LGD-Net: Leader-Guided Cross-Modal Dynamics for Hyperspectral and Panchromatic Image Fusion

Rajkumar[1], Balasubramanian Raman[1], and Pravendra Singh[1]

Department of Computer Science and Engineering,
Indian Institute of Technology Roorkee, Uttarakhand 247667, India
{raj_k,pravendra.singh,bala}@cs.iitr.ac.in

Abstract. Hyperspectral-panchromatic fusion reconstructs high-resolution hyperspectral (HRHS) images by combining low-resolution hyperspectral (LRHS) observations with high-resolution panchromatic (PAN) imagery. Although recent deep architectures achieve strong spectral–spatial reconstruction performance, most models rely on stacked layers in which global cross-modal interactions and local spatial refinement are updated jointly at every stage. Consequently, the influence of PAN information is implicitly coupled to network depth, limiting independent control over cross-modal interaction and reducing flexibility in balancing spectral preservation and spatial enhancement. We propose the Leader-Guided Dynamics Network (LGD-Net), a framework that explicitly decouples global cross-modal interaction from local multi-scale reconstruction. A global leader representation first aggregates modality-specific features to summarize global context. Instead of propagating interactions through repeated stacking, only the leader state evolves through a controlled, continuous-depth dynamic process, while local feature nodes focus on spatial reconstruction. The refined leader is subsequently injected back into the multi-scale feature hierarchy via structured residual modulation, providing global guidance without increasing architectural depth. By separating global state evolution from local reconstruction, LGD-Net enables explicit control over cross-modal interactions while maintaining computational efficiency. Experiments on multiple hyperspectral benchmarks demonstrate improved spectral–spatial reconstruction quality compared to conventional stacked fusion architectures.

Keywords: Hyperspectral imaging · Image fusion · Remote Sensing

1 Introduction

Hyperspectral (HS) imaging captures detailed spectral information across hundreds of contiguous bands and supports applications such as environmental monitoring, resource exploration, and precision navigation [2, 29]. However, due to physical acquisition limits and signal-to-noise constraints, hyperspectral sensors typically provide low spatial resolution. This limitation makes it difficult to capture fine structural details. Hyperspectral image fusion addresses this problem

by combining low-resolution HS (LRHS) imagery with high-resolution panchromatic (PAN) images to generate high-resolution hyperspectral (HRHS) images that retain spectral fidelity while improving spatial detail [19, 24, 32].

Deep learning has greatly improved hyperspectral image fusion by learning joint spectral–spatial features [12, 22, 46]. Convolutional networks extract local spatial structures, while transformer-based models capture long-range dependencies through attention mechanisms [1, 4, 10, 31, 49]. Hybrid architectures attempt to combine these advantages to balance spatial enhancement and spectral preservation [7, 8, 33, 38, 43]. In most existing methods, however, fusion is performed through stacked layers where spatial and spectral interactions are updated together at every stage of the network [10, 16, 27].

In stacked architectures, global cross-modal information is accumulated implicitly as features propagate through successive layers. Strengthening global interaction, therefore, typically requires increasing network depth or introducing heavier interaction modules. However, depth simultaneously controls the intensity of local feature transformations. As a result, global interaction and local reconstruction are governed by the same architectural parameter, limiting the flexibility to control the extent to which PAN information influences HS image. This coupling is particularly restrictive in regions where spectral preservation and spatial enhancement require different levels of emphasis [6].

To address this limitation, we introduce an explicit global state that summarizes global context for each modality prior to cross-modal interaction. Instead of relying solely on stacking depth to shape global interactions, this global state evolves through a dedicated continuous dynamic transformation across continuous integration steps. Local feature nodes are updated separately to focus on spatial reconstruction. By separating global state updates from local feature refinement, the model enables more direct control over cross-modal interaction strength without increasing architectural depth.

Based on this idea, we propose a leader-guided dynamic framework. A leader representation is first constructed for each modality to summarize global multi-scale context. This leader is updated via a learnable continuous dynamic operator that models depth-wise global-state transformations. The leader state undergoes iterative dynamic updates during the continuous evolution process, keeping the computation efficient while preserving long-range modeling capacity. The updated leader subsequently guides the refinement of multi-scale feature nodes, enabling detailed spatial reconstruction while maintaining spectral consistency. This design separates global interaction modeling from local refinement.

The main contributions of this work are summarized as follows:

1. We propose a leader-guided dynamic framework that explicitly separates global cross-modal interaction from local multi-scale reconstruction for hyperspectral image fusion.
2. We introduce a global leader state that evolves through continuous dynamics while interacting with feature nodes via structured graph refinement.
3. Extensive experiments and ablation studies demonstrate consistent improvements in spectral–spatial reconstruction quality on benchmark datasets.

2 Related Work

Deep learning has significantly advanced hyperspectral image fusion by enabling hierarchical spectral–spatial representation learning [6, 22, 46, 58]. Early convolutional neural network (CNN) approaches focused on extracting local spatial structures while preserving spectral fidelity through repeated local feature aggregation across network depth. HyperPNN [12] pioneered CNN-based hyperspectral fusion by strengthening deep spectral representations tailored to fusion tasks, demonstrating that end-to-end spectral feature learning can effectively improve reconstruction quality.

Subsequent methods extended this paradigm to capture broader contextual dependencies and multiscale information. Multi-branch and multi-scale designs were introduced to enhance representation capability and progressive reconstruction. For instance, Dong *et al.* [6] proposed a dual-branch pyramid network that exploits hierarchical feature extraction and progressive refinement to better integrate LRHS and PAN information. PanNet [46] adopted an unfolding-based strategy inspired by iterative optimization, improving both reconstruction accuracy and interpretability by embedding model-based priors into a deep architecture. In addition, multi-task formulations [42] jointly optimized fusion and downstream classification objectives through collaborative learning strategies, enhancing feature generalization and semantic consistency.

To overcome the locality limitations of CNNs, transformer-based architectures have been introduced to model long-range spectral–spatial and cross-modal dependencies. HyperTransformer [1] leveraged self-attention mechanisms to explicitly capture global interactions between LRHS and PAN features, enabling more effective cross-modal information exchange. Recent advances have incorporated increasingly sophisticated modulation and attention mechanisms to refine feature alignment. FMPM-DNet [39] introduced a wave function-based feature modulation stage combined with a learnable probabilistic mask to guide reconstruction, improving fusion accuracy and robustness. Similarly, dual-level attention correction networks [40] employed hierarchical spatial and spectral attention strategies to progressively refine fused representations, achieving superior spectral preservation and spatial enhancement.

More recently, diffusion-based image fusion and pansharpening frameworks have demonstrated promising reconstruction performance by modeling the generation process through iterative denoising and probabilistic refinement. These approaches progressively reconstruct high-resolution representations through multiple diffusion steps, enabling strong spatial detail recovery and robustness to complex degradations. However, their fusion process is typically formulated as iterative generative reconstruction, often requiring substantial computational cost during inference [30, 44]. In contrast, our work focuses on controlled cross-modal interaction through a compact dynamical state evolution mechanism, providing an alternative perspective that explicitly separates global interaction modeling from local structural refinement.

