3D-CmT: 3D-CNN meets Transformer for Hyperspectral Image Classification

Sunita Arya[1], Shiv Ram Dubey[2], S Manthira Moorthi[1], Debajyoti Dhar[1], and Satish Kumar Singh[2]

[1] Space Applications Centre, Ahmedabad, India
s {sunita33,smmoorthi,deb}@sac.isro.gov.in
[2] Indian Institute of Information Technology, Allahabad, India
{srdubey,sk.singh}@iiita.ac.in

Abstract. In recent years, the combined use of Vision Transformer (ViT) and Convolutional Neural Network (CNN) has shown promising results in tasks related to satellite imagery. In our study, we propose a 3D-CmT (3D-CNN meets Transformer) model for Hyperspectral Image Classification. This model leverages the unique capabilities of both 3D-CNN and ViT to effectively classify images captured by hyperspectral imaging. To learn the local features of the narrow and contiguous electromagnetic spectrum of the hyperspectral images, we utilize a 3D-CNN under the spectral feature extraction (SFE) module. Subsequently, a transformer encoder (TE) module is applied on top of the 3D-CNN to incorporate global attention and model long-range dependencies for spatial information in the images. We conducted experiments using commonly used hyperspectral image datasets and performed various ablation studies, such as evaluating the impact of image patch size and different percentages of training samples. The performance of our proposed model is comparable to that of other CNN-based, transformer-based, and hybrid CNN-Transformer-based models in terms of model parameters and accuracy. In addition, we conducted quantitative and qualitative analyses to assess the performance of our model.

Keywords: Remote Sensing · Hyperspectral Image (HSI) Classification · 3D Convolutional Neural Network (CNN) · Vision Transformer (ViT)

1 Introduction

Hyperspectral image (HSI) acquired using space-borne or air-borne instruments plays a very significant role in geological studies, mineral mapping, and other applications that use its unique capability of narrow spectral bands in the electromagnetic spectrum. With its rich spatial and spectral information, an HSI can be used for various applications, such as mineral studies [18], precision agriculture [19], food safety [20], biomedical imaging [21], and military applications [22].

Along with computer vision and natural language processing, rapid advances in deep learning approaches have pushed the development of signal and image processing techniques in a range of domains [28]. Deep learning-based models

have been widely used for many HSI applications [29] [34], including HSI classification [30] [35] and image fusion [31]. To process the abundant spatial and spectral information of HSI, architectures based on the convolutional neural network (CNN) [33] and transformer [32] have been extensively utilized for HSI analysis and processing. Among these, land use and land cover information classification have attracted much attention using the HSI dataset [43]. Accurate identification and classification of considered targets opens a new field of studies for accurate mapping of minerals and other geological targets. For HSI classification, various works have been done using CNN and transformer-based networks. In addition to that, many researchers have effectively used hybrid models to leverage the use of CNN and transformer features together for HSI classification.

Most transformer-based models use small patches of input image before passing it to the encoder of the transformer network but inspired by the performance of CMT: convolution neural network meets vision transformers [1] model for image classification of Red-Green-Blue (RGB) images. We grounded the transformer model of our proposed model using one of the stages of the CMT transformer. CMT architecture is currently suitable for images with three channels only. Therefore, we proposed a hybrid network of a 3D convolution neural network and a CMT block as a transformer for the classification of hyperspectral images.

The main contributions of this paper are summarized as follows:

1. A simple 3D-CNN and ViT-based hybrid module is proposed in our 3D-CmT network for handling hyperspectral images' spatial and spectral data effectively for pixel-level classification.
2. We devise two simple but effective modules in 3D-CmT, i.e., the 3D-CNN module assisted by Principal Component Analysis (PCA) [17] to extract the most relevant band information through extraction and learning of spectral features, and the second module is the Transformer Encoder which is based on the CMT network [1], to learn global spatial representations using the self-attention mechanism of the vision transformer. Using PCA, we have reduced the dimensionality of hyperspectral data for a shorter computation time before using the 3D-CNN module.
3. We quantitatively and qualitatively evaluate the classification efficacy of the proposed 3D-CmT model on three representative hyperspectral (HS) datasets, i.e., Salinas Scene, Indian Pines, and Pavia University.

The remainder of this paper is structured as follows. The Related Work is discussed in Section 2. Section 3 introduces the proposed 3D-CmT model together with the experimental data sets used for this work. The experimental setting and the results are discussed in Section 4. The conclusion is presented in Section 5.

2 Related Work

This section gives an overview of the work done using CNN and Transformer models for hyperspectral image classification.

2.1 CNN Based

In the last decade, deep learning has produced significant technological advancements for hyperspectral image processing and analysis. For computer vision, CNNs are a very popular and widely used architecture for automatic feature extraction and learning. In recent years, CNN-based models have been extensively explored for the HSI classification. In [6], the authors have used 2D-CNN along with multilayer perceptron to learn the spatial and spectral information of pixels for the HSI classification. They demonstrated the capability of CNN on different variants of the support vector machine (SVM). The hierarchical deep spatial features have been extracted by off-the-shelf CNN architecture in [23] for HSI classification. A deep feature fusion network (DFFN) has been proposed by [24] to explore the hierarchical layers of CNN for HSI classification. In [25], a dual-path network (DPN) based HSI classification model has been proposed. Along with the wide usage of 2D-CNN, 3D-CNNs have also been utilized by various researchers to handle the spectral data of HSIs. In [7] [40], a 3D-CNN model has been proposed to jointly learn the spatial and spectral features of hyperspectral images. 3D-CNN has also been exploited with 2D-CNN by HybridSN model for HSI classification [2]. A deep pyramidal residual network spectral and spatial information of hyperspectral image classification has been proposed in [26].

