CALSRN: CLUSTERING ATTENTION-BASED LOW-DIMENSIONAL SPECTRAL RECONSTRUCTION NETWORK FOR HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral image single-image super-resolution (HSI-SISR) aims to reconstruct a high-resolution hyperspectral image (HR-HSI) with high spatial resolution and high spectral fidelity from only a low-resolution observation. Existing methods still face two key challenges when processing hyperspectral images(HSIs). On the one hand, HSIs contain a large number of continuous spectral bands with severe spectral redundancy, and directly modeling features in the original high-dimensional spectral space introduces heavy computational burden. On the other hand, pixels belonging to the same land-cover category may be spatially scattered while sharing similar spectral responses. Existing attention mechanisms based on fixed and rectangle windows, spectral-dimensional attention, or Top-k selection are still limited in directly modeling such pixel-level non-local self-similarity under relatively low computational complexity. To address these issues, this paper proposes a clustering attention-based low-dimensional spectral reconstruction network (CALSRN). First, a parameter-free spectral compression module is designed to compress the original high-dimensional HSI into a compact spectral representation, thereby reducing spectral redundancy and the computational cost of subsequent attention modeling. Second, a clustering attention block (CAB) is proposed. Specifically, clustering is performed on the compressed spectral HSI at the $\times 1$ scale level to pre-generate a pixel membership matrix, which is then used to guide attention computation. In this way, spatially distant pixels with similar spectral responses can effectively exchange information. To address the difficulty of parallel computation caused by unequal cluster sizes, CAB further adopts a category-sorting-based fixed-size grouping strategy and an adjacent-group expansion strategy. Meanwhile, pixel-to-category-center inter-group cross-attention is introduced to incorporate global category-level contextual information. Third, hierarchical reconstruction loss and dynamic weight balancing are introduced to supervise intermediate reconstruction results and stabilize progressive multi-scale training. Experimental results demonstrate that the proposed CALSRN effectively improves spatial detail recovery and spectral fidelity for hyperspectral image super-resolution.

1 INTRODUCTION

Hyperspectral images (HSIs) record the spectral responses of target scenes across tens or hundreds of continuous narrow bands. Compared to conventional RGB images, HSIs provide richer material attributes and highly discriminative spectral information. Consequently, they have been extensively applied in land-cover classification, target detection, agricultural monitoring, and environmental analysis (Plaza et al., 2009). However, constrained by the physical limitations of imaging sensors, hyperspectral imaging inherently suffers from a strict trade-off between spatial and spectral resolutions. To achieve finer spectral sampling, spatial resolution is frequently compromised, resulting in an insufficient representation of textures, edges, and fine-grained spatial structures (Mei et al., 2017), (Li et al., 2020). Therefore, reconstructing HR-HSIs with both enhanced spatial resolution and strict spectral fidelity from low-resolution observations remains a critical task in hyperspectral

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

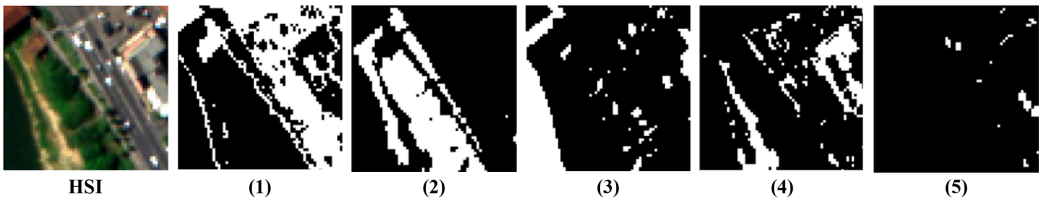


Figure 1: K-means regions, where white pixels share the same spectral category.

image processing (Sidorov & Hardeberg, 2019), (Yang et al., 2024). Depending on the incorporation of external auxiliary images, HSI super-resolution can be broadly categorized into fusion-based HSI-SR and single-image HSI-SR. Fusion-based methods utilize high-resolution RGB, multispectral, or panchromatic images to supplement spatial information of HSIs (Khader et al., 2022), (Huang et al., 2022). However, these approaches necessitate additional sensors and accurate cross-modal coregistration. In contrast, HSI-SISR reconstructs the HR-HSI solely from a low-resolution observation (LR-HSI), thereby increasing its practicality for real-world deployment. Nevertheless, in the absence of external spatial guidance, HSI-SISR must effectively exploit the spatial-spectral correlations and non-local self-similarity inherent within the low-resolution observation (Sidorov & Hardeberg, 2019).

Existing HSI-SISR methods are mainly based on convolutional networks or Transformer architectures. While convolutional models efficiently extract local spatial-spectral features (Mei et al., 2017), (Li et al., 2018), their fixed receptive fields restrict their capacity to model distant pixels exhibiting similar spectral responses. Transformer self-attention mechanisms can capture long-range dependencies (Zhang et al., 2023), (Chen et al., 2024), however, standard global attention is computationally prohibitive for HSIs (Zhang et al., 2023), (Liu et al., 2021). To reduce computational complexity, existing methods have adopted window attention (Liu et al., 2021), (Liang et al., 2021), rectangular-window attention (Chen et al., 2024), (Li et al., 2025b), spectral attention (Chen et al., 2023b), kernelized attention (Zhang et al., 2023), and Top-k token selection (Xiao et al., 2024). Although these mechanisms improve efficiency from various perspectives, they typically rely on predefined geometric regions or attention-score-based token selection, thereby hindering the direct characterization of pixel-level non-local relationships among spatially scattered but spectrally similar HSI pixels.

HSIs possess two crucial intrinsic priors. First, pixels belonging to the same land-cover category often exhibit similar spectral responses despite being spatially distant. As shown in Figure 1, following the clustering of HSI pixels, those with similar spectral signatures are typically grouped into the same categorical regions. This observation indicates that the non-local self-similarity of HSIs is not strictly constrained by regular spatial windows, but rather depends on the intrinsic categorical affiliations of pixels. Second, continuous spectral bands typically exhibit high correlation, and adjacent bands contain significant redundancy (Wang et al., 2019), (Bioucas-Dias & Nascimento, 2008). As shown in Figure 2, the reference HSIs exhibit clear high-similarity structures among spectral bands. Directly modeling features within the original high-dimensional spectral space not only imposes a substantial computational burden, but also risks introducing interference from noisy bands and redundant spectral information (Wang et al., 2019), (Ji et al., 2022). Consequently, a more effective strategy is to first compress the HSI into a compact representation, then model pixel relationships based on spectral similarity, and finally restore the full spectral dimension.

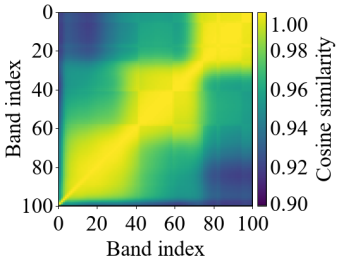


Figure 2: Band-wise cosine-similarity matrix of Pavia Center; warmer colors indicate higher similarity.

Based on the above analysis, CALSRN is proposed. Initially, a parameter-free spectral compression module maps the high-dimensional HSI into a low-dimensional spectral representation, thereby reducing spectral redundancy and the computational cost of subsequent attention modeling. Subsequently, the CAB is designed. Within the CAB, K-means clustering is executed on the compressed spectral representation at the original scale to generate a pixel membership matrix, which then guides the attention computation. In contrast to fixed-window or score-selection-based attention mech-

anisms, the CAB adaptively constructs attention regions based on pixel-level spectral similarity, enabling effective interactions among spatially distant pixels exhibiting similar spectral responses. To reduce computational overhead while ensuring parallel efficiency, the CAB employs a category-sorting-based fixed-size grouping strategy, complemented by an adjacent-group expansion mechanism to prevent pixel truncation. Integrated with a progressive multiscale reconstruction framework, hierarchical reconstruction loss, and dynamic weight balancing, the proposed CALSRN effectively enhances both spatial detail recovery and spectral fidelity under a low computational complexity. The main contributions of this paper are summarized as follows.

- A CALSRN is proposed. Feature learning is executed on a low-dimensional spectral representation obtained via spectral compression, thereby effectively reducing spectral redundancy and computational complexity.
- The CAB is designed as a clustering-prior-guided attention mechanism for HSI-SISR, dynamically grouping spectrally similar pixels into content-aware windows to exploit non-local self-similarity.
- A category-sorting-based fixed-size grouping strategy, complemented by adjacent-group expansion, is developed to enhance the parallel efficiency of intra-group attention. Furthermore, pixel-to-category-center inter-group cross-attention is introduced to incorporate global category-level context.
- A hierarchical reconstruction loss and dynamic weight balancing are incorporated to improve the stability of the progressive multi-scale reconstruction.

