# Rotation is All You Need: Cross Dimensional Residual Interaction for Hyperspectral Image Classification

Xin Qiao, *Student Member, IEEE*, Swalpa Kumar Roy, *Student Member, IEEE*, and Weimin Huang, *Senior Member, IEEE*

***Abstract*—The performance of deep convolutional neural networks has been significantly improved in recent years as a result of additional attention mechanisms applied to the standard networks. Numerous experiments conducted have demonstrated that spectral-spatial attention enhances the network's categorization ability. The three attention modules that currently use spatial attention, spectral attention, and channel attention are isolated from each other and their interrelationships are not fully considered. To solve this problem and establish the dependencies among different channels, spectral bands, spatial height, and width simultaneously, in this article, a new cross attention module called quadlet is proposed, which can capture information using simultaneous interaction of the channel, spectral depth and spatial location to improve the classification accuracy of hyperspectral images. By incorporating the quadlet attention module, a cross-dimensional residual network (QuadNet) is proposed for HSIs classification. A series of experiments conducted on four publicly available hyperspectral datasets showed that the proposed cross-attention residual network can effectively establish the dependencies among different dimensions of input tensor and achieve 98.22%, 99.88%, 99.10%, and 96.46% overall accuracy on IN, UP, SA, and UH datasets, respectively.**

***Index Terms*—Hyperspectral image classification, multibranches cross-attention, multibranches cross-attention residual network.**

## I. Introduction

Hyperspectral images (HSIs) can efficiently distinguish objects with similar appearance through contiguous spectral signatures, which provide abundant and detailed spectral information [1], [2]. They have been widely used in various fields of Earth observation, such as agriculture, forestry, land management, and military monitoring [3]. One of the fundamental research areas of HSIs processing is classification, which aims to classify each pixel in HSIs [4].

Although hundreds of bands can provide rich spectral features, they also cause severe band redundancy and the spectral curse of dimensionality [5]. To solve these problems and extract feature bands efficiently, many band selection and band extraction methods have been applied to HSIs over the past decades, such as principal component analysis and search-based or clustering methods. [6], [7]. However, the feature extraction methods with manual intervention cannot achieve expected results with a good generalization ability. Therefore, discriminative feature extraction from HSIs remains challenging.

With the application of deep learning techniques, especially deep convolutional neural networks (CNNs), HSIs classification performance has made great progress [4]. According to the difference of input features, CNN-based methods could be roughly divided into spectral-based methods and spectral-spatial-based methods. Spectral-based methods [8], [9] utilize the spectral signatures of each pixel as input, without considering the spatial information. Spectral-spatial-based methods [10], [11], [12] extract patches that consist of the central target pixel and its neighboring pixels to effectively integrate both spectral and spatial features. Besides CNNs, recurrent neural networks (RNNs) [13], gated recurrent unit network (GRU) [14], long short-time memory [15], and generative adversarial networks (GANs) [16], [17], [18] have also been widely explored for HSIs classification.

The abovementioned models extract features using deep neural networks but without attention modules. Attention mechanisms have also been introduced to improve the image classification results [19]. The attention mechanism is an emerging technique in recent years to simulate the signal processing mechanism unique to the human vision system, and it quickly acquires the target regions that need to be focused on [20], [21]. The aim of attention is to create dependencies among different channels within feature maps and capture the meaningful information encoded in channel dimensions. The attention mechanism has been widely studied in the field of computer vision [22], [23], as well as in HSIs classification tasks [24], [25].

However, most of the attention modules applied in the existing networks establish interrelationships between the spectral channels and the spatial features, together or separately, for HSIs classification, while ignoring the importance of cross-dimensional interaction with the number of obtained feature maps [26]. 3-D convolutional layers create three-dimensional

Manuscript received 17 March 2023; revised 24 May 2023; accepted 2 June 2023. Date of publication 6 June 2023; date of current version 20 June 2023. The work was supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grants under Grant NSERC RGPIN-2017-04508. *(Corresponding author: Weimin Huang.)*

Xin Qiao and Weimin Huang are with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada (e-mail: xqiao@mun.ca; weimin@mun.ca).

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

Digital Object Identifier 10.1109/JSTARS.2023.3283342

feature maps from inputs with spatial and spectral dimensions and the feature maps correspond to the learned representation of the input data. Incorporating the number of feature maps in an attention module of 3-D CNNs can selectively highlight the most informative features and suppress both irrelevant and noisy features. Therefore, incorporating the number of feature maps in an attention module can improve the quality of the learned features and increases the accuracy of the model. To simultaneously model the interactions among the number of feature maps, spectral depth, and spatial locations, i.e., height and width, inspired by the triplet attention mechanism [27], a new quadlet attention is proposed in this article for accelerating the learning of discriminative spectral and spatial features during models training. Quadlet attention constructs the relationship among different dimensions of the input tensor, i.e., the number of feature maps, spectral channels, and spatial locations to extract the cross dimensional attention weights by capturing cross dimension interaction using a four-branch parallel architecture.

Consider the shape of input tensor $(B, C, D, H, W)$ where the batch size $B$, the number of feature maps $C$, spectral depth $D$, spatial height $H$, and width $W$ are generated during the forward propagation of CNNs. The corresponding four independent branches of quadlet attention can be modeled as $(D, H, W)$, $(C, H, W)$, $(D, C, W)$, and $(D, H, C)$, to establish the dependencies between channels, bands, spatial height, and width, respectively. Quadlet attention encodes interchannels and spatial information for a given input tensor and develops interdimensional interdependence through permutation operation, followed by a cost effective residual connection. The proposed quadlet attention module is utilized to design a simple and effective cross dimensional spectral-spatial residual interaction network for HSI classification. The main contributions of this article are summarized as follows.

1) We integrate a simple and effective cross dimensional attention called triplet attention in HSI classification. Moreover, we consider one additional dimension to further propose the quadlet attention, which could establish the dependencies between any three dimensions among the number of feature maps, spectral depth, the spatial height and width of input tensor.
2) The quadlet attention module is integrated with an improved SSRN which enables learning of cross dimensional spectral-spatial feature representation for the HSIs classification task.

The rest of this article is organized as follows. Section II introduces related work in detail. The proposed quadlet attention module and the developed QuadNet architecture are described in detail in Section III. The experimental setup and results are provided in Section IV. Finally, Section V concludes this article.

## II. Related Work

### A. HSIs Classification Using Conventional Networks

Many conventional networks without attention mechanisms have been widely explored for HSIs classification. From the 1-D spectral perspective, Hu et al. [8] directly classified HSIs from the spectral domain using a 1-D CNN. Gao et al. [9] extracted spectral information and transformed the 1-D spectral array to a 2-D feature map. Then, they classified the HSIs by stacking convolutional layers with kernel sizes of $1 \times 1$ and $3 \times 3$. For 2-D spatial frameworks, Yu et al. [28] took the original data as input and employed a 2-D CNN for HSIs classification. Ding et al. [29] trained a 2-D CNN framework where the kernel size was adaptively learned from the data to classify HSIs. More common approaches involve extracting spectral and spatial features jointly for HSIs classification. For instance, Roy et al. [30] proposed a hybrid spectral convolutional neural network (HybridSN), which includes spectral-spatial 3-D-CNN and spatial 2-D-CNN. The former learns the joint spectral-spatial feature representations, and the latter extracts more abstract spatial information. Zhong et al. [31] developed a supervised deep learning framework called spectral-spatial residual network (SSRN) for HSI classification. SSRN includes four consecutive residual blocks to capture discriminative features from spectral signatures and spatial contexts. Paoletti et al. [32] proposed a deep pyramidal residual network to extract deeper spectral-spatial representations through more convolutional filters of the network. Zhang et al. [33] designed a multiscale dense network to combine and make full use of different scale features. Mou et al. [34] proposed a fully end-to-end conv–deconv network for unsupervised spectral-spatial feature learning. The proposed conv–deconv network can largely alleviate the reliance on training sample data with labels and solve the problem of a limited number of hyperspectral remote sensing image samples.

In addition to CNNs, other types of networks, such as RNNs GANs, have also been applied for HSIs classification. Mou et al. [35] considered HSIs as sequenced data and explored the efficient RNN for HSIs classification. In addition, Hang et al. [36] considered spectral signatures to be sequences and used GRUs to create a cascaded RNN model to separate the important representation from the redundant data. To deal with the issue of limited sample data and the challenge of gathering ground-truth labels, Hang et al. [37] proposed a multitask generative adversarial network (MTGAN). The proposed MTGAN consists of a generator network for hypercube reconstruction and classification, and a discriminator network to discriminate between the real and reconstructed data. Similarly, Roy et al. [38] introduced a generative model which can efficiently tackle the problem of classwise imbalanced training samples for HSIs classification.

### B. HSIs Classification Using Attention-Aided Networks

In recent studies, attention module has been introduced to establish the dependencies within spectral bands or spatial locations. Paoletti et al. [39] designed an attention-aided capsule network to increase hyperspectral classification performance and computational efficiency. The attention mechanisms could help extract and identify the most representative and meaningful features of the images. Yu et al. [40] presented a feedback attention-guided spectral-spatial dense CNN to address the problem of information redundancy and inefficient representations of spectral-spatial features for hyperspectral classification tasks.

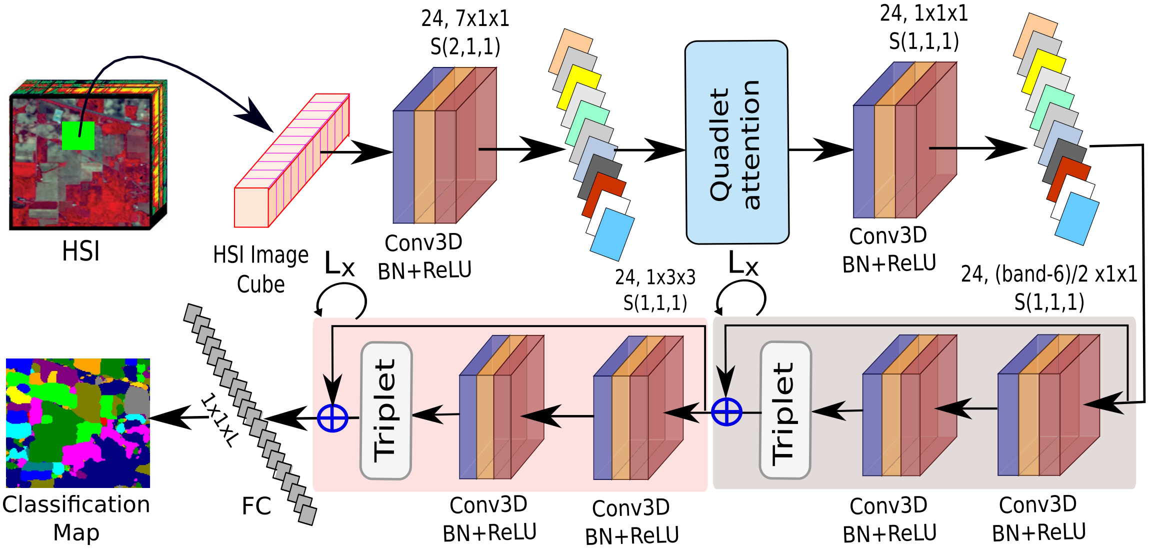


Fig. 1. Overall architecture of the proposed multibranch cross attention residual network.

Yang et al. [41] proposed a cross-attention spectral-spatial network to solve the problem of the high sensitivity of convolutional features extracted from the HSIs. However, it still exhibited a poor classification performance for the pixels near the edges. Hang et al. [42] designed an attention-aided CNN model to fully explore the discriminative features by focusing on the spectral bands and spatial positions within small hypercubes. Zhu et al. [43] proposed a residual spectral-spatial attention network (RSSAN) for HSIs classification. Nevertheless, a notable limitation of RSSAN was the lack of utilization of 3-D CNN for discriminative spectral-spatial feature extraction. Haut et al. [44] incorporated the attention mechanisms to the residual networks for characterizing the spectral-spatial information. This approach resulted in an improved classification performance but the quality of the captured features needed to be enhanced further. Mou et al. [45] developed a spectral attention module to selectively highlight the most important spectral bands in HSIs using the gating mechanisms. Despite their method achieved promising results, it lacked an interpretation for assessing the importance of the spectral bands obtained through the spectral attention module. Li et al. [46] developed a double-branch dual-attention mechanism network (DBDA), which enables learning of complementary spectral and spatial features by utilizing channel attention and spatial attention separately. Mei et al. [47] introduced a novel approach named the spectral-spatial attention network for HSI classification. The network utilized a combination of RNNs with attention to capture spectral correlations within a continuous spectrum, and CNNs with attention to model the spatial relevance between neighboring pixels in the spatial domain. However, the generalization of these methods to complex scenarios was not taken into consideration. Wu et al. [48] constructed a 3-D CNN-based residual group channel and spatial attention network for HSIs classification. The attention modules selectively strengthened the informative features in the input data, enhancing both the spatial as well as channelwise representations. The developed method improved the classification accuracy but also resulted in an increase in the number of network parameters.

Furthermore, the self-attention transformers have also been applied in the field of HSIs classification [49], [50], [51], [52]. Liu et al. [53] designed a scaled dot-product central attention module to extract spectral-spatial information from the central pixels and their adjacent pixels. Based on the proposed attention module, a central attention network was developed, which achieved superior classification performance. Zhao et al. [54] presented a graph transformer network with graph attention mechanism to learn node features on heterogeneous graphs. The proposed method can solve the problem of zero weight edge in heterogeneous graph through assigning weights for edges. Nonetheless, most of the aforementioned methods primarily focus on establishing dependencies among spectral and spatial dimensions, ignoring the potential interaction across different dimensions, such as the number of feature maps.

## III. Proposed Method

Fig. 1 shows the overall architecture of the proposed multibranched cross attention residual network (QuadNet). The proposed method mainly consists of four crucial steps as follows.

