WaveFormer: Spectral–Spatial Wavelet Transformer for Hyperspectral Image Classification

Muhammad Ahmad, Usman Ghous, Muhammad Usama, and Manuel Mazzara

***Abstract*— Transformers have proven effective for hyperspectral image classification (HSIC) but often incorporate average pooling that results in information loss. This letter presents WaveFormer, a novel transformer-based approach that leverages wavelet transforms for invertible downsampling. This preserves data integrity while enabling attention learning. Specifically, WaveFormer unifies downsampling with wavelet transforms to decompress feature maps without loss. This provides an efficient tradeoff between performance and computation. Furthermore, the wavelet decomposition enhances the interaction between structural and shape information in image patches and channel maps. To evaluate WaveFormer, we conducted extensive experiments on two benchmark hyperspectral datasets. Our results demonstrate that WaveFormer achieves state-of-the-art classification accuracy, obtaining overall accuracies of 95.66% and 96.54% on the Pavia University and the University of Houston datasets, respectively. By integrating wavelet transforms, WaveFormer presents a new transformer architecture for hyperspectral imagery that achieves superior classification without information loss from average pooling.**

***Index Terms*— Hyperspectral image classification (HSIC), spatial–spectral feature, spatial–spectral transformers (SSTs), wavelet transformer (WaveFormer).**

I. Introduction

HYPERSPECTRAL imaging (HSI) has emerged as a powerful remote sensing technique, capturing contiguous spectral information across various wavelengths. Its applications span diverse fields, including remote sensing [1], [2], earth observation [3], urban planning [4], agriculture [5], forestry [6], target/object detection [7], mineral exploration [8], environmental monitoring [9], [10], and climate change [11]. Notably, HSI excels in capturing detailed spatial and spectral information, although its sensors, characterized by high spectral resolution, may face challenges in achieving optimal spatial resolution, especially in complex scenarios.

Hyperspectral image classification (HSIC) involves categorizing pixels based on their spectral and spatial characteristics, facilitating object identification. Traditionally, HSIC used both traditional machine learning (TML) and deep learning (DL) techniques [12]. DL, particularly convolutional neural networks (CNNs), addresses challenges in handling multimodal data. However, CNNs struggle with capturing long-term dependencies for spectrally similar classes. Recurrent neural networks (RNNs) can model these dependencies but lack the ability for simultaneous model training, a vital consideration for the extensive HSI data comprising numerous samples [13].

On the other hand, transformers, using self-attention mechanisms [14], represent state-of-the-art networks for handling dependencies and enabling parallel training. Specifically tailored for HSIC challenges, vision transformers (ViTs) have emerged as promising candidates. ViTs use self-attention to capture relationships in image sequences represented as patches, proving beneficial for modeling long-range dependencies across the entire HSI cubes [15], [16]. However, ViTs encounter challenges related to scale invariance and texture features. Their fixed-size input may hinder recognizing objects at different scales, and their emphasis on global context may limit their effectiveness in extracting fine-grained texture details [17]. In addition, the extensive data requirements for ViTs may pose challenges in scenarios where a large number of training samples are not readily available for HSIC [18].

To address the aforementioned issues, this letter proposes WaveFormer, a novel approach that combines the strengths of wavelets and transformers for improved HSIC. It introduces spatial–spectral wavelet convolution within a transformer architecture, enhancing the interaction between structural and shape information of image tokens. This leads to more accurate classification compared with the traditional transformer-based models. WaveFormer extracts wavelet-based multiscale spatial–spectral features from the HSI data, which are then input into a classification model. The combination of wavelets and transformers allows WaveFormer to capture both local and global relationships in the data, resulting in improved classification accuracy. In this letter, we made the following contributions.

1) This work introduces WaveFormer, which integrates wavelet transformation and transformers for HSIC. Within the transformer architecture, the WaveFormer model incorporates a trainable spatial–spectral wavelet network, thereby improving the interaction between the structural and shape information of HSI tokens and class tokens.

