Multimodal Fusion Transformer for Remote Sensing Image Classification

Swalpa Kumar Roy, *Studen Member, IEEE,* Ankur Deria, Danfeng Hong, *Senior Member, IEEE,* Behnood Rasti, *Senior Member, IEEE,* Antonio Plaza, *Fellow, IEEE,* and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—Vision transformers (ViTs) have been trending in image classification tasks due to their promising performance when compared to convolutional neural networks (CNNs). As a result, many researchers have tried to incorporate ViTs in hyperspectral image (HSI) classification tasks. To achieve satisfactory performance, close to that of CNNs, transformers need fewer parameters. ViTs and other similar transformers use an external classification (CLS) token which is randomly initialized and often fails to generalize well, whereas other sources of multimodal datasets, such as light detection and ranging (LiDAR) offer the potential to improve these models by means of a `CLS`. In this paper, we introduce a new multimodal fusion transformer (MFT) network which comprises a multihead cross patch attention (`mCrossPA`) for HSI land-cover classification. Our `mCrossPA` utilizes other sources of complementary information in addition to the HSI in the transformer encoder to achieve better generalization. The concept of tokenization is used to generate `CLS` and HSI patch tokens, helping to learn a distinctive representation in a reduced and hierarchical feature space. Extensive experiments are carried out on widely used benchmark datasets i.e., the University of Houston, Trento, University of Southern Mississippi Gulfpark (MUUFL), and Augsburg. We compare the results of the proposed MFT model with other state-of-the-art transformers, classical CNNs, and conventional classifiers models. The superior performance achieved by the proposed model is due to the use of multihead cross patch attention. The source code will be made available publicly at https://github.com/AnkurDeria/MFT.**

***Index Terms*—Vision transformer, convolutional neural networks, multihead cross patch attention, remote sensing.**

I. Introduction and Contributions

PHENOMENA such as climate change or desertification has led to a drastic growth in the popularity of Earth Observation (EO) via remote sensing (RS). These tasks include (but are not limited to) land cover classification [1]–[3], forestry [4], mineral exploration and mapping, object/target detection [5], [6], environmental monitoring [7], urban planning [8], biodiversity conservation, and disaster response and management. All of these tasks have been explored in the past few decades using data coming from single EO sensors, i.e., hyperspectral imaging (HSI) instruments, which can simultaneously provide rich spectral and spatial information [9]. However, such single-sensor data tend not sufficient to identify and recognize objects of interest.

Recent advances in RS technology have increased the availability of multi-sensor data, allowing for multiple representations of the same geographical region. Depending on the sensors' characteristics, the captured data can provide information with different characteristics for the same observed land-cover region. For example, synthetic-aperture radar (SAR) data provide the amplitude and phase geometrical information, while light detection and ranging (LiDAR) collects depth and intensity information, measuring the elevation of an object on the Earth's surface. Multispectral (MS) sensors measure the reflected light within specific wavelength ranges across the electromagnetic spectrum. Integrating these data with different modalities provides unique complementary information, allowing us to further obtain complete feature representations [10].

However, from the perspective of an RS imaging system, the compatibility of spatial and spectral resolution is often questionable [11]. The very high spectral resolution of HSI data often limits spatial resolution. As a consequence, spectral aliasing becomes a phenomenon that adversely affects land-cover classification in complex scenes [12]. In contrast, LiDAR provides elevation information that allows distinguishing land-cover objects with identical spectral signatures but different elevations, such as roads and roofs with built-in cement. For instance, HSI data can accurately differentiate between water and grass. But, for instance, roads and roofs built using the same materials cannot be differentiated. Therefore, the complementary elevation information obtained from LiDAR can be greatly beneficial for classification purposes [13]. Khodadadzadeh *et. al* extracted geometric representations from LiDAR data for land-cover classification [14]. Many works have explored to fully exploit the complementary information between HSI and other sources of multimodal data such as LiDAR, SAR, MS or digital surface models (DSMs) [10], [15]. In this context, researchers used extended morphological profiles (EPs) and attribute profiles (APs) for joint feature extraction and classification of HSI and LiDAR data [12],

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401. (*Corresponding author: Danfeng Hong*)

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

A. Deria is with the Department of Informatics, Technical University of Munich, 85748 Garching bei München, Germany (e-mail: i.am.ankur.deria@tum.de).

D. Hong is with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: hongdf@aircas.ac.cn)

B. Rasti is with Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg 09599, Germany. (e-mail: b.rasti@hzdr.de)

A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain (e-mail: aplaza@unex.es).

J. Chanussot is with the Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France (e-mail: jocelyn@hi.is)

[16], [17]. Rasti *et al.* improved the joint extraction of EPs by applying total variation component analysis for feature fusion [18]. Merentitis *et al.* introduced an automatic fusion technique using the random forest classifier for joint HSI and LiDAR data classification [19].

To adequately exploit the information in both HSI and other sources of multimodal data like SAR, DSM, LiDAR, and DSM, we propose a new yet efficient multimodal fusion transformer (MFT) network which uses multihead cross patch attention mechanism to fuse other sources of complementary information and HSI patch tokens for land cover classification. Our newly developed MFT incorporates the best properties of transformers encoder and multihead cross patch attention for the fusion of RS classification tasks without introducing much computational overhead. The major contributions of this paper can be summarized as follows:

- We propose a new multimodal fusion transformer (MFT) network to improve RS data fusion for land cover classification, where HSIs are used along with other multimodal data sources (e.g., LiDAR, MSI, SAR and DSM) in a transformer network to enhance the classification performance.
- We introduce a novel multihead cross patch attention (`mCrossPA`) mechanism for RS data fusion. The class token, which also contains supplementary information, is derived from multimodal data (e.g., LiDAR, MSI, SAR, and DSM), which are fed to the transformer network along with HSI patch tokens.
- The newly developed `mCrossPA` utilizes the widely used attention mechanism that can efficiently fuse the information from HSI patch tokens and existing `CLS` tokens into a new token that integrates multimodal features.
- We conduct extensive experiments on four public hyperspectral datasets, i.e., Houston, Trento, University of Southern Mississippi Gulfpark (MUUFL), and Augsburg. These experiments reveal the effectiveness of the proposed method. To illustrate the advantages of our approach, we use HSI data alone and also combine it with other sources of multimodal data.

The remaining of paper is organized as follows. Section II discusses the traditional methods, classic deep learning-based methods, and transformer-based methods used for HSI classification. The pre-processing mechanisms for HSI and other sources of multimodal data and the components of the proposed MFT model, including a novel LiDAR and HSI cross-patch attention mechanism for transformer-based deep feature fusion and image classification are described in Section III. Extensive experiments are then conducted, with an analysis of hyperparameter sensitivity and a discussion of the obtained results in Section IV. The paper concludes with some remarks and hints at plausible future research lines in Section V.

II. Related Works

A. Traditional Methods

Conventional methods have been widely used for HSI classification, even with limited training samples [20]–[22]. Generally speaking, these methods incorporate two steps. First, they represent the HSI data in feature space to reduce the dimensionality and extract a few highly informative features. Then, the extracted features are sent to a spectral classifier [1], [23]–[26]. Among conventional methods, support vector machines (SVMs) with non-linear kernels are very popular, especially when the training data are limited [27]. The extreme learning machine (ELM) has also been widely used for unbalanced feature classification, and it was used in HSI classification. It has been shown that ELMs were able to achieve better performance when compared to SVMs. The random forest has been used due to its discriminative power when classifying unlabeled HSIs [28]. However, conventional methods encounter a performance bottleneck when the training data becomes complex, due to their limitations in terms of data fitting and representation ability. The aforementioned spectral classifiers consider HSIs as a collection of spectral measurements without considering their spatial arrangement. Spatial-spectral classifiers considerably boost the performance of spectral classifiers by incorporating spatial information such as the shape and size of different objects and textures. It is also worth mentioning that the spectral classifiers are less robust to noise compared to the spatial-spectral classifiers [1], [29].

B. Conventional Deep Learning Methods

In recent years, multimodal data integration for RS data classification using deep learning (DL) methods has attracted significant attention [30]. DL methods can efficiently learn spectral and spatial features from vast amounts of fragmented data. In the case of HSI-LiDAR data integration, both supervised and unsupervised DL methods have shown efficient performance in deep feature representation. Hong *et al.* [31] developed a new technique for classifying multi-source RS data (HSI and LiDAR), which can effectively extract compact feature representations of multimodal RS data. Very recently, Hong *et al.* first proposed a general and unified multimodal deep learning framework for RS image classification [32] which provides an effective solution with respect to multi-source and multimodal RS products.

Unlike unsupervised methods, which learn the feature representation using observed data, supervised methods rely on ground reference training samples. Therefore, DL-based supervised techniques often demonstrate superiority in RS image classification tasks. One-dimensional convolution (CNN1D) [33], two-dimensional convolution (CNN2D) [34], and three-dimensional convolution (CNN3D) [35] have shown success in HSI and LiDAR data classification. Recent studies show that DL methods can balance algorithm accuracy and robustness. On the other hand, shallow learning methods are dependent on the prior information obtained from training or observed data. He *et al.* proposed a residual network (ResNet) with minimum information loss after each convolutional operation to redeem the vanishing gradient problem [36]. Zhong *et al.* designed a spectral-spatial residual network (SSRN) that can better utilize both spectral and spatial information for enhanced classification [37]. Roy *et. al* explored the light-weighted paradigm by modeling spectral and spatial features extracted through the squeeze-and-excitation residual network,

which can be combined with bag-of-features learning for accurate land-use and land-cover classification [38], [39]. To boost the performance of SSRNs, Zhu *et al.* included another spatial and channel attention layer in the SSRN architecture to extract discriminative representation [40]. To take advantage of the residual behavior of the network, one can extend it to form an even more complex model, i.e., the lightweight spectral-spatial squeeze-and-excitation attention [38] with adaptive kernels [20] and pyramidal residual networks [41]. Lightweight heterogeneous kernel convolution [42], rotation equivariant CNN [43] and gradient centralized convolution [3], [44] enable efficient feature extraction and classification. On the other hand, the generative adversarial network also provides a solution for the classification of HSIs containing a class-wise imbalanced number of samples [45], [46].

The networks mentioned above (and their variants) are generally insufficient for detecting subtle discrepancies among spectral dimensions. CNNs, which have shown their capacity in the task of extracting spatial-contextual information from HSIs, barely capture the sequence attributes, especially the middle and long-term dependencies. That leads to a decrease in performance, especially when the data comprise many classes with similar spectral signatures, creating problems in extracting diagnostic spectral attributes. On the other hand, recursive neural networks (RNNs) can accurately model spectral signatures from HSIs by accumulating them band by band in an orderly manner. As such, the order of spectral bands is essential in learning long-term dependencies and preventing gradient vanishing problems [47], which might further complicate the interpretation of spectrally salient changes. This is because HSIs contains limited samples, and RNNs cannot train the model concurrently, which limits the classification performance. In this work, we rethink HSI data classification using transformers to address the aforementioned limitations.

C. Transformer-based Methods

Unlike CNNs and RNNs, transformers are one of the most cutting-edge backbone networks due to their adoption of self-attention techniques, that are efficient for processing and analyzing sequential (or time series) data [48]. Inspired by the success of transformers in computer vision [49], several new transformer models have been developed in the last few years [50]. Although the transformer excels at capturing the information containing in spectral signatures, it cannot equally characterize local semantic elements and fails to make adequate use of spatial information. The original transformer [51] is a model based on the self-attention mechanism that is mainly used in natural language processing (NLP). A series of tokens are used as model's input and multi-head attention is employed to draw global correlations in the input character sequence. Hong et al. recently introduced SpectralFormer [22], which can learn spectral representation information using neighboring bands and can construct a cross-layer transformer encoder module. The spectral-spatial feature tokenization (SSFT) transformer model [52] was suggested to make use of the transformer's capacity to extract local spatial semantic information and represent the link between neighboring sequences. The major drawback of this model is the number of parameters that are used. SSFT performs well with HSI data; however, it cannot include other multimodal data sources during the classification task. Ding *et al.* introduced a global–local transformer network (GLT-Net) to fully exploit the potential of transformer architectures for modelling the long-range dependencies as well as the convolution operator, which characterizes locally correlated features in joint classification of HSI and LiDAR data [53]. To fuse multisource heterogeneous information and enhance the joint classification performance, a novel dual-branch network was presented by Zhao *et al.* that consists of a transformer network and a hierarchical CNN [54]. Yu *et al.* created a cross-context capsule vision transformer (CapViT) to take advantage of long-range global feature interactions at various context scales for land cover classification with MS-LiDAR data [55]. In order to capture the relationship between HSI and its complementary information such as LiDAR, as well as to fuse the multisource features and prevent the generation of redundant information, the local information interaction transformer (LIIT) model was introduced for land cover classification using a combination of HSI and LiDAR data [56]. A multi-attentive hierarchical fusion network (MAHiDFNet) was also developed in [57] to implement feature-level fusion and classification of HSIs using LiDAR data. The connection between the structure and shape information of the HSI token and the CLS token was enhanced via spectral and spatial morphological convolution procedures in conjunction with a self-attention mechanism (morphFormer) for land cover classification in [58].

III. Methodology

A. Pre-processing of HSI and LiDAR Data

Let us denote an HSI as $\mathbf{X_H} \in \mathcal{R}^{M\times N\times B}$ and a corresponding LiDAR image as $\mathbf{X_L} \in \mathcal{R}^{M\times N}$ where M and N respectively represent two spatial dimensions and B refers to the number of spectral bands of the HSI. All the pixels are classified into C land-cover classes denoted by $C = (y_1, y_2, \dots y_c)$. A single pixel can be defined as $\mathbf{x}_{i,j} \in \mathbf{X_H}$, where $i = 1, \dots M$ and $j = 1, \dots N$ with $\mathbf{x}_{i,j} = [x_{i,j,1}, \dots x_{i,j,B}] \in \mathcal{R}^B$ containing B spectral bands. Patch extraction is performed in a pre-processing step, where a spectral-spatial cube $\mathbf{x}_h^{i,j} \in \mathcal{R}^{k\times k\times B}$ is obtained from the normalized HSI data $\mathbf{X_H}$ and a spatial image $\mathbf{x}_l^{i,j} \in \mathcal{R}^{k\times k}$ is obtained from the LiDAR data $\mathbf{X_L}$, with neighboring regions of size $(k \times k)$ and centered at pixel (i, j).