Graph-based strategies have been widely explored to model nonlocal spectral–spatial relationships in hyperspectral imagery. In such approaches, graphs

are constructed over pixels, superpixels, or spectral bands to enable structured message passing among relational nodes [54]. A graph-based superpixel segmentation method was introduced for image-driven fusion to facilitate the integration of PAN and multispectral information through graph structures [9].

Several subsequent works have further strengthened relational reasoning mechanisms. Jin et al. [56] combines multiscale convolutional feature extraction with graph attention to model spectral-node dependencies and iteratively refine node embeddings for enhanced nonlocal feature representation. Jia et al. [23]introduced the graph-in-graph architecture, which operates on superpixel-level nodes and captures both intra- and inter-superpixel relationships to align local and global contextual information [37]. More recently, Zhang et al. [52] proposed a spectral–spatial dual graph unfolding Network, derived from graph-regularized restoration models, to explicitly model spectral band correlations and spatial nonlocal similarities through dual graph structures. In image fusion, Keyu et al. [45] GCPNet proposed incorporates graph-based relational reasoning to enhance cross-resolution alignment and reconstruction fidelity.

In parallel with architectural developments, dynamical system perspectives have increasingly influenced deep learning research. Residual networks have been interpreted as discretizations of differential equations, and Neural Ordinary Differential Equations (Neural ODEs) provide a continuous-depth modeling framework [5, 18, 21, 25]. ODE-inspired models have been applied to hyperspectral classification and super-resolution tasks [26, 53], demonstrating advantages in interpretability and controlled feature evolution. However, the explicit modeling of a persistent global dynamical state for cross-modal fusion in hyperspectral image fusion remains largely unexplored.

Despite these advances across CNNs, transformers, diffusion-based methods, graph-based frameworks, most hyperspectral image fusion methods rely on hierarchical feature aggregation across network depth. Global spectral–spatial representations emerge implicitly through stacked transformations, and cross-modal interactions are largely governed by the architecture rather than explicitly controlled. Similarly, ODE-inspired and dynamical system models offer continuous-depth feature evolution, but they rarely incorporate a persistent global state for regulating cross-modal fusion. These observations highlight a key limitation: existing methods lack mechanisms to explicitly separate global interaction modeling from local structural refinement, motivating approaches that can explicitly regulate global cross-modal dynamics while preserving fine-grained spatial details.

3 Methodology

In this section, we present the proposed Leader-Guided Dynamics Network (LGD-Net), whose architecture is illustrated in Figure 1. The framework is designed to bridge the gap between local spatial refinement and global cross-modal fusion through a hierarchical reasoning process.

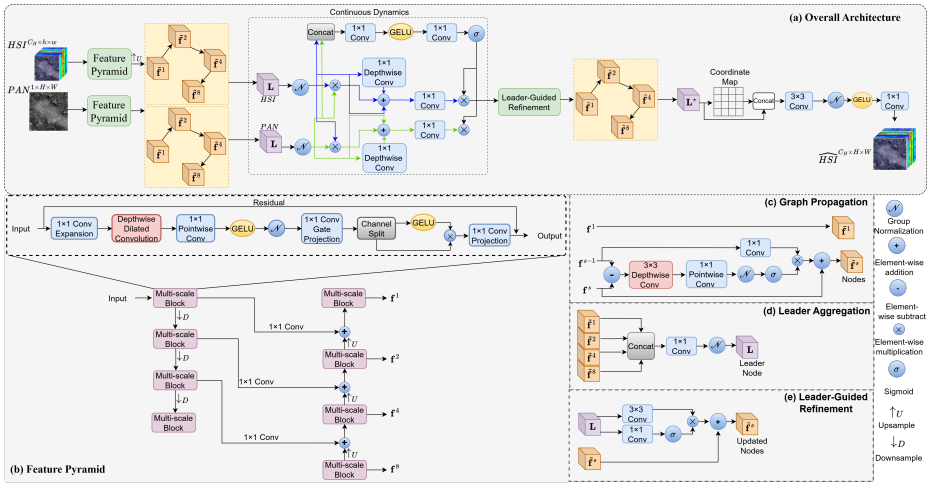

Fig. 1: Overview of the proposed fusion framework: (a) Overall architecture, (b) Feature pyramid encoder, (c) Graph propagation, (d) Leader aggregation, and (e) Leader-guided refinement.

First, we describe the gated feature pyramid encoder, which extracts multi-scale features by leveraging gated depthwise-dilated convolutions to capture both fine-grained spatial structures and broad semantic context. Next, we introduce a leader-guided dynamic module that combines multi-scale refinement with continuous cross-modal evolution. Graph propagation first performs discrepancy-aware inter-scale updates to refine intra-modal feature representations. The refined features are then aggregated into modality-specific leaders, which subsequently evolve through a continuous cross-modal dynamical system.

Finally, we provide a comprehensive overview of the complete LGD-Net pipeline, detailing the leader-guided refinement mechanism and the coordinate-conditioned decoding process for high-resolution hyperspectral image reconstruction.

3.1 Feature Pyramid

To extract multi-scale representations that balance local spatial fidelity with global semantic awareness, we introduce a gated feature pyramid encoder. Unlike standard pyramids that rely on simple convolutions, our architecture utilizes gated depthwise-dilated blocks to adaptively modulate information flow across varying receptive fields.

Let $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}}$ denote the input feature. The encoder first projects the input into an intermediate latent space using a 1×1 convolution to align the channel dimension to a hidden width C, producing the feature $\mathbf{Z}_e$. To model spatial dependencies while maintaining computational efficiency, we employ a depthwise dilated convolution followed by a pointwise convolution to enable

inter-channel mixing. The resulting feature is normalized using channel-wise normalization:

$$\mathbf{Z}_n = \mathcal{N}\left(\mathrm{Conv}_{\mathrm{pw}}\big(\mathrm{Conv}_{\mathrm{dw}}^{(k,d)}(\mathbf{Z}_e)\big)\right), \tag{1}$$

where k denotes the square kernel size and d represents the dilation rate, which expands the receptive field without increasing the number of trainable parameters.

Channel-wise modulation is introduced through a gated mechanism. The normalized feature $\mathbf{Z}_n$ is expanded to $2C$ channels via a linear transformation $\mathcal{W}_g$ and split into two components, $\mathbf{G}_1$ and $\mathbf{G}_2$. The gated response is defined as:

$$\mathbf{Z}_g = \mathbf{G}_1 \odot \mathrm{GELU}(\mathbf{G}_2), \tag{2}$$

where $\odot$ denotes the element-wise product. The final output of the block $\mathbf{Z}_o$ is obtained after a linear projection and a residual connection, ensuring stable gradient flow during training.

