A Center-Masked Convolutional Transformer for Hyperspectral Image Classification

Yifan Wang[1], Shuguo Jiang[1], Meng Xu[1], Shuyu Zhang[1] and Sen Jia[1,*]

[1]College of Computer Science and Software Engineering, Shenzhen University, China

Abstract

Hyperspectral images (HSIs) have a wide field of view and rich spectral information, where each pixel represents a small area of the earth's surface. The pixel-level classification task of HSI has become one of the research hotspots in hyperspectral image processing and analysis. More and more deep learning methods have been proposed in recent years, among which convolutional neural network (CNN) is the most influential. However, it is difficult for CNN-based models to obtain the global receptive field in HSI classification task. Besides, most of the self-supervised training methods are based on sample reconstruction, and it is not easy to achieve effective use of unlabeled samples. In this paper, we propose a novel convolutional embedding module, combined with the Transformer blocks, which successfully improves the context-awareness while retaining the local feature extraction capability. Moreover, a new self-supervised task is designed to make more efficient use of unlabeled data. Our proposed pre-training task only masks the central token and reconstructs the central pixel from a learnable vector. It allows the model to capture the patterns between the central object and surrounding objects without labels.

Keywords

Deep learning, Masked autoencoder, Transformer, Hyperspectral image classification.

1. Introduction

Hyperspectral images are generally composed of dozens to hundreds of bands and have the characteristics of low spatial resolution and high spectral resolution. The spectral information provides the possibility to distinguish the corresponding land covers, which has spawned various research fields. Among them, pixel-level hyperspectral image classification is the most concerned one in the community. Its main task is to assign a class label to each pixel, somewhat like semantic segmentation in the computer vision (CV) field. Different from RGB image, hyperspectral image is high-dimensional data. In order to avoid the curse of dimensionality, principal component analysis (PCA) [1] and independent component analysis [2] are widely used for redundancy elimination.

So far, many hyperspectral image classification methods have been proposed, but deep learning methods have taken the lead. According to the different techniques used, it can be divided into traditional methods and deep learning-based methods. In early research, people mostly selected a single pixel and all its spectral information as the training sample and rely on the traditional classifiers, such as logistic regression [3], decision tree [4], random forest [5], and support vector machine (SVM) [6], to classify the ground objects through spectral information. However, the imaging distance of HSI is far away, and there are many interference factors in this process, so that the spectral curve of different surface objects is not always easy to distinguish. This creates difficulties for these methods to achieve good performance in complex scenes. In recent years, deep learning methods have gradually become popular, in which CNN-based methods are dominant. Hu *et al.* [7] made a preliminary attempt that several 1-D convolutional layers are stacked to extract local spectral information, and many classical data augmentation methods in CV have been introduced. Roy *et al.* [8] combined 3D-CNN and 2D-CNN to achieve hierarchical feature learning. In addition, other neural networks have also achieved good performance. Zhou *et al.* [9] designed a two-branch Long Short-Term Memory network (LSTM) to extract spectral information and spatial information respectively. He *et al.* [10] proposed a pure multilayer perceptron (MLP) network, proving that the MLP network still has potential. Hong *et al.* [11] designed a mini-batch graph neural network. It is worth mentioning that the recently prevalent Transformer model has also been introduced. Hu *et al.* [12] used 1-D convolution as an embedding layer combined with Transformer Block. Hong *et al.* [13] analyzed the difference between Transformer and other classical neural networks in detail and proposed a ViT-based SpectralFormer for spectral information learning. Zhong *et al.* [14] proposed a spatial–spectral Transformer network and a model structure search framework. Dang *et al.* [15] combined spectral-spatial attention module with densely connected Transformer blocks. Besides, self-attention

CDCEO 2022: 2nd Workshop on Complex Data Challenges in Earth Observation, July 25, 2022, Vienna, Austria

*Corresponding author.

2070276050@email.szu.edu.cn (Y. Wang); shuguoj@foxmail.com (S. Jiang); m.xu@szu.edu.cn (M. Xu); shuyu-zhang@szu.edu.cn (S. Zhang); senjia@szu.edu.cn (S. Jia)

network is also used to address adversarial attacks that may be encountered in hyperspectral classification tasks [16].