2.2 Transformer Based

In addition to CNN, in recent years, various vision transformer (ViT) [8] based architectures have shown impressive performance in computer vision due to its global attention and model long-range dependencies for spatial information in the images. There has been abundant work completed to explore transformer-based models for HSI classification. A SpectralFormer model proposed by [9] used a transformer to take advantage of hyperspectral data from a sequential perspective. Two modules are devised in their model, the first module is capable of learning local sequential information of spectrally correlated hyperspectral bands and providing the group-wise spectral embeddings (GSE) and the second module is cross-layer adaptive fusion (CAF) which carries memory components from shallow layers to deep layers. In CSiT [12], a multiscale vision transformer model is proposed. They have fused the two branches of vision transformer which are individually learning the pixel-wise features at different scales. They proposed two modules namely multiscale spectral embedding (MSSE) and cross-spectral attention fusion (CSAF) module for HSI classification. The spatial-spectral transformer (SST) model [13] combines CNN, DenseTransformer, and multilayer perceptron for HSI classification.

2.3 CNN-Transformer Based

The hybrid combination of CNN with the transformer block is performing impressively in HSI classification. Hyperspectral image transformer (HiT) proposed in [4] also comprises both CNN and ViT encoder, in which they pro-

posed a spectral-adaptive 3D convolution projection (SACP) module and Conv-Permutator module to capture and learn the spectral-spatial feature information of hyperspectral images. In the FusionNet [14] model, a fusion of CNN and Transformer network for HSI classification is proposed. In [15], authors developed a convolution and transformer adaptive fusion (CTAFNet) strategy for pixel-wise classification of hyperspectral images. To capture the local high-frequency information they have used a convolution module and to handle sequential and global low-frequency they have used a transformer module. In addition to HSI data for classification, many researchers have used other modalities data sources like LiDAR, SAR, and MSI to enhance the capability of a classification model. In [3], a spectral-spatial feature tokenization transformer (SSFTT) based model is proposed. The latest work described by DBCTNet [39] also used a hybrid network of convolutional and transformer based model for hyperspectral classification, they used double branch convolutional transformer network for parallel combination convolution and self-attention instead of their serial combination. In this work, authors proposed three modules for HSI classification the first module comprises a spectral-spatial feature extraction module for low-level features, the second module uses a Gaussian weighted feature tokenizer for feature transformation and lastly the third module comprises a transformer encoder used for feature learning. As HSI is providing more information through the narrow bands, various works were done for spectral dimension.

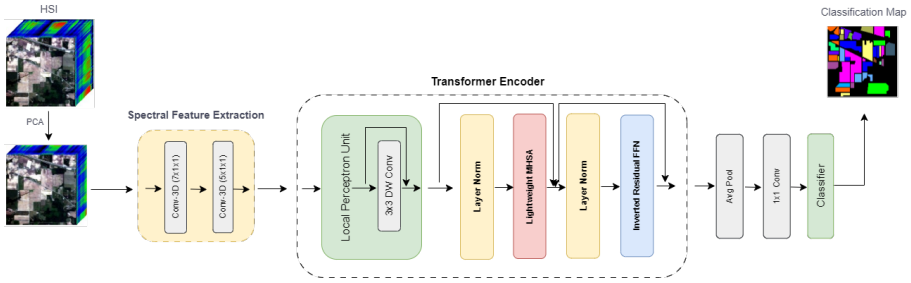
Inspired by the previous work on hyperspectral image classification, we also investigate and demonstrate the potential of a 3D convolutional layer for spectral feature extraction and transformer network grounded by the CMT model for global feature learning with self-attention mechanism.

3 Methodology

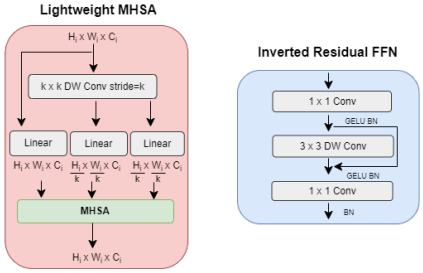
3.1 Overall Architecture

We aim to build a hybrid model using the capability of 3D convolution and transformers for the classification of HSI. Fig. 1 shows the overall framework of HSI classification using the proposed 3D-CmT model which consists of two key modules, i.e., HSI spectral feature extraction module (SFE) using the 3D-convolution and Transformer Encoder (TE) module as illustrated in Fig. 1b. As most of the transformer-based classification model splits an input image to nonoverlapping patches, however, this may ignore the intraobject relationship and representation. Therefore, to handle this limitation, we do not split the input image into nonoverlapping patches before the transformer encoder module. However, the spectral dimension of HSI is reduced using PCA transformation and 3D-Convolution for less computation overhead. Each module has been described in the following subsections.

Spectral Feature Extraction Module The Hyperspectral data denoted as $I \in \mathbb{R}^{m \times n \times c}$, where m, n, and c are the height, width, and number of channels of an input image, respectively. HS imagery comes with an abundance of



(a) Proposed 3D-CmT architecture with CMT [1] block as Transformer Encoder module.



(b) Lightweight MHSA and Inverted Residual FFN blocks.

Fig. 1: Overall framework of the proposed 3D-CmT model for the HSI classification.

spectral data in the form of large spectral channels. To reduce computational cost and memory consumption during model training, PCA transformation has been applied before the training process. In addition to that, PCA has been used to extract the most relevant band information from spectral data. Reduces c channels to b channels. So, after PCA transformation the data is expressed as $I_{\text{pca}} \in \mathbb{R}^{m \times n \times b}$, and then given as input to the model.