By jointly exploiting the spectral-spatial information, we can increase the discriminative power of the feature learning network. For this reason, spectral-spatial cubes $\mathbf{x}_h^{(i,j)}$ are extracted from the original HSIs and stacked into X_H prior to the feature extraction process. Similarly, for the LiDAR data, spatial patches of the same size $\mathbf{x}_l^{(i,j)}$ are extracted and stacked into X_L. Finally, the training and test samples in each class for X_H and X_L can be represented by

$$\begin{aligned} D^{train} &= \{(x_h, x_l), y^{(i)} | i = 1, \dots P\} \\ D^{test} &= \{(x_h, x_l), y^{(i)} | i = 1, \dots Q\} \end{aligned} \tag{1}$$

where $x_h \in X_H$ and $x_l \in X_L$ are randomly chosen from the HSI and LiDAR data, P and Q represent the number of

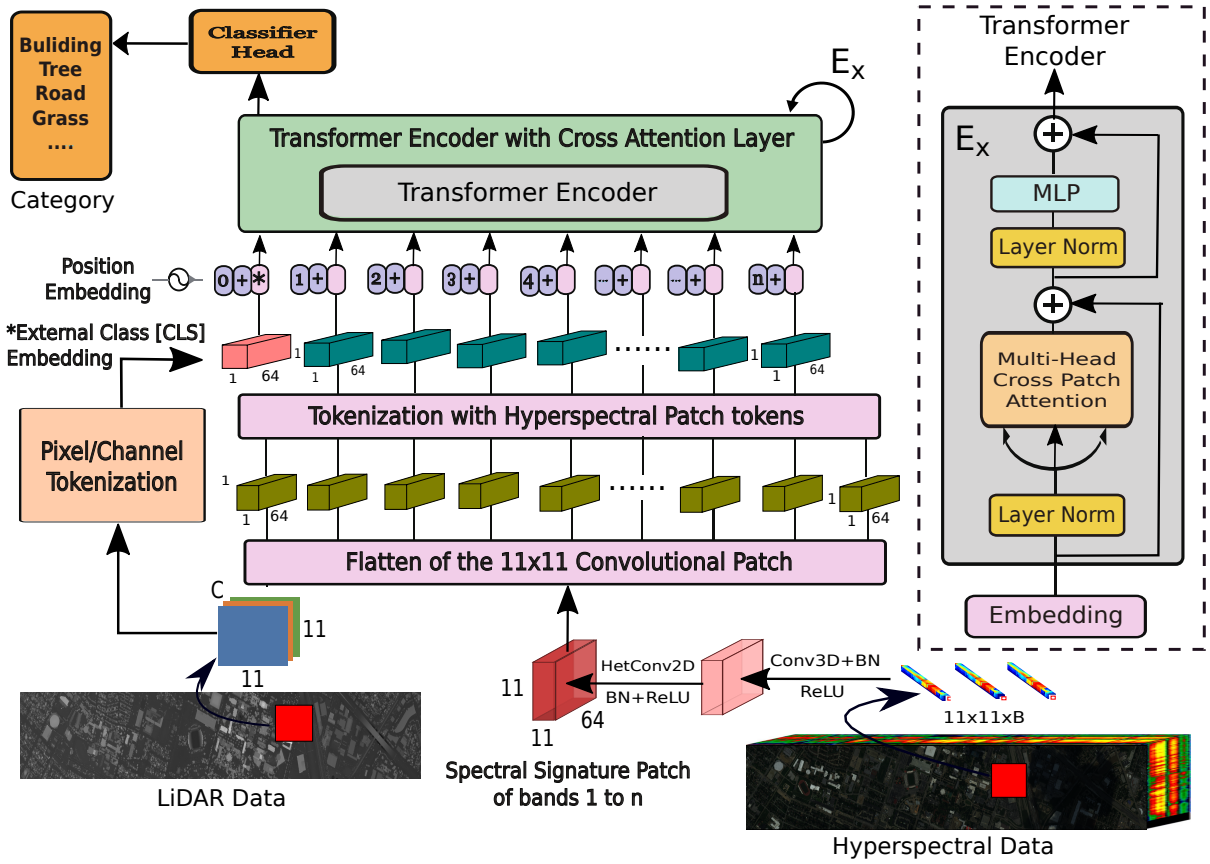


Fig. 1: Graphical representation of the proposed multimodal fusion transformer network for HSI and LiDAR data fusion.

training and test samples, and $y^{(i)}$ is the actual class for the i^{th} land-cover pixel.

B. Proposed Multimodal Fusion Transformer

Fusion is crucial for effectively learning multimodal feature representations of RS data since complementary information of a model can enhance the learning of discriminative features. If we incorporate this technique into conventional transformer models, we may exponentially increase the number of parameters. For instance, in conventional ViT models, if we use the HSI data as input, we increase the linear projection's complexity due to the fact that the input has numerous spectral bands, which may lead to chances of over-fitting. If we assume that other multimodal data (e.g., LiDAR, SAR, or DSM) are concatenated to the HSIs to learn complementary information, the aforementioned problem is aggravated due to the increase in the number of bands. Hence, a transformer model can effectively learn multimodal information without increasing computational overhead of linear projection.

To address the aforementioned challenges, we introduce a multimodal fusion transformer (MFT) where the input patches are taken from the HSIs. Additionally, the `CLS` token is generated from a patch that describes the same spatial region as the HSI patch, taken from the corresponding LiDAR image (or other modalities like SAR, MSI, DSM etc). In addition, we propose a simple yet effective multi-head cross patch attention (`mCrossPA`) module to fuse the LiDAR (as `CLS`) token and the HSI patch tokens. Fig. 1 depicts the proposed multimodal fusion transformer (MFT) network architecture for multimodal RS image fusion to improve the land cover classification performance. For the sake of simplicity, only LiDAR data from the University of Houston dataset are considered, but the concept can be extended to other multimodal data as well. The results for different kinds of multimodal data (i.e., MS, SAR, and DSM) are reported in the results section. The objective of the proposed multimodal fusion transformer model is to learn the spectral-spatial patch embeddings instead of the band-wise embeddings of the input HSIs, in addition to enriching the abstract description of the `CLS` token by considering LiDAR as an external class embedding without introducing any computational overhead. The components of the proposed multimodal fusion transformer are discussed step by step as follows.

1) HSI and LiDAR Feature Learning via CNNs: The ability of automatic contextual modeling among features helps CNNs to make strong inferences and exhibit promising performance in HSI classification tasks. The availability of a large number of spectral bands in HSIs enables us to utilize the benefits of CNNs and control the depth of the output feature maps. It has already been shown that CNNs can extract high-level abstract features, which are also invariant to the source of data modality. Hence, the CNNs adopted in our proposed model extract high-level abstract features that are used as input in the transformer and reduce the HSI bands down to a suitable

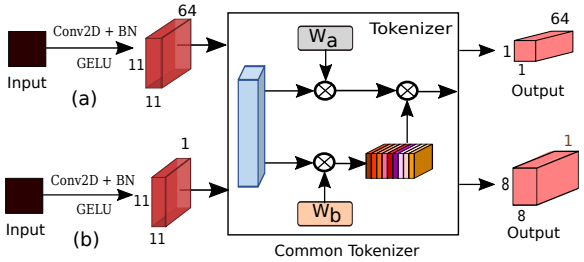


Fig. 2: Graphical representation of: (a) channel and (b) pixel tokenizer.

number.

In order to extract robust and discriminative features from raw HSIs, sequential layers of Conv3D and `HetConv2D` are used. Fig. 1 shows a graphical representation of the `Conv3D` and `HetConv2D` network for HSIs. The HSI cube $\mathbf{X}_H$ of size $(11 \times 11 \times B)$ is first unsqueezed into shape $(1 \times 11 \times 11 \times B)$, passed through the first `Conv3D` layer to produce $X_{in} = Conv3D(X_H)$ and the size of the considered kernels of that layer is chosen to be $(3 \times 3 \times 9)$ while the padding is $(1 \times 1 \times 0)$ to keep the spatial height and width of the output image equal to those of the input. On the other hand, the `HetConv2D` [59] utilizes two `Conv2D` layers in parallel, one of which performs a group-wise convolution (with kernel size = 3, groups = 4 and padding = 1) and the other one performs a point-wise convolution (with kernel size = 1, groups = 1 and padding = 0). The outputs from both these convolutions are added in element-wise ($\oplus$) fashion and returned from the `HetConv2D` layer. These can be defined as follows:

$$X_{out} = \begin{cases} Conv2D(X_{in}, k = (3,3), s = 1, g = 4, p = 1) \\ \oplus \\ Conv2D(X_{in}, k = (1,1), s = 1, g = 1, p = 1) \end{cases} \tag{2}$$

The shapes of the output feature maps after the `Conv3D` layer and the `HetConv2D` layer are $(8 \times 11 \times 11 \times (B-8))$ and $(11 \times 11 \times 64)$, respectively. Both of these layers are followed by their respective batch normalization (BN) [60] and ReLU activation layers. The BN layer is used to address the issue of overfitting (due to the small number of training samples) and also to accelerate the training performance. ReLU plays an important role by introducing non-linearities into the output features maps in order to achieve efficient and smooth propagation of the loss gradient.

2) HSI and LiDAR Tokenization: Spectral and spatial features contain highly discriminative information, leading to higher accuracy in HSI classification tasks. HSI cubes of size $\{(11 \times 11) \times 64\}$ are flattened into feature patch tokens with shape (1×64) each. n such patches are selected from 121 patches using tokenization [52]. The tokenization operation uses two learnable weights, i.e., W_a and W_b, which are multiplied by the input to extract the key features. The following equation summarizes the process:

$$\begin{aligned} X^H_{flat} &= T(\text{Flatten}(X_{out})) \\ X^H_{W_a} &= \text{softmax}(T(X^H_{flat}.W_{aH})) \\ X^H_{W_b} &= X^H_{flat}.W_{bH} \\ X^H_{patch} &= X^H_{W_a}.X^H_{W_b} \end{aligned} \tag{3}$$

where $T(.)$ is a transpose function, $X^H_{flat} \in \mathcal{R}^{121\times64}$, $W_a \in \mathcal{R}^{64\times n}$ and $W_b \in \mathcal{R}^{64\times64}$, $X^H_{W_a} \in \mathcal{R}^{n\times121}$ and $X^H_{W_b} \in \mathcal{R}^{121\times64}$ and $X^H_{patch} \in \mathcal{R}^{n\times64}$.

Focusing on the LiDAR data, a patch of size $(11 \times 11 \times C)$ extracted from a LiDAR image is first passed through either a pixel tokenization or a channel tokenization operation shown in Fig. 1. Both operations output a `CLS` token of shape (1×64) generated from the input LiDAR patch by passing the it through a `Conv2D` layer, followed by a BN layer, a `GELU` activation function and a tokenizer, respectively. Fig. 2 illustrates the difference between the pixel and the channel tokenization operations. The `Conv2D` layer may reduce the number of channels to 1 or increase it to 64 depending on the variation used. The convolutional layer extracts different spatial elevation information of height and shape like roads and roof built-in cement having different heights, which facilitates the exchange of complementary information among the HSI patch tokens (X^H_{patch}). The LiDAR data may have only one channel to begin with, but other multimodal types of data like SAR, DSM, MSI, etc. may have several channels which need to be reduced to one, since the aim of the network is to use the LiDAR data as the `CLS` token which should have only one channel. The input and output spatial heights and widths are kept the same through the convolutional layer using a kernel of size 3 and padding of size 1. The LiDAR feature embeddings are now suitable to be used as the external classification (`CLS`) token. The whole step can be summarized as follows:

$$\begin{aligned} X^L_{conv} &= T(\text{Flatten}(Conv2D(X^L_{patch}))) \\ X^L_{W_a} &= \text{softmax}(T(X^L_{conv}.W_{aL})) \\ X^L_{W_b} &= X^L_{conv}.W_{bL} \\ X^L_{cls} &= X^L_{W_a}.X^L_{W_b} \end{aligned} \tag{4}$$

where $X^L_{patch} \in \mathcal{R}^{11\times11\times C}$, $T(.)$ is a transpose function and, in the case of the pixel tokenizer, $X^L_{conv} \in \mathcal{R}^{121\times1}$, $W_{aL} \in \mathcal{R}^{1\times1}$ and $W_{bL} \in \mathcal{R}^{1\times64}$. In the case of the channel tokenizer, $X^L_{conv} \in \mathcal{R}^{121\times64}$, $W_{aL} \in \mathcal{R}^{64\times1}$ and $W_{bL} \in \mathcal{R}^{64\times64}$. In both cases, $X^L_{W_a} \in \mathcal{R}^{1\times121}$ and $X^L_{W_b} \in \mathcal{R}^{121\times64}$ and $X^L_{cls} \in \mathcal{R}^{1\times64}$.

The `CLS` token, i.e., the LiDAR feature embeddings are then concatenated to the n HSI patch embeddings, making a total of $(n+1)$ patches from n as shown in Eq. (5). A size of 64 is used because it is a power of 2, which helps in calculating the head dimension.

$$\widehat{X}^{HL} = [X^L_{cls} \ \| \ X^H_{patch}] \tag{5}$$

To retain the positional information, trainable position embeddings are added to the patch embeddings to preserve the semantic textural information of the image cube within the image patch tokens. For visual understanding, we refer to

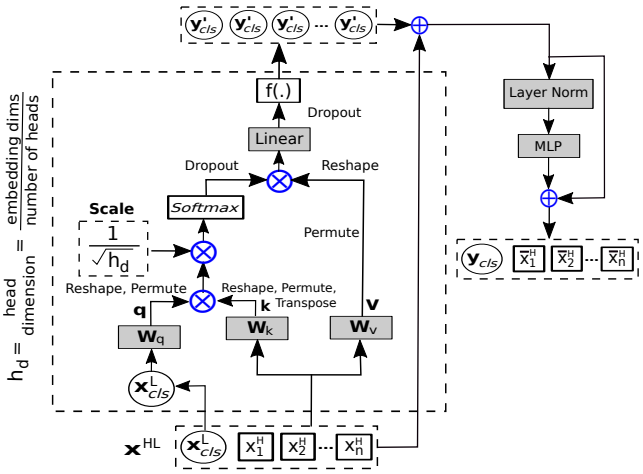


Fig. 3: Multihead cross patch attention taking example of LiDAR and HSI data.

Fig. 1 where the position embeddings are added in element-wise fashion to all the patches (1 to $n+1$). This process is followed by a dropout layer which helps in reducing the effects of the overfitting of the network. The above process can be summarized by the following equation:

$$\begin{aligned} X^{HL} &= \mathcal{DP}(X^{L}_{cls} \oplus \mathcal{PE} \;\; \| \;\; X^{H}_{patch} \oplus \mathcal{PE}) \\ &= (\widehat{X}^{L}_{cls} \;\; \| \;\; \widehat{X}^{H}_{patch}) \end{aligned} \tag{6}$$

where $\mathcal{DP}$ is a dropout layer with a value of 0.1 and $\mathcal{PE}$ is a learnable position embedding.

Algorithm 1: Transformer Encoder with `mCrossPA`

Input: $\mathbf{X^{HL}}$, $\mathbf{X^{L}_{cls}}$, $\mathbf{X^{HL}_{patch}}$, h_d
Output: $\mathbf{y_{cls}}$
Multihead Cross Patch Attention (Begin)
Step 1. $\mathbf{Q} = \mathbf{X^{L}_{cls}W_q}$, $\mathbf{K} = \mathbf{X^{HL}W_k}$, $\mathbf{V} = \mathbf{X^{HL}W_v}$,
Step 2. $\mathbf{Z} = \mathtt{softmax}(\mathbf{QK^T}/\sqrt{\mathbf{h_d}})$, (7a)
$\mathtt{mCrossPA}(\mathbf{X^{HL}}) = \mathcal{DP}(\mathbf{W_l ZV})$ (Eq. (7b))
Multihead Cross Patch Attention (End)
Step 3. $\mathbf{y'_{cls}} = \mathtt{mCrossPA}(\mathbf{X^{HL}}))$ (Eq. (8a))
Step 4. $\mathbf{y_k} = f(y'_{cls}) \oplus ([\mathbf{\widehat{X}^{L}_{cls}} \;\; \| \;\; \mathbf{\widehat{X}^{H}_{patch}}])$ (Eq. (8b)),
$\mathbf{X^{HL}_k} = \mathbf{y_k} + \mathtt{MLP}(\mathtt{LN}(\mathbf{y_k}))$ (Eq. (9)),
$\mathbf{y_{cls}} = \mathbf{X^{HL}_k}(1,:)$

3) LiDAR and HSI Cross Patch Attention: The `CLS` (X^{L}_{cls}) token plays a vital role in learning the abstract representation of the entire HSI patch by exchanging the information between the patch tokens and itself. This entire process takes place in the transformer encoder blocks, where each transformer encoder block comprises a residual multihead cross patch attention block and a residual `MLP` block, respectively. The residual attention block starts with a layer normalization (LN) operation followed by a self-attention layer, after which the output is added in element-wise ($\oplus$) fashion to the input of the LN (shown in Fig. 1). Similar to the attention block, the `MLP` block also starts with an LN operation, but the input of the LN is added element-wise ($\oplus$) to the output of `MLP` layer, as shown in Fig. 1.

