Boosting Hyperspectral Image Classification with a 3D CNN and Vision Transformer Hybrid Architecture

Ghazala Hcini[1] and Imen Jdey[1]

***Abstract*— In recent years, Vision Transformer (ViT) models for hyperspectral image (HSI) classification have increased popularity due to their excellence in modeling long-range spatial-spectral dependencies, which have achieved state-of-the-art performance in classification, object detection, and segmentation tasks. However, ViTs are likely to miss local spatial features with fine granularity, while convolutional neural networks (CNNs) excel at extracting local patterns but cannot model long-range dependencies. We propose a hybrid framework that combines a 3D Residual CNN and a ViT module to counter these limitations. The CNN component possesses the ability to extract fine-grained local spectral and spatial information, which is passed through the ViT for extracting global contextual dependencies. Experimental evaluation on three benchmark HSI datasets, Indian Pines (IP), Pavia University (PU), and Salinas (SA), confirms the superiority of our approach. Experimental results show that our method outperforms the state-of-the-art methods in Hyperspectral image classification tasks, demonstrating good performance and potential application prospects.**

I. Introduction

Hyperspectral imaging is a powerful remote sensing technique that collects data in a broad range of the electromagnetic spectrum, including infrared and ultraviolet spectra [1]. Hyperspectral images can thus display close examination of the spectral characteristics of objects on the ground [2], and have significantly facilitated applications in agriculture [3], environmental monitoring [4], and military applications [5] [6]. Figure 1 illustrates the publication trends retrieved from Google Scholar by searching the terms ’hyperspectral image’ and ’classification’ in combination with application specific keywords such as ’agriculture’, ’medical diagnosis’, ’environment’, ’military’, and ’water management’, highlighting the growing interest in diverse domains.

However, the challenge lies in the high similarity and overlap between the spectral signatures of different objects, making it difficult to distinguish between them.

With these conditions in mind, examining strong spectral-spatial methods that integrate spatial information into hyperspectral image classification has increased immensely. Deep Learning (DL) methods have shown the ability to distinguish very subtle spectral variations in hyperspectral images. Among these methods, convolutional neural networks (CNNs) are the most representative [7]. Generally, CNNs consist of multiple layers with many parameters, which are learned adaptively.

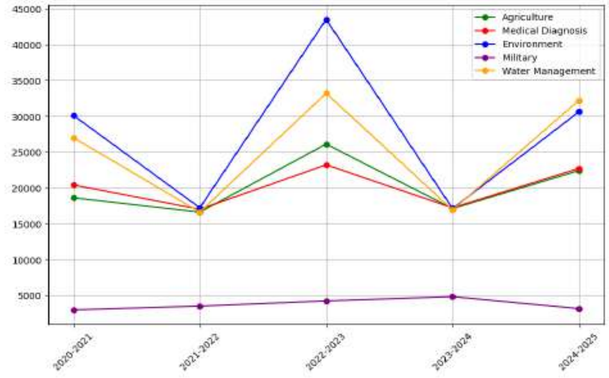


Fig. 1. Evolution of Hyperspectral Image Classification Applications (2020–2025) across different domains

Furthermore, due to the specific convolution operation, CNNs have a strong potential to capture local correlations. Consequently, CNNs can extract subtle spectral-spatial features to discriminate between different objects [8]. Numerous efforts have been made to improve the performance of CNNs for hyperspectral image classification by enhancing CNN backbones and their variants [9]. However, due to the architecture of CNNs, they cannot capture long-range correlations, which are crucial for hyperspectral image processing since these images typically contain hundreds of bands.
The transformer has emerged as an alternative to CNNs and has become a popular choice for DL architectures [10] [11]. It incorporates a self-attention mechanism, positional encoding, residual connections, and other components to capture long-range dependencies and relationships.

The Transformer’s ability to extract global features has made it widely used in natural language processing. Various variants, such as Vision Transformer, Swin Transformer, and Data-efficient image Transformers (DeiT) [12]

In this paper, we aim to address the limitations of existing hyperspectral image classification methods by introducing a set of novel DL architectures. The key contributions of our work are as follows:

- We propose a customized 3D Residual CNN tailored for HSI classification, effectively capturing spectral–spatial features.
- We develop a dedicated ViT model that leverages global attention mechanisms to enhance long-range dependency modeling in HSI data.
- We design a hybrid architecture that combines the

*This work was supported by ReGIM-Lab (REsearch Groups in Intelligent Machines)

[1]Ghazala Hcini is with ReGIM-Lab. REsearch Groups in Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, Tunisia `ghazala.hcini.doc@enetcom.usf.tn`

[1]Imen Jdey with ReGIM-Lab. REsearch Groups in Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, Tunisia `imen.jdey@fsegs.usf.tn`

strengths of the customized 3D Residual CNN and ViT, achieving both local feature extraction and global context understanding.