Manuscript received 3 November 2023; revised 5 January 2024; accepted 8 January 2024. Date of publication 15 January 2024; date of current version 7 February 2024. *(Corresponding author: Muhammad Ahmad.)*

Muhammad Ahmad, Usman Ghous, and Muhammad Usama are with the Department of Computer Science, National University of Computer and Emerging Sciences, (NUCES), Islamabad 38000, Pakistan (e-mail: mahmad00@gmail.com; usman.ghous@nu.edu.pk; m.usama@nu.edu.pk).

Manuel Mazzara is with the Institute of Software Development and Engineering, Innopolis University, 420500 Innopolis, Russia (e-mail: m.mazzara@innopolis.ru).

Digital Object Identifier 10.1109/LGRS.2024.3353909

2) The proposed method extracts multiscale spatial–spectral features based on wavelets from the HSI cube. These features are subsequently input into a classification model, to capture a broader range of information beneficial for HSIC.
3) This work showcases the combined impact of wavelet-based feature extraction and transformer-based modeling of global relationships, indicating the potential for improved accuracy compared with the traditional transformer approaches for HSIC.

II. Problem Formulation

Let us consider HSI data comprising B spectral bands, each with a spatial resolution of $M \times N$ pixels. The HSI data cube $X \in \mathbb{R}^{(M\times N\times B)}$ is first divided into overlapping 3-D patches. Each patch is centered at a spatial location (α, β) and covers a spatial extent of $S \times S$ pixels across all B-bands. The total number of 3-D patches (m) extracted from X (i.e., $X \in \mathbb{R}^{(S\times S\times B)}$) is $(M-S+1)\times(N-S+1)$. A patch located at (α, β) is denoted as $P_{\alpha,\beta}$ and spans spatially from $\alpha-(S-1/2)$ to $\alpha+(S-1/2)$ in width and $\beta-(S-1/2)$ to $\beta+(S-1/2)$ in height. The labeling of these patches is determined by the label assigned to the central pixel within each patch. The transformed patches (as explained in Algorithm1) are then processed by the baseline 3DCNN model. Specifically, the activation value at spatial location (x, y, z) in the jth feature map of the ith layer, $v_{i,j}^{x,y,z}$, is computed as

$$v_{i,j}^{x,y,z} = \phi\left(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\lambda=-v}^{v} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho,\lambda} \times v_{i-1,\tau}^{x+\sigma,y+\rho,z+\lambda}\right)$$

where ϕ is the activation function, $b_{i,j}$ is the bias for the jth feature map at the ith layer, d_{l-1} is the number of feature maps in the (l-1)th layer, and $w_{i,j}$ is the depth of the kernel for the jth feature map at the ith layer. $2\gamma+1$, $2\delta+1$, and $2v+1$ define the width, height, and depth of the kernel, respectively, along the spatial–spectral dimension.

Algorithm 1 Wavelet Transformation

Input: Image Patch X_{patch}

```
Set L = 4 and Initialize empty output array O;
for i = 1 to S do
    for j = 1 to B do
        Coffs = wavedec2(X_patch[i, :, :, j], haar, L);
        O[i, :, :, j] = waverec2(Coffs, haar);
    end
end
```