Then, the first component of our proposed 3D-CmT is two layers of the 3D convolutional block, which has given impressive results, especially in handling the channels of 3D images. To learn the spectral representation, 3D-Conv layers are the perfect choice as discussed in the literature review Section. 2. Hence, spectral feature extraction from the input HSI is termed a spectral feature extraction module. The kernel size of the first 3D-Conv layer is kept as 7 and for the subsequent second layer, it is chosen as 3. The overall computation of the SFE module is represented using Eq. 1,

$$SFE(I_{\text{pca}}) = Conv_2(Conv_1(I_{\text{pca}}, k_1 = 7), k_2 = 5). \tag{1}$$

Transformer Encoder Module The features extracted from the SFE module that extracts the spectral features from HSI served as input to the TE module

that learns the high-level spatial representations for input and output mapping. Our TE component is grounded on the CMT-Tiny block of the CMT network [1], which consists of a local perceptron unit (LPU) as shown in Fig. 1a, a lightweight multiheaded self-attention module (LWMHSA) and an inverted residual feedforward network (IRFFN) as presented in Fig. 1b.

LPU : In the LPU block, depthwise convolution (DWConv) is used to keep the local relation and structural information with less computational cost. It is described using Eq. 2 which is discussed in [1]: where $I \in \mathbb{R}^{h \times w \times d}$, h × w is the spatial dimension of the input at the current stage, d represents the dimension of features, and $DWConv$ indicates the depthwise convolution.

$$LPU(I) = DWConv(I) + I. \tag{2}$$

LWMHSA : To decrease the computational overhead of the original self-attention module, we use a similar light weight multi-head self-attention as mentioned in [1], where $k \times k$ depthwise convolution with stride k is used to reduce the spatial dimension of key $\mathbf{K}$ and value $\mathbf{V}$ before the attention step with a relative bias of $\mathbf{B}$ as shown in Fig. 1b. The light-weight attention function is defined as:

$$LWMHSA(Q, K, V) = softmax(\frac{QK^{'\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}} + B)V^{'}. \tag{3}$$

where query $\mathbf{Q} \in \mathbb{R}^{h \times w \times d_{\mathrm{k}}}$, key $\mathbf{K'} = DWConv(\mathbf{K}) \in \mathbb{R}^{\frac{h \times w}{k^2} \times d_{\mathrm{k}}}$ and value $\mathbf{V'} = DWConv(\mathbf{V}) \in \mathbb{R}^{\frac{h \times w}{k^2} \times d_{\mathrm{v}}}$ and bias $\mathbf{B} \in \mathbb{R}^{(h \times w) \times \frac{h \times w}{k^2}}$.

IRFFN : The inverted residual feed-forward network is similar to [1] consisting of an expansion layer and depth-wise convolution followed by a projection layer.

$$IRFFN(I) = Conv(DWConv(Conv(I)) + Conv(I)), \tag{4}$$

The 3D-CmT model can be formulated using the above modules and components:

$$I_{\mathrm{i}} = LPU(SFE(I_{\mathrm{i\text{-}1}})) \tag{5}$$

$$I^{'}_{\mathrm{i}} = LWMHSA(LN(I_{\mathrm{i}})) + I_{\mathrm{i}} \tag{6}$$

$$O_{\mathrm{i}} = IRFFN(LN(I^{'}_{\mathrm{i}})) + I^{'}_{\mathrm{i}} \tag{7}$$

where $\mathrm{I_i}$ and $\mathrm{I'_i}$ represent the output features of SFE followed by LPU and LWMHSA module for the i^{th} block, respectively. LN denotes the layer normalization [42].

To classify the pixels of the input image, the softmax function is applied to the output O_{i} from the TE module. The label with the highest probability value is the category of the pixel. The reason to combine both 3D-Convolution and Vision Transformer networks is their capability to learn the representation of

image spatial data along with its rich spectral information. The main advantage of HS imageries is their narrow spectrum rich data which tells the unique spectral signature of the target object. Therefore, initially, 3D-Conv layers extract the spectral features from HSI data. Then, after extracting and learning the spectral representation from HSI data, the incorporation of the Transformer module is used to understand the global spatial features and also to handle the long-range dependencies.

Table 1: Ground Truth classes of the **Salinas Scene**, **Indian Pines** and **University of Pavia** Datasets with their respective samples number.

	Salinas Scene		Indian Pines		University of Pavia	
Class No.	Class Name	Samples	Class Name	Samples	Class Name	Samples
1	Brocol_green_weeds_1	2009	Alfalfa	46	Asphalt	6631
2	Brocoli_green_weeds_1	3726	Corn-notill	1428	Meadows	18649
3	Fallow	1976	Corn-mintill	830	Gravel	2099
4	Fallow_rough_plow	1394	Corn	237	Trees	3064
5	Fallow_smooth	2678	Grass-pasture	483	Painted metal sheets	1345
6	Stubble	3959	Grass-trees	730	Bare Soil	5029
7	Celery	3579	Grass-pasture-mowed	28	Bitumen	1330
8	Grapes_untrained	11271	Hay-windrowed	478	Self-Blocking Bricks	3682
9	Soil_vinyard_develop	6203	Oats	20	Shadows	947
10	Corn_senesced_green	3278	Soybean-notill	972		
11	Lettuce_romaine_4wk	1068	Soybean-mintill	2455		
12	Lettuce_romaine_5wk	1927	Soybean-clean	593		
13	Lettuce_romaine_6wk	916	Wheat	205		
14	Lettuce_romaine_7wk	1070	Woods	1265		
15	Vinyard_untrained	7268	Building-Grass-Trees-Drives	386		
16	Vinyard_vertical	1807	Stone-Steel-Towers	93		

4 Experiments and Results

4.1 Dataset Description

In our work, three commonly and widely used openly available HSI datasets are selected for the experiments, including the Salinas Scene (SA), Indian Pines (IP), and Pavia University (UP) datasets.