The hierarchy is constructed through two coupled pathways. The bottom-up pathway generates multi-scale features across S pyramid levels with scales $\{1, 2, 4, 8\}$. At each level s, spatial resolution is reduced via average pooling while feature channels are projected to the desired width using convolutional layers. To facilitate top-down integration, each level is passed through a lateral 1×1 convolution to obtain a uniform feature dimension, producing lateral representations $\{\mathbf{l}_s\}_{s=1}^{S}$.

The top-down pathway refines these representations by propagating global context from coarser to finer levels. Starting from the deepest pyramid level, features are progressively upsampled via bilinear interpolation and merged with the corresponding lateral features through element-wise addition. Each fused state is then processed by a refinement block to produce the final refined pyramid:

$$\mathcal{F}_m = \{\mathbf{f}_m^{(1)}, \mathbf{f}_m^{(2)}, \ldots, \mathbf{f}_m^{(S)}\}, \tag{3}$$

where $m \in \{\mathrm{HS}, \mathrm{PAN}\}$ denotes the modality. This structure allows $\mathbf{f}_m^{(1)}$ to retain high-resolution structural details, while $\mathbf{f}_m^{(S)}$ captures condensed semantic information, providing an effective foundation for subsequent graph-based reasoning.

3.2 Leader-Guided Dynamics

The multi-scale pyramid provides rich spatial and spectral representations, but effective fusion requires controlled interaction across scales and modalities. To address this, we introduce a leader-guided dynamic mechanism that combines intra-modal refinement with stable cross-modal continuous evolution. The design enables controlled information exchange across scales while maintaining stable nonlinear interaction between modalities.

Graph Propagation While the feature pyramid provides multi-scale representations $\{\mathbf{f}_m^{(s)}\}_{s=1}^{S}$ for each modality $m \in \{\mathrm{HS}, \mathrm{PAN}\}$, these features remain

hierarchically separated. The multi-scale feature representations define the graph nodes, while interactions between adjacent pyramid levels define the graph edges. After encoding, the HS and PAN nodes are upsampled to PAN resolution to ensure spatial alignment. Inter-scale propagation is performed through a discrepancy-aware residual update that promotes structural consistency while preserving local spectral and spatial details.

We define a residual difference-gated update between adjacent scales. For each $s > 1$, the refined node is computed as:

$$\widetilde{\mathbf{f}}_m^{(s)} = \mathbf{f}_m^{(s)} + \mathbf{g}_m^{(s)} \odot \phi\Big(\mathbf{f}_m^{(s-1)}\Big), \tag{4}$$

where $\phi(\cdot)$ denotes a learnable 1×1 linear projection that aligns channel representations across scales.

The propagation strength is modulated by a channel-wise gating function $\mathbf{g}_m^{(s)}$ that adaptively controls the contribution of features from adjacent scales based on the inter-scale discrepancy:

$$\mathbf{g}_m^{(s)} = \sigma\Big(\mathcal{N}\left(\psi\Big(\mathbf{f}_m^{(s)} - \mathbf{f}_m^{(s-1)}\Big)\right)\Big), \tag{5}$$

where $\psi(\cdot)$ denotes depthwise spatial filtering followed by pointwise projection, $\mathcal{N}(\cdot)$ is channel wise Group Normalization, and $\sigma(\cdot)$ is the sigmoid function. The graph propagation operator is implemented through the discrepancy-aware gated residual update between adjacent pyramid-scale nodes, where the propagation strength is adaptively regulated according to the inter-scale discrepancy.

If the structural discrepancy between adjacent scales is minimal, the gate $\mathbf{g}_m^{(s)}$ converges toward a neutral equilibrium state, facilitating a steady-state integration of features that prevents over-correction. Conversely, if the discrepancy is significant, the gate dynamically modulates the propagation of residual features to enforce structural alignment. This discrepancy-aware gating allows inter-scale propagation while preserving local structures through bounded residual updates.

The resulting refined set $\{\widetilde{\mathbf{f}}_m^{(s)}\}_{s=1}^{S}$ is then processed once per forward pass through the discrete graph block for intra-modal refinement.

Leader Aggregation After discrete graph refinement, the multi-scale representations $\{\widetilde{\mathbf{f}}_m^{(s)}\}_{s=1}^{S}$ of modality $m \in \{\mathrm{HS}, \mathrm{PAN}\}$ are fused into a single leader representation. This operation aggregates the refined multi-scale representations into a unified modality-specific descriptor.

The leader is defined as

$$\mathbf{L}_m = \mathcal{N}\left(\mathbf{W}_{\mathrm{agg},m}\,\mathrm{Concat}\left[\widetilde{\mathbf{f}}_m^{(1)}, \ldots, \widetilde{\mathbf{f}}_m^{(S)}\right]\right), \tag{6}$$

where $\mathbf{W}_{\mathrm{agg},m}$ denotes a 1×1 linear projection that reduces the concatenated multi-scale feature to C channels, and $\mathcal{N}(\cdot)$ represents channel-wise normalization. The resulting leader $\mathbf{L}_m \in \mathbb{R}^{C \times H \times W}$ provides a scale-balanced representation with controlled feature magnitude, ensuring that subsequent cross-modal interaction.

Cross-Modal Continuous Dynamics Following leader aggregation, each modality is represented by $\mathbf{L}_{\mathrm{HSI}}$ and $\mathbf{L}_{\mathrm{PAN}}$. To enable structured fusion, we model their interaction as a continuous dynamical system. To avoid the high computational cost of evolving all multi-scale feature nodes, the continuous dynamics are applied only to the compact leader representations. This design captures global cross-modal context while maintaining efficient computation. At each time step t, both leaders are first normalized:

$$\mathbf{L}_m^{(t)} \leftarrow \mathcal{N}\Big(\mathbf{L}_m^{(t)}\Big). \tag{7}$$

We define the cross-modal interaction terms as

$$\begin{cases} \mathbf{Z}_{\mathrm{HS}}^{(t)} = \mathbf{L}_{\mathrm{PAN}}^{(t)} + \mathbf{L}_{\mathrm{HS}}^{(t)} \odot \mathbf{L}_{\mathrm{PAN}}^{(t)} + \mathbf{D}_{\mathrm{PAN}}\Big(\mathbf{L}_{\mathrm{PAN}}^{(t)}\Big), \\ \mathbf{Z}_{\mathrm{PAN}}^{(t)} = \mathbf{L}_{\mathrm{HS}}^{(t)} + \mathbf{L}_{\mathrm{HS}}^{(t)} \odot \mathbf{L}_{\mathrm{PAN}}^{(t)} + \mathbf{D}_{\mathrm{HS}}\Big(\mathbf{L}_{\mathrm{HSI}}^{(t)}\Big), \end{cases} \tag{8}$$

where $\mathbf{D}_m(\cdot)$ denotes depthwise spatial filtering.

These interaction features are projected through dynamic linear mappings:

$$\begin{cases} \mathbf{F}_{\mathrm{HS}}^{(t)} = \mathbf{W}_{\mathrm{dyn,HS}}\left(\mathbf{Z}_{\mathrm{HS}}^{(t)}\right), \\ \mathbf{F}_{\mathrm{PAN}}^{(t)} = \mathbf{W}_{\mathrm{dyn,PAN}}\left(\mathbf{Z}_{\mathrm{PAN}}^{(t)}\right). \end{cases} \tag{9}$$

where $\mathbf{W}_{\mathrm{dyn},m}$ denotes the 1×1 projection used in the dynamic stage.

