QTN: Quaternion Transformer Network for Hyperspectral Image Classification

Xiaofei Yang, Weijia Cao, Yao Lu, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— Numerous state-of-the-art transformer-based techniques with self-attention mechanisms have recently been demonstrated to be quite effective in the classification of hyperspectral images (HSIs). However, traditional transformer-based methods severely suffer from the following problems when processing HSIs with three dimensions: (1) processing the HSIs using 1D sequences misses the 3D structure information; (2) too expensive numerous parameters for hyperspectral image classification tasks; (3) only capturing spatial information while lacking the spectral information. To solve these problems, we propose a novel Quaternion Transformer Network (QTN) for recovering self-adaptive and long-range correlations in HSIs. Specially, we first develop a band adaptive selection module (BASM) for producing Quaternion data from HSIs. And then, we propose a new and novel quaternion self-attention (QSA) mechanism to capture the local and global representations. Finally, we propose a new and novel transformer method, *i.e.*, QTN by stacking a series of QSA for hyperspectral classification. The proposed QTN could exploit computation using Quaternion algebra in hypercomplex spaces. Extensive experiments on three public datasets demonstrate that the QTN outperforms the state-of-the-art vision transformers and convolution neural networks.**

***Index Terms*— Hyperspectral image classification, convolution neural network, transformer network, quaternion transformer network (QTN).**

Manuscript received 14 February 2023; revised 6 May 2023; accepted 27 May 2023. Date of publication 6 June 2023; date of current version 7 December 2023. This work was supported in part by the Science and Technology Development Fund, Macau, under File 0049/2022/A1; in part by the University of Macau under File MYRG2022-00072-FST; in part by the Macao Young Scholars Program under Grant AM2020012; in part by the National Natural Science Foundation of China (NSFC) Fund under Grant 62206073 and Grant 62002122; in part by the Guangdong Shenzhen Joint Youth Fund under Grant 2021A151511074; in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515010893; in part by the Shenzhen Key Technical Project under Grant 2022N063; and in part by the Shenzhen Doctoral Initiation Technology Plan under Grant RCBS20221008093222010. This article was recommended by Associate Editor D. Gragnaniello. *(Corresponding authors: Weijia Cao; Yicong Zhou.)*

Xiaofei Yang is with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou 510182, China, and also with the Department of Computer and Information Science, University of Macau, Zhuhai, Macau, China (e-mail: xiaofeiyang@gzhu.edu.cn).

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100045, China, also with the Department of Computer and Information Science, University of Macau, Zhuhai, Macau, China, and also with Yangtze Three Gorges Technology and Economy Development Company Ltd., Beijing 100038, China (e-mail: caowj@aircas.ac.cn).

Yao Lu is with the Department of Computer Science and Technology, and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: luyao2021@hit.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Zhuhai, Macau, China (e-mail: yicongzhou@um.edu.mo).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3283289.

Digital Object Identifier 10.1109/TCSVT.2023.3283289

I. Introduction

REMOTE sensing improvements have resulted in the availability of hyperspectral images (HSIs) that contains hundreds of narrow and contiguous wavelength bands. HSIs can provide detailed spectral properties and rich spatial information for recognizing objects [1], [2]. Since HSIs could offer both spectral and spatial information, they are frequently employed in a variety of fields, including land cover classification [3], marine monitoring [4], and urban planning [5].

Although the massive spectral and spatial information of HSIs makes it possible for accurate identification of the ground materials, they bring some challenges to hyperspectral image classification tasks. On the one hand, the high dimensionality of HSIs and limited labeled samples may lead to the Hughes phenomenon [6]. On the other hand, due to some internal and external factors (such as sensor parameters, and atmospheric conditions, etc.), there are some questions in hyperspectral image classification, including the same spectrum of foreign matter and the same objects but different spectrum in surface cover. All these problems will unavoidably degrade classification performance. In order to solve these problems, feature extraction is introduced to extract some appropriate features for improving the classification performance.

Traditional hyperspectral image classification approaches, such as Support Vector Machines (SVM) [7], K-nearest Neighbor (KNN) classifiers [8], and low rank-based models [9], [10] are proposed for hyperspectral image classification only using the spectral information. However, they have a failure to accurately distinguish different materials by only using spectral information. Recent studies have demonstrated that considering spatial contextual information could improve classification performance. Thus, it is important to explore how to extract effective spatial contextual information for hyperspectral image classification tasks. For example, Fauvel et al. [11] employed an extended morphological profile (EMP) for extracting spatial contextual features and then fused them with spectral information to improve the classification accuracy. Wang et al. [12] proposed multinomial logistic regression and designed the Locality Adaptive Discriminant Analysis methods for hyperspectral image classification. However, these above approaches perform unsatisfactorily while processing complex data. The possible reason is that these traditional approaches could not fit and represent numerous complex data.

Due to the strong ability for extracting abundant local spatial features, Convolutional Neural Network (CNN) has

achieved great success in natural image recognition. Recent studies have demonstrated that applying CNN to hyperspectral classification tasks can greatly outperform satisfactory performance [13], [14], [15], [16], [17]. For example, Rasti et al. [18] summarized the datasets and the toolbox of hyperspectral feature extraction, and provided a detailed and organized overview of hyperspectral image classification. Lee and Kwon [19] proposed a deep and wide CNN architecture to extract spatial-spectral information for hyperspectral classification. Cao et al. [20] proposed a new model by integrating the CNN and Markov Random Fields (MRF), in which the CNN was used to extract the spatial-spectral information and MRF was utilized to further exploit spatial information. Li et al. [21] proposed to use a CNN architecture to learn deep Pairwise Potential Functions (PPFs) from the labeled and unlabeled data for hyperspectral image classification. These CNNs have achieved significant classification results, which demonstrates that CNN can extract desirable local spatial contextual and spectral features for hyperspectral image classification. However, these 2D CNN-based methods limit the further improvement of the performance, which only extracts features by separating the spatial and spectral dimensions. To solve this issue, some 3D CNN-based methods have been proposed to capture abundant features by using spatial and spectral information together. For example, Ben Hamida et al. [22] proposed a spectral-spatial 3D CNN for hyperspectral image classification. In order to retrieve rich spectral-spatial features, it employed a 3D convolution layer to replace the pooling layer for reducing the dimension. In references [22], [23], [24], [25], [26], [27], [28], [29], [30], CNNs were utilized to extract spatial contextual information. By introducing spatial contextual information, these CNN-based approaches achieved outstanding classification performance compared to the traditional approaches that only used spectral information. Unfortunately, many of these CNN-based approaches are heavily parameterized, which may cause the Hughes phenomenon. Moreover, CNN-based approaches are adept in extracting the local spatial contextual information, yet not capturing subtle spectral discrepancies among the spectral bands.

Recently, the self-attention models (*i.e.*, Transformer) have quickly become the dominant architecture in natural language processing. For example, Dosovitskiy et al. [31] first introduced the transformer architecture into computer vision and proposed the vision transformer (ViT). The proposed ViT performed comparable results to the state-of-the-art CNNs on image classification tasks. Benefiting from its powerful modeling capabilities in the sequences, transformer-based approaches are quickly applied to various tasks, including image classification [32], [33], [34], semantic segmentation [35], etc. Although these transformer-based approaches have achieved remarkable success, self-attention mechanisms still have issues. The self-attention mechanism is originally designed for 1D NLP tasks, it always treats 2D images as 1D sequences, which neglects the 2D structure information of the images. In order to integrate the local spatial information, the convolution techniques are embedded in transformers [36], [37], [38], [39], [40], [41]. For example, Wu et al. [42] integrated the convolution layers into transformer, and proposed a new method called convolution vision transformer (CvT) for image classification. However, all these transformers are designed for natural RGB images. There are few studies on applying transformers to hyperspectral image classification tasks. For example, He et al. [43] directly employed the transformer network for hyperspectral image classification and proposed the HSI-BERT by utilizing a bidirectional encoder representation network (BERT). He et al. [44] proposed a spatial-spectral transformer network for hyperspectral image classification. However, it is also difficult to directly apply the transformers in processing high-resolution images because of their numerous parameters, and memory overhead. On the other hand, the self-attention mechanism is specially designed to consider the adaptability in the spatial dimension, while ignoring the adaptability in the spectral dimension of HSIs.

To solve these problems, this paper aims to investigate a novel way of greatly enhancing the performance of existing transformers-based methods while simultaneously reducing the parameter cost. In order to achieve this goal, we transcend physical space and investigate the self-attention mechanism in the Quaternion domain (*i.e.*, Hypercomplex numbers). In the Quaternion domain, the hypercomplex numbers contain a real part and three imaginary components (*e.g.*, i, j, k). As such, quaternion algebra could effectively process 3D data [45], [46]. This is mainly because quaternions can capture the mutual information among these components using the Hamilton product, and remain the physical meaning of the original data [47], [48]. Due to the structural characteristics and advantages of the quaternions, quaternion has been applied to many fields [49], [50]. Nevertheless, quaternions have not been widely used in hyperspectral image classification tasks because of the incorporation of multiple spectral bands in HSIs. Moreover, most of the current methods [51], [52] use Principal component analysis (PCA) to generate the quaternion data, which ignores the relationship between neighbor spectral bands. There also have some feature extraction methods to reduce the bands of the hyperspectral image, for example, non-negative matrix factorization (NMF) [53] and Linear Discriminant Analysis (LDA) [54]. Wang et al. [55] proposed a band selection method, namely optimal neighborhood reconstruction (ONR) to exploit the strong correlation between neighboring bands.

However, these band selection methods are implemented on the feature extraction, and could not be suitable for deep learning networks, especially quaternion-based deep learning networks. To this end, this paper first presents a band adaptive selection module (BASM) to select the bands for generating the quaternion data. The performance will be maintained/improved while reducing the parameters by using BASM to generate the quaternion data. Moreover, the existing multi-head-self-attention (MHSA) module could not capture the global representation of the spectral dimension and has lots of parameters. For solving this question, we propose a new and straightforward quaternion self-attention (QSA) module. It can engross the advantages of convolution operation and self-attention mechanism, including local 2D spatial information, long-rang spectral dependence, and adaptability in

spatial-spectral dimension, while using a few parameters. Finally, we propose a new transformer network called the quaternion transformer network (QTN) to model spatial-spectral information for hyperspectral image classification by stacking a BASM layer and a series of QSA. The proposed QTN could take advantage of quaternions and the self-attention mechanism to identify the HSIs. Different from the previous hyperspectral image classification methods, the proposed QTN is an efficient and small transformer-based method. Moreover, it is an extension of the transformer on the quaternion domain.

Therefore, the significant contributions of this paper can be described as follows:

- we rethink hyperspectral image classification on the quaternion domain, and extend the transformer-based hyperspectral image classification methods to the quaternion domain;
- we propose a novel band adaptive selection module (BASM) to generate the quaternion data from HSIs, which can adaptively select bands for each class leveraging both spatial and spectral information;
- we also propose a new and novel quaternion self-attention (QSA) module for hyperspectral image classification that is capable of convolution as well as self-attention with a small number of parameters;
- we finally propose a new and straightforward transformer-based hyperspectral image classification network based on QSA and BASM, namely Quaternion Transformer Network (QTN), which could analyze the HSIs on the quaternion field and extract the local context features and global representation;
- we demonstrate that QTN outperforms state-of-the-art CNNs-and transformers-based approaches using extensive experiments on three hyperspectral datasets, including the Indian Pines scene, the University of Pavia, and Houston2013 scene datasets.