Fig. 3 illustrates the cross patch attention module for the LiDAR data (as a `CLS`) and for the HSI data (as a patch token). More specifically, the generated patch tokens from the HSI feature maps are collected and then concatenated with the external `CLS` tokens from the LiDAR data from the same spatial region corresponding to the HSI patch, followed by the combination of position embeddings and dropout, as shown in Eqs. (5) and (6) and illustrated in Fig. 1. After that, the module performs cross patch attention (`CrossPA`) between X^{L}_{cls} and X^{HL}_{patch}, where query ($\mathbf{Q} = \mathbf{X}^{L}_{cls}\mathbf{W}_q$ where $\mathrm{W_q} \in \mathcal{R}^{64\times64}$) along with the key ($\mathbf{K} = \mathbf{X}^{HL}\mathbf{W}_k$ where $\mathrm{W_k} \in \mathcal{R}^{64\times64}$) and the value ($\mathbf{V} = \mathbf{X}^{HL}\mathbf{W}_v$ where $\mathrm{W_v} \in \mathcal{R}^{64\times64}$) are fused to generate the `CLS` token (y_{cls}) at the end of the `CrossPA` block. A linear projection layer ($\mathbf{W}_l \in \mathcal{R}^{64\times64}$) is applied to the final fused `CLS` token, which precedes a dropout layer ($\mathcal{DP}$) with a value of 0.1. Mathematically, the cross attention can be formulated as follows:

$$Z = \mathrm{softmax}(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{\mathrm{h_d}}}) \tag{7a}$$

$$\mathrm{CrossPA}(\mathrm{X^{HL}}) = \mathcal{DP}(\mathbf{W}_l\mathbf{ZV}) \tag{7b}$$

where $Z \in \mathcal{R}^{1\times64}$ and h_d = embedding dimension/number of heads. As shown in the self-attention module, if the number of heads is more than one, then the `CrossPA` will become a multi-head cross attention and, upon doing so, it can be represented as `mCrossPA`. The connections between the various patch tokens and CLS token are further strengthened by applying cross patch attention with numerous heads. The output y_k of the `mCrossPA` module for a given embedding X^{HL}_{k-1} with layer normalization (LN) and residual shortcut, where k is the k^{th} transformer encoder block, is defined below:

$$y'_{cls} = \mathrm{mCrossPA}(\mathrm{LN}(X^{HL}_{k-1})) \tag{8a}$$

$$y_k = f(y'_{cls}) \oplus [\widehat{X}^{L}_{cls} \;\; \| \;\; \widehat{X}^{H}_{patch}] \tag{8b}$$

where $y_k \in \mathcal{R}^{64\times64}$ and $f(.)$ is a function for broadcasting the dimension of y'_{cls} that is used to make the dimension the same as that of input X^{HL}_{k-1}, as shown in Fig. 3. This output y_k is then passed through multi-layer perceptron (`MLP`) block after layer normalization, followed by a residual shortcut to generate X^{HL}_{k}, which can be treated as input to $(k+1)$th transformer encoder block, as shown in Fig. 3, and can be defined as follows:

$$X^{HL}_{k} = y_k \oplus \mathrm{MLP}(\mathrm{LN}(y_k)) \tag{9}$$

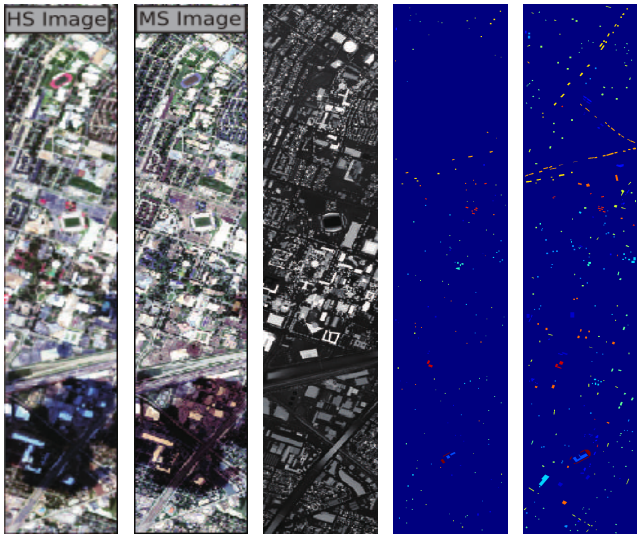
The number of heads h_d is set to 8 in the proposed model. Finally, the output from the kth transformer encoder blocks $y_{cls} = X^{HL}_{k}(1,:)$ is fed into the classifier head to get the final classification results.

IV. Experiments

A. HSI Datasets

In this section, we consider four different HSIs and their associated multimodal sources of data (LiDAR, MS, SAR, and DSM) to evaluate the performance of the proposed multimodal fusion transformer network. The experimental datasets include the University of Houston (UH), Trento, MUUFL Gulfport, and Augsburg scenes.

- The **University of Houston** dataset was collected by the Compact Airborne Spectrographic Imager (CASI) and was



(a) Pseudo-color Map (b) MSI Map (c) LiDAR Map (d) Training Samples (e) Test Samples

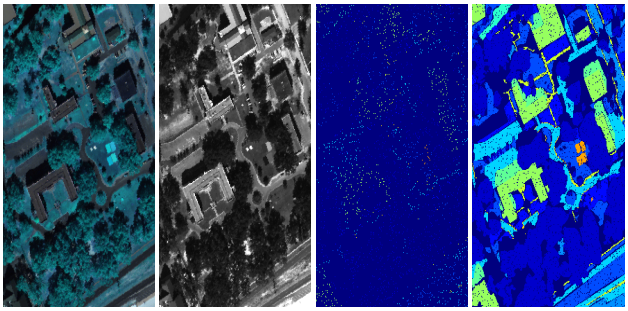
Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	662013	652648		Grass-healthy	198	1053
	Grass-stressed	190	1064		Grass-synthetic	192	505
	Tree	188	1056		Soil	186	1056
	Water	182	143		Residential	196	1072
	Commercial	191	1053		Road	193	1059
	Highway	191	1036		Railway	181	1054
	Parking-lot1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

Fig. 4: the University of Houston (UH) scene. (a) Pseudo-color image from the HSI data using bands 64, 43, and 22, respectively. (b) Grayscale image from the MSI data (c) Grayscale image from the LiDAR data, (d) Ground truth of disjoint training samples, and (e) Ground truth of disjoint test samples. The table represents class-specific land-cover types and the number of disjoint training and test samples.

provided by IEEE Geoscience and Remote Sensing Society in 2013 as part of its Data Fusion Contest. The dataset comprises an HSI, an MS image and a LiDAR image. All the images are made up of 340×1905 pixels and the HSI has 144 bands, whereas the MS image has 8 spectral bands. This dataset has a spatial resolution of 2.5 meters per pixel and wavelength ranging between $0.38 - 1.05$ μm. The ground truth has 15 different land-cover and land-use classes. Furthermore, the samples of the 15 land-cover classes are divided into fixed size training and testing samples. Fig. 4 represents the 15 different categories of land-cover and land-use and the training and test samples associated with each.

• The **MUUFL Gulfport** scene was collected over the campus of the University of Southern Mississippi in November 2010 using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor [61], [62]. There are 325×220 pixels with 72 spectral bands in the HSI of this dataset. The LiDAR image of this dataset contains elevation data of 2 rasters. The 8 initial and final bands were removed due to noise, giving a total of 64 bands. The data depicts 11 urban land-cover classes containing 53687 ground truth pixels. Fig. 5 displays the distribution of 5% of the samples randomly selected from every class.

• AISA Eagle sensors were used to collect HSI data over rural regions in the south of **Trento**, Italy, where the Optech ALTM 3100EA sensors collected LiDAR data. There are 63 bands in each HSI with wavelength ranging from $0.42-0.99 \mu m$, and 1 raster in the LiDAR data that provides elevation information. The spectral resolution is 9.2 nm, and the



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

Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	68817	20496		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	13	256

Fig. 5: MUUFL scene. (a) True-color image from the HSI data using bands 40, 20, and 10, respectively (b) Grayscale image from the LiDAR data and (c) Ground truth of MUUFL scene. The table represents class-specific land-cover types and the number of randomly selected 5% training and the remaining 95% test samples.

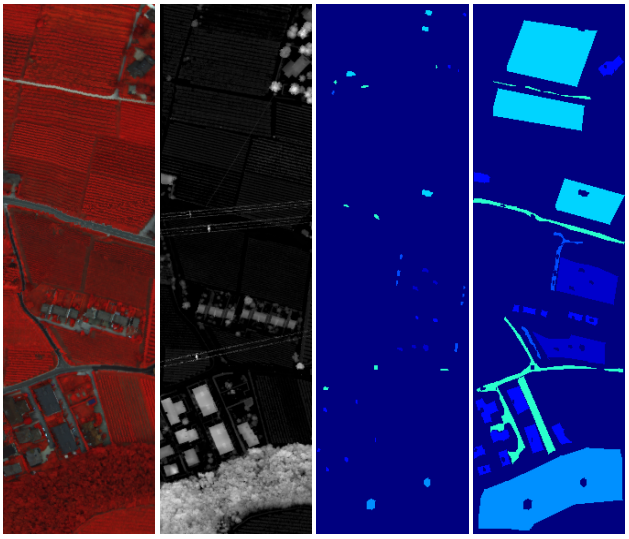
spatial resolution is 1 meters per pixel. The scene comprises 6 vegetation land-cover classes that are mutually exclusive and a pixel count of 600×166. Moreover, the training and test samples are disjoint. Fig. 6 provides information about the training and testing samples for each class.

• There are three types of data in **Augsburg scene** which include an HSI, a dual-Pol SAR image, and a DSM image [63]. SAR data are collected from the Sentinel-1 platform, while HS and DSM data are captured by DAS-EOC, DLR over the city of Augsburg, Germany. The collection is done by the HySpex sensor [64], the Sentinel-1 sensor, and the DLR-3 K system [65], respectively. The spatial resolutions of all images are down-sampled to a unified spatial resolution of 30 m ground sampling distance (GSD) for adequately managing the multimodal fusion. For the HSI, there are 332×485 pixels and 180 spectral bands ranging between $0.4 - 2.5$ μm. The DSM image has a single band, whereas the SAR image has 4 bands. The four bands indicate VV intensity, VH intensity, the real component, and the imaginary component of the PolSAR covariance matrix's off-diagonal element. There are 15 distinct land-cover classes in the ground truth. Fig. 7 shows detailed information on the train and test sets.

A class-specific summary of the UH, MUUFL, and Trento scenes is shown in Figs. 4, 5, 6 and 7, respectively. Each dataset includes its corresponding ground truth, the type associated with the land-cover classes, and the number of available labeled samples per class.

B. Experimental Setup

In order to study the effectiveness of the proposed multimodal fusion transformer model, we have performed extensive experiments and compared it with traditional as well as state-of-the-art methods. The compared methods include traditional classifiers i.e., KNN [66], RF [1], and SVM [27], as well as classical CNN models i.e., CNN1D [33], CNN2D [34], CNN3D [35] and RNN [67]. We also considered a few state-of-the-art transformer models, i.e., ViT [49] and Spectral-



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

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: Trento data. (a) True color image from the HSI using bands 40, 20, and 10, respectively. (b) Grayscale image from the LiDAR data. (c) Ground truth of disjoint training samples. (d) Ground truth of disjoint test samples. The table represents class-specific land-cover types and the number of disjoint training and test samples.

Former [22]. The experiments have been performed using both HSI data and other sources of multimodal data i.e., LiDAR, MSI, SAR and DSM, combined. It can be noted that, for other compared models, the HSI and the other complementary data were concatenated band-wise before feeding the data to the models.

Configuration: All the tests have been performed on a Red Hat Enterprise Server (Release 7.6) with a CPU having **ppc64le** architecture and a total of 40 cores with 4 threads per core and 377 GB of RAM. The GPU used is a single **Nvidia Tesla V100** with 32510 MB of VRAM. The number of HSI patch tokens (n) that is received from the tokenization operation is set to 4 in our experiments.

A **batch size** of 64 and 500 has been used for training and testing the performance of the considered models, where patches of size $11 \times 11 \times B$ are extracted from the HSI and $11 \times 11 \times C$ from other sources of multimodal data. All the models except KNN, RF, SVM and RNN have been trained with **Adam** optimizer [68], [69] with **learning rate** set to $5e^{-4}$ and **weight decay** of $5e^{-3}$. For the RNN, no weight decay was used and a higher learning rate of $1e^{-3}$ was adopted. These models (including the RNN) also utilized a **step scheduler** with step size $=$ 50 and gamma $=$ 0.9, while training has been conducted using 500 **epochs**. Each experiment has been repeated 3 times and the average and standard deviations are reported. The source code of the proposed multimodal fusion transformer was implemented using PyTorch 1.5.0 and Python 3.7.7. The parameters and computational complexities for the considered models with respect to the dataset are shown in Fig. 16.

Evaluation Matrix: Several widely used quantitative measures, such as overall accuracy (OA), average accuracy (AA) and statistical Kappa (κ) coefficients, are used to assess the performance of the proposed network and compare it with other methods. OA represents the proportion of correctly classified test samples versus all test samples, whereas AA represents the average of class-wise accuracy. The kappa value reflects the degree of agreement between the generated classification maps of the considered model and the provided ground truth.

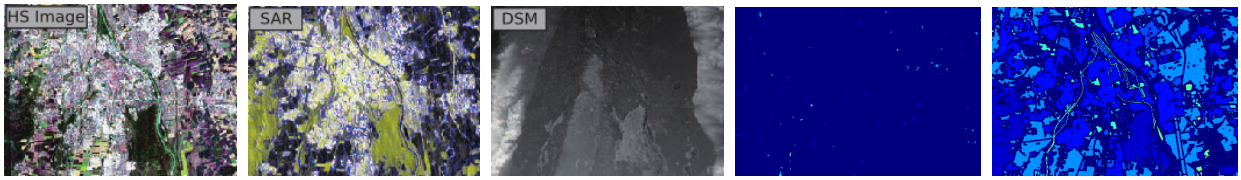
The experimental process has been conducted in three different settings, 1) Disjointed (spatially and spectrally) training and test samples, i.e., the intersection between the training samples and testing samples remains empty. 2) Different percentages of training samples have been randomly selected and used to validate the performance of the proposed network. 3) Different variations of the proposed model have been compared on the same disjointed datasets.

C. Performance Evaluation with Disjoint Train/Test Samples

Table I, II and III reports the quantitative OAs, AAs, κs, and each class accuracy using the proposed model as well as the other methods for the HSI, the HSI with LiDAR data, and HSI with MS data, respectively, for the disjoint Houston dataset. The best results are shown in bold. The evaluation data indicate that the proposed method outperforms the rest by obtaining the highest OAs, AAs, and κs and outperforms others in most class-wise classification accuracies.

Overall, the transformer networks outperform the conventional networks and classifiers. The proposed model dominates the class-wise accuracies and surpasses the other methods in terms of OA, AA, and Kappa in all three cases for the Houston dataset, i.e., HSI, HSI with LiDAR, and HSI with MS image. Combining HSI data with LiDAR data, even using the traditional fusion method, leads to performance gains in almost all models. Though the proposed model stands out on top in the case of Houston using HSI with LiDAR data, the increase in performance (1.35% OA, 1.46% AA, and 1.47% κ) is not significantly better than that achieved by the other models. The new transformer based fusion technique used in the proposed model provides the best result for Houston using HSI data with MS image, where almost all the other models that use the traditional fusion method have lower OAs, AAs, and κs when compared with those in the case Houston using HSI data only, but the gains obtained by the proposed model remain consistent.

It can be seen from Table I that, among the conventional classifiers, RF performs the best, having mean OA, AA and κ 74.87%, 77.94% and 72.93% with standard deviations of 0.09%, 0.14% and 0.09%, respectively. RF can also beat all the classical networks shown in the table but is not as good when compared with transformer networks. The proposed transformer network provides the best performance among the three transformer-based networks giving mean OA, AA, and κ 88.45%, 90.05% and 87.46% with standard deviations of 0.28%, 0.14% and 0.09% respectively, followed by ViT and SpectralFormer. The standard deviations of the proposed



(a) HS image (b) SAR Image (c) DSM Image (d) Training Image (e) Test Image

Color	Land cover	Train	Test	Color	Land cover	Train	Test
	Background	68816	20497		Forest	1162	22084
	Residential-Area	344	6538		Industrial-Area	91	1735
	Low-Plants	334	6353		Allotment	23	443
	Commercial-Area	112	2121		Water	312	5928

Fig. 7: Visualization of the Augsburg scene. (a) True-color image for the HS image data over bands 40, 20, and 10, respectively (b) Grayscale image for the LiDAR data and (c) Ground truth of Augsburg scene. The table represents class-specific land-cover types and the number of randomly selected 5% training and remaining 95% test samples.