Cross-modal exchange is spatially regulated by

$$\boldsymbol{\gamma}^{(t)} = \sigma\left(\mathbf{G}\left(\mathrm{Concat}\left[\mathbf{L}_{\mathrm{HS}}^{(t)}, \mathbf{L}_{\mathrm{PAN}}^{(t)}\right]\right)\right), \tag{10}$$

where $\boldsymbol{\gamma}^{(t)} \in (0,1)^{1\times H\times W}$ is a spatial gate shared across channels. The function $\mathbf{G}(\cdot)$ denotes a two-layer convolutional network composed of a 1×1 convolution for channel reduction, a GELU activation, and a second 1×1 convolution that produces a single-channel gating response.

The continuous dynamics are therefore defined as

$$\frac{d\mathbf{L}_{\mathrm{HS}}}{dt} = \boldsymbol{\gamma}^{(t)} \odot \mathbf{F}_{\mathrm{HS}}^{(t)}, \qquad \frac{d\mathbf{L}_{\mathrm{PAN}}}{dt} = \boldsymbol{\gamma}^{(t)} \odot \mathbf{F}_{\mathrm{PAN}}^{(t)}. \tag{11}$$

Additional details on the dynamical system are provided in the *Supplementary Material*. The system is integrated using a second-order Runge-Kutta (RK2) scheme. Let $\eta = \Delta t \cdot s$ denote the effective step size, where $s \in (0, s_{\max})$ is a learnable scalar bounded through a sigmoid function.

For each modality $m \in \{\mathrm{HS}, \mathrm{PAN}\}$, the RK2 update is given by

$$\mathbf{L}_m^{\left(t+\frac{1}{2}\right)} = \mathbf{L}_m^{(t)} + \frac{\eta}{2}\,\boldsymbol{\gamma}^{(t)} \odot \mathbf{F}_m^{(t)}, \tag{12}$$

$$\mathbf{L}_m^{(t+1)} = \mathbf{L}_m^{(t)} + \eta\,\boldsymbol{\gamma}^{(t)} \odot \mathbf{F}_m^{\left(t+\frac{1}{2}\right)}. \tag{13}$$

Because $\boldsymbol{\gamma}^{(t)} \in (0,1)$ and s is bounded, the magnitude of each update remains controlled, ensuring stable cross-modal refinement.

Leader-Guided Refinement After T iterations, the refined HS leader is injected back into the multi-scale HSI nodes:

$$\widehat{\mathbf{f}}_{\mathrm{HS}}^{(s)} = \widetilde{\mathbf{f}}_{\mathrm{HS}}^{(s)} + \boldsymbol{\beta} \odot \mathbf{R}\left(\mathbf{L}_{\mathrm{HS}}^{(T)}\right), \tag{14}$$

where $\mathbf{R}$ is a 3×3 linear projection and $\boldsymbol{\beta} \in (0,1)^{1\times H\times W}$ is a spatial gating map computed from the normalized final HSI leader. The bounded range of $\boldsymbol{\beta}$ ensures that the update magnitude remains controlled, preserving the stability of the feature hierarchy.

3.3 Overall Architecture

Given a low-resolution hyperspectral image $\mathbf{HS} \in \mathbb{R}^{C_H\times h\times w}$ and a high-resolution panchromatic image $\mathbf{PAN} \in \mathbb{R}^{1\times H\times W}$, where $(H, W) = (rh, rw)$, LGD-Net reconstructs a high-resolution hyperspectral image $\widehat{\mathbf{HSI}}$.

First, modality-specific feature pyramid encoders extract multi-scale representations $\{\mathbf{f}_m^{(s)}\}_{s=1}^{S}$ for $m \in \{\mathrm{HS}, \mathrm{PAN}\}$. All feature maps are spatially aligned to the PAN resolution (H, W) via bilinear interpolation. The aligned nodes are then processed once through the graph propogation block, yielding refined multi-scale features $\{\widetilde{\mathbf{f}}_m^{(s)}\}_{s=1}^{S}$.

The refined nodes are first aggregated into modality-specific leaders according to Eq. 6. The leaders $\mathbf{L}_{\mathrm{HS}}$ and $\mathbf{L}_{\mathrm{PAN}}$ evolve through the continuous cross-modal dynamical system described (Sec 3.2), producing the refined HS leader $\mathbf{L}_{\mathrm{HS}}^{(T)}$ after T integration steps.

The evolved HS leader is injected back into the HS feature hierarchy through the residual update defined in Equation (14).

The refined multi-scale HS features are re-aggregated using the same aggregation operator:

$$\mathbf{L}_{\mathrm{HS}}^{\star} = \mathcal{N}\left(\mathbf{W}_{\mathrm{agg,HSI}}\,\mathrm{Concat}\left[\widehat{\mathbf{f}}_{\mathrm{HS}}^{(1)}, \ldots, \widehat{\mathbf{f}}_{\mathrm{HS}}^{(S)}\right]\right). \tag{15}$$

Since convolutional decoders are inherently translation-invariant, we concatenate a normalized coordinate grid $\mathbf{C} \in [-1,1]^{2\times H\times W}$ with the final leader to provide explicit spatial coordinates for reconstruction:

$$\tilde{\mathbf{L}}_{\mathrm{HS}} = \mathrm{Concat}\left(\mathbf{L}_{\mathrm{HS}}^{\star}, \mathbf{C}\right). \tag{16}$$

A lightweight convolutional decoder then maps the coordinate-conditioned representation to the HRHS output using a 3×3 convolution, normalization, nonlinearity, and a final 1×1 projection. The overall architecture combines intra-modal refinement, continuous cross-modal dynamics, and leader-guided feature modulation to reconstruct HRHS images.

3.4 Loss Function

We adopt the **L1 loss** as the reconstruction objective, which measures the absolute difference between the predicted and target HRHS. It is defined as

$$\mathcal{L} = |x_{\text{ref}} - x|_1 , \tag{17}$$

where x_{ref} denotes the ground-truth HRHS image and x represents the reconstructed HRHS image [20].

4 Experiment

4.1 Datasets and Evaluation Protocol

We evaluate the proposed LGD-Net on three widely adopted hyperspectral benchmarks: Pavia Center [28], Botswana [34], and Chikusei [48]. These datasets differ in spatial resolution, spectral dimensionality, and scene complexity, providing a comprehensive assessment across urban, vegetation-dominated, and mixed land-cover scenarios.

Following the data preparation procedure described in the Supplementary Material, we extract non-overlapping hyperspectral patches from each dataset to serve as reference HRHS images (x_{ref}). The resulting patch dimensions are $102\times160\times160$ for Pavia Center, $145\times128\times128$ for Botswana, and $128\times256\times256$ for Chikusei, where the first dimension denotes the number of spectral bands.