1) Extraction of HSIs small hypercubes and low-level representations.
2) Establishment of dependencies among the number of feature maps, spectral channels, spatial height and width through the quadlet attention module.
3) Extraction of spectral-spatial features via triplet attention-aided spectral-spatial residual blocks.
4) Classification by fully connected layer and softmax function.

### A. Preprocessing of HSI Data

Suppose the HSI data are represented as $\mathbf{H} \in \mathbf{R}^{h,w,b}$, where $h, w$ denote the spatial dimension, i.e., height and width, and $b$

TABLE I
PARAMETERS CONFIGURATION OF QUADNET FOR THE IN DATASET

| Layer / Block | Channels | Kernel Size | Padding | Stride | Output Shape |
|---|---|---|---|---|---|
| Input Layer | Input Shape= (11, 11, 200) | | | | |
| Conv3D<br>BN | 24 | (7,1,1) | (0,0,0) | (2,1,1) | (B,24,97,11,11) |
| Quadlet Attention | 24 | / | / | / | (B,24,97,11,11) |
| Conv3D<br>BN<br>ReLU | 24 | (1,1,1) | (0,0,0) | (1,1,1) | (B,24,97,11,11) |
| Residual block-1<br>Residual block-2 | 24 | (7,1,1) | (3,0,0) | (1,1,1) | (B,24, 97, 11,11) |
| Conv3D<br>BN | 128 | (97,1,1) | (0,0,0) | (1,1,1) | (B, 128, 1, 11, 11) |
| Permute | / | / | / | / | (B, 1, 128, 11, 11) |
| Conv3D<br>BN | 24 | (128,3,3) | (0,0,0) | (1, 1, 1) | (B, 24, 1, 9, 9) |
| Residual block-3<br>Residual block-4 | 24 | (1,3,3) | (0,1,1) | (1, 1, 1) | (B, 24, 1, 9, 9) |
| Average Pooling 3D | / | / | / | / | (B, 24, 1, 1, 1) |
| Linear | 16 | / | / | / | (B, 16) |

denotes its spectral dimension, namely, the number of bands of the hyperspectral data. Among all pixels, suppose there are $N$ pixels with category labels $\{x_1, x_2, \ldots, x_n\} \in \mathbf{R}^{1\times 1\times b}$, and the number of land-cover categories is $c$, then the corresponding true values of $N$ pixels are $\{y_1, y_2, \ldots, y_n\} \in \mathbf{R}^{1\times 1\times c}$. To consider both the spectral and spatial information from HSIs, the central pixel with ground truth label and their neighborhood pixels in a certain range of spatial dimensions are extracted simultaneously, thus forming a small hypercube of $\mathbf{X} \in \mathbf{R}^{p\times p\times b}$, where $p$ denotes the patch size. All the $N$ samples with their associated labels are randomly divided into training sets ($\mathbf{X}_{\text{train}}$), validation sets ($\mathbf{X}_{\text{val}}$), and test sets ($\mathbf{X}_{\text{test}}$), respectively. $\mathbf{X}_{\text{train}}$ is used to train the model and optimize the model parameters, $\mathbf{X}_{\text{val}}$ is used for model selection during training, and $\mathbf{X}_{\text{test}}$ is used for final model evaluation.

### B. Overall Classification Framework

Let us consider Indian pine (IN) dataset with the size of $200 \times 145 \times 145$ as an example. The overlapping patches are extracted to create small hypercube of the size $200 \times 11 \times 11$. The input data of shape $(B, 1, 200, 11, 11)$ are given to the initial 3-D convolution layer, where $B$ represents the batch size, 1 is the number of feature maps and 200, 11, and 11 denote the spectral bands, height, and width, respectively, of the extracted hypercube of the IN dataset, as shown in Fig. 1. Table I shows the model parameters for each block. The network takes a small hypercube as input and the first 3-D convolution layer is applied to extract low-level features by considering the convolution operations in both spectral and spatial dimensions with the help of a 3-D convolutional kernel of size $(7, 1, 1)$, and stride of $(2, 1, 1)$, followed by a batch normalization (BN) layer. We consider the input small hypercubes as the initial feature map with the number of 1, and the number of kernels in the first layer, i.e., the number of output feature maps is set to 24. The first 3-D convolutional layer will increase the number of feature maps from 1 to 24, decrease the spectral dimension from 200 to 97, and keep the same spatial dimension. BN is applied after every convolutional layer to prevent the model from overfitting.

To establish the cross dimensional interaction among different dimensions of feature maps, i.e., number of feature maps, spectral depth, and spatial locations, a quadlet attention module, which can effectively extract the meaningful feature representation using the interaction of all the dimensions of the feature maps is introduced and explained step by step in the following section. The quadlet attention module does not change the shape of its input feature map, so the shape of the output feature map obtained after the attention module is still $(B, 24, 97, 11, 11)$. After exploring the interaction in different dimensions of the feature maps, the feature extraction is further performed using a 3-D convolution with a kernel size of $(1, 1, 1)$ and stride of $(1, 1, 1)$, followed by a BN, and a ReLU activation layer. Then four successive triplet attention aided spectral-spatial residual blocks are used to further extract spatial and spectral features. Finally, two fully connected layers are used to obtain the predicted label. Fig. 1 shows the framework of the proposed QuadNet network, which is described in the following sections.

*1) 3-D Convolution:* Suppose the input feature map of the $l$th 3-D convolutional layer is defined by $(B, C^{l-1}, D^{l-1}, H^{l-1}, W^{l-1})$, where $C^{l-1}$ is the input channel, i.e., the number of input feature maps, and $D^{l-1}, H^{l-1}$, and $W^{l-1}$ represent the spectral depth, spatial height, and width of the $(l-1)$th layer output feature maps. The $l$th convolutional layer has $C^l$ convolutional kernels of size $(k_1^l, k_2^l, k_3^l)$, and subsampling strides of $(s_1, s_2, s_3)$. Zero padding is also employed to keep the shape of output feature maps unchanged. Then the $l$th convolutional layer generates an output feature map of shape $(C^l, D^l, H^l, W^l)$, where the number of output feature maps is equal to the number of kernels $C^l$. The spectral depth $D^l$ equals to $1 + (D^{l-1} - k_1^l)/s_1$, spatial height $H^l$ equals to $1 + (H^{l-1} - k_2^l)/s_2$, and the spatial width is similar to its height. The $l$th 3-D convolutional layer with BN operation could be expressed by as follows:

$$\mathbf{X}_p^l = \left( \sum_{j=1}^{C^{l-1}} \mathbf{ReLU}\left( \hat{\mathbf{X}}_j^{l-1} \right) * \mathbf{W}_p^l + b_p^l \right) \tag{1}$$

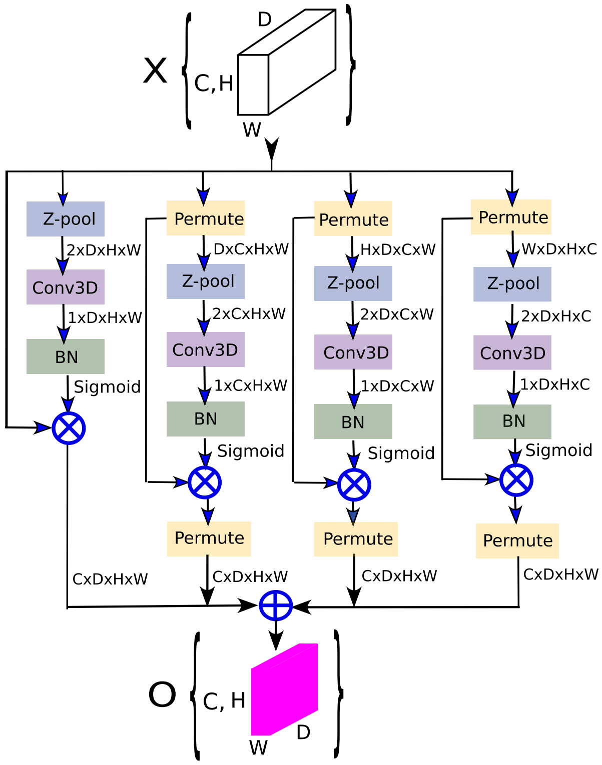


Fig. 2. Proposed quadlet cross attention module.

$$\hat{\mathbf{X}}^{l-1} = \frac{\mathbf{X}^{l-1} - \mu\left(\mathbf{X}^{l-1}\right)}{\sqrt{\sigma^2(\mathbf{X}^{l-1}) + \epsilon}} \cdot \gamma + \beta \tag{2}$$

where, $*$ denotes the convolution operation, $\mathbf{X}_j^{l-1}$ is the $j$th input tensor, $\mathbf{W}_p^l$ and $b_p^l$ are the weights and additive inductive bias of the $p$th filter bank in the $l$th convolution layer. BN is represented in (2), where $\mu$ and $\sigma^2$ represent the expectation and variance, $\gamma$ and $\beta$ are the learnable parameters during the training process. $R$ is activation function and calculated as $\text{ReLU}(x) = \max(0, x)$.

*2) Quadlet Attention:* This attention aims to model the cross dimensional interaction of input tensor, which include the number of feature maps, spectral depth, spatial height and width, respectively. The quadlet attention module $\mathcal{F}_{\text{quad}}(\cdot)$ takes the convolutional feature map with shape $(B, C, D, H, W)$ as input, where $C$ denotes the number of feature maps, $D$ is the spectral depth, and $H$, $W$ are the spatial height and width, respectively, and produces a calibrated feature map $\mathcal{O}$ of the same shape as the input

$$\mathcal{O} = \mathcal{F}_{\text{quad}}(\mathbf{X}; \theta) \tag{3}$$

where, $\theta$ represents the learnable parameters of the attention function. Fig. 2 illustrates the structure of the quadlet attention module which considers the convolution operation in four different branches to learn the dependencies between the dimensions $(D, H, W)$, $(C, H, W)$, $(D, C, W)$, and $(D, H, C)$ of the input tensor. Here, we ignore the batch size dimension and only consider the last four dimensions of the feature maps, i.e., $(C, D, H, W)$.

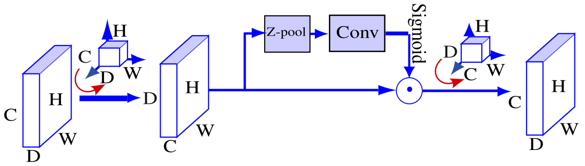


Fig. 3. Single branch in quadlet attention module, take the second branch as an example.

The quadlet attention mechanism is implemented via four independent branches for extraction of the cross-dimensional feature information, and they are named $F_c(\cdot)$, $F_d(\cdot)$, $F_h(\cdot)$, and $F_w(\cdot)$, respectively, according to their corresponding rotational dimensions. To establish the cross-dimensional relationship of input feature maps with a shape of $(B, C, D, H, W)$, the extracted features are aggregated using an elementwise addition operation as

$$F_{\text{quad}}(\mathbf{X}; \theta) = \frac{F_c(\mathbf{X}; \theta_c) \oplus F_d(\mathbf{X}; \theta_d) \oplus F_h(\mathbf{X}; \theta_h) \oplus F_w(\mathbf{X}; \theta_w)}{4} \tag{4}$$

where, $\theta_c, \theta_d, \theta_h$, and $\theta_w$ represent trainable parameters of the four branches of the quadlet attention module, i.e., $F_c$, $F_d$, $F_h$, and $F_w$, respectively. $\oplus$ is the elementwise addition operation. Each branch of the quadlet attention module is explained next

$$F_c(\mathbf{X}; \theta_c) = \sigma(\text{BN}(\text{Conv3D}(Z\text{−pool}(\mathbf{X})))) \otimes I(\mathbf{X}) \tag{5}$$

$$Z - \text{pool}(\mathbf{X}) = [\text{MaxPool}(\mathbf{X}), \text{AvgPool}(\mathbf{X})]. \tag{6}$$

The $F_c(\mathbf{X}; \theta_c)$ branch is used to build interaction among the spectral depth, spatial height and width of the feature maps. To do this, global maximum pooling and average pooling are first performed in the number of feature maps dimensions, as shown in (6), where MaxPool and AvgPool represent max pooling and average pooling, respectively. The results produce a feature map with a shape of $(2, D, H, W)$. After that, feature extraction is performed using 3-D convolution followed by a BN operation to produce the intermediate feature of dimensions $(1, D, H, W)$. To obtain the cross-dimensional attention weights, the intermediate features are passed through a sigmoid $(\sigma)$ function, and finally the dot product $(\otimes)$ with the input features denoted by identity function $I(\mathbf{X})$ is conducted to obtain the output features

$$F_d(\mathbf{X}; \theta_d) = \sigma(\text{BN}(\text{Conv3D}(Z\text{−pool}(\bar{\mathbf{X}}_{D,C,H,W})))) \otimes \bar{\mathbf{X}}_{D,C,H,W}. \tag{7}$$

The second branch $F_d(\mathbf{X}; \theta_d)$ is used to perform the interactions among the number of feature maps, the spatial height and width. First, the positional relationship is created between feature maps $C$ and spectral depth $D$ in the feature map by rotating the input tensor, and this produces feature maps $\bar{\mathbf{X}}_{D,C,H,W}$. Then global pooling and average pooling are performed in the spectral depth dimension and the resultant feature is concatenated to obtain the output feature map of shape $(2, C, H, W)$. The term "rotation" means the permutation of the dimensions of the input tensor. In Fig. 3, the red arrow represents the dimension permutation operation, which can be viewed as a rotation operation. Then, the 3-D convolutional layer, BN blocks are utilized in the last three dimensions, including the number of feature maps, height, and width, to capture the cross dimensional interaction. Finally, the obtained feature maps are passed through the sigmoid function to produce the corresponding attention weights. The obtained attention weights are produced with the original permuted feature maps, and the dimensions of the spectra and channels are exchanged again to obtain the final output of shape $(C, D, H, W)$, as shown in Fig. 3.

$$F_h(\mathbf{X};\theta_h) = \sigma(\mathrm{BN}(\mathrm{Conv3D}(Z\text{−pool}(\bar{\mathbf{X}}_{H,D,C,W})))) \otimes \bar{\mathbf{X}}_{H,D,C,W}. \tag{8}$$

In the third branch $F_h(\mathbf{X};\theta_h)$, the cross-dimensional attention weights between the spectral depth $D$, the number of feature maps $C$, and the spatial width $W$ are constructed. Similar to the second branch, the input feature map is firstly permuted between the number of feature maps $C$ and the spatial height $H$ to obtain a feature map $\bar{\mathbf{X}}_{H,D,C,W}$. Next, the height $H$ dimension is globally and averaged pooled and concatenated along the second dimension, i.e., spatial $H$ dimension, to obtain a feature map with shape $(2, D, C, W)$. The output is passed through a 3-D convolution layer, followed by BN and a sigmoid $(\sigma)$ function, to calculate the attention weights of spectral, channel, and width dimensions. Finally, the output is further rotated to keep same with its original input feature shape $(C, D, H, W)$

$$F_w(\mathbf{X};\theta_w) = \sigma(\mathrm{BN}(\mathrm{Conv3D}(Z\text{−pool}(\bar{\mathbf{X}}_{W,D,H,C})))) \otimes \bar{\mathbf{X}}_{W,D,H,C}. \tag{9}$$

The fourth branch $F_w(\mathbf{X};\theta_w)$ is similar to the second and third branches of quadlet attention to capture the relationship information among spectral, height, and channel. The input features are first permuted in $C$ and $W$ dimensions to obtain a feature map of shape $(W, D, H, C)$. Then, global and average pooling, followed by 3-D convolution and BN, are applied, which is then passed through the sigmoid activation function to obtain the cross-dimension attention weights. After the elementwise dot product between attention weights and input tensor, we obtain the output feature map having the same shape as its input.