2 METHOD

2.1 OVERALL ARCHITECTURE

As illustrated in Figure 3, CALSRN adopts a coarse-to-fine progressive pyramid framework. Given a low-resolution hyperspectral image I_{LR}, the network progressively restores spatial details at multiple scales. For $\times 4$ super-resolution, the reconstruction consists of three stages, with each stage utilizing a reconstruction module of the same structure to achieve image reconstruction at magnification scales of $r = 1$, 2, and 4. The input I^r_{in} and output I^r_{SR} at scale r are formulated as:

$$I^r_{\mathrm{in}} = \begin{cases} I_{\mathrm{LR}}, & r = 1, \\ \mathcal{U}_2\left(I^{r/2}_{\mathrm{SR}}\right) + \mathcal{U}_r(I_{\mathrm{LR}}), & r > 1, \end{cases} \qquad I^r_{\mathrm{SR}} = \mathcal{H}^r_{\mathrm{level}}(I^r_{\mathrm{in}}), \tag{1}$$

where $\mathcal{U}_r(\cdot)$ denotes bicubic upsampling by a factor of r, and $\mathcal{H}^r_{\mathrm{level}}(\cdot)$ represents the reconstruction module at scale r. At each stage, the upsampled result from the preceding stage is fused with the LR-HSI upsampled to the current resolution. This design repeatedly introduces the structural prior of the LR input throughout progressive reconstruction, ensuring the spectral fidelity of the reconstruction results.

At scale r, the input is first mapped to a compact spectral space through a spectral compression module. The CAB and window attention block (WAB) (Liang et al., 2021) then model clustering-guided non-local interactions and local spatial dependencies, respectively. Finally, the spectral dimension is restored through a spectral reconstruction module, and two residual refinement blocks generate the current-scale output:

$$I^r_{\mathrm{SR}} = \mathcal{H}^r_{\mathrm{RBS}}[\mathcal{R}(\mathcal{H}^r_{\mathrm{WAB}}(\mathcal{H}^r_{\mathrm{CAB}}(\mathcal{C}(I^r_{\mathrm{in}}), L^r)), I^r_{\mathrm{in}})], \tag{2}$$

where $\mathcal{H}^r_{\mathrm{X}}$ ($\mathrm{X} \in \{\mathrm{CAB}, \mathrm{WAB}, \mathrm{RBS}\}$) denotes the corresponding block at scale r, $\mathcal{C}(\cdot)$ and $\mathcal{R}(\cdot)$ are spectral compression and reconstruction, and L^r is the K-means membership matrix of the compressed representation. The reconstructed results after the final stage are obtained through a residual refinement block and a global residual connection:

$$I_{\mathrm{SR}} = \mathcal{H}_{\mathrm{RBS}}(I^s_{\mathrm{SR}}) + \mathcal{U}_s(I_{\mathrm{LR}}). \tag{3}$$

This progressive framework jointly exploits non-local spectral similarity and local spatial structure in a compact spectral space, and can be readily adapted to $\times 2$, $\times 4$, and $\times 8$ super-resolution tasks.

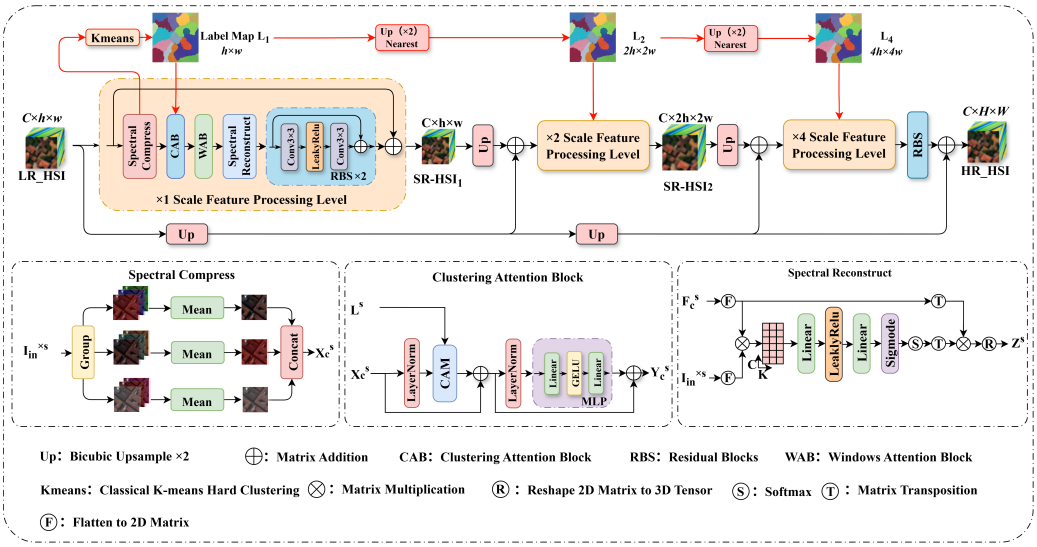


Figure 3: Overall architecture of CALSRN and its core modules, including spectral compression, clustering attention, and spectral reconstruction.

2.2 Spectral Compression and Clustering-Prior Generation

HSIs contain many highly correlated bands, causing severe spectral redundancy. As shown in Figure 2, adjacent bands are highly similar, so the spectra admit a compact representation. We thus compress the spectral dimension using a parameter-free module, as shown in the spectral compression block in Figure 3, reducing attention cost and noise interference.

Given the input at scale r, $I_{\text{in}}^{r} \in \mathbb{R}^{B\times C\times H_r\times W_r}$, its C spectral bands are divided into K non-overlapping groups $\{\mathcal{G}_k\}_{k=1}^{K}$. The compressed representation is obtained by performing a mean operation on the bands within each group:

$$X_c^r(b,k,i,j) = \frac{1}{|\mathcal{G}_k|}\sum_{c\in\mathcal{G}_k} I_{\text{in}}^r(b,c,i,j), \tag{4}$$

where $X_c^r \in \mathbb{R}^{B\times K\times H_r\times W_r}$ denotes the compressed representation and $K \ll C$. Without introducing additional parameters, the module enables subsequent feature learning to be performed efficiently in a compact low-dimensional spectral space.

Clustering-Prior Generation. The compact spectrum obtained above also serves as the clustering prior. At the $\times 1$ scale, the LR-HSI is compressed by (4), giving $X_c^1 = \mathcal{C}(I_{\text{LR}}) \in \mathbb{R}^{B\times K\times h\times w}$. Hard K-means is then applied to each compressed spectral vector $x_{b,i,j} = X_c^1(b,:,i,j)$:

$$z_{b,i,j} = \underset{m\in\{1,\dots,M\}}{\arg\min} \|x_{b,i,j} - \mu_m\|_2^2, \tag{5}$$

where M denotes the number of clusters and μ_m the m-th cluster center. This yields the pixel membership matrix L^1 using only the LR input, without accessing the HR ground truth. At higher-resolution levels, the labels are propagated by nearest-neighbor interpolation rather than recomputed:

$$L^r = \text{Up}_{\text{nearest}}(L^1) \in \{1,\dots,M\}^{B\times H_r\times W_r}. \tag{6}$$

2.3 Clustering Attention Block

As illustrated in Figure 3, CAB consists of layer normalization, the clustering attention mechanism (CAM), residual connections, and an MLP. Given X_c^r and L^r, CAB is formulated as:

$$F^r = LN(X_c^r), \qquad Z = \mathcal{A}_{\text{CAM}}(F^r, L^r), \qquad Y_c^r = MLP(LN(X_c^r + Z)) + Z, \tag{7}$$

where $\mathcal{A}_{\text{CAM}}(\cdot)$ denotes CAM. Since spectrally similar pixels may be spatially distant, CAM uses the pixel clustering priors to construct content-adaptive non-local interaction regions. As shown in Figure 4, CAM consists of intra-group attention and inter-group attention.

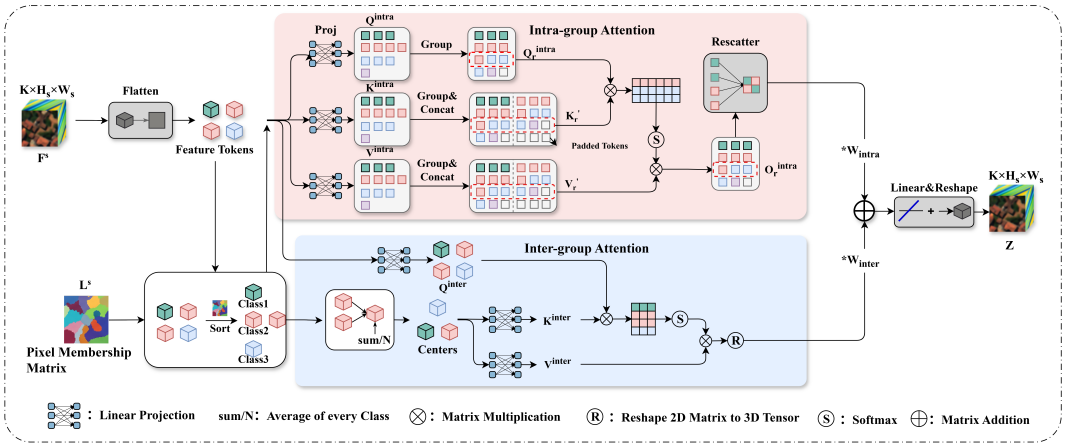


Figure 4: Clustering attention mechanism.