The 3-D feature maps $v_{i,j}^{x,y,z}$ are transformed into $\hat{X} = X\ E_d$ with reduced channel dimensions via embedding matrix $E_d \in R^{D\times(D/4)}$, i.e.,$\hat{X} \in R^{M\times N\times(D/4)}$ (decomposed into four wavelet subbands) which is downsampled through wavelet transformation. Note that here we used the classical Haar wavelet as expressed in [19] and [20]. Concretely, the wavelet transformation is applied using a low-pass filter, denoted as $f_L = \left((1/(2)^{1/2}), (1/(2)^{1/2})\right)$, and a high-pass filter, denoted as $f_H = \left((1/(2)^{1/2}), -(1/(2)^{1/2})\right)$, along the rows of the input data $\hat{X}$. This process results in the creation of two subbands, namely, $\hat{X}L$ and $\hat{X}H$. Subsequently, the same low-pass filter f_L and high-pass filter f_H are used, this time along the columns of the derived subbands $\hat{X}L$ and $\hat{X}H$. This leads to the formation of four subbands in total: $\hat{X}LL$, $\hat{X}LH$, $\hat{X}HL$, and $\hat{X}HH$. Each of these wavelet subbands can be viewed as a downsampled version of the original input $\hat{X}$. They collectively preserve all the input details without any loss of information. The four wavelet subbands are created: $\hat{X}LL$, $\hat{X}LH$, $\hat{X}HL$, and $\hat{X}HH$. These feature maps are then linearly transformed into downsampled keys $K^w \in \mathbb{R}^{m\times D}$ and values $V^w \in \mathbb{R}^{m\times D}$, where $m = (M/2) \times (N/2)$ denotes the total number of patches. Multiheaded self-attention learning (Attention) is then performed on the queries and their respective downsampled keys and values for each attention head

$$\begin{aligned}\text{head}_J &= \text{Attention}(Q_j, K_j^w, V_j^w)\\ &= \text{Softmax}\left(\frac{Q_j K^{w_j T}}{\sqrt{D_h}}\right) V_j^w\end{aligned}$$

where K_j^w and V_j^w represent the downsampled keys and values specific to the jth head, respectively. It is worth noting that the collective output of self-attention learning for each head can be understood as the incorporation of long-range contextualized information from the input data. Finally, the features are fed into a fully connected layer for classification, and the softmax function is applied to generate the class probability distributions from which the final ground-truth maps are generated. The WaveFormer captures the essence of the wavelet for HSIC. By integrating spatial–spectral information through attention mechanisms and linear projections, WaveFormer can effectively process the HSI cube with reduced computational complexity, making it suitable for resource-constrained environments. Fig. 1 explains the complete model in detail.

III. Experimental Results and Discussion

A. Comparison With CNN-Based Networks

A uniform experimental methodology is imperative when evaluating CNN-based approaches. It is crucial to uphold consistency in the distribution of samples assigned to training, validation, and testing. Each comparative model underwent training and validation using 5% of the samples, respectively, while the remaining 90% were used for classification based on 10×10 pixel patches. The performance evaluation of WaveFormer is conducted on the University of Houston dataset to compare the performance against several models: 3D CNN [21], Hybrid Inception Net [22], 3-D Inception Net [23], 2-D Inception Net [24], 2D CNN [25], and Hybrid CNN [26]. Fig. 2 presents a graphical representation of accuracy and loss convergence over 50 training epochs for both the training and validation sets. In addition, the results highlight the advantage of decoupling spatial–spectral information, proving to be a superior approach in approximating information within an HSI cube compared with the alternative strategy of convolutions at distinct layers of the architectures. When contrasted with the 3-D models, Fig. 3 illustrates that the proposed WaveFormer