The rest of this paper is organized as follows. In Section II, we will introduce the related work about deep learning-based hyperspectral image classification, transformer-based classification, and quaternion-based hyperspectral image classification. In Section III, we will give a brief introduction to the quaternion. In Section IV, we will present the proposed method and its components. In Section V, we first give an illustration of three benchmark HSIs datasets and experimental settings, and then present the results and the analyses of the experiment. Finally, We conduct a conclusion and the future work in Section VI.

II. Related Work

A. Convolution Neural Networks for Hyperspectral Image Classification

The goal of hyperspectral image classification is to recognize the center pixel, thus it is also called the pixel classification task. Due to their strong ability for extracting local feature extraction, Convolution Neural Networks (CNNs) are frequently used to extract local spatial contextual information for hyperspectral image classification. CNNs and their variants often achieve outstanding classification accuracy [56], [57], [58], [59]. For example, Sharma et al. [27] established a two-dimensional (2D) CNN model to classify the HSIs. It is a simple 2D CNN through stacking 2D convolution layers to capture the local spatial context information. Within these 2D CNNs, 2D convolution operation separately handles the spatial and spectral information. However, it may decrease performance. To address this issues, Yang et al. [23] and Chen et al. [26] introduced the three-dimensional (3D) convolution operation to process the spatial and spectral information together, and presented 3D CNN approaches for hyperspectral image classification. Roy et al. [60] proposed a HybridSN for hyperspectral image classification by stacking several 3D and 2D convolution layers to extract the spatial and spectral features. There also have some studies of attention mechanism application in hyperspectral image classification. For example, Lorenzo et al. [28] first employed an attention mechanism to select the spectral bands, and then fed them into the CNN model to recognize the HSIs. There are some works for introducing CNN into other networks to improve the performance in hyperspectral image classification [61], [62]. For example, Hong et al. [61] integrated the CNN and Graph Convolution Network (GCN) to present a new minibatch GCN to overcome the limitations of traditional GCNs in large-scale remote sensing problems.

Although these CNNs and their variants have achieved outstanding performance, they also have some issues. (1) CNNs are excessively concerned with the local spatial information, while ignoring the relationship between features; (2) CNNs have a failure in capturing long-range dependence. These issues will limit them from achieving higher performance in hyperspectral image classification tasks.

B. Transformers Network for Hyperspectral Image Classification

Now, transformers are the dominant solutions in natural language processing (NLP). Recently, numerous studies on resolving image classification tasks using transformers have been presented. Dosovitskiy et al. [31] employed the transformer for the image classification task and proposed the vision image transformer (ViT) network. ViT is the first study that applies the transformer network to computer vision to rethink computer vision from a sequence perspective, and achieves comparable results to the state-of-the-art CNNs. Unfortunately, the above transformers often treat the input image as one-dimensional data, which neglects the 2D structure of the input image. This limits transformers further performing outstandingly. As such, the introduction of local spatial contextual information in transformers would certainly have a promising performance [36], [37], [38], [39], [40], [41], [42]. For example, Graham et al. [37] and Heo et al. [41] employed the convolution layers to extract local spatial information for improving the performance of transformers. Liu et al. [32] introduced the shift window manner to generate the local attention features and presented a new transformer network called SwinT. Zhou et al. [33] recalculated the attention value by using a learnable matrix and presented a very

deep transformer network called DeepViT. Many of these transformers can learn useful feature representations from the inputs, and provide a helpful inductive bias for achieving satisfactory results.

Notably, progresses on transformer-based hyperspectral image classification are still in its infancy, and accordingly, most studies on this topic are very recent. For example, He et al. [43] firstly introduced the transformer (i.e., BERT [63]) in hyperspectral image classification and presented a transformer network, namely for HSI-BERT hyperspectral image classification. Hong et al. [64] proposed a new neural network called SpectralFormer for hyperspectral image classification. Within SpectralFormer, the transformers are used to learn spectrally local sequence information from neighboring bands of hyperspectral images. However, they did not consider the local spatial information. Yang et al. [65] proposed a special transformer-based method, namely Hyperspectral image Transformer (HiT) for Hyperspectral image classification to extract the global and local features. The HiT consists of two modules: one dynamic 3D convolution projection module and four attention-based stages. The dynamic 3D convolution projection module is used for extracting the local spatial-spectral information by using two 3D convolution layers. Each attention-based stage consists of two Layernorm layers, one Conv-Permutator module, and one channel-MLP layer. The Conv-Permutator module is comprised of three depth-wise convolution layers, and separately encodes the spatial-spectral representations along the height, width, and spectral dimensions, respectively. However, these transformer-based methods are implemented by using the multi-head-self-attention module (MHSA), which could not completely capture the global spatial-spectral representation and occupy lots of computing resources.

C. Quaternion for Hyperspectral Image Classification

Quaternion representations for deep learning have drawn attention, there are many studies on this topic recently [66]. For example, Gaudet and Maida [67] presented a deep Quaternion network for image classification. Zhu et al. [68] proposed Quaternion CNNs and applied them to image classification tasks. All these works can be demonstrated that Quaternion representations are helpful and provide effectiveness over real-valued representations. Some researchers introduce quaternion to hyperspectral image classification [51], [52]. For example, Li et al. [51] firstly used PCA to construct the 3-channel PCA features and then analyzed the features on the quaternion field. However, these methods only analyze features in the quaternion field, rather than reserving the raw channel information and using deep learning.

So, the goal is to design a new and novel deep learning network for hyperspectral image classification, which could take advantage of quaternion, CNN and transformer. We apply the quaternion convolution (QConv) filters to the self-attention mechanism and present a quaternion self-attention (QSA), which could extract the local spatial information and encourage interactions between the real and imaginary components. Moreover, the QSA can reduce the parameter size because of parameter saving in the Hamilton product. As such, the proposed QTN built with multi-QSA would achieve a satisfactory performance for hyperspectral image classification, while using much fewer parameters.

III. Background on Quaternion Algebra

This section introduces the background for this paper. Here, a brief introduction to quaternion algebra is given. Actually, quaternions are an extension of the complex field $\mathbb{C}$. A quaternion Q, comprising one real and three imaginary components, which can be written in the form

$$Q = r + q_1 i + q_2 j + q_3 k, \;\; r, q_1, q_2, q_3 \in \mathbb{R}, \tag{1}$$

where $\mathbb{R}$ denotes the real field, and i, j, and k are orthogonal imaginary units, which attend the following rules:

$$\begin{aligned} ijk = i^2 = j^2 = k^2 = -1, \\ ij = -ji = k, \\ jk = -kj = i, \\ ki = -ik = j. \end{aligned} \tag{2}$$

In Eq.1, Q is called the pure quaternion if $r = 0$. Operations on quaternions are defined in the following.

Addition The addition of two quaternions is defined by:

$$\begin{aligned} Q + P = \quad & Q_r + P_r + (Q_{q1} + P_{p1})i \\ & + (Q_{q2} + P_{p2})j + (Q_{q3} + P_{p3})k, \end{aligned} \tag{3}$$

where Q and P with subscripts are the real value and imaginary components of quaternion Q and P.

Scalar Multiplication scalar α multiplies across all components, *i.e.,*

$$\alpha Q = \alpha r + \alpha q_1 i + \alpha q_2 j + \alpha q_3 k. \tag{4}$$

Conjugate The conjugate of Q is defined by

$$Q^* = r - (q_1 i + q_2 j + q_3 k). \tag{5}$$

Modulus The Modulus of Q is defined as:

$$|Q| = (r^2 + q_1^2 + q_2^2 + q_3^2)^{1/2}. \tag{6}$$

Inverse The Inverse of Q can be formulated as:

$$Q^{-1} = Q^*/|Q|^2. \tag{7}$$

Norm The normalized of unit quaternion $Q^\triangleleft$ is expressed as:

$$Q^\triangleleft = \frac{Q}{\sqrt{r^2 + q_1^2 + q_2^2 + q_3^2}}. \tag{8}$$

Hamilton product Following [69], the Hamilton product of two quaternions Q and P is computed as:

$$\begin{aligned} Q \bigotimes P = & (Q_r P_r - q_1 p_1 - q_2 p_2 - q_3 p_3) \\ & + (Q_r p_1 + q_1 P_r + q_2 p_3 - q_3 p_2) i \\ & + (Q_r p_2 - q_1 p_3 + q_2 P_r + q_3 p_1) j \\ & + (Q_r p_3 + q_1 p_2 - q_2 p_1 + q_3 P_r) k. \end{aligned} \tag{9}$$

In this work, we use Hamilton products extensively for matrix transformations in quaternion convolution layers.

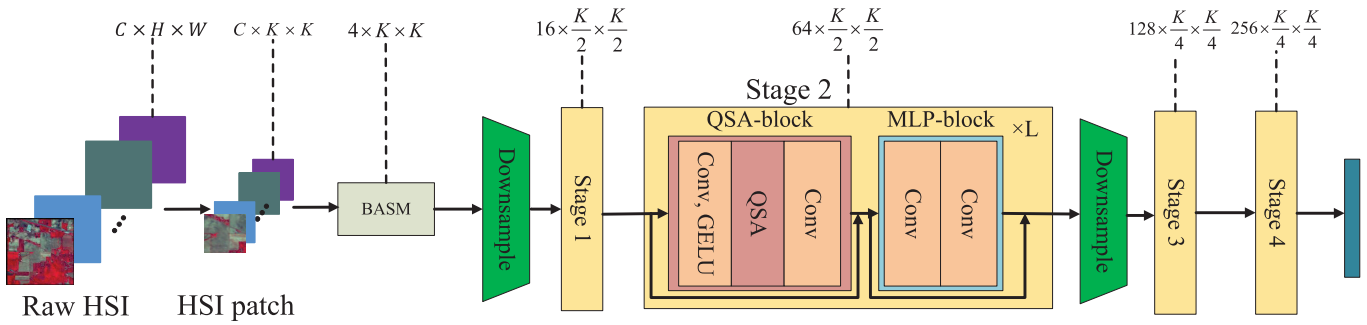


Fig. 1. Overall architecture of the proposed QTN. The QTN consists of two parts: BASM and quaternion transformer block. The BASM is a band adaptive selection module for generating the quaternion data from HSI. The quaternion transformer block is built on four stages, each of which comprises of QSA-block and MLP-block. It is noted that "Conv" denotes the Layer Norm and convolution operations.