The calibrated features of shape $(C, D, H, W)$ generated by each branch of quadlet attention module are then aggregated using elementwise addition and the result is divided by the total number of branches, as shown in (4). It can be seen from (4) that the triplet attention is a special case of the quadlet attention (shown in Fig. 2), which ignores the cross dimensional interaction in the channel dimension of the input tensor, and hence, we can rewrite the (4) as

$$\mathcal{O} = \mathcal{F}_{\mathrm{TA}}(\mathbf{X};\theta) = \frac{F_d(\mathbf{X};\theta_d) \oplus F_h(\mathbf{X};\theta_h) \oplus F_w(\mathbf{X};\theta_w)}{3} \tag{10}$$

where, $\mathcal{F}_{\mathrm{TA}}(\mathbf{X};\theta)$ denotes the triplet attention applied on input tensor $X$.

*3) Triplet Attention Aided Spectral-Spatial Residual Block:* The residual network has been well designed to deal with the gradient disappearance problem that occurs during the training of hyperspectral classification tasks and has generated great interest in the remote sensing research community. In order to extract more robust spectral and spatial information, the triplet attention aided SSRN, which incorporates a triplet attention layer after every residual block is introduced. The output of quadlet attention is then passed through a 3-D convolution followed by BN and the ReLU activation function to perform feature normalization with the help of $(1, 1, 1)$ the pointwise convolutional kernel. The gray shading structures in Fig. 1 show the triplet attention aided spectral-spatial residual block. Why choosing triplet rather than quadlet in the spectral-spatial residual block is mainly based on the tradeoff between classification performance and computational cost. In spectral-spatial residual blocks, both the spectral and spatial residual blocks are repeated twice, which requires us to use the attention module four times. In comparison to the triplet attention module, the quadlet attention module has more parameters and operations. If it is also used in subsequent spectral-spatial residual blocks, the computational cost of the network will be increased significantly.

To learn the robust representation, the normalized convolutional feature input is passed through four consecutive triplet attention aided spectral-spatial residual blocks. Each residual block consists of a Conv3D layer followed by a BN and a ReLU activation layers, and these three primitive steps are repeated twice in a residual block for enhancing the feature extraction as well as forward and backward propagation of information. Suppose $\mathbf{X}^{k-1}$ represents the input feature map of the spectral-spatial residual blocks, which is parameterized with $F_{\mathrm{RN}}(\mathbf{X}^{k-1};\omega_1,\omega_2)$ and the output feature map is then passed through the triple attention layer $F_{\mathrm{TA}}(\cdot)$, and the final feature representation can be calculated as follows:

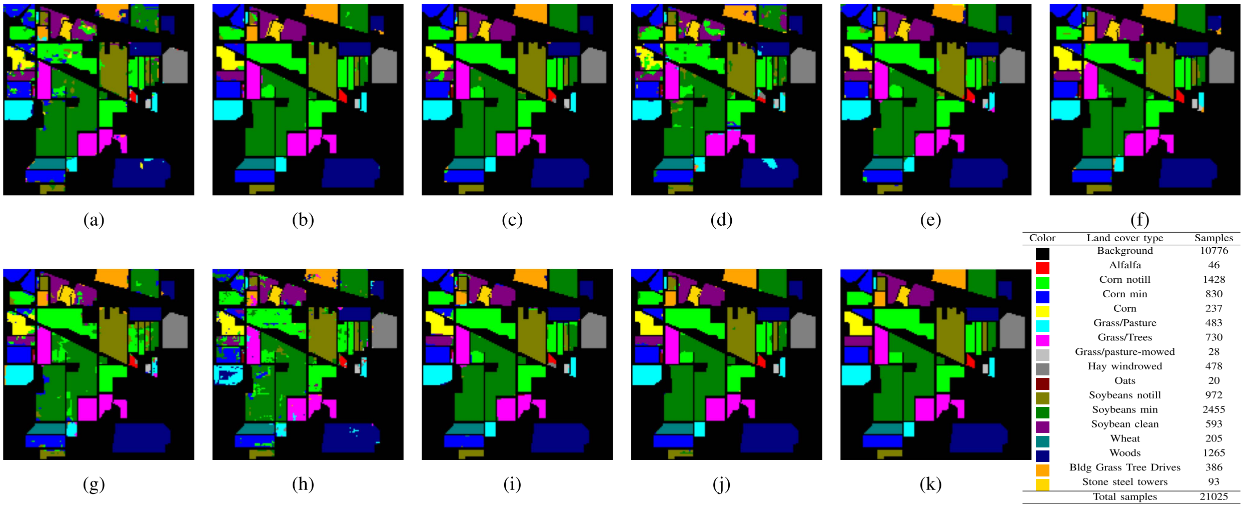
$$\mathbf{X}^{k+1} = I(\mathbf{X}^{k-1}) + F_{\mathrm{TA}}(F_{\mathrm{RN}}(\mathbf{X}^{k-1};\omega_1,\omega_2)) \tag{11}$$

$$F_{\mathrm{RN}}(\mathbf{X}^{k-1};\omega_1,\omega_2) = \mathrm{ReLU}(\mathrm{BN}(\mathbf{X}^{k} * \mathbf{W}^{k+1} + \mathbf{b}^{k+1}) \tag{12}$$

$$\mathbf{X}^{k} = \mathrm{ReLU}(\mathrm{BN}(\mathbf{X}^{k-1} * \mathbf{W}^{k} + \mathbf{b}^{k})) \tag{13}$$

where, $I(\mathbf{X}^{k-1})$ is the skip function connects input to the output of a residual unit, $\mathbf{X}^{k+1}$ is the output of the triple attention aided spectral-spatial residual block. $\omega_1, \omega_2$ are the parameters of the triple attention aided spectral-spatial residual blocks, and $\omega_1 = \{\mathbf{W}^k, \mathbf{W}^{k+1}\}$, and $\omega_2 = \{\mathbf{b}^k, \mathbf{b}^{k+1}\}$ denote the weight matrix and bias associated with the $k$th and $(k+1)$th 3-D convolutional layers, respectively.

In Fig. 1, we consider $L_x$ as 2, which means two units of both spectral and spatial residual blocks used sequentially are investigated in order to improve the discriminative capability of each residual block. One may readily distinguish between spectral feature learning and spatial feature learning based on the convolutional operation used to perform convolution in depth or spatially to extract the robust feature representation. Each residual block having 24 kernels shown in Fig. 1 and extracting spectrally focused spatial features is done with the last two


(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k)

| Color | Land cover type | Samples |
|---|---|---|
| ■ | Background | 10776 |
| ■ | Alfalfa | 46 |
| ■ | Corn notill | 1428 |
| ■ | Corn min | 830 |
| ■ | Corn | 237 |
| ■ | Grass/Pasture | 483 |
| ■ | Grass/Trees | 730 |
| ■ | Grass/pasture-mowed | 28 |
| ■ | Hay windrowed | 478 |
| ■ | Oats | 20 |
| ■ | Soybeans notill | 972 |
| ■ | Soybeans min | 2455 |
| ■ | Soybean clean | 593 |
| ■ | Wheat | 205 |
| ■ | Woods | 1265 |
| ■ | Bldg Grass Tree Drives | 386 |
| ■ | Stone steel towers | 93 |
|  | Total samples | 21025 |

Fig. 4. Classification maps generated by different models over IN dataset. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet. (k) Ground truth.

ResBlocks whereas extracting spatially focused spectral features is done with the first two ResBlocks. Table I shows the shapes of the applied kernels and strides for both the spectral and spatial residual blocks. The values of the hyperparameters (the number of feature maps, the patch size and convolutional kernel size) are determined by refereeing to the widely used setting for A2S2K-ResNet [25] and SSRN [31]. As shown by many existing research, the 3-D convolutional operation enables the model to capture spatial and spectral dependencies among the input data, which is essential for achieving high performance on HSIs classification. Thus, 3-D convolutional operation is also employed here. The receptive field of the convolutional filters, which regulates the amount of features to be learned by the model, is determined by the kernel size in a 3-D convolutional operation. In our method, the kernel size is selected by referring to the widely accepted setting, such as those in SSRN [31]. The kernel of size (7,1,1) with a stride of (2,1,1) in the first convolutional layer is selected to reduce the spectral dimension for saving computation resources. A kernel size of (7,1,1) is used in the first two residual blocks to extract spectral features whereas a kernel size of (1,3,3) is applied in the last two residual blocks to extract spatial features.

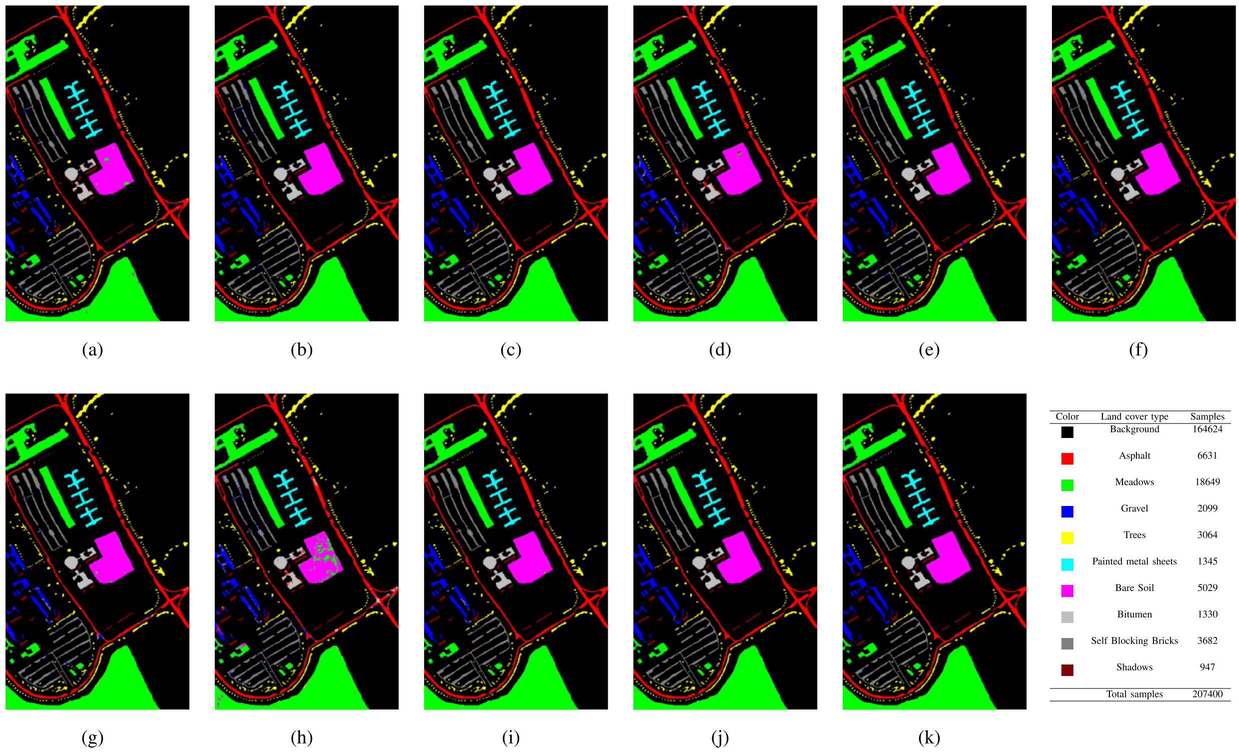
As a result, the discriminative capability of the proposed model is increased by cooperative learning of spectral and spatial information. In addition, the triple attention is embedded in the residual blocks to achieve spectral and spatial cross dimensions interaction of the residual feature representation. The triplet attention in the spectral-spatial residual block captures the spectral and spatial cross dimensional relationship in spectral depth, spatial height and width, respectively, ignoring the dimension of the number of channels.

## IV. Experiments and Discussion

### A. Datasets

The experiments are conducted on three widely used HSI datasets, including IN, University of Pavia (UP), Salinas (SA), and University of Houston (UH). The details of each dataset are explained in the following.

