Pushing Boundaries in Hyperspectral Image Classification: A Comparative Analysis of CTMixer, SF-SMF, and MAEST

Ying-Tzu Chen [1], Hsuan Ren.[2*]

[1] Center for Space and Remote Sensing Research: Student, National Central University, Taiwan

[2] Center for Space and Remote Sensing Research: Associate Professor, National Central University, Taiwan

* hren@csrsr.ncu.edu.tw (*Corresponding author's email only)

1. Introduction

With advancements in remote sensing (RS) and deep learning (DL), interest in classifying materials on and below the Earth's surface has grown significantly. Hyperspectral (HS) images, which are data cubes containing spatial-spectral information, capture data across various electromagnetic wavelengths. Recent transformer-based architectures for classifying these images have achieved notable accuracy. However, variations in datasets and parameters, such as the number of layers and learning rate, require deeper investigation into their differences and computational efficiency. This study compares three architectures: Convolution Transformer Mixer (CTMixer), SpectralFormer enhanced by the Spectrum Motion Feature (SF-SMF), and Masked Auto Encoding Spectral-spatial Transformer (MAEST), using the Indian Pines, Pavia University, and Houston 2013 datasets. Indian Pines, mainly covering crops and natural vegetation, presents a challenging classification task due to limited samples. Preliminary results show that MAEST, even with optimal parameters, has lower accuracy and kappa than CTMixer and SF-SMF. Future work includes the Pavia University and Houston 2013 datasets. Parameters such as learning rate, epochs, and patch size will be standardized for all methods. Computational speed and performance will be compared, and classification results will be visualized to highlight differences in boundary sharpness and smoothness.

2. MATERIALS AND METHODS

Hyperspectral (HS) Images collect and store information across electromagnetic waves, which also are data cubes with spatial-spectral information. Due to the high spectral resolution and rich data content of HS images, they are applied in classification tasks across numerous domains. However, the high data dimensionality of HS images requires complex algorithms for processing and consumes a significant amount of computational resources and storage space. Deep learning has developed rapidly in recent years, and

their powerful fitting ability can extract features from multivariate data. Recent developments in transformer-based architectures for classifying hyperspectral images have shown outstanding accuracy in classification tasks. In our research, we are going to compare 3 brand-new transformer-based methods, which are Convolution Transformer Mixer, SpectralFormer enhanced by the Spectrum Motion Feature, and Masked Auto Encoding Spectral-spatial Transformer.

2.1 Introduction of three transformer-based methods

2.1.1 Convolution Transformer Mixer (CTMixer)

CTMixer is mainly composed of a Group Parallel Residual Block, a transformer encoder with convolution (TEC) branch and a CNN branch. Initially, the GPRB module extracts preliminary features from HSI patches, focusing on both spectral and spatial details. This is followed by the combined efforts of the TEC and CNN branches to capture detailed local and broad-scale information. To refine the accuracy further, the novel local–global multihead attention mechanism integrates convolutional and attention mechanisms, focusing on both localized and generalized data aspects. Instead of using the traditional class token found in ViT, this approach employs an average pooling layer to better integrate convolutional actions. The classification is finalized through a straightforward linear classifier.

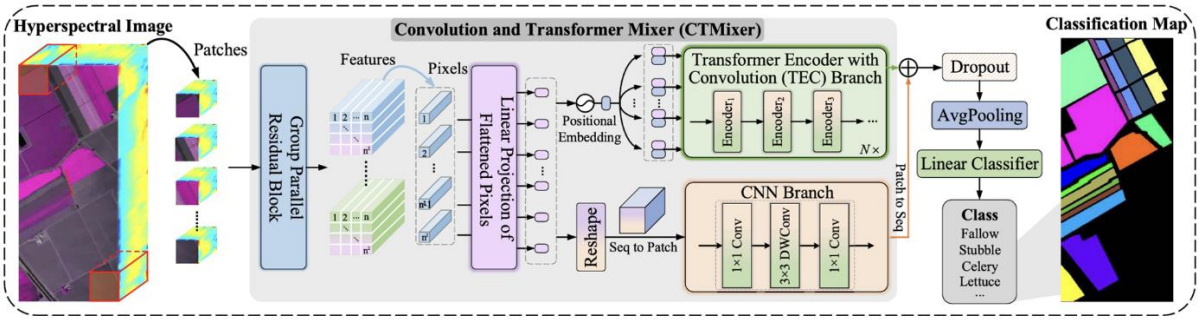

Figure 1: Overall framework of the proposed CTMixer for HS classification.

Source: Junjie Zhang et al. (2022). Convolution Transformer Mixer for Hyperspectral Image Classification.

In the original research, the Salinas, and Botswana datasets primarily focus on the classification of natural land cover types such as vegetation, crops, or water bodies, where

CTMixer shows excellent performance in handling vegetation type classification tasks. However, urban land cover types were not included in the original data selection, hence the effectiveness of these methods in urban area classification has not yet been validated.

