HSIFORMER: AN EFFICIENT VISION TRANSFORMER FRAMEWORK FOR ENHANCED HYPERSPECTRAL IMAGE CLASSIFICATION USING LOCAL WINDOW ATTENTION

Mohammed Q. Alkhatib [1*], *Ali Jamali* [2]

[1] College of Engineering and IT, University of Dubai, Dubai, UAE
[2] Department of Geography, Simon Fraser University, Burnaby, Canada
[*] mqalkhatib@ieee.org

ABSTRACT

Convolutional neural networks (CNNs) have recently gained significant attention in image classification due to their exceptional performance in computer vision. Building on this success, researchers are now investigating the potential of transformers in Earth observation applications. However, transformers face a significant challenge: they require substantially more training data than CNN classifiers. This makes their application in remote sensing, particularly with Hyperspectral Image (HSI) data, difficult due to the limited availability of labeled data. In this paper, we will repurpose the PolSARFormer model for hyperspectral image classification. Originally designed for polarized SAR image classification, the model's initial parameters have been fine-tuned to better meet the requirements of hyperspectral data. The PolSARFormer model employs a vision transformer (ViT)-based framework that utilizes 3D and 2D CNNs as feature extractors and incorporates local window attention (LWA) for effective HSI data classification. Extensive experimental results show that the model, HSIFormer, achieves better classification accuracy than the state-of-the-art Swin Transformer and ViT algorithms. HSIFormer outperformed the Swin Transformer and ViT by 2.31% and 3.24% in overall accuracy (OA) on the Pavia University benchmark dataset. Additionally, results on the Salinas dataset demonstrated that HSIFormer surpasses several other algorithms, including HybridSN (96.89%), Tri-CNN (97.05%), Vision Transformer (94.68%), 3D-CNN (96.77%), and Swin Transformer (95.62%), with a kappa index (KI) of 98.23%. The code will be made publicly available at https://https://github.com/mqalkhatib/HSIFormer

***Index Terms*—** HSI classification, Attention mechanism, convolutional neural networks (CNNs), local window attention (LWA), Vision Transformers (ViT)

1. INTRODUCTION

Hyperspectral image (HSI) data has been available since the 1980's [1]. It can provide massive amounts of spectral and spatial information in hundreds of narrow contiguous spectral bands ranging from visible to infrared wavelengths, which in turn makes challenging fine-grained remote sensing tasks possible. At present, HSI classification has become an interesting topic in the field of hyperspectral remote sensing, as HSI is used in a wide variety of earth observation applications, such as land cover mapping and environmental monitoring. HSI classification accuracy is an important criteria for such applications. In order to achieve precise and accurate classification results, it is essential to obtain effective spatial and spectral features.

With the rapid development in Deep Learning (DL) technology, researchers in the field of computer vision began to use DL approaches, particularly Deep Convolutional Neural Networks (DCNNs) for classifying HSI features as they have shown superiority over other traditional classification methods [2, 3]. The work of [4] presented 1D-CNN method to extract the spectral features. However, spectral information is not enough to get accurate classification results. To tackle this problem, researchers in [5] presented 2D-CNN approach to learn the spatial information for HSI classification. However, these methods did not exploit the full advantage of the 3D nature of HSI cube. Thus, researchers in [6] proposed a 3D-CNN approach in which both spectral and spatial information are utilized to boost the classification performance. Roy et al. developed Hybrid Spectral CNN (HybridSN) [7], by fusing 2D-CNN with 3D-CNN, where 3D-CNN facilitates the joint spatial–spectral feature representation at the early stage, 2D-CNN is used for extracting more abstract-level spatial representation. The literature is rich with other similar examples [2, 8].