1) The IN dataset was collected by the airborne visible/infrared imaging spectrometer (AVIRIS) sensor over the test site in northwest India in 1992. The IN dataset has a size of 145 × 145 pixels in the spatial dimension and contains 224 bands in the spectral dimension. Twenty four bands were excluded due to the effect of water vapor absorption. The wavelength range is 400–2500 nm, and the spatial resolution of each pixel point is 20 m. Out of 21 025 pixels, a total of 10 249 pixels that contain 16 different kinds of vegetation classes are selected. 10% of the selected samples are used for training, 10% for validation, and 80% for testing. Fig. 4 shows the distribution of various categories in the IN dataset and the colour representation for each land cover category in the IN dataset.
2) The UP dataset was acquired by the reflective optics system imaging spectrometer sensor in 2001 at the UP in northern Italy. It contains 610 × 340 pixels, each with a spatial resolution of 1.3 m. The spectral dimension is 103 with wavelengths in the 430–860 nm range. All the pixels with labels were classified into nine different urban land cover types. The color representation and the number of instances for each category are illustrated in Fig. 5. For the UP dataset, 5% of the samples are used as training samples, 10% as validation samples, and 85% used as test samples.
3) The SA dataset was acquired by the AVIRIS sensor in SA Valley, California. The dataset has 512 × 217 pixels and 204 spectral bands with a spatial resolution of 3.7 m. The 54 129 pixels with labels were divided into a total of 16 different terrestrial categories. Fig. 6 shows the ground truth and the number of labeled samples in the SA dataset. During the experiment, 5% of the samples are selected as training samples, 10% as validation samples, and the remaining 85% are considered as test set.



| Color | Land cover type | Samples |
|---|---|---|
| ⬛ | Background | 164624 |
| 🟥 | Asphalt | 6631 |
| 🟩 | Meadows | 18649 |
| 🟦 | Gravel | 2099 |
| 🟨 | Trees | 3064 |
| 🟦 | Painted metal sheets | 1345 |
| 🟪 | Bare Soil | 5029 |
| ⬜ | Bitumen | 1330 |
| ⬛ | Self Blocking Bricks | 3682 |
| 🟫 | Shadows | 947 |
|  | Total samples | 207400 |

Fig. 5. Classification maps generated by different models on UP datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet. (k) Ground truth.

4) The UH dataset was collected over the campus of the UH using a compact airborne spectrographic imager with a spatial resolution of 2.5 m and a spectral range of 380–1050 nm. The dataset consists of 144 bands and covers an area of 340×1905 pixels. It contains 15 land cover classes. Fig. 7 shows the ground truth and the number of labeled samples in the SA dataset. During the experiment, 5% of the samples are selected as training samples.

## B. Experiment Setup

To evaluate the effectiveness of the proposed QuadNet method, we compare it with various state-of-the-art models using these three adopted datasets. The reference models include a HybridSN [30], a deep SSRN [31], and its variants with different attention mechanisms, e.g., A2S2K-ResNet [25], DBMA [33], DBDA [46], RSSAN [43], and SPANet [55], respectively. Besides, two representative transformer-based methods, vision transformer (ViT) [56] and multimodal fusion transformer (MFT) [52], are selected for comparison. It should be noted that the MFT method used in our experiments is a modified version where only the hyperspectral branch of that in [30] is kept and the LiDAR part is removed. This is because this study only involves HSI classification.

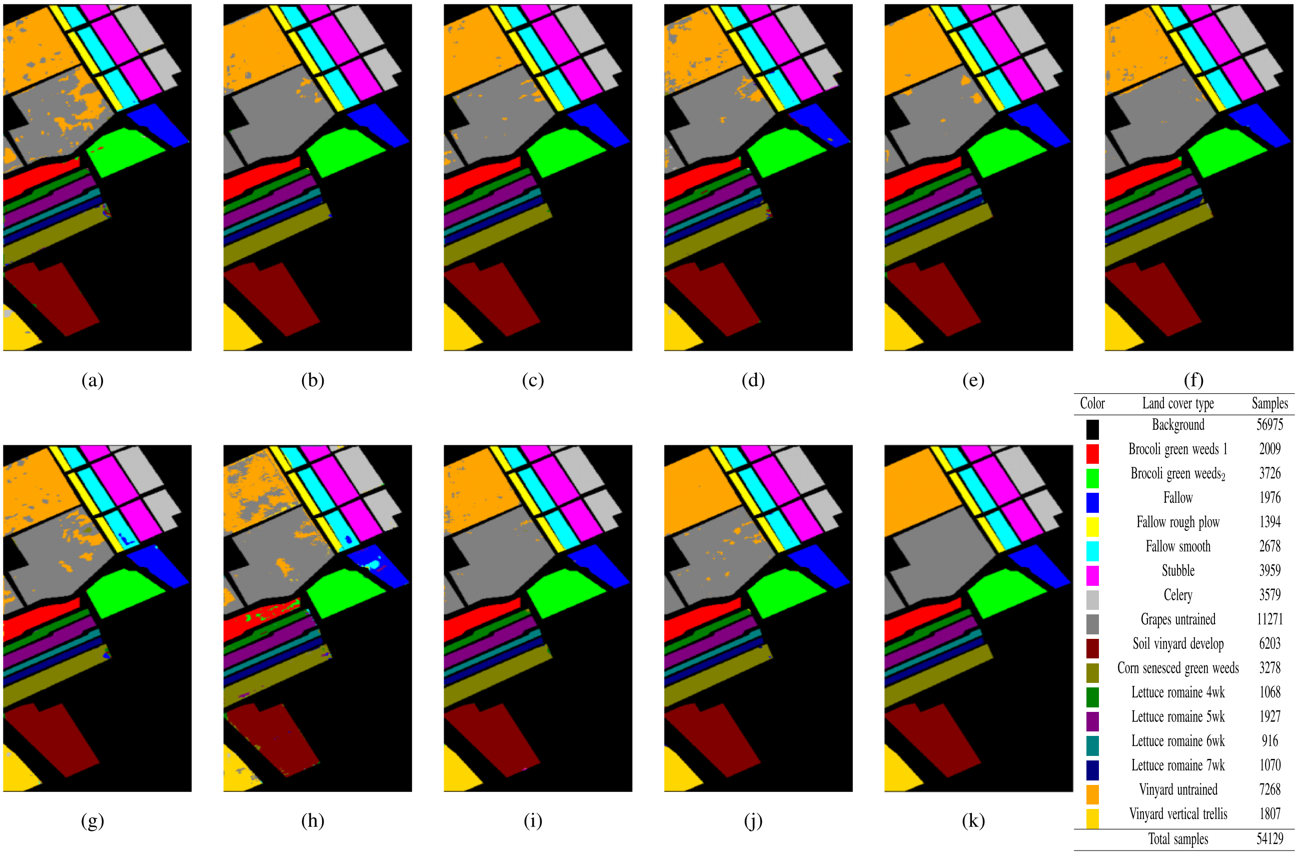
For all the models, cross-entropy is used as a loss function to measure the classification effect of the model parameters during training, and the Adam optimizer is chosen to back-propagate the error gradient and update the model weights in the network with a learning rate 0.001. The models are trained for 200 epochs in each experiment. In addition, in order to prevent the model from overfitting on the training set, an early stop strategy is used. When the loss value of the model on the validation set does not decrease for 50 consecutive times, the model training is terminated and the weights with the lowest loss value on the validation set are saved. The model weights corresponding to the lowest loss values on the validation set are used to evaluate the test set. The whole experiment is repeated three times, and the AA and standard deviation of the three experiments are obtained to avoid the randomness that may exist in a single run. All experiments are performed on the compute Canada server with 64 GB memory.

For the input tensor (i.e., extracted small patches), the min–max scaling normalization is implemented before feeding it into the deep learning models used in our experiments. Min–max scaling normalization is a commonly used technique to scale the input features to a fixed range of values, [–0.5, 0.5] here. It can also help reduce the impact of different scales on the learning process and improve the performance of the model.

## C. Classification Results

In this section, the qualitative and quantitative experimental results are analyzed. Three different evaluation metrics, including overall accuracy (OA), average accuracy (AA), and kappa coefficients (Kappa) are used for model evaluation.

*1) Results on Random Split Datasets:* We first conduct experiments over the IN, UP, and UH datasets to compare the performance with and without quadlet attention, the results are



| Color | Land cover type | Samples |
|---|---|---|
| ■ | Background | 56975 |
| ■ | Brocoli green weeds 1 | 2009 |
| ■ | Brocoli green weeds$_2$ | 3726 |
| ■ | Fallow | 1976 |
| ■ | Fallow rough plow | 1394 |
| ■ | Fallow smooth | 2678 |
| ■ | Stubble | 3959 |
| ■ | Celery | 3579 |
| ■ | Grapes untrained | 11271 |
| ■ | Soil vinyard develop | 6203 |
| ■ | Corn senesced green weeds | 3278 |
| ■ | Lettuce romaine 4wk | 1068 |
| ■ | Lettuce romaine 5wk | 1927 |
| ■ | Lettuce romaine 6wk | 916 |
| ■ | Lettuce romaine 7wk | 1070 |
| ■ | Vinyard untrained | 7268 |
| ■ | Vinyard vertical trellis | 1807 |
| | Total samples | 54129 |

Fig. 6. Classification maps generated by different models on SA datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet. (k) Ground truth.

TABLE II
Classification Results With Triplet and Quad Attention Module

| | Metrics | IN | UP | UH |
|---|---|---|---|---|
| Triplet | OA | 97.54±0.17 | 99.86±0.03 | 95.45±0.37 |
| | AA | **98.45±0.17** | 99.82±0.05 | 96.08±0.34 |
| | Kappa | 97.19±0.19 | 99.82±0.04 | 95.09±0.40 |
| quad | OA | **98.22±0.28** | **99.88±0.05** | **96.46±0.85** |
| | AA | 98.18±0.52 | **99.83±0.08** | **96.83±0.68** |
| | Kappa | **97.97±0.32** | **99.85±0.06** | **96.17±0.92** |

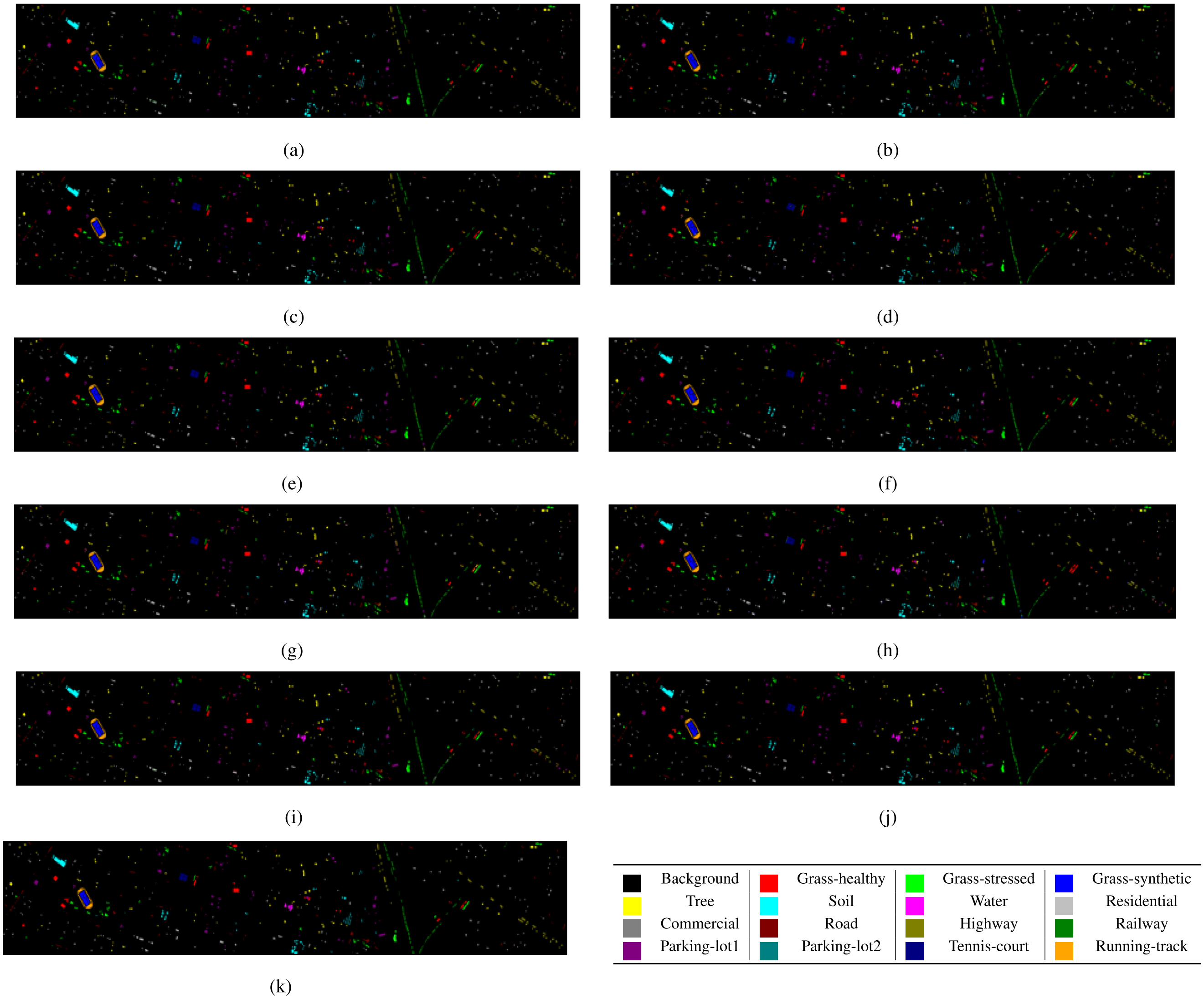
The bold values denote the best accuracy.

shown in Table II. The table shows that the classification performances in terms of OA, AA, and Kappa have been improved after incorporating the additional dimension, i.e., the number of feature maps.