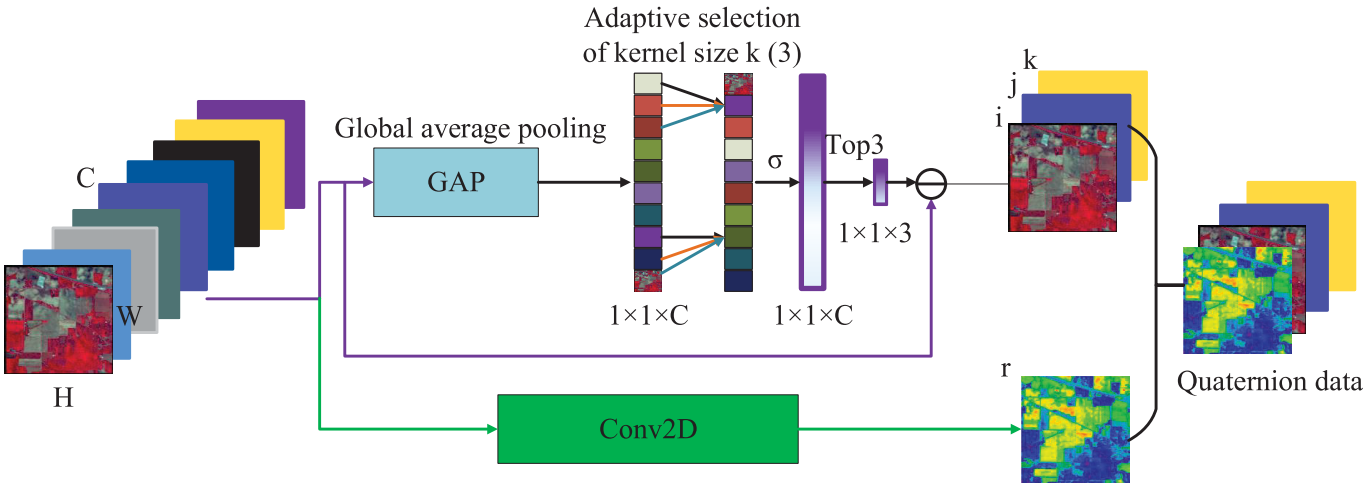


Fig. 2. Diagram of band adaptive select module (BASM). There are two parts for generating the quaternion data, including imaginary components and real part processes. Given the aggregated features gained by global average pooling (GAP), it generates the real part by using a 2D convolution (Conv2D for short in the figure) and the imaginary components by using the channel attention module. $\ominus$ is the pick-up function.

IV. Proposed Approach

In this section, we introduce the proposed QTN. First, we present the overview of the proposed QTN in Section IV-A. Secondly, we introduce the details of the BASM module in Section IV-B. Thirdly, we describe the proposed QSA in detail in Section IV-C. And finally, we introduce the two QTN with different parameters in Section IV-D.

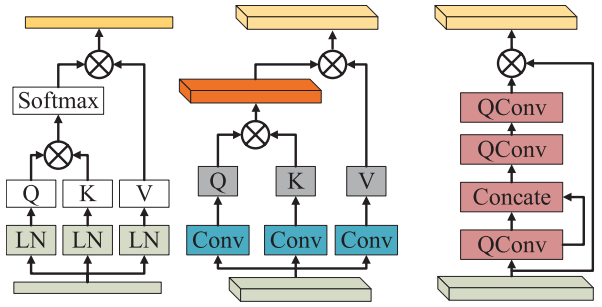
A. Overview

A hyperspectral image is a 3D cube data, which consists of spatial and spectral information. The hyperspectral image can be defined by a three-dimensional tensor $H \in R^{W \times H \times C}$, where W and H are spatial dimensions, and C denotes the spectral dimension. Thus, there are a total of $N = W \times H$ pixels in the HSIs. The hyperspectral classification is also pixel classification. Since both spatial and spectral information of H can help to identify the pixel $x \in \mathbb{R}^C$, they should be considered [14], [23], [27]. The categories of adjacent pixels are likely to be the same, thus, the adjacent pixels of the pixel x could be used to supplement spatial information for identifying the pixel x. Therefore, a square region with a size $K \times K$ is cropped from the raw HSIs H to construct the patch X. The cropped $X \in \mathbb{R}^{K \times K \times C}$ is fed into the proposed method. Finally, the proposed QTN is exploited to extract features and identify the category of the pixel x based on the constructed patch X. It is worth noting that there are N patches.

Figure 1 shows the detailed proposed QTN architecture for hyperspectral image classification. We can see that the proposed QTN is composed of a BASM, two downsample layers, and four transformer stages. It is worth noting that each transformer stage contains a Quaternion self-attention (QSA) block and a Multilayer Perceptron (MLP) block. We first crop the raw HSIs and create the HSIs patch X. Then, we perform a BASM to select the bands and generate the quaternion data for the HSIs patch X. The generated quaternion data then will be fed into a series of transformer stages to extract the local and global features. The downsample layer is used to reduce the spatial size. Finally, we re-weight the spatial-spectral fusion features, and then feed them into a classifier to perform hyperspectral image classification.

B. Band Adaptive Selection Module

HSIs are characterized by the inclusion of numerous spectral bands, which increases the computational complexity and brings challenges for quaternion application in hyperspectral image classification. To address this issue, many methods are used to select the bands, such as Principal component analysis (PCA) [70], Linear Discriminant Analysis (LDA) [54] and Non-negative Matrix Factorization (NMF) [53]. However, the selected band order is fixed for each category, which is not suitable. To this end, we present a band adaptive



(a) Self-attention (b) Non-local attention (c) Q-Self-attention

Fig. 3. Structure of different modules: (a) the self-attention module; (b) the non-local attention module; (c) the proposed quaternion self-attention (QSA) module. It is noted that the difference between (b) and (c) is the element-wise multiply and quaternion. QConv denotes the quaternion convolution operation. The LN and Conv denote the Layer Norm and convolution operations.

selection module (BASM) to select the appropriate band for each category. The proposed BASM consists of two modules: (1) the Top-3 efficient channel attention (ECA) module; and (2) the 2D convolution module. The Top-3 ECA is used to select three bands, which will be regarded as the imaginary components (*i.e.*, i, j, k). And the 2D convolution module is utilized to render the real part (*i.e.*, r).

As shown in Figure 2, there are two parts to generate the quaternion data, such as imaginary components and real part processes. Then, the band selection process can be briefly expressed as follows.

For generating the real part, a 2D convolution module (Conv2D for short in Figure 2) is utilized to calculate the attention weight of each spectral band and then sum them together. The real part F_r can be formulated as:

$$F_r = F_{Conv}(X). \tag{10}$$

where F_{Conv} denotes a 2D convolution operation with outputting 1 feature map.

For the imaginary components process, ECA firstly generates the attention weight of spectral bands from the neighbor spectral and spatial information. We then select the Top-3 weights to generate 3 bands for representing the input HSIs. It can be calculated as follows:

$$\begin{aligned} Attention &= f_{top}(\sigma f(g(X))), \\ F &= Attention \ominus X. \end{aligned} \tag{11}$$

where $g(X) = \frac{1}{KK}\sum_{j=1,j=1}^{K,K} X_{ij}$ is channel-wise global average pooling (GAP). σ is a Sigmoid function. f denotes a 1D convolution operation with the size of 3. $\ominus$ is the pick-up function, and f_{top} denotes the Top-3 function.

Finally, we concatenate the real part and imaginary components together to generate the quaternion data for representing the HSIs.

C. Quaternion Self-Attention Module

Actually, the self-attention mechanism is an adaptive selection technique, which can extract the discriminative features and ignore the noise from the input features. The critical step of the self-attention mechanism is creating an attention map to distinguish the importance of different regions. The self-attention mechanism is adopted for capturing the long-range dependence, however, there are three obvious issues for application in computer vision that have been shown in Section I.

To address the above-listed issues, we present a quaternion self-attention module to capture the local spatial information and the long-range relationship. As shown in Figure 3 (c), the attention map is obtained by using three quaternion convolution layers. Assume X_Q is the input feature, the QSA module can be written as

$$\begin{aligned} y &= QConv(X_Q), \\ Attention &= QConv_{1\times1}(QConv(concate(y, y))), \\ output &= Attention \bigodot X_Q. \end{aligned} \tag{12}$$

Here, $Attention$ denotes the attention map. $\bigodot$ is the element-wise product. With this special design, the proposed QSA combines the advantages of convolution and self-attention, while reducing the parameters. It takes the local spatial contextual information, long-range spectral dependence, and latent interaction between features into consideration.

1) Quaternion Convolution: We first rethink the convolution operation before introducing the quaternion convolution operation. For classic convolution operation, its process is defined in the real-valued domain, which can be formulated as follows:

$$WX = \sum_{i=0}^{H}\sum_{j=0}^{W} w_{ij} X_{i,j}, \tag{13}$$

where H and W indicate the height and the width, respectively. w_{ij} is the value at position (i, j) of the convolution kernel, and $X_{i,j}$ denotes the value at position (i, j) of the input HSIs, and $X \in R^{4\times H\times W}$.

For quaternion convolution operation, it is a convolution operation using quaternion algebra. It is performed as a multiplication between a quaternion filter matrix and a quaternion vector, in which the Hamilton product is used for computation. Assume $W_Q = r + q_1 i + q_2 j + q_3 k$ denotes a quaternion weight filter matrix, and the input X could be represented by $X_Q = r_x + x_1 i + x_2 j + x_3 k$(a quaternion input vector). Then, the quaternion convolution operation *i.e.*, the Hamilton product $W_Q \bigotimes X_Q$ can be formulated as follows:

$$\begin{aligned} W_Q \bigotimes X_Q =& (rr_x - q_1x_1 - q_2x_2 - q_3x_3) \\ &+ (rr_x + q_1x_1 + q_2x_3 - q_3x_2)i \\ &+ (rx_2 - q_1x_3 + q_2r_x + q_3x_1)j \\ &+ (rx_3 + q_1x_2 - q_2x_1 + q_3r_x)k. \end{aligned} \tag{14}$$

We can also express it in a matrix form as follows:

$$W_Q \bigotimes X_Q = \begin{bmatrix} r & -q_1 & -q_2 & -q_3 \\ q_1 & r & -q_3 & q_2 \\ q_2 & q_3 & r & q_1 \\ q_3 & -q_2 & q_1 & r \end{bmatrix} * \begin{bmatrix} x_r \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{15}$$

TABLE I
DETAILED SETTING OF DIFFERENT VERSIONS OF QTN

stage	output size	QTN-Tiny	QTN-Samll
1	$\frac{K}{2} \times \frac{K}{2}$	$\begin{bmatrix} C=16 \\ L=3 \end{bmatrix}$	$\begin{bmatrix} C=16 \\ L=2 \end{bmatrix}$
2	$\frac{K}{2} \times \frac{K}{2}$	$\begin{bmatrix} C=64 \\ L=3 \end{bmatrix}$	$\begin{bmatrix} C=64 \\ L=2 \end{bmatrix}$
3	$\frac{K}{4} \times \frac{K}{4}$	$\begin{bmatrix} C=128 \\ L=5 \end{bmatrix}$	$\begin{bmatrix} C=128 \\ L=3 \end{bmatrix}$
4	$\frac{K}{4} \times \frac{K}{4}$	$\begin{bmatrix} C=256 \\ L=2 \end{bmatrix}$	$\begin{bmatrix} C=256 \\ L=2 \end{bmatrix}$
Parameters (MB)		38.21	30.98
FLOPs (G)		3.89	3.96

D. Quaternion Transformer Network

The proposed Quaternion Transformer Network (QTN) has a simple hierarchical structure, *i.e.*, a sequence of four stages with decreasing output spatial resolution, *i.e.*, $\frac{K}{2} \times \frac{K}{2}$, $\frac{K}{2} \times \frac{K}{2}$, $\frac{K}{4} \times \frac{K}{4}$, and $\frac{K}{4} \times \frac{K}{4}$ respectively. Here, K and K denote the height and width of the input HSIs patch. With the decreasing of spatial resolution, the number of channels increases. The change of output channels is listed in Table I. It is noted that C denotes the output channel, and L is the repeat times.