To simulate realistic fusion conditions, we adopt Wald's protocol [36, 51]. Specifically, each reference HS is first convolved with an 8×8 Gaussian kernel to model sensor point spread effects, followed by spatial downsampling with a scaling factor of 4 to generate the LR-HS. The corresponding PAN image is synthesized from the reference HS according to the standard Wald formulation. This protocol ensures a controlled setting in which the fused output can be quantitatively compared against the ground-truth reference.

For each dataset, approximately 80% of the extracted patches are randomly assigned to the training set, while the remaining 20% are reserved for testing. The split is performed at the patch level to prevent spatial overlap between training and evaluation samples.

For visualization purposes, RGB composites are synthesized using dataset-specific spectral bands. We employ bands $(26, 24, 14)$ for Botswana, $(60, 18, 15)$ for Chikusei, and $(55, 35, 15)$ for Pavia Center as the red, green, and blue channels, respectively.

We assess reconstruction quality using both spectral and spatial fidelity metrics. Specifically, we report Cross-Correlation (CC) [47], Spectral Angle Mapper (SAM) [50], Root Mean Square Error (RMSE) [15], Peak Signal-to-Noise Ratio (PSNR) [14], Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [35], and Structural Similarity Index Measure (SSIM) [41]. These metrics jointly evaluate spectral fidelity, spatial detail preservation, and overall reconstruction accuracy.

Table 1: Quantitative comparison on the Botswana, Chikusei, and Pavia datasets. ↑ indicates higher is better, while ↓ indicates lower is better.

Method	Botswana						Chikusei						Pavia					
	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓
HyperPNN	37.430	0.944	2.454	0.969	0.019	1.544	42.890	0.974	5.160	0.981	0.008	3.480	39.060	0.967	4.612	0.986	0.012	2.459
FPFNet	37.750	0.957	3.520	0.977	0.018	1.481	41.030	0.969	5.589	0.977	0.010	3.947	39.320	0.966	6.115	0.986	0.013	2.451
HyperDSNet	38.010	0.947	2.308	0.972	0.018	1.421	43.020	0.974	5.099	0.982	0.008	3.403	39.110	0.967	4.675	0.986	0.012	2.446
HspeNet	38.150	0.947	2.214	0.972	0.018	1.402	42.840	0.975	4.917	0.982	0.008	3.401	38.370	0.965	4.382	0.984	0.013	2.625
HyperRefiner	39.200	0.959	2.094	0.978	0.015	1.249	43.560	0.978	4.862	0.983	0.008	3.339	40.090	0.973	4.213	0.987	0.011	2.318
Lformer	39.210	0.957	2.087	0.978	0.015	1.258	44.890	0.983	3.635	0.987	0.007	2.862	40.730	0.975	3.979	0.988	0.010	2.182
DHP-Darn	40.150	0.965	1.906	0.981	0.013	1.141	45.400	0.985	4.431	0.988	0.006	2.724	41.670	0.978	3.820	0.990	0.009	2.018
MCIFNet	39.237	0.958	2.068	0.977	0.015	1.259	45.101	0.984	4.514	0.988	0.007	2.814	41.365	0.976	3.903	0.989	0.010	2.070
LGD-Net(Ours)	41.856	0.979	1.695	0.987	0.011	1.016	47.814	0.991	3.996	0.992	0.005	2.240	44.830	0.984	3.297	0.992	0.007	1.693

4.2 Implementation Details

The network is trained for 1600 epochs using the Adam optimizer [17]. The initial learning rate is set to 5×10^{-4} and decayed using a cosine annealing schedule to a minimum of 1×10^{-5}. All experiments are conducted on an NVIDIA GeForce RTX A6000 GPU with 48 GB memory. The batch size is set to 8 for Botswana and Pavia Center, and 4 for Chikusei due to its larger spatial resolution.

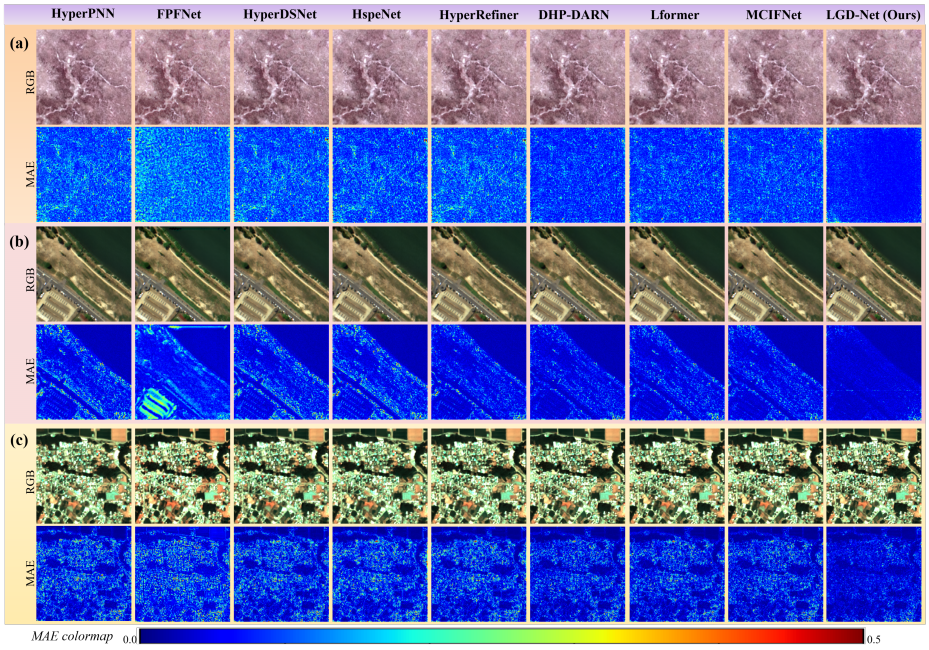

Fig. 2: Qualitative comparison of hyperspectral image fusion results on (a) Botswana (30th patch), (b) Pavia (37th patch), and (c) Chikusei (47th patch). Each panel shows RGB reconstructions from different methods alongside MAE maps representing per-pixel errors averaged across all spectral bands.

Table 2: Analysis illustrating the impact of model complexity on inference time.

Model	Param (M)	FLOPs (G)	Inference Time (ms)
HyperPNN	0.14	2.26	2.17
HyperDSNet	0.25	0.27	4.94
HspeNet	0.12	1.96	3.185
DHP-Darn	0.47	7.59	4.456
HyperRefiner	4.99	13.33	70.34
FPFNet	47.34	321.47	75.19
Lformer	0.81	13.98	19.38
MCIFNet	1.92	15.97	14.08
LGD-Net	2.39	25.73	34.18

Table 3: Ablation study showing the impact of each architectural component on fusion performance.