TABLE I: OA, AA and Kappa values on the University of Houston dataset (in %) by considering HSI data only.

Class No.	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	77.87	82.81 ± 00.08	79.77	81.10 ± 00.43	80.53 ± 00.23	81.70 ± 00.38	80.22 ± 02.80	82.40 ± 00.39	82.49 ± 00.25	**82.53 ± 00.13**
2	77.44	82.86 ± 00.36	82.42	80.23 ± 01.51	83.90 ± 00.09	80.55 ± 01.84	78.51 ± 00.19	80.29 ± 00.79	**89.13 ± 06.36**	85.03 ± 00.04
3	96.83	63.10 ± 01.71	59.41	53.73 ± 00.09	57.49 ± 02.10	96.57 ± 00.09	52.94 ± 17.72	97.43 ± 01.32	69.77 ± 12.29	**98.55 ± 01.08**
4	75.28	91.95 ± 00.15	83.81	83.74 ± 00.71	89.46 ± 00.24	78.54 ± 02.82	83.81 ± 07.71	90.40 ± 00.25	91.73 ± 03.42	**94.60 ± 02.53**
5	90.72	99.78 ± 00.12	95.27	87.06 ± 01.51	92.36 ± 01.20	98.48 ± 00.28	85.61 ± 09.73	99.24 ± 00.13	96.78 ± 00.76	**99.84 ± 00.16**
6	66.43	96.97 ± 00.33	67.13	52.45 ± 01.14	64.10 ± 02.38	73.89 ± 00.66	70.16 ± 07.18	91.38 ± 00.87	85.31 ± 07.99	**95.10 ± 00.57**
7	76.96	85.23 ± 00.50	83.21	71.42 ± 01.46	71.39 ± 04.04	82.77 ± 01.45	73.01 ± 01.31	**86.10 ± 01.54**	80.25 ± 02.23	85.79 ± 00.94
8	30.96	42.58 ± 00.36	29.53	41.12 ± 01.70	44.95 ± 06.09	38.30 ± 01.09	43.84 ± 08.75	73.95 ± 00.29	62.74 ± 11.68	**79.87 ± 09.29**
9	69.50	85.36 ± 00.13	75.45	60.25 ± 00.08	62.45 ± 02.04	65.94 ± 02.86	68.84 ± 02.54	85.33 ± 02.24	70.57 ± 01.62	**89.05 ± 02.47**
10	42.95	35.81 ± 00.70	46.62	39.12 ± 02.03	49.94 ± 01.66	43.28 ± 07.46	37.52 ± 01.15	50.42 ± 05.59	48.17 ± 05.04	**62.45 ± 02.95**
11	56.17	63.03 ± 00.39	45.07	42.06 ± 00.78	44.53 ± 00.78	33.59 ± 02.72	49.65 ± 09.61	80.80 ± 03.06	62.75 ± 04.12	**98.04 ± 01.45**
12	75.79	66.63 ± 00.50	70.03	62.98 ± 04.58	53.92 ± 06.25	67.85 ± 02.10	64.07 ± 00.28	81.91 ± 01.77	79.09 ± 02.16	**95.84 ± 00.48**
13	60.35	87.60 ± 00.17	68.42	42.11 ± 01.52	47.13 ± 02.33	77.54 ± 01.98	53.92 ± 08.72	89.47 ± 01.74	63.63 ± 05.56	**91.81 ± 00.33**
14	76.92	99.73 ± 00.19	75.30	83.94 ± 00.38	82.46 ± 03.07	92.58 ± 01.06	81.38 ± 12.63	99.33 ± 00.95	93.66 ± 04.30	**99.73 ± 00.19**
15	88.37	85.62 ± 01.38	49.89	34.46 ± 02.04	42.92 ± 05.11	93.52 ± 01.55	44.12 ± 13.23	**99.72 ± 00.26**	77.24 ± 07.34	92.60 ± 02.42
OA	69.48	74.87 ± 00.09	68.13	63.04 ± 00.08	65.85 ± 00.09	70.26 ± 00.14	65.20 ± 02.74	83.23 ± 00.47	76.35 ± 01.97	**88.45 ± 00.28**
AA	70.84	77.94 ± 00.14	67.42	61.05 ± 00.10	64.50 ± 00.11	73.67 ± 00.17	64.51 ± 02.77	85.88 ± 00.33	76.89 ± 02.31	**90.05 ± 00.29**
$\kappa(\times 100)$	67.08	72.93 ± 00.09	65.56	60.01 ± 00.08	63.04 ± 00.09	67.91 ± 00.16	62.43 ± 02.87	81.88 ± 00.51	74.42 ± 02.14	**87.46 ± 00.30**

TABLE II: OA, AA and Kappa values on the University of Houston dataset (in %) by considering HSI and LiDAR data.

Class No.	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
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**	82.34 ± 00.47
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	**88.78 ± 05.27**
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	**98.15 ± 01.08**
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	**94.35 ± 02.40**
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	99.12 ± 01.25
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	**99.30 ± 00.00**
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	**88.56 ± 01.16**
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	**86.89 ± 05.33**
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	**87.91 ± 03.90**
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	**64.70 ± 00.92**
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	**98.64 ± 00.70**
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	**94.24 ± 02.19**
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	**90.29 ± 00.66**
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	99.73 ± 00.38
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	99.58 ± 00.17
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	**89.80 ± 00.53**
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	**91.51 ± 00.40**
$\kappa(\times 100)$	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	**88.93 ± 00.59**

model are also lower than those of the other transformer networks, since our model can represent long-range dependencies between the feature maps better than ViT and SpectralFormer. Conventional classifiers, which rely completely on the spectral representation of HSI data, cannot capture the spatial information and hence perform worse than transformer networks.

Table II indicates a substantial rise in the accuracies of all the given classical networks, where the accuracies of the best-performing classical networks are close to those of the best-performing conventional classifiers. Table I confirms that RF outperforms the conventional classifiers as well as the classical networks due to its better discriminative feature learning, but still cannot compete with the transformer networks. The proposed model again beats the rest by quite a margin due to its superior learning of spectral and spatial information, with mean OA, AA, and κ 89.80%, 91.51% and 88.93% with minimal standard deviations of 0.53%, 0.40% and 0.59% respectively.

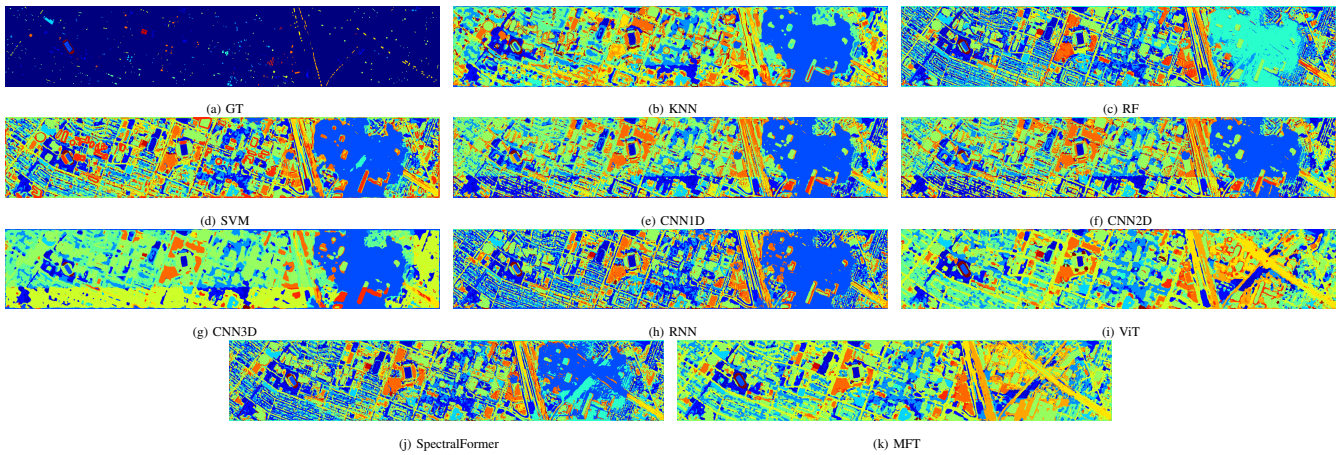


Fig. 8: (a) Ground truth and classification maps obtained for the UH data set by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) RNN, (h) ViT, (i) SpectralFormer, and (j) MFT using disjoint training samples.

TABLE III: OA, AA and Kappa values on the University of Houston dataset (in %) by considering HSI and MS image data.

Class	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
No.	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	78.06	78.82 ± 00.40	79.87	84.52 ± 02.26	88.86 ± 04.90	82.15 ± 00.68	79.23 ± 00.39	**89.02 ± 05.03**	81.54 ± 00.65	82.72 ± 00.41
2	75.28	72.65 ± 02.53	79.14	81.05 ± 02.70	**86.18 ± 05.21**	84.62 ± 00.04	81.42 ± 00.51	82.11 ± 00.67	80.20 ± 02.86	85.09 ± 00.09
3	96.63	96.63 ± 00.49	61.25	38.75 ± 04.96	45.35 ± 03.93	49.70 ± 04.53	38.75 ± 01.22	**99.21 ± 00.16**	59.27 ± 11.24	98.55 ± 00.76
4	68.47	77.15 ± 00.27	85.23	81.09 ± 03.63	75.47 ± 06.68	86.68 ± 00.50	88.35 ± 01.54	81.72 ± 01.08	90.59 ± 00.31	**95.99 ± 02.39**
5	89.77	90.03 ± 00.89	87.22	72.92 ± 04.99	87.12 ± 04.46	92.20 ± 03.75	90.91 ± 01.01	93.15 ± 00.97	95.01 ± 00.96	**99.78 ± 00.09**
6	71.33	86.48 ± 06.49	61.54	63.64 ± 02.97	72.03 ± 04.67	79.95 ± 02.87	77.39 ± 02.87	**99.77 ± 00.33**	73.43 ± 02.06	97.20 ± 01.98
7	72.57	77.89 ± 03.73	81.53	58.18 ± 01.21	68.13 ± 05.26	78.17 ± 04.71	65.76 ± 03.53	72.45 ± 05.50	78.42 ± 00.62	**86.32 ± 00.69**
8	30.10	41.63 ± 00.25	18.80	26.34 ± 16.50	23.36 ± 15.68	27.95 ± 06.36	35.61 ± 01.58	77.87 ± 02.72	48.40 ± 01.57	**81.16 ± 08.04**
9	68.18	74.66 ± 00.40	68.27	54.93 ± 06.23	59.05 ± 04.45	83.41 ± 00.76	72.05 ± 03.16	83.47 ± 07.68	73.25 ± 01.69	**87.76 ± 00.74**
10	42.28	38.42 ± 00.21	48.46	28.73 ± 04.69	32.37 ± 01.22	69.76 ± 04.74	34.04 ± 05.16	51.03 ± 01.04	50.87 ± 01.76	**74.71 ± 15.98**
11	53.80	58.16 ± 00.54	38.71	42.50 ± 01.88	43.33 ± 01.90	60.25 ± 04.92	36.78 ± 02.50	76.03 ± 01.53	55.03 ± 00.43	**93.71 ± 06.01**
12	68.68	80.24 ± 00.09	57.54	43.84 ± 13.82	35.57 ± 11.98	65.10 ± 15.37	58.44 ± 03.47	75.47 ± 02.24	71.41 ± 03.96	**96.16 ± 00.87**
13	57.89	66.08 ± 01.47	65.96	48.65 ± 04.63	52.28 ± 04.22	47.49 ± 10.75	52.63 ± 06.82	83.63 ± 04.12	57.08 ± 00.17	**92.51 ± 00.17**
14	79.35	80.30 ± 01.34	83.40	63.16 ± 02.17	70.58 ± 12.48	96.22 ± 03.34	81.24 ± 07.20	99.60 ± 00.33	88.26 ± 08.02	**100.0 ± 00.00**
15	86.68	**94.15 ± 00.20**	48.41	55.60 ± 02.85	45.24 ± 09.15	58.42 ± 08.47	40.94 ± 15.84	91.97 ± 01.70	46.51 ± 05.61	86.82 ± 08.99
OA	67.16	71.51 ± 00.46	64.18	56.63 ± 00.61	59.06 ± 00.57	71.49 ± 01.07	62.61 ± 01.80	80.49 ± 01.19	70.95 ± 00.58	**89.15 ± 00.96**
AA	69.27	74.22 ± 00.06	64.36	56.26 ± 00.93	58.99 ± 00.25	70.80 ± 01.16	62.24 ± 02.51	83.77 ± 01.15	69.95 ± 00.63	**90.56 ± 00.93**
$\kappa(\times 100)$	64.61	69.30 ± 00.42	61.30	53.07 ± 00.68	55.70 ± 00.60	69.02 ± 01.16	59.64 ± 01.98	78.88 ± 01.29	68.59 ± 00.64	**88.22 ± 01.04**

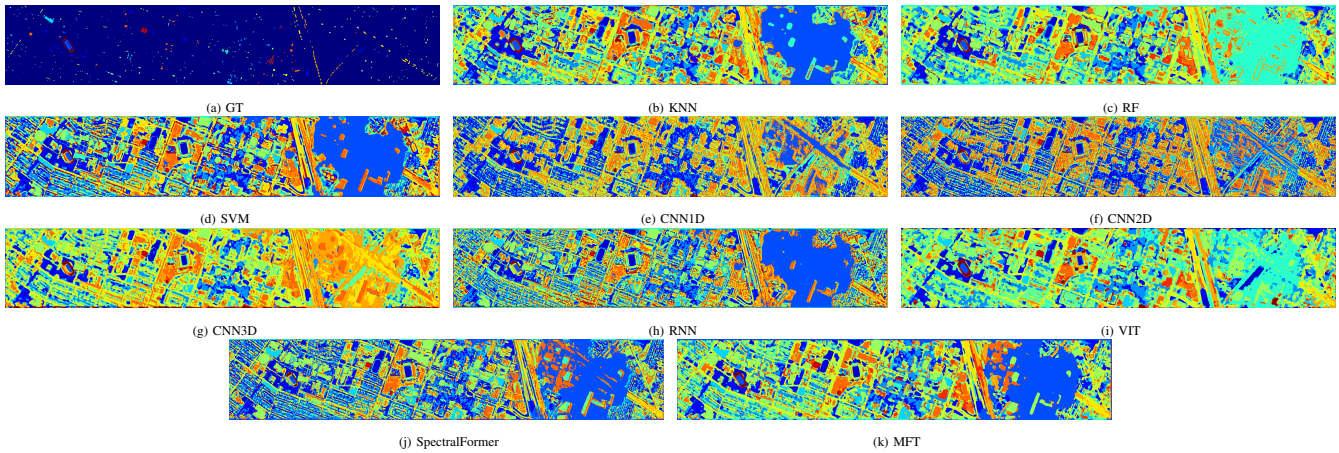


Fig. 9: (a) Ground truth and classification maps obtained by HSI and LiDAR data for the UH dataset by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) RNN, (h) ViT, (i) SpectralFormer, and (j) MFT using disjoint training samples.

Table III reveals an opposite trend than Table II in all but two networks. The networks other than CNN3D and the proposed network perform worse with the fusion of both HSI and MS image using the traditional fusion method when compared to that of HSI only (Table I). Unsurprisingly, the novel transformer-based fusion technique allows the proposed model to increase its performance resulting in mean OA, AA, and κ of 89.15%, 90.56%, and 88.22% with acceptable standard deviations of 0.96%, 0.93% and 1.04% respectively.

