspectral, and rgb data for urban land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 366–370, 2021.

[16] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.

[17] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[18] G. Gao, L. Yao, W. Li, L. Zhang, and M. Zhang, "Onboard information fusion for multisatellite collaborative observation: Summary, challenges, and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 2, pp. 40–59, 2023.

[19] N. He, M. E. Paoletti, J. M. Haut, L. Fang, S. Li, A. Plaza, and J. Plaza, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 755–769, 2019.

[20] J. Yang, B. Du, C. Wu, and L. Zhang, "Automatically adjustable multiscale feature extraction framework for hyperspectral image classification," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3649–3652.

[21] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[22] S. K. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, and J. Chanussot, "Hyperspectral and lidar data classification using joint cnns and morphological feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[23] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for hsi and lidar data joint classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[24] Y. Fan, Y. Qian, Y. Qin, Y. Wan, W. Gong, Z. Chu, and H. Liu, "Mslaenet: Multiscale learning and attention enhancement network for fusion classification of hyperspectral and lidar data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 10 041–10 054, 2022.

[25] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[26] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and lidar data using morphological features," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 552–556, 2015.

[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[30] 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.

[31] 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.*, 2022.

[32] Y. Yu, T. Jiang, J. Gao, H. Guan, D. Li, S. Gao, E. Tang, W. Wang, P. Tang, and J. Li, "Capvit: Cross-context capsule vision transformers for land cover classification with airborne multispectral lidar data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 111, p. 102837, 2022.

[33] Y. Zhang, Y. Peng, B. Tu, and Y. Liu, "Local information interaction transformer for hyperspectral and lidar data classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[34] 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," *Information Fusion*, 2021.

[35] 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*, 2023.

[36] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and lidar data classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 3095–3110, 2022.

[37] Y. Zhang, S. Xu, D. Hong, H. Gao, C. Zhang, M. Bi, and C. Li, "Multimodal transformer network for hyperspectral and lidar classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[38] 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*, vol. 61, pp. 1–20, 2023.

[39] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 60–88, 2020.

[40] M. E. Paoletti, O. Mogollon-Gutierrez, S. Moreno-Álvarez, J. C. Sancho, and J. M. Haut, "A comprehensive survey of imbalance correction techniques for hyperspectral data classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5297–5314, 2023.

[41] S. K. Roy, J. M. Haut, M. E. Paoletti, S. R. Dubey, and A. Plaza, "Generative adversarial minority oversampling for spectral–spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[42] P. Singh, V. K. Verma, P. Rai, and V. P. Namboodiri, "Hetconv: Heterogeneous kernel-based convolutions for deep cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*. PMLR, 2015, pp. 448–456.

[44] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel resnet for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7831–7843, 2021.

[45] 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*, vol. 7, no. 6, pp. 2405–2418, 2014.

[46] 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.

[47] X. Du and A. Zare, "Scene label ground truth map for muufl gulfport data set," *Dept. Elect. Comput. Eng., Univ. Florida, Gainesville, FL, USA, Tech. Rep*, 2017.

[48] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 968–999, 2022.

[49] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[50] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. and Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[51] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IGARSS*. IEEE, 2015, pp. 4959–4962.

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

[53] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[54] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

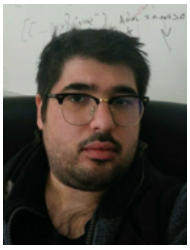
Swalpa Kumar Roy (Senior Member, IEEE) received the bachelor's and the master's degree both in Computer Science and Engineering from West Bengal University of Technology, Kolkata, India, in 2012, and Indian Institute of Engineering Science and Technology (IIEST), Shibpur, Howrah, India, in 2015 and also the Ph.D. degree in Computer Science and Engineering from University of Calcutta, Kolkata in 2021.