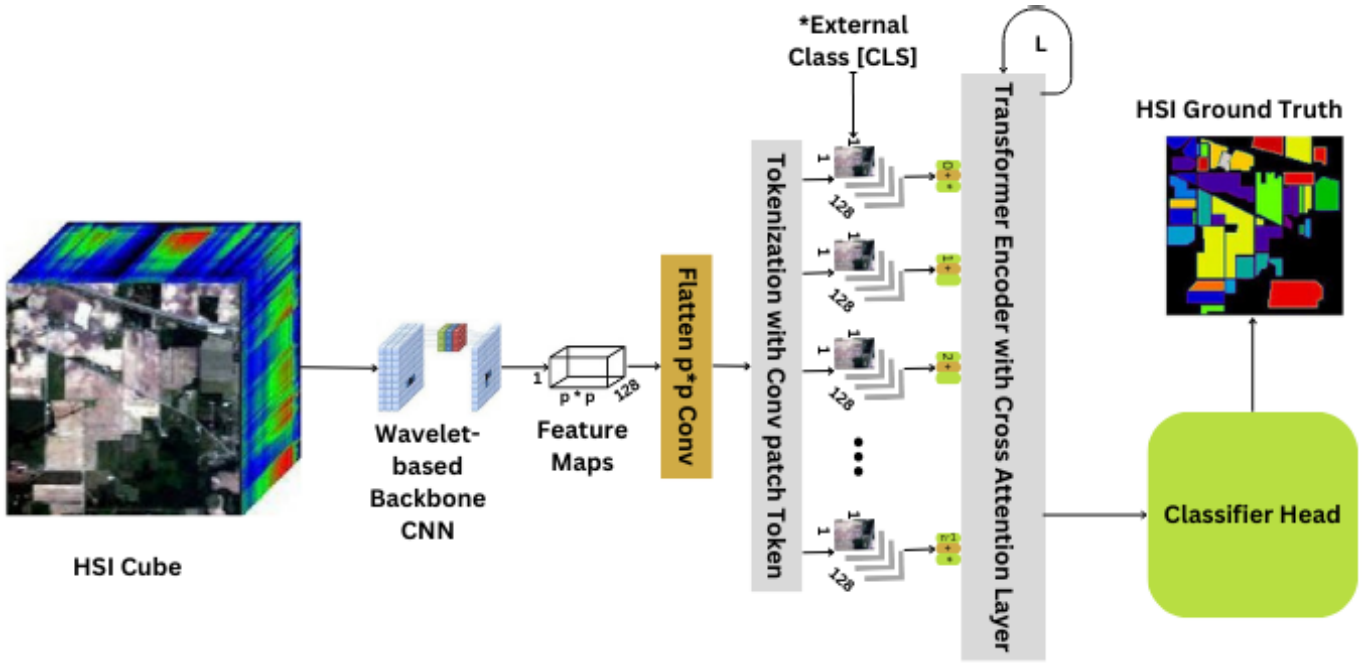

Fig. 1. HSI cube was initially partitioned into overlapping 3-D patches, each of which was centered at a spatial point and covered a $S \times S$ pixel extent over all the spectral bands. Wavelet transform was applied to these patches using Haar wavelets, resulting in four subbands that captured different frequency components and spatial features. The subbands were then concatenated to generate a new 3-D representation. Locally contextualized feature maps were produced through a 3-D convolution to define spatial locality within this representation. To incorporate long-range contextualized information, these feature maps were further translated into downsampled keys and values, and multiheaded attention learning was performed.

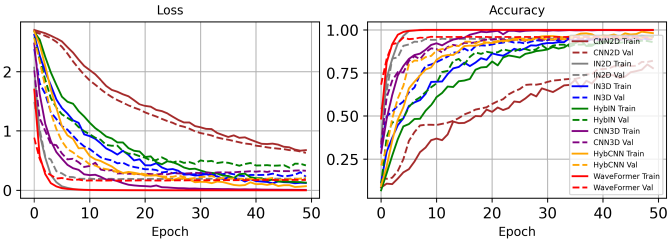

Fig. 2. Accuracy and loss trends on the University of Houston dataset.

achieves comparable results, notably achieving impressive scores of 97% (approx). In particular, models using 2-D convolutions exhibit varying performance on the UH dataset, with 2-D CNN achieving OA, AA, and κ scores of 78%, and the 2-D Inception Net attaining OA, AA, and κ scores of 95%. The marginal difference in accuracy between the proposed model and comparative methods is noteworthy and can be attributed to the computational efficiency of WaveFormer, its ability to mitigate overfitting, and its enhanced capacity to model spatial and spectral dependencies effectively.