- We evaluate our models on three benchmark hyperspectral datasets, demonstrating superior classification accuracy and robustness compared to state-of-the-art methods.

The structure of this paper is as follows: The Background of the Study presents a background of CNNs and ViTs, highlighting their use for image classification. The Related Works section presents recent studies. The Materials and Methods section outlines the three datasets used and the methodology in focus. In the Results and Discussion, we present the results achieved as well as ablation experiments, and afterwards we discuss in depth the model's performance, its strengths, and its weaknesses. Finally, the Conclusion presents the main conclusions and proposes potential directions for future work.

II. Background of the Study

A. CNNs Architecture

Convolutional Neural Networks (CNNs) have shown promising results in image classification [13] [**?**], recognition [15], and localization tasks. We focus on image classification, the most important task in image processing [16] [17]. Multilayer neural networks attempt to learn representations from input images, enabling us to achieve specific results such as segmentation or classification without relying on handcrafted features. An example of CNN architecture for our classification task is shown in figure 2.

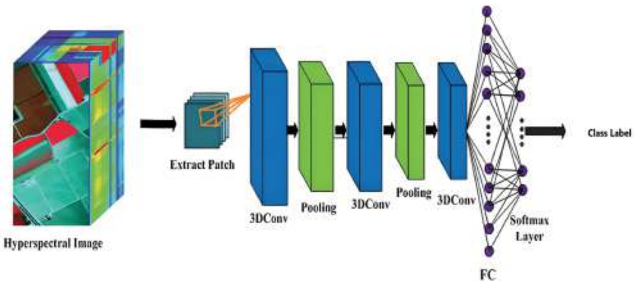


Fig. 2. CNNs Architecture

B. Vision Transformer Architecture

Vision Transformers (ViTs) use a transformer model based solely on self-attention mechanisms [18] [19] without convolutional layers. ViTs project input images into patches and compute uniform self-attention over them to learn global context in effect [12] (see figure 3). The non-hierarchical flat model allows ViTs to excel on those tasks where a good understanding of the whole input is required, such as image classification and natural language processing. ViTs have a more straightforward design, which facilitates effective training but requires high computational power.

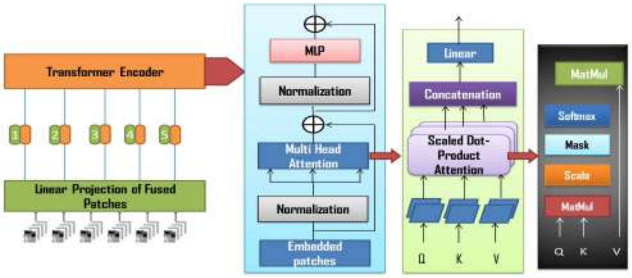


Fig. 3. Vision Transformer Architecture

III. Related Works

The HSI classification has been a widely studied area in remote sensing. One of the major challenges in this area is the scarcity of training labels. This area has made significant advances over the past few years, with DL techniques being used mainly.

The authors suggested in [20] a DL-based method for HSI classification that incorporates Squeeze-and-Excitation (SE) blocks into CNNs, resulting in the SE-CNN model. This architecture improves feature extraction by emphasizing essential channel-specific information, allowing for more efficient learning of spatial and spectral properties from HSI data. The model was evaluated on three benchmark datasets—Pavia University, Pavia Centre, and Salinas—and outperformed cutting-edge deep transfer learning models, with overall accuracies of 96.05%, 98.94%, and 96.33%, respectively.

The authors proposed in [21] a HSI classification method called Tri-CNN, which leverages a multi-scale 3D-CNN combined with a three-branch feature fusion strategy. Unlike traditional 2D-CNNs, Tri-CNN captures both spectral and spatial features at multiple scales, addressing the limitations of insufficient feature extraction and limited training samples. The extracted features from three branches are concatenated and passed through fully connected layers, followed by a softmax classifier. Experiments on the Pavia University, Salinas, and GulfPort datasets demonstrate that Tri-CNN achieves superior performance in terms of Overall Accuracy (OA), Average Accuracy (AA), and Kappa score compared to existing approaches.