However, limited by the size of the receptive field, it is difficult for CNN models to capture the global relationship. Meanwhile, deep learning models are data-driven which means that more labeled data leads to better model performance. But, obtaining such a large number of labeled samples in practical applications is expensive. How to effectively use unlabeled data has become an urgent need. The self-supervised pre-training method in HSI classification task is still stuck in the autoencoder-based sample reconstruction [17]. This article proposes a band-grouping-based 3D convolutional Transformer (BG3DCT) and a new self-supervised task for model pre-training. The main contributions are listed as follows:

- A novel band-grouping-based 3D convolutional Transformer is designed for HSI classification. We replace the commonly used linear embedding module with a well-designed 3D convolutional embedding module, combined with the spectral segmentation strategy, to achieve efficient spatial-spectral feature embedding in each sub-band.
- According to the characteristics of hyperspectral data, a new pre-training task is proposed. In the process of masking and reconstructing the center pixel, the model's ability to capture the relationship between the center pixel and surrounding pixels is improved. Compared to the overall sample reconstruction task, center-masked pre-training task is more efficient for the representation of center area in the pre-training stage.
- A series of comparative experiments and ablation experiments demonstrate the effectiveness of our proposed pre-training method and BG3DCT network. In particular, our proposed pre-training method can alleviate the instability of results caused by random sampling in the limited training samples scenario.

The rest of this paper is organized as follows. Our proposed method will be introduced in detail in section II. The descriptions of the comparative experiments and result analysis will be provided in section III, and section IV presents the conclusions.

2. Methodology

2.1. BG3DCT Network

The BG3DCT network has three parts, the band-grouping-based 3D convolutional embedding (BG3DCE) module, Transformer encoder, and MLP head. The specific design is as follows.

Table 1
The Detail Description of BG3DCE Module

Layer	Kernel Size	Output Shape	Groups	Param
Input	-	(8, 13, 13, 10)	-	-
Conv3D-1	(3, 3, 3)	(32, 11, 11, 8)	8	896
BatchNorm3d-1	-	-	-	64
Conv3D-2	(3, 3, 3)	(32, 9, 9, 6)	8	3,488
BatchNorm3d-2	-	-	-	64
Reshape	-	(192, 9, 9)	-	
Conv2D	(1,1)	(128, 9, 9)	8	3,200
Reshape	-	(81, 128)	-	-

2.1.1. BG3DCE Module

Considering the differences between RGB images and HSIs, the ViT network is not well compatible with hyperspectral image. When the training sample is limited, the linear embedding module cannot sufficiently characterize the spatial-spectral features. Meanwhile, CNN-based module is more adaptable to this situation, while also being able to capture the local texture information. So we design a band-grouping-based 3D convolutional embedding module for HSI embedding. Firstly, we perform PCA processing on the input samples and employ a spectral partition strategy to divide the spectra into several sub-bands of equal length. Because the spectral curves of objects often have local differences, the extraction of 3D-CNN on sub-band is more efficient than full-band. Then, paralleled 3D convolution extraction are performed on each sub-band twice, and the 3D batch normalization operation is followed to unify the feature deviations generated from each sub-band. Finally, we concatenate the features to maintain the relative positional relationship between sub-bands and a lightweight 2D-CNN is used for feature fusion and compression. In particular, paralleled 3D convolution operation have a simple implementation, and the convolution function includes a grouped convolution option. The detailed description of BG3DCE module is listed in table 1.

2.1.2. Transformer Encoder

The context-aware ability of CNN often needs to make the model go deeper, but HSI data is limited, so it is difficult for us to stack modules as simply as the model in CV task. In contrast, The multi-head attention module can make up for the shortcomings of CNN here and effectively model the relationship between ground objects. So, the combination of CNN and Transformer is complementary and powerful. A standard Transformer mainly comprises position encoding, multi-head attention, and feedforward layers. Since the convolution features contains position information, the positional encoding is not

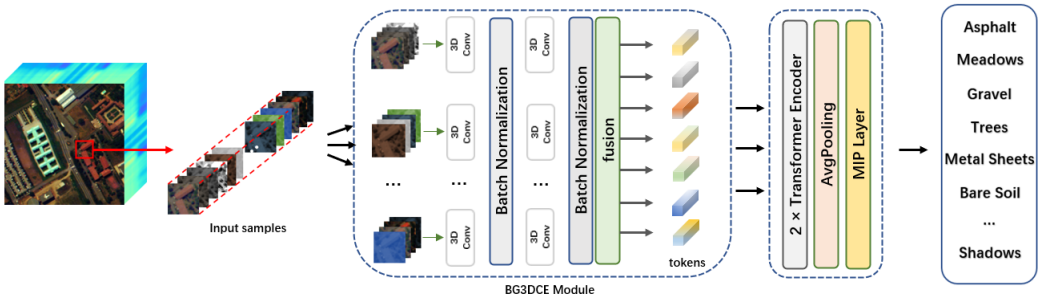

Figure 1: Architecture of the proposed band-grouping-based 3D convolutional Transformer for HSI classification.