For each stage as shown in Figure 1, we first down-sample the quaternion input. After down-sampling, all other layers in a stage remain the same output size. Secondly, the L groups of 1×1 Conv, GELU activation function [71], quaternion self-attention, Conv layer, and MLP are stacked in one sequence to extract local spatial and long-range dependencies features. It is noted that the convolution operation is integrated into the MLP rather than the layer norm operation. Finally, we apply one-layer normalization at the end of QTN. We devise two architectures QTN-Small and QTN-Tiny according to the parameters, computation cost, and HSIs.

E. Discussions

1) Difference to HiT: There are three differences between the proposed QTN and the previous transformer-based methods, *e.g.,* HiT [65]. First, the network architecture is different. Although QTN is an evolution of HiT, it has a BASM to generate the quaternion data from the original hyperspectral image. Additionally, the proposed QTN utilizes the QSA module to capture and aggregate the local context features and the global representations. The second difference is the attention mechanism. Within QTN, the attention mechanism QSA is built on the quaternion convolution layer. Moreover, the proposed QSA uses quaternion convolution layers in series to capture the global representation, while HiT uses convolution layers in parallel. The third one is the domain. The proposed QTN is implemented on the quaternion domain, but the HiT is built on the real domain.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets and Experimental Settings

In this section, we conduct extensive experiments to evaluate the performance of the proposed QTN based on three widely used real HSIs datasets, including the Indian Pines scene (IndianPines), the University of Pavia scene (PaviaU), and the Houston2013 scene (Houston).

TABLE II
NUMBER OF TRAINING/TESTING PIXELS ON INDIAN PINES DATASET

Class No.	Name	Training	Testing
1	Alfalfa	5	41
2	Corn-notill	143	1285
3	Corn-mintill	83	747
4	Corn	24	213
5	Grass-pasture	48	435
6	Grass-trees	73	657
7	Grass-pasture-mowed	3	25
8	Hay-windrowed	48	430
9	Oats	2	18
10	Soybean-notill	97	875
11	Soybean-mintill	246	2209
12	Soybean-clean	59	534
13	Wheat	21	184
14	Woods	127	1138
15	Buildings-Grass-Trees-Drives	39	347
16	Stone-Steel-Towers	9	84
Total		1027	9222

TABLE III
NUMBER OF TRAINING/TESTING PIXELS ON PAVIAU DATASET

Class No.	Name	Training	Testing
1	Asphalt	663	5968
2	Meadows	1865	16784
3	Gravel	210	1889
4	Trees	306	2758
5	Painted metal sheets	135	1210
6	Bare Soil	503	4526
7	Bitumen	133	1197
8	Self-Blocking Bricks	368	3314
9	Shadows	95	852
Total		4278	38498

1) IndianPines Dataset: It was obtained by AVIRIS sensors over the Indian Pines test site in North-western India in 1992. This HSI data contains 145×145 in the spatial dimension, and 224 spectral bands. In the experiment, only 200 bands are applied with discarding the noise bands. It has 16 classes, including Alfalfa, Corn, Woods, etc. Table II reports the number of pixels on the IndianPines dataset.

2) PaviaU Dataset: The second scene was acquired by the ROSIS sensor at the University of Pavia, Italy. This image has 610×340 pixels and 103 spectral bands. There are 9 classes of the PaviaU dataset, including Asphalt, Gravel, trees, etc. The number of training and testing pixels are listed in Table III.

3) Houston Dataset: The third scene was captured in Houston by the ITRES-CASI 1500 sensor in 2012. It was supplied at the 2013 IEEE GRSS Data Fusion Contest. Houston dataset consists of 349×1905 pixels and 144 bands. There are 15 categories of land cover in the Houston dataset, including trees, soil, and water. Table IV shows the detailed number of pixels for training and testing.

4) Comparison Methods: In order to demonstrate the effectiveness of the proposed QTN, we choose four state-of-the-art CNN-based approaches (such as 2D CNN [23], 3D CNN [43], HybridSN [60] and SyCNN [14]), and newly transformers-based approaches (*e.g.*, ViT [31], DeepViT [33], CvT [42], SwinT [32], and HiT [65]). Details of comparison methods are introduced as follows:

- 2D CNN [23]: Three 2D convolution layers are stacked to extract spatial-spectral features for hyperspectral image classification.

TABLE IV
NUMBER OF TRAINING/TESTING PIXELS ON HOUSTON DATASET

Class No.	Name	Training	Testing
1	Healthy Grass	125	1126
2	Stressed grass	125	1129
3	Synthetic Grass	70	627
4	Trees	124	1120
5	Soil	124	1118
6	Water	33	292
7	Residential	127	1141
8	Commercial	124	1120
9	Road	125	1127
10	Highway	123	1104
11	Railway	124	1111
12	Parking Lot 1	123	1110
13	Parking Lot 2	47	422
14	Tennise Court	43	385
15	Running Track	66	594
Total		1503	13526

- 3D CNN [43]: Three 3D convolution layers are stacked to analyze the spatial-spectral together and extract features.
- HybridSN [60]: HybridSN consists of several 3D and 2D convolution layers to extract the spatial and spectral features.
- SyCNN [14]: A hybrid network that consists of a 2D CNN branch and a 3D CNN branch for extract different spatial-spectral features.
- ViT [31]: A classic transformer network that transposes the 2D inputs into 1D sequence inputs and uses a multi-head self-attention module to extract features.
- DeepViT [33]: It is an improvement transformer network of ViT, which suggests recalculating the attention of each head.
- CvT [42]: It stacks three mixed blocks, each of which consists of convolution and attention. In the mixed block, convolution is used to embed and downsample the inputs.
- SwinT [32]: Four swin transformer blocks are stacked to extract the local and global representations for hyperspectral image classification. Each Swin transformer block has a W-MSA and an SW-MSA. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively.
- HiT [65]: A special Hyperspectral image classification method, which is a transformer-based method by using the convolution operation to capture the local context features and global representation.

5) Metrics: In order to quantitatively evaluate the performances of different approaches, we adopt two metrics, including the overall accuracy (OA), and Kappa coefficient (κ). The OA is the ratio of the number of correctly classified samples to the total number of test samples, and the κ denotes a statistical measure of the degree of agreement. These two evaluation metrics are formulated as follows:

$$OA = \left(\frac{1}{n}\sum_{k}\frac{TP+TN}{TP+TN+FN+FP}\right),$$
$$\kappa = \frac{N\sum_{i=1}^{n} x_{ii} - \sum_{i=1}^{n}(x_{i+}\times x_{+i})}{N^2 - \sum_{i=1}^{n}(x_{i+}\times x_{+i})}. \quad (16)$$

where TP, FP, and FN are the true positive value, the false positive value, and the false negative value, respectively. n denotes the number of categories, and N is the total number of data samples. x_{ii} is the number of categories on the diagonal of the confusion matrix, x_{i+} denotes the total number of the $i-th$ row, and x_{+i} represents the total number of the $i-th$ column.

6) Implementation Details: All approaches are implemented on a desktop PC with an Intel Core 7 Duo CPU (at 3.40 GHz), 64 GB of RAM, and one GTX 1080ti GPU (11 GB of ROM). We adopt the Adam gradient descent optimizer with an initial learning rate of 0.6 to train the models. The mini-batch size is set to 100. The dropout ratio and weight decay rate are respectively set to 0.5 and $5e^{-4}$. It should be noted that all experiments were conducted 10 times.

B. Experiment Results on the IndianPines, PaviaU, and Houston Datasets

In this section, we compare the classification performance of the proposed QTN with other traditional approaches, CNNs, and Transformers. We randomly choose 10% labelled samples from each class as the training samples, and the remaining as the testing samples. Table V, VI and VII present the quantitative experimental results about OA, κ, with the best results are marked in bold.

According to Tables V, VI and VII, we can find that the CNN-based methods perform satisfactory results, which attributes to their strong ability to extract local context features. Among the CNN-based methods, 3D-CNN-based methods, such as 3D CNN, HybridSN, and SyCNN, are composed of 3D convolution operations, which could capture spectral-spatial features. However, it can be seen that 2D CNN outperforms those 3D-CNN-based methods. The reason is possible because the 3D convolution operations make these methods need many samples to be trained. Therefore, 3D CNN, HybridSN and SyCNN perform worse than 2D CNN. For example, the OAs of 2D CNN, 3D CNN, HybridSN and SyCNN on the IndianPines dataset are 91.67% $\pm$ 0.51, 76.49% $\pm$ 2.79, 63.69% $\pm$ 20.15 and 87.65%$\pm$1.14, respectively. It is noted that bold characters represent the best results in the tables.

But for transformer-based methods, their performances are very different. For example, the transformers-based approaches (*e.g.*, CvT, SwinT, and HiT) of introducing convolution operation achieve better results than the classic transformers-based approaches (*e.g.*, ViT and DeepViT). It can be seen that the convolution-based transformer methods (especially HiT and SwinT) outperform classic transformer methods (such as ViT and DeepViT). It demonstrates that the convolution operation could introduce the local context features into transformers. Moreover, the proposed QTN achieves the best classification results, in which the OA and κ obtained are invariably outstanding to other comparison methods. In the following discussion, we will choose a representative network in each category.

DeepViT is a special classic transformer architecture that fully utilizes the self-attention mechanism. Compared QTN with DeepViT, QTN overtakes DeepViT by 22.76% (95.50%$\pm$0.27 vs. 71.60%$\pm$3.20) on the IndianPines dataset and 1.66% (98.85% $\pm$0.19 vs. 97.06%$\pm$0.32) on the PaviaU

TABLE V

COMPARE WITH THE STATE-OF-THE-ART CNNS AND TRANSFORMERS ON INDIAN PINES SCENE DATASET (10% TRAINING SAMPLES)