In [22], the authors proposed an enhanced spectral fusion network (ESFNet) for hyperspectral image classification, addressing challenges related to spectral continuity and information loss. The model integrates two main components: a multi-scale fused spectral attention module (FsSE) to enrich spectral information through weighted attention, and a spectral-stride fusion 3D CNN (SSFCNN) that effectively learns spectral features by leveraging different spectral resolutions. This combination allows ESFNet to preserve spectral continuity while enhancing feature learning. Experimental results on the Indian Pines and Pavia University datasets show that ESFNet surpasses existing models in both classification accuracy and generalization ability.

To address the difficulty of limited training samples, the authors, in [23], suggested a deep-LSTM-based technique for Land Use/Land Cover (LULC) classification using hyperspectral images. The methodology consists of an autoencoder for feature extraction, a ranking-based algorithm for band selection, and a deep-LSTM network for final classification. When tested on the Pavia University, Kennedy Space Centre, and Indian Pines datasets, the proposed method outperforms existing techniques in terms of overall accuracy, average accuracy, and Kappa coefficient.

The authors, in [24], proposed an end-to-end deep spectral–spatial residual attention network (DSSpRAN) for HSI classification to alleviate the issues of overfitting and weak feature discrimination due to high dimensionality and limited training samples. The model visualizes HSI data in the form of a 3D cube and uses two modules: a spectral residual attention network (SRAN) that adaptively selects meaningful spectral features, and a SpRAN that attends to consistent labeling of nearby pixels. DSSpRAN is tested on five datasets and outperforms the state-of-the-art in accuracy and robustness for different land use and land cover scenarios.

IV. Materials and Methods

A. Datasets

The datasets used in this study are the standard benchmarks in the hyperspectral image analysis community. Their wide range of features, such as variation in spatial and spectral resolution, geographical area, and class distribution, provides a solid foundation for establishing the performance and applicability of the suggested classification models.

1) *Indian Pines* The Indian Pines dataset captured by the AVIRIS sensor features 224 spectral reflectance bands in the wavelength range 0.4 to $2.5times10^{-6}$ meters, and has 145 × 145 pixels, from `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Indian_Pines`. The dataset includes a variety of land cover classes and is a useful baseline for assessing classification methods. It contains 10,249 images with 16 classes.
2) *Pavia University* The ROSIS-03 sensor provided a hyperspectral dataset of ”Pavia University” from the University of Pavia, Italy, `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University`, is used in hyperspectral image classification and remote sensing studies due to its complexity and spectral diversity. The dataset comprises 103 bands of spectral resolution from 430 to 860 nm, with a spatial resolution of 1.3 meters per pixel. The image size is 610 x 340 pixels. It contains 7,456 images with 9 classes.
3) *Salinas scene* The 224 bands AVIRIS sensor captured a high-resolution view of California’s Salinas Valley, covering 217 samples and 512 lines, from `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Salinas`. The image, containing vineyards, barren soil, and vegetables, was obtained using radiance data at the sensor level. It contains 54,131 images with 16 classes.

B. Setup

Table I presents the experimental environment and hyperparameter settings used for model training. The values for the optimizer, learning rate, number of epochs, and batch size were determined and fixed after conducting several preliminary tests to ensure optimal model performance and stability during training. Additionally, a 90-10 training/validation split was adopted after extensive experimentation with various ratios.

TABLE I
Experimental Environment and Parameter Settings

Component	Details
Framework	PyTorch
Acceleration	CUDA (GPU-accelerated computing)
GPU	Nvidia GeForce RTX 3050 (24 GB VRAM)
Optimizer	AdamW
Learning Rate	0.0001
Epochs	100
Batch Size	9
Training/Validation Split	90% Training / 10% Validation

C. Followed Approach

1) 3D Residual CNN : The customized 3D Residual CNN model (figure 4) consists of 11 convolutional layers. It begins with an initial convolutional layer that has a kernel size of 1×1×7, spanning 7 spectral bands while preserving spatial layout. This layer is followed by batch normalization and a ReLU activation function to enhance numerical stability.