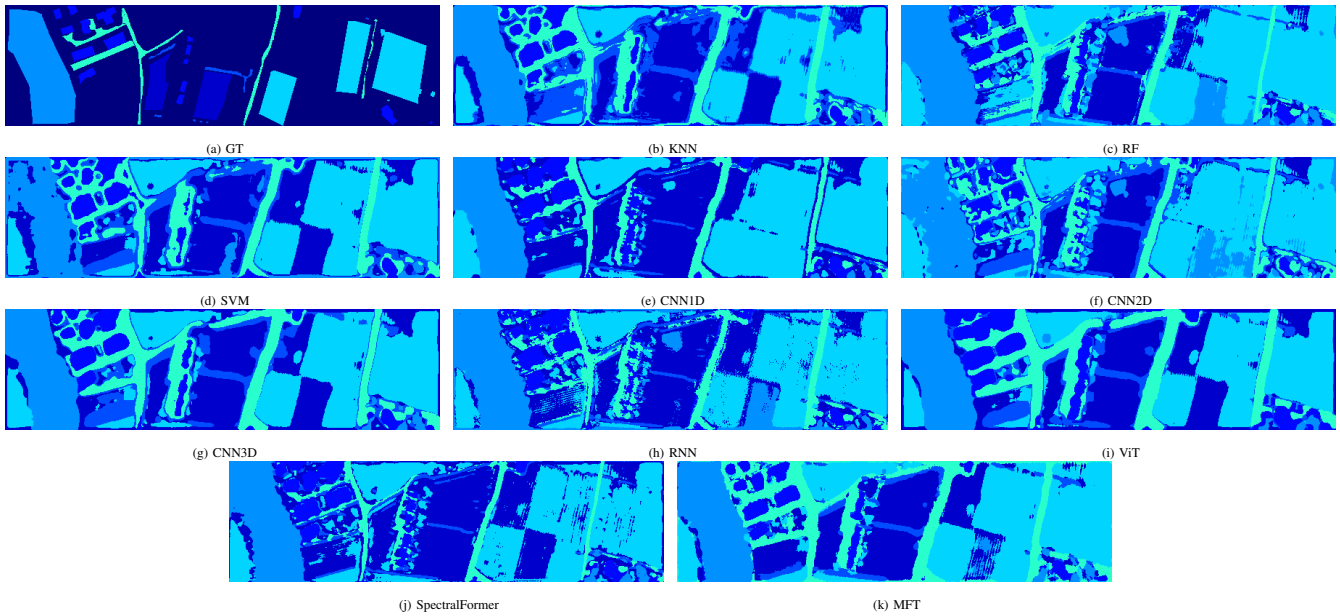
To demonstrate the generalization ability of the proposed model for remote sensing image classification methods, we have considered Trento and MUUFL datasets (HSI data only and HSI with LiDAR data) with disjoint training and test samples. Fig. 6 shows the pseudo color map, corresponding LiDAR data, and the disjoint train and test ground truth for the Trento dataset. Fig. 6 also provides 6 different land-cover

TABLE IV: OA, AA, and Kappa values on the University of Trento dataset (in %) by considering HS image only.

Class No.	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	89.07	96.65 ± 00.36	97.21	97.29 ± 00.20	96.76 ± 00.24	92.22 ± 00.66	88.06 ± 04.36	89.07 ± 00.86	94.01 ± 03.77	**98.20 ± 00.64**
2	88.59	89.38 ± 00.74	94.17	86.73 ± 01.60	84.17 ± 01.05	91.42 ± 00.69	79.11 ± 06.15	84.74 ± 06.22	91.26 ± 01.23	**91.65 ± 01.43**
3	87.17	73.71 ± 01.61	56.15	50.62 ± 00.13	50.53 ± 02.84	**96.88 ± 00.91**	39.22 ± 19.36	92.25 ± 01.65	46.52 ± 05.46	94.92 ± 00.95
4	80.24	99.92 ± 00.02	84.78	99.17 ± 00.14	96.12 ± 00.24	99.62 ± 00.10	83.31 ± 04.10	**99.63 ± 00.21**	85.48 ± 11.89	99.45 ± 00.25
5	95.15	99.96 ± 00.01	98.13	99.18 ± 00.02	98.76 ± 00.16	98.99 ± 00.12	97.96 ± 00.64	98.23 ± 00.10	97.60 ± 01.45	**99.72 ± 00.10**
6	69.59	66.73 ± 01.52	55.05	59.62 ± 01.47	66.19 ± 00.90	85.47 ± 00.62	70.87 ± 15.09	84.04 ± 06.23	61.42 ± 01.79	**91.52 ± 01.72**
OA	86.42	94.73 ± 00.14	88.55	93.02 ± 00.06	92.31 ± 00.05	96.14 ± 00.02	86.83 ± 01.91	94.62 ± 00.21	88.42 ± 03.45	**97.76 ± 00.40**
AA	84.97	87.73 ± 00.43	80.91	82.10 ± 00.08	82.09 ± 00.46	94.10 ± 00.14	76.42 ± 04.46	91.33 ± 00.22	79.38 ± 00.96	**95.91 ± 00.41**
$\kappa(\times 100)$	82.21	92.92 ± 00.18	84.83	90.65 ± 00.09	89.69 ± 00.07	94.83 ± 00.02	82.49 ± 02.53	92.81 ± 00.28	84.68 ± 04.41	**97.00 ± 00.53**

TABLE V: OA, AA, and Kappa values on the University of Trento dataset (in %) by considering HS image and LiDAR data.

Class No.	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
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	**98.23 ± 00.38**
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	99.34 ± 00.02
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	89.84 ± 09.00
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	99.82 ± 00.26
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	**99.93 ± 00.05**
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	**88.72 ± 00.94**
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	**98.32 ± 00.25**
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	**95.98 ± 01.64**
$\kappa(\times 100)$	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	**97.75 ± 00.33**



(a) GT (b) KNN (c) RF (d) SVM (e) CNN1D (f) CNN2D (g) CNN3D (h) RNN (i) ViT (j) SpectralFormer (k) MFT

Fig. 10: (a) Ground truth and classification maps obtained by HSI and LiDAR data for the Trento data set by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) RNN, (h) ViT, (i) SpectralFormer, and (j) MFT using disjoint training samples.

TABLE VI: OA, AA, and Kappa values on the MUUFL dataset (in %) by considering HS image data only.

Class No.	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	92.74	97.97 ± 00.07	96.64	95.21 ± 00.05	95.94 ± 00.20	94.72 ± 00.65	96.06 ± 00.19	97.40 ± 00.25	94.63 ± 00.89	**97.61 ± 00.15**
2	47.46	77.86 ± 00.19	59.37	69.31 ± 01.06	70.87 ± 01.50	62.98 ± 03.76	80.34 ± 02.85	77.34 ± 01.62	77.12 ± 03.77	**92.51 ± 00.87**
3	69.10	84.33 ± 00.23	81.49	74.31 ± 00.60	78.65 ± 00.78	71.15 ± 03.46	80.65 ± 00.87	86.10 ± 01.49	74.26 ± 02.81	**92.12 ± 01.66**
4	53.43	86.09 ± 00.37	74.35	78.71 ± 01.27	82.96 ± 01.17	64.86 ± 02.15	86.09 ± 01.95	**93.35 ± 00.84**	84.59 ± 01.18	92.83 ± 01.30
5	83.68	92.10 ± 00.08	83.80	75.84 ± 01.93	78.76 ± 01.09	77.95 ± 02.15	90.15 ± 00.39	**94.82 ± 00.77**	89.38 ± 02.06	94.31 ± 01.26
6	01.13	68.77 ± 00.65	15.35	46.35 ± 00.74	48.53 ± 02.60	75.24 ± 90.92	53.42 ± 04.33	82.69 ± 01.31	60.42 ± 09.08	**88.56 ± 02.86**
7	43.33	80.47 ± 00.42	77.09	78.45 ± 00.58	78.96 ± 00.30	48.44 ± 01.87	80.39 ± 02.75	87.44 ± 00.97	83.77 ± 03.58	**92.68 ± 01.42**
8	73.48	91.57 ± 00.17	87.11	68.28 ± 02.36	68.86 ± 00.37	69.48 ± 01.91	88.67 ± 02.04	**97.41 ± 00.75**	88.91 ± 02.78	97.08 ± 00.87
9	04.94	45.87 ± 00.31	21.50	39.97 ± 00.27	45.64 ± 00.48	04.79 ± 03.39	58.94 ± 00.73	58.92 ± 07.76	55.02 ± 01.43	**59.80 ± 01.56**
10	00.00	04.98 ± 03.05	00.00	08.81 ± 00.72	13.03 ± 00.27	00.00 ± 00.00	22.80 ± 04.51	**32.76 ± 07.01**	17.62 ± 02.12	12.45 ± 04.79
11	58.59	48.31 ± 02.12	61.33	23.83 ± 03.24	24.74 ± 00.97	61.20 ± 02.17	**80.99 ± 04.01**	66.67 ± 04.94	45.44 ± 07.29	71.09 ± 01.66
OA	75.80	89.85 ± 00.03	84.30	81.17 ± 00.09	82.95 ± 00.12	77.59 ± 00.05	88.60 ± 00.69	91.99 ± 00.35	86.68 ± 00.93	**94.18 ± 00.09**
AA	47.99	70.76 ± 00.29	59.82	59.92 ± 00.52	62.45 ± 00.22	50.58 ± 00.34	74.41 ± 01.40	79.54 ± 01.93	70.11 ± 01.52	**81.00 ± 00.55**
$\kappa(\times 100)$	67.34	86.44 ± 00.04	78.89	74.95 ± 00.12	77.30 ± 00.16	69.81 ± 00.10	84.90 ± 00.92	89.37 ± 00.46	82.43 ± 01.19	**92.30 ± 00.12**

classes and the number of samples belonging to the train and test set for each class in a table. Fig. 5 shows the pseudo color map corresponding to the LiDAR image, and the disjoint train and test ground truth for the MUUFL dataset. It also gives the 11 different land-cover classes and the number of training and testing samples per class.

TABLE VII: OA, AA and Kappa values on the MUUFL dataset (in %) by considering HS image and LiDAR data.

Class	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
No.	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
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	**97.90 ± 00.39**
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	**92.11 ± 01.58**
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	**91.80 ± 00.82**
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	91.59 ± 02.25
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	**95.60 ± 01.21**
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	**88.19 ± 03.49**
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	**90.27 ± 02.13**
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	97.26 ± 00.53
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	**61.35 ± 03.80**
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	17.43 ± 04.63
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	72.79 ± 09.25
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	**94.34 ± 00.07**
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	**81.48 ± 00.70**
$\kappa(\times 100)$	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	**92.51 ± 00.10**

Tables IV and VI show the experimental results for Trento and MUUFL datasets using HS images only. Among conventional classifiers and classical networks, CNN3D performs the best for Trento. In the case of MUUFL, both RF and RNN perform similarly and better than the other conventional classifiers and classical networks. In the case of transformer-based models, the proposed model takes the lead with OA =$97.76\pm00.40\%$, AA =$95.91\pm00.41\%$ and κ =$97.00\pm00.53\%$ for Trento and OA =$94.18 \pm 00.09\%$, AA =$81.00 \pm 00.55\%$ and κ =$92.30 \pm 00.12\%$ for MUUFL.

Table IV and IV show the experimental results for Trento and MUUFL datasets using combined HS image and LiDAR data. In Trento, except for RF, it is quite evident from the better accuracies of all the models that using LiDAR and HS images improves all models' performance. But, even with this significant increase in performance, the proposed model (in Tables IV and VI) performs better than the rest of the models while the proposed model (in Tables IV and IV) performs even better. For the Trento data using HS image and LiDAR, the gain in the proposed model's OA is 0.56%, AA is 0.07%, and κ is 0.75%. For MUUFL using HS image and LiDAR data, the gain in the proposed model's OA, AA, and κ are 0.26%, 0.48%, and 0.21%, respectively. Among conventional classifiers and classical networks, SVM and CNN3D perform the best for Trento, while MUUFL, RF and CNN2D perform the best.

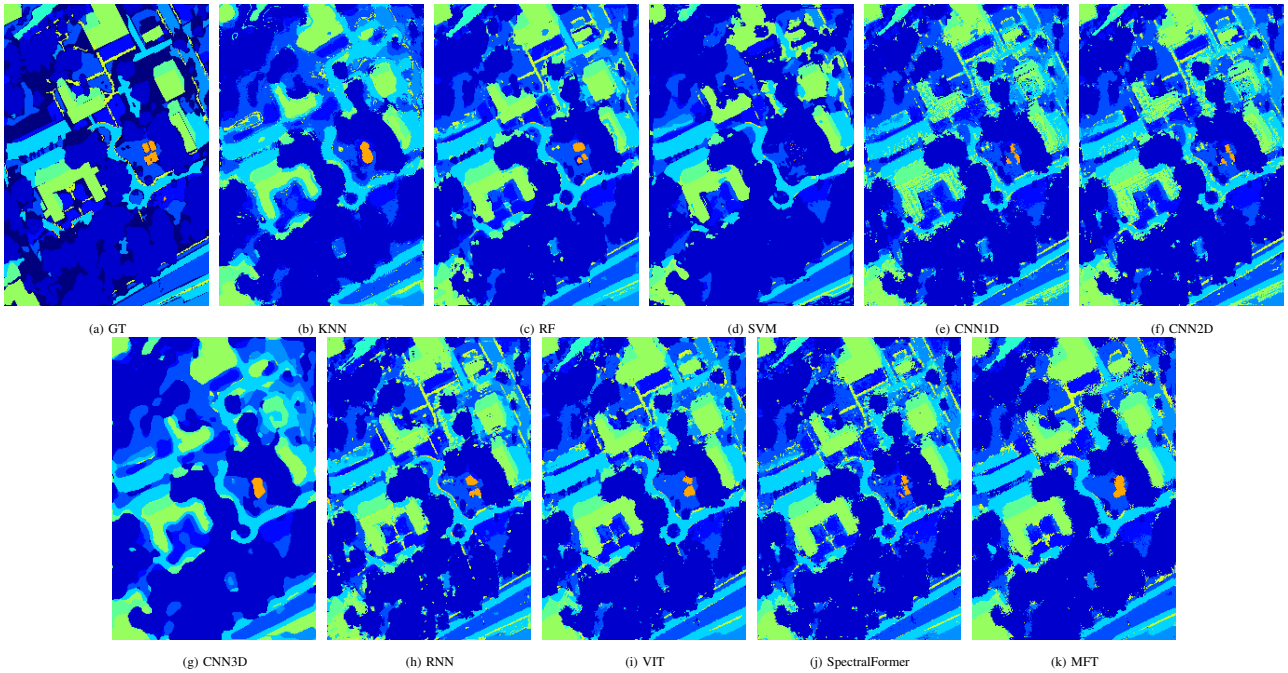
The proposed MFT is not only limited to HS image, MS image, and LiDAR data but can also work with SAR and DSM data. To illustrate this capability of the proposed MFT, we have considered the Augsburg dataset (using HS image data only, HS image and SAR data, and HS image and DSM data) with disjoint training and test samples. Fig. 7 shows the pseudo color map, corresponding SAR data, corresponding DSM data, and the disjoint train and test ground truth for the Augsburg dataset. It also provides the 7 different land-cover classes and the number of samples belonging to the train and test set per class.

After observing the AA in Table VIII, it is evident that Augsburg is a considerably harder dataset than the previous ones. Consequently, the gain in performance in most models with fusion is not significant. However, the proposed model gives the highest class-wise classification accuracies for most classes in all three cases. It also provides the highest OAs, AAs, and κs with the exception in the case of the Augsburg dataset using HS and SAR data. Table IX indicates that ViT has 0.76% more AA than the proposed model. However, the proposed MFT has 4.39% and 6.20% more OA and κ than ViT, respectively. In Table X, the AA of the proposed MFT reaches $64.70 \pm 00.44\%$ which is the highest recorded for all three cases in all models.