Class	2D CNN	3D CNN	HybridSN	SyCNN	ViT	DeepViT	SwinT	CvT	HiT	QTN
1	23.29 ±17.33	0.00 ±0.00	4.34 ±7.19	66.39 ±25.19	12.47 ±15.07	6.85 ±14.89	77.07 ±19.43	33.87 ±35.19	33.78 ±20.93	**93.32 ±2.00**
2	90.32 ±1.34	67.36 ±3.79	56.80 ±26.69	84.23 ±2.97	68.92 ±5.49	80.40 ±4.82	91.07 ±1.49	56.27 ±30.51	88.79 ±2.45	**95.18 ±0.59**
3	88.38 ±2.57	69.81 ±8.58	46.63 ±24.53	82.11 ±2.96	60.95 ±13.04	72.66 ±5.51	87.20 ±3.33	49.52 ±34.44	87.98 ±2.64	**92.87 ±0.39**
4	88.66 ±4.60	30.43 ±19.30	42.66 ±39.47	88.12 ±5.27	82.00 ±10.71	86.59 ±5.25	93.98 ±2.88	48.35 ±37.71	93.54 ±3.12	**99.25 ±0.46**
5	92.92 ±1.02	84.71 ±9.45	49.02 ±39.53	88.06 ±2.05	41.28 ±14.68	50.45 ±8.84	78.64 ±13.55	75.00 ±20.17	86.49 ±10.17	**95.39 ±1.05**
6	93.64 ±0.98	85.64 ±5.01	62.53 ±28.64	90.25 ±1.63	82.95 ±3.34	80.85 ±2.80	91.65 ±1.87	81.27 ±22.24	90.93 ±1.59	**96.99 ±0.87**
7	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	18.03 ±29.45	24.96 ±31.40	0.00 ±0.00	**66.98 ±20.22**
8	84.82 ±1.18	98.52 ±1.24	67.79 ±41.91	96.74 ±1.01	99.09 ±1.09	98.97 ±0.54	98.35 ±0.92	93.10 ±7.64	97.89 ±1.29	**99.66 ±0.30**
9	95.47 ±0.63	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	12.28 ±18.29	21.83 ±29.47	0.00 ±0.00	**53.21 ±40.82**
10	84.66 ±2.23	70.88 ±2.45	37.78 ±34.26	79.17 ±2.25	44.47 ±9.80	61.00 ±4.17	84.83 ±1.71	55.37 ±33.49	82.77 ±2.63	**88.93 ±0.73**
11	92.39 ±5.25	79.83 ±2.69	76.00 ±15.15	90.30 ±0.87	74.03 ±4.10	81.51 ±3.91	94.03 ±1.01	71.58 ±19.16	93.37 ±1.34	**97.16 ±0.39**
12	98.30 ±0.53	61.95 ±6.17	40.47 ±34.49	84.20 ±4.17	41.91 ±14.66	70.58 ±8.69	85.33 ±3.59	60.62 ±16.36	84.92 ±3.44	**93.14 ±0.62**
13	92.19 ±4.02	50.97 ±26.04	41.26 ±36.96	93.26 ±3.09	87.30 ±5.84	86.71 ±5.49	90.04 ±3.63	60.41 ±36.86	93.06 ±5.19	**96.70 ±1.10**
14	98.30 ±0.53	92.63 ±2.07	76.09 ±27.01	94.27 ±1.49	87.52 ±1.91	89.97 ±1.40	95.07 ±3.11	89.84 ±9.17	96.03 ±2.12	**98.79 ±0.20**
15	92.19 ±4.02	67.47 ±15.90	36.35 ±36.15	90.30 ±6.08	81.32 ±8.36	88.60 ±5.11	94.07 ±3.25	64.14 ±32.86	93.00 ±2.91	**97.91 ±0.67**
16	6.80 ±13.07	0.00 ±0.00	0.00 ±0.00	59.83 ±17.59	1.12 ±3.37	25.97 ±24.80	27.45 ±23.53	39.15 ±37.40	25.13 ±26.36	**63.32 ±6.63**
OA	91.67 ±0.51	76.49 ±2.79	63.69 ±20.15	87.65 ±1.14	71.60 ±3.20	78.82 ±2.82	90.52 ±1.21	70.40 ±16.40	90.20 ±1.05	**95.50 ±0.27**
kappa	90.46 ±0.58	72.71 ±3.29	56.33 ±26.32	85.91 ±1.32	66.75 ±3.90	75.61 ±3.25	89.17 ±1.40	66.14 ±18.53	88.80 ±1.19	**94.85 ±0.30**

TABLE VI

COMPARE WITH THE STATE-OF-THE-ART CNNS AND TRANSFORMERS ON UNIVERSITY OF PAVIA SCENE DATASET (10% TRAINING SAMPLES)

Class	SVM	KNN	2D CNN	3D CNN	SyCNN	ViT	DeepViT	CvT	SwinT	QTN
1	98.47 ±0.36	91.38 ±2.23	90.24 ±4.61	95.15 ±1.03	96.83 ±0.54	96.55 ±1.08	97.71 ±0.37	98.30 ±1.11	97.20 ±0.61	**99.03 ±0.56**
2	99.85 ±0.01	99.62 ±0.11	99.41 ±0.37	99.57 ±0.10	99.18 ±0.19	99.47 ±0.13	99.72 ±0.11	99.80 ±0.07	99.68 ±0.07	**99.89 ±0.02**
3	98.64 ±0.47	93.31 ±3.02	84.26 ±13.81	93.22 ±1.29	94.10 ±1.35	94.06 ±2.40	98.09 ±0.56	98.10 ±1.34	97.91 ±0.59	**99.33 ±0.21**
4	94.55 ±0.78	87.24 ±1.14	78.75 ±6.97	86.35 ±3.37	90.52 ±1.35	90.23 ±1.78	92.55 ±1.52	93.05 ±4.00	92.06 ±0.63	**96.36 ±1.37**
5	93.88 ±0.58	93.72 ±0.64	92.97 ±0.93	91.53 ±0.44	93.86 ±0.93	**95.61 ±1.33**	93.53 ±0.25	94.76 ±2.33	94.32 ±0.96	94.38 ±1.09
6	99.89 ±0.05	98.30 ±0.72	99.45 ±0.42	99.23 ±0.25	98.35 ±0.51	99.01 ±0.27	99.76 ±0.29	99.91 ±0.09	99.14 ±0.21	**99.99 ±0.01**
7	99.20 ±0.21	90.83 ±4.56	87.85 ±6.36	96.05 ±1.72	97.08 ±1.09	96.49 ±3.09	98.74 ±0.25	98.13 ±2.24	97.51 ±1.51	**98.67 ±0.31**
8	98.93 ±0.17	91.17 ±3.95	77.53 ±19.03	93.56 ±2.78	96.78 ±1.10	96.41 ±1.48	98.83 ±0.14	97.03 ±4.54	97.28 ±0.68	**99.42 ±0.14**
9	84.23 ±1.21	73.93 ±8.60	73.47 ±4.66	79.15 ±0.96	81.86 ±2.08	85.93 ±3.20	81.31 ±1.90	86.23 ±3.31	83.12 ±1.86	**86.07 ±1.98**
OA(%)	98.60 ±0.12	95.32 ±0.85	92.82 ±3.17	96.22 ±0.61	97.06 ±0.32	97.28 ±0.57	98.14 ±0.23	98.28 ±1.06	97.85 ±0.17	**98.95 ±0.19**
κ(%)	98.14 ±0.15	93.80 ±1.13	90.49 ±4.19	95.00 ±0.81	96.10 ±0.43	96.39 ±0.75	97.54 ±0.31	97.72 ±1.40	97.14 ±0.23	**98.61 ±0.25**

TABLE VII

COMPARE WITH THE STATE-OF-THE-ART CNNS AND TRANSFORMERS ON HOUSTON2013 SCENE DATASET (10% TRAINING SAMPLES)

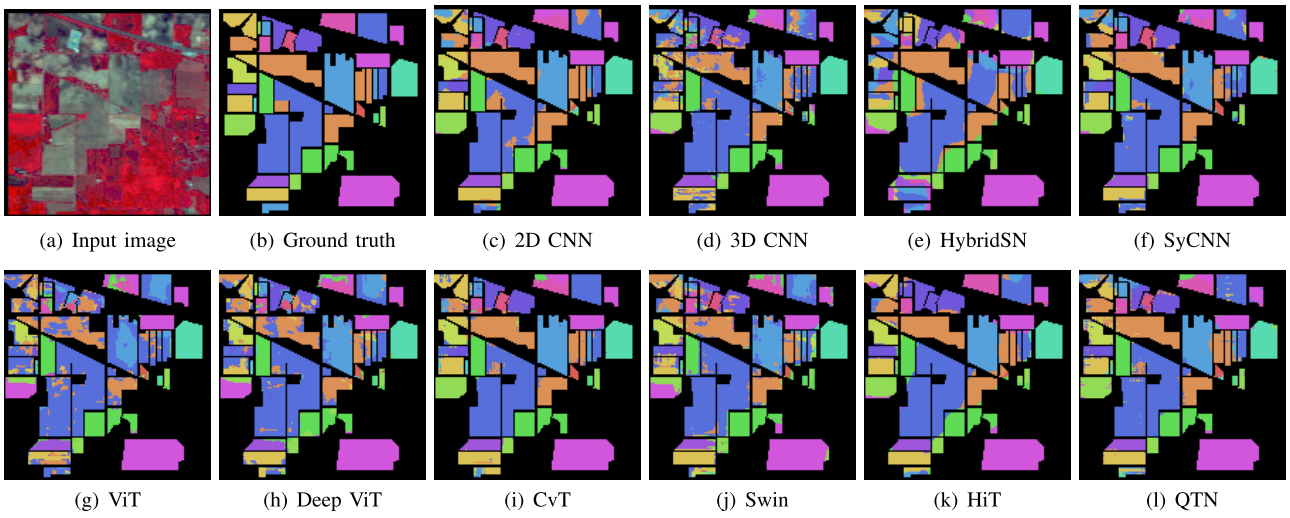
Class	2D CNN	3D CNN	HybridSN	SyCNN	ViT	DeepViT	CvT	SwinT	HiT	QTN
1	89.59 ± 1.00	89.08 ± 1.98	87.34 ± 5.14	91.01 ± 1.42	84.13 ± 1.29	84.11 ± 3.87	91.37 ± 1.49	88.62 ± 9.02	92.55 ± 0.83	**94.28 ± 2.04**
2	92.73 ± 1.10	85.15 ± 4.63	85.20 ± 3.22	91.03 ± 1.37	67.38 ± 4.37	74.45 ± 5.89	90.85 ± 2.33	89.24 ± 6.97	89.36 ± 2.71	**94.95 ± 2.38**
3	98.88 ± 0.28	96.67 ± 1.81	97.33 ± 1.66	90.97 ± 2.44	96.55 ± 1.12	96.85 ± 1.37	98.42 ± 0.49	98.59 ± 0.71	98.45 ± 0.35	**99.28 ± 0.23**
4	88.82 ± 1.56	83.56 ± 3.97	73.78 ± 5.71	87.51 ± 3.21	74.56 ± 5.79	69.78 ± 10.64	89.03 ± 1.99	88.43 ± 5.06	79.34 ± 2.77	**89.77 ± 4.37**
5	99.60 ± 0.30	93.32 ± 3.89	93.74 ± 2.49	97.05 ± 1.09	96.65 ± 1.76	97.75 ± 1.27	**99.28 ± 0.45**	97.81 ± 2.77	99.12 ± 0.53	98.43 ± 1.24
6	87.72 ± 2.78	79.51 ± 9.28	85.49 ± 2.95	84.81 ± 4.61	81.49 ± 3.00	86.17 ± 3.26	90.35 ± 2.75	**92.22 ± 2.13**	86.29 ± 1.39	88.06 ± 4.46
7	90.65 ± 2.04	81.21 ± 2.83	76.65 ± 10.02	86.01 ± 1.71	77.54 ± 2.58	76.19 ± 3.74	91.48 ± 1.98	89.92 ± 4.11	81.47 ± 5.16	**91.72 ± 4.02**
8	93.09 ± 1.61	78.76 ± 4.93	85.69 ± 2.51	86.59 ± 2.30	78.20 ± 2.38	83.74 ± 3.31	95.21 ± 1.31	92.81 ± 5.70	91.67 ± 2.06	**96.05 ± 1.58**
9	86.63 ± 2.83	77.02 ± 3.54	59.85 ± 13.88	82.67 ± 2.28	78.24 ± 1.38	77.61 ± 4.85	89.69 ± 1.74	84.95 ± 3.81	84.54 ± 2.61	**93.83 ± 2.45**
10	94.26 ± 1.93	79.50 ± 7.46	89.04 ± 7.23	95.27 ± 1.19	83.10 ± 3.14	87.45 ± 2.90	97.14 ± 1.72	91.87 ± 7.10	95.08 ± 1.78	**99.69 ± 0.23**
11	94.39 ± 1.60	77.41 ± 4.85	66.04 ± 17.27	85.77 ± 2.13	70.70 ± 7.89	80.45 ± 5.86	96.75 ± 2.49	93.41 ± 7.21	92.52 ± 1.76	**99.23 ± 0.88**
12	94.57 ± 2.03	81.93 ± 4.27	88.89 ± 5.27	94.53 ± 0.83	72.81 ± 3.73	83.87 ± 4.47	96.19 ± 2.49	87.16 ± 11.43	93.20 ± 1.31	**98.17 ± 0.62**
13	95.15 ± 0.90	75.80 ± 5.19	62.22 ± 10.03	79.97 ± 3.55	71.02 ± 11.02	80.45 ± 7.86	92.04 ± 4.23	91.60 ± 5.43	89.88 ± 3.80	**96.65 ± 2.14**
14	99.94 ± 0.06	95.40 ± 3.86	87.75 ± 9.67	96.17 ± 2.25	94.46 ± 3.08	92.68 ± 3.75	99.07 ± 0.75	98.34 ± 3.58	99.59 ± 0.59	**99.83 ± 0.41**
15	96.90 ± 1.57	91.84 ± 3.43	86.26 ± 5.52	86.66 ± 4.44	86.82 ± 5.55	93.19 ± 3.99	97.75 ± 0.61	96.06 ± 2.34	97.03 ± 1.20	**99.17 ± 0.31**
OA(%)	93.07 ± 0.76	83.94 ± 2.21	81.64 ± 3.41	89.32 ± 0.81	79.79 ± 1.76	83.21 ± 2.19	94.09 ± 1.12	91.39 ± 3.77	90.74 ± 0.92	**95.90 ± 1.39**
κ(%)	92.50 ± 0.82	82.63 ± 2.39	80.19 ± 3.66	88.46 ± 0.88	78.14 ± 1.90	81.84 ± 2.37	93.61 ± 1.21	90.69 ± 4.08	89.99 ± 0.99	**95.57 ± 1.50**