Next, the model includes 8 convolutional layers arranged into four residual blocks, each composed of two convolutional layers with skip connections. These skip connections help prevent vanishing gradients and ensure that information is not lost by adding the input of each block to its final layer. The first two residual blocks focus on extracting spectral features by processing neighboring spectral bands while maintaining spatial information. The last two residual blocks shift their focus to spatial feature extraction while refining and preserving spectral details.

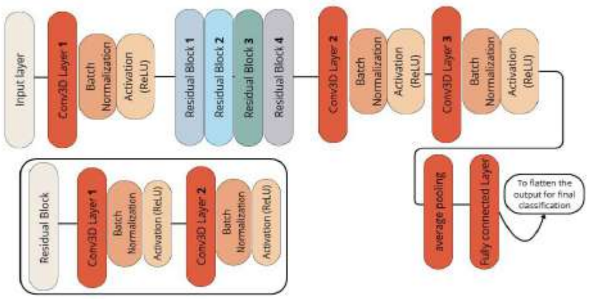


Fig. 4. 3D Residual CNN

After the residual blocks, a secondary convolutional layer condenses spectral features while retaining spatial information, followed by batch normalization and ReLU activation.

A third convolutional layer then integrates spatial and spectral features. The model concludes with an average pooling layer that reduces feature dimensions for classification, followed by a fully connected layer that outputs a localized feature map—a matrix representation containing rich local spectral-spatial information for final classification

The Patch Embedding procedure involves splitting a feature map into fixed-size patches, linearly embedding each patch, adding positional embeddings, and feeding the resulting sequence to the Transformer encoder (figure 5).

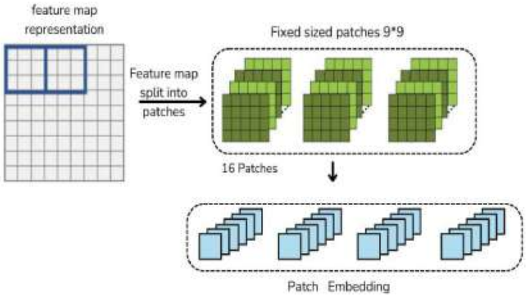


Fig. 5. Illustration of feature map partitioning into fixed-size patches

2) Vision Transformer: The ViT processes the local feature map representation produced by the preceding model. This 2D matrix, rich in localized spectral-spatial information, is divided into fixed-size patches as defined in the preprocessing phase. This is a key advantage of the hybrid model, as it uses feature-rich matrices instead of raw and standard pixel values as input. Each patch, representing a submatrix of local features, is flattened into a 1D vector and treated as a token. Positional embeddings are added to these tokens to preserve information about their location within the original feature map. The tokens, along with a special classification token (CLS), are then passed through a linear projection layer.

The embedded sequence is subsequently fed into the multi-head self-attention mechanism, which scans the sequence to learn and model global relationships between patches. This process allows the ViT to capture global context and dependencies across the entire feature map, enabling robust classification.

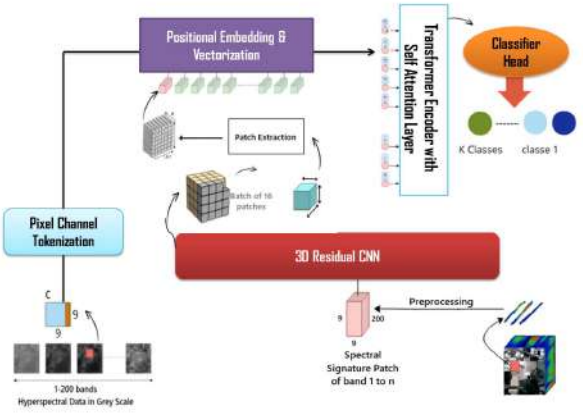
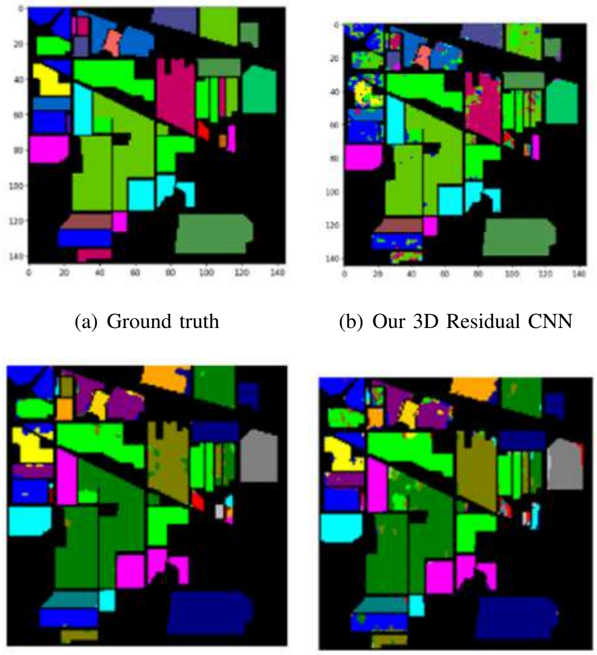