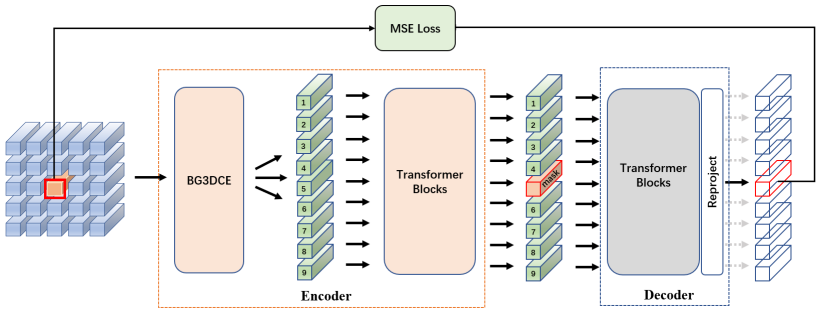

Figure 2: Architecture of the proposed center-mask pre-training task.

used here, and the embedded spatial-spectral features are directly input into the Transformer blocks. Finally, we add an average pooling layer to achieve the global representation and get the classification results through a MLP layer.

2.2. Center-mask Pre-training Task

Today, most hyperspectral image classification methods are patch-based. The model's input not only is the spectral curve of the centre pixel but also contains its neighbour region, which is generally a square area and makes the input more distinctive. Inspired by the form of the training sample, we propose the center-masked pre-training task, which is similar to MAE [18] but our method is easier to implement.

The flowchart of our proposed pre-training task is shown in Fig. 2. The Encoder is a BG3DCT network but removes the average pooling layer and MLP layer. The decoder consists of two layers of standard Transformer encoders, which are only used in the pre-training stage. Given an input sample $\mathbf{X}$ and center pixel vector v_c, the latent representation of the input sample is $\mathbf{E}$ (embedded by the BG3DCE module). Unlike self-supervised pre-training in the CV field, RGB images cannot directly find areas that need to be focused on, but the neighborhood areas of HSI samples serve for the central pixel. So, our masking target can select the essential part in the training sample, namely the center pixel. Therefore, we replace the token in the middle of the sequence $\mathbf{E}$ with a learnable vector. Then, the masked sequence is input into the decoder, and the pixel-level reconstruction is performed by a MLP head to obtain the reconstruction result $\hat{v_c}$ of the center pixel. The target of the center-masked pre-training task is to reconstruct the centre pixel as efficiently as possible so that the encoder can better learn the relationship between the centre pixel and the neighbour pixels without labels. The reconstruction target can be formulated as:

$$T(v_c, \hat{v_c}) = \min |v_c - \hat{v_c}|_2 \tag{1}$$

where T is the similarity function. In the deep learning framework, function T is equivalent to the mean squared error (MSE) loss function.

3. Experiment

To fully evaluate our proposed pre-training method and BG3DCT network, we conduct comparative and ablation experiments on two public datasets, Salinas and Yellow River Estuary (YRE). The detailed information and the

Table 2
Number of Training and Testing Samples on The Salinas Dataset

Class	Class Name	Training	Testing
1	Brocoli green weeds 1	5	2004
2	Brocoli green weeds 2	5	3721
3	Fallow	5	1971
4	Fallow rough plow	5	1389
5	Fallow smooth	5	2673
6	Stubble	5	3954
7	Celery	5	3574
8	Grapes untrained	5	11266
9	Soil vinyard develop	5	6198
10	Corn senesced green weeds	5	3273
11	Lettuce romaine 4wk	5	1063
12	Lettuce romaine 5wk	5	1922
13	Lettuce romaine 6wk	5	911
14	Lettuce romaine 7wk	5	1065
15	Vinyard untrained	5	7263
16	Vinyard vertical trellis	5	1802
	Total	80	54129