TABLE VIII

EXPERIMENT RESULTS WITH LARGE TRAINING SAMPLES (70%)

Methods	IndianPines		PaviaU		Houston	
	OA(%)	κ(%)	OA(%)	κ(%)	OA(%)	κ(%)
2D CNN	93.66 ± 1.55	92.76 ± 1.77	99.01 ± 0.13	98.69 ± 0.18	97.67 ± 0.57	97.48 ± 0.62
3D CNN	91.08 ± 0.58	89.80 ± 0.67	97.30 ± 0.35	96.42 ± 0.46	90.50 ± 0.92	89.73 ± 0.99
SyCNN	85.97 ± 11.06	84.01 ± 12.53	97.91 ± 0.28	97.24 ± 0.37	94.76 ± 0.61	94.33 ± 0.66
HybridSN	93.04 ± 0.60	92.07 ± 0.67	96.07 ± 1.12	94.79 ± 1.48	92.47 ± 4.29	91.86 ± 4.64
ViT	93.40 ± 0.54	92.45 ± 0.62	98.29 ± 0.23	97.73 ± 0.30	94.61 ± 0.62	94.18 ± 0.67
DeepViT	94.22 ± 0.55	93.39 ± 0.63	98.48 ± 0.22	97.98 ± 0.29	95.48 ± 0.48	95.11 ± 0.52
CvT	84.22 ± 11.26	82.20 ± 12.28	98.69 ± 0.76	98.27 ± 1.00	97.89 ± 0.38	97.71 ± 0.41
Swin	94.13 ± 0.39	93.29 ± 0.44	98.88 ± 0.11	98.51 ± 0.15	98.01 ± 0.38	97.88 ± 0.42
HiT	93.39 ± 0.38	92.45 ± 0.43	98.67 ± 0.12	98.24 ± 0.15	95.80 ± 0.61	95.46 ± 0.66
QTN	**95.61 ± 0.27**	**94.99 ± 0.31**	**99.04 ± 0.17**	**98.72 ± 0.22**	**98.18 ± 0.77**	**98.03 ± 0.84**

dataset. This is mainly because of the introduction of local spatial information and channel adaptability by using QSA. CvT is a well-known ViT variant, which introduces the convolution operation. Due to the QTN being friendly for 3D structural information and latent interaction between features, QTN surpasses CvT by 25.10% (95.50%±0.27 vs. 70.40%±16.40) on the IndianPines dataset and 8.92% (98.85% ±0.19 vs. 90.43%) on the PaviaU dataset. Moreover, the proposed QTN displays strong robustness and achieves the best classification results on the IndianPines, PaviaU,



(a) Input image (b) Ground truth (c) 2D CNN (d) 3D CNN (e) HybridSN (f) SyCNN

(g) ViT (h) Deep ViT (i) CvT (j) Swin (k) HiT (l) QTN

Fig. 4. Classification maps obtained by different methods on the Indian Pines scene dataset (with 10% training samples).

and Houston datasets. Furthermore, the proposed QTN outperforms other comparison methods in the categories with few samples. For example, it performs better classification results in the "Alfalfa" and "Oats" categories on the IndianPines dataset. We can also observe that the CNNs and transformers approaches can not recognize the special class, which only has a few samples. It demonstrates the superiority and effectiveness of the proposed QTN.

Figures. 4, 5, and 6 show the classification maps obtained by different approaches, along with the false-color HSIs and its corresponding ground truth. According to Figures. 4, 5, and 6, we can observe that there has obvious noise in the classification maps of some comparison approaches, and the proposed QTN produces more exact classification maps. In particular, the proposed QTN generates more accurate homogeneous regions than other compared approaches. For transformer-based methods, the classic transformer-based methods (such as ViT and DeepViT) use the linear-project operation to embed the inputs, which neglects the 3D structure of HSIs. Therefore, they fail to capture the local context information. Moreover, the misclassified pixels are mostly located in the region central. Since introduced to capture the local information, the transformers (such as CvT, SwinT, HiT, and QTN) result in better classification maps.

For CNN-based methods, they focus on extracting the local spatial-spectral information. Therefore, they could not capture the long-range dependencies, and their misclassified pixels are mostly located at the region edges.

C. Experiment Results With Large Training Samples

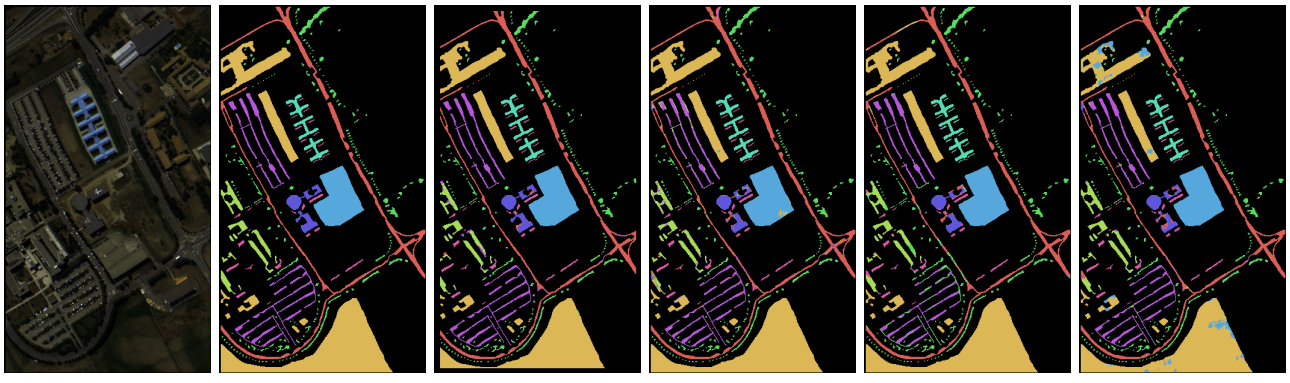
In this section, we will argue the performance of the proposed QTN with large samples (such as 70% training samples). Table VIII reports the results of all methods on 70% training samples. First, we can see that the proposed QTN outperforms all the comparison methods when all methods are trained with more training samples, especially with complex HSI datasets (such as IndianPines and Houston 2013). This demonstrates the superiority and effectiveness of the proposed QTN with 70% training samples. Second, we can also find that CNN-based methods outperform transformer-based methods on the PaviaU dataset. This is possibly because CNN-based methods are better at extracting local spatial information. Finally, we can observe that transformer-based methods achieve better results than CNN-based methods on complex datasets. This is mainly because transformer-based methods are capable of capturing long-range dependencies to improve performance.

D. Complexity Analysis

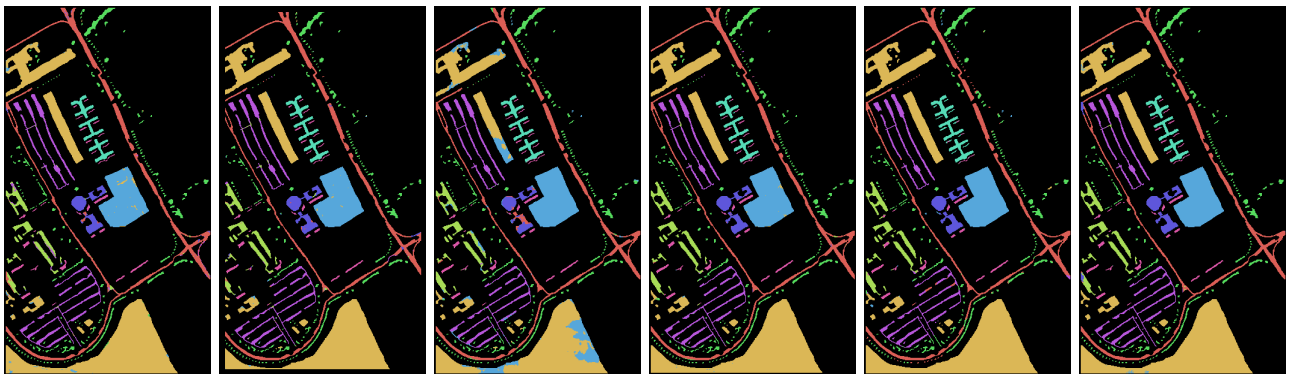
We analyze the complexity of the proposed QTN and comparison methods on the IndianPines dataset in terms of FLOPS, Parameters, Training and Testing times, and Throughput, and report the results in Table IX. It is worth noting that all the methods are re-implemented in a fair environment, which has been introduced in Section V-A. In Table IX, the "F", "P", and "TP" are short for "FLOPS", "Parameters", and "Throughput", respectively.

According to the results of Table IX, we can see that most CNN-based methods have achieved high computational efficiency in terms of FLOPS, Parameters, Training and Testing times, and Throughput. Specifically, 2D CNN takes 2D convolution layers as the basic operations, which results in high computational efficiency in terms of FLOPS, Parameters, Training and Testing times, and Throughput. Since 3D CNN is composed of three 3D convolution layers, it urgently needs many samples to be trained. This results in a low classification result (e.g., OA). Moreover, SyCNN consists of two branches, i.e., 2D CNN and 3D CNN, which leads to its low computational efficiency among CNN-based methods. Although CNN-based methods have high computational efficiency, they fail to capture the long-range dependencies features, which limits further improving the performance.