Fig. 6. Overview of the proposed Architecture



(a) Ground truth (b) Our 3D Residual CNN

(c) Our Vit (d) our hybrid model

Fig. 7. Ablation experimental research on Indian Pines Dataset

V. Results and Discussion

A. Obtained results and Ablation studies

The table II summarizes the classification results of three models: 3D Residual CNN, ViT, and a Hybrid model across three hyperspectral datasets: IP, PU, and SA.

TABLE II
Results of each proposed method on different datasets

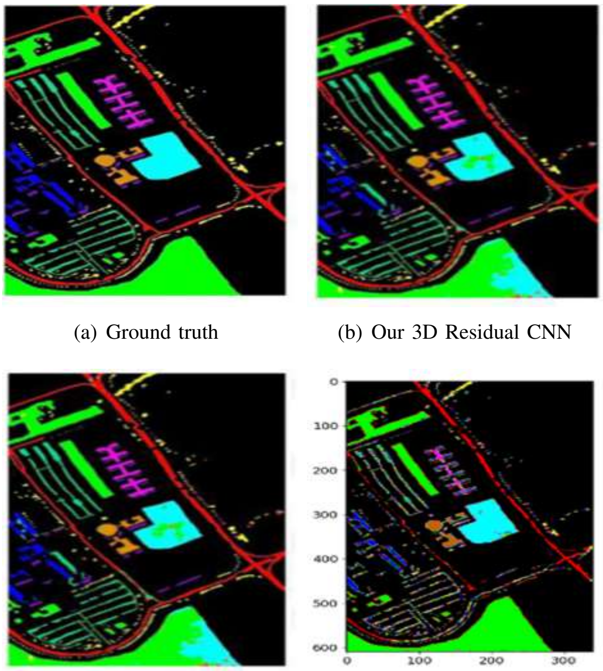
Method	IP Dataset	PU Dataset	SA Dataset

3D Residual CNN	OA=93.06%,
	AA=93.05%
		OA=96.05%,
		AA=96.24%
			OA=95.71%,
			AA=94.23%

ViT	OA=92.10%,
	AA=0.9265
		OA=90.29%,
		AA=90.00%
			OA=96.40%,
			AA=95.28%

Hybrid	OA=98.88%,
	AA=98.69%
		OA=99.51%,
		AA=99.56%
			OA=98.66%,
			AA=98.60%

Figures 7, 8, and 9 illustrate the results of ablation experiments conducted on three benchmark datasets: Indian Pines, Pavia University, and Salinas, respectively. Each figure presents a visual comparison of segmentation outcomes across different models, including the ground truth, the 3D Residual CNN, the ViT, and a hybrid model that combines both approaches. These comparisons highlight the effectiveness of each architecture in capturing spatial and spectral information, providing insights into the contributions of individual components within the hybrid model.