B. Comparison With Transformer-Based Networks

For comparison, we selected several state-of-the-art networks including Attention Is All You Need (transformer) [14], Spectralformer [16], hyperspectral image transformer classification networks (HiTs) [27], and CSiT: a multiscale ViT for HSIC [28]. The transformer architecture uses a standard ViT [29] adapted for pixelwise HSI input with five encoder blocks. Spectralformer retains the original five transformer encoder architecture with pixelwise embeddings and cross-layer fusion. HiT incorporates two depthwise convolution

TABLE I

PAVIA UNIVERSITY: PROPOSED WAVEFORMER VERSUS COMPARATIVE MODELS. ALL THESE MODELS ARE EVALUATED USING 8×8 PATCH SIZE WITH 5/5/90% TRAINING/VALIDATION/TEST SAMPLES, RESPECTIVELY

Class	ViT [14]	SF [16]	HiT [27]	CSiT [28]	CSiT-CSAF [28]	WaveFormer
Asphalt	95.67%	92.67%	95.21%	93.84%	95.25%	**96.21%**
Meadows	88.37%	92.79%	92.54%	95.23%	96.64%	**99.33%**
Gravel	73.71%	90.60%	**91.18%**	88.79%	81.74%	81.31%
Trees	98.03%	**98.15%**	97.21%	96.19%	95.50%	96.77%
Painted	99.01%	98.28%	100.0%	99.18%	99.59%	**100%**
Soil	89.26%	93.29%	**99.93%**	91.99%	93.43%	91.55%
Bitumen	79.40%	83.01%	**95.75%**	92.06%	89.24%	89.55%
Bricks	85.54%	84.50%	**97.59%**	82.25%	88.81%	89.34%
Shadows	99.65%	**99.77%**	99.47%	99.19%	99.53%	96.47%
OA	89.32%	92.30%	91.35%	93.35%	94.48%	**95.66%**
AA	87.39%	88.86%	85.07%	90.48%	93.15%	**93.39%**
κ	86.60%	89.66%	88.94%	91.13%	92.67%	**94.22%**

layers for spatial processing and one pointwise convolution layer for spectral processing, relying solely on patchwise input. CSiT uses a consistent backbone transformer architecture in each branch—four heads in the small-scale branch and six heads in the large-scale branch. Token sequences are input into two cross-attention transformers, each with a four-head attention layer and multilayer perceptron layer. CSiT is evaluated with and without cross-spectral attention fusion (CSAF) modules. All the results use the specified parameter configurations from the original papers to enable direct comparison.

The comprehensive results of the earlier mentioned models are provided in Tables I and II. In summary, the proposed WaveFormer model demonstrates remarkable performance, outperforming the state-of-the-art ViT-based models across a spectrum of evaluation metrics, encompassing overall accuracy (OA), average accuracy (AA), and the κ coefficient. A thorough analysis of the quantitative performance reveals

