RESPONSE-PAIR SPHERICAL ATTENTION FOR HYPERSPECTRAL AND MULTISPECTRAL IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral and multispectral image fusion (HMF) combines spatially coarse hyperspectral observations with fine-scale multispectral guidance to reconstruct a high-resolution hyperspectral image. In attention-based fusion, reconstruction quality depends on how neighboring responses are weighted relative to each target. Yet conventional dot-product attention couples candidate and target response amplitudes solely through their product. Consequently, response pairs with the same amplitude product and directional agreement receive identical affinity scores even when their relative amplitude allocations differ. We introduce response-pair spherical attention (RPSA), which decomposes each learned candidate–target response pair into overall effective amplitude, relative amplitude allocation, and orientation difference. Its affinity preserves sensitivity to joint response strength while using amplitude balance and directional agreement to regulate candidate contributions. We embed RPSA in RPSA-Net, an alternating cross-resolution architecture that alternates high-resolution feature refinement under multispectral guidance with feedback that updates a low-resolution hyperspectral state. Experiments on three simulated benchmarks and a real-world dataset demonstrate that RPSA-Net achieves state-of-the-art spectral fidelity and spatial reconstruction while maintaining fast inference.

1 INTRODUCTION

Hyperspectral and multispectral image fusion (HMF) aims to reconstruct a high-resolution hyperspectral image (HR-HSI) from observations with complementary spatial and spectral resolutions (Fang et al., 2026; Li et al., 2025). The low-resolution hyperspectral image (LR-HSI) provides dense spectral measurements on a coarse spatial grid (Li et al., 2024), whereas the high-resolution multispectral image (HR-MSI) captures fine spatial structures in a few broad bands (Li et al., 2026c; Huang et al., 2024). Together, these observations constrain the desired image but do not fully determine the fine-scale spatial variation within each hyperspectral band (Zhu et al., 2025; Liu et al., 2025a). Near a material boundary, for example, a neighboring feature may provide a useful spatial cue (Zhang et al., 2026) while carrying information more representative of the adjacent material. Accurate fusion therefore requires weighting neighboring features according to their relevance to the spectrum being reconstructed at each target location (Liu et al., 2025b).

In attention-based HMF networks (Wu et al., 2025; Xu et al., 2026a), learned affinities determine how neighboring features contribute to reconstruction at each target location (Jia et al., 2023; Ma et al., 2024; Fang et al., 2024). Dot-product attention (Vaswani et al., 2017) combines response magnitudes with directional agreement, allowing a larger candidate amplitude to compensate for poorer alignment with the target response. To examine whether this preference agrees with spectral correspondence, we compare the candidate preferred by standard dot-product attention with a better-aligned alternative for the same target. The preferred candidate has the highest dot-product score among nonself candidates in the attention window, whereas the alternative has a smaller response amplitude and better directional alignment with the target response but receives less attention. Across all 12 CAVE test scenes, the alternative is spectrally closer to the target in 58.24% of the selected comparisons and has a smaller mean reference spectral angle (Fig. 1). This discrepancy shows that the weight advantage gained through response magnitude does not necessarily reflect closer spectral correspondence.

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.

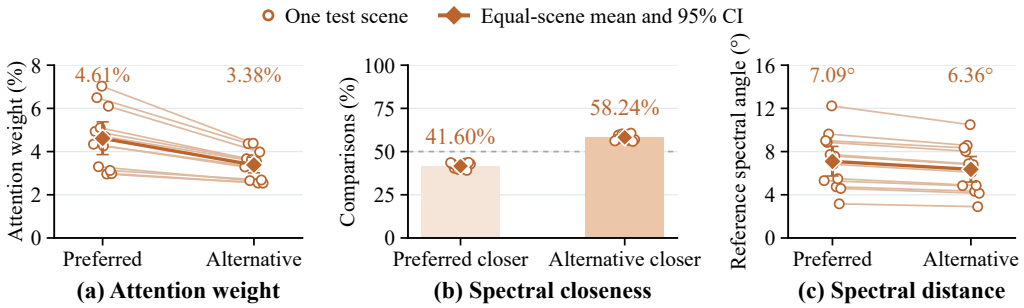


Figure 1: Candidate weighting and spectral correspondence under standard dot-product attention on the 12 CAVE test scenes. The preferred candidate has the highest nonself dot-product score, while the alternative has a lower response amplitude and better directional alignment with the target. (a) Softmax weights normalized over the full attention window. (b) Fractions of selected comparisons in which the preferred or alternative candidate is spectrally closer to the target. (c) Mean candidate-target spectral angles for the same comparisons. Spectral angles are computed between ground-truth spectra, with smaller values indicating closer spectral correspondence.