Table 3
Number of Training and Testing Samples on The YRE Dataset

Class	Class Name	Training	Testing
1	Building	10	523
2	River	10	5366
3	Salt Marsh	10	4985
4	Shallow Sea	10	17540
5	Deep Sea	10	18667
6	Intertidal Saltwater Marsh	10	2333
7	Tidal Flat	10	1782
8	Pond	10	1777
9	Sorghum	10	636
10	Corn	10	1499
11	Lotus Root	10	2709
12	Aquaculture	10	8009
13	Rice	10	5498
14	Tamarix Chinensis	10	1210
15	Freshwater Herbaceous Marsh	10	1407
16	Suaeda Salsa	10	864
17	Spartina Alterniflora	10	570
18	Reed	10	1960
19	Floodplain	10	337
20	Locus	10	65
	Total	200	77737

partition of the training set and testing set are shown in the table 2 and table 3, respectively. We use three metrics to evaluate the classification results, overall accuracy (OA), classwise average accuracy (AA), and kappa coefficient (κ). All the experiments are conducted on a computer with an Intel Xeon Platinum 8260 CPU, 64-GB RAM and an NVIDIA Tesla P100-16GB GPU. The model structure and parameter settings of the comparison methods comply with open source codes or corresponding papers. For our proposed model, the patchsize is set to 13, the spectral dimension is 80 after PCA, and the number of sub-bands is 10. The embedding size of each token is set to 128. The learning rate is set to 0.001, and Adam is adopt as the gradient descent optimizer. Meanwhile, all the experiments are repeated ten times to smooth out errors caused by random sampling. The setting of the center-masked pre-training is the same.

3.1. Datasets Description

3.1.1. Salinas Dataset

The salinas dataset, collected by the AVIRS sensor in the Salinas Valley, USA, has an image size of 512 × 217 and a spatial resolution of 3.7 meters. After noise band removal, 204 bands are remained. There are 16 kinds of ground objects in the dataset, with 56,975 samples that can be used for pixel-level classification.

3.1.2. YRE Dataset

YRE dataset is a large scene dataset captured by the Gaofen-5 satellite in the yellow river estuary region of Shandong Province, China. Its size is 1400 × 1400, and the spatial resolution of each pixel is 30 meters, leaving 180 bands after removing noise bands. The surface objects are mainly wetland vegetation, there are 20 kinds of objects, and the total number of labeled samples is 77,937.

3.2. Comparative Experiment

To demonstrate the superiority of our proposed method, we select five state-of-the-art methods on two public datasets, Salinas and YRE, including four CNN-based methods and one classical Transformer network. They are CNNHSI [19], FC3D [20], HybridSN [21], TwoCNN [22], and Vision Transformer (ViT) [23]. Among them, several methods based on 2D-CNN are distinguished in the size of the convolution kernel and the structure design. CNNHSI stacks several 2-D Convolution layers with 1×1 kernel size. TwoCNN is a dual-branch CNN with a 2D-CNN and a 1D-CNN to extract spatial information and spectral information, respectively. FC3D is a pure 3D-CNN network, and HybridSN uses 3D convolution and 2D convolution successively for hierarchical feature extraction. ViT divides the input samples into equal-sized patches, obtains the embedded tokens through linear embedding module, and then inputs them into the Transformer encoder.

The results of the comparative experiments are shown in table 4 and table 5. Our method has obtained obvious advantages and achieved the best or second-best results in each class, reflecting our approach's superiority and robustness. Under the setting of training with limited samples, CNNHSI achieves excellent classification results due to its lightweight network structure. Limited by the large model size, HybridSN, FC3D, and TwoCNN fail to