Intra-Group Attention. The feature and membership matrix at the current scale are flattened into $f^r \in \mathbb{R}^{B\times N_r\times K}$ and $l^r \in \{1,\dots,M\}^{B\times N_r}$, respectively, where f^r denotes the token features, l^r the cluster label of each token, and $N_r = H_r W_r$ the number of tokens. To place tokens from the same cluster close to one another, the tokens are sorted according to their cluster labels:

$$\pi_b = \mathrm{Argsort}(l_b^r), \qquad \widetilde{f}_b^r = f_{b,\pi_b,:}^r. \tag{8}$$

where π_b indexes the tokens by cluster label, and $f_{b,\pi_b,:}^r$ gathers the rows of f_b^r in that order. The sorted features are then linearly projected into queries, keys, and values:

$$Q_{\text{intra}} = \widetilde{f}^r W_{\text{intra}}^Q, \qquad K_{\text{intra}} = \widetilde{f}^r W_{\text{intra}}^K, \qquad V_{\text{intra}} = \widetilde{f}^r W_{\text{intra}}^V. \tag{9}$$

Since different clusters contain different numbers of tokens, the sorted sequence is divided into fixed-size subgroups of size g_r. The number of subgroups is

$$R_r = \left\lceil \frac{N_r}{g_r} \right\rceil, \tag{10}$$

and the required number of padding tokens is

$$p_r = R_r g_r - N_r. \tag{11}$$

Fixed-size grouping may split tokens from the same cluster across adjacent subgroups. To alleviate this issue, the queries of the t-th subgroup are obtained from the current subgroup, whereas its keys and values are constructed from both the current and subsequent subgroups:

$$Q_t^{\text{intra}} = Q_{t,:,:}^{\text{intra}}, \qquad K_t' = [K_{t,:,:}^{\text{intra}}; K_{t+1,:,:}^{\text{intra}}], \qquad V_t' = [V_{t,:,:}^{\text{intra}}; V_{t+1,:,:}^{\text{intra}}]. \tag{12}$$

The corresponding intra-group attention is computed as:

$$O_t^{\text{intra}} = \mathrm{MaskedSoftmax}\left(\frac{Q_t^{\text{intra}}(K_t')^{\mathsf{T}}}{\sqrt{K}}\right) V_t', \tag{13}$$

where $O_{t,b}^{\text{intra}}$ is the attention output of the t-th subgroup of the b-th sample and the mask excludes padded positions. After attention is computed, the padding tokens are removed, and the inverse sorting permutation restores the tokens to their original spatial order:

$$Z_b^{\text{intra}} = \left[\mathrm{Unpad}\Big(\mathrm{Concat}_{t=1}^{R_r}\big(O_{t,b}^{\text{intra}}\big)\Big)\right]_{\pi_b^{-1},:}, \tag{14}$$

$$Z_{\text{intra}} = \mathrm{Stack}_{b=1}^{B}\big(Z_b^{\text{intra}}\big) \in \mathbb{R}^{B\times N_r\times K}, \tag{15}$$

where $\mathrm{Unpad}(\cdot)$ removes the p_r padding tokens introduced in Eq. (11), while $\mathrm{Stack}(\cdot)$ stacks the restored token sequences of all samples along the batch dimension.

Cluster-based sorting places spectrally similar tokens close to one another, fixed-size grouping enables parallel batch computation, and adjacent-group expansion alleviates the truncation of same-cluster tokens at subgroup boundaries.

Inter-Group Attention. To incorporate global category-level information, the inter-group branch first computes the center of each cluster. The token set of the m-th cluster in the b-th sample is

defined as $\Omega^r_{b,m} = \{n \mid l^r_{b,n} = m\}$, and the set of valid clusters is $\mathcal{M}^r_b = \{m \mid |\Omega^r_{b,m}| > 0\}$. The corresponding cluster center is calculated as:

$$c^r_{b,m} = \begin{cases} \dfrac{1}{|\Omega^r_{b,m}|} \displaystyle\sum_{n \in \Omega^r_{b,m}} f^r_{b,n}, & m \in \mathcal{M}^r_b, \\ \mathbf{0}, & m \notin \mathcal{M}^r_b \end{cases}. \tag{16}$$

All cluster centers are concatenated as $C^r_b = [c^r_{b,1}, \ldots, c^r_{b,M}]$. The inter-group branch employs pixel-to-cluster-center cross-attention, in which the queries are generated from the pixel tokens, while the keys and values are generated from the cluster centers:

$$Q_{\text{inter}} = f^r W^Q_{\text{inter}}, \qquad K_{\text{inter}} = C^r W^K_{\text{inter}}, \qquad V_{\text{inter}} = C^r W^V_{\text{inter}}. \tag{17}$$

Each pixel attends to all valid cluster centers to obtain global category-level context:

$$Z_{\text{inter}} = \text{MaskedSoftmax}\left(\frac{Q_{\text{inter}} K^{\mathsf{T}}_{\text{inter}}}{\sqrt{K}}\right) V_{\text{inter}}. \tag{18}$$

Invalid centers corresponding to empty clusters are excluded by the attention mask. Finally, the intra-group and inter-group outputs are fused as:

$$Z = \mathcal{P}(\alpha Z_{\text{intra}} + (1-\alpha) Z_{\text{inter}}), \tag{19}$$

where $\mathcal{P}$ is the output projection layer and α is the fusion weight. The intra-group branch models fine-grained non-local relationships among similar pixels, while the inter-group branch provides global category-level context.

The computational complexity of standard global self-attention is $\mathcal{O}(N_r^2 K)$. The complexities of the proposed intra-group and inter-group attention are approximately $\mathcal{O}(2N_r g_r K)$ and $\mathcal{O}(N_r M K)$, respectively. Therefore, the overall complexity of CAB is approximately $\mathcal{O}(2N_r g_r K + N_r M K)$, which is substantially lower than that of global self-attention. By determining the attention range according to content similarity rather than fixed spatial windows, CAB efficiently captures both fine-grained non-local relationships and global category-level information.

2.4 Spectral reconstruction module

As shown in Figure 3, the spectral reconstruction module uses channel attention to restore the features learned in the low-dimensional spectral space to the original spectral dimension. Specifically, the compact feature F^r_c at the current scale and the input I^r_{in} are flattened along the spatial dimension, yielding $F^r_{c,\text{flat}} \in \mathbb{R}^{B \times K \times N_r}$ and $I^r_{\text{in,flat}} \in \mathbb{R}^{B \times C \times N_r}$. The spectral mapping process is formulated as:

$$A^r_0 = I^r_{\text{in,flat}} \left(F^r_{c,\text{flat}}\right)^{\mathsf{T}}, \quad A^r = \text{Softmax}(\sigma(W_2 \delta(W_1 A^r_0)), \dim = 1), \quad Z^r_{\text{flat}} = A^r F^r_{c,\text{flat}}, \tag{20}$$

where W_1 and W_2 are learnable projection matrices, while $\delta(\cdot)$ and $\sigma(\cdot)$ denote LeakyReLU and Sigmoid, respectively. The attention matrix A^r adaptively learns the mapping relationship between each original spectral band and the compressed spectral components. Finally, Z^r_{flat} is reshaped into $Z^r \in \mathbb{R}^{B \times C \times H_r \times W_r}$. This module enables efficient feature learning in the low-dimensional space while recovering complete spectral information.

2.5 Loss function

To generate reconstructed HR-HSIs with enhanced spatial details and spectral fidelity, a joint loss function is defined to train CALSRN. This loss function includes four parts: the reconstruction loss $\mathcal{L}_{\text{rec}}$, the spectral-fidelity loss $\mathcal{L}_{\text{spe}}$, the gradient loss $\mathcal{L}_{\text{gra}}$, and the hierarchical reconstruction loss $\mathcal{L}_{\text{scale}}$. $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{spe}}$, and $\mathcal{L}_{\text{gra}}$ are common loss terms in HSI-SISR and are detailed in Appendix D. Instead of manually tuned weights, each auxiliary term is balanced by a dynamic normalization, and the total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \Psi(\mathcal{L}_{\text{spe}}) + \Psi(\mathcal{L}_{\text{gra}}) + \Psi(\mathcal{L}_{\text{scale}}), \tag{21}$$

where $\Psi(\cdot)$ is a dynamic normalization strategy defined as:

$$\Psi(\mathcal{L}_i) = \frac{\mathcal{L}_i}{\text{SG}\left(\frac{\mathcal{L}_i + \epsilon}{\mathcal{L}_{\text{rec}} + \epsilon}\right)}, \tag{22}$$

in which $\mathrm{SG}(\cdot)$ denotes the stop-gradient operator and ϵ is a small stability constant. The hierarchical reconstruction loss aggregates the reconstruction supervision over all intermediate scales:

$$\mathcal{L}_{\mathrm{scale}} = \sum_{\substack{r \in \mathcal{S} \\ r < s}} \mathcal{L}_{\mathrm{rec}}^{r}, \tag{23}$$

where $\mathcal{L}_{\mathrm{rec}}^{r}$ denotes the L_1 loss between the output I_{SR}^{r} at scale r and the correspondingly downsampled ground truth. This dynamic strategy avoids manually specified loss weights, balances the gradient scales of different supervision terms, and improves the training stability of progressive reconstruction.