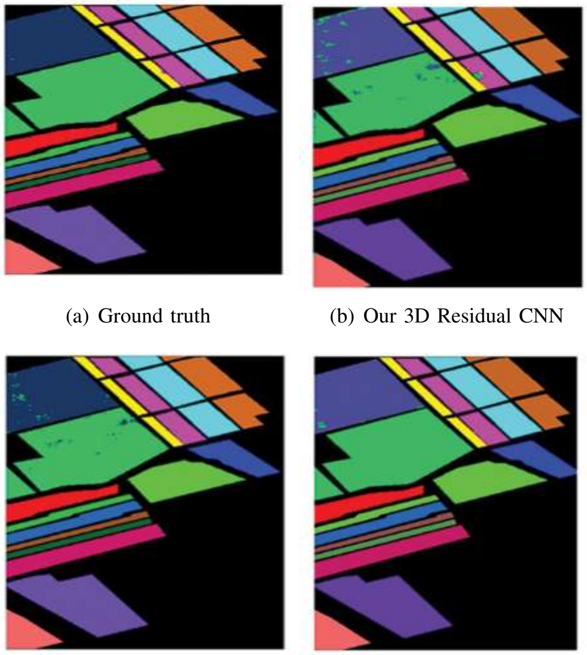
(a) Ground truth

(b) Our 3D Residual CNN

(c) Our Vit

(d) our hybrid model

Fig. 8. Ablation experimental research on Pavia University Dataset



(a) Ground truth

(b) Our 3D Residual CNN

(c) Our Vit

(d) our hybrid model

Fig. 9. Ablation experimental research on Salinas Dataset

B. Discussion and Advantages and Limitations

In this study, we propose three methods for hyperspectral image (HSI) classification: a CNN-based model, a ViT-based model, and a hybrid model that combines both. The first approach leverages a 3D Residual CNN tailored for small patch analysis (9×9×200), focusing on extracting local spectral-spatial features. Residual blocks play a critical role in maintaining training stability while effectively capturing both spatial and spectral patterns. This patch-based strategy enhances computational efficiency and reduces memory usage, making it suitable for high-dimensional HSI data.

The second approach employs a Vision Transformer (ViT), which processes raw HSI images by dividing them into patches, flattening them, and embedding them into tokens with positional information. Through self-attention mechanisms, the ViT captures both local and global dependencies. While powerful, this method typically demands large datasets and may not effectively capture fine-grained spectral-spatial details in complex data such as hyperspectral images.

To address the limitations of the individual methods, we introduce a third, hybrid approach. This model feeds local features extracted by a 3D Residual CNN into a ViT, allowing the transformer to focus on modeling global relationships based on rich, high-level spectral-spatial representations. By doing so, the hybrid model significantly reduces the data requirements and computational overhead typically associated with ViTs. More importantly, it combines the strengths of CNNs in local feature extraction with the ViT's ability to model long-range dependencies. This hybrid architecture is designed to achieve superior performance and robustness, particularly for the complex and high-dimensional nature of HSI data.

Table III presents a comparative analysis of our proposed model against several state-of-the-art approaches across three benchmark hyperspectral datasets: IP, PU, and SA. The proposed model consistently outperforms or matches existing methods in terms of Overall Accuracy (OA) and Average Accuracy (AA), achieving top results on the PU and SA datasets and competitive performance on the IP dataset. This demonstrates the robustness and effectiveness of our model across diverse hyperspectral image classification tasks.

TABLE III
COMPARATIVE STUDY

Model	**Indian Pines (IP)**	**Pavia University (PU)**	**Salinas (SA)**
SE-CNN [20]	-	OA=98.94%	OA=96.33%
Tri-CNN [21]	-	OA=92.66% ± 2.24, AA=90.65% ± 2.37	OA=96.68% ± 2.11, AA=97.69% ± 1.09
SSFCNN [22]	OA=90.125%	OA=96.044%	-
deep-LSTM [23]	OA=93.99%, AA=92.92%	OA=99.01%, AA=98.46%	-
DSSpRAN [24]	**OA=99.03%**, AA=98.08%	OA=98.76%, AA=98.12%	-
Proposed Model	**OA = 98.88%, AA = 98.69%**	**OA = 99.51%, AA = 99.56%**	**OA = 98.66%, AA = 98.60%**

1) Advantages:

- Hybrid Strengths: Combines CNN's local focus with ViT's global context modeling for improved performance.

- State-of-the-Art Results: Achieves top or competitive accuracy on IP, PU, and SA datasets.
- Explainability via Attention: The attention mechanism in the ViT enhances model interpretability by highlighting informative regions.

2) Limitations:

- Hyperparameter Sensitivity: Requires careful tuning of CNN and Transformer settings.
- Preprocessing Dependence: Performance can be affected by the quality of patch extraction and data normalization.

VI. Conclusion

In this work, we proposed a hybrid framework for HSI by combining a 3D deep residual CNN with a ViT. The 3D CNN captures spatial-spectral features, while the ViT models global dependencies through attention, resulting in improved classification accuracy.

Experiments on IP, PU, and SA datasets demonstrate the model's strong performance.

However, the model is sensitive to patch size and computationally demanding due to its depth. Future work will explore adaptive patch size selection using self-attention, enhance efficiency with automated hyperparameter tuning, and extend the model for real-time applications in remote sensing and agriculture.