Experiment Configuration	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓
Baseline	44.896	0.984	3.288	0.992	0.007	1.685
w/o Bilinear interaction	44.201	0.983	3.375	0.992	0.008	1.754
w/o Graph Propagation	43.972	0.982	3.423	0.992	0.008	1.774
w/o Edge gating	44.385	0.983	3.353	0.992	0.007	1.740
w/o Leader refinement	43.818	0.982	3.462	0.992	0.009	1.894
w/o Continuous dynamics	32.725	0.908	4.854	0.958	0.024	4.437

4.3 Quantitative Results

To evaluate the proposed LGD-Net, we compare it with several representative hyperspectral image fusion methods, including HyperPNN [12], HSpNet [11], HyperDSNet [58], DHP-DARN [55], FPFNet [6], HyperRefiner [3], LFormer [13], and MCIFNet [57]. These methods span multiple architectural paradigms, including convolutional networks, refinement-based models, transformer-based designs, and recent state-space architectures, providing a diverse set of baselines for comparison. Table 1 presents the quantitative comparison of LGD-Net with representative hyperspectral image fusion methods on the Botswana, Chikusei, and Pavia datasets. Across all datasets, the proposed method consistently achieves the best overall performance among the compared approaches. On Botswana, LGD-Net improves upon previous methods such as DHP-Darn and Lformer with noticeable gains in reconstruction quality while simultaneously reducing spectral distortion. Similar trends are observed on the large-scale Chikusei dataset, where LGD-Net surpasses existing convolutional, transformer-based, and hybrid models, indicating more effective cross-modal interaction between hyperspectral and panchromatic observations. On the Pavia dataset, the proposed framework again achieves the strongest results, outperforming recent methods including MCIFNet and DHP-Darn and demonstrating improved preservation of both spatial structures and spectral signatures in complex urban scenes. These results demonstrate that LGD-Net consistently improves reconstruction quality across diverse hyperspectral scenes.

4.4 Qualitative Comparison

Figure 2 presents visual comparisons of the reconstructed RGB images and their corresponding spatial MAE maps for the Botswana, Pavia, and Chikusei datasets. Across all scenes, LGD-Net produces reconstructions with clearer structural details and reduced reconstruction error compared with competing methods. In the Botswana example (Figure 2(a)), most existing approaches exhibit noticeable reconstruction artifacts and scattered error patterns in the MAE maps, whereas LGD-Net produces a significantly smoother error distribution with substantially lower magnitude. On the Pavia scene (Figure 2(b)),

Table 4: Analysis of the effect of RK2 solver steps (T).

T	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓
1	36.142	0.954	0.973	2.361	0.022	1.649
2	**41.856**	**0.979**	**0.987**	**1.695**	**0.011**	**1.016**
3	41.242	0.977	0.985	1.762	0.012	1.064
4	40.562	0.975	0.984	1.841	0.012	1.151

Table 5: Analysis of the effect of scale factor (s) with $T = 2$.

s	PSNR↑	SSIM↑	SAM↓	CC↑	RMSE↓	ERGAS↓
0.8	39.268	0.971	0.985	1.915	0.015	1.245
1.0	41.056	0.975	0.985	1.803	0.012	1.082
1.2	**41.856**	**0.979**	**0.987**	**1.695**	**0.011**	**1.016**
1.4	41.201	0.976	0.985	1.775	0.012	1.076
1.8	39.803	0.973	0.984	1.989	0.014	1.190

competing methods struggle to accurately recover fine spatial structures along boundaries and road patterns, which results in concentrated high-error regions. In contrast, LGD-Net better preserves these structural details, leading to visibly reduced error concentrations. A similar trend is observed for the Chikusei dataset (Figure 2(c)), where LGD-Net maintains sharper spatial structures and produces consistently lower error across the scene. The corresponding MAE maps further confirm that the proposed approach suppresses reconstruction errors over both homogeneous regions and fine structures, indicating improved spatial consistency in the fused hyperspectral images.

4.5 Model Complexity and Inference Efficiency

Table 2 reports the number of parameters, FLOPs, and inference time of representative hyperspectral fusion models. Lightweight convolutional models such as HyperPNN and HspeNet exhibit relatively low computational cost, while larger architectures (e.g., FPFNet) involve higher FLOPs and latency.

LGD-Net achieves an efficient balance between model capacity and computational cost, with 2.39M parameters and 25.73G FLOPs. Despite its moderate complexity, it maintains competitive inference time and delivers superior reconstruction accuracy compared with several methods across different computational budgets (Table 1). These results demonstrate that the proposed leader-guided dynamic formulation enhances representational efficiency and reconstruction performance without requiring excessive model scaling.

4.6 Effect of Solver Steps and Step Scale

Table 4 analyzes the effect of the number of RK2 solver steps. A single update step ($T = 1$) yields substantially lower performance across all metrics, indicating that one leader update is insufficient to fully capture cross-modal interactions. Increasing the evolution depth to $T = 2$ significantly improves reconstruction quality, producing the highest PSNR, SSIM, and CC together with the lowest SAM, RMSE, and ERGAS. However, further increasing the number of solver steps gradually degrades performance, suggesting that excessive evolution may introduce unnecessary feature mixing and reduce reconstruction fidelity. Table 5 evaluates the effect of the step scale factor controlling the magnitude of each leader update. Smaller scale factors ($s = 0.8$ and $s = 1.0$) result in weaker

cross-modal interaction and lower reconstruction quality. Performance improves consistently as the scale factor increases, reaching the optimum at $s = 1.2$. Beyond this point, larger update magnitudes lead to gradual degradation across all metrics. These observations indicate that both the evolution depth and update magnitude play important roles in regulating cross-modal interaction, with moderate values yielding the best balance between spatial enhancement and spectral preservation.

4.7 Ablation Study

To evaluate the contribution of each architectural component, we perform ablation studies in which individual modules are removed or modified while keeping the training protocol unchanged. The results are reported in Table 3.

Bilinear Interaction. We replace the multiplicative cross-modal term in Eq. 8 with a single linear interaction. This modification consistently degrades reconstruction quality, indicating that the bilinear coupling enables richer cross-modal feature alignment than linear fusion alone.

Graph Propagation. Removing the discrepancy-aware inter-scale propagation results in consistent performance degradation. This suggests that explicit inter-scale refinement enhances structural coherence prior to leader aggregation.

Edge Gating. Disabling the adaptive gating while retaining residual propagation yields a modest decrease in performance, indicating that learned gating stabilizes feature exchange across scales.

Leader Refinement. Omitting the leader-guided refinement step reduces both spatial and spectral fidelity. This demonstrates that global feedback from the leader representation improves multi-scale consistency before decoding.

Continuous Dynamics. Replacing the continuous-depth dynamic module with simple residual stacking leads to substantial degradation across all metrics. This highlights the importance of controlled continuous cross-modal evolution in the proposed framework.