3 Experiments

3.1 Datasets and settings

We evaluate CALSRN on three public hyperspectral image benchmark datasets: Pavia Center (Plaza et al., 2009), Botswana (Ungar, 2002), and Washington DC Mall (Aldrich et al., 1996). Following the Wald protocol (Yang et al., 2024), the original high-resolution HSI is treated as the ground-truth image and is Gaussian-blurred before being downsampled to $1/2$, $1/4$, and $1/8$ of its original spatial size, thereby constructing the $\times 2$, $\times 4$, and $\times 8$ super-resolution tasks.

After removing noisy bands, Pavia Center retains 102 bands and is divided into 138 training blocks and 49 testing blocks. Botswana retains 145 bands after removing uncalibrated and noisy bands, and is divided into 69 training blocks and 23 testing blocks. Washington DC Mall contains 191 bands; after cropping, 60 data blocks are used for training and 20 data blocks are used for testing. For fairness, all learning-based methods are retrained or tested under the same experimental protocol. CALSRN is trained with the Adam optimizer. In the final setting, the number of compressed spectral bands is $K_c = 3$, the number of K-means clusters is 32, and the intra-group attention weight is set to $\alpha = 0.85$. The main text reports PSNR, SAM, ERGAS, SSIM, and CC, while RMSE is retained in Appendix C.

3.2 Comparison with state of the art

We compare CALSRN with eleven representative HSI-SISR methods: Bicubic, DHP (Sidorov & Hardeberg, 2019; Ulyanov et al., 2018), MCNet (Li et al., 2020), ERCSR (Li et al., 2021), MSDFormer (Chen et al., 2023b), ESSAFormer (Zhang et al., 2023), CST (Chen et al., 2024), SCPSN (Yang et al., 2024), ReS^3 (Hou et al., 2024), SSRMamba (Li et al., 2025a), and LDERT (Li et al., 2025b). Quantitative results on the three datasets are shown in Tables 1–3, where bold and underlined values indicate the best and second-best results.

Overall, CALSRN achieves the best or highly competitive performance on all three datasets at $\times 2$/$\times 4$/$\times 8$, outperforming existing methods in most PSNR comparisons while maintaining stable advantages in SAM, ERGAS, SSIM, and CC. The gains persist at large scaling factors, indicating stronger modeling of complex spatial structures and spectra, and confirming that clustering-guided compression and reconstruction improve spatial accuracy and spectral preservation simultaneously.

Figure 5 shows visual results on Pavia Center at $\times 4$: CALSRN recovers clearer boundaries and produces smaller errors, consistent with the quantitative results in Tables 1–3. Additional visual results and error analyses on the other datasets are provided in Appendix E.

3.3 Ablation Study

We conduct ablations on Pavia Center under $\times 4$ to evaluate the main components and key designs of CALSRN. The full model uses $K_c = 3$, 32 clusters, and $\alpha = 0.85$, evaluated by PSNR, SSIM, and SAM.

Component Ablation. Each reconstruction level of CALSRN consists of the spectral compression and reconstruction component, the CAB, and the WAB. As shown in Table 4, removing CAB causes a clear performance drop, confirming the importance of long-range dependency modeling among semantically similar pixels. Removing WAB also degrades performance, indicating that local spatial

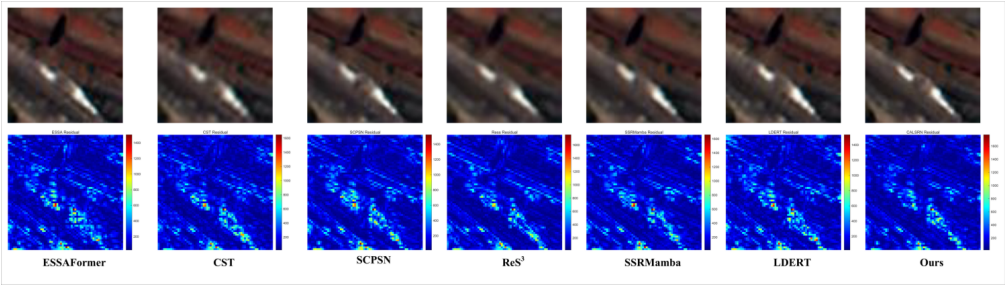


Figure 5: Visual comparison on Pavia Center at $\times 4$. The upper row shows pseudo-color reconstructions and the lower row the corresponding error maps.

Table 1: Quantitative comparison with SOTA methods on the Pavia Center dataset.

Methods	2 × Scaling					4 × Scaling					8 × Scaling				
	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC
Bicubic	34.0790	4.7695	4.6193	0.9010	0.9446	28.8448	7.0668	8.2805	0.6987	0.8194	25.8229	9.5930	11.7798	0.5369	0.6008
DHP	36.9808	4.3986	3.5156	0.9509	0.9605	30.6850	6.4407	6.6482	0.8149	0.8831	26.6353	9.5091	10.9616	0.6178	0.7097
MCNet	36.5016	4.2348	3.5927	0.9454	0.9619	30.8185	6.4178	6.6225	0.7946	0.8813	27.1181	8.8605	10.0555	0.6078	0.7246
ERCSR	37.1148	4.1091	3.3461	0.9516	0.9661	30.9526	6.4497	6.5168	0.8008	0.8836	26.7494	8.9714	10.5264	0.5846	0.6979
MSDFormer	37.2392	4.2080	3.3908	0.9541	0.9633	30.8406	6.4899	6.5763	0.8046	0.8827	26.9689	8.8060	10.2657	0.6097	0.7145
ESSAFormer	37.5543	4.1269	3.3007	0.9555	0.9646	31.4355	6.0923	6.1593	0.8307	0.8958	27.0852	8.5905	10.0458	0.6278	0.7269
CST	37.1358	4.1769	3.4625	0.9538	0.9624	31.3724	6.1833	6.2520	0.8266	0.8944	27.2145	8.5733	9.9359	0.6188	0.7309
SCPSN	37.7238	3.9015	3.1983	0.9578	0.9679	31.5756	5.6875	6.0688	0.8339	0.8989	27.3280	8.2716	9.8175	0.6359	0.7396
ReS^3	36.9248	3.9264	3.4089	0.9510	0.9652	31.1522	5.9333	6.3818	0.8073	0.8885	27.2694	8.4336	9.7992	0.6175	0.7242
SSRMamba	37.2934	4.0413	3.3430	0.9539	0.9658	31.2248	5.9479	6.3251	0.8195	0.8919	27.2760	8.1738	9.8998	0.6329	0.7369
LDERT	37.3805	3.9371	3.2906	0.9546	0.9665	31.1859	5.9000	6.3179	0.8158	0.8909	27.3019	8.4721	9.9763	0.6310	0.7381
CALSRN	**37.9104**	**3.8069**	**3.1292**	**0.9592**	**0.9690**	**31.9204**	**5.4464**	**5.8574**	**0.8471**	**0.9056**	**27.4061**	**7.9302**	**9.7525**	**0.6390**	**0.7437**

Table 2: Quantitative comparison with SOTA methods on the Botswana dataset.