The classification results obtained by eight different methods with 10% IN training data are given in Table III. From the table, it can be seen that the HybridSN and RSSAN networks obtain lower OA, i.e., HybridSN = 87.83% and RSSAN = 85.57%. Among the six CNN-based networks incorporated different attention modules, the method proposed in this article achieves the best classification results with OA = 98.22%, AA = 98.18%, and Kappa = 97.97%. The accuracy of class 9 reaches 100%. The classification results of QuadNet outperform all other networks that contain different spectral and spatial attention modules, including A2S2K-ResNet (OA = 97.81%), DBMA (OA = 96.23%), DMDA (OA = 96.46%), and SPANet (OA = 91.16%). This is because the proposed multibranches cross-attention can simultaneously establish the dependencies among four different dimensions, i.e., the number of feature maps, spectrum, spatial height, and spatial width, thus achieving higher classification results. Besides, the accuracy of ViT is relatively lower than that of QuadNet, while MFT shows small decrease in performance compared with QuadNet.

Fig. 4 shows the classification results of different methods on IN dataset. It can be seen that a large amount of noise appeared in the classification maps obtained using the HybridSN, RSSAN, and ViT methods, indicating that a large number of pixels are misclassified. The classification maps obtained by the A2S2K-ResNet network have some confusion between Alfalfa (red) and Hay-Windrowed (dark gray), and the classification maps obtained by the DBMA and SSRN methods have confusion between soybean-notill (yellow-brown) and corn-notill (light green). The classification map obtained by the QuadNet network proposed in this article is the closest to the true map, thus proving its superiority over the other seven methods. However, a little misclassification happened in the boundary region between corn-notill (light green) and soybean-notill (dark green).

To verify the classification performance sensitivity of different sample sizes for these models, the classification results of the proposed QuadNet and other models are compared using



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

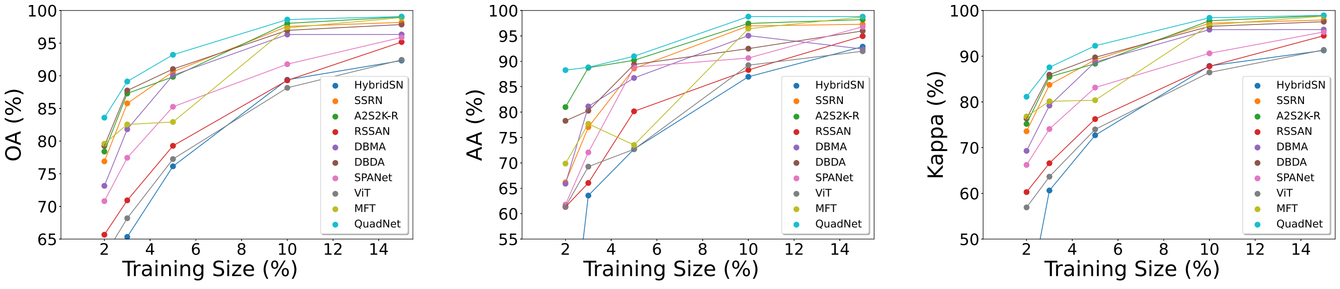
(k)

Fig. 7. Classification maps generated by different models on UH datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet. (k) Ground truth.

TABLE III
Classification Results of Different Methods With 10% Training Samples on the in Datasets

| Class | HybridSN | SSRN | A2S2K-R | RSSAN | DBMA | DBDA | SPANet | ViT | MFT | QuadNet |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 97.57±1.77 | 91.27±12.35 | 96.30±5.24 | 83.97±11.36 | 92.98±5.22 | **100±0** | 90.49±9.15 | 80.56±6.74 | 95.65±6.15 | 98.09±1.35 |
| 2 | 82.93±3.55 | 96.76±1.85 | 96.65±1.56 | 81.61±4.61 | 94.12±0.67 | 95.27±1.57 | 84.80±3.92 | 80.26±1.93 | 95.47±0.91 | **98.62±0.57** |
| 3 | 77.49±3.31 | 96.22±0.51 | **97.49±0.34** | 84.39±4.87 | 96.06±1.54 | 97.25±0.41 | 89.20±3.78 | 79.11±9.04 | 96.05±0.50 | 96.30±2.30 |
| 4 | 80.35±3.41 | **98.54±1.67** | 90.48±4.13 | 77.28±12.59 | 95.56±3.00 | 94.27±1.60 | 88.05±4.62 | 89.68±7.71 | 97.63±1.33 | 97.27±2.39 |
| 5 | 91.45±0.62 | 98.54±0.52 | 99.54±0.14 | 83.54±4.97 | 94.76±0.72 | 97.19±0.76 | 94.89±3.86 | 85.82±4.50 | 96.47±1.03 | **99.91±0.12** |
| 6 | 93.78±1.76 | 97.76±0.07 | 98.82±0.36 | 97.28±2.76 | 97.44±0.48 | **99.65±0.38** | 97.20±1.06 | 90.24±3.02 | 98.53±0.42 | 98.87±0.56 |
| 7 | 80.63±11.67 | **100±0** | 96.75±2.30 | 96.67±4.71 | 85.08±7.01 | 60.27±5.33 | 78.33±20.95 | 71.81±6.59 | 98.55±2.05 | 98.48±2.14 |
| 8 | 99.14±0.64 | 99.05±0.68 | 96.39±0.64 | 95.12±0.91 | **99.91±0.12** | **99.91±0.12** | 96.34±2.28 | 95.68±1.68 | 97.97±1.64 | 99.74±0.21 |
| 9 | 45.36±17.80 | **100±0** | 95.24±6.73 | 82.75±14.90 | 90.45±3.38 | 82.74±12.57 | 85.71±11.66 | 93.33±9.43 | 73.12±9.28 | **100±0** |
| 10 | 88.56±1.95 | 97.29±1.62 | 97.09±0.44 | 91.31±2.06 | 92.71±1.13 | 96.59±0.50 | 82.18±9.39 | 91.87±1.28 | 97.35±0.88 | **98.88±0.33** |
| 11 | 88.92±2.84 | 97.09±1.25 | **98.72±0.45** | 85.88±6.69 | 98.00±0.56 | 97.22±0.54 | 94.23±1.28 | 86.03±2.10 | 98.06±0.53 | 97.12±1.00 |
| 12 | 84.09±4.15 | 96.56±1.10 | 98.02±0.45 | 66.70±3.23 | 97.56±1.12 | 93.19±2.62 | 89.66±8.21 | 83.85±2.85 | 93.94±2.45 | **98.32±0.28** |
| 13 | 83.34±3.83 | 99.19±0.57 | **100±0** | 88.25±7.81 | 94.05±1.11 | 96.88±2.38 | 98.99±0.56 | 97.01±1.32 | 98.19±1.48 | 99.40±0.85 |
| 14 | 93.85±1.57 | **99.80±0.21** | 98.85±0.05 | 89.02±5.35 | 97.59±0.86 | 98.26±0.84 | 98.62±0.90 | 87.40±2.43 | 98.12±0.20 | 99.48±0.26 |
| 15 | 85.77±13.06 | 96.92±1.12 | 97.47±1.33 | 84.57±7.79 | 96.20±1.86 | 90.85±2.26 | 92.49±1.22 | 87.36±6.57 | 97.55±1.34 | **99.14±0.54** |
| 16 | 93.83±5.70 | 81.73±5.17 | **97.31±1.07** | 91.68±6.37 | 88.72±2.74 | 87.62±1.30 | 83.85±11.97 | 91.78±4.32 | 91.95±0.50 | 91.24±3.33 |
| OA | 87.83±1.79 | 97.35±0.15 | 97.81±0.25 | 85.57±2.64 | 96.23±0.09 | 96.46±0.42 | 91.16±0.74 | 86.03±1.52 | 97.06±0.33 | **98.22±0.28** |
| AA | 85.82±1.51 | 96.67±0.34 | 97.19±0.24 | 86.25±1.56 | 94.45±0.87 | 92.95±0.61 | 90.31±0.74 | 86.99±1.59 | 95.29±0.84 | **98.18±0.52** |
| Kappa | 86.09±2.06 | 96.98±0.17 | 97.51±0.28 | 83.49±3.05 | 95.70±0.10 | 95.97±0.48 | 89.85±0.82 | 84.01±1.74 | 96.64±0.38 | **97.97±0.32** |

The bold values denote the best accuracy.



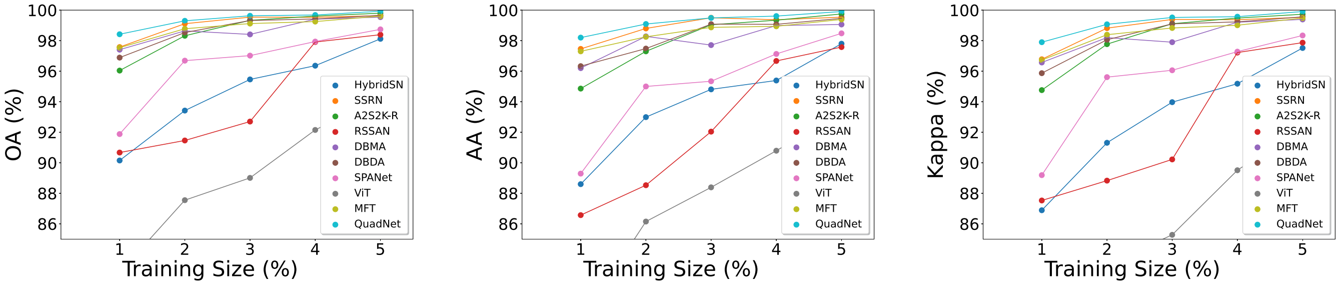
(a) (b) (c)

Fig. 8. OA, AA, and Kappa values over IN dataset under different training samples. (a) OA. (b) AA. (c) Kappa.

TABLE IV
CLASSIFICATION RESULTS OF DIFFERENT METHODS WITH 5% TRAINING SAMPLES ON THE UP DATASETS

| Class | HybridSN | SSRN | A2S2K-R | RSSAN | DBMA | DBDA | SPANet | ViT | MFT | QuadNet |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 95.00±1.66 | 96.07±5.01 | 99.54±0.17 | 94.86±1.84 | 99.86±0.08 | 99.49±0.12 | 99.13±0.33 | 94.48±0.82 | 99.83±0.04 | **99.91±0.04** |
| 2 | 98.27±0.90 | **99.98±0.02** | 99.94±0.02 | 98.84±0.49 | 99.78±0.09 | 99.97±0.01 | 99.78±0.13 | 93.49±3.05 | 99.91±0.05 | 99.97±0.02 |
| 3 | 92.40±3.43 | 97.68±2.31 | 99.46±0.54 | 83.97±12.10 | 97.07±0.88 | 99.55±0.04 | 95.42±2.24 | 82.39±3.58 | **100±0** | 99.72±0.40 |
| 4 | 99.79±0.16 | 99.34±0.53 | **99.85±0.11** | 99.12±0.23 | 98.84±0.30 | 97.74±0.56 | 99.25±0.23 | 97.20±0.76 | 98.32±0.13 | 99.81±0.11 |
| 5 | 99.85±0.21 | 99.97±0.04 | **100±0** | 98.97±0.99 | 99.85±0.04 | 99.62±0.25 | 99.97±0.04 | 98.63±1.16 | 99.71±0.08 | 99.88±0.04 |
| 6 | 95.62±2.20 | **100±0** | 99.98±0.01 | 96.20±2.81 | 99.84±0.09 | 99.92±0.06 | 99.67±0.10 | 93.19±1.88 | 99.84±0.15 | 99.97±0.02 |
| 7 | 93.52±2.56 | **100±0** | 99.91±0.13 | 92.43±4.20 | 99.79±0.23 | **100±0** | 98.23±0.78 | 76.63±4.16 | 99.77±0.33 | 99.88±0.17 |
| 8 | 93.04±1.86 | 98.82±0.92 | 99.10±0.19 | 91.85±1.62 | 99.29±0.37 | 99.25±0.26 | 92.34±1.84 | 87.99±6.00 | 97.66±0.26 | **99.41±0.06** |
| 9 | 99.41±0.47 | 99.83±0.16 | 99.75±0.10 | 98.89±1.39 | 96.78±0.38 | 98.82±0.16 | 99.55±0.29 | 98.85±0.70 | 99.38±0.27 | **99.92±0.12** |
| OA | 96.76±1.29 | 99.04±0.80 | 99.78±0.02 | 96.32±1.61 | 99.49±0.03 | 99.61±0.04 | 98.70±0.04 | 92.40±2.07 | 99.56±0.04 | **99.88±0.05** |
| AA | 96.33±1.41 | 99.08±0.44 | 99.73±0.02 | 95.01±2.00 | 99.01±0.08 | 99.37±0.07 | 98.15±0.24 | 91.43±1.42 | 99.38±0.06 | **99.83±0.08** |
| Kappa | 95.70±1.72 | 98.72±1.07 | 99.70±0.02 | 95.12±2.14 | 99.33±0.04 | 99.48±0.05 | 98.27±0.05 | 89.92±2.86 | 99.42±0.05 | **99.85±0.06** |

The bold values denote the best accuracy.



(a) (b) (c)

Fig. 9. OA, AA, and Kappa values over UP dataset under different training samples. (a) OA. (b) AA. (c) Kappa.

different proportions of training data. Table 8 shows the corresponding OA, AA, and Kappa results. For the IN dataset, 2%, 3%, 5%, 10%, and 15% are chosen as the training data, and the corresponding OA, AA, and kappa values of the QuadNet model are depicted by the light blue curves in Fig. 8. It can be seen from the figure that the classification performance of various models increases with the number of training samples. The proposed QuadNet has the best classification performance for ranging amount of training samples. Under the condition of limited training sizes, QuadNet also obtains highest classification accuracy, whereas the rest of the models, such as HybridSN and RSSAN, are relatively worse in terms of OA, AA, and Kappa.