Given the substantial success of transformer models in language processing, researchers are now exploring their potential in computer vision and Earth observation [9, 10]. These advanced models have recently demonstrated effectiveness across a range of applications, including hyperspectral imagery analysis [11]. However, a significant challenge with transformers is their requirement for more training data compared to CNNs. This makes their application in remote sensing, especially in hyperspectral scenarios with limited labeled

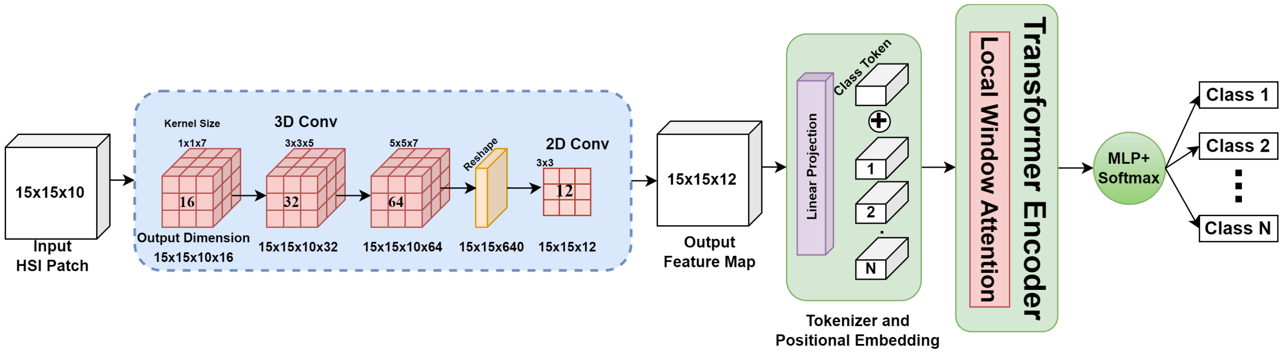

Fig. 1: Overall Architecture of the model used for HSI classification.

data, quite challenging. To tackle this issue, we have adapted the model introduced by [12], which features an efficient vision transformer (ViT) utilizing neighborhood attention for accurate classification of Polarized Synthetic Aperture Radar (PolSAR) images. Our objective is to develop a ViT model that achieves high precision in classifying hyperspectral data.

In this paper, we adapt the PolSARFormer [12] model, originally designed for PolSAR image classification, to hyperspectral classification. This model combines CNNs and Vision Transformers (ViTs) to deliver precise hyperspectral image classification. It utilizes local window attention (LWA) instead of the more computationally intensive self-attention mechanism, enhancing feature generalization in local regions and substantially reducing computational costs compared to traditional ViTs. Additionally, the model is evaluated on two benchmark hyperspectral image datasets: Pavia University (PU) and Salinas (SA).

The rest of the paper is organized as follows: Section 2 explains the architecture and building blocks of the model used in the paper, experimental results and comparisons against state-of-the-art models are discussed in Section 3, and finally, Section 4 summarizes the paper and states the future direction of this research.

2. NETWORK ARCHITECTURE

CNNs have already been established as high-level feature extractors and have been successfully applied to numerous computer vision tasks. In this section, we present HSIFormer, a robust, reliable, and scalable hierarchical ViT-based encoder network designed for Hyperspectral imagery classification. We explore a flexible and straightforward attention mechanism called the local attention transformer (LAT) to effectively utilize extracted features for classification. Therefore, the network consists of two key components: the feature extractor and the self-attention mechanism. Figure 1 illustrates the framework for Hyperspectral image classification.

2.1. Feature Extraction using 3D-2D CNN

A hyperspectral image can be represented as $X \in \mathbb{R}^{W\times H\times D}$, containing two spatial dimensions, width W and height H, along with a spectral dimension D. All pixels within the region of interest are classified into c land-cover classes, denoted by $Y = (y_1, y_2, \ldots, y_c)$. First, the hyperspectral image cube will undergo dimensionality reduction using Principal Component Analysis (PCA), reducing the spectral dimension from D to N, where N $\ll$ D. This will create a new cube with a reduced spectral dimension X_{Reduced}. Class-wise land-cover regions of size 15×15 are sampled from the reduced hyperspectral data X_{Reduced} to create the training and validation datasets. To leverage the capabilities of CNNs as feature extractors, we employed a hierarchical architecture combining 3D and 2D CNNs as the backbone network [7]. This approach aims to utilize 3D Convolution to extract both spectral and spatial features, while 2D Convolution refines the prominent spatial features among the spectral data, ensuring the backbone feature extractor remains computationally efficient. The feature extractor consists of three 3D convolutional layers with 16, 32, and 64 kernels of different scales as shown in Figure 1, followed by a 2D convolutional layer with 12 kernels of size 3×3. To establish long-range dependencies among the extracted feature maps, we employ a simple attention mechanism, LWA, which effectively localizes each query's receptive field to its nearest neighboring pixels within a local window.