Salinas Scene: The images in the SA dataset have a spatial size of 512 × 217 and 224 spectral bands that span the electromagnetic wavelength range of 360 to 2500 *nm*. This dataset consists of a total of 16 classes.

Indian Pines: The IP dataset contains images having 224 spectral bands with spatial size of 145 × 145 each. This dataset covers the hyperspectral imaging wavelength range from 400 to 2500 *nm*. Their ground truth is provided for 16 different classes of vegetation.

University of Pavia: Images with a spatial resolution of 610 × 340 and 103 spectral bands between 430 and 860 *nm* are included in the UP dataset. There are nine classifications of urban land cover in the ground truth.

The publicly available hyperspectral datasets[1] are downloaded for the experiments. Table. 1 represents the description of the four datasets taken in our study including the total number of classes along with the class type and total number of samples corresponding to each class.

4.2 Evaluation Metrics

To measure the classification accuracy of the proposed model, we used three commonly used classification evaluation metrics, i.e., *Average Accuracy (AA)*, *Overall Accuracy (OA)* and *Kappa Coefficient (Kappa)*. AA is expressed as the percentage of the average of classwise classification accuracies; OA is represented as a percentage of the number of precisely classified samples divided by the total test samples; and Kappa is a statistical measure between the ground samples map and classification map that gives mutual information about a strong agreement between them. Additionally, the total number of Floating Point Operations (FLOPs) and model parameters are also considered to compare the classification accuracy of the proposed model.

4.3 Implementation Details

The 3D-CmT model is implemented on a system with Intel ® Xeon ® Platinum 8180 CPU @2.50GHz, 1TB of RAM, and NVIDIA Tesla V100 GPU, 32GB of RAM. Python v3.7.4 and PyTorch [5] based environment is used to do all the experiments.

The Salinas Scene dataset is selected as an example to illustrate the 3D-CmT model. The Adam [27] optimizer has been chosen as the optimizer taking the initial learning rate value as 1e-3 and 1e-5 as weight decay. For batch training, the size of each batch is set to 32 with total training epochs of 100. The number of PCA components taken is 30 with an image size of 64×64 for training. The overall steps of the proposed 3D-CmT method are shown in Algorithm. 1

To demonstrate the efficacy of the proposed model, several CNN and Transformer networks are considered for comparative analysis: CNN-based models include 2D-CNN [38], 3D-CNN [37] with their original implementations and HybridSN [2] with same parameters excluding PCA bands which is selected as 30 for fair comparison. Transformer and hybrid model-based networks include ViT [8], where ViT-Base model configuration is chosen. SpectralFormer [9], the model is re-trained and the accuracies are similar to the results reported in the original paper for IP and UP Datasets. But, for Salinas Scene Datasets, all the model parameters have been set as per the Indian Pines dataset except the number of epochs and training samples i.e., 100 and 0.3 respectively for a fair comparison. For SSFTT [3], for a fair comparison we have taken an image with a patch size of 65×65 and the model has been trained for 100 epochs. DBCTNet [40] model is used as per its original implementation.

[1] Openly Available HSI Datasets `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`

Algorithm 1 3D-CmT Model

Input: HSI data $X \in \mathbb{R}^{m \times n \times c}$ as input and $Y \in \mathbb{R}^{m \times n}$ as ground truth with a patch size of 64 and number of PCA bands of 30.
Output: Predicted classification labels on the test datasets.

1: Set Adam (learning rate: 1e-3) as optimizer, epochs number *total_epoch* to 100 with batch-size of 32.
2: After PCA transformation, obtain the I_{pca} components.
3: Divide I_{pca} into the training dataset and validation dataset and then generate data loader for training and validation data.
4: **for** epoch = 1 to *total_epoch* **do**
5: Perform 3D convolution of SFE module.
6: Perform Transformer Encoding on TE module.
7: Use the softmax function to identify the class labels.
8: **end for**
9: Use the test dataset with the trained model to get the predicted class labels.

4.4 Classification Results

Quantitative Analysis This section presents the quantitative results of the proposed method in terms of metrics mentioned in Section. 4.2. An extensive comparative analysis is performed for various state-of-the-art models on the SA dataset in Table. 2, the UP dataset in Table. 3, and the IP dataset in Table. 4. For the SA dataset, the 3D-CmT model outperforms all other models in terms of *OA*, *AA*, and *kappa* values. However, in terms of training time, number of Params, and number of FLOPs, performs satisfactorily. For the IP dataset, in terms of *OA* and *kappa* values, the proposed model is performing second best as the HybridSN model is performing best in terms of accuracy metrics. Similarly,

Table 2: Comparative Results for **Salinas Scene Dataset** where training samples taken as 30% and patch size is 64×64. The bold and underlined text shows the best and second best performance.