Different from CNN-based methods, transformer-based methods contain a series of multi-head self-attention (MHSA) modules to capture the long-range dependencies representation. However, MHSA modules lead to the low computational efficiency of transformers. From Table IX, we can see that all the existing transformer-based methods result in low



(a) Input image (b) Ground truth (c) 2D CNN (d) 3D CNN (e) HybridSN (f) SyCNN



(g) ViT (h) Deep ViT (i) CvT (j) Swin (k) HiT (l) QTN

Fig. 5. Classification maps obtained by different methods on the University of Pavia scene dataset (with 10% training samples).

TABLE IX

COMPLEXITY ANALYSIS OF ALL METHODS ON INDIAN PINES DATASET

Methods	F(G)	P(MB)	Training(S)	Testing(S)	TP	OA(%)
2D CNN	0.07	0.49	13.57	1.36	1658	91.67 ± 0.51
3D CNN	0.27	1.45	96.69	5.98	565	76.49 ± 2.79
HybridSN	5.31	4.32	91.22	5.46	12.34	63.69 ± 20.15
SyCNN	3.88	45.82	215.53	13.61	57	87.65 ± 1.14
ViT	2.71	52.22	93.69	6.12	42	71.60 ± 3.20
DeepViT	2.71	52.22	103.86	6.56	40	78.82 ± 2.82
CvT	23.44	45.89	1243.31	63.92	11	90.52 ± 1.21
SwinT	1.57	43.41	70.95	4.12	74	70.40 ± 16.40
HiT	2.33	51.18	112.04	6.70	20	90.20 ± 1.05
QTN	3.89	38.21	277.41	37.47	26	**95.50 ± 0.27**

computational efficiency in terms of FLOPS, Parameters, Training and Testing times, and Throughput. Moreover, some methods (e.g., ViT, DeepViT, HiT, and CvT) perform worse classification results than CNN-based methods. Compared to other transformers, however, both QTN, HiT, and SwinT perform better classification results. This demonstrates that the jointing of convolution operation in the transformers could improve their performance. Furthermore, the proposed QTN has few parameters. A possible explanation for this might be that QTN is built on QSA, which is a lightweight self-attention module. However, calculating the attention maps with convolution operations leads to the low computational efficiency of QTN in terms of FLOPS, Training and Testing times, and Throughput. This is the limitation of the proposed QTN. My future work continues to design a lightweight transformer network in the quaternion field.

TABLE X

ABLATION STUDY OF DIFFERENT BAND SELECTION MODULES IN QUATERNION TRANSFORMER NETWORK

Methods	IndianPines		PaviaU		Houston	
	OA(%)	κ(%)	OA(%)	κ(%)	OA(%)	κ(%)
PCA [72]	87.53	85.90	91.08	88.60	83.66	82.40
NMF [53]	81.84	79.60	91.23	88.80	92.89	92.30
LDA [54]	86.18	84.40	90.86	88.40	74.08	72.00
FNGBS [73]	93.98	93.11	99.03	98.72	91.23	90.52
OCF [74]	93.15	92.18	98.98	98.64	88.41	87.47
ONR [74]	93.75	92.85	99.10	98.81	89.95	89.13
BASM	**95.00**	**94.29**	**99.35**	**99.13**	**96.16**	**95.85**

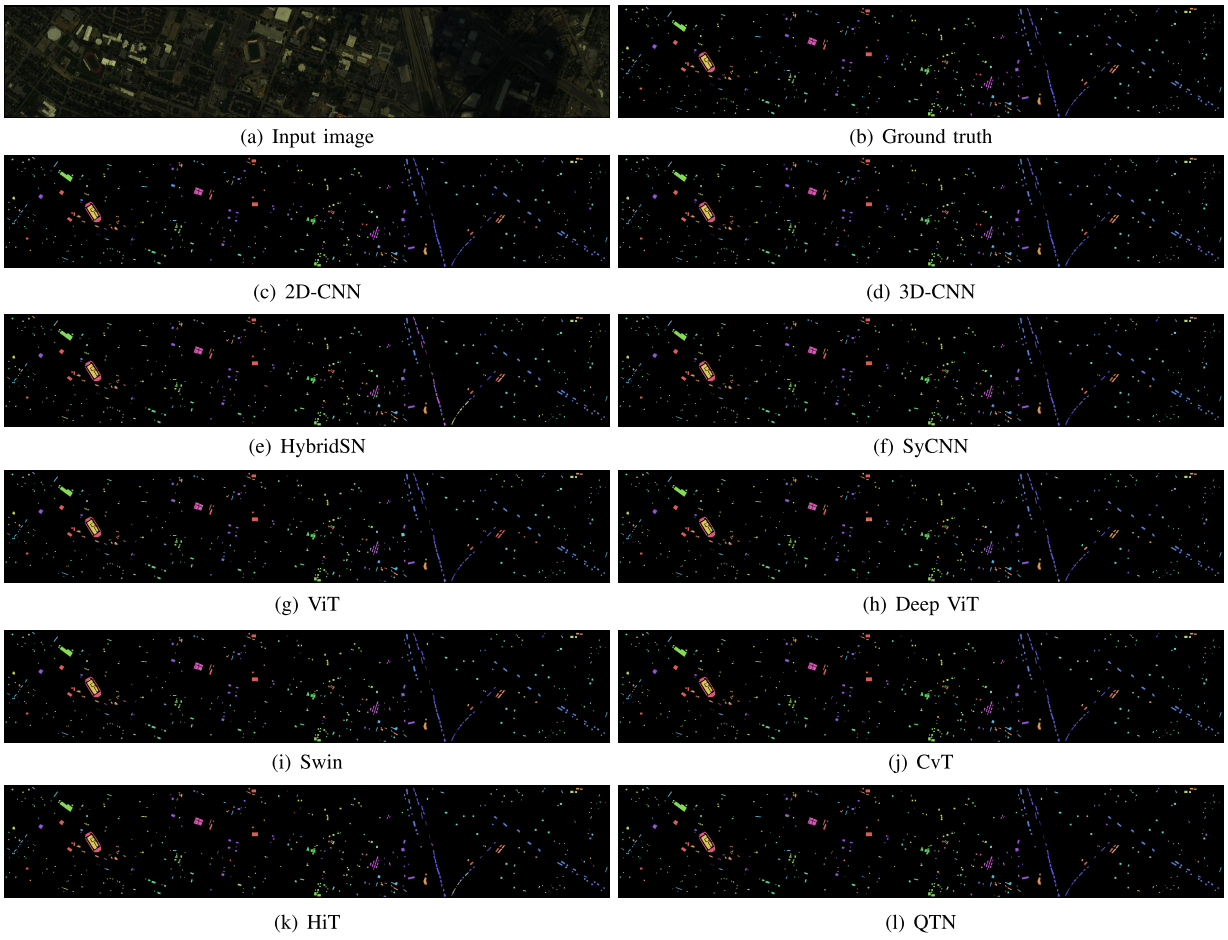
TABLE XI

ABLATION STUDY OF DIFFERENT COMPONENTS IN BAND ADAPTIVE SELECTION MODULE

Methods	IndianPines		PaviaU		Houston	
	OA(%)	κ(%)	OA(%)	κ(%)	OA(%)	κ(%)
PCA	87.53	85.90	91.08	88.60	83.66	82.40
w/imaginary	91.59	90.38	97.80	97.09	90.96	90.23
w/real	92.34	91.23	98.24	97.67	90.89	90.15
w/BASM	**95.00**	**94.29**	**99.35**	**99.13**	**96.16**	**95.85**

E. Ablation Studies

1) Ablation Study of Different Band Selection Modules: In this study, we investigate the band adaptive selection module (BASM). We conduct the experiment with different band selection modules on three datasets, including IndianPines, PaviaU and Houston. We choose three feature extraction methods (such as PCA [72], LDA [54] and NMF [53]) and three band selection methods (*e.g.*, ONR [55],Fast Neighborhood Grouping Band Selection(FNGBS) [73] and Optimal Clustering Framework (OCF) [74]). Table X presents the



(a) Input image (b) Ground truth (c) 2D-CNN (d) 3D-CNN (e) HybridSN (f) SyCNN (g) ViT (h) Deep ViT (i) Swin (j) CvT (k) HiT (l) QTN

Fig. 6. Classification maps obtained by different methods on the Houston2013 Dataset (with 10% training samples).

TABLE XII
ABLATION STUDY OF DIFFERENT ATTENTION MODULES IN QTN

Methods	IndianPines			PaviaU			Houston		
	OA(%)	FLOPS(G)	Parameters(MB)	OA(%)	FLOPS(G)	Parameters(MB)	OA(%)	FLOPS(G)	Parameters(MB)
w/SA	85.34	0.50	1.86	97.19	0.50	1.86	80.16	0.50	1.86
w/CAM	92.50	37.69	58.89	98.94	37.69	58.89	94.41	37.69	58.89
w/QSA	**95.00**	3.89	38.21	**99.35**	3.89	38.21	**96.16**	3.89	38.21

TABLE XIII
ABLATION STUDY OF DIFFERENT PARAMETERS IN QTN

Methods	QTN_Small			QTN_Tiny		
	OA(%)	F(G)	P(MB)	OA(%)	F(G)	P(MB)
IndianPines	93.35	3.96	30.98	**95.00**	3.89	38.21
PaviaU	98.46	3.96	30.98	**99.35**	3.89	38.21
Houston	93.34	3.96	30.98	**96.16**	3.89	38.21

results of QTN with different band selection modules. We can see that the dimensionality reduction methods (in particular, PCA, NMF, and LDA) lead to low classification results, especially the performance of LDA. A possible explanation for this might be that these methods could not exploit the correlation between neighboring bands. By exploiting the strong correlation between neighboring bands, the FNGBS, OCF, and ONR are implemented independently to select bands and perform better results than the dimensionality reduction ones. However, they may reserve inapplicable bands for QTN. Finally, we can observe that the QTN with BASM outperforms other comparison band selection methods. There are two possible explanations for this result. First, BASM is built with ECA to fully exploit the correlation between neighboring bands, resulting in adaptively selecting and reserving the applicable bands for QTN. Secondly, BASM consists of a convolution layer to keep the long dependencies of the spectral domain. It is worth noting that we chose one result from the experiments 10 times.

2) Ablation Study of Band Adaptive Selection Modules: In this part, we analyze the different components of BASM, such as the real part and the imaginary part. The experimental results are reported in Table XI. The QTN with PCA is chosen as the baseline, which performs a poor classification result. When we use one part to generate the quaternion data, resulting in a better performance than the one with PCA. It is demonstrated that different parts will lead to satisfactory quaternion data. Finally, the BASM is utilized to generate the quaternion dataset, resulting in the highest classification results. A possible explanation for this might be that BASM consists of real and imaginary parts to exploit the strong correlation between different bands and reserve the long-dependence spectral information.