2.1.2 SpectralFormer Enhanced by the Spectrum Motion Feature (SF-SMF)

The SF-SMF method enhances SpectralFormer with Spectrum Motion Feature, aiming to leverage the spectrum's discriminative potential fully. Despite SpectralFormer's innovative approach to encoding spectrum sequences, it falls short against advanced spectral-spatial methods. To address this, the authors incorporate an efficient sparse-to-dense optical flow estimation to track spectrum variations. These variations, termed spectrum motion features, boost the spectrum's discriminative capacity. Finally, SpectralFormer encodes these enhanced spectrum sequences for classification, improving accuracy and performance.

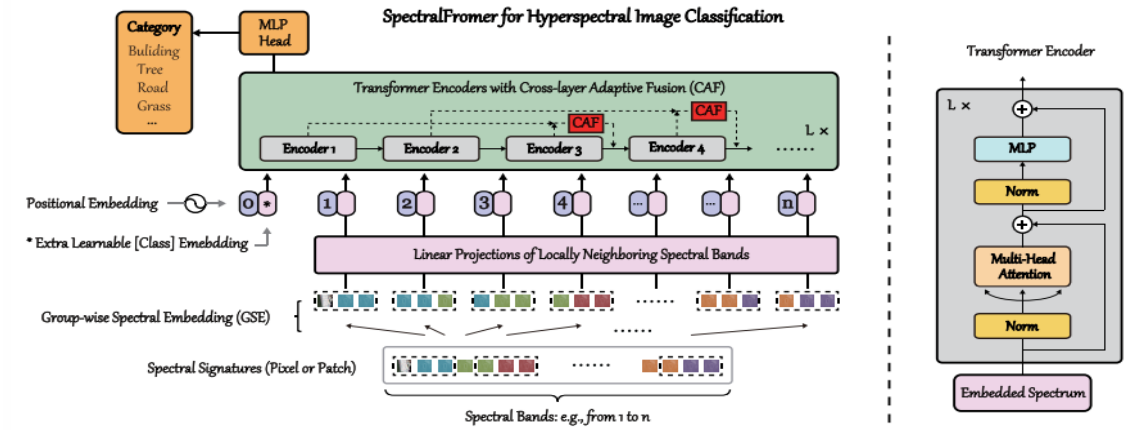

Figure 2: Overview illustration of the SpectralFormer network.

Source: Danfeng Hong et al. (2022). SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers.

For encoding the enhanced spectrum sequence, they followed the original SpectralFormer literature settings, adopting pixel-wise mode and CAF with a neighboring band width of 3. SF-SMF accurately reflects the true distribution of ground objects and avoids additional spatial information interference, showcasing its practical value in creating detailed whole-domain classification maps. However, SF-SMF's performance degrades with a small number of training samples due to insufficient training of SpectralFormer.

2.1.3 Masked Auto Encoding Spectral-spatial Transformer (MAEST)

MAEST consists of three main blocks: the reconstruction encoder (RE) and decoder (RD), and a classification encoder (CE). Each of these blocks is designed to specifically learn information from HS image data through specialized modules.

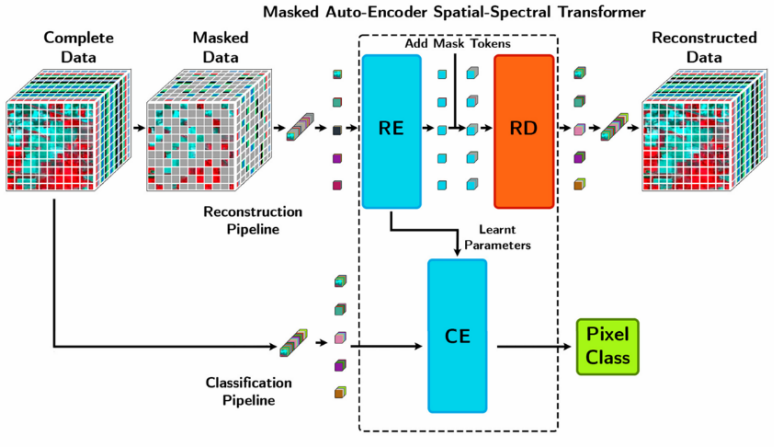

Figure 3: General scheme of MAEST, including two pipelines.

Source: Daimian Ibañez et al. (2022). Masked Auto-Encoding Spectral–Spatial Transformer for Hyperspectral Image Classification.