Model	Training Time	#Params	#FLOPs	OA	AA	Kappa
2D-CNN[38]	3m37s	1.67M	32.80M	<u>99.99</u>	<u>99.99</u>	99.89
3D-CNN[37]	53s	995K	**24.2K**	99.34	99.75	99.26
HybridSN[2]	2h44m	51.76M	2.39G	<u>99.99</u>	<u>99.99</u>	<u>99.99</u>
ViT[8]	2h41m	125.16M	<u>967.21K</u>	99.52	99.44	99.47
SpectralFormer[9]	6m14s	378.13K	17.19M	88.41	93.25	87.15
SSFTT[3]	22m39s	<u>153.22K</u>	508.61M	99.92	99.93	99.93
DBCTNet[40]	8m38s	**30.88K**	12.81M	94.44	97.34	93.82
3D-CmT	1h8m	7.55M	120.68M	**100.0**	**100.0**	**100.0**

for the UP dataset, the proposed model is performing best among all other CNN and Transformer models in terms of all three accuracy metrics. Based on the comparative results with CNN and Transformer models, it is worth mentioning

Table 3: Comparative Results for **University of Pavia Dataset** where training samples taken as 30% and patch size is 64 × 64. The bold and underlined text shows the best and second best performance.

Model	Training Time	#Params	#FLOPs	OA	AA	Kappa
2D-CNN[38]	1m47s	1.61M	19.70M	99.95	99.94	99.94
3D-CNN[37]	39s	994K	**24.1K**	99.74	99.72	99.66
HybridSN[2]	2h10m	51.76M	2.39G	99.80	99.51	99.74
ViT[8]	1h13m	105.3M	961.84K	97.22	94.52	96.32
SpectralFormer[9]	22m8s	183.65K	4.47M	90.24	89.05	86.89
SSFTT[3]	39m25s	153.22K	508.61M	98.72	96.28	98.30
DBCTNet[40]	2m17s	**16.36K**	6.37M	98.53	97.79	98.06
3D-CmT	55m57s	7.54M	120.67M	**99.98**	**99.96**	**99.97**

that the 3D-CmT model is outperforming all the models for the SA dataset as represented in Table. 2 and UP dataset as shown in Table. 3. However, for the IP dataset, the proposed model is the second best performing as shown in Table. 4.

Table 4: Comparative Results for **Indian Pines Dataset** where training samples taken as 30% and patch size is 64 × 64. The bold and underlined text shows the best and second best performance.

Model	Training Time	#Params	#FLOPs	OA	AA	Kappa
2D-CNN[38]	33s	1.66M	32.28M	84.55	91.36	82.08
3D-CNN[37]	13s	995K	**24.2K**	98.39	97.76	98.16
HybridSN[2]	31m26s	51.76M	2.39G	**99.62**	**99.61**	**99.57**
ViT[8]	26m09s	124.37M	967.21K	89.86	84.90	88.43
SpectralFormer[9]	3m18s	355.57K	16.53M	76.31	84.48	73.31
SSFTT[3]	8m9s	153.22K	508.61M	97.10	92.09	96.69
DBCTNet[40]	20m38s	**30.3K**	12.55M	98.14	98.38	97.88
3D-CmT	12m48s	7.55M	120.68M	98.50	95.61	98.29

Qualitative Analysis This section presents the visual results of the proposed method along with the results of comparative models and their ground-truth maps. Fig. 3, Fig. 2, and Fig. 4 represent the results of classification maps generated on the SA, IP, and UP datasets, respectively. By visual comparisons, it is worth noting that 3D-CmT model-generated labels are very close and accurate to ground-truth labels compared to the other state-of-the-art models. For the SA data set, the results of 3D-CNN, ViT, SpectralFormer, and DBCTNet are not accurate for some classes, as they are misclassified. However, our proposed model can accurately predict each class, demonstrating its good performance. For the IP dataset, the visual results of HybridSN are good compared to all other models. There are misclassifications of pixels for 2D-CNN, ViT, and SpectralFormer models. However, our proposed model is comparable to HybridSN performance.

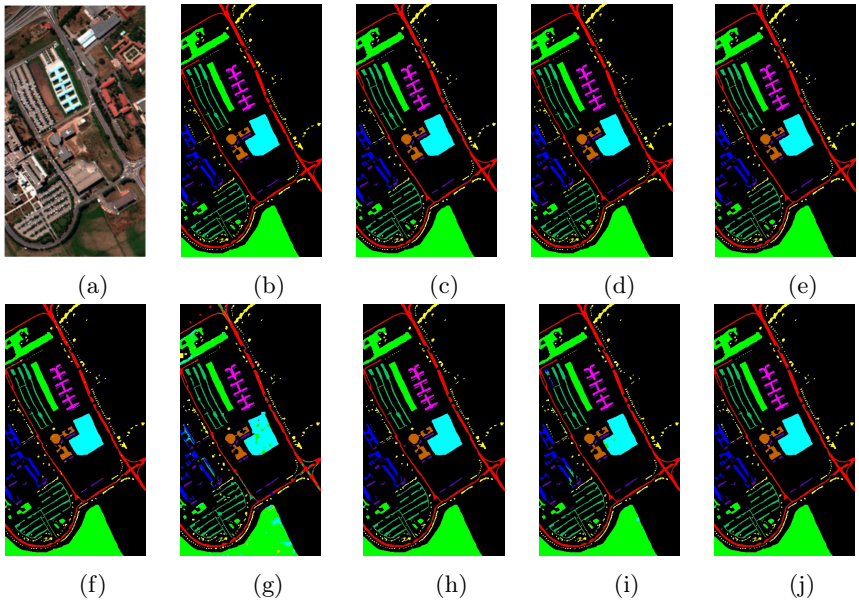


Fig. 2: Classification Results of the UP dataset. (a) Input. (b) Ground Truth. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (j) DBCTNet. (j) 3D-CmT.