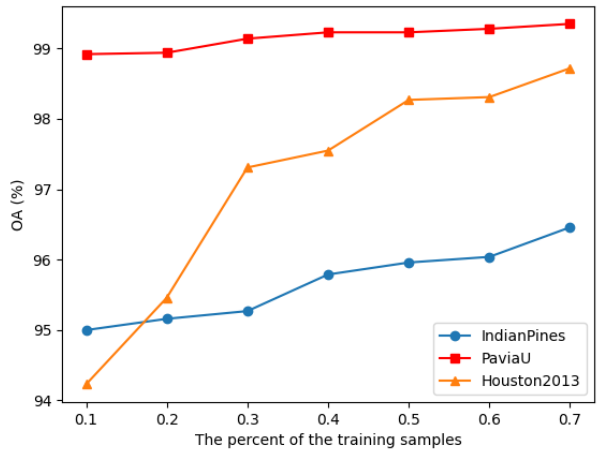


Fig. 7. Ablation study of the percent of training samples on the three datasets.

3) Ablation Study of Quaternion Self-Attention: Table XII presents the ablation investigation on the effects of the self-attention module (with layer norm), the convolution attention module (CAM), and the quaternion self-attention module (QSA). It is noted that the CAM is implemented by using the convolution operation instead of the quaternion convolution operation. From Table XII, we can see that the QTN with the self-attention module results in high computation efficiency in terms of FLOPS and Parameters. However, it achieves poor results on the three datasets. For example, the OAs of IndianPines, PaviaU, and Houston are 85.34%, 97.19%, and 80.16%, respectively. A possible explanation for this might be that the self-attention module is built with layer norm, which ignores the 3D structure of the HSIs. On the other hand, integrating the convolution operation will increase the performance, especially the quaternion convolution operation. Attributing to the integration of convolution operation, QTN with CAM performs better classification results than the one with the self-attention module. However, it results in low computation efficiency in terms of FLOPS and Parameters. Finally, the QTN takes QSA as the basic attention module, which results in the best performance but with fewer parameters. This result could be attributed to the special design of QSA and the quaternion convolution layers. In Table XII, "w/SA" is short for with self-attention, "w/CAM" and "w/QSA" denote with convolution attention module and quaternion self-attention module, respectively.

4) Ablation Study of the Parameters: In this section, we conduct an ablation study to investigate the impact of the number of parameters. The experimental results of different methods are reported in Table XIII. It is worth noting that "F" and "P" are short for FLOPS and Parameters, respectively. We can observe that the OA increases with the growth of the number of parameters. It demonstrates that deep networks could learn a better feature representation and achieve high performance. However, it will result in a low computation efficiency in terms of FLOPS and Parameters. Considering the performance and the parameters, we choose the QTN-Tiny as the baseline network in this paper.

5) Ablation Study of the Percent of Training Samples: In this section, we conduct an ablation study to investigate the percent of training samples. The experimental results of the proposed QTN are reported in Fig. 7. We vary the percent of training samples from 0.1 to 0.7 on the IndianPines and Houston, and from 0.01 to 0.1 on the PaviaU dataset. From Fig. 7, we can observe that the OA increases with the percent of the training samples. A possible explanation for this might be that more training samples could lead the proposed QTN to learn a good representation of hyperspectral image classification.

VI. Conclusion

In this paper, we investigate a novel deep vision transformer network for hyperspectral image classification on the quaternion domain. In hyperspectral image classification, different objects have different spectral curves, and the spatial features are also different. Based on this phenomenon, we exploit a novel transformer architecture to model the hyperspectral image, namely the quaternion transformer network (QTN) to handle the long-range dependencies along the spectral dimension and the local spatial features. In order to transfer the hyperspectral image to quaternion data, we presented a band selection module, namely the band adaptive selection module (BASM) for adaptive selecting of the spectral bands from HSIs and transforming them into quaternion data. Thereafter, utilizing the powerful ability of the transformer, we further propose a novel self-attention mechanism to exploit the long-range dependencies along spectral and spatial dimensions. Therefore, the proposed QTN can extract more discriminative characteristic features for hyperspectral image classification. Finally, extensive experiment results of three groups of HSI classification demonstrated that the QTN outperforms state-of-the-art CNNs and transformers. And the ablation studies certify the effectiveness and superior performance of the proposed modules. Our future works will continue to develop this concept and establish a lightweight and more efficient transformer network in the quaternion field.

References

[1] J. Liu, Y. Feng, W. Liu, D. Orlando, and H. Li, "Training data assisted anomaly detection of multi-pixel targets in hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 68, pp. 3022–3032, 2020.

[2] J. Peng, Y. Zhou, W. Sun, Q. Du, and L. Xia, "Self-paced nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1501–1515, Feb. 2020.

[3] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[4] V. E. Brando and A. G. Dekker, "Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1378–1387, Jun. 2003.

[5] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.

[6] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[7] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[8] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[9] L. Sun et al., "Low rank component induced spatial–spectral kernel method for hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3829–3842, Oct. 2020.

[10] H. Liu, Y. Jia, J. Hou, and Q. Zhang, "Global-local balanced low-rank approximation of hyperspectral images for classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2013–2024, Apr. 2022.

[11] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[12] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectral–spatial classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2077–2081, Nov. 2017.

[13] C. Tao, W. Lu, J. Qi, and H. Wang, "Spatial information considered network for scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 984–988, Jun. 2021.

[14] X. Yang et al., "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 2033, Jun. 2020.

[15] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[16] B. Liu, X. Yu, P. Zhang, X. Tan, A. Yu, and Z. Xue, "A semi-supervised convolutional neural network for hyperspectral image classification," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 839–848, Sep. 2017.

[17] B. Xi et al., "DGSSC: A deep generative spectral–spatial classifier for imbalanced hyperspectral imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1535–1548, Apr. 2022.

[18] B. Rasti et al., "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, Dec. 2020.

[19] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[20] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[21] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[22] 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, Aug. 2018.

[23] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[24] H. Lee and H. Kwon, "Contextual deep CNN based hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3322–3325.

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

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

[27] V. Sharma, A. Diba, T. Tuytelaars, and L. Van Gool, "Hyperspectral CNN for image classification & band selection, with application to face recognition," 2016.

[28] P. R. Lorenzo, L. Tulczyjew, M. Marcinkiewicz, and J. Nalepa, "Band selection from hyperspectral images using attention-based convolutional neural networks," 2018, *arXiv:1811.02667*.

[29] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.

[30] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–11.

[32] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[33] D. Zhou et al., "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[34] L. Yuan et al., "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.

[35] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[36] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 32–42.

[37] B. Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12239–12249.

[38] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[39] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

[40] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[41] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11916–11925.

[42] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[43] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

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

[45] M. Xiang, B. S. Dees, and D. P. Mandic, "Multiple-model adaptive estimation for 3-D and 4-D signals: A widely linear quaternion approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 72–84, Jan. 2019.

[46] B. Chen, M. Yu, Q. Su, and L. Li, "Fractional quaternion cosine transform and its application in color image copy-move forgery detection," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8057–8073, Apr. 2019.

[47] Y. Liu, Y. Zheng, J. Lu, J. Cao, and L. Rutkowski, "Constrained quaternion-variable convex optimization: A quaternion-valued recurrent neural network approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 1022–1035, Mar. 2020.

[48] E. C. Mengüç, N. Acir, and D. P. Mandic, "Widely linear quaternion-valued least-mean kurtosis algorithm," *IEEE Trans. Signal Process.*, vol. 68, pp. 5914–5922, 2020.

[49] J. Flamant, S. Miron, and D. Brie, "Quaternion non-negative matrix factorization: Definition, uniqueness, and algorithm," *IEEE Trans. Signal Process.*, vol. 68, pp. 1870–1883, 2020.

[50] M. Xiang, Y. Xia, and D. P. Mandic, "Performance analysis of deficient length quaternion least mean square adaptive filters," *IEEE Trans. Signal Process.*, vol. 68, pp. 65–80, 2020.

[51] H. Li, H. Li, and L. Zhang, "Quaternion-based multiscale analysis for feature extraction of hyperspectral images," *IEEE Trans. Signal Process.*, vol. 67, no. 6, pp. 1418–1430, Mar. 2019.

[52] H. Li, H. Huang, Z. Ye, and H. Li, "Hyperspectral image classification using adaptive weighted quaternion Zernike moments," *IEEE Trans. Signal Process.*, vol. 70, pp. 701–713, 2022.

[53] R. Rajabi and H. Ghassemian, "Multilayer structured NMF for spectral unmixing of hyperspectral images," in *Proc. 6th Workshop Hyperspectral Image Signal Processing: Evol. Remote Sens. (WHISPERS)*, Jun. 2014, pp. 1–4.

[54] Q. Tian, T. Arbel, and J. J. Clark, "Task dependent deep LDA pruning of neural networks," *Comput. Vis. Image Understand.*, vol. 203, Feb. 2021, Art. no. 103154.

[55] Q. Wang, F. Zhang, and X. Li, "Hyperspectral band selection via optimal neighborhood reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8465–8476, Dec. 2020.

[56] S. L. Al-Khafaji, J. Zhou, X. Bai, Y. Qian, and A. W. Liew, "Spectral–spatial boundary detection in hyperspectral images," *IEEE Trans. Image Process.*, vol. 31, pp. 499–512, 2022.

[57] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 246–259, Jan. 2021.

[58] S. Huang, H. Zhang, and A. Pižurica, "Subspace clustering for hyperspectral images via dictionary learning with adaptive regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524017.
[59] W. Yao, C. Lian, and L. Bruzzone, "ClusterCNN: Clustering-based feature learning for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1991–1995, Nov. 2021.
[60] 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.
[61] 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.
[62] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
[63] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[64] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
[65] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.
[66] Q. Yin, J. Wang, X. Luo, J. Zhai, S. Kr. Jha, and Y. Shi, "Quaternion convolutional neural network for color image classification and forensics," *IEEE Access*, vol. 7, pp. 20293–20301, 2019.
[67] C. J. Gaudet and A. S. Maida, "Deep quaternion networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.
[68] X. Zhu, Y. Xu, H. Xu, and C. Chen, "Quaternion convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 631–647.
[69] S. Zhang, L. Yao, L. V. Tran, A. Zhang, and Y. Tay, "Quaternion collaborative filtering for recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4313–4319.
[70] A. Mackiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.*, vol. 19, pp. 303–342, Mar. 1993.
[71] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[72] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[73] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5028–5039, Jun. 2021.
[74] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018.

Xiaofei Yang received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, China, in 2014 and 2019, respectively. He was a Post-Doctoral with the Department of Computer and Information Science, University of Macau, Macau, China, from 2020 to 2023. Currently, he is with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou, China. His research interests are in the areas of semi-supervised learning, deep learning, remote sensing, transfer learning, and graph mining.

Weijia Cao received the master's and Ph.D. degrees in computer science from the University of Macau, Macau, China, in 2013 and 2017, respectively. She is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her main research interests revolve around machine learning and remote-sensing image processing.

Yao Lu received the B.S. degree in computer science and technology from Huaqiao University, Xiamen, China, in 2015, and the Ph.D. degree in computer applied technology from the Harbin Institute of Technology at Shenzhen, Shenzhen, China, in 2020. She was a Post-Doctoral Fellow with the University of Macau, Macau, China, from 2020 to 2021. She is currently an Assistant Professor with the Biocomputing Research Center, Harbin Institute of Technology. Her research interests include pattern recognition, deep learning, computer vision, and relevant applications.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is currently a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE) and was recognized as one of "Highly Cited Researchers" in 2020 and 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.