Table 4
Classification Accuracy (%) and Kappa Measure for The Salinas Dataset

Class	ViT	HybridSN	FC3D	CNNHSI	TwoCNN	Ours
1	97.83	99.93	99.95	93.64	93.44	**99.98**
2	97.21	98.31	92.92	79.30	89.86	**99.92**
3	82.83	97.04	97.84	84.62	90.63	**99.99**
4	92.91	93.18	96.66	**99.40**	97.45	99.11
5	85.29	**98.08**	86.54	77.31	95.23	97.59
6	98.58	98.46	94.08	98.45	99.74	**99.95**
7	98.01	99.76	99.78	99.13	**100.00**	**100.00**
8	69.92	59.81	53.82	65.56	**74.59**	63.87
9	97.38	96.84	95.46	94.70	99.67	**99.77**
10	82.71	93.33	90.32	46.59	**95.17**	95.08
11	90.61	97.65	89.21	90.40	95.71	**99.96**
12	98.10	94.88	91.28	97.85	99.20	**99.43**
13	90.71	84.63	87.82	99.19	**99.28**	97.08
14	96.15	98.97	92.99	92.94	96.39	**99.31**
15	62.93	71.60	65.73	40.12	66.80	**83.49**
16	94.19	79.97	88.53	82.69	96.38	**99.08**
OA (%)	84.68	85.24	81.65	76.41	87.99	**89.66**
AA (%)	89.71	91.40	88.93	83.87	93.10	**95.85**
κ	83.01	83.70	79.78	73.74	86.65	**88.54**

Table 5
Classification Accuracy (%) and Kappa Measure for The YRE Dataset

Class	ViT	HybridSN	FC3D	CNNHSI	TwoCNN	Ours
1	49.78	82.35	78.84	**90.66**	70.48	84.23
2	95.45	**100.00**	99.93	98.83	99.96	97.36
3	55.45	61.34	72.20	74.55	**85.93**	78.15
4	78.22	71.04	72.78	73.08	90.44	**92.36**
5	85.89	76.74	90.49	84.31	97.59	**99.55**
6	79.48	81.77	83.37	81.59	82.65	**85.85**
7	56.38	51.53	57.30	59.95	**63.63**	60.30
8	73.16	73.94	73.68	**78.55**	60.17	73.23
9	75.10	85.90	86.11	86.57	82.70	**90.24**
10	57.06	70.82	62.15	88.20	72.22	**88.49**
11	63.65	82.55	83.38	88.73	75.71	**90.84**
12	72.82	76.36	**79.69**	77.62	73.76	76.96
13	71.27	87.47	84.49	**92.63**	87.79	89.78
14	67.11	75.32	76.50	87.81	**88.63**	79.28
15	59.32	64.65	74.65	82.00	**96.32**	71.77
16	74.27	93.02	89.89	92.66	92.20	**95.08**
17	81.29	93.98	89.47	**95.08**	93.72	94.40
18	44.74	58.11	62.74	65.86	67.59	**70.65**
19	60.53	82.89	68.25	**88.72**	68.84	71.60
20	87.69	91.28	71.79	85.84	64.00	**92.15**
OA (%)	75.58	76.14	80.84	84.61	87.45	**89.01**
AA (%)	69.43	78.05	77.88	83.26	80.72	**84.11**
κ	71.87	72.96	78.09	82.30	85.48	**87.29**

obtain superior classification results. The performance of the ViT model is not stable. When the distribution of ground objects in the dataset is more complex, the linear embedding module drags down the model performance. Therefore, this proves the necessity of a well-designed embedding layer for the Transformer network in HSI classification. In addition, all the methods cannot discriminate the Vinyard untrained class well, which may be caused by the large variability of this land cover. It is a problem we need to solve in the future. The classification results of each method on the YRE dataset are similar to the Salinas dataset. The YRE dataset is a large scene dataset so that the classification task is more complicated. Hence, the classification performance of each method is slightly lower than that of the Salinas dataset. It is worth mentioning that TwoCNN achieves good classification results, which may benefit from its spectral feature extraction branch.

Table 6
Ablation Study Results Toward The Center-Masked Pretraining Pretask on The salinas and YRE Datasets

Case	YRE			Salinas		
	OA	AA	κ	OA	AA	κ
BG3DCT w/ pretrain	**89.01**	**84.11**	**87.30**	**89.99**	**95.40**	**88.87**
BG3DCT w/o pretrain	86.24	82.19	84.15	88.60	85.42	86.46

3.3. Ablation Study

In this section, we only conduct ablation experiments on our proposed pre-training task, considering that the experimental results compared with the ViT model can intuitively demonstrate the effectiveness of our proposed BG3DCT module. The results are shown in table 6 that the OA of the model finetuned on pre-trained model outperforms the model without pre-training by 2.77% and 1.33%, on the Salinas and YRE datasets, respectively. This undoubtedly proves the superiority and robustness of our pre-training task.