He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, Alipurduar Government Engineering and Management College, West Bengal, India. Previously he was associated with the same position at Jalpaiguri Government Engineering College, India. From July 2015 to March 2016, he was a Project Linked Person with the Optical Character Recognition (OCR) Laboratory, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata. Swalpa was nominated for the Indian National Academy of Engineering (INAE) engineering teachers mentoring fellowship program by INAE Fellows in 2021-2022 and 2023-24 also a recipient of the Outstanding Paper Award in second Hyperspectral Sensing Meets Machine Learning and Pattern Analysis (HyperMLPA) at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021. He serves as Associate Editor of the journal of Springer Nature Computer Science (SNCS) and also an topic editor of the frontiers journal of Advanced Machine Learning Techniques for Remote Sensing Intelligent Interpretation. He has served as a reviewer for the IEEE Transactions on Geoscience and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters. His research interests include computer vision, deep learning and remote sensing. web: https://swalpa.github.io

Atri Sukul is currently pursuing his bachelor's degree in Computer Science and Engineering from Jalpaiguri Government Engineering College, West Bengal, India.

Mr. Sukul was nominated for the Indian National Academy of Engineering (INAE) Engineering Students Mentoring Fellowship by INAE Fellows in academic tenure 2023–2024. His research interests include computer vision, and deep learning for remote sensing applications.

Ali Jamali received the doctor of philosophy (Ph.D.) degree in geoinformatics from Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia, in 2017. He is an experienced researcher with a demonstrated history of working in higher education, skilled in statistics, GIS, remote sensing, machine learning, and algorithm optimization. His current research interest includes remote sensing image processing based on advanced mathematical and statistical algorithms. He has developed various shallow and deep learning algorithms for satellite image processing which are mainly focused on ecological, flood, deforestation, and wetlands mapping/modelling etc. He is currently a postdoctoral fellow at Simon Fraser University (SFU), Barnaby, Canada, focusing on the development of cutting-edge machine learning techniques for advancing agriculture of British Columbia (BC), Canada. For more information, please visit https://www.researchgate.net/profile/Ali-Jamali.



Juan M. Haut (Senior member, IEEE) is a professor with the Department of Computers and Communications at the University of Extremadura, Caceres, Spain. Also, he is a member of the Hyperspectral Computing Laboratory (HyperComp) at the Department of Technology of Computers and Communications, University of Extremadura, where he received the B.Sc and M.Sc. degrees in computer engineering in 2011 and 2014, respectively, and the Ph.D. degree in Information Technology in 2019 supported by an University Teacher Training Programme from the Spanish Ministry of Education. Dr. Haut was a recipient of the Outstanding Ph.D. Award at the University of Extremadura in 2019. His research interests include remote sensing data processing and high dimensional data analysis, applying machine (deep) learning and cloud computing approaches. In this sense, he has authored/co-authored more than 70 JCR journal articles (more than 40 in IEEE journals) and more than 40 peer-reviewed conference proceeding papers. Some of his contributions have been recognized as hot-topic publications for their impact on the scientific community. Also, he was a recipient of the Outstanding Paper Award in the 2019 and 2021 IEEE WHISPERS conferences. From his experience as a reviewer, it is worth mentioning his active collaboration in more than 10 scientific journals, such as the IEEE Transactions on Geoscience and Remote Sensing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters, and he has been awarded with the Best Reviewer recognition of the IEEE Geoscience and Remote Sensing Letters and IEEE Transactions on Geoscience and Remote Sensing in 2018 and 2020 respectively. Furthermore, he has guest-edited special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of the IEEE Gesocience and Remote Sensing Letters and IEEE Journal on Miniaturization for Air and Space Systems. web: https://mhaut.github.io

Pedram Ghamisi (Senior Member) obtained his Ph.D. in electrical and computer engineering from the University of Iceland in 2015.

He currently serves as (1) the head of the machine learning group at Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany, and (2) visiting full professor at Lancaster University, UK. He has previously held positions as Senior PI, research professor, and group leader of AI4RS at the Institute of Advanced Research in Artificial Intelligence (IARAI), Austria. Additionally, he is a co-founder of VasoGnosis Inc., with branches in San Jose and Milwaukee, USA. His research interests primarily revolve around deep learning, particularly in the domain of remote sensing applications. For more information, please visit http://www.ai4rs.com.