RNN performs the worst in Tables VIII and X but gets a significant boost in performance in Table IX. Comparing Tables IX and X, conventional classifiers like KNN and SVM show barely noticeable changes in performance. On the other hand, RF gives higher accuracy using SAR data than DSM data. In the case of classical networks, CNN1D performs better-using DSM, CNN2D and RNN perform better using SAR, and CNN3D performs similarly in both cases.

D. Visual Comparison

We make a qualitative evaluation by visualizing the classification maps obtained by different methods. Figs. 8, 9, 10, 11, 12, and 13 depict the obtained classification maps for the hyperspectral datasets i.e., Houston, Trento, MUUFL and Augsburg along with their corresponding other modalities like LiDAR, MS image, SAR and DSM data, respectively. Although the conventional classifiers, i.e., KNN, RF, and SVM, produce rich classification maps, the maps still have salt and pepper noise in the boundary regions of the land-cover classes since they only rely on the spectral information from HS image data. The imperfections in the generated classification maps can be slightly reduced by considering the complementary information from other sources of modalities. However, the DL models successfully capture the nonlinear relationship between the input and output feature maps due to their learnable feature extraction capabilities. Hence, CNN1D, CNN2D, and CNN3D produce relatively smooth classification maps where land-use and land-cover class boundaries are well separable. ViT was found effective for HS classification due to its ability to extract high-level sequential abstract representations from HS data resulting. Therefore, classification maps are of better visual quality compared to classical networks. By enhancing the neighboring spectral information and conveying positional information across layers more effectively, the proposed MFT obtains highly desirable classification maps, especially in terms of texture and edge details compared to ViT, SpectralFormer. As observed from the generated classification maps, distinct class boundaries and



(a) GT (b) KNN (c) RF (d) SVM (e) CNN1D (f) CNN2D

(g) CNN3D (h) RNN (i) VIT (j) SpectralFormer (k) MFT

Fig. 11: (a) Ground truth and classification maps obtained by HS image and LiDAR data for the MUUFL data set by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) RNN, (h) ViT, (i) SpectralFormer, and (j) MFT using disjoint training samples.

TABLE VIII: OA, AA and Kappa values on the Augsburg dataset (in %) by considering HS image data only.

Class	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
No.	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	83.27	90.18 ± 00.34	91.51	77.79 ± 04.77	82.77 ± 02.64	78.66 ± 01.26	21.21 ± 29.99	88.62 ± 03.36	83.23 ± 04.20	**93.93 ± 00.62**
2	86.38	97.41 ± 00.14	86.18	87.40 ± 01.55	86.19 ± 01.41	96.10 ± 00.98	**98.99 ± 01.39**	94.90 ± 00.84	86.63 ± 02.81	96.97 ± 00.49
3	36.79	04.59 ± 00.33	10.34	30.30 ± 08.46	56.08 ± 00.99	49.25 ± 23.59	00.00 ± 00.00	68.15 ± 03.93	27.75 ± 05.25	**68.34 ± 01.97**
4	49.72	76.16 ± 00.99	62.75	65.25 ± 01.80	64.95 ± 02.05	84.96 ± 01.88	05.35 ± 07.56	84.40 ± 02.26	60.96 ± 04.83	**93.14 ± 01.80**
5	38.62	18.16 ± 00.62	23.14	15.74 ± 04.72	10.26 ± 02.98	10.13 ± 09.16	00.00 ± 00.00	34.29 ± 02.60	37.54 ± 07.18	**46.02 ± 05.76**
6	08.30	00.12 ± 00.09	00.00	**20.35 ± 13.47**	14.12 ± 08.17	08.45 ± 03.66	00.00 ± 00.00	17.83 ± 05.12	03.24 ± 00.90	14.59 ± 04.29
7	05.37	08.03 ± 00.99	10.68	14.36 ± 02.30	24.62 ± 02.92	11.70 ± 04.50	00.00 ± 00.00	**45.90 ± 00.92**	13.49 ± 08.87	26.45 ± 15.31
OA	67.27	79.96 ± 00.23	71.60	72.00 ± 01.70	73.59 ± 00.59	82.89 ± 00.78	40.26 ± 02.03	85.90 ± 00.26	70.81 ± 01.47	**90.26 ± 00.91**
AA	44.07	42.09 ± 00.21	40.66	44.45 ± 01.41	48.43 ± 01.99	48.46 ± 04.17	14.94 ± 00.86	62.01 ± 01.18	44.69 ± 02.31	**62.78 ± 02.31**
$\kappa(\times 100)$	53.86	70.04 ± 00.33	58.57	59.96 ± 02.40	62.16 ± 01.10	75.08 ± 01.35	02.65 ± 03.60	79.88 ± 00.27	58.05 ± 02.01	**85.98 ± 01.27**

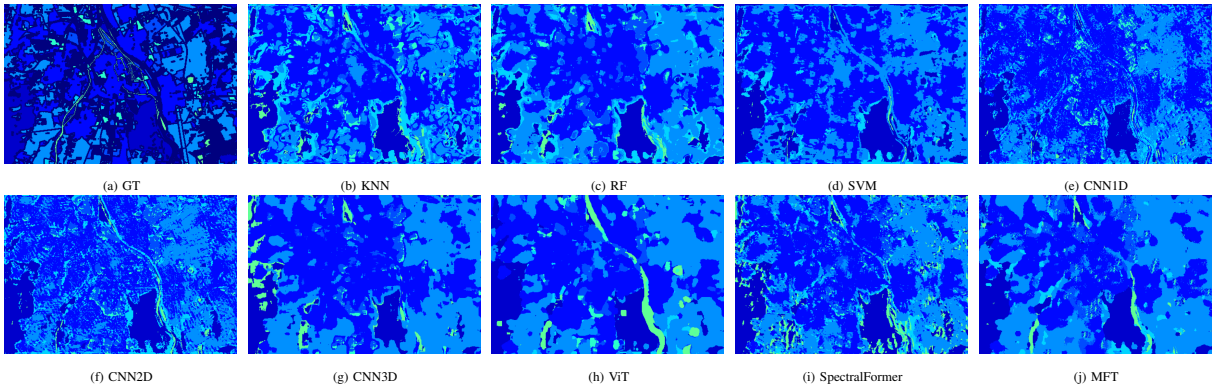
TABLE IX: OA, AA and Kappa values on the Augsburg dataset (in %) by considering HS image and SAR data.

Class	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
No.	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	81.98	87.15 ± 00.77	91.51	81.26 ± 02.95	90.96 ± 01.32	78.21 ± 00.56	87.60 ± 06.19	90.01 ± 00.48	92.04 ± 01.87	**94.65 ± 00.62**
2	86.17	95.82 ± 00.10	86.32	88.32 ± 01.78	82.07 ± 00.74	93.82 ± 00.67	95.26 ± 02.62	94.27 ± 00.78	93.60 ± 02.29	**96.90 ± 00.62**
3	29.66	64.62 ± 02.34	10.03	42.43 ± 14.40	48.31 ± 10.85	63.11 ± 11.45	51.35 ± 72.62	64.62 ± 05.57	43.28 ± 07.08	**69.80 ± 06.43**
4	51.28	84.33 ± 00.80	62.78	67.42 ± 02.97	61.05 ± 04.86	86.24 ± 01.07	91.14 ± 02.98	85.53 ± 00.91	77.08 ± 04.11	**93.98 ± 01.65**
5	36.90	27.72 ± 01.02	22.75	09.88 ± 06.07	18.48 ± 07.96	11.15 ± 08.03	10.52 ± 11.42	29.00 ± 00.63	**45.32 ± 05.54**	32.70 ± 11.81
6	13.80	13.06 ± 02.27	00.00	10.26 ± 05.48	12.11 ± 02.80	09.67 ± 04.92	00.00 ± 00.00	**20.41 ± 01.26**	06.96 ± 01.01	10.52 ± 04.31
7	07.37	11.44 ± 00.39	11.28	09.93 ± 03.22	29.13 ± 01.00	09.42 ± 05.13	24.33 ± 34.41	**44.17 ± 02.06**	29.44 ± 04.77	23.98 ± 07.85
OA	67.29	85.00 ± 00.28	71.65	73.96 ± 01.17	71.79 ± 00.65	83.04 ± 00.47	83.42 ± 01.14	86.10 ± 00.26	81.77 ± 02.30	**90.49 ± 00.20**
AA	43.88	54.88 ± 00.28	40.67	44.21 ± 02.81	48.88 ± 02.61	50.23 ± 01.09	40.75 ± 02.04	**61.14 ± 00.85**	55.39 ± 01.98	60.36 ± 02.36
$\kappa(\times 100)$	54.13	78.62 ± 00.38	58.62	62.20 ± 01.14	59.84 ± 00.63	75.49 ± 00.62	75.26 ± 01.90	80.06 ± 00.33	74.06 ± 03.04	**86.26 ± 00.23**

TABLE X: OA, AA and Kappa values on the Augsburg dataset (in %) by considering HS image and DSM data.

Class	Conventional Classifiers			Classical Convolutional Networks				Transformer Networks		
No.	KNN	RF	SVM	CNN1D	CNN2D	CNN3D	RNN	ViT	SpectralFormer	MFT
1	81.99	84.94 ± 00.97	91.51	87.75 ± 00.49	78.92 ± 01.14	77.41 ± 04.03	00.00 ± 00.00	90.39 ± 00.85	90.12 ± 02.42	**94.77 ± 02.29**
2	86.16	88.48 ± 00.06	86.21	93.69 ± 00.22	85.50 ± 01.49	95.47 ± 00.93	**99.92 ± 00.12**	95.04 ± 01.07	86.98 ± 01.18	96.03 ± 01.84
3	29.66	64.50 ± 02.06	10.23	28.96 ± 04.51	56.41 ± 19.63	61.22 ± 06.55	00.00 ± 00.00	59.49 ± 03.02	37.31 ± 03.93	**72.52 ± 10.54**
4	51.29	73.27 ± 00.55	62.80	89.25 ± 00.62	57.72 ± 02.86	86.15 ± 00.07	16.91 ± 23.92	86.34 ± 01.78	58.17 ± 11.82	**93.52 ± 01.96**
5	36.90	26.83 ± 00.59	22.94	18.61 ± 01.86	10.45 ± 04.10	09.31 ± 02.24	00.00 ± 00.00	21.16 ± 14.48	47.74 ± 08.32	**51.31 ± 05.13**
6	13.80	**17.97 ± 01.99**	00.00	15.85 ± 05.68	08.85 ± 08.48	09.22 ± 03.42	00.00 ± 00.00	15.53 ± 11.50	02.32 ± 00.18	13.49 ± 09.44
7	07.37	10.20 ± 00.17	10.75	23.60 ± 01.14	23.65 ± 04.78	09.67 ± 01.79	00.00 ± 00.00	**44.72 ± 01.77**	20.86 ± 03.16	31.28 ± 10.80
OA	67.29	78.04 ± 00.07	71.62	84.43 ± 00.18	70.07 ± 01.11	83.40 ± 00.76	38.80 ± 00.04	86.34 ± 00.40	71.84 ± 03.87	**90.49 ± 00.54**
AA	43.88	52.31 ± 00.17	40.64	51.10 ± 00.61	45.93 ± 02.28	49.78 ± 00.59	14.30 ± 00.02	58.95 ± 03.61	49.07 ± 01.38	**64.70 ± 00.44**
$\kappa(\times 100)$	54.14	68.91 ± 00.09	58.60	77.21 ± 00.26	57.06 ± 01.91	75.90 ± 01.14	06.22 ± 08.80	80.31 ± 00.68	60.15 ± 04.79	**86.35 ± 00.77**

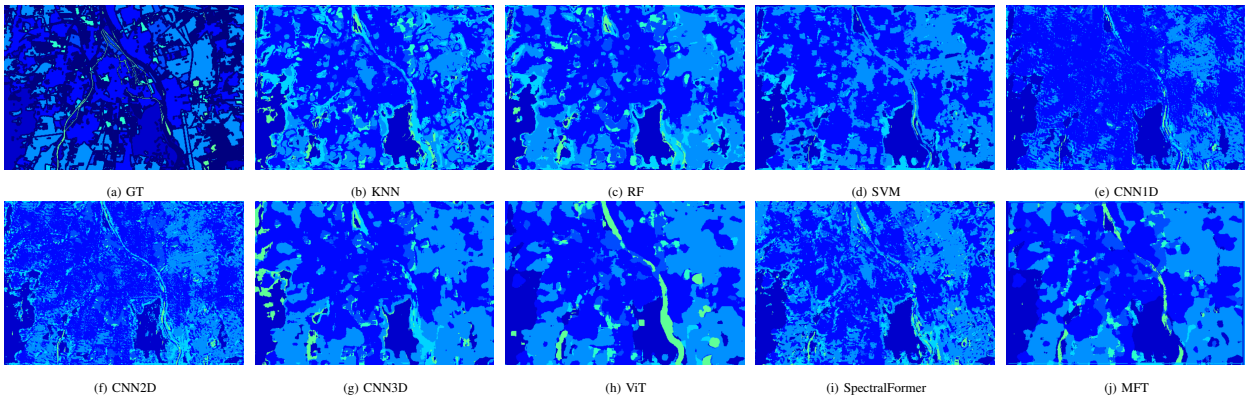
remarkable improvements in the visual quality are obtained by considering the other sources of modalities during feature extraction and learning of all the networks. More noticeable improvements can be seen for the proposed MFT network due



(a) GT (b) KNN (c) RF (d) SVM (e) CNN1D

(f) CNN2D (g) CNN3D (h) ViT (i) SpectralFormer (j) MFT

Fig. 12: (a) Ground truth and classification maps obtained by HSI and SAR data for the Augsburg data set by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) ViT, (h) SpectralFormer, and (i) MFT using disjoint training samples.



(a) GT (b) KNN (c) RF (d) SVM (e) CNN1D

(f) CNN2D (g) CNN3D (h) ViT (i) SpectralFormer (j) MFT

Fig. 13: (a) Ground truth and classification maps obtained by HSI and DSM data for the Augsburg dataset by: (b) KNN, (c) RF, (d) CNN1D, (e) CNN2D, (f) CNN3D, (g) ViT, (h) SpectralFormer, and (i) MFT using disjoint training samples.

TABLE XI: OA, AA and Kappa values achieved by MFT with different tokenization i.e. pixel tokenizer (PT) and channel tokenizer (CT) for UH (HSI+LiDAR), UT (HSU+LiDAR), MUUFL (HSI+LiDAR), UH (HSI+MS), Augsburg (HSI+SAR), and Augsburg (HSI+DSM) datasets.