Using directional agreement alone would remove this amplitude dependence, but would also assign identical affinity to weak and strong response pairs sharing the same relative orientation. We therefore seek to preserve response strength while using the target response as a reference for interpreting candidate amplitude. This calls for separating overall strength from relative amplitude allocation within a response pair. In dot-product attention, however, candidate and target amplitudes enter only through their product. Consequently, response pairs sharing the same amplitude product and directional agreement receive identical affinity scores despite different amplitude allocations. Representing relative amplitude allocation alongside overall strength and directional agreement would allow the target response to regulate how candidate strength contributes to the affinity.

To address this issue, we introduce response-pair spherical attention (RPSA), in which the radius, latitude, and longitude of a spherical representation encode overall effective amplitude, relative amplitude allocation, and orientation difference, respectively. RPSA constructs a nonnegative affinity by modulating the effective-amplitude product with factors for amplitude balance and directional agreement. The amplitude product retains response-strength information, while the balance factor regulates its contribution according to the relative amplitudes of the candidate and target. For an aligned pair, jointly scaling up both effective amplitudes preserves their balance and increases the unnormalized affinity. In contrast, amplifying only the candidate at a fixed target response and relative direction drives the affinity toward a finite, target-dependent limit. RPSA thus preserves sensitivity to joint response strength while limiting the advantage of amplifying a candidate alone.

Building on this mechanism, we develop RPSA-Net, an alternating cross-resolution architecture for HMF. The network initializes a low-resolution feature state from the LR-HSI, repeatedly lifts it to the high-resolution grid for HR-MSI-guided refinement, and feeds the refined representation back to update the state. Throughout this cycle, RPSA uses the target response to regulate neighboring feature contributions. This conditions the transfer of HR-MSI spatial cues on the evolving hyperspectral state at each location, supporting both fine-detail reconstruction and spectral fidelity. Our main contributions are summarized as follows.

- We propose RPSA, a novel attention mechanism that jointly models overall response strength, relative amplitude allocation, and orientation difference, enabling it to distinguish response pairs with identical amplitude products and directional agreement but different amplitude allocations.
- We provide a geometric analysis of RPSA, revealing how amplitude allocation shapes candidate competition and how the target response regulates the affinity's sensitivity to candidate amplification.
- We develop RPSA-Net by integrating RPSA into an alternating cross-resolution fusion architecture, achieving state-of-the-art spectral fidelity and spatial reconstruction with fast inference on three simulated benchmarks, with qualitative validation on a real-world dataset.

2 Related Work

2.1 Hyperspectral and Multispectral Image Fusion

Recent HMF methods improve reconstruction through cross-modal interaction and structured priors. MSST (Jia et al., 2023) combines spatial and spectral Transformer branches with multiscale embeddings to extract complementary features. Reciprocal Transformer (Ma et al., 2024) introduces bidirectional cross-attention for spatial and spectral information exchange. SRLF (Liu et al., 2025b) uses the observation model to identify and selectively refine spatially or spectrally distorted features. More recently, TRT (Xu et al., 2026b) employs multilinear tensor-ring interactions as a structured prior for residual reconstruction. PIF-Net (Li et al., 2026a) combines invertible Mamba with fusion-aware low-rank adaptation to preserve information and calibrate spatial-spectral features. DPFormer (Liu et al., 2026) separates spectral and spatial transfer into successive phases to coordinate cross-modal interaction. LSO (Li et al., 2026b) learns the fusion mapping in a compact latent spectral space through trigonometric basis functions. While these methods strengthen feature representation and information exchange, they do not directly specify how candidate strength should be evaluated relative to the receiving spectral state. RPSA incorporates amplitude balance into the affinity, allowing candidate strength to be interpreted relative to the target response.

2.2 Attention Affinity Design

Attention affinities govern how response magnitude and direction influence aggregation. Dot-product attention (Vaswani et al., 2017) couples query and key magnitudes through their product, whereas scaled cosine attention (Liu et al., 2022) removes both norms from the content-similarity term. ESSAformer (Zhang et al., 2023) constructs a kernelized spectral-correlation affinity for single-image hyperspectral super-resolution. PolaFormer (Meng et al., 2025) models same-signed and opposite-signed interactions separately and improves selectivity through learnable power rescaling. MALA (Fan et al., 2025) restores query-magnitude dependence lost through normalization in linear attention. QUEST