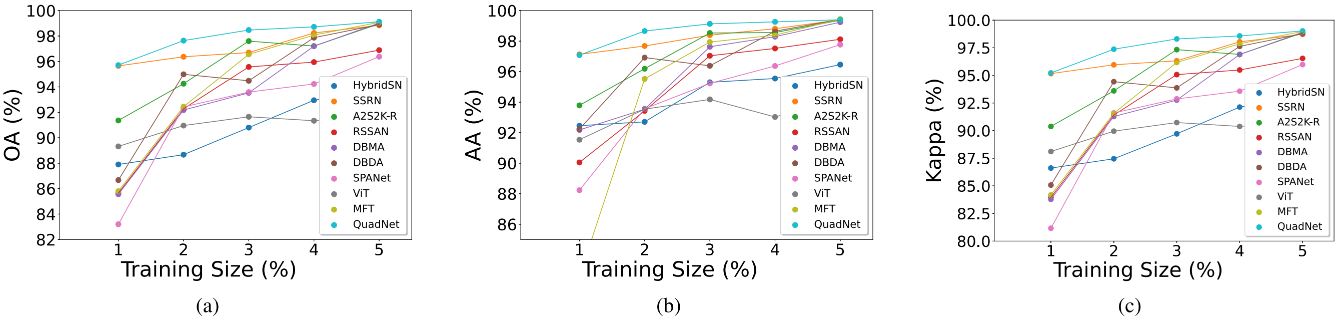
The classification results of different models on the UP dataset are illustrated in Table IV. The UP dataset has more samples with ground truth than the IN dataset, therefore, the OA of all methods is higher than 95% with only 5% as training samples. Among all the methods, the HybridSN, RSSAN, ViT similarly show the lower OA in classification. Nevertheless, the QuadNet achieves the highest OA (99.88%), AA (99.83%), and kappa (99.85%).

Fig. 5 shows the classification graphs obtained by different models when 5% data are used for training. It can be seen that large differences are observed between the ground truth and the classification maps obtained by HybridSN, RSSAN. For example, there is an obvious confusion of pixels between bare soil (pink) and meadows (light green) in the maps. Since the training data are sufficient, SSRN, A2S2K ResNet, DBDA, DBMA, and the QuadNet proposed in this article, all achieve an accuracy greater than 99.5%. Fig. 9 shows the OA, AA, and kappa obtained under different number of training samples. Due to the availability of sufficient labeled samples in UP datasets, most methods demonstrate good classification performance. However, to assess the model's capability with fewer samples,

TABLE V
Classification Results of Different Methods With 5% Training Samples on the SA Datasets

| Class | HybridSN | SSRN | A2S2K-R | RSSAN | DBMA | DBDA | SPANet | ViT | MFT | QuadNet |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 99.61±0.43 | **100±0** | **100±0** | 99.50±0.59 | **100±0** | **100±0** | 99.90±0.03 | 99.94±0.05 | **100±0** | 99.94±0.08 |
| 2 | 99.38±0.08 | **100±0** | 99.88±0.16 | 99.68±0.27 | 99.90±0.15 | 99.72±0.08 | 99.32±0.96 | 87.95±10.21 | 99.98±0.03 | **100±0** |
| 3 | 97.02±0.76 | 94.16±6.13 | 98.06±2.75 | 98.15±0.86 | 99.90±0.10 | 99.47±0.75 | 97.40±2.71 | 92.95±2.91 | 99.68±0.37 | **100±0** |
| 4 | 98.82±1.06 | 98.45±0.20 | 99.69±0.17 | 98.09±0.95 | 97.85±0.80 | 99.46±0.20 | **99.86±0.11** | 97.24±1.33 | 99.00±0.49 | 99.28±0.42 |
| 5 | 90.16±12.56 | **100±0** | 99.74±0.30 | 98.89±0.74 | 99.63±0.27 | 99.61±0.34 | 96.77±3.23 | 92.13±0.92 | 98.87±0.70 | 99.99±0.02 |
| 6 | 99.88±0.13 | **100±0** | 99.98±0.03 | 99.35±0.27 | **100±0** | 99.99±0.01 | 99.97±0.04 | 99.99±0.01 | 99.88±0.07 | **100±0** |
| 7 | 97.82±0.66 | **100±0** | **100±0** | 99.53±0.49 | 99.98±0.03 | 99.91±0.12 | 98.65±1.01 | 97.84±2.12 | 99.98±0.02 | **100±0** |
| 8 | 95.13±0.38 | 96.63±0.91 | 96.95±3.01 | 90.80±2.72 | 99.09±0.22 | 97.33±0.22 | 96.13±0.55 | 86.74±3.73 | 98.11±0.68 | **99.63±0.27** |
| 9 | 98.93±0.74 | 99.93±0.08 | 97.49±3.00 | 99.14±0.29 | 99.84±0.16 | 99.89±0.05 | 99.09±0.53 | 98.61±1.37 | **99.96±0.06** | 99.87±0.04 |
| 10 | 98.02±1.03 | 99.64±0.18 | 99.42±0.07 | 98.56±0.47 | 99.57±0.38 | 99.19±0.29 | 96.16±1.10 | 95.34±0.41 | 98.80±0.39 | **99.95±0.05** |
| 11 | 92.89±1.75 | 98.01±1.01 | 97.69±2.38 | 94.98±1.71 | 97.66±1.04 | 98.41±0.62 | 93.75±1.00 | 94.47±1.54 | **98.50±0.79** | 97.60±0.40 |
| 12 | 97.81±1.75 | **100±0** | 99.60±0.36 | 99.12±0.45 | 99.68±0.16 | 99.98±0.03 | 99.41±0.37 | 96.11±2.36 | 99.86±0.16 | **100±0** |
| 13 | 97.61±1.05 | 99.79±0.16 | 99.32±0.32 | 99.74±0.21 | 98.38±1.43 | 99.96±0.06 | 98.10±1.73 | 95.17±1.34 | 99.97±0.10 | **100±0** |
| 14 | 98.50±0.19 | **100±0** | 99.27±0.81 | 99.34±0.18 | 99.85±0.14 | 99.59±0.34 | 98.97±0.76 | 97.47±1.29 | 98.83±0.28 | 99.34±0.23 |
| 15 | 74.40±1.44 | 89.81±10.39 | 94.11±0.88 | 91.33±1.77 | 94.04±1.49 | **95.18±2.04** | 84.43±2.07 | 74.66±7.69 | 93.93±3.30 | 94.88±0.30 |
| 16 | 98.94±0.43 | **100±0** | 99.96±0.61 | 99.24±0.49 | **100±0** | **100±0** | 99.54±0.44 | 87.72±0.75 | 99.87±0.14 | **100±0** |
| OA | 93.21±0.62 | 97.32±1.92 | 98.03±1.11 | 96.24±0.56 | 98.75±0.28 | 98.60±0.28 | 95.93±0.33 | 90.17±1.47 | 98.50±0.41 | **99.10±0.01** |
| AA | 95.93±0.75 | 98.59±0.98 | 98.82±0.76 | 97.84±0.20 | 99.08±0.12 | 99.23±0.19 | 97.34±0.41 | 93.40±0.70 | 99.17±0.28 | **99.40±0.02** |
| Kappa | 92.46±0.69 | 97.02±2.13 | 97.80±1.24 | 95.81±0.62 | 98.61±0.31 | 98.44±0.31 | 95.47±0.36 | 89.06±1.62 | 98.33±0.46 | **99.00±0.01** |

The bold values denote the best accuracy.



(a) (b) (c)

Fig. 10. OA, AA, and Kappa values over SA dataset under different training samples. (a) OA. (b) AA. (c) Kappa.

we evaluate the classification accuracy using a smaller number of samples. For the UP dataset, the training sample proportions of 1%, 2%, 3%, 4%, and 5% are considered, and it can be seen from the figure that the proposed model still achieves better results with small training sample proportions. When only 1% are used for training, QuadNet achieves the highest OA (98.42%), which is better than SSRN (97.58%), A2S2K ResNet (96.04%) DBMA (97.41%), and DBDA (96.89%), as shown in Fig. 9(a). In addition, it can be seen that HybirdSN and RSSAN are less effective under small sample conditions.

Table V lists the average OA, AA, Kappa, and their test standard deviations based on three runs using 5% of the SA dataset. Similar to the IN and UP datasets, the proposed method outperforms other network models. Specifically, the QuadNet method achieves an OA of 99.10%, while SSRN, DBMA, and DBDA are 98.03%, 98.75%, and 98.60%, respectively. HybridSN, RSSAN, ViT are relatively less effective. In terms of AA and Kappa, the proposed QuadNet also display the highest scores compared with SSRN, DBMA, and A2S2K ResNet. In addition, the deviations obtained from the three experiments show that QuadNet has the lowest deviations for OA (0.01%), AA (0.02%), and Kappa (0.01%), which is lower than A2S2K ResNet, DBMA, and DMDA. This indicates that the proposed network has higher stability.

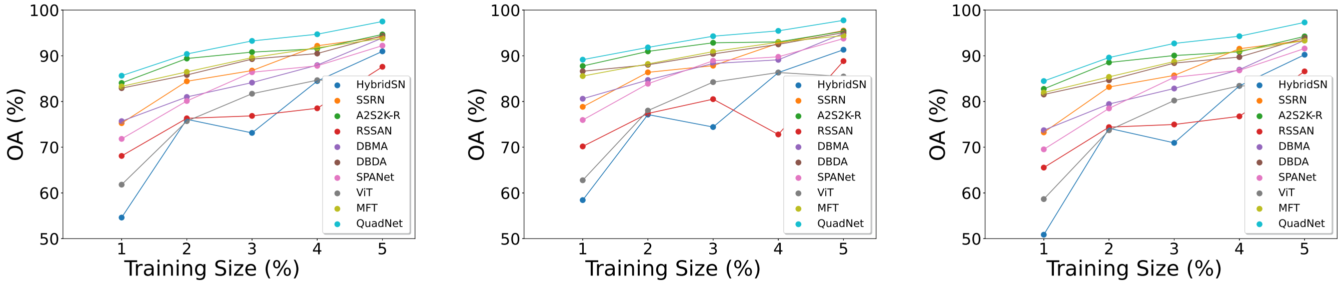
Fig. 6 illustrates the classification maps generated by different methods and the ground truth. It can be seen that the classification maps obtained by QuadNet are the closest to the ground truth, while other methods, such as DBMA, RSSAN, and SSRN, display more confusion between Vineyard untrained (orange) and grapes untrained (dark gray), resulting in a lower classification accuracy. Similar to UP datasets, the SA dataset also provides sufficient labeled samples. Therefore, we conduct experiments using fewer training dataset proportions of 1%, 2%, 3%, 4%, and 5%, as shown in Fig. 10. Again, QuadNet achieves better results than all other models at different amounts of training sets.

To further demonstrate the effectiveness and robustness of the proposed model, a relatively new and advanced dataset—the UH dataset is also employed. Table VI displays the classification results of various methods on the UH dataset. As evident from the table, the proposed QuadNet outperforms all other methods in terms of OA, AA, Kappa, and for the majority of the classes, demonstrating its superiority. On the other hand, ViT, HybridSN, and RSSAN exhibit lower accuracies when compared to other CNN or transformer-based methods. Fig. 7 presents the classification maps generated by different methods and provides a visual representation of the classification performance of each method on the UH dataset. The maps clearly demonstrate that the proposed method outperforms other methods, as it generates more clear and distinct boundaries between different land cover categories. Fig. 11 displays the accuracy corresponding to different percentage (1%, 2%, 3%,