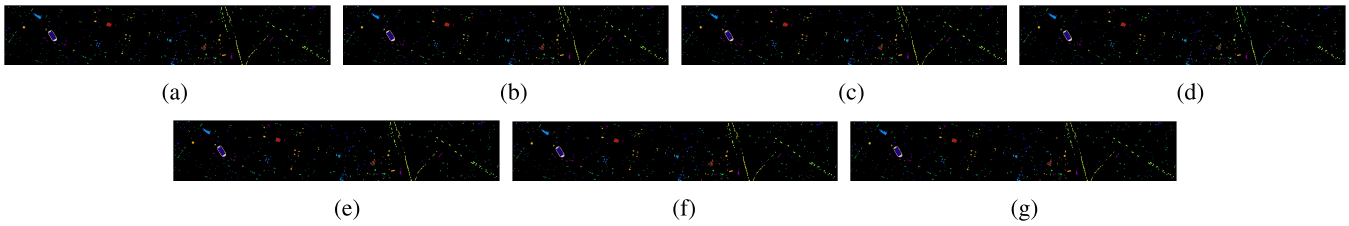

Fig. 3. Proposed WaveFormer achieves 96.6881% on the UH dataset, showing competitive performance compared with the recent state-of-the-art CNN models for HSIC. (a) Two-dimensional Inception Net [24]: training time = 263.4120; test time = 20.0542; OA = 95.3426; AA = 94.3591; and Kappa = 94.9626. (b) Three-dimensional inception net [23]: training time = 623.1804; test time = 56.1943; OA = 96.5180; AA = 96.3571; Kappa = 96.2351. (c) Hybrid Inception Net [22]: training time = 607.5244; test time = 55.2575; OA = 95.1504; AA = 94.8054; and Kappa = 94.7554. (d) 2D CNN [25]: training time = 22.1890; test time = 3.0396; OA = 77.3859; AA = 77.7413; and Kappa = 75.5432. (e) 3D CNN [21]: training time = 202.6686; test time = 14.4914; OA = 95.7122; AA = 94.9145; and Kappa = 95.3622. (f) Hybrid CNN [26]: training time = 159.9331; test time = 20.9201; OA = 95.5792; AA = 95.5704; and Kappa = 95.2199. (g) WaveFormer: training time = 208.9219; test time = 21.4688; OA = 96.6881; AA = 95.8371; and Kappa = 96.4176.

TABLE II

UNIVERSITY OF HOUSTON: PROPOSED WAVEFORMER VERSUS COMPARATIVE MODELS. ALL THESE MODELS ARE EVALUATED USING 2×2 PATCH SIZE WITH 10/10/80% TRAINING/VALIDATION/TEST SAMPLES, RESPECTIVELY

Class	ViT [14]	SF [16]	HiT [27]	CSiT [28]	WaveFormer
Healthy grass	90.28%	93.31%	97.26%	93.39%	**98.90%**
Stressed grass	98.21%	97.81%	97.29%	**99.54%**	97.60%
Synthetic grass	96.90%	**100.00%**	98.74%	**100.00%**	99.82%
Trees	**100.00%**	**100.00%**	95.78%	98.09%	99.19%
Soil	96.89%	98.16%	98.41%	97.70%	**99.79%**
Water	98.21%	**100.00%**	91.36%	**100.00%**	98.07%
Residential	82.82%	87.83%	94.60%	90.96%	**90.64%**
Commercial	79.83%	85.91%	91.82%	89.18%	**97.38%**
Road	76.88%	75.33%	92.39%	90.62%	**97.40%**
Highway	80.31%	82.52%	90.61%	93.22%	**97.75%**
Railway	83.17%	79.19%	89.09%	87.91%	**96.76%**
Parking Lot 1	69.42%	72.76%	94.18%	83.15%	**97.56%**
Parking Lot 2	63.08%	79.49%	82.51%	**84.11%**	65.33%
Tennis Court	90.36%	93.90%	91.55%	97.21%	**98.83%**
Running Track	94.40%	97.50%	96.72%	**100.00%**	99.05%
OA	86.45%	88.45	93.06%	93.09%	**96.54%**
AA	86.60%	87.81	86.61%	92.06%	**95.60%**
κ	85.35%	87.50	92.50%	92.53%	**96.26%**

that WaveFormer consistently achieves superior results across diverse categories, exhibiting substantial improvement in accuracy, as evidenced in Table II. Notably, while performance differences remain relatively modest in the PU dataset due to the abundance of samples, the UH dataset poses a significant challenge to modeling capabilities. For example, when assessing the challenging UH dataset, WaveFormer outperforms the baseline ViT by more than 10%, and it surpasses SpectralFormer by approximately 8%. Furthermore, the AA achieved by the WaveFormer exceeds that of both ViT and SpectralFormer by margins ranging from 10%, emphasizing the potential efficacy of spatial–spectral feature extraction. In a comparative context with the most recent spatial–spectral transformer and CSiT models, the WaveFormer consistently presents results, showcasing its proficiency in both the spectral and spectral–spatial feature extraction tasks. It is worth noting that while HiT excels in identifying land-cover or land-use classes with spectral–spatial information, the WaveFormer approaches similar levels of performance. To sum up, these findings emphasize the robustness and effectiveness of the WaveFormer model in the field of HSIC, particularly in scenarios where the extraction of spatial–spectral information holds importance, especially when considering the limited availability of training samples.