4.8 Spectral Error Analysis

Figure 3 illustrates the spectral mean absolute error (MAE) across wavelength bands for different methods on the Botswana, Pavia, and Chikusei datasets. Across all datasets, LGD-Net consistently achieves the lowest error profile over most spectral bands, indicating improved spectral reconstruction accuracy. On the Botswana dataset (Figure 3(a)), competing methods exhibit noticeable error fluctuations around several mid-range spectral bands, whereas LGD-Net maintains a smoother and consistently lower error curve, suggesting better preservation of spectral signatures. A similar pattern is observed on the Pavia dataset

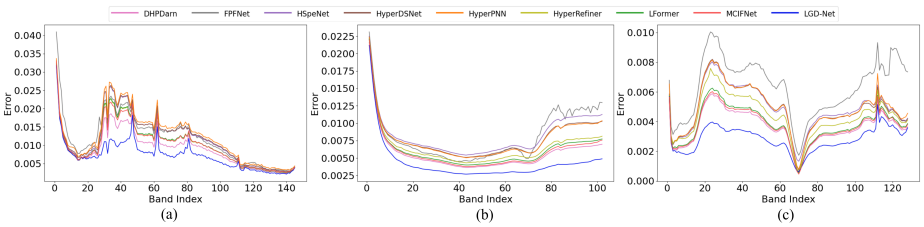

Fig. 3: Spectral MAE across bands for different methods on (a) Botswana, (b) Pavia, and (c) Chikusei datasets.Lower curves correspond to smaller reconstruction errors and therefore better spectral fidelity.

(Figure 3(b)), where LGD-Net produces the lowest error across nearly the entire spectral range while other approaches show increased deviations at higher band indices. On the Chikusei dataset (Figure 3(c)), which contains complex spectral variations, LGD-Net again maintains a stable and lower error trajectory compared to other methods. Overall, the reduced magnitude and smoother distribution of spectral errors indicate that the proposed approach reconstructs spectral information more faithfully across wavelengths. Additional experiments are provided in the supplementary material.

5 Conclusion

We propose LGD-Net, a leader-guided dynamic framework for hyperspectral image fusion that explicitly separates global cross-modal interaction from local multi-scale reconstruction. By evolving a global leader representation through a controlled continuous dynamic process while local nodes handle spatial refinement, the proposed design enables explicit cross-modal interaction without increasing architectural depth. Experimental results demonstrate that this structured global–local decoupling improves spectral–spatial consistency over conventional stacked fusion architectures. Ablation and complexity analyses further indicate that the performance gains primarily stem from the proposed interaction mechanism rather than increased model capacity, while maintaining a favorable balance between reconstruction quality and computational cost. Overall, LGD-Net demonstrates that separating global cross-modal interaction from local multi-scale reconstruction provides an effective and scalable design principle for hyperspectral image fusion.

References

1. Bandara, W.G.C., Patel, V.M.: Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1767–1777 (2022)
2. Bhargava, A., Sachdeva, A., Sharma, K., Alsharif, M.H., Uthansakul, P., Uthansakul, M.: Hyperspectral imaging and its applications: A review. Heliyon **10**(12), e33208 (2024)

3. Bo Zhou, Xianfeng Zhang, X.C.M.R., Feng, Z.: Hyperrefiner: a refined hyperspectral pansharpening network based on the autoencoder and self-attention. International Journal of Digital Earth **16**(1), 3268–3294 (2023). https://doi.org/10.1080/17538947.2023.2246944
4. Cai, J., Huang, B.: Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing **59**(6), 5206–5220 (2020)
5. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
6. Dong, W., Yang, Y., Qu, J., Li, Y., Yang, Y., Jia, X.: Feature pyramid fusion network for hyperspectral pansharpening. IEEE Transactions on Neural Networks and Learning Systems (2023)
7. Guan, P., Lam, E.Y.: Multistage dual-attention guided fusion network for hyperspectral pansharpening. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2021)
8. Guarino, G., Ciotola, M., Vivone, G., Poggi, G., Scarpa, G.: Pca-cnn hybrid approach for hyperspectral pansharpening. IEEE Geoscience and Remote Sensing Letters (2023)
9. Hallabia, H.: A graph-based superpixel segmentation approach applied to pansharpening. Sensors **25**(16) (2025), https://www.mdpi.com/1424-8220/25/16/4992
10. He, L., Ye, H., Xi, D., Li, J., Plaza, A., Zhang, M.: Tree-structured neural network for hyperspectral pansharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023)
11. He, L., Zhu, J., Li, J., Meng, D., Chanussot, J., Plaza, A.: Spectral-fidelity convolutional neural networks for hyperspectral pansharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 5898–5914 (2020). https://doi.org/10.1109/JSTARS.2020.3025040
12. He, L., Zhu, J., Li, J., Plaza, A., Chanussot, J., Li, B.: Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(8), 3092–3100 (2019). https://doi.org/10.1109/JSTARS.2019.2917584
13. Hou, J., Cao, Z., Zheng, N., Li, X., Chen, X., Liu, X., Cong, X., Hong, D., Zhou, M.: Linearly-evolved transformer for pan-sharpening. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 1486–1494. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3680979, https://doi.org/10.1145/3664647.3680979
14. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics Letters **44**, 800–801 (2008). https://doi.org/10.1049/el:20080522
15. Jagalingam, P., Hegde, A.V.: A review of quality metrics for fused image. Aquatic Procedia **4**, 133–142 (2015). https://doi.org/https://doi.org/10.1016/j.aqpro.2015.02.019, iNTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15)
16. Jiao, J., Wang, C.: Hyperspectral pansharpening based on detail injection and hypertransformer. In: Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning. pp. 430–436 (2024)
17. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

18. Koch, J., Forland, B., Doster, T., Emerson, T.: A neural differential equation formulation for modeling atmospheric effects in hyperspectral images. In: Velez-Reyes, M., Messinger, D.W. (eds.) Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX. vol. 12519, p. 125190E. International Society for Optics and Photonics, SPIE (2023). https://doi.org/10.1117/12.2663958, https://doi.org/10.1117/12.2663958
19. Kumar, R., Roy, S.K., Vivone, G., Das, S., Raman, B., Singh, P.: From classical image fusion to deep representation learning: A survey of the advances in hyperspectral image pan-sharpening. Information Fusion **127**, 103834 (2026). https://doi.org/https://doi.org/10.1016/j.inffus.2025.103834
20. Loncan, L., de Almeida, L.B., Bioucas-Dias, J.M., Briottet, X., Chanussot, J., Dobigeon, N., Fabre, S., Liao, W., Licciardi, G.A., Simões, M., Tourneret, J.Y., Veganzones, M.A., Vivone, G., Wei, Q., Yokoya, N.: Hyperspectral pansharpening: A review. IEEE Geoscience and Remote Sensing Magazine **3**(3), 27–46 (2015)
21. Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: International conference on machine learning. pp. 3276–3285. PMLR (2018)
22. Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G.: Pansharpening by convolutional neural networks. Remote Sensing **8**(7), 594 (2016)
23. Mullakaeva, K., Cosmo, L., Kazi, A., Ahmadi, S.A., Navab, N., Bronstein, M.M.: Graph-in-graph (gig): Learning interpretable latent graphs in non-euclidean domain for biological and healthcare applications. arXiv preprint arXiv:2204.00323 (2022)
24. Pallas Enguita, S., Chen, C.H., Kovacic, S.: A review of emerging sensor technologies for tank inspection: A focus on lidar and hyperspectral imaging and their automation and deployment. Electronics **13**(23) (2024). https://doi.org/10.3390/electronics13234850
25. Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A.: Solving deep neural networks with ordinary differential equations for remotely sensed hyperspectral image classification. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 576–579 (2019). https://doi.org/10.1109/IGARSS.2019.8899012
26. Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A.: Neural ordinary differential equations for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **58**(3), 1718–1734 (2020)
27. Pereira-Sánchez, I., Sans, E., Navarro, J., Duran, J.: Multi-head attention residual unfolded network for model-based pansharpening. arXiv preprint arXiv:2409.02675 (2024)
28. Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., Trianni, G.: Recent advances in techniques for hyperspectral image processing. Remote Sensing of Environment **113**, S110–S122 (2009). https://doi.org/https://doi.org/10.1016/j.rse.2007.07.028, imaging Spectroscopy Special Issue
29. Ram, B.G., Oduor, P., Igathinathane, C., Howatt, K., Sun, X.: A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. Computers and Electronics in Agriculture **222**, 109037 (2024). https://doi.org/https://doi.org/10.1016/j.compag.2024.109037
30. Rui, X., Cao, X., Pang, L., Zhu, Z., Yue, Z., Meng, D.: Unsupervised hyperspectral pansharpening via low-rank diffusion model. Information Fusion **107**, 102325