ACKNOWLEDGMENT

This work would not have been possible without the support of the Ministry of Higher Education and Scientific Research of Tunisia (LR11ES48).

References

[1] Bhargava, Anuja, et al. "Hyperspectral imaging and its applications: A review." *Heliyon* 10.12 (2024).

[2] Wang, Zhengyang, and Shufang Tian. "Ground object information extraction from hyperspectral remote sensing images using deep learning algorithm." *Microprocessors and Microsystems* 87 (2021): 104394.

[3] Lu, Bing, et al. "Recent advances of hyperspectral imaging technology and applications in agriculture." *Remote Sensing* 12.16 (2020): 2659.

[4] Stuart, Mary B., Andrew JS McGonigle, and Jon R. Willmott. "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems." *Sensors* 19.14 (2019): 3071.

[5] Jdey, Imen, et al. "Fuzzy fusion system for radar target recognition." International Journal of Computer Applications & Information Technology 1.3 (2012): 136-142.

[6] Shimoni, Michal, Rob Haelterman, and Christiaan Perneel. "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques." *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019): 101-117.

[7] Yu, Shiqi, Sen Jia, and Chunyan Xu. "Convolutional neural networks for hyperspectral image classification." *Neurocomputing* 219 (2017): 88-98.

[8] Kong, Fanqiang, et al. "A spectral-spatial feature extraction method with polydirectional CNN for multispectral image compression." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022): 2745-2758.

[9] Yao, Jing, et al. "Semi-active convolutional neural networks for hyperspectral image classification." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-15.

[10] Chitty-Venkata, Krishna Teja, et al. "Neural architecture search for transformers: A survey." *IEEE Access* 10 (2022): 108374-108412.

[11] Chen, Chun-Fu Richard, Quanfu Fan, and Rameswar Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[12] Hcini, Ghazala, Imen Jdey, et al. "Hyperparameter optimization in customized convolutional neural network for blood cells classification". *J. Theor. Appl. Inf. Technol*, 2021, vol. 99, p. 5425-5435.

[13] Tripathi, Milan. "Analysis of convolutional neural network based image classification techniques." *Journal of Innovative Image Processing (JIIP)* 3.02 (2021): 100-117.

[14] Jlassi, Sinda, Imen Jdey, and Hela Ltifi. "Bayesian hyperparameter optimization of deep neural network algorithms based on ant colony optimization." *Document Analysis and Recognition–ICDAR* 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16. Springer International Publishing, 2021.

[15] Hijazi, Samer, Rishi Kumar, and Chris Rowen. "Using convolutional neural networks for image recognition." Cadence Design Systems Inc.: San Jose, CA, USA 9.1 (2015).

[16] Elngar, Ahmed A., et al. "Image classification based on CNN: a survey." *Journal of Cybersecurity and Information Management* 6.1 (2021): 18-50.

[17] Hcini, Ghazala, Imen Jdey, and Hela Ltifi. "HSV-Net: a custom cnn for malaria detection with enhanced color representation." 2023 International Conference on Cyberworlds (CW). IEEE, 2023.

[18] Mehrani, Paria, and John K. Tsotsos. "Self-attention in vision transformers performs perceptual grouping, not attention." *Frontiers in Computer Science* 5 (2023): 1178450.

[19] Slimani, Nawel, Imen Jdey, and Monji Kherallah. "Improvement of Satellite Image Classification Using Attention-Based Vision Transformer." ICAART (3). 2024.

[20] Rajendran, T., et al. "Hyperspectral image classification model using squeeze and excitation network with deep learning." *Computational intelligence and neuroscience 2022* (2022): 9430779.

[21] Alkhatib, Mohammed Q., et al. "Tri-CNN: A three branch model for hyperspectral image classification." Remote Sensing 15.2 (2023): 316.

[22] Zhou, Junbo, et al. "An enhanced spectral fusion 3D CNN model for hyperspectral image classification." *Remote Sensing* 14.21 (2022): 5334.

[23] Tejasree, Ganji, and L. Agilandeeswari. "Land use/land cover (LULC) classification using deep-LSTM for hyperspectral images." *The Egyptian Journal of Remote Sensing and Space Sciences* 27.1 (2024): 52-68.

[24] Chhapariya, Koushikey, Krishna Mohan Buddhiraju, and Anil Kumar. "A deep spectral–spatial residual attention network for hyperspectral image classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024): 15393-15406.