Methods	2 × Scaling					4 × Scaling					8 × Scaling				
	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC
Bicubic	40.0174	1.8687	2.1869	0.9005	0.9075	35.8657	2.8237	3.6900	0.7846	0.7496	33.6759	3.5515	4.8633	0.7193	0.5525
DHP	41.7317	1.7135	1.8493	0.9333	0.9339	36.5655	2.7133	3.5472	0.8196	0.8033	34.1217	3.4746	4.8313	0.7240	0.6426
MCNet	41.8825	1.5840	1.7622	0.9305	0.9364	36.4421	2.5950	3.5115	0.8211	0.7992	33.8576	3.4378	4.8036	0.7291	0.6153
ERCSR	41.1373	1.6867	1.9199	0.9300	0.9255	36.8997	2.6505	3.2308	0.8111	0.8142	33.8045	3.4962	4.8226	0.7148	0.5772
MSDFormer	41.7145	1.6412	1.8582	0.9348	0.9368	36.5454	2.6720	3.3940	0.8206	0.7917	34.1404	3.4251	4.6757	0.7383	0.6510
ESSAFormer	41.9300	1.6571	1.7788	0.9353	0.9374	37.5031	2.4944	3.0039	0.8281	0.8366	34.2975	3.4387	4.5252	0.7389	0.6392
CST	42.2682	1.5864	1.6754	0.9369	0.9406	37.6646	2.4000	2.9327	0.8282	0.8414	34.8612	3.2843	4.3313	0.7285	0.6876
SCPSN	42.5988	1.4990	1.5826	0.9371	0.9444	37.7980	2.3189	2.8759	0.8306	0.8449	35.1350	3.1120	4.1310	0.7421	0.7073
ReS^3	42.2351	1.5123	1.6358	0.9319	0.9404	37.0715	2.4588	3.1631	0.8272	0.8096	34.5852	3.2060	4.3538	0.7404	0.6489
SSRMamba	42.0777	1.5633	1.6746	0.9292	0.9387	36.9736	2.5074	3.1688	0.8212	0.8117	34.5399	3.2332	4.3659	0.7405	0.6506
LDERT	42.6514	1.4826	1.5742	0.9372	0.9454	37.6526	2.3318	2.9364	0.8291	0.8422	35.1646	3.0705	4.1282	0.7419	0.7076
CALSRN	**42.8445**	**1.4501**	**1.5297**	**0.9395**	**0.9469**	**37.9122**	**2.2683**	**2.8377**	**0.8345**	**0.8491**	**35.2401**	**3.0342**	**4.0552**	**0.7432**	**0.7127**

Table 3: Quantitative comparison with SOTA methods on the Washington DC Mall dataset.

Methods	2 × Scaling					4 × Scaling					8 × Scaling				
	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC	PSNR	SAM	ERGAS	SSIM	CC
Bicubic	49.1959	3.8568	4.6439	0.9527	0.9427	44.2918	7.0875	8.2807	0.8686	0.8025	41.6946	9.8181	12.6389	0.8114	0.5581
DHP	48.6770	4.1999	5.4657	0.9608	0.9337	44.2277	6.0418	11.9932	0.9087	0.8009	41.6603	8.9784	12.9644	0.8343	0.6283
MCNet	49.3561	3.2657	6.7956	0.9631	0.9336	44.8932	6.0256	9.4150	0.9017	0.8342	41.8294	8.7748	13.0103	0.8364	0.6504
ERCSR	50.6414	3.0378	4.5261	0.9702	0.9517	45.5203	5.8795	7.4903	0.9053	0.8515	41.7427	8.8516	14.5249	0.8339	0.6419
MSDFormer	50.3477	3.2030	4.5819	0.9709	0.9499	45.3878	6.2597	7.2885	0.8958	0.8498	42.5329	9.0397	10.1801	0.8361	0.6821
ESSAFormer	51.2354	2.7663	3.9563	0.9730	0.9607	45.9792	5.2684	6.9037	0.9161	0.8689	42.1159	8.7289	10.8149	0.8323	0.6568
CST	51.5148	2.8717	4.0666	0.9723	0.9624	46.2649	5.8130	6.5021	0.9179	0.8775	42.5333	8.8424	10.1331	0.8326	0.6822
SCPSN	51.4377	2.6749	3.6909	0.9778	0.9627	46.2784	5.0383	6.4044	0.9188	0.8795	42.6109	8.2052	10.0698	0.8366	0.6939
ReS^3	51.2863	2.8648	3.7171	0.9697	0.9625	45.9711	5.5918	6.7373	0.9067	0.8668	42.5629	8.5790	10.0898	0.8351	0.6830
SSRMamba	50.9183	3.0802	3.8015	0.9702	0.9592	45.7699	5.7821	6.9338	0.9045	0.8618	42.5897	8.7965	10.1190	0.8360	0.6874
LDERT	51.0725	2.7879	4.1628	0.9718	0.9578	45.7322	5.5846	7.0234	0.9071	0.8597	42.6091	8.3799	10.0200	0.8378	0.6854
CALSRN	**52.1198**	**2.5614**	**3.3136**	**0.9782**	**0.9690**	**46.5135**	**4.9888**	**6.3617**	**0.9219**	**0.8828**	**42.6607**	**8.2024**	**10.0065**	**0.8399**	**0.6940**

Table 4: Component ablation.

Variant	Params (M)	FLOPs (G)	PSNR	SSIM	SAM
w/o CAB	1.3132	3.0525	31.7087	0.8405	5.5524
w/o SCR	2.0715	4.4648	31.8352	0.8459	5.5690
w/o WAB	1.3127	3.0520	31.7611	0.8425	5.5076
Full model	1.3133	3.0530	**31.9204**	**0.8471**	**5.4464**

Table 5: Internal CAB ablation.

Variant	Params (M)	FLOPs (G)	PSNR	SSIM	SAM
w/o inter-cl. att.	1.3133	3.0530	31.7730	0.8430	5.5116
w/o intra-cl. att.	1.3133	3.0530	31.7577	0.8415	5.5246
w/o cat.sort.grp.	1.3133	3.0530	31.8913	0.8469	5.5001
Full CAB	1.3133	3.0530	**31.9204**	**0.8471**	**5.4464**

refinement is still necessary after non-local aggregation. Without SCR, both model efficiency and reconstruction accuracy become worse, showing that spectral redundancy suppression is beneficial for HSI reconstruction.

Internal Design of CAB. CAB contains intra-cluster attention and inter-cluster attention. The former models long-range interactions among pixels with similar spectral-semantic responses, while the latter introduces category-level global context through cluster centers. As shown in Table 5, removing either branch reduces performance, confirming that the two branches are complementary. Removing the category-sorting-based grouping strategy causes only marginal accuracy fluctuation but loses parallel computation across cluster groups; as shown in Figure 6, the runtime gap grows as H and W increase.

Loss Function Ablation. The objective combines $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{spe}}$, $\mathcal{L}_{\text{gra}}$, and $\mathcal{L}_{\text{scale}}$. As shown in Table 6, each auxiliary term consistently improves performance over $\mathcal{L}_{\text{rec}}$ alone, confirming the benefits of spectral consistency, spatial gradient preservation, and intermediate-scale supervision.

Number of K-means Clusters. This controls the semantic grouping granularity in CAB. With $K_c = 3$ fixed, Table 7 shows that 32 clusters achieve the best PSNR, SSIM, and SAM: too few clusters yield coarse grouping, while too many over-fragment related pixels and weaken long-range aggregation.

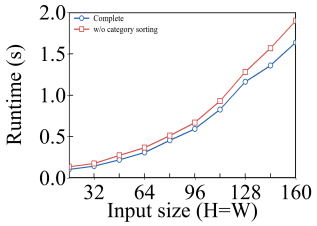


Figure 6: End-to-end runtime per HSI versus input size (H=W).

Table 6: Loss-function ablation ($\times 4$).

$\mathcal{L}_{rec}$	$\mathcal{L}_{spe}$	$\mathcal{L}_{gra}$	$\mathcal{L}_{scale}$	PSNR	SSIM	SAM
✓				31.5326	0.8331	5.9115
✓		✓	✓	31.8193	0.8442	5.6073
✓	✓		✓	31.7577	0.8421	5.5538
✓	✓	✓		31.6232	0.8372	5.7026
✓	✓	✓	✓	**31.9204**	**0.8471**	**5.4464**

Table 7: Number of K-means clusters ($\times 4$).

Clusters	Params (M)	FLOPs (G)	PSNR	SSIM	SAM
1	1.3133	3.0527	31.7616	0.8420	5.5108
8	1.3133	3.0528	31.7817	0.8421	5.5016
16	1.3133	3.0529	31.8564	0.8451	5.4889
32	1.3133	3.0530	**31.9204**	**0.8471**	**5.4464**
64	1.3133	3.0532	31.7531	0.8423	5.5136

3.4 Computational Complexity Comparison

We compare the parameter count and FLOPs of CALSRN with representative SOTA methods on Pavia Center at $\times 4$, including the additional K-means clustering cost. As shown in Table 8, CALSRN has only 1.31M parameters, fewer than all compared learning-based methods; although SSRMamba has the lowest FLOPs, CALSRN attains higher accuracy at 3.05G FLOPs. Compared with SCPSN, the second-best method in PSNR, CALSRN improves PSNR by 0.3448 dB with fewer parameters and lower FLOPs.

Table 8: Complexity comparison on Pavia Center ($\times 4$).