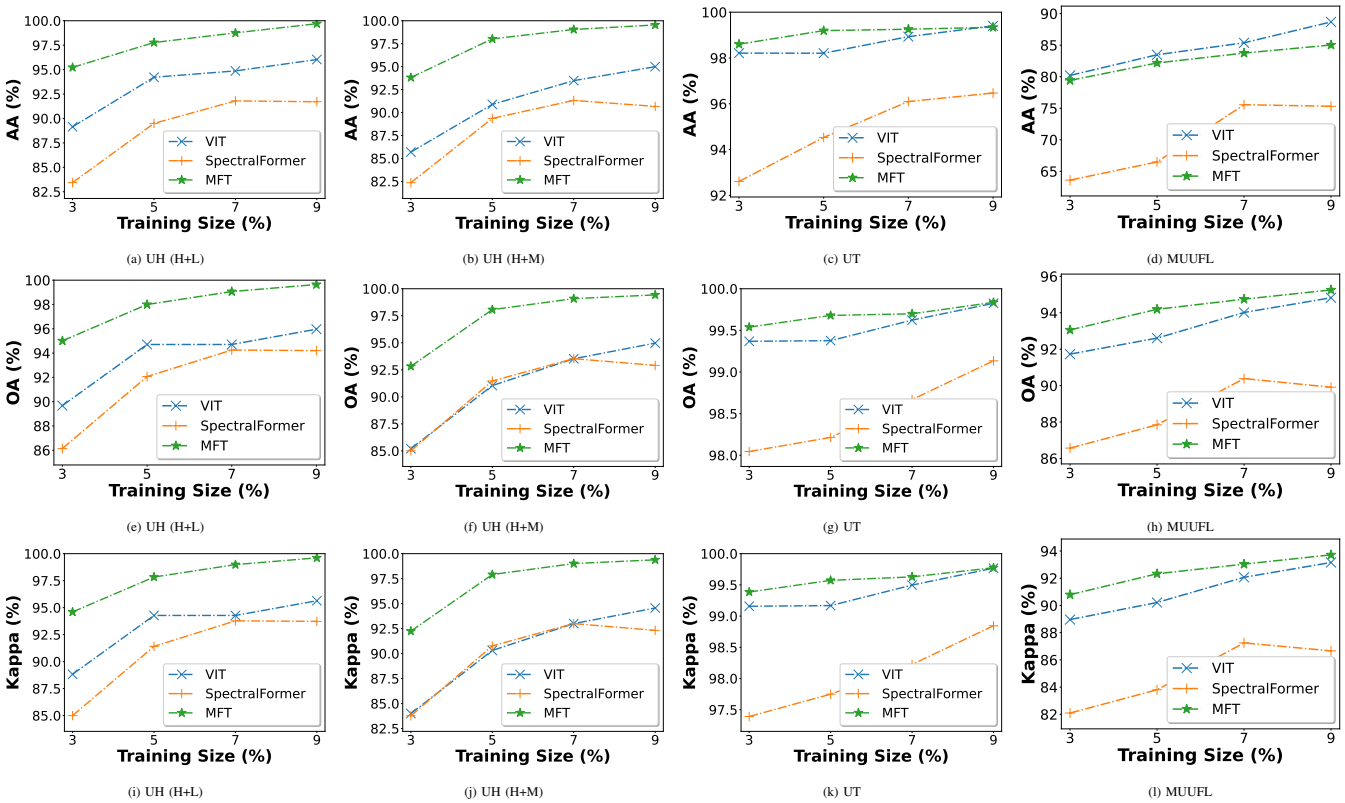
Class	Houston (HSI+LiDAR)		Trento (HSI+LiDAR)		MUUFL (HSI+LiDAR)		Houston (HSI+MS)		Augsburg (HSI+SAR)		Augsburg (HSI+DSM)	
No.	PT	CT	PT	CT	PT	CT	PT	CT	PT	CT	PT	CT
1	82.34 ± 00.47	**84.90 ± 03.02**	**98.71 ± 00.80**	98.23 ± 00.38	97.52 ± 00.38	**97.90 ± 00.39**	82.72 ± 00.41	**87.75 ± 07.12**	94.65 ± 00.62	**95.53 ± 00.69**	**95.38 ± 00.38**	94.77 ± 02.29
2	**88.78 ± 05.27**	84.71 ± 00.56	**99.50 ± 00.12**	99.34 ± 00.02	**92.35 ± 02.13**	92.11 ± 01.58	85.09 ± 00.09	**85.15 ± 00.00**	**96.90 ± 00.62**	96.81 ± 00.91	**97.21 ± 00.09**	96.03 ± 01.84
3	98.15 ± 01.08	**98.61 ± 01.01**	80.12 ± 07.78	**89.84 ± 09.00**	**93.21 ± 01.30**	91.80 ± 00.82	**98.55 ± 00.76**	98.42 ± 00.32	**69.80 ± 06.43**	66.98 ± 02.92	55.10 ± 07.22	**72.52 ± 10.54**
4	94.35 ± 02.40	**96.88 ± 01.45**	**99.96 ± 00.03**	99.82 ± 00.26	**93.37 ± 01.85**	91.59 ± 02.25	**95.99 ± 02.39**	92.90 ± 02.89	93.98 ± 01.65	**94.31 ± 01.81**	92.42 ± 00.34	**93.52 ± 01.96**
5	99.12 ± 01.25	**99.97 ± 00.04**	99.82 ± 00.23	**99.93 ± 00.05**	95.18 ± 00.13	**95.60 ± 01.21**	**99.78 ± 00.09**	99.21 ± 00.39	**32.70 ± 11.81**	24.54 ± 06.34	43.28 ± 11.66	**51.31 ± 05.13**
6	**99.30 ± 00.00**	90.44 ± 07.58	84.29 ± 01.40	**88.72 ± 00.94**	84.65 ± 02.48	**88.19 ± 03.49**	97.20 ± 01.98	**97.67 ± 01.74**	10.52 ± 04.31	**12.31 ± 02.19**	**14.06 ± 04.71**	13.49 ± 09.44
7	**88.56 ± 01.16**	81.84 ± 05.59			89.93 ± 02.23	**90.27 ± 02.13**	86.32 ± 00.69	**87.28 ± 01.38**	**23.98 ± 07.85**	15.28 ± 07.14	16.46 ± 04.41	**31.28 ± 10.80**
8	86.89 ± 05.33	**89.78 ± 01.83**			97.24 ± 00.48	**97.26 ± 00.53**	**81.16 ± 08.04**	77.71 ± 00.98				
9	87.91 ± 03.90	**91.94 ± 01.20**			58.89 ± 05.71	**61.35 ± 03.80**	87.76 ± 00.74	**88.29 ± 01.72**				
10	64.70 ± 00.92	**65.03 ± 12.68**			11.69 ± 04.51	**17.43 ± 04.63**	**74.71 ± 15.98**	65.67 ± 02.15				
11	98.64 ± 00.70	**99.81 ± 00.20**			**74.87 ± 04.06**	72.79 ± 09.25	93.71 ± 06.01	**96.90 ± 00.47**				
12	**94.24 ± 02.19**	89.95 ± 01.93					96.16 ± 00.87	**97.05 ± 00.99**				
13	**90.29 ± 00.66**	88.42 ± 03.49					**92.51 ± 00.17**	91.23 ± 00.76				
14	**99.73 ± 00.38**	99.73 ± 00.38					**100.0 ± 00.00**	99.46 ± 00.76				
15	**99.58 ± 00.17**	94.08 ± 07.07					86.82 ± 08.99	**95.07 ± 02.46**				
OA	**89.80 ± 00.53**	89.39 ± 00.78	97.82 ± 00.24	**98.32 ± 00.25**	94.26 ± 00.20	**94.34 ± 00.07**	**89.15 ± 00.96**	88.97 ± 00.36	**90.49 ± 00.20**	90.40 ± 00.17	89.48 ± 00.27	**90.49 ± 00.54**
AA	**91.51 ± 00.40**	90.41 ± 01.10	93.73 ± 01.52	**95.98 ± 01.64**	80.81 ± 01.02	**81.48 ± 00.70**	90.56 ± 00.93	**90.65 ± 00.29**	**60.36 ± 02.36**	57.97 ± 02.33	59.13 ± 02.80	**64.70 ± 00.44**
$\kappa(\times 100)$	**88.93 ± 00.59**	88.48 ± 00.85	97.08 ± 00.33	**97.75 ± 00.33**	92.41 ± 00.27	**92.51 ± 00.10**	**88.22 ± 01.04**	88.02 ± 00.39	**86.26 ± 00.23**	86.07 ± 00.28	84.80 ± 00.43	**86.35 ± 00.77**

to the better exchange of information between HS patch and other modalities. The learning of complementary information helps find strong nonlinear sequential relations from the HS image data. Hence, the generated classification maps contain more realistic and finer details inside the land-use and land-cover region.

E. Performance Over Varying Training Ratio

Here, we evaluate the performance of the proposed model with respect to the number of labeled training samples. In the case of inadequate or unreliable training samples, under-fitting (also called the Hughes Phenomena) or over-fitting (non-reliable) issues can occur. Therefore, selecting the appropriate number of training samples is another crucial factor for classification performance. The purpose of this section is to analyze the experimental results on different percentages of training samples where 3%, 5%, 7%, and 9% training samples are randomly selected to train the models (both the one presented in this paper and the comparative methods), and the remaining samples are used for testing. Other parameters remain the same, as discussed in the previous sections.

Figs. 14(a)-(d), 14(e)-(h), and 14(i)-(l) compare the detailed classification performances of the transformer models using AA, OA, and κ metrics respectively, on four datasets, i.e., Houston (HSI + LiDAR), Trento (HSI + LiDAR), MUUFL (HSI + LiDAR) and Houston (HSI + MS), with different percentages (3%, 5%, 7% and 9%) of randomly selected training samples. The y-axes of the graphs show the metrics

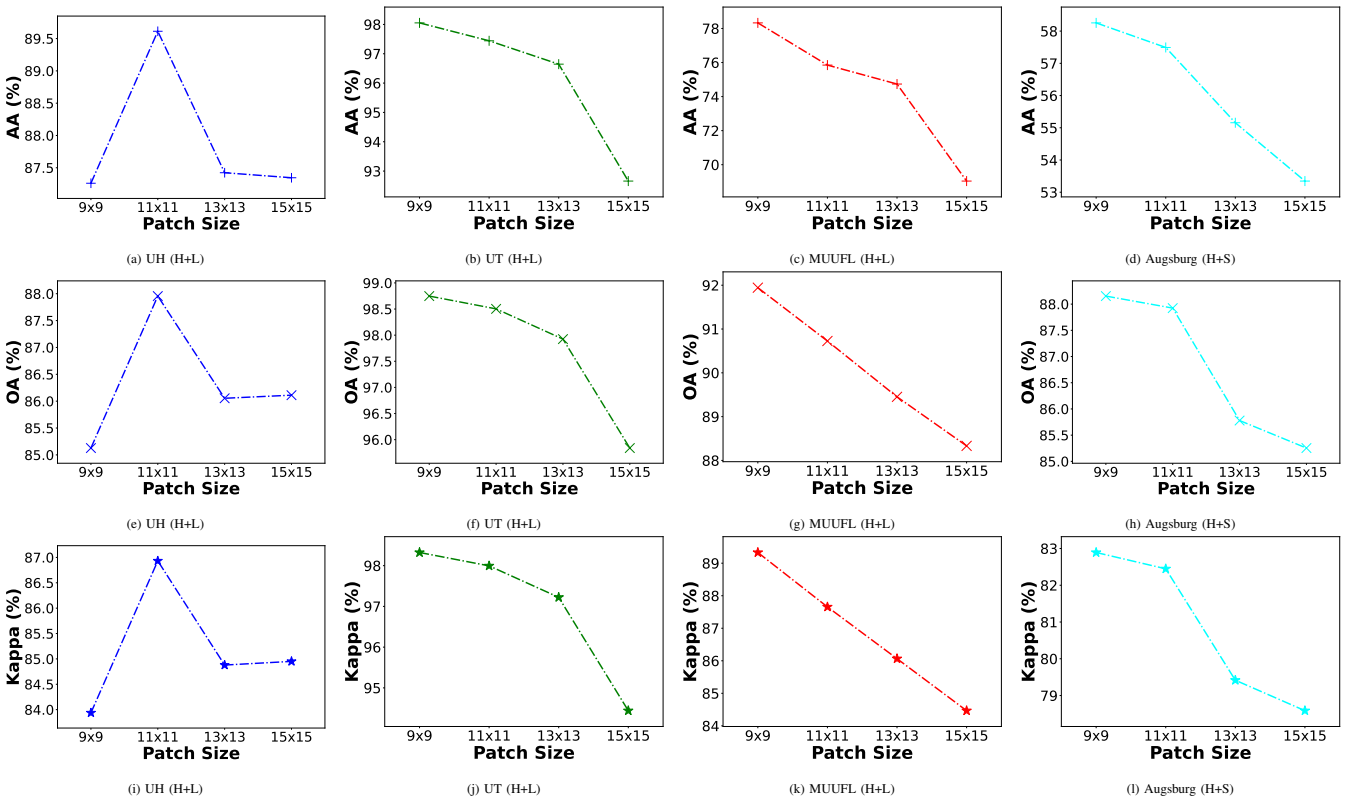


(a) UH (H+L) (b) UH (H+M) (c) UT (d) MUUFL

(e) UH (H+L) (f) UH (H+M) (g) UT (h) MUUFL

(i) UH (H+L) (j) UH (H+M) (k) UT (l) MUUFL

Fig. 14: Average accuracy (AA), overall accuracy (OA) and kappa coefficient (κ) achieved by different methods with varying training sample sizes which are randomly taken from (a) UH (HSI+LiDAR) (b) UH (HSI+MS) (c) Trento (HSI+LiDAR) and (d) MUUFL (HSI+LiDAR) datasets.



(a) UH (H+L) (b) UT (H+L) (c) MUUFL (H+L) (d) Augsburg (H+S)

(e) UH (H+L) (f) UT (H+L) (g) MUUFL (H+L) (h) Augsburg (H+S)

(i) UH (H+L) (j) UT (H+L) (k) MUUFL (H+L) (l) Augsburg (H+S)

Fig. 15: Average accuracy (AA), overall accuracy (OA) and kappa coefficient (κ) achieved by the MFT network with varying patch sizes which are taken from (a) UH (HSI+LiDAR) (b) Trento (HSI+LiDAR) (c) MUUFL (HSI+LiDAR) and (d) Augsburg (HSI+SAR) datasets.

(OA, AA or κ) while the x-axes show the training sample size in percent. Each model is denoted with a different marker and color which are consistent throughout Fig. 14. The proposed MFT is shown in green.

The graphs illustrate that the experimental results obtained with 3% training samples are inferior to those obtained with 9% training samples for all four datasets. In Houston, (HSI + LiDAR) and Houston (HSI + MS) the proposed MFT performs approximately 4% and 6% better than the second best model i.e. ViT, respectively, in all 3 accuracies for all percentages of training samples. In other datasets, though the proposed MFT performs better than ViT, the improvement is not much for larger training sample sizes. It is quite evident that even with the decrease in training samples, the performance of the proposed MFT is significantly better than the other transformer networks.

F. Model Stability Analysis

1) **Ablation Study:** Regardless of the source of multimodal data, the dimension to tokenize plays a crucial role in the final classification performance and should be fine-tuned in addition to learning parameters and hyper-parameters. Therefore, it is essential to compare the model's performance with respect to pixel tokenization and channel tokenization. Hence, we investigate the classification performance of both variations on all four datasets, along with their respective combined multimodal data. Table XI shows the comparison of the classification accuracies of both methods in terms of OA, AA, and κ for all the datasets. The optimal values are marked in bold. Overall, the performances of both methods are quite similar, but channel tokenization narrowly exceeds pixel tokenization. The reason may be that the channels represent the features better than pixels after convolution operations.

2) **Impact of Patch Size:** The performance of the proposed MFT network is also influenced by the spatial size of the input HSI patches. As shown in Fig. 15, the best and more stable AAs, OAs and Kappa's are achieved by the proposed model using a window of size (11×11) for all the experimentally evaluated datasets. In all these experiments, disjointed train/test samples are utilized as explained in Section IV-A. It can also be observed from the figures that, even the higher values in terms of AAs, OAs and Kappa's can be achieved over the UT (H+L), MUUFL (H+L), and Augsburg (H+S) datasets using a patch size of (9×9), but the performance is not satisfactory for the UH (H+L) multi-modalities. The results for patches of size 13×13 and 15×15 shows that performance can significantly be reduced when increasing the patch size. It is worth mentioning that increasing the spatial window size will lead to an increase in redundancy. Hence we have considered 11×11 HSI patches for the entire experimental evaluation.

3) **Hyperparameter Sensitivity Analysis:** The proposed MFT is not only effective but also fairly efficient in terms of computational complexity. The parameters and computations of the proposed MFT against other transformer networks are compared in Fig. 16(a)-(f). They illustrate the overall accuracies, parameters, and computations (FLOPs) for UH (HSI+LiDAR), UT (HSI+LiDAR), MUUFL (HSI+LiDAR), UH (HSI+MS), Augsburg (HSI+SAR), and Augsburg (HSI+DSM) datasets, respectively. The radii of circles indicate the computations (FLOPs). The efficiency of the proposed MFT is quite evident in Trento (HSI+LiDAR) and MUUFL (HSI+LiDAR), where it uses fewer parameters and FLOPS than ViT (the model with second-best accuracies). In other cases, though, the parameters and FLOPS of the proposed MFT are higher than others, the increase in performance makes up for it. As we can see with Houston (H + MS), the proposed MFT exhibits an impressive 8.66% increase in OA than the next best model, i.e., ViT. Such a massive gain in classification accuracy makes the parameter trade-off justified.