TABLE VI
CLASSIFICATION RESULTS OF DIFFERENT METHODS WITH 5% TRAINING SAMPLES ON THE UH DATASETS

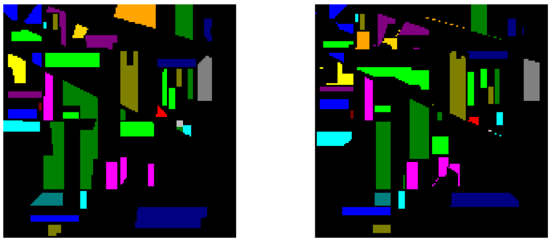
| Class | HybridSN | SSRN | A2S2K-R | RSSAN | DBMA | DBDA | SPANet | ViT | MFT | QuadNet |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 93.23±0.48 | 95.72±1.57 | 95.02±1.77 | 88.76±6.01 | 97.67±0.26 | 92.35±2.78 | 95.93±3.50 | 82.99±0.90 | 93.33±1.16 | **99.39±0.12** |
| 2 | 94.28±1.13 | 99.35±0.65 | 99.62±0.27 | 94.69±1.96 | 98.54±1.84 | 91.46±5.48 | 91.01±5.76 | 90.86±0.73 | 92.75±3.82 | **99.56±0.16** |
| 3 | 79.16±13.33 | **100±0** | 99.83±0.24 | 90.68±3.54 | 99.70±0.39 | 99.89±0.16 | 96.39±5.11 | 87.70±3.34 | 98.46±1.10 | **100±0** |
| 4 | 98.00±1.81 | 98.99±1.02 | 97.82±0.77 | 99.12±0.18 | 78.35±15.64 | 97.27±2.15 | 96.42±1.49 | 95.85±2.42 | **99.66±0.40** | 97.99±0.42 |
| 5 | 95.22±1.88 | 97.82±1.17 | 98.10±0.67 | 95.28±1.00 | 96.92±1.17 | 94.24±0.75 | 96.58±0.33 | 94.34±0.87 | **99.49±0.65** | 98.58±1.03 |
| 6 | 86.35±8.75 | 99.68±0.45 | **100±0** | 94.79±0.35 | 97.58±0.34 | 99.56±0.63 | 96.25±0.33 | 92.30±4.79 | 97.27±1.89 | 99.38±0.46 |
| 7 | 88.15±4.18 | 95.03±1.45 | 90.62±1.72 | 89.81±2.03 | 85.67±7.46 | 94.48±0.65 | 95.54±1.14 | 89.57±1.39 | 93.72±2.44 | **97.30±0.38** |
| 8 | 84.17±2.62 | 98.24±0.53 | 92.79±3.78 | 73.99±4.76 | 97.09±0.67 | 94.10±1.71 | 94.50±2.82 | 81.73±0.83 | 92.41±1.62 | **99.29±0.40** |
| 9 | 69.99±8.23 | 92.17±1.51 | 90.47±1.58 | 90.20±1.16 | 86.47±5.98 | 93.76±3.56 | 93.47±3.14 | 86.20±2.36 | 89.11±4.99 | **95.52±1.45** |
| 10 | 82.53±5.69 | 87.06±0.35 | 89.16±5.78 | 69.01±12.84 | 83.60±1.68 | 86.23±2.38 | 68.04±4.36 | 65.70±3.42 | 87.97±2.18 | **90.78±2.43** |
| 11 | 75.72±10.30 | 83.56±4.24 | 93.41±3.28 | 75.75±2.24 | 79.60±14.15 | 91.18±6.74 | 96.01±1.06 | 77.39±3.71 | 93.46±1.17 | **97.46±0.82** |
| 12 | 71.34±7.82 | 83.73±1.51 | 87.52±4.13 | 78.64±1.42 | 90.58±1.99 | **91.10±0.54** | 80.66±1.91 | 72.94±4.92 | 88.15±2.36 | 87.02±5.38 |
| 13 | 75.64±17.44 | **98.08±0.67** | 89.44±6.56 | 93.71±0.15 | 94.88±1.18 | 92.90±3.18 | 97.29±0.11 | 84.22±2.22 | 92.46±2.25 | 94.75±2.12 |
| 14 | 75.71±6.28 | **100±0** | 97.02±2.12 | 82.08±3.73 | 99.91±0.13 | 97.38±2.36 | 89.78±3.61 | 87.54±2.89 | 98.91±0.97 | 97.01±1.92 |
| 15 | 98.88±0.80 | **99.30±0.66** | 98.31±0.85 | 87.46±7.21 | 97.19±1.49 | 98.30±0.08 | 97.90±0.38 | 89.59±1.08 | 96.23±0.35 | 98.36±0.32 |
| OA | 84.24±4.77 | 93.93±0.16 | 93.87±0.57 | 85.58±2.72 | 89.07±3.42 | 93.28±0.80 | 90.65±1.57 | 84.02±0.57 | 93.49±0.31 | **96.46±0.85** |
| AA | 84.56±4.96 | 95.25±0.19 | 94.61±0.60 | 86.91±2.27 | 92.25±1.88 | 94.28±0.56 | 92.39±1.31 | 85.21±0.46 | 94.22±0.25 | **96.83±0.68** |
| Kappa | 82.95±5.15 | 93.44±0.18 | 93.37±0.61 | 84.41±2.94 | 88.17±3.71 | 92.74±0.86 | 89.89±1.70 | 82.72±0.62 | 92.93±0.34 | **96.17±0.92** |

The bold values denote the best accuracy.



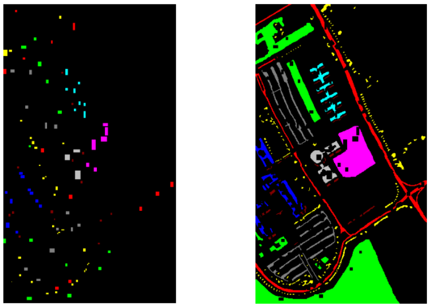
(a) (b) (c)

Fig. 11. OA, AA, and Kappa values over UH dataset under different training samples. (a) OA. (b) AA. (c) Kappa.



(a) (b)

Fig. 12. Spatial distribution of disjointed IN pine dataset. (a) Disjointed training set. (b) Disjointed test set.



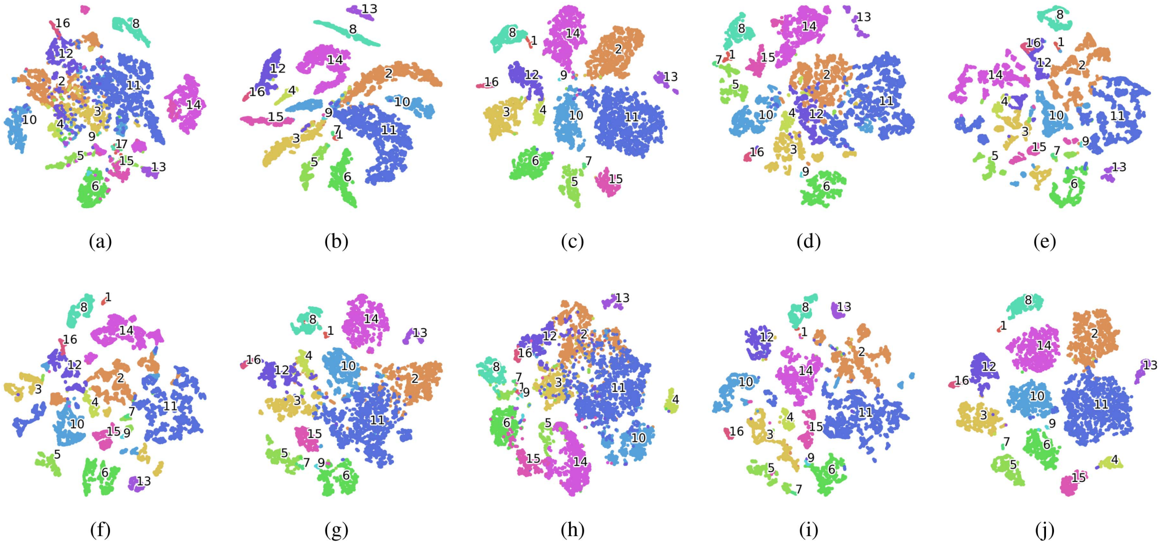
(a) (b)

Fig. 13. Spatial distribution of DUP dataset. (a) Disjointed training set. (b) Disjointed test set.

4%, and 5%) of training samples. As shown in the graph, the proposed method (see the light blue curve) achieves the highest accuracy even with a limited number of training samples, demonstrating its superior generalization capacity.

Figs. 14–17 present the distribution of the extracted features from four datasets for different methods using T-distributed stochastic neighbor embedding (t-SNE). For the proposed method, the same classes of samples are clustered together and there is a significant difference among different categories, further demonstrating the strong classification capacity of the proposed method.

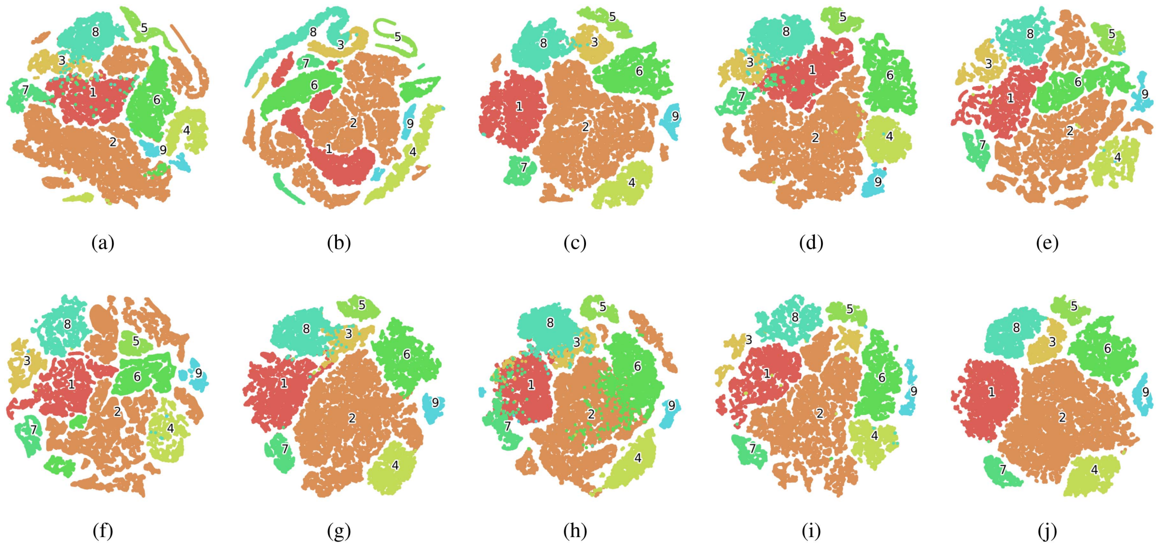
*2) Results on Disjointed Datasets:* The sampling method using random selection of training data for HSIs is prone to the problem that the training and test sets are similar. For classifying any pixel, a patch with it as center is used as input. In the random sampling method, the extracted patches for training and testing always overlap in some extent. For example, two adjacent pixels with one belonging to the training dataset and the other for testing, a large portion of overlap exists between their corresponding patches, thus the training and testing datasets are not completely separated. To avoid this issue, disjointed datasets that are sampled from nonoverlapping regions, as illustrated in the figure below are used. This ensures that the training and testing sets are entirely disjoint, thus avoiding any potential spatial overlap between them for better evaluating the robustness of the models. The training and test data distribution of disjointed



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 14. T-SNE visualization of extracted features for IN datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet.



(a) (b) (c) (d) (e)

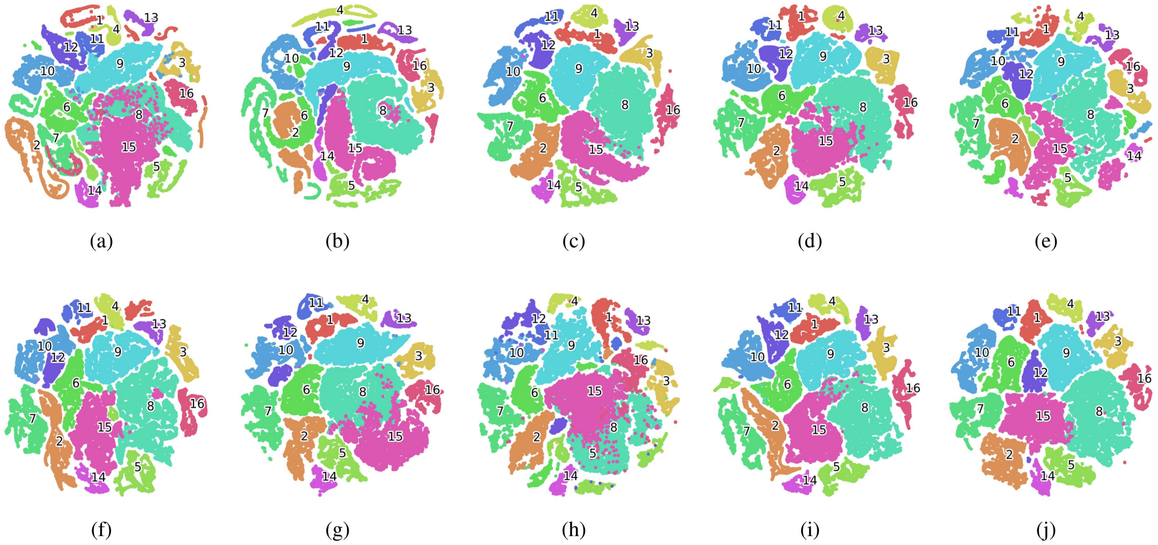
(f) (g) (h) (i) (j)

Fig. 15. T-SNE visualization of extracted features for UP datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet.

TABLE VII
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON DISJOINTED DATASETS

| Dataset | DIP | | | DUP | | |
|---|---|---|---|---|---|---|
| Methods | OA | AA | Kappa | OA | AA | Kappa |
| HybridSN | 61.53±3.36 | 61.46±4.83 | 55.69±3.83 | 78.79±1.62 | 76.56±0.94 | 72.78±1.63 |
| SSRN | 82.25±4.07 | 73.02±5.69 | 79.68±4.66 | 84.86±0.15 | 90.28±0.49 | 80.63±0.19 |
| A2S2K-R | 81.03±1.86 | 73.00±2.74 | 78.37±2.11 | 85.75±1.14 | 89.39±1.39 | 81.67±1.36 |
| RSSAN | 68.63±1.01 | 60.81±2.55 | 64.13±1.20 | 85.60±0.58 | 83.26±1.14 | 80.64±0.70 |
| DBMA | 84.60±0.52 | 72.03±0.60 | 82.53±0.58 | 85.28±1.51 | 89.39±0.24 | 81.11±1.80 |
| DBDA | 80.22±1.24 | 70.29±1.68 | 77.51±1.39 | 86.60±1.30 | 88.82±1.28 | 82.66±1.59 |
| SPANet | 80.56±2.23 | 72.98±1.37 | 77.85±2.53 | 87.94±0.98 | 86.48±1.71 | 83.78±1.32 |
| ViT | 66.29±1.57 | 60.03±0.19 | 61.44±1.83 | 81.71±0.60 | 78.71±0.91 | 75.29±0.70 |
| MFT | 82.91±2.38 | 70.83±2.57 | 80.56±2.79 | **91.39±4.49** | 90.98±4.15 | **88.78±5.64** |
| QuadNet | **84.94±1.13** | **76.77±1.22** | **82.84±1.30** | 88.64±0.42 | **91.40±0.51** | 85.13±0.56 |

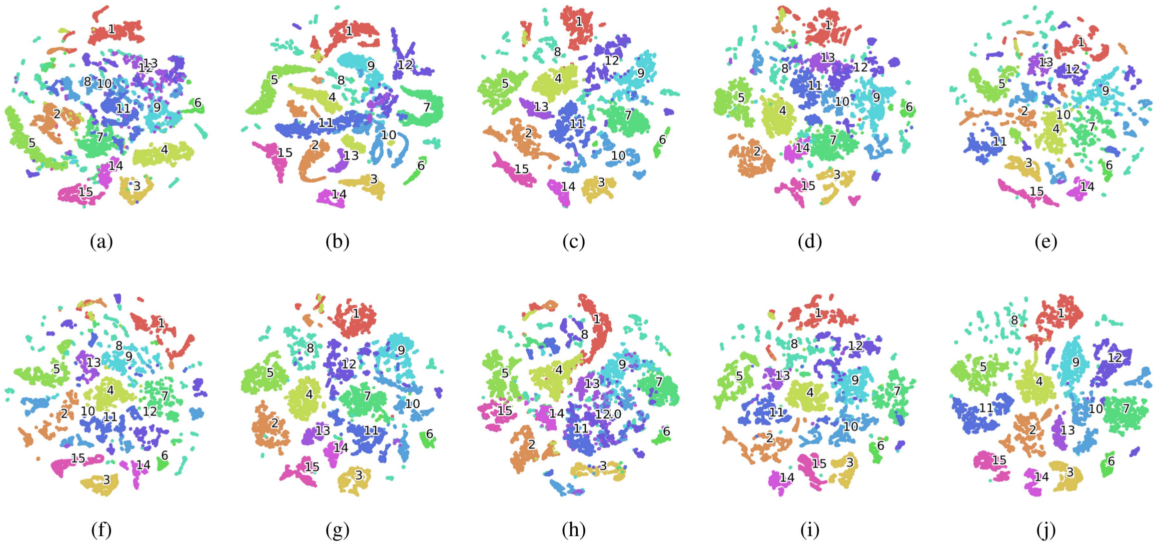
The bold values denote the best accuracy.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 16. T-SNE visualization of extracted features for SA datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 17. T-SNE visualization of extracted features for UH datasets. (a) HybridSN. (b) SSRN. (c) A2S2K-R. (d) RSSAN. (e) DBMA. (f) DBDA. (g) SPANet. (h) ViT. (i) MFT. (j) QuadNet.