Metric	DHP	MCNet	ERCSR	MSDFormer	ESSAFormer	ReS^3	SCPSN	CST	LDERT	SSRMamba	CALSRN
Params (M)	8.58	2.17	1.59	32.99	11.52	4.74	3.81	10.99	92.78	8.31	**1.31**
FLOPs (G)	4.65	230.51	229.30	24.07	50.19	193.09	6.37	4.97	15.62	**1.75**	3.05
PSNR	30.69	30.82	30.95	30.84	31.44	31.15	<u>31.58</u>	31.37	31.19	31.22	**31.92**

4 Conclusion

We presented CALSRN for HSI-SISR. A parameter-free module first averages within band groups to obtain a compact spectrum, reducing redundancy and attention cost. CAB then clusters the compressed LR pixels into a membership map and uses it to build content-adaptive regions rather than fixed windows; category-sorted grouping with adjacent-group expansion keeps batched computation efficient for unequal clusters, while pixel-to-center cross-attention adds global category context. A channel-attention module restores the full spectrum, while a progressive pyramid refines the result scale by scale. On three benchmarks, CALSRN gains consistently at $\times 2$/$\times 4$/$\times 8$ with only 1.31M parameters and 3.05G FLOPs. The main limitation is reliance on hard LR memberships, unreliable for severe mixed pixels or unknown sensor degradation; future work will pursue confidence-aware clustering and adaptive grouping.

AI USE STATEMENT

In this work, the research ideas, the proposed method, the experimental design, the implementation, and the analysis of the reported results are entirely the authors' own work. Generative AI tools (large language model based assistants) were used only in the ways described below, and all AI-assisted output was manually reviewed and verified by the authors.

Tasks with required disclosure. We have not used generative AI tools for any task in the required-disclosure list. Specifically, no generative AI tool was used to generate synthetic data, to build the theoretical model or conceptual framework, to propose or prove mathematical statements, to formulate or refine research hypotheses, to design the experiments or suggest experimental protocols, to implement the proposed method, to translate the manuscript, to clean or reformat datasets, to support qualitative analysis, or to interpret the research results.

Tasks with recommended disclosure. We used generative AI tools for (i) summarizing and organizing the existing literature during the related-work study, with all cited claims checked against the original sources; (ii) writing and modifying software code, including numerical implementations and debugging scripts, where every piece of AI-suggested code was read, tested, and validated by the authors against the reported results; (iii) polishing the readability of the manuscript, restricted to grammar, spelling, and phrasing, and without introducing, removing, or altering any technical content, claim, or conclusion; and (iv) assisting with the LaTeX typesetting and formatting of the paper, including table layout, figure placement, and reference formatting.

We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

W. S. Aldrich, M. E. Kappus, R. G. Resmini, and P. A. Mitchell. HYDICE postflight data processing. In *Algorithms for Multispectral and Hyperspectral Imagery II*, volume 2758, pp. 354–363, 1996.

J. M. Bioucas-Dias and J. M. P. Nascimento. Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445, 2008.

C. Chen, Y. Wang, N. Zhang, Y. Zhang, and Z. Zhao. A review of hyperspectral image super-resolution based on deep learning. *Remote Sensing*, 15(11):2853, 2023a.

Shuo Chen, Lefei Zhang, and Liangpei Zhang. MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14, 2023b.

Shuo Chen, Lefei Zhang, and Liangpei Zhang. Cross-scope spatial-spectral information aggregation for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 33:5878–5891, 2024.

J. Hou, Z. Zhu, J. Hou, H. Liu, H. Zeng, and D. Meng. Deep diversity-enhanced feature representation of hyperspectral images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8123–8138, 2024.

T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Transactions on Computational Imaging*, 8: 201–214, 2022.

H. Ji, Z. Zuo, and Q.-L. Han. A divisive hierarchical clustering approach to hyperspectral band selection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

A. Khader, J. Yang, and L. Xiao. NMF-DuNet: Nonnegative matrix factorization inspired deep unrolling networks for hyperspectral and multispectral image fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5704–5720, 2022.

B. Li, X. Wang, and H. Xu. SSRMamba: Efficient visual state space model for spectral super-resolution. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025a.

M. Li, Y. Fu, T. Zhang, J. Liu, D. Dou, C. Yan, and Y. Zhang. Latent diffusion enhanced rectangle transformer for hyperspectral image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):549–564, 2025b.

Qing Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, 12(10):1660, 2020.

Qing Li, Qi Wang, and Xuelong Li. Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59 (10):8693–8703, 2021.

Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang. Single hyperspectral image super-resolution with grouped deep recursive residual network. In *IEEE International Conference on Multimedia Big Data*, pp. 1–4, 2018.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844, 2021.

Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Shaohui Mei, Xin Yuan, Jingyu Ji, Yifan Zhang, Shuai Wan, and Qian Du. Hyperspectral image spatial super-resolution via 3d full convolutional neural network. *Remote Sensing*, 9(11):1139, 2017.

Antonio Plaza et al. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009.

Oleksii Sidorov and Jon Yngve Hardeberg. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3844–3851, 2019.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Stephen G. Ungar. Overview of the earth observing one (EO-1) mission. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 1, pp. 568–571, 2002.

Q. Wang, Q. Li, and X. Li. Hyperspectral band selection via adaptive subspace partition strategy. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12): 4940–4950, 2019.

Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang. TTST: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33: 738–752, 2024.

Y. Yang, A. Zhao, S. Huang, X. Wang, and Y. Fan. SCPSN: Spectral clustering-based pyramid super-resolution network for hyperspectral images. In *ACM International Conference on Multimedia*, pp. 2662–2670, 2024.

M. Zhang, C. Zhang, Q. Zhang, J. Guo, X. Gao, and J. Zhang. ESSAformer: Efficient transformer for hyperspectral image super-resolution. In *IEEE/CVF International Conference on Computer Vision*, pp. 23016–23027, 2023.

A Appendix Outline

In this supplementary material, we provide more details of CALSRN as follows:

- Sec. B collects the generic operators used throughout the paper; content-specific symbols are defined where they first appear in the main text.
- Sec. C defines the evaluation metrics adopted in this work and reports the complete RMSE results on all three datasets at $\times 2$, $\times 4$, and $\times 8$.
- Sec. D details the mathematical formulations of the three conventional loss terms used in our training phase: the reconstruction, spectral-fidelity, and gradient losses.
- Sec. E presents additional visual comparison results and mean spectral difference maps on the remaining datasets.
- Sec. F reports supplementary ablation and hyperparameter studies that are not included in the main paper.
- Sec. G presents the downstream classification experiments, which verify that the reconstructed SR-HSIs preserve discriminative land-cover information.

B Notation

Content-specific symbols are defined where they first appear in the main text. The generic operators used throughout the paper are collected here for reference.

Operator	Definition
$LN(\cdot)$	layer normalization over the feature dimension
$MLP(\cdot)$	feed-forward network $MLP(x) = W_2\,\mathrm{GELU}(W_1 x)$, with two linear layers and a GELU in between
$\delta(\cdot), \sigma(\cdot)$	LeakyReLU and Sigmoid activations
$\mathrm{Argsort}(\cdot)$	indices that sort the entries of a vector in ascending order
$\mathrm{Up}_{\mathrm{nearest}}(\cdot)$	nearest-neighbor upsampling of discrete label maps
$\lceil\cdot\rceil$	ceiling operator (smallest integer no smaller than the argument)
$(\cdot)^{\mathsf{T}}$	matrix transpose
$\mathrm{MaskedSoftmax}(\cdot)$	row-wise softmax with masked (padded or invalid) positions excluded
$\mathrm{Concat}(\cdot)$	concatenation along the token dimension
$\mathrm{Unpad}(\cdot)$	removal of the padding tokens appended in Eq. (11)
$\mathrm{Stack}(\cdot)$	stacking the per-sample results along the batch dimension
$\mathrm{SG}(\cdot)$	stop-gradient operator (the argument is treated as a constant in backpropagation)
$\mathcal{O}(\cdot)$	asymptotic computational complexity

C Evaluation metrics

Following the standard evaluation protocol for HSI-SISR, the reconstructed HR-HSI is compared with the ground truth in terms of pixel fidelity, structural consistency, and spectral fidelity. Six metrics are adopted: PSNR and RMSE measure pixel-level reconstruction error, with PSNR higher being better and RMSE lower being better; SSIM evaluates the preservation of local luminance, contrast, and structure; SAM measures the angular discrepancy between the reconstructed and reference spectral vectors, and is insensitive to illumination scaling; ERGAS is a global synthesis error that jointly reflects spatial and spectral distortion; and CC measures the linear correlation between the reconstructed and reference spectra. Among them, SSIM and CC are higher-is-better metrics, while SAM, ERGAS, and RMSE are lower-is-better metrics. The main paper reports PSNR, SAM, ERGAS, SSIM, and CC; RMSE is additionally reported in Table 9 for all three datasets and all scale factors.

Table 9: RMSE (lower is better) on all three datasets at $\times 2$, $\times 4$, and $\times 8$. Best and second-best results are bold and underlined, respectively.