IV. CONCLUSION

This letter introduced "WaveFormer" which combines the power of wavelet transforms and ViT for HSIC. By extracting multiscale spatial–spectral features using wavelets and feeding them into a transformer encoder, WaveFormer can capture both the local texture patterns and global contextual relationships in an end-to-end trainable model. A notable innovation is the use of wavelet convolution within the transformer's attention mechanism, allowing for enhanced integration of spectral and structural information. Extensive experiments demonstrate WaveFormer achieves the state-of-the-art performance, particularly for challenging datasets with limited training data where its multiscale extraction of spatial–spectral cues proves valuable. Beyond superior classification accuracy, WaveFormer attains robustness and generalizability that hold promise for addressing real problems in remote sensing. Future work could explore self-supervised pretraining and network optimizations to maximize WaveFormer's potential when data are scarce.

REFERENCES

[1] D. G. Manolakis, R. B. Lockwood, and T. W. Cooley, *Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[2] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.

[3] V. Lodhi, D. Chakravarty, and P. Mitra, "Hyperspectral imaging for Earth observation: Platforms and instruments," *J. Indian Inst. Sci.*, vol. 98, no. 4, pp. 429–443, Dec. 2018.

[4] S. Roessner, K. Segl, U. Heiden, and H. Kaufmann, "Automated differentiation of urban surfaces based on airborne hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1525–1532, Jul. 2001.

[5] B. Lu, P. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, p. 2659, Aug. 2020.

[6] T. Adão et al., "Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sens.*, vol. 9, no. 11, p. 1110, Oct. 2017.

[7] C.-I. Chang, H. Ren, and S.-S. Chiang, "Real-time processing algorithms for target detection and classification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 4, pp. 760–768, Apr. 2001.

[8] E. Bedini, "The use of hyperspectral remote sensing for mineral exploration: A review," *J. Hyperspectral Remote Sens.*, vol. 7, no. 4, pp. 189–211, Dec. 2017.

[9] C. Weber et al., "Hyperspectral imagery for environmental urban planning," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1628–1631.

[10] Stuart, McGonigle, and Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, p. 3071, Jul. 2019.

[11] C. B. Pande and K. N. Moharir, "Application of hyperspectral remote sensing role in precision farming and sustainable agriculture under climate change: A review," in *Climate Change Impacts on Natural Resources, Ecosystems and Agricultural Systems*. Cham, Switzerland: Springer, 2023, pp. 503–520. [Online]. Available: https://doi.org/10.1007/978-3-031-19059-9_21

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

[13] M. K. Khlifi, W. Boulila, and I. R. Farah, "Graph-based deep learning techniques for remote sensing applications: Techniques, taxonomy, and applications—A comprehensive review," *Comput. Sci. Rev.*, vol. 50, Nov. 2023, Art. no. 100596.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[15] L. Scheibenreif, M. Mommert, and D. Borth, "Masked vision transformers for hyperspectral image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2165–2175.

[16] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[17] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-ViT: Unifying wavelet and transformers for visual representation learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 328–345.

[18] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[19] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.

[20] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," 2018, *arXiv:1805.08620*.

[21] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-D CNN for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] H. Fırat, M. E. Asker, M. Bayındır, and D. Hanbay, "Hybrid 3D/2D complete inception module and convolutional neural network for hyperspectral remote sensing image classification," *Neural Process. Lett.*, vol. 55, no. 2, pp. 1087–1130, Apr. 2023.

[23] X. Zhang, "Improved three-dimensional inception networks for hyperspectral remote sensing image classification," *IEEE Access*, vol. 11, pp. 32648–32658, 2023.

[24] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2647–2650.

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

[26] S. Ghaderizadeh, D. Abbasi-Moghadam, A. Sharifi, N. Zhao, and A. Tariq, "Hyperspectral image classification using a hybrid 3D-2D convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7570–7588, 2021.

[27] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[28] W. He, W. Huang, S. Liao, Z. Xu, and J. Yan, "CSiT: A multiscale vision transformer for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9266–9277, 2022.

[29] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.