2.2. Local Window Attention

It is considered as a localized self-attention mechanism that integrates inductive biases similar to convolution-like operations, removing the need for additional overheads like pixel shifts found in advanced ViT models such as the Swin Transformer [13]. LWA confines the receptive field of each query token to fixed-sized neighboring pixels. The goal of LWA is to establish a local neighborhood window; smaller neighboring regions receive more concentrated local attention, while larger neighboring regions receive wider global attention [14]. Consequently, the LWA mechanism more effectively controls the receptive fields while maintaining a balance between translational invariance and equivariance properties compared to other ViTs.

It is important to note that self-attention allows each token to interact with all other tokens, while LWA limits each token's receptive field to its surrounding area. Consequently, LWA has the advantage of directly restricting each pixel to its neighboring area without additional computational costs, eliminating the need for pixel transitions to incorporate cross-window interactions. Furthermore, unlike window attention, LWA is not restricted by the window size of the input. If the size of the pixel's neighborhood is greater than or equal to the size of the feature map, self-attention and neighborhood attention will yield results similar to those of the input map.

2.3. HSIFormer

In the developed model, the spectral and spatial features from the hyperspectral data will be fed into the feature extractor as described previously. The input size of the hyperspectral data will remain unchanged throughout the feature extraction process. The output from the feature extractor will then be processed by the LWA. The LWA integrates the feature extractor's output with two consecutive 3×3 convolutional layers using strides of 2×2, resulting in a spatial size one-fourth of the original hyperspectral data input. Unlike non-overlapping convolutions, the LWA employs overlapping convolutions. The model is structured into two levels: the first level contains three LWA blocks, while the second level includes four blocks. Similar to the Swin Transformer, multiple levels with varying numbers of LWA blocks can be implemented. Notably, the output from each level is passed to the subsequent level.

3. EXPERIMENTS AND ANALYSIS

To demonstrate the performance of the proposed approach shown in Figure 1, the proposed methodology is compared with 3D-CNN [6], HybridSN [7], Tri-CNN [2], Vision Transformer [15] and Swin Transfomer [13]. The Overall Accuracy (OA), the Average Accuracy (AA), and the Kappa statistic (Kappa) are reported to evaluate the performance of the model. The classification accuracy of each class is also provided. The experiments were conducted and repeated 10 times. For all algorithms, only the classification results with the highest accuracy in 10 trials are recorded. For OA, AA, and Kappa the average and standard deviation of all 10 trials are recorded. The model is evaluated using two widely used hyperspectral datasets: Pavia University (PU) and Salinas (SA). Figure 2 shows the reference data for both datasets. Full description of both datasets is available in [16]

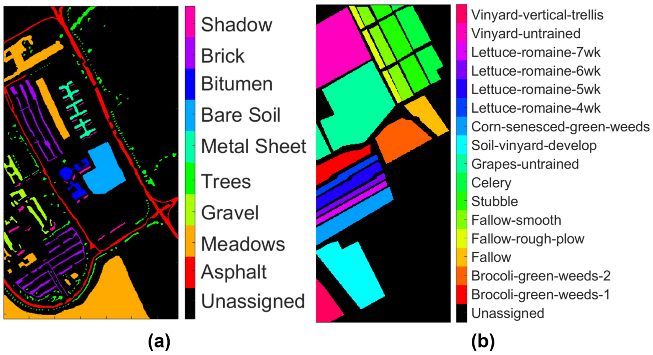

Fig. 2: Reference Data: (a) PU; (b) SA

For both datasets, image patches were randomly divided into 1% for training, and the remaining 99% are used for testing and evaluation, we used a patch size of 15×15 and the number of principal components was set to 10, multiple attempts were made through trials to discover the most suitable values, leading to the determination of these optimal choices. The model is trained for 100 epochs with batch size of 128. During the model training, early stopping strategy is adopted. Specifically, if the model's performance did not improve over 10 consecutive epochs, the training process was terminated and the model was restored to its best weights. The optimization algorithm is Adam. The learning rate is set to 5×10^{-4}. All models are implemented using Python, Keras framework that runs with TensorFlow back-end. To ensure a fair comparison, all models were trained under the same conditions.

Table 1 list the classification accuracies for PU dataset. It can be seen that the proposed model can achieve superior results than other classification models. the table reveals that the HSIFormer model outperforms several existing algorithms, including 3D-CNN (92.87%), HybridSN (94.40%), Tri-CNN (94.68%), ViT (90.90%), and Swin Transformer (92.12%), achieving a notable kappa of 95.21%. The HSIFormer model surpasses the Swin Transformer by 2.31% in overall accuracy, demonstrating its advanced capability in hyperspectral image classification. Besides, the results show less variability or fluctuation compared to other models. This indicates that the used model is performing well and producing reliable and accurate predictions.

Table 2 presents the classification results, while Figure 4 shows the visual classification maps for the SA dataset. The results indicate that the HSIFormer model achieves an average accuracy (AA) of 98.51%, outperforming the 3D-CNN (98.01%), HybridSN (97.93%), Tri-CNN (96.95%), Vision Transformer (97.05%), and Swin Transformer (97.38%). Despite being a leading ViT model, the Swin Transformer shows lower classification accuracy due to its higher requirement for training data compared to CNN classifiers. In contrast, the HSIFormer model demonstrates significantly improved clas-

Table 1: Classification performance of different methods for the PU dataset. Bold indicates the best result

Class	3DCNN	HybridSN	Tri-CNN	ViT	SWIN	HSIFormer
Asphalt	97.26	98.82	**98.87**	93.30	91.15	96.58
Meadows	97.99	99.42	97.60	98.49	99.52	**99.73**
Gravel	**88.52**	78.23	83.04	88.18	63.03	86.94
Trees	86.46	88.28	**93.54**	81.95	92.07	92.89
Metal Sheet	99.41	99.05	98.85	91.60	93.38	**99.55**
Bare Soil	98.33	95.70	90.86	94.25	94.83	**99.60**
Bitumen	79.25	96.84	**99.32**	97.44	87.14	88.41
Brick	82.40	92.56	83.87	82.16	92.83	**94.10**
Shadow	92.61	64.73	93.67	67.69	**94.61**	85.63
OA (%)	94.62±1.93	95.81±0.50	96.00±0.74	93.15±1.19	94.08±0.91	**96.39±0.40**
AA (%)	91.63±3.29	90.50±2.51	93.23±2.25	88.34±2.07	89.49±1.43	**93.71±0.42**
Kappa x 100	92.87±2.60	94.40±2.05	94.68±2.11	90.90±1.58	92.12±1.23	**95.21±0.54**

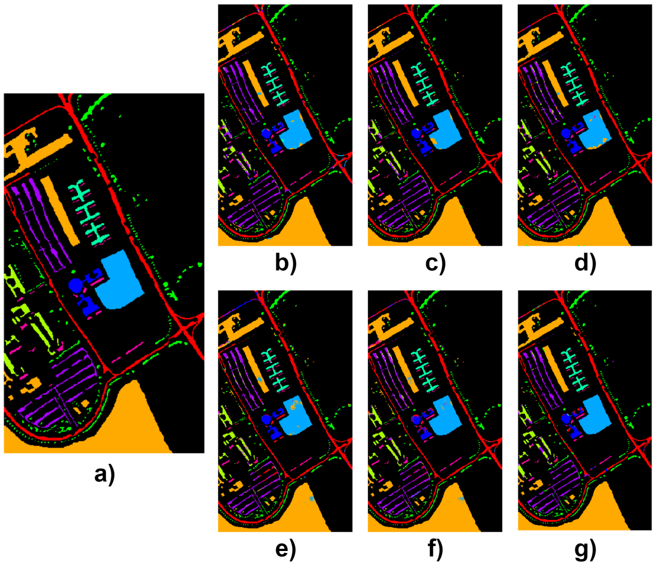

Fig. 3: Classification maps of PU Dataset. (a) Reference Data; (b) 3D-CNN; (c) HybridSN; (d) Tri-CNN (e) ViT; (f) SWIN; (g) HSIFormer

sification accuracy for hyperspectral imagery, surpassing the Swin Transformer by 1.13% in terms of average accuracy.

To conduct the ablation study, we evaluated the data on CNN only and Transformer network. The outcomes revealed that the CNN stream achieved an accuracy of 95.81% for PU and 97.21% for Salinas. On the other hand, the Transformer achieved scores of 93.15% for Pavia and 95.21% for Salinas. However, when both parts are combined together together, the model's performance improved, resulting in an increase in accuracies of more than 1% for both datasets, as reported in Tables 1 and 2.

4. CONCLUSIONS

This paper introduces a ViT-based framework for hyperspectral image classification that utilizes local window attention (LWA) to enhance local feature representation while significantly reducing annotation costs and hardware requirements. Results from two hyperspectral benchmark datasets reveal that the model, HSIFormer, outperforms the current state-of-the-art ViT models, including the Swin Transformer and others. On the Pavia University (PU) benchmark dataset, HSIFormer exceeds the Swin Transformer by a margin of 2.31% in overall accuracy (OA). Additionally, on the Salinas dataset, HSIFormer achieves an average accuracy (AA) of 98.51%, surpassing several other algorithms such as 3D-CNN (98.01%), HybridSN (97.93%), Tri-CNN (96.95%), Vision Transformer (97.05%), and Swin Transformer (97.38%).

For Future Work, we will investigate the integration of other advanced deep learning techniques, such as self-supervised learning or generative models, to further improve classification performance and feature extraction.

Table 2: Classification performance of different methods for the SA dataset. Bold indicates the best result

Class	3DCNN	HybridSN	Tri-CNN	ViT	SWIN	HSIFormer
Brocoli-green-weeds-1	99.3	98.26	**100.00**	99.95	**100.00**	99.50
Brocoli-green-weeds-2	**100.00**	**100.00**	99.62	**100.00**	**100.00**	**100.00**
Fallow	98.28	**100.00**	**100.00**	97.47	99.49	**100.00**
Fallow-rough-plow	98.71	99.28	99.36	95.12	99.00	**99.50**
Fallow-smooth	99.74	99.85	99.81	97.35	**100.00**	99.74
Stubble	99.56	99.85	99.35	99.67	99.95	**100.00**
Celery	99.63	99.77	99.33	99.58	99.26	**99.86**
Grapes-untrained	95.44	89.59	98.03	85.92	97.23	**99.65**
Soil-vinyard-develop	99.94	**100.00**	**100.00**	99.97	**100.00**	**100.00**
Corn-senesced-green-weeds	96.34	98.93	97.68	96.52	94.54	**99.21**
Lettuce-romaine-4wk	97.47	89.98	98.22	98.44	98.41	**98.78**
Lettuce-romaine-5wk	**100.00**	**100.00**	**100.00**	91.49	99.12	**100.00**
Lettuce-romaine-6wk	95.96	96.72	76.97	**98.58**	98.14	91.92
Lettuce-romaine-7wk	**98.04**	95.89	92.52	97.94	96.92	97.57
Vinyard-untrained	89.56	**99.24**	89.28	93.42	80.28	91.18
Vinyard-vertical-trellis	99.67	98.73	**100.00**	**100.00**	95.07	99.28
OA (%)	97.10±1.26	97.21±0.68	97.35±1.20	95.21±1.10	96.07±1.07	**98.41±0.48**
AA (%)	98.01±1.76	97.93±1.13	96.95±0.72	97.05±1.77	97.38±0.76	**98.51±0.97**
Kappa x 100	96.77±1.41	96.89±0.75	97.05±1.33	94.68±1.22	95.62±1.18	**98.23±0.54**