Method	Pavia Center			Botswana			Washington DC Mall		
	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$
Bicubic	0.0240	0.0450	0.0643	0.0190	0.0301	0.0386	0.0115	0.0203	0.0274
DHP	0.0164	0.0348	0.0548	0.0149	0.0260	0.0357	0.0109	0.0169	0.0246
MCNet	0.0179	0.0348	0.0541	0.0154	0.0270	0.0366	0.0097	0.0171	**0.0241**
ERCSR	0.0168	0.0348	0.0560	0.0166	0.0268	0.0377	0.0091	0.0168	<u>0.0242</u>
MSDFormer	0.0162	0.0347	0.0547	0.0148	0.0271	0.0354	0.0092	0.0176	0.0244
ESSAFormer	0.0156	0.0325	0.0538	0.0147	0.0248	0.0355	<u>0.0084</u>	0.0161	0.0250
ReS^3	0.0171	0.0344	0.0535	0.0152	0.0263	0.0349	0.0091	0.0167	0.0244
CST	0.0162	0.0325	0.0530	0.0144	0.0246	0.0348	0.0084	<u>0.0160</u>	0.0244
SCPSN	<u>0.0153</u>	<u>0.0322</u>	<u>0.0528</u>	<u>0.0143</u>	<u>0.0243</u>	0.0334	0.0087	0.0161	0.0243
LDERT	0.0159	0.0336	0.0533	0.0143	0.0246	<u>0.0332</u>	0.0091	0.0167	0.0242
SSRMamba	0.0160	0.0334	0.0531	0.0152	0.0265	0.0351	0.0094	0.0170	0.0243
CALSRN	**0.0150**	**0.0308**	**0.0525**	**0.0140**	**0.0241**	**0.0330**	**0.0082**	**0.0157**	0.0243

D LOSS FUNCTIONS

The main paper introduces the hierarchical reconstruction loss and the dynamic weighting strategy. This section details the conventional loss terms that are only briefly mentioned there. Numerous studies have shown that L_1 and L_2 losses are effective in SR tasks. However, it has been proven that the L_1 loss performs better as a reconstruction loss compared to the L_2 loss. Therefore, we use the L_1 loss to calculate the pixel-level difference between the reconstructed HR-HSI I^{SR} and the corresponding ground truth (GT) image I^{GT}, which can be expressed as:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{B}\sum_{b=1}^{B}\left\|I_b^{\mathrm{GT}} - I_b^{\mathrm{SR}}\right\|_1, \tag{24}$$

where B represents the number of images in a training batch and $\|\cdot\|_1$ denotes the L_1 norm.

To ensure the spectral fidelity of the reconstructed HR-HSI, a spectral-fidelity loss is defined to measure the spectral similarity between I^{SR} and I^{GT}. The spectral-fidelity loss is expressed as:

$$\mathcal{L}_{\mathrm{spe}} = \frac{1}{BHW}\sum_{b=1}^{B}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{1}{\pi}\arccos\frac{\left\langle x_{b,i,j}^{\mathrm{GT}},\, x_{b,i,j}^{\mathrm{SR}}\right\rangle}{\left\|x_{b,i,j}^{\mathrm{GT}}\right\|_2\left\|x_{b,i,j}^{\mathrm{SR}}\right\|_2}, \tag{25}$$

where $x_{b,i,j}^{\mathrm{GT}}$ and $x_{b,i,j}^{\mathrm{SR}}$ are the ground-truth and reconstructed spectral vectors at position (i,j), $\arccos(\cdot)$ denotes the arccosine function, and $\|\cdot\|_2$ denotes the L_2 norm. This loss constrains the angular discrepancy between the reconstructed spectral vector and the ground-truth spectral vector, encouraging spectral consistency.

To ensure that the reconstructed HR-HSI possesses rich high-frequency features, a gradient loss is defined to constrain the consistency of spatial gradients between the reconstructed image and the real image. This loss reinforces the model's ability to recover edge and texture details by calculating the difference in gradient magnitude in the horizontal and vertical directions. Specifically, the Sobel operator is adopted as the gradient operator, with convolution kernels as follows:

$$G_x = \begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}, \qquad G_y = \begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}. \tag{26}$$

For an image I, its gradient magnitude map $G(I)$ can be calculated as:

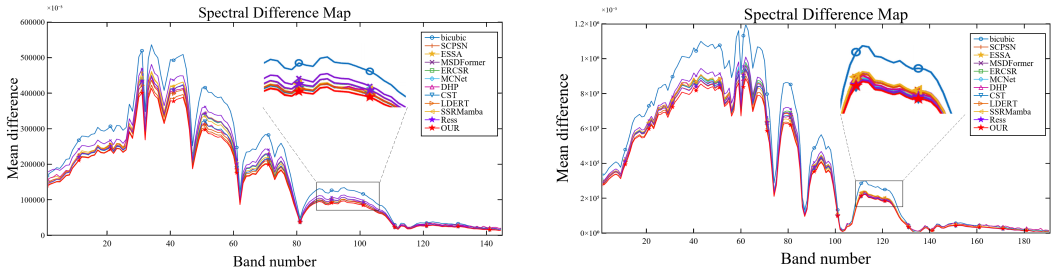
$$G(I) = \sqrt{(I * G_x)^2 + (I * G_y)^2}, \tag{27}$$

where $*$ denotes the convolution operation. The gradient loss $\mathcal{L}_{\mathrm{gra}}$ is defined as the L_1 norm difference between the gradient magnitude maps of the reconstructed image I^{SR} and the real image I^{GT}:

$$\mathcal{L}_{\mathrm{gra}} = \frac{1}{B}\sum_{b=1}^{B}\left\|G\big(I_b^{\mathrm{GT}}\big) - G\big(I_b^{\mathrm{SR}}\big)\right\|_1. \tag{28}$$

This loss encourages the network to recover sharper edges, textures, and local structures. The hierarchical loss, the dynamic normalization, and the final objective are given in the main paper.

E ADDITIONAL VISUAL RESULTS



(a) Botswana (b) Washington DC Mall

Figure 7: Mean spectral difference maps at $\times 4$.

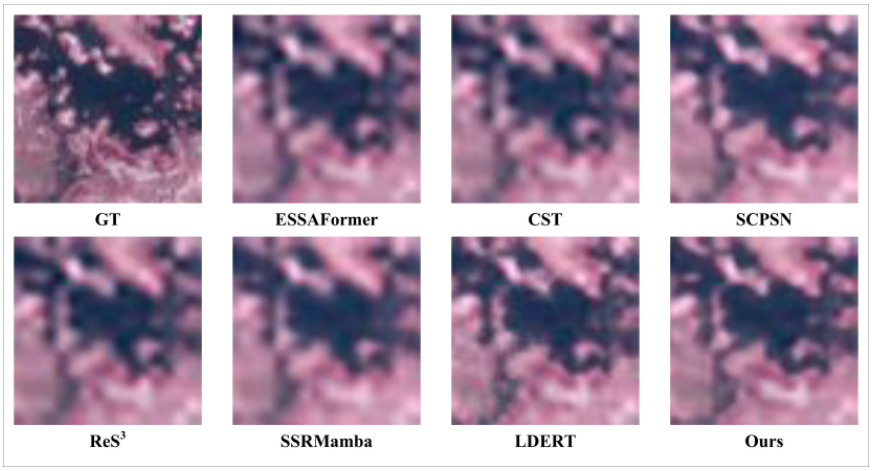


Figure 8: Visual comparison on the Botswana dataset under $\times 4$ super-resolution. For each compared method, the upper image shows the reconstructed pseudo-color HSI and the lower image the corresponding reconstruction error map.

As shown in Figure 8, CALSRN preserves clearer boundaries and produces fewer visible artifacts than the competing methods. The mean spectral difference curves in Figure 7 also show lower spectral errors across most bands, confirming that CALSRN better maintains spectral fidelity while enhancing spatial resolution. The consistent gains on both urban and natural land-cover scenes suggest that the clustering-guided grouping prior is not tied to a specific dataset appearance.

F ADDITIONAL ABLATION AND HYPERPARAMETER ANALYSIS

To verify the effectiveness of the proposed modules and the rationality of the internal designs of CALSRN, ablation experiments are conducted on the Pavia Center dataset under the $\times 4$ super-resolution setting. All ablation variants are evaluated under the same single-seed experimental protocol, and the complete CALSRN entry uses the final configuration with $K_c = 3$, 32 clusters, and a $0.85/0.15$ intra/inter attention ratio. This section reports the ablations and hyperparameter studies that are not included in the main paper.

Internal ablation of the spectral compression module. To validate the parameter-free spectral compression strategy, only the compression module is replaced while the rest of the architecture is kept unchanged. The proposed strategy is compared with PCA-based spectral dimensionality reduction (Chen et al., 2023a) and the SCPSN-style compression strategy (Yang et al., 2024) based on spectral clustering and hyperchannel selection.