(2024). https://doi.org/https://doi.org/10.1016/j.inffus.2024.102325, https://www.sciencedirect.com/science/article/pii/S1566253524001039
31. Shang, Y., Liu, J., Yang, J., Wu, Z.: A model-inspired approach with transformers for hyperspectral pansharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 7187–7202 (2022)
32. Stuart, M.B., McGonigle, A.J.S., Willmott, J.R.: Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems. Sensors **19**(14) (2019). https://doi.org/10.3390/s19143071
33. Sun, Q., Sun, Y., Pan, C.: Aidb-net: An attention-interactive dual-branch convolutional neural network for hyperspectral pansharpening. Remote Sensing **16**(6), 1044 (2024)
34. Ungar, S., Pearlman, J., Mendenhall, J., Reuter, D.: Overview of the earth observing one (eo-1) mission. IEEE Transactions on Geoscience and Remote Sensing **41**(6), 1149–1159 (2003). https://doi.org/10.1109/TGRS.2003.815999
35. Wald, L.: Data fusion: definitions and architectures: fusion of images of different spatial resolutions. Presses des MINES (2002)
36. Wald, L., Ranchin, T., Mangolini, M.: Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogrammetric engineering and remote sensing **63**(6), 691–699 (1997)
37. Wan, S., Gong, C., Zhong, P., Du, B., Zhang, L., Yang, J.: Multiscale dynamic graph convolutional network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **58**(5), 3162–3177 (2020). https://doi.org/10.1109/TGRS.2019.2949180
38. Wang, H., Zhang, J., Huo, L.Z.: Multi-scale hyperspectral pansharpening network based on dual pyramid and transformer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024)
39. Wang, X., Yang, Y., Huang, S., Lu, H., Wan, W., Zhao, A.: Fmpm-dnet: hyperspectral pansharpening dynamic network based on feature modulation and probability mask. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025)
40. Wang, X., Yang, Y., Huang, S., Wan, W., Liu, Z., Zhang, L., Zhao, A.: Dfcfn: Dual-stage feature correction fusion network for hyperspectral pansharpening. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025). https://doi.org/10.1109/TGRS.2025.3527568
41. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
42. Wu, X., Feng, J., Shang, R., Wu, J., Zhang, X., Jiao, L., Gamba, P.: Multi-task multi-objective evolutionary network for hyperspectral image classification and pansharpening. Information Fusion **108**, 102383 (2024)
43. Wu, X., Feng, J., Shang, R., Zhang, X., Jiao, L.: Multiobjective guided divide-and-conquer network for hyperspectral pansharpening. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–17 (2022)
44. Xiao, J.L., Huang, T.Z., Deng, L.J., Lin, G., Cao, Z., Li, C., Zhao, Q.: Hyperspectral pansharpening via diffusion models with iteratively zero-shot guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12669–12678 (June 2025)

45. Yan, K., Zhou, M., Liu, L., Xie, C., Hong, D.: When pansharpening meets graph convolution network and knowledge distillation. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022). https://doi.org/10.1109/TGRS.2022.3168192
46. Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J.: Pannet: A deep network architecture for pan-sharpening. In: Proceedings of the IEEE international conference on computer vision. pp. 5449–5457 (2017)
47. Yang, Y., Wu, L., Huang, S., Sun, J., Wan, W., Wu, J.: Compensation details-based injection model for remote sensing image fusion. IEEE Geoscience and Remote Sensing Letters **15**(5), 734–738 (2018). https://doi.org/10.1109/LGRS.2018.2810219
48. Yokoya, N., Iwasaki, A.: Airborne hyperspectral data over chikusei (05 2016)
49. Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L.: A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **11**(3), 978–989 (2018)
50. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm (1992)
51. Zeng, Y., Huang, W., Liu, M., Zhang, H., Zou, B.: Fusion of satellite images in urban area: Assessing the quality of resulting images. In: 2010 18th International Conference on Geoinformatics. pp. 1–4 (2010). https://doi.org/10.1109/GEOINFORMATICS.2010.5568105
52. Zhang, K., Yan, J., Zhang, F., Ge, C., Wan, W., Sun, J., Zhang, H.: Spectral–spatial dual graph unfolding network for multispectral and hyperspectral image fusion. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–18 (2024)
53. Zhang, X., Song, C., You, T., Bai, Q., Wei, W., Zhang, L.: Dual ode: Spatial–spectral neural ordinary differential equations for hyperspectral image super-resolution. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024)
54. Zhao, X., Ma, J., Wang, L., Zhang, Z., Ding, Y., Xiao, X.: A review of hyperspectral image classification based on graph neural networks. Artificial Intelligence Review **58**(6), 172 (2025)
55. Zheng, Y., Li, J., Li, Y., Guo, J., Wu, X., Chanussot, J.: Hyperspectral pansharpening using deep prior and dual attention residual network. IEEE Transactions on Geoscience and Remote Sensing **58**(11), 8059–8076 (2020). https://doi.org/10.1109/TGRS.2020.2986313
56. Zhou, H., Luo, F., Zhuang, H., Weng, Z., Gong, X., Lin, Z.: Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
57. Zhu, C., Deng, S., Song, X., Li, Y., Wang, Q.: Mamba collaborative implicit neural representation for hyperspectral and multispectral remote sensing image fusion. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–15 (2025). https://doi.org/10.1109/TGRS.2025.3537638
58. Zhuo, Y.W., Zhang, T.J., Hu, J.F., Dou, H.X., Huang, T.Z., Deng, L.J.: A deep-shallow fusion network with multidetail extractor and spectral attention for hyperspectral pansharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 7539–7555 (2022). https://doi.org/10.1109/JSTARS.2022.3202866