V. Conclusions

In this paper, we propose a new multimodal fusion transformer network based on a multihead cross patch attention (`mCrossPA`) module to fuse HSI and other sources of multimodal data for land cover classification. Instead of using conventional feature fusion techniques, other sources of multimodal data are used as an external classification (`CLS`) token that is incorporated into the `mCrossPA` block in the transformer encoder, which helps learning long-range feature dependencies. As the `CLS` token captures information that is complementary to that of the HSI data, it also helps in achieving a better generalization of the model and improves classification accuracy. The `CLS` token can be generated from multiple sources of easily available data, such as MSIs, SAR, DSM, and LiDAR. As a result, the model's performance is significantly increased with minimal effort.

Fusing multimodal data (such as HSIs, MSIs, SAR, DSMs, and Lidar) using recently developed transformers is challenging and requires significant effort. However, the extensive experimental results presented in this work confirm that the proposed MFT can successfully fuse multimodal data using transformers, which are inherently better at classification tasks than classical convolutional models. The proposed MFT performs better than all the other tested models with all the considered datasets. Therefore, we believe that the proposed MFT exhibits the capacity to perform multimodal fusion tasks easily in remote sensing and Earth Observation due to its superior fusion capabilities.

References

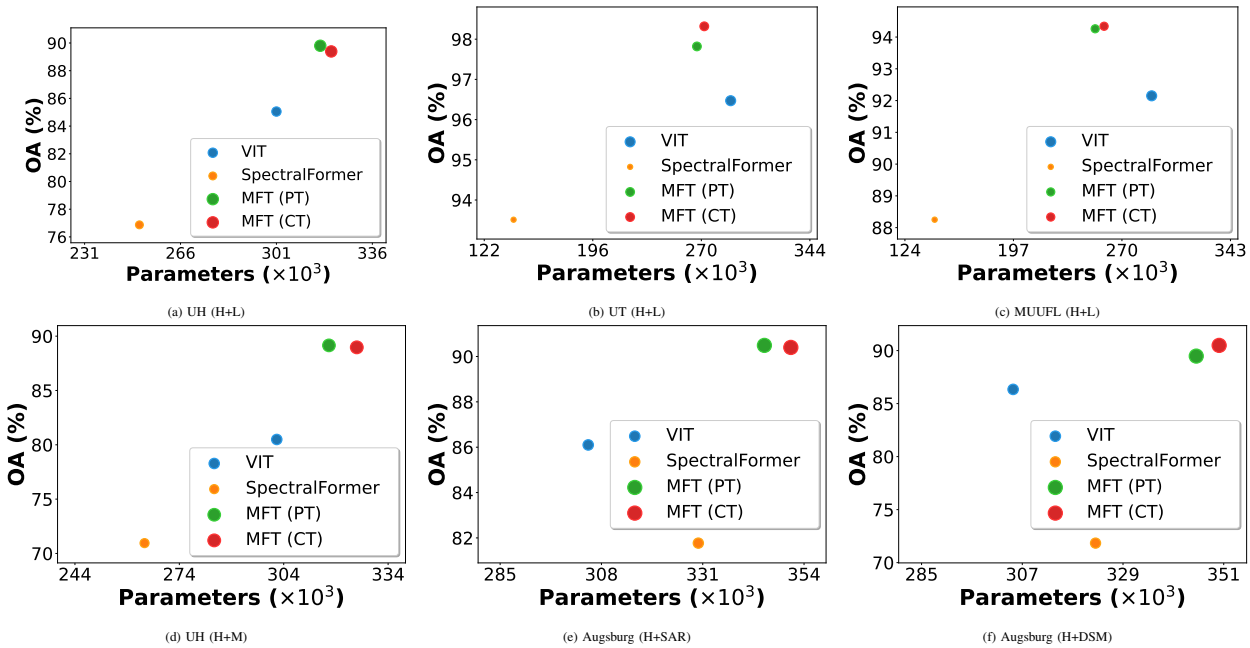
[1] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.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

[2] E. Bartholome and A. S. Belward, "Glc2000: a new approach to global land cover mapping from earth observation data," *International Journal of Remote Sensing*, vol. 26, no. 9, pp. 1959–1977, 2005.

[3] S. K. Roy, P. Kar, D. Hong, X. Wu, A. Plaza, and J. Chanussot, "Revisiting deep hyperspectral feature extraction networks via gradient centralized convolution," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[4] B. Koetz, F. Morsdorf, S. Van der Linden, T. Curt, and B. Allgöwer, "Multi-source land cover classification for forest fire management based on imaging spectrometry and lidar data," *Forest Ecology and Management*, vol. 256, no. 3, pp. 263–271, 2008.

[5] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 2, pp. 302–306, 2019.



(a) UH (H+L) (b) UT (H+L) (c) MUUFL (H+L)

(d) UH (H+M) (e) Augsburg (H+SAR) (f) Augsburg (H+DSM)

Fig. 16: Comparison of the transformer networks in terms of overall classification accuracy, network parameters, and FLOPs, indicated by radii of circles taken from (a) UH (HSI+LiDAR) (b) UT (HSI+LiDAR) (c) MUUFL (HSI+LiDAR) (d) UH (HSI+MS) (e) Augsburg (HSI+SAR) and (f) Augsburg (HSI+DSM) datasets.

[6] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2023.

[7] S. L. Ustin, *Manual of remote sensing, remote sensing for natural resource management and environmental monitoring*, vol. 4. John Wiley & Sons, 2004.

[8] C. Chen, J. Yan, L. Wang, D. Liang, and W. Zhang, "Classification of urban functional areas from remote sensing images and time-series user behavior data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1207–1221, 2020.

[9] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, and X. Zhu, "Interpretable hyperspectral artificial intelligence: When non-convex modeling meets hyperspectral remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 52–87, 2021.

[10] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, *et al.*, "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.

[11] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, 2019.

[12] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and lidar data," *International Journal of Image and Data Fusion*, vol. 6, no. 3, pp. 189–215, 2015.

[13] U. Heiden, W. Heldens, S. Roessner, K. Segl, T. Esch, and A. Mueller, "Urban structure type characterization using hyperspectral remote sensing and height information," *Landsc. Urban Plan.*, vol. 105, no. 4, pp. 361–375, 2012.

[14] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and lidar remote sensing data using multiple feature learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2971–2983, 2015.

[15] B. Rasti and P. Ghamisi, "Remote sensing image classification using subspace sensor fusion," *Information Fusion*, vol. 64, pp. 121 – 130, 2020.

[16] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, 2010.

[17] W. Liao, R. Bellens, A. Pizurica, S. Gautama, and W. Philips, "Graph-based feature fusion of hyperspectral and lidar remote sensing data using morphological features," in *Proc. IGARSS*, pp. 4942–4945, 2013.

[18] 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.

[19] A. Merentitis, C. Debes, R. Heremans, and N. Frangiadakis, "Automatic fusion and classification of hyperspectral and lidar data using random forests," in *2014 IEEE Geoscience and Remote Sensing Symposium*, pp. 1245–1248, IEEE, 2014.

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

[21] S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza, "Morphological convolutional neural networks for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8689–8702, 2021.

[22] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[23] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 2, pp. 277–281, 2020.

[24] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-d cnn for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, 2020.

[25] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal gans: Toward crossmodal hyperspectral-multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, 2021.

[26] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (jpsa) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3602–3615, 2021.

[27] 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.

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

[29] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sensing*, vol. 10, no. 3, 2018.

[30] 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*, vol. 82, pp. 1–18, 2022.

[31] 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. 10.1109/LGRS.2020.3017414.

[32] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, D. Qian, and B. Zhang, "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.

[33] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, 2021.

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

[35] 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.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[37] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, 2017.

[38] S. K. Roy, S. Chatterjee, S. Bhattacharyya, B. B. Chaudhuri, and J. Platoš, "Lightweight spectral–spatial squeeze-and-excitation residual bag-of-features learning for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5277–5290, 2020.

[39] S. K. Roy, S. R. Dubey, S. Chatterjee, and B. B. Chaudhuri, "Fusenet: Fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification," *IET Image Processing*, vol. 14, no. 8, pp. 1653–1661, 2020.

[40] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[41] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, 2018.

[42] S. K. Roy, D. Hong, P. Kar, X. Wu, X. Liu, and D. Zhao, "Lightweight heterogeneous kernel convolution for hyperspectral image classification with noisy labels," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[43] M. E. Paoletti, J. M. Haut, S. K. Roy, and E. M. Hendrix, "Rotation equivariant convolutional neural networks for hyperspectral image classification," *IEEE Access*, vol. 8, pp. 179575–179591, 2020.

[44] S. Roy, M. Paoletti, J. Haut, E. Hendrix, and A. Plaza, "A new max-min convolutional network for hyperspectral image classification," in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–5, IEEE, 2021.

[45] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, 2018.

[46] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[47] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[48] 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.

[49] 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.

[50] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, *et al.*, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2209.01206*, 2022.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[52] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral-spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2022.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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 Trans. Geosci. Remote Sens.*, 2023.

[59] 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*, pp. 4835–4844, 2019.

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

[61] 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.

[62] 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.

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

[64] A. Baumgartner, P. Gege, C. Köhler, K. Lenhard, and T. Schwarzmaier, "Characterisation methods for the hyperspectral sensor hyspex at dlr's calibration home base," in *Sensors, Systems, and Next-Generation Satellites XVI*, vol. 8533, p. 85331H, International Society for Optics and Photonics, 2012.

[65] F. Kurz, D. Rosenbaum, J. Leitloff, O. Meynberg, and P. Reinartz, "Real time camera system for disaster and traffic monitoring," in *Proceedings of International Conference on SMPR 2011*, pp. 1–6, 2011.

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

[67] 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.

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

[69] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, "diffgrad: An optimization method for convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Swalpa Kumar Roy (S'15) received the bachelor's and the master's degree in Computer Science and Engineering from West Bengal University of Technology, Kolkata, India, in 2012, and Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, (IIEST Shibpur) in 2015 and also the Ph.D. degree in Computer Science and Engineering from University of Calcutta, Kolkata in 2021.

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. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, West Bengal, India. Dr. Roy was nominated for the Indian National Academy of Engineering (INAE) engineering teachers mentoring fellowship program by INAE Fellows in 2021 and 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 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.

Ankur Deria received a bachelor's degree in Computer Science and Engineering from Jalpaiguri Government Engineering College, Jalpaiguri, India in 2022. He is currently pursuing an M.Sc. degree in Informatics: Games Engineering, at the Technical University of Munich, Germany. He was nominated for the Indian National Academy of Engineering (INAE) engineering students mentoring fellowship by INAE fellows in academic tenure 2022-23. His research interests include computer vision, 3D shape, and volume reconstruction.

Danfeng Hong (S'16–M'19–SM'21) received the M.Sc. degree (summa cum laude) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, the Dr. -Ing degree (summa cum laude) from the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences (CAS). Before joining CAS, he has been a Research Scientist and led a Spectral Vision Working Group at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He was also an Adjunct Scientist at GIPSA-lab, Grenoble INP, CNRS, Univ. Grenoble Alpes, Grenoble, France. His research interests include artificial intelligence, remote sensing big data analysis, multimodal interpretation, and their applications in Earth Vision.

Dr. Hong is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing (TGRS), an Editorial Board Member of Remote Sensing, an Editorial Advisory Board Member of the ISPRS Journal of Photogrammetry and Remote Sensing. He was a recipient of the Best Reviewer Award of the IEEE TGRS in 2021 and 2022, and the Best Reviewer Award of the IEEE JSTARS in 2022, the Jose Bioucas Dias Award for recognizing the outstanding paper at WHISPERS in 2021, the Remote Sensing Young Investigator Award in 2022, the IEEE GRSS Early Career Award in 2022, and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) in 2022.

Behnood Rasti (M'12–SM'19) received the B.Sc. and M.Sc. degrees both in electronics- electrical engineering from the Electrical Engineering Department, University of Guilan, Rasht, Iran, in 2006 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2014. In 2015 and 2016, he worked as a Post-Doctoral Researcher with the Electrical and Computer Engineering Department, University of Iceland. From 2016 to 2019, he has been a Lecturer with the Center of Engineering Technology and Applied Sciences, Department of Electrical and Computer Engineering, University of Iceland. Dr. Rasti was a Humboldt research fellow in 2020 and 2021. He is currently a Principal Research Associate with Helmholtz-Zentrum Dresden-Rossendorf (HZDR). His research interests include signal and image processing, machine/deep learning, remote sensing, and artificial intelligence.

Dr. Rasti was the Valedictorian as an M.Sc. Student in 2009. He won the Doctoral Grant of The University of Iceland Research Fund "The Eimskip University Fund," and the "Alexander von Humboldt Research Fellowship Grant" in 2013 and 2019, respectively. He serves as an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL).

Antonio Plaza (M'05-SM'07-F'15) is a Full Professor and the Head of the Hyperspectral Computing Laboratory at the Department of Technology of Computers and Communications, University of Extremadura, where he received the M.Sc. degree in 1999 and the Ph.D. degree in 2002, both in Computer Engineering. His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has authored more than 600 publications in this field, including 406 JCR journal papers, 25 book chapters, and 330 peer-reviewed conference proceeding papers. He has guest edited 17 special issues on hyperspectral remote sensing for different journals. Prof. Plaza is a Fellow of IEEE "for contributions to hyperspectral data processing and parallel computing of Earth observation data," a Fellow of the Asia-Pacific Artificial Intelligence Association (AAIA), and an elected member of Academia Europaea, The Academy of Europe. He is a recipient of the recognition of Best Reviewers of the IEEE Geoscience and Remote Sensing Letters (in 2009) and a recipient of the recognition of Best Reviewers of the IEEE Transactions on Geoscience and Remote Sensing (in 2010), for which he served as Associate Editor in 2007-2012. He is also an Associate Editor for IEEE Access (receiving the recognition of Outstanding Associate Editor for the journal in 2017) and was also a member of the Editorial Board of the IEEE Geoscience and Remote Sensing Newsletter (2011-2012) and the IEEE Geoscience and Remote Sensing Magazine (2013). He was also a member of the steering committee of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS). He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015, the 2013 Best Paper Award of the JSTARS journal, and the most highly cited paper (2005-2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, the IEEE International Conference on Space Technology, and the IEEE Symposium on Signal Processing and Information Technology. He served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) in 2011-2012, and as President of the Spanish Chapter of IEEE GRSS in 2012-2016. He is currently serving as Chair of the Publications Awards Committee of IEEE GRSS and as a Vice-Chair of the Fellow Evaluations Committee of IEEE GRSS. He has reviewed more than 500 manuscripts for over 50 different journals. He served as the Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing journal for five years (2013-2017) and as the Editor-in-Chief of the IEEE Journal on Miniaturization for air and Space Systems (2019-2020). He has been included in the 2018, 2019, 2020, 2021, and 2022 Highly Cited Researchers List (Clarivate Analytics). Additional information: http://sites.google.com/view/antonioplaza

Jocelyn Chanussot (M'04–SM'04–F'12) (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA, the KTH Royal Institute of Technology, Stockholm, Sweden, and the National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor at the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor at the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008. He was a member of the Institut Universitaire de France from 2012 to 2017. He was the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He is also a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters). He was the Chair and the Co-Chair of the GRS Data Fusion Technical Committee from 2009 to 2011 and 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society and was in charge of meetings and symposia from 2017 to 2019. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014, he served as a Guest Editor for the IEEE Signal Processing Magazine. He is also an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE.