HSRNet: Hyperspectral Retentive Network for Hyperspectral Image Classification

Junbo Zhou
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
imjbzhou@my.swjtu.edu.cn

Guoqiang Gao *
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
xnjdggq@163.com

Chengyu Zhang
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
zcy991115@my.swjtu.edu.cn

Yujun Guo
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
yjguo@swjtu.edu.cn

Xueqin Zhang
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
xq_zhang@swjtu.edu.cn

Song Xiao
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
xiaosong@home.swjtu.edu.cn

Guangning Wu
Southwest Jiaotong University
School of Electrical Engineering
Chengdu, Sichuan, China
hv-swjtu@163.com

***Abstract*—Hyperspectral image (HSI) classification faces a persistent challenge: modeling long-range semantic context while preserving fine-grained local structures. Existing paradigms often struggle to balance these needs; Convolutional Neural Networks (CNNs) are limited by local receptive fields, whereas Transformer-style attention can mix irrelevant regions and overlook the sequential nature of spectral signatures. Moreover, treating HSI as a unified volume often blurs the distinct roles of the spectral and spatial axes. To address these issues, this paper proposes the Hyperspectral Retention Network (HSRNet), which factorizes feature learning into two stages: a spectral retention encoder that captures continuous band correlations in a pixel-wise spectral token sequence, followed by a spatial retention encoder that aggregates geometric context over the flattened spatial tokens. A distance-decayed retention operator with head-wise learnable decay factors controls the extent of contextual mixing, allowing different heads to emphasize short-range interactions or preserve per-token details, while a lightweight local context enhancement module injects explicit 2D neighborhood cues. Experiments on the Pavia University and Houston 2013 datasets demonstrate the effectiveness of HSRNet, achieving Overall Accuracies of 98.9% and 98.3%, respectively, and producing more coherent classification maps than recent state-of-the-art baselines.**

***Keywords*—Hyperspectral Image Classification, Retentive Network, Dual-Axis Factorization, Spectral-Spatial Feature Extraction**

I. INTRODUCTION

Unlike traditional imaging systems that rely on broad spectral bands, hyperspectral imaging (HSI) captures scene information across hundreds of narrow, contiguous channels [1]. This process generates a volumetric "data cube" that assigns a unique spectral fingerprint to every pixel, enabling the precise discrimination required for applications such as precision agriculture and environmental monitoring [2], [3], [4]. Consequently, significant research has focused on HSI classification to accurately map land-cover labels based on spectral-spatial signatures [5]. However, HSI's high dimensionality often leads to the "curse of dimensionality," where significant data redundancy and strong inter-band correlations complicate discriminative feature extraction. Early research addressed these challenges through manual feature engineering, utilizing classifiers like Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors as primary baselines [6], [7], [8]. While these methods were eventually enhanced by integrating spatial context to improve stability, they remain limited by their reliance on shallow, hand-crafted features. Such descriptors require extensive domain expertise and often generalize poorly to complex or spatially heterogeneous scenes, necessitating more robust representation techniques.

The shift toward deep learning has redefined HSI classification by replacing manual descriptors with hierarchical representation learning. Convolutional Neural Networks (CNNs) initially led this transition, evolving into 3D and hybrid architectures to jointly encode spectral-spatial textures [9], [10]. Despite their efficiency in capturing local features, CNNs are hindered by restricted receptive fields that limit their ability to model long-range dependencies. In response, Vision Transformers (ViTs) have gained prominence for their ability to facilitate global interaction through self-attention [11]. Contemporary research often explores hybrid systems, such as SSFTT or Swin-based models, to combine the inductive biases of CNNs with the global reasoning of Transformers [12], [13]. Nevertheless, achieving an optimal balance between local granularity and global context remains a persistent challenge [14].

Despite the advancements in hybrid architecture, fundamentally optimizing the modeling mechanism remains an open challenge, largely due to two structural limitations in contemporary designs. First, existing backbones struggle to flexibly balance local granularity and global coherence. While CNNs are constrained by rigid, static windows, standard Transformers adopt an isotropic attention pattern that tends to correlate all tokens uniformly. This weak selectivity can introduce “attention drift,” where irrelevant background cues are mixed into the representation of homogeneous regions, thereby undermining spatial continuity. Second, treating HSI data as a unified 3D volume or a simple multi-channel input is

This work is supported in part by National Natural Science Foundation of Sichuan Province of China (2024NSFTD0009). (*Corresponding Author)

often overly simplistic. The spectral axis forms a continuous sequential signature, whereas the spatial axis describes 2D geometric structure. Modeling these two domains in a single coupled operation may weaken diagnostic cues and limit effective feature learning.

To address these limitations, we are inspired by RetNet [15], and propose the Hyperspectral Retention Network (HSRNet), a novel architecture designed to introduce both elasticity and specificity into the feature learning process.

(1) We propose a distance-aware retention mechanism with learnable decay rates to replace fixed kernels and indiscriminate attention. By applying distance-dependent weighting across 1D spectral and 2D spatial dimensions, the model develops head-specialized receptive fields. This allows specific heads to focus on near-field details or broader context, effectively preserving sharp boundaries while maintaining intra-class consistency.

(2) We introduce a dual-axis factorization framework that decouples feature extraction into sequential spectral and geometric spatial stages. This approach respects the unique properties of each domain and minimizes cross-axis interference. By processing these axes independently, the network produces more stable and interpretable representations than traditional coupled modeling techniques.

The remainder of this paper is organized as follows. Section II details the proposed HSRNet framework. Section III presents the experimental results and discussion. Finally, Section IV concludes the paper.

II. Hyperspectral Retention Network

The proposed Hyperspectral Retention Network (HSRNet) is designed to effectively classify hyperspectral imagery by decoupling spectral correlations and spatial dependencies. The architecture is shown in Figure 1.

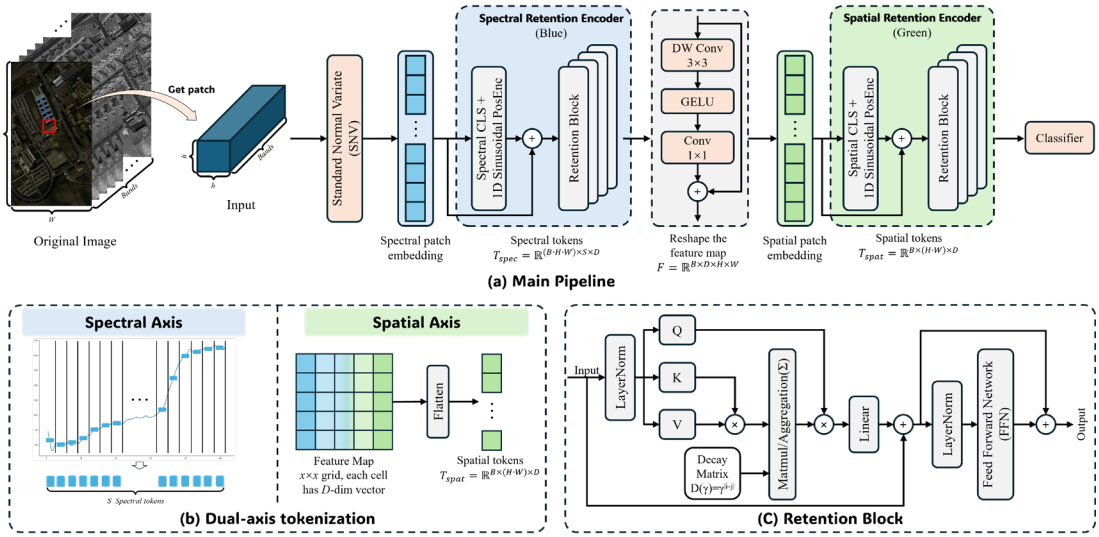

Fig. 1. The structure of HSRNet

A. Preprocessing and Spectral Tokenization

Let $\boldsymbol{X}_{raw} \in \mathbb{R}^{C\times H\times W}$ denote the input hyperspectral data cube, where C is the number of spectral bands, and H, W represent the spatial dimensions.

To mitigate the spectral variability caused by scattering effects and baseline shifts, we first apply pixel-wise SNV. For a given pixel vector $\boldsymbol{x} \in \mathbb{R}^{C}$, the normalized vector $\hat{\boldsymbol{x}}$ is computed as:

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}-\mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})+\epsilon} \tag{1}$$

where $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ are the mean and standard deviation of the spectral intensities for that pixel.

Hyperspectral data contains high redundancy in adjacent bands. Instead of treating every band as an independent token, we employ a grouping strategy. The spectral dimension C is partitioned into segments of size P. These segments are linearly projected to an embedding dimension D, resulting in a sequence of spectral tokens $\boldsymbol{T}_{spec} \in \mathbb{R}^{S\times D}$, where $S = \lceil C/P \rceil$.

Since the retention mechanism is permutation-invariant, we inject positional information using standard sinusoidal encodings. For a position p and dimension i, the encoding is defined as:

$$PE_{(p,2i)} = sin(p/10000^{2i/D}), \quad PE_{(p,2i+1)} = cos(p/10000^{2i/D}) \tag{2}$$

This is added element-wise to both spectral and spatial tokens before encoding.

B. Dual-Axis Retention Mechanism

The core novelty of HSRNet lies in the Dual-Axis Retention framework, which factorizes the 3D context modeling into two orthogonal steps using a unified retention mechanism.

1) The Generalized Retention Operator

We adopt a parallel retention formulation with a spatial decay mask. Let $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{L\times D}$ be the query, key, and value matrices derived from the input sequence via linear projections. The retention output is computed as:

$$\text{Retention}(\mathbf{X}) = \mathbf{Q} \odot (\mathbf{D}(\mathbf{K} \odot \mathbf{V})) \tag{3}$$

where $\odot$ denotes the Hadamard product. The matrix $\boldsymbol{D} \in \mathbb{R}^{L\times L}$ introduces geometric decay to enforce locality and relative position awareness. For the h-th head, the decay mask is parameterized by a scalar $\gamma_h \in (0,1)$:

$$\boldsymbol{D}_{nm} = \gamma_h^{|n-m|} \tag{4}$$

To capture multi-scale features, we utilize a Decay Mixture strategy. Different heads are assigned distinct γ values. Lower γ values focus on high-frequency local variations, while higher γ values (or $\gamma = 1$) retain global low-frequency context.

The output is processed by a Gated Feed-Forward Network (FFN) with GELU activation and Layer Normalization (LN) to form the final block output:

$$\boldsymbol{Y} = \mathrm{FFN}(\mathrm{LN}(\mathbf{X} + \mathrm{Retention}(\mathrm{LN}(\mathbf{X})))) \tag{5}$$

2) Axis-Specific Encoding

We first apply the retention encoder along the spectral dimension. The sequence $\boldsymbol{T}_{spec}$ is augmented with a learnable class token $[CLS]_{spec}$. The encoder aggregates intra-pixel spectral correlations into this token, producing a compact pixel-wise feature representation $\boldsymbol{F}_{pixel} \in \mathbb{R}^{HW\times D}$.

Subsequently, we model the spatial context. The feature map is flattened into spatial tokens $\boldsymbol{T}_{spatial} \in \mathbb{R}^{HW\times D}$. A spatial class token $[CLS]_{spatial}$ is appended. The retention mechanism is then applied to this sequence. Although flattened, the 1D decay mask $\boldsymbol{D}$ effectively approximates 2D spatial distances, allowing each pixel to attend to global context with distance-aware weighting.

C. Local Context Enhancement and Classification

Before the global spatial modeling described above, we insert a Local Context Enhancement (LCE) module to capture fine-grained local patterns that might be overlooked by the global retention mechanism.

The LCE is formulated using a depthwise separable convolution structure on the 2D feature map. For input features $\boldsymbol{Z}$:

$$\boldsymbol{Z}' = \boldsymbol{Z} + \mathrm{Conv}_{1\times1}(\sigma(\mathrm{DWConv}_{3\times3}(\boldsymbol{Z}))) \tag{6}$$

This inverted bottleneck design enhances local spatial features with minimal computational overhead.

Finally, for classification, we utilize a Normed Linear Head to mitigate the dominance of easy samples and improve convergence. Both the final feature embedding $\boldsymbol{z}$ (from $[CLS]_{spatial}$) and the classifier weights $\boldsymbol{W}$ are L_2-normalized before the dot product:

$$\boldsymbol{logits} = s \cdot \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2} \cdot \frac{\boldsymbol{W}^T}{\|\boldsymbol{W}\|_2} \tag{7}$$

where s is a learnable scaling factor.

III. Experiments and Discussions

A. Baselines and Implementation

To rigorously assess the effectiveness of the proposed framework, we benchmark it against six diverse comparison methods. These baselines were selected to represent a broad spectrum of HSI classification strategies, ranging from traditional shallow learning classifiers to advanced deep learning architectures, including pure CNNs, hybrid models, and Transformer-based networks.

1) Support Vector Machine (SVM) [16]: Serving as a classical baseline, this method employs a Radial Basis Function (RBF) kernel to handle spectral non-linearity. It evaluates pixel-wise classification limits in the absence of spatial context.

2) Spatial Spectral Pyramid Network (SSPN) [17]: A representative 3D-CNN that utilizes a spatial-spectral pyramid network. It extracts hierarchical features across multiple scales, serving as a benchmark for modeling volumetric dependencies in hyperspectral data.

3) Double Branch Convolution-Transformer Network (DBCTNet) [18]: A hybrid framework merging a dual-branch CNN for local extraction with a Transformer for global modeling. It tests the efficacy of combining convolutional inductive bias with long-range attention mechanisms.

4) Dual Selective Fusion Transformer Network (DSFormer) [19]: A state-of-the-art pure Transformer utilizing a dual selective fusion module. It dynamically integrates spatial and spectral tokens, providing a rigorous comparison against architectures relying solely on self-attention mechanisms.

5) Mixed CNN–Transformer with Graph Contrastive Learning (MCTGCL) [20]: A lightweight graph contrastive learning network designed for efficient feature aggregation. Its dual-branch structure captures structural relationships with low computational complexity, offering a distinct perspective from heavy deep learning models.

6) Deformable Convolution-Enhanced Hierarchical Transformer With Spectral-Spatial Cluster Attention (SClusterFormer) [14]: This method serves as a highly complex baseline integrating deformable convolutions with a specialized cluster attention mechanism. SClusterFormer focuses on capturing fine-grained details by utilizing spectral angle and Euclidean distance metrics to differentiate between local homogeneous regions and non-local structures.

The proposed method and all comparative baselines were implemented using the PyTorch framework. All experimental trials were conducted on a high-performance workstation powered by an NVIDIA GeForce RTX 4090 GPU. The software environment consisted of Python 3.9 and PyTorch 2.2.1, accelerated by CUDA 11.8.

B. The results of all models

We used two normally used datasets. The quantitative assessment in Tables I and II reveals that while competitive baselines such as DSFormer and DBCTNet already operate in a high-accuracy regime, distinctive disparities remain in how these architectures handle spectral ambiguity and structural heterogeneity at the class level. On the Pavia University dataset, the primary challenge involves differentiating spectrally correlated materials, a task where the spectral-only SVM baseline demonstrates limited effectiveness by yielding 0% accuracy on Bitumen and roughly 4.4% on Gravel. Although SSPN, DBCTNet, DSFormer, and MCTGCL substantially alleviate this issue, their behavior is not fully uniform: for example, DSFormer and MCTGCL show a more visible drop on Shadows than DBCTNet, suggesting that preserving spatial smoothness in low-texture regions remains nontrivial even when overall metrics are high.

TABLE I

THE ACCURACY OF ALL MODELS ON PAVIA UNIVERSITY DATASET

Class Name	SVM	SSPN	DBCTNet	DSFormer	MCTGCL	SClusterFormer	HSRNet
Asphalt	83.7	99.1	99.7	**99.2**	99.3	73.0	99.6
Meadows	88.6	**99.2**	**99.2**	99.1	99.1	36.9	**99.2**
Gravel	4.4	98.2	98.9	95.2	98.6	80.4	**99.6**
Trees	88.9	97.7	97.4	96.7	96.4	82.7	**98.2**
Painted metal sheets	99.1	99.4	99.8	98.8	98.5	94.7	99.7
Bare soil	32.6	99.9	**100.0**	98.5	99.9	53.7	**100.0**
Bitumen	0.0	97.3	99.8	99.7	99.8	83.5	**100.0**
Self-blocking bricks	75.7	98.0	99.4	98.9	98.7	91.1	99.1
Shadows	**100.0**	97.9	98.8	94.7	95.8	76.6	97.5
OA	79.4	98.5	98.4	98.2	98.5	62.6	**98.9**
Kappa	0.713	0.981	0.980	0.977	0.981	0.565	**0.986**

HSRNet attains the best Overall Accuracy of 98.9% and reaches perfect recognition on Bitumen and Bare soil, indicating that factorized spectral–spatial modeling can better protect subtle spectral cues from being diluted by geometric variations. Conversely, the Houston 2013 dataset presents a rigorous multiscale test where SClusterFormer exhibits marked performance degradation with an Overall Accuracy of 56.5%, including near-collapse on categories such as Highway and Railway, highlighting its vulnerability in cluttered urban layouts. In contrast, DSFormer performs strongly overall but is less stable on small, easily confused targets such as Water, while DBCTNet and MCTGCL show improved robustness on several structured categories yet still trail the best result in OA.

TABLE II

THE ACCURACY OF ALL MODELS ON HOUSTON 2013 DATASET

Class Name	SVM	SSPN	DBCTNet	DSFormer	MCTGCL	SClusterFormer	HSRNet
Healthy grass	89.8	97.6	94.7	**96.6**	95.9	39.1	96.6
Stressed grass	91.7	95.3	98.2	99.3	97.6	54.7	**100.0**
Synthetic grass	0.3	99.5	99.8	98.6	98.5	25.6	**99.8**
Trees	95.5	96.4	93.7	**97.3**	95.0	94.3	96.4
Soil	93.2	98.9	98.8	**100.0**	99.5	33.0	**100.0**
Water	91.2	96.5	99.0	90.8	93.8	21.8	**100.0**
Residential	56.1	95.8	**97.5**	96.5	95.4	86.0	95.3
Commercial	56.7	90.1	91.7	98.5	95.0	68.5	**99.2**
Road	63.0	92.0	84.9	**96.5**	89.6	72.6	96.2
Highway	44.5	97.5	93.9	97.9	95.1	0.0	**99.1**
Railway	49.3	97.1	96.0	97.8	90.3	1.2	**98.7**
Parking Lot 1	34.5	94.3	93.5	**99.4**	92.5	67.0	**99.4**
Parking Lot 2	1.4	92.1	91.7	91.3	92.2	79.5	**98.3**
Tennis Court	85.2	99.9	97.5	**100.0**	**100.0**	99.5	**100.0**
Running Track	99.1	96.2	**99.7**	98.5	98.9	77.7	99.0
OA	67.5	95.7	94.8	96.7	94.9	56.5	**98.3**
Kappa	0.648	0.954	0.944	0.965	0.946	0.534	**0.982**

HSRNet further improves the overall performance to 98.3% and maintains balanced precision across both compact objects and large land covers, which is consistent with the distance-decayed retention design that mitigates excessive context mixing while retaining long-range coherence. The classification maps of all models

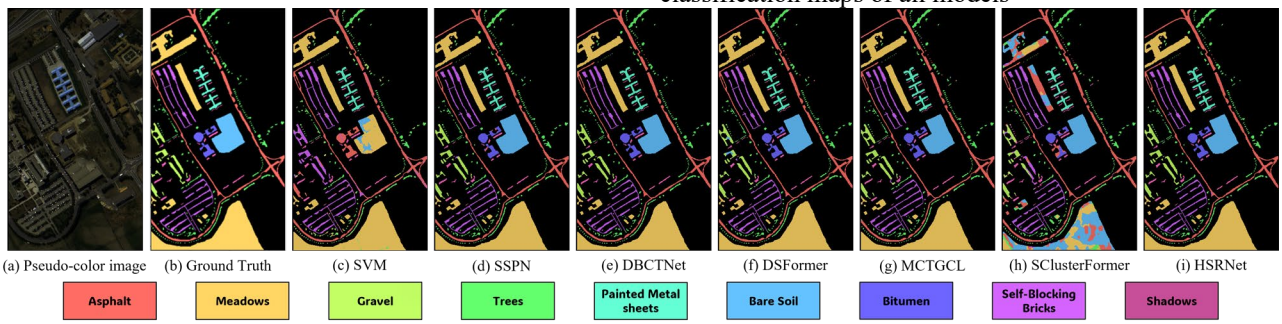

Fig. 2. The classification maps of all models on Pavia University Dataset

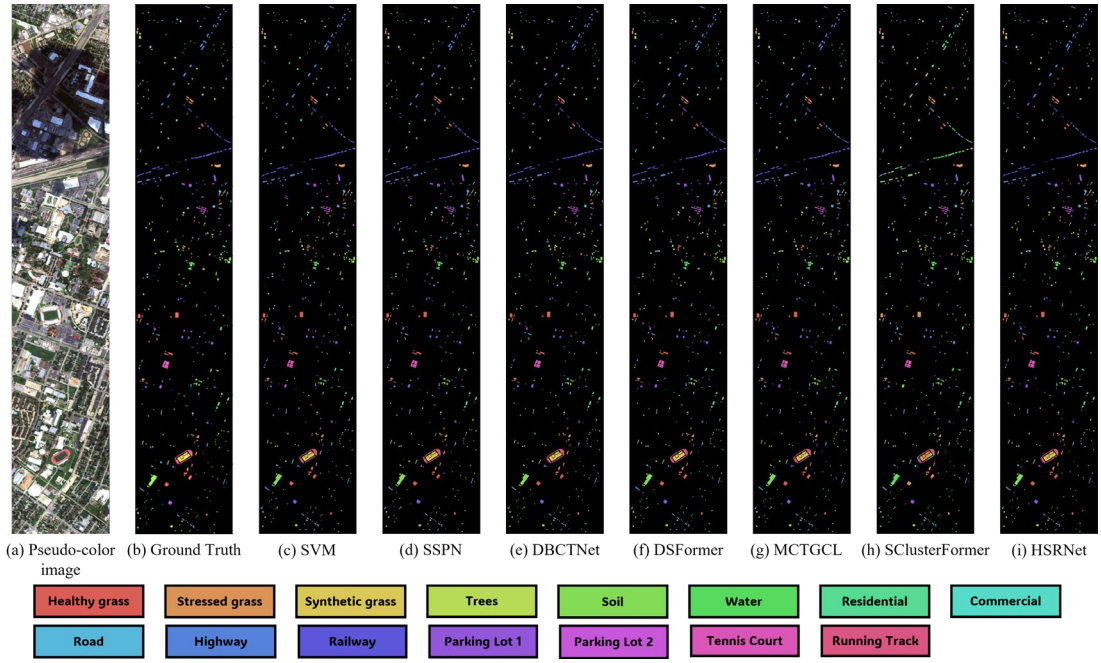

Fig. 3. The classification maps of all models on Pavia University Dataset

Figures 2 and 3 provide a qualitative visualization of the classification results for all comparison models. On the Pavia University dataset, the most noticeable weakness of the baseline pipeline appears as speckle-like misclassifications that

break the continuity of large homogeneous regions. This artifact is pronounced for SVM, where spectrally correlated classes such as Meadows, Gravel, and Bare Soil are easily interchanged, producing fragmented patches and occasional label flipping inside otherwise uniform parcels. Modern spatial–spectral networks, including SSPN, DBCTNet, DSFormer, and MCTGCL, substantially suppress this noise and recover clearer object extents, yet subtle differences remain: some predictions still show mild boundary smoothing around thin structures and sporadic confusion near high-contrast borders, where Shadows can trigger small isolated errors. SClusterFormer is visually less stable on this scene, with a conspicuous region-level mismatch in the lower part of the map that disrupts the dominant land-cover layout. Overall, HSRNet produces the cleanest and most coherent parcels, with sharper boundary transitions and fewer isolated outliers. The Houston 2013 dataset presents a different difficulty characterized by thin linear structures and small, scattered man-made objects.

In many comparative maps, roads and rail-related trajectories become discontinuous, and compact categories such as Parking Lot and Tennis Court are partially missed or broken into short segments. DSFormer and MCTGCL preserve the main corridors more reliably than earlier baselines, whereas SClusterFormer shows a clear topological distortion by bleeding non-road labels along the long highway-like structures. In contrast, HSRNet better maintains the linear continuity of transportation networks and retains the structural integrity of dense urban patterns, yielding a map that more closely matches the ground truth in both topology and region completeness.

C. *The visualization of feature embeddings*

To qualitatively inspect the representation learned by HSRNet, Fig. 4 shows a 3D t-SNE projection extracted from the spatial encoder.

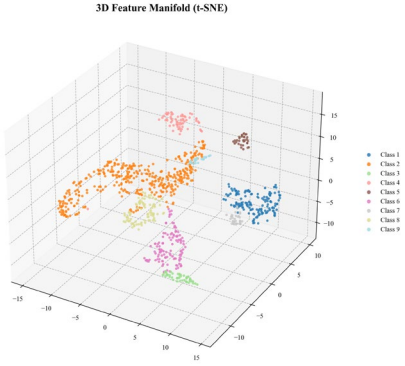

Fig. 4. The visualization of feature embeddings on Pavia University dataset

Despite the severe spectral mixing that characterizes raw HSI signatures, the learned CLS features display a structured distribution in which most classes form compact groups with limited overlap, while a few categories exhibit a more elongated spread, reflecting residual intra-class variability rather than random confusion.

This overall organization is consistent with the strong quantitative results at Pavia University and suggests that the model captures class-discriminative cues in a stable latent space. Such behavior aligns with our factorized pipeline: the spectral retention encoder first processes each pixel as a spectral token sequence and models long-range band dependencies with head-wise learnable decay, producing robust pixel features under spectral variability. These features are then enhanced by the LCE module, which injects genuine 2D neighborhood priors, before the spatial retention stage aggregates contextual evidence over the flattened spatial token sequence using learnable decay and identity-style heads to avoid excessive smoothing. Together, these components tend to tighten intra-class feature distributions while maintaining clear margins between classes, leading to more reliable high-level embeddings for classification.

D. *Visualizing Retention Patterns in HSRNet*

To interpret how HSRNet aggregates information along the two axes, Fig. 5 visualizes the normalized retention score maps in the spectral and spatial stages. In the spectral view, the retention surface shows a clear preference for nearby spectral tokens, as indicated by stronger responses along the diagonal, consistent with the continuity of hyperspectral signatures across adjacent bands. At the same time, several meaningful off-diagonal responses remain, indicating that the spectral retention stage also allows interactions among distant spectral tokens when they are supported by the learned feature similarity, rather than collapsing into a purely local filter. In the spatial view, the center-pixel map reveals a selective context pattern: the model assigns relatively higher scores to a subset of locations around the query pixel while suppressing others, suggesting that contextual aggregation is neither uniform nor indiscriminately global. This behavior matches our implementation, where spatial retention mixes tokens on the flattened grid sequence under head-wise learnable decay, while the LCE module injects genuine 2D neighborhood priors before spatial aggregation. Overall, these visualizations support the motivation for dual-axis factorization: spectral retention stabilizes band-sequence dependencies at the pixel level, and spatial retention consolidates context through restrained mixing, thereby preserving regional consistency without oversmoothing fine structures.

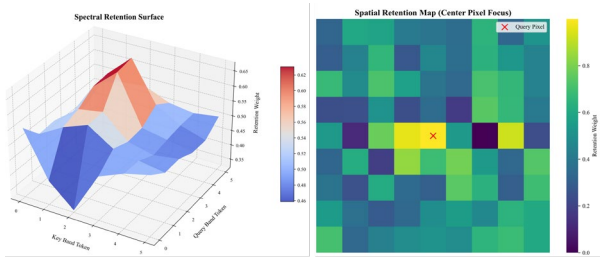

Fig. 5. The visualization of Dual-Axis Retention Mechanism on Pavia University dataset

E. *Decay Rates per Head*

To examine the multi-scale behavior of HSRNet, Fig. 6 visualizes the head-wise decay factors learned in the spatial retention encoder. In our retention block, the decay matrix is defined as $\boldsymbol{D}_{ij} = \gamma^{|i-j|}$, which controls how quickly contributions attenuate with token distance along the flattened

spatial sequence: smaller γ yields faster decay and a more localized effective context, whereas larger γ produces slower attenuation. The learned values span roughly 0.90 to 0.99, indicating that different heads adopt distinct mixing strengths rather than sharing a single fixed receptive field. In particular, heads with lower γ emphasize short-range aggregation and help preserve sharp transitions around object boundaries. Meanwhile, identity-style heads are retained in our implementation to limit unnecessary spatial mixing and reduce over-smoothing. This complementary combination of restrained aggregation and detail-preserving paths is consistent with the cleaner; more coherent predictions observed on the Pavia University dataset.

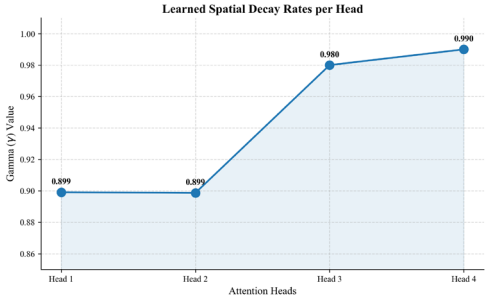

Fig. 6. Learned Spatial Decay Rates per Head

IV. Conclusion

In this paper, we proposed HSRNet for hyperspectral image classification, aiming to address the difficulty of balancing long-range contextual modeling with the preservation of local structures. Different from coupled spectral–spatial processing, HSRNet adopts a dual-axis factorization pipeline, where spectral retention first encodes each pixel spectrum as a token sequence and then spatial retention aggregates the resulting features over spatial tokens. The key component is a distance-decayed retention operator with head-wise learnable decay factors, together with identity-style paths, which enables multi-scale contextual mixing while limiting unnecessary smoothing. Experiments on the Pavia University and Houston 2013 datasets demonstrate that HSRNet achieves Overall Accuracies of 98.9% and 98.3%, respectively, and produces more coherent classification maps with cleaner regions and clearer boundaries than recent competitive baselines. Overall, these results suggest that decay-controlled retention provides a practical alternative to isotropic attention for reducing excessive context mixing while maintaining useful long-range dependencies in HSI scenes.

References

[1] X. Zhao, J. Ma, L. Wang, Z. Zhang, Y. Ding, and X. Xiao, "A review of hyperspectral image classification based on graph neural networks," *Artif Intell Rev*, vol. 58, no. 6, p. 172, Mar. 2025, doi: 10.1007/s10462-025-11169-y.

[2] X. Wang, K. Tan, P. Du, B. Han, and J. Ding, "A capsule-vectored neural network for hyperspectral image classification," *Knowledge-Based Systems*, vol. 268, p. 110482, May 2023, doi: 10.1016/j.knosys.2023.110482.

[3] B. Fu *et al.*, "Cross-scenario transfer learning for estimating mangrove nitrogen and phosphorus content from field hyperspectral data to SDGSAT-1 and Sentinel-2 images," *Remote Sensing of Environment*, vol. 329, p. 114923, Nov. 2025, doi: 10.1016/j.rse.2025.114923.

[4] T. Zhang, C. Xuan, F. Cheng, Z. Tang, X. Gao, and Y. Song, "CenterMamba: Enhancing semantic representation with center-scan Mamba network for hyperspectral image classification," *Expert Systems with Applications*, vol. 287, p. 127985, Aug. 2025, doi: 10.1016/j.eswa.2025.127985.

[5] V. Kumar, R. S. Singh, M. Rambabu, and Y. Dua, "Deep learning for hyperspectral image classification: A survey," *Computer Science Review*, vol. 53, p. 100658, Aug. 2024, doi: 10.1016/j.cosrev.2024.100658.

[6] H. Song *et al.*, "Compact staring-type underwater spectral imaging system utilizing k-Nearest neighbor-based interpolation for spectral reconstruction," *Optics & Laser Technology*, vol. 181, p. 111880, Feb. 2025, doi: 10.1016/j.optlastec.2024.111880.

[7] N. Rahmani, M. Sekandari, A. B. Pour, H. Ranjbar, H. Nezamabadi pour, and E. J. M. Carranza, "Evaluation of support vector machine classifiers for lithological mapping using PRISMA hyperspectral remote sensing data: Sahand–Bazman magmatic arc, central Iran," *Remote Sensing Applications: Society and Environment*, vol. 37, p. 101449, Jan. 2025, doi: 10.1016/j.rsase.2025.101449.

[8] M. Chen *et al.*, "Classification and Recognition of Soybean Quality Based on Hyperspectral Imaging and Random Forest Methods," *Sensors*, vol. 25, no. 5, Mar. 2025, doi: 10.3390/s25051539.

[9] X. Ma *et al.*, "An Ultralightweight Hybrid CNN Based on Redundancy Removal for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024, doi: 10.1109/TGRS.2024.3356524.

[10] M. Wang *et al.*, "Quantitative Inversion of Oil Film Thickness Based on Airborne Hyperspectral Data Using the 1DCNN_GRU Model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023, doi: 10.1109/TGRS.2023.3325805.

[11] X. Zhang, Y. Su, L. Gao, L. Bruzzone, X. Gu, and Q. Tian, "A Lightweight Transformer Network for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023, doi: 10.1109/TGRS.2023.3297858.

[12] Y. Xu *et al.*, "Spatial–Spectral 1DSwin Transformer With Groupwise Feature Tokenization for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023, doi: 10.1109/TGRS.2023.3294424.

[13] F. Ullah *et al.*, "Squeeze-SwinFormer: Spectral Squeeze and Excitation Swin Transformer Network for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 21400–21418, 2025, doi: 10.1109/JSTARS.2025.3595434.

[14] Y. Fang, L. Sun, Y. Zheng, and Z. Wu, "Deformable Convolution-Enhanced Hierarchical Transformer With Spectral-Spatial Cluster Attention for Hyperspectral Image Classification," *IEEE Trans. on Image Process.*, vol. 34, pp. 701–716, 2025, doi: 10.1109/TIP.2024.3522809.

[15] Y. Sun *et al.*, "Retentive Network: A Successor to Transformer for Large Language Models," Aug. 09, 2023, *arXiv*: arXiv:2307.08621. doi: 10.48550/arXiv.2307.08621.

[16] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, Jul. 2003, pp. 288–290 vol.1. doi: 10.1109/IGARSS.2003.1293752.

[17] J. Zhou, S. Zeng, G. Gao, Y. Chen, and Y. Tang, "A Novel Spatial–Spectral Pyramid Network for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–14, 2023, doi: 10.1109/TGRS.2023.3303338.

[18] R. Xu, X.-M. Dong, W. Li, J. Peng, W. Sun, and Y. Xu, "DBCTNet: Double Branch Convolution-Transformer Network for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024, doi: 10.1109/TGRS.2024.3368141.

[19] Y. Xu, D. Wang, L. Zhang, and L. Zhang, "Dual selective fusion transformer network for hyperspectral image classification," *Neural Networks*, vol. 187, p. 107311, Jul. 2025, doi: 10.1016/j.neunet.2025.107311.

[20] B. Xi *et al.*, "MCTGCL: Mixed CNN–Transformer for Mars Hyperspectral Image Classification With Graph Contrastive Learning," *IEEE Trans. Geosci. Remote Sensing*, vol. 63, pp. 1–14, 2025, doi: 10.1109/TGRS.2025.3529996.