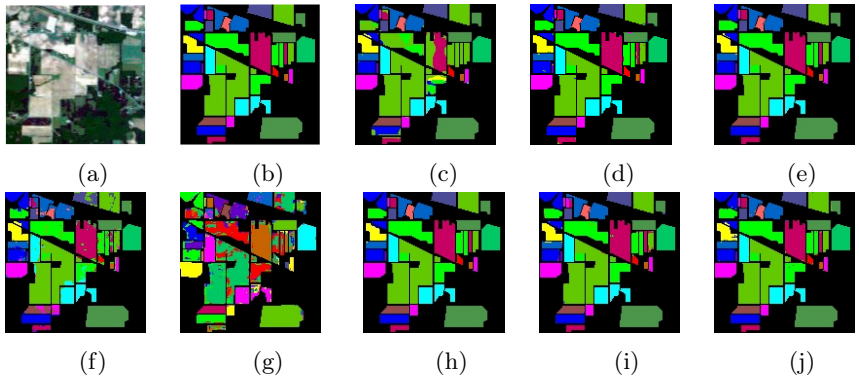


Fig. 3: Classification Results of the IP dataset. (a) Input. (b) Ground Truth. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (j) DBCTNet. (j) 3D-CmT.

For the UP dataset, there is slightly lower visual performance of the 3D-CNN and SpectralFormer model, but all other models including the proposed model can give good visual results.

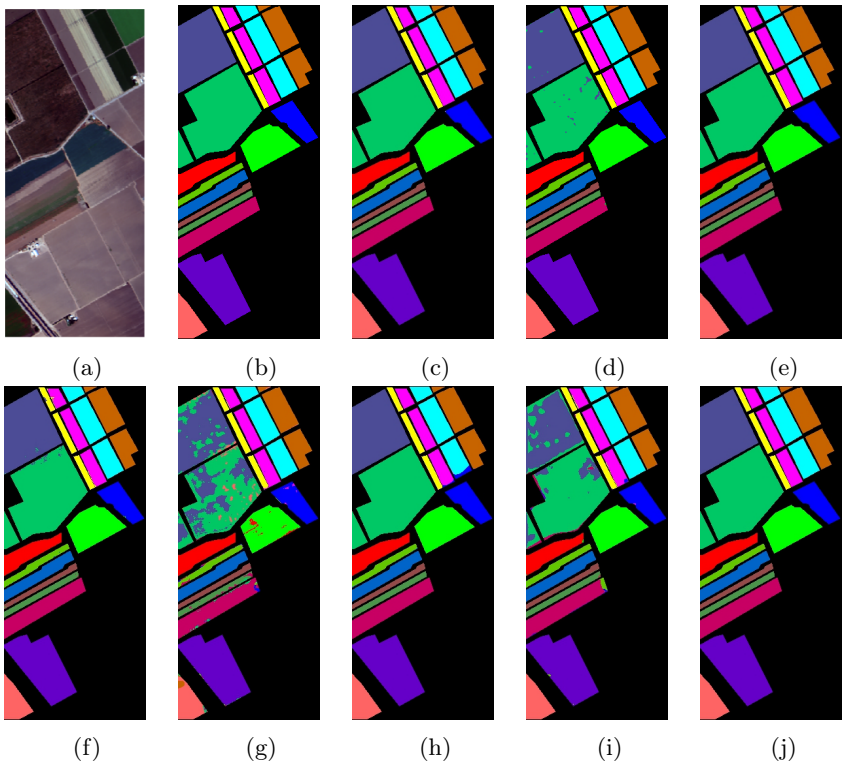


Fig. 4: Classification Results of the SA dataset. (a) Input. (b) Ground Truth. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (j) DBCTNet. (j) 3D-CmT.

4.5 Ablation Study

We analyze and evaluate the classification performance of the 3D-CmT model in terms of classification evaluation metrics as discussed in Section. 3. Extensive ablation studies on the Salinas Scene Dataset include experimentation with different combinations of PCA components, image patch size, and percentage of training samples. Although our base model is a CMT-Tiny network of [1] that consists of 4 stages of CMT blocks, therefore, we experimented with different combinations of CMT blocks in our proposed model.

Table 5: Classification performance analysis of effect of **with and without PCA** in 3D-CmT model for 30% training samples with patch size of 64 × 64 on Salinas Dataset.

PCA	#Params	#FLOPs	OA	AA	Kappa
w/o	7.63M	856.19M	99.98	99.98	99.98
10	7.54M	**36.13M**	99.99	99.98	99.98
20	7.54M	78.41M	99.99	99.99	99.99
30	**7.54M**	57.27M	**100.0**	**100.0**	**100.0**

Effect of PCA Components on Training Table. 5 represents the accuracy of the model while we trained the model with and without PCA components. Extensive experiments with different numbers of PCA components show that there is very little difference in the accuracies of all the cases, but PCA components with a value of 30 give the best accuracies in terms of all metrics. Moreover, we also experiment with the case where PCA is not applied before the training, in this case, model training time is very high and accuracies are also decreasing. So, the value of 30 is used as the optimal number of PCA components for our proposed model in all three datasets.

Effect of Patch Size and Percentage of Training Samples To check the effects of the patch size of an input image, we experimented with different patch sizes along with the model training with 30% and 10% training samples. The results of the 30% training samples are shown in Table. 6, where the patch size of 64 × 64 gives the best accuracies among all the other cases, but its FLOPs are higher than those of the other cases. Although patch sizes of 8 × 8, 16 × 16, and 32 × 32 have fewer FLOPs, their accuracies are slightly lower than the 64 × 64 case. Therefore, we used a large patch size for our experiments, that is, 64 × 64. Similarly, Table. 7 represents the experimental results for 10% training samples. Patch size of 16 × 16 is the best-performing case in terms of accuracy. But the 64 × 64 case is second best among all other cases. The results of the experiment show that a patch size of 64 × 64 is the optimal patch size for our proposed model.