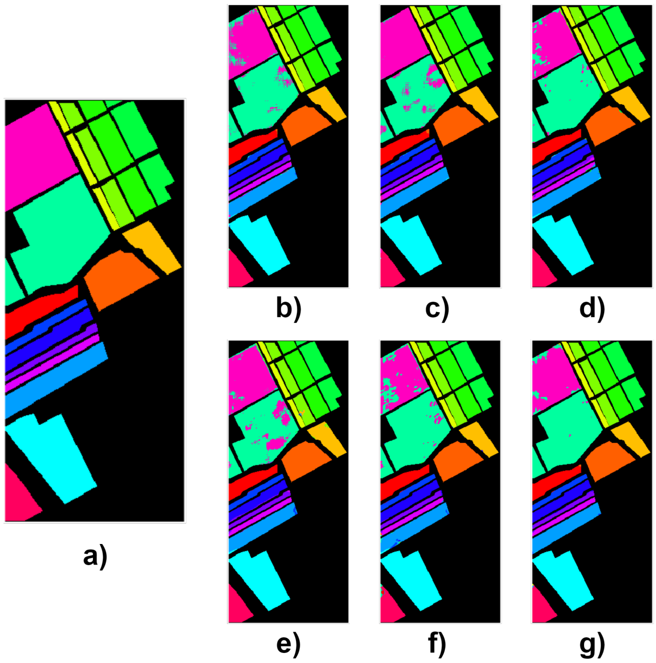

Fig. 4: Classification maps of SA Dataset. (a) Reference Data; (b) 3D-CNN; (c) HybridSN; (d) Tri-CNN (e) ViT; (f) SWIN; (g) HSIFormer

5. REFERENCES

[1] Shenming Qu, Xiang Li, and Zhihua Gan, “A review of hyperspectral image classification based on joint spatial-spectral features,” in *Journal of Physics: Conference Series*. IOP Publishing, 2022, vol. 2203, p. 012040.

[2] Mohammed Q Alkhatib, Mina Al-Saad, Nour Aburaed, Saeed Almansoori, Jaime Zabalza, Stephen Marshall, and Hussain Al-Ahmad, “Tri-cnn: a three branch model for hyperspectral image classification,” *Remote Sensing*, vol. 15, no. 2, pp. 316, 2023.

[3] Muhammad Ahmad, Adil Mehmood Khan, Manuel Mazzara, Salvatore Distefano, Mohsin Ali, and Muhammad Shahzad Sarfraz, “A fast and compact 3-d cnn for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[4] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, “Deep convolutional neural networks for hyperspectral image classification,” *Journal of Sensors*, vol. 2015, pp. 1–12, 2015.

[5] Konstantinos Makantasis, Konstantinos Karantzalos, Anastasios Doulamis, and Nikolaos Doulamis, “Deep supervised learning for hyperspectral data classification through convolutional neural networks,” in *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, 2015, pp. 4959–4962.

[6] Amina Ben Hamida, Alexandre Benoit, Patrick Lambert, and Chokri Ben Amar, “3-d deep learning approach for remote sensing image classification,” *IEEE Transactions on geoscience and remote sensing*, vol. 56, no. 8, pp. 4420–4434, 2018.

[7] Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, and Bidyut B Chaudhuri, “Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2019.

[8] Chunyan Yu, Rui Han, Meiping Song, Caiyu Liu, and Chein-I Chang, “A simplified 2d-3d cnn architecture for hyperspectral image classification based on spatial–spectral fusion,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2485–2501, 2020.

[9] Pengyuan Lv, Wenjun Wu, Yanfei Zhong, Fang Du, and Liangpei Zhang, “Scvit: A spatial-channel feature preserving vision transformer for remote sensing image scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[10] Durong Cai and Peng Zhang, “T^{3}SR: Texture transfer transformer for remote sensing image superresolution,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7346–7358, 2022.

[11] Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot, “Multimodal fusion transformer for remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[12] Ali Jamali, Swalpa Kumar Roy, Avik Bhattacharya, and Pedram Ghamisi, “Local window attention transformer for polarimetric sar image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[14] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi, “Neighborhood attention transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6185–6194.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[16] “Hyperspectral remote sensing scenes,” https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes, Accessed: 14-July-2024.