Indian pine (DIP) and disjointed University of Pavia (DUP) datasets are shown in Figs. 12–13. Table VII shows the classification results of different models on DIP and DUP datasets. It can be seen that the classification results of all models are degraded by different degrees due to the spatial separation of the training and test sets. However, the proposed method in this paper still achieves the best classification results on both datasets. On the DIP dataset, QuadNet produces OA=81.70%, AA=79.29%, and Kappa=85.99%. For the DUP dataset, the OA, AA, and Kappa are 87.88%, 90.82%, and 84.31%, respectively.

TABLE VIII
ACCURACY ANALYSIS IN TERMS OF OA, AA, AND KAPPA FOR DIFFERENT MODULES OF THE PROPOSED FRAMEWORK

| Metrics | Datasets | TA-Residual | Quadlet-Residual | QuadNet |
|---|---|---|---|---|
| OA | IN | 97.83±0.25 | 97.66±0.30 | **98.22±0.28** |
| AA |  | 97.02±0.11 | 97.91±0.60 | **98.18±0.52** |
| Kappa |  | 97.52±0.28 | 97.33±0.34 | **97.97±0.32** |
| OA | UP | 99.80±0.03 | 99.84±0.02 | **99.88±0.05** |
| AA |  | 99.73±0.05 | 99.78±0.05 | **99.83±0.08** |
| Kappa |  | 99.73±0.04 | 99.79±0.03 | **99.85±0.06** |
| OA | SA | 98.66±1.45 | 98.64±0.17 | **99.10±0.01** |
| AA |  | 99.24±0.64 | 99.22±0.10 | **99.40±0.02** |
| Kappa |  | 98.52±1.62 | 98.48±0.19 | **99.00±0.01** |

The bold values denote the best accuracy.

### *D. Ablation Study*

Ablation studies are conducted to further validated the effectiveness of different modules in the proposed QuadNet model. Three different modules, i.e., triplet attention aided residual network (TA-Residual), quadlet attention aided residual network

TABLE IX
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON DISJOINTED DATASETS

| | HybridSN | SSRN | A2S2K-R | RSSAN | DBMA | DBDA | SPANet | QuadNet |
|---|---|---|---|---|---|---|---|---|
| Params | 3.498M | 355.633K | 361.949K | 164.877K | 591.896K | 592.682K | 415.723K | 358.363K |
| FLOPs $\times$ $10^6$ | 301.958 | 449.262 | 486.365 | 22.396 | 691.133 | 692.128 | 478.892 | 551.281 |

(Quadlet-Residual) and the proposed QuadNet model are compared. In the configuration of TA-Residual, quadlet attention module is removed from the proposed QuadNet model and the remaining model keeps the triplet attention aided spectral and spatial residual blocks unchanged. The second model is Quadlet-Residual, which removes the triple attention from the residual blocks and remains the quadlet attention module in proper position of the QuadNet model. The final scenario is the proposed QuadNet model discussed in Section III.

Table VIII depicts the OA, AA, and Kappa classification results of the ablation experiments over the IN, UP, and SA datasets. Overall, it can be seen from the table that QuadNet incorporating both attentions, i.e., quadlet and triplet can improve the classification results in terms of OA, AA, and Kappa for three datasets and achieves the best performance. Therefore, it demonstrates that the cross dimensional interaction among different dimensions, i.e., the number of feature maps, the spectral depth, spatial height and width helps emphasize discriminative power of features extraction by suppressing useless or redundant information.

### E. Computational Cost Analysis

Table IX illustrates the number of trainable weights and computational cost during the training process of the proposed QuadNet as well as other comparison networks. From the table, one can see that HybridSN has the largest number of parameters due to the 3-D convolution operation with large kernel sizes. The DBMA and DBDA methods have similar parameter numbers because of the use of multiscale kernel in the feature extraction process. The proposed method QuadNet has similar number of parameters as SSRN, and fewer than A2S2K-ResNet. RSSAN has the least number of model parameters. In terms of floating point operations (FLOPs), proposed QuadNet has nearly $550 \times 10^6$, less than DBMA, DBDA, and SPANet, but more than other models, such as SSRN and A2S2K-ResNet.

## V. CONCLUSION

In this article, a cross-attention module named quadlet is proposed for capturing the dependencies of HSIs across different dimensions during the forward propagation of the network. The designed quadlet attention can build the relationships among the number of feature maps, spectral bands, spatial height and width. Besides, triplet attention is incorporated to spectral-spatial residual blocks to enhance the learning of spectral-spatial features. Based on the quadlet cross-attention module and improved spectral-spatial residual blocks, a quadlet cross-attention aided residual network is further built for the HSI classification task. With the help of generalized triple attention, the developed network can extract more discriminative features and boost the classification performance. A series of experiments are conducted and the results show that the proposed Quadlet-Residual can achieve higher classification accuracy with limited samples due to the extracted cross dimensional dependencies and discriminative power of feature representation.

However, the constraint also needs to be noted even though the proposed strategy yields encouraging results in the experiments. The presented approach uses 3-D convolutional processes, which could be computationally expensive when compared with 2-D convolution. Moreover, the attention module involves an additional dimension—the number of feature maps, this also increases the computation complexity. In the future, the effect of different attention mechanisms, especially the self-attention mechanisms, on the classification performance of HSIs will be investigated.

## REFERENCES

[1] N. Audebert, B. Le Saux, and S. Lefevre, “Deep learning for classification of hyperspectral data: A comparative review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[2] 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, Jul. 2021.

[3] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, “Advanced spectral classifiers for hyperspectral images: A review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[4] M. Ahmad et al., “Hyperspectral image classification—traditional to deep models: A. survey for future prospects,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.

[5] B. Kumar, O. Dikshit, A. Gupta, and M. K. Singh, “Feature extraction for hyperspectral image classification: A review,” *Int. J. Remote Sens.*, vol. 41, no. 16, pp. 6248–6287, 2020.

[6] W. Sun and Q. Du, “Hyperspectral band selection: A review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[7] S. S. Sawant, P. Manoharan, and A. Loganathan, “Band selection strategies for hyperspectral image classification based on machine learning and artificial intelligent techniques–Survey,” *Arab. J. Geosci*, vol. 14, no. 7, pp. 1–10, 2021.

[8] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” *J. Sensors*, vol. 2015, pp. 1–12, 2015.

[9] H. Gao, Y. Yang, C. Li, H. Zhou, and X. Qu, “Joint alternate small convolution and feature reuse for hyperspectral image classification,” *ISPRS Int. J. Geo- Inf*, vol. 7, no. 9, 2018, Art. no. 349.

[10] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, “Hyperspectral image classification using random occlusion data augmentation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1751–1755, Nov. 2019.

[11] J. Yue, W. Zhao, S. Mao, and H. Liu, “Spectral–spatial classification of hyperspectral images using deep convolutional neural networks,” *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015.

[12] Y. Li, H. Zhang, and Q. Shen, “Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network,” *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 67.

[13] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, “Spatial sequential recurrent neural network for hyperspectral image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.

[14] E. Pan, X. Mei, Q. Wang, Y. Ma, and J. Ma, “Spectral-spatial classification for hyperspectral image based on a single GRU,” *Neurocomputing*, vol. 387, pp. 150–160, 2020.

[15] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[16] 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, Sep. 2018.

[17] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "Caps-tripleGAN: GAN-assisted capsNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7232–7245, Sep. 2019.

[18] J. Feng et al., "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1149.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[21] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[22] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2021, pp. 1–11.

[23] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust.*, *Speech Signal Process.*, 2021, pp. 2235–2239.

[24] 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, Aug. 2020.

[25] 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, Sep. 2021.

[26] U. Nandi, S. K. Roy, D. Hong, X. Wu, and J. Chanussot, "TAttMSRecNet: Triplet-attention and multiscale reconstruction network for band selection in hyperspectral images," *Expert Syst. Appl.*, vol. 212, 2023, Art. no. 118797.

[27] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3138–3147.

[28] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017.

[29] C. Ding, Y. Li, Y. Xia, W. Wei, L. Zhang, and Y. Zhang, "Convolutional neural networks based hyperspectral image classification method with adaptive kernels," *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 618.

[30] 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, Feb. 2020.

[31] 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, Feb. 2018.

[32] 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, Feb. 2019.

[33] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[34] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[35] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[36] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[37] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, "Classification of hyperspectral images via multitask generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1424–1436, Feb. 2021.

[38] 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, 2022.

[39] M. E. Paoletti, S. Moreno-Álvarez, and J. M. Haut, "Multiple attention-guided capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.

[40] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[41] K. Yang, H. Sun, C. Zou, and X. Lu, "Cross-attention spectral–spatial network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[42] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.

[43] 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.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[44] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[45] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

[46] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 582.

[47] X. Mei et al., "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 963.

[48] P. Wu, Z. Cui, Z. Gan, and F. Liu, "Residual group channel and space attention network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2035.

[49] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[50] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 498.

[51] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[52] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plazza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," 2022, *arXiv:2203.16952*.

[53] H. Liu, W. Li, X.-G. Xia, M. Zhang, C.-Z. Gao, and R. Tao, "Central attention network for hyperspectral imagery classification," *IEEE Trans. Neural Netw. Learn. Syst*., early access, Mar. 10, 2022, doi: 10.1109/TNNLS.2022.3155114.

[54] X. Zhao, J. Niu, C. Liu, Y. Ding, and D. Hong, "Hyperspectral image classification based on graph transformer network and graph attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 19, pp. 1–5, 2022.

[55] S. Fang, K. Li, and Z. Li, "Salient positions based attention network for image classification," 2021, *arXiv:2106.04996*.

[56] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

**Xin Qiao** (Student Member, IEEE) received the B.Eng. degree in navigation from Dalian Maritime University, Dalian, China, in 2018, and the M.Eng. degree in naval architecture and ocean engineering from the Zhejiang University, Zhejiang, China, in 2021. He is currently working toward the Ph.D. degree in electrical engineering with the Memorial University of Newfoundland, St. John's, NL, Canada.

His research focuses on hyperspectral image classification.

**Swalpa Kumar Roy** (Student Member, IEEE) received the bachelor's and the master's degrees in computer science and engineering from the West Bengal University of Technology, Kolkata, India, in 2012, and Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, in 2015, respectively, and the Ph.D. degree in computer science and engineering from the University of Calcutta, Kolkata, India, in 2021.

From July 2015 to March 2016, he was a Project Linked Person with the Optical Character Recognition Laboratory, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. He is an Assistant Professor with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, India. His research interests include computer vision, deep learning, and remote sensing.

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 is an Associate Editor for the journal of *Springer Nature Computer Science* (SNCS) and also an Editor for the frontiers journal of *Advanced Machine Learning Techniques for Remote Sensing Intelligent Interpretation*. He was a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

**Weimin Huang** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in radio physics from Wuhan University, Wuhan, China, in 1995, 1997, and 2001, respectively, and the M.Eng. degree in electrical engineering from the Memorial University of Newfoundland, St. John's, NL, Canada, in 2004.

From 2008 to 2010, he was a Design Engineer with Rutter Technologies, St. John's, NL, Canada. Since 2010, he has been with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, where he is currently a Professor. He has authored more than 260 research articles. His research interests include the mapping of oceanic surface parameters via high-frequency ground wave radar, X-band marine radar, synthetic aperture radar, and global navigation satellite systems.

Dr. Huang was a Technical Program Committee Member. He was the Technical Program Co-Chair for the IEEE Newfoundland Electrical and Computer Engineering Conference in 2012 and 2013. He is an Editor for the book Ocean Remote Sensing Technologies: High Frequency, Marine, and GNSS-Based Radar. He is also an Area Editor for IEEE CANADIAN JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING, an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE JOURNAL OF OCEANIC ENGINEERING, *Remote Sensing, Frontiers in Marine Science*, and he has been a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and five other journals. He is a Reviewer for more than 100 international journals and a Reviewer for many IEEE international conferences, such as RadarCon, International Conference on Communications, IEEE Global Communications Conference, IEEE International Geoscience and Remote Sensing Symposium, and Oceans. From 2018–2021, he was a Member and Co-Chair of the Electrical and Computer Engineering Evaluation Group for Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants. He was a recipient of Postdoctoral Fellowship from the Memorial University of Newfoundland; the Discovery Accelerator Supplements Award from NSERC in 2017; and the IEEE Geoscience and Remote Sensing Society 2019 Letters Prize Paper Award as well as some other teaching and research awards.