Table 10: Spectral-compression alternatives on Pavia Center ($\times 4$).

Method	Time (s)	FLOPs (G)	PSNR	SSIM	SAM
PCA-based	0.2417	**3.0530**	31.7942	0.8425	5.4940
SCPSN-based	0.0525	3.0532	31.8429	0.8451	5.4938
Ours	**0.0423**	3.0530	**31.9204**	**0.8471**	**5.4464**

Compared with both strategies, the proposed method achieves lower end-to-end runtime and the best reconstruction performance, indicating that parameter-free adjacent-band averaging effectively reduces spectral redundancy while preserving useful information.

Internal ablation of the spectral reconstruction module. For the spectral reconstruction module, the channel-attention-based strategy is compared with a simple 1×1 convolution that directly restores all spectral channels from the compressed representation.

Table 11: Spectral-reconstruction alternatives on Pavia Center ($\times 4$).

Method	Time (s)	FLOPs (G)	PSNR	SSIM	SAM
Conv	0.0701	3.0544	31.4808	0.8317	5.7357
Ours	**0.0423**	**3.0530**	**31.9204**	**0.8471**	**5.4464**

Using a 1×1 convolution leads to much worse reconstruction accuracy, showing that a simple convolution cannot capture the relationship between compressed spectral channels and the full spectral channels, whereas the channel-attention-based strategy recovers spectral information more effectively.

Comparison with other attention mechanisms. To further demonstrate the effectiveness of clustering attention, CAB is replaced with several representative attention mechanisms for long-range dependency modeling.

Table 12: Comparison with other attention mechanisms on Pavia Center ($\times 4$).

Method	Params (M)	FLOPs (G)	PSNR	SSIM	SAM
ESSA attention	**1.3133**	**3.0527**	31.7754	0.8425	5.5150
Rectangle attention	1.3133	3.0527	31.6964	0.8392	5.5326
Top-k attention	1.3133	3.0527	31.6893	0.8396	5.5376
Window attention	1.3138	3.0535	31.7209	0.8398	5.5323
Clustering attention	1.3133	3.0530	**31.9204**	**0.8471**	**5.4464**

Compared with ESSA attention, rectangle attention, Top-k attention, and window attention, the proposed clustering attention achieves the best overall performance, indicating that constructing content-aware pixel groups according to spectral similarity is more suitable for HSI-SISR than fixed geometric partitions or score-based token selection.

Progressive upsampling. To validate the progressive reconstruction architecture, CALSRN is compared with a direct $\times 4$ upsampling variant in which the intermediate upsampling stages are removed.

Table 13: Progressive versus direct upsampling on Pavia Center ($\times 4$).

Method	PSNR	SSIM	SAM
Direct upsampling	31.5192	0.8285	5.6811
Progressive upsampling	**31.9204**	**0.8471**	**5.4464**

Direct upsampling causes a substantial performance drop, demonstrating that decomposing a large-scale reconstruction task into hierarchical subproblems helps preserve spatial-spectral structures.

Dynamic loss weighting. Fixed loss weighting is compared with the proposed dynamic weighting strategy.

Table 14: Fixed versus dynamic loss weighting on Pavia Center ($\times 4$).

Method	PSNR	SSIM	SAM
Fixed loss weights	31.7098	0.8389	5.6781
Dynamic loss weights	**31.9204**	**0.8471**	**5.4464**

Dynamic weighting achieves better performance on all reported metrics, indicating that adaptively balancing different supervision terms reduces the sensitivity to manually selected loss weights and improves the stability of multi-objective optimization.

Number of compressed spectral bands. The spectral compression module divides the original C bands into K_c adjacent groups and averages the bands within each group. Different values of K_c are evaluated with the number of clusters fixed to 32.

Table 15: Effect of the compressed spectral dimension on Pavia Center ($\times 4$).

K_c	Params (M)	FLOPs (G)	PSNR	SSIM	SAM
1	**1.3125**	**3.0515**	31.7630	0.8421	5.5258
3	1.3133	3.0530	**31.9204**	**0.8471**	**5.4464**
6	1.3158	3.0573	31.8789	0.8460	5.5067
9	1.3196	3.0640	31.8138	0.8425	5.5154
18	1.3395	3.0987	31.7810	0.8424	5.5202
102	2.1346	4.4793	31.6898	0.8401	5.5390

Using all spectral bands leads to the worst PSNR and SSIM, confirming that redundant spectral information introduces unnecessary computation and interferes with reconstruction. A small compressed spectral dimension performs better, and $K_c = 3$ gives the best PSNR, SSIM, and SAM; it is therefore adopted in the final configuration.

Ratio between intra-cluster and inter-cluster attention. The outputs of intra-cluster and inter-cluster attention are fused by weighted summation under different intra/inter ratios.

Table 16: Effect of the intra/inter attention ratio on Pavia Center ($\times 4$). The inter-group weight is $1 - \alpha$.

α (intra)	PSNR	SSIM	SAM
0.05	31.7276	0.8402	5.5655
0.15	31.8007	0.8425	5.5048
0.30	31.7859	0.8414	5.5179
0.50	31.7749	0.8414	5.5143
0.70	31.7772	0.8425	5.5138
0.85	**31.9204**	**0.8471**	**5.4464**
0.95	31.7529	0.8411	5.5587

The model performs best when intra-cluster attention receives the larger weight, indicating that long-range interaction among semantically similar pixels plays a dominant role in HSI reconstruction; inter-cluster attention remains necessary because it provides complementary global category-level context.

G Downstream classification

To further evaluate whether the reconstructed SR-HSIs preserve discriminative land-cover information for downstream tasks, classification experiments are conducted on the Pavia Center $\times 4$ super-resolution results. Following the classification setting of the reference experiment, the iterative self-organizing data analysis technique algorithm (ISODATA) in ENVI is adopted to classify the SR-HSIs generated by different methods. Different from the reference setting based on the $\times 8$ Pavia super-resolution task, the classification experiment in this work is performed on the $\times 4$ task, with the number of classes set to 5 and the maximum number of iterations set to 100.

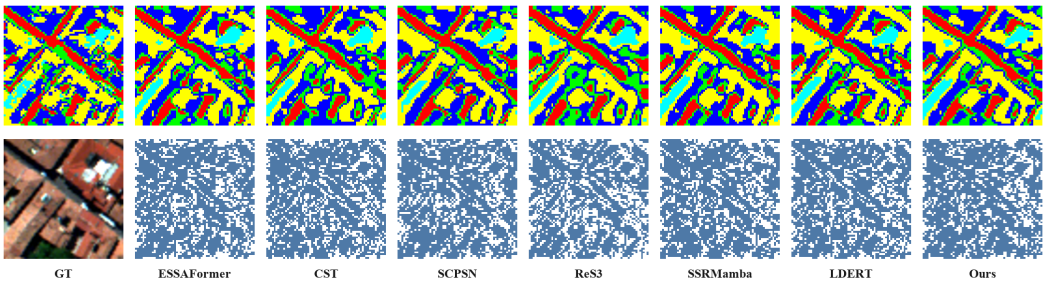


Figure 9: Qualitative classification comparison on the Pavia Center dataset under $\times 4$ super-resolution. The first row shows the GT label map and the classification maps of representative methods; the second row shows the GT pseudo-color RGB composite and the corresponding classification residual maps. Blue and white pixels indicate correct and incorrect predictions with respect to the GT labels, respectively.

Figure 9 provides a qualitative assessment of the downstream classification behavior. Compared with competing reconstructions, CALSRN yields a classification map that is more spatially coherent and better aligned with the GT label structure, with fewer fragmented regions around boundaries and fine-grained land-cover transitions. The residual maps further show that the errors of CALSRN are sparser and less spatially structured, indicating that the reconstructed spectra retain more discriminative information for pixel-level target recognition.

For objective evaluation, overall accuracy (OA) and the kappa coefficient (Kappa) are reported below. CALSRN achieves the best OA of 71.6797% and the best Kappa of 0.6231; compared with the second-best CST, it improves OA by 1.0498 percentage points and Kappa by 0.0149, further demonstrating the benefit of the proposed reconstruction strategy for downstream HSI classification.

Table 17: Classification accuracy on Pavia Center ($\times 4$).

Method	OA(%)	Kappa
Bicubic	51.6357	0.3651
DHP	67.9199	0.5707
ERCSR	64.7217	0.5331
MSDFormer	66.4063	0.5528
ESSAFormer	69.4824	0.5919
ReS^3	63.7695	0.5213
LDERT	69.3604	0.5917
CST	<u>70.6299</u>	<u>0.6082</u>
SCPSN	70.3125	0.6020
MCNet	63.9160	0.5206
SSRMamba	69.2627	0.5898
Ours	**71.6797**	**0.6231**