4. Conclusion

In this article, we creatively propose a band-grouping-based convolutional embedding module to extract spatial-spectral information in each sub-bands. The Transformer module is used to model the global relationship between surface objects. Additionally, for effective use of unlabeled data, we design a new unsupervised pre-training task for hyperspectral classification. Through the mask and reconstruction process of the token generated from the central area, the model can initialize the backbone network without labeled data and provide a more stable model performance. To fully evaluate our proposed methods, we conducted a series of comparative experiments and ablation experiments on two public datasets, Salinas and YRE. The experimental results prove the effectiveness and superiority of our method.

5. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 41971300, and Grant 61901278; in part by the Key Project of Department of Education of Guangdong Province under Grant 2020ZDZX3045; in part by the Guangdong Basic and Applied Basic Research Foundation under

Grant 2022A1515011290; in part by the Natural Science Foundation of Guangdong Province under Grant 2021A1515011413; in part by Shenzhen Scientific Research and Development Funding Program under Grant 20200803152531004.

References

[1] M. D. Farrell, R. M. Mersereau, On the impact of pca dimension reduction for hyperspectral detection of difficult targets, IEEE Geoscience and Remote Sensing Letters 2 (2005) 192–195.

[2] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, J. A. Benediktsson, On the decomposition of mars hyperspectral data by ica and bayesian positive source separation, Neurocomputing 71 (2008) 2194–2208.

[3] Y. Qian, M. Ye, J. Zhou, Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features, IEEE Transactions on Geoscience and Remote Sensing 51 (2012) 2276–2291.

[4] S. Kuching, The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis, Journal of Computer Science 3 (2007) 419–423.

[5] J. Xia, P. Ghamisi, N. Yokoya, A. Iwasaki, Random forest ensembles and extended multiextinction profiles for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 56 (2017) 202–216.

[6] M. Chi, R. Feng, L. Bruzzone, Classification of hyperspectral remote-sensing data with primal svm for small-sized training dataset problem, Advances in space research 41 (2008) 1793–1799.

[7] W. Hu, Y. Huang, L. Wei, F. Zhang, H. Li, Deep convolutional neural networks for hyperspectral image classification, Journal of Sensors 2015 (2015).

[8] S. K. Roy, G. Krishna, S. R. Dubey, B. B. Chaudhuri, Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 17 (2019) 277–281.

[9] F. Zhou, R. Hang, Q. Liu, X. Yuan, Hyperspectral image classification using spectral-spatial lstms, Neurocomputing 328 (2019) 39–47.

[10] X. He, Y. Chen, Modifications of the multi-layer perceptron for hyperspectral image classification, Remote Sensing 13 (2021) 3547.

[11] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, J. Chanussot, Graph convolutional networks for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 59 (2020) 5966–5978.

[12] X. Hu, W. Yang, H. Wen, Y. Liu, Y. Peng, A lightweight 1-d convolution augmented transformer with metric learning for hyperspectral image classification, Sensors 21 (2021) 1751.

[13] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, J. Chanussot, Spectralformer: Rethinking hyperspectral image classification with transformers, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15. doi:10.1109/TGRS.2021.3130716.

[14] Z. Zhong, Y. Li, L. Ma, J. Li, W.-S. Zheng, Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework, IEEE Transactions on Geoscience and Remote Sensing (2021).

[15] L. Dang, L. Weng, W. Dong, S. Li, Y. Hou, Spectral-spatial attention transformer with dense connection for hyperspectral image classification, Computational Intelligence and Neuroscience 2022 (2022).

[16] Y. Xu, B. Du, L. Zhang, Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification, IEEE Transactions on Image Processing 30 (2021) 8671–8685. doi:10.1109/TIP.2021.3118977.

[17] Y. Chen, Z. Lin, X. Zhao, G. Wang, Y. Gu, Deep learning-based classification of hyperspectral data, IEEE Journal of Selected topics in applied earth observations and remote sensing 7 (2014) 2094–2107.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, arXiv preprint arXiv:2111.06377 (2021).

[19] S. Yu, S. Jia, C. Xu, Convolutional neural networks for hyperspectral image classification, Neurocomputing 219 (2017) 88–98.

[20] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, M. S. Sarfraz, A fast and compact 3-d cnn for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters (2020).

[21] S. K. Roy, G. Krishna, S. R. Dubey, B. B. Chaudhuri, Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 17 (2020) 277–281.

[22] J. Yang, Y.-Q. Zhao, J. C.-W. Chan, Learning and transferring deep joint spectral–spatial features for hyperspectral classification, IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 4729–4742.

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