In the first pipeline, RE extracts a latent representation for unmasked segments of the spectral signature of each pixel, and the RD reconstructs the masked data from this latent representation. Also, to train RE and RD, the authors used the unlabeled training data in a self-supervised way. The second pipeline is responsible for supervised classification. In this branch, the complete labeled training data are used as input by a single block, the CE. This encoder exploits the robust feature extraction learned parameters in the reconstruction pipeline to categorize pixels after a short fine-tuning of the encoder and the classification layer.

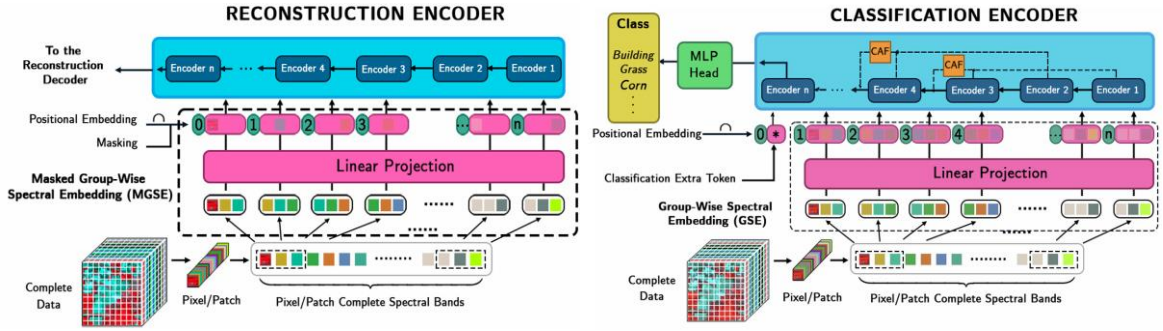

Figure 4: Detailed diagram of RE and CE.

Source: Daimian Ibañez et al. (2022). Masked Auto-Encoding Spectral–Spatial Transformer for Hyperspectral Image Classification.

To study the performance of the proposed MAEST, the researchers used three well-known HS datasets, the Indian Pines, the Pavia University, and the Houston2013 dataset. In this experiment, every datasets were set to different epochs.

2.2 Datasets

The methods mentioned above analyze different datasets, so in our upcoming study, we plan to use the same three sets of data for comparative analysis.

Table 1: Key Features of the Indian Pines, Pavia University and HOUSTON 2013

Dataset	Indian Pines	Pavia University	HOUSTON 2013
Spatial Resolution	20 m	1.3 m	2.5 m
Spectral Resolution	220	103	144
Sensor	Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)	Reflective Optics System Imaging Spectrometer (ROSIS)	ITRES CASI-1500
Categories	16	9	15

2.2.1 Indian Pines

Indian Pines mainly covers types of crops and natural vegetation, and make the classification task more challenging due to fewer samples.

2.2.2 Pavia University

Pavia University mostly includes roads, buildings, trees, etc., and it is commonly used to test performance in urban land cover classification.

2.2.3 HOUSTON 2013

With its large size and diverse types of land cover, HOUSTON 2013 is widely used in research on hyperspectral image processing.

3. Results and Discussion

We have already applied Indian Pines Dataset to the methods mentioned above and found that, even when set to their respective optimal parameters, the accuracy and kappa of MAEST remain lower than the other two.

Table 2: Preliminary result of Indian Pines on 3 methods

Indian Pines (IP)	CTMixer	SF-SMF	MAEST (Patch)
OA(%)	98.70	99.07	85.02
AA(%)	97.70	97.65	91.73
Kappa	0.985	0.986	0.832

4. Conclusion and Recommendation

In the future, we will use Pavia University, and Houston 2013. We also plan to set the parameters such as learning rate, epochs, and patch size to the same combination for these three methods and compare their computational speed and performance under these conditions. Also, we will visualize the classification results of these three methods to facilitate the comparison of their differences on the map, such as the sharpness or smoothness of the boundaries.

References

Daimian Ibañez, Ruben Fernandez-Beltran, Filiberto Pla, Naoto Yokoya (2022). Masked Auto-Encoding Spectral–Spatial Transformer for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 60, doi: 10.1109/TGRS.2022.3217892

Danfeng Hong, ZhuHan, JingYao, Lianru Gao, Bing Zhang, Antonio Plaza, Jocelyn Chanussot (2022). SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, doi: 10.1109/TGRS.2021.3130716

Junjie Zhang, Zhe Meng, Feng Zhao, Hanqiang Liu, Zhenhui Chang (2022). Convolution Transformer Mixer for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters, 19, doi: 10.1109/LGRS.2022.3208935

Yifan Sun, Bing Liu, Xuchu Yu, Anzhu Yu, Pengqiang Zhang, Zhixiang Xue (2023). Exploiting Discriminative Advantage of Spectrum for Hyperspectral Image Classification: SpectralFormer Enhanced by Spectrum Motion Feature. *IEEE Geoscience and Remote Sensing Letters, 20*, doi: 10.1109/LGRS.2022.3228531