Table 6: Classification performance for **30% training samples** on Salinas Scene Dataset.

Patch Size	#Params	#FLOPs	OA	AA	Kappa
8 × 8	7.47M	**1.90M**	99.93	99.92	99.93
16 × 16	**7.47M**	7.54M	99.99	99.99	99.99
32 × 32	7.48M	30.13M	99.96	99.95	99.96
64 × 64	7.54M	57.27M	**100.0**	**100.0**	**100.0**

Table 7: Classification performance for **10% training samples** on Salinas Scene Dataset.

Patch Size	#Params	#FLOPs	OA	AA	Kappa
8 × 8	7.47M	**1.90M**	99.35	99.67	99.28
16 × 16	**7.47M**	7.54M	**99.88**	**99.86**	**99.87**
32 × 32	7.48M	30.13M	99.77	99.72	99.75
64 × 64	7.55M	120.68M	99.79	99.76	99.76

Effect of CMT-Stage and 3D-Conv Layer Experiments have been conducted to select an optimal number of CMT stages and 3D-Conv layers. Table 8

Table 8: Classification performance analysis between **different CMT stage** in 3D-CmT model for 30% training samples with patch size of 64 × 64 on Salinas Dataset.

#CMT-Stage	#Params	#FLOPs	OA	AA	Kappa
4	7.96M	443.47M	100.0	100.0	100.0
3	7.72M	295.67M	100.0	100.0	100.0
2	7.60M	121.98M	99.96	99.92	99.95
1	**7.54M**	**57.27M**	**100.0**	**100.0**	**100.0**

Table 9: Classification performance analysis between **different 3D-Conv layer** in 3D-CmT model for 30% training samples with patch size of 64 × 64 on Salinas Dataset.

#3D-Conv Layer	#Params	#FLOPs	OA	AA	Kappa
1 (k_1=(7x1x1))	7.55M	60.12M	99.99	99.98	99.99
2 (k_1=(7x1x1),k_2=(5x1x1))	**7.54M**	**57.27M**	**100.0**	**100.0**	**100.0**
3 (k_1=(7x1x1),k_2=(5x1x1),k_3=(3x1x1))	7.54M	216.62M	99.99	99.99	99.99

shows the experimental results for different numbers of CMT-Stages used. No major improvement is observed in the performance of the model if we increase the number of stages. However, CMT-Stage with only one block is performing best in terms of FLOPs and accuracy. Therefore, we have considered only one CMT block in the proposed model. The experimental results of different sets of 3D-Conv layers with one CMT block are shown in Table .9, where two 3D-Conv layers are the best-performing case among all accuracy metrics.

5 Conclusion

A 3D-CNN and Vision Transformer hybrid network named 3D-CmT has been proposed for hyperspectral image classification. It uses 3D-CNN for local feature learning along with the narrow spectral information, and Vision Transformer for global feature representation. Experiments have been carried out for different patch sizes of images, along with the different percentages of training samples. The proposed model is comparable with other comparative models in terms of both quantitative and visual results. From this study, we can conclude that to handle spatial and spectral HSI information, a hybrid network of 3D-CNN and ViT can be effectively used for the classification of hyperspectral images. For future work, we would like to add other image modalities such as Synthetic Aperture Radar (SAR), Light Detection and Ranging (LiDAR), and Digital Elevation Model (DEM) data to further verify the efficacy of the model.

References

1. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y. & Xu, C. Cmt: Convolutional neural networks meet vision transformers. Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition. pp. 12175-12185 (2022)
2. Roy, S., Krishna, G., Dubey, S. & Chaudhuri, B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. IEEE Geoscience And Remote Sensing Letters. 17, 277-281 (2019)
3. Sun, L., Zhao, G., Zheng, Y. & Wu, Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Transactions On Geoscience And Remote Sensing. 60 pp. 1-14 (2022)
4. Yang, X., Cao, W., Lu, Y. & Zhou, Y. Hyperspectral image transformer classification networks. IEEE Transactions On Geoscience And Remote Sensing. 60 pp. 1-15 (2022)
5. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances In Neural Information Processing Systems 32. pp. 8024-8035 (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
6. Makantasis, K., Karantzalos, K., Doulamis, A. & Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. 2015 IEEE International Geoscience And Remote Sensing Symposium (IGARSS). pp. 4959-4962 (2015)
7. Hamida, A., Benoit, A., Lambert, P. & Amar, C. 3-D deep learning approach for remote sensing image classification. IEEE Transactions On Geoscience And Remote Sensing. 56, 4420-4434 (2018)
8. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv Preprint ArXiv:2010.11929. (2020)
9. Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A. & Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Transactions On Geoscience And Remote Sensing. 60 pp. 1-15 (2021)
10. Qiao, X., Roy, S. & Huang, W. Multiscale Neighborhood Attention Transformer With Optimized Spatial Pattern for Hyperspectral Image Classification. IEEE Transactions On Geoscience And Remote Sensing. 61 pp. 1-15 (2023)
11. Roy, S., Deria, A., Hong, D., Rasti, B., Plaza, A. & Chanussot, J. Multimodal fusion transformer for remote sensing image classification. IEEE Transactions On Geoscience And Remote Sensing. (2023)
12. He, W., Huang, W., Liao, S., Xu, Z. & Yan, J. Csit: A multiscale vision transformer for hyperspectral image classification. IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing. 15 pp. 9266-9277 (2022)
13. He, X., Chen, Y. & Lin, Z. Spatial-spectral transformer for hyperspectral image classification. Remote Sensing. 13, 498 (2021)
14. Yang, L., Yang, Y., Yang, J., Zhao, N., Wu, L., Wang, L. & Wang, T. FusionNet: a convolution–transformer fusion network for hyperspectral image classification. Remote Sensing. 14, 4066 (2022)
15. Li, J., Xing, H., Ao, Z., Wang, H., Liu, W. & Zhang, A. Convolution-Transformer Adaptive Fusion Network for Hyperspectral Image Classification. Applied Sciences. 13, 492 (2022)

16. Yang, H., Yu, H., Zheng, K., Hu, J., Tao, T. & Zhang, Q. Hyperspectral Image Classification Based on Interactive Transformer and CNN with Multilevel Feature Fusion Network. IEEE Geoscience And Remote Sensing Letters. (2023)
17. Abdi, H. & Williams, L. Principal component analysis. WIREs Computational Statistics. 2, 433-459 (2010), https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.101
18. Wang, J., Zhang, L., Tong, Q. & Sun, X. The Spectral Crust project—Research on new mineral exploration technology. 2012 4th Workshop On Hyperspectral Image And Signal Processing: Evolution In Remote Sensing (WHISPERS). pp. 1-4 (2012)
19. Gevaert, C., Suomalainen, J., Tang, J. & Kooistra, L. Generation of spectral–temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications. IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing. 8, 3140-3146 (2015)
20. Fong, A., Shu, G. & McDonogh, B. Farm to table: Applications for new hyperspectral imaging technologies in precision agriculture, food quality and safety. CLEO: Applications And Technology. pp. AW3K-2 (2020)
21. Noor, S., Michael, K., Marshall, S., Ren, J., Tschannerl, J. & Kao, F. The properties of the cornea based on hyperspectral imaging: Optical biomedical engineering perspective. 2016 International Conference On Systems, Signals And Image Processing (IWSSIP). pp. 1-4 (2016)
22. Ardouin, J., Lévesque, J. & Rea, T. A demonstration of hyperspectral image exploitation for military applications. 2007 10th International Conference On Information Fusion. pp. 1-8 (2007)
23. Cheng, G., Li, Z., Han, J., Yao, X. & Guo, L. Exploring hierarchical convolutional features for hyperspectral image classification. IEEE Transactions On Geoscience And Remote Sensing. 56, 6712-6722 (2018)
24. Song, W., Li, S., Fang, L. & Lu, T. Hyperspectral image classification with deep feature fusion network. IEEE Transactions On Geoscience And Remote Sensing. 56, 3173-3184 (2018)
25. Kang, X., Zhuo, B. & Duan, P. Dual-path network-based hyperspectral image classification. IEEE Geoscience And Remote Sensing Letters. 16, 447-451 (2018)
26. Paoletti, M., Haut, J., Fernandez-Beltran, R., Plaza, J., Plaza, A. & Pla, F. Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. IEEE Transactions On Geoscience And Remote Sensing. 57, 740-754 (2018)
27. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. ArXiv Preprint ArXiv:1412.6980. (2014)
28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature. 521, 436-444 (2015)
29. Signoroni, A., Savardi, M., Baronio, A. & Benini, S. Deep learning meets hyperspectral image analysis: A multidisciplinary review. Journal Of Imaging. 5, 52 (2019)
30. Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P. & Benediktsson, J. Deep learning for hyperspectral image classification: An overview. IEEE Transactions On Geoscience And Remote Sensing. 57, 6690-6709 (2019)
31. Jia, S., Jiang, S., Lin, Z., Li, N., Xu, M. & Yu, S. A survey: Deep learning for hyperspectral image classification with few labeled samples. Neurocomputing. 448 pp. 179-204 (2021)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. Advances In Neural Information Processing Systems. 30 (2017)
33. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. Backpropagation applied to handwritten zip code recognition. Neural Computation. 1, 541-551 (1989)

34. Petersson, H., Gustafsson, D. & Bergstrom, D. Hyperspectral image analysis using deep learning—A review. 2016 Sixth International Conference On Image Processing Theory, Tools And Applications (IPTA). pp. 1-6 (2016)
35. Paoletti, M., Haut, J., Plaza, J. & Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. ISPRS Journal Of Photogrammetry And Remote Sensing. 158 pp. 279-317 (2019)
36. Hu, W., Huang, Y., Wei, L., Zhang, F. & Li, H. Deep convolutional neural networks for hyperspectral image classification. Journal Of Sensors. 2015 pp. 1-12 (2015)
37. He, M., Li, B. & Chen, H. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. 2017 IEEE International Conference On Image Processing (ICIP). pp. 3904-3908 (2017)
38. Liu, B., Yu, X., Zhang, P., Tan, X., Yu, A. & Xue, Z. A semi-supervised convolutional neural network for hyperspectral image classification. Remote Sensing Letters. 8, 839-848 (2017)
39. Xu, R., Dong, X., Li, W., Peng, J., Sun, W. & Xu, Y. DBCTNet: Double Branch Convolution-Transformer Network for Hyperspectral Image Classification. IEEE Transactions On Geoscience And Remote Sensing. (2024)
40. Ahmad, M., Khan, A., Mazzara, M., Distefano, S., Ali, M. & Sarfraz, M. A fast and compact 3-D CNN for hyperspectral image classification. IEEE Geoscience And Remote Sensing Letters. 19 pp. 1-5 (2020)
41. Ren, Q., Tu, B., Liao, S. & Chen, S. Hyperspectral image classification with iformer network feature extraction. Remote Sensing. 14, 4866 (2022)
42. Ba, J., Kiros, J. & Hinton, G. Layer normalization. ArXiv Preprint ArXiv:1607.06450. (2016)
43. Moharram, M. & Sundaram, D. Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions. Neurocomputing. 536 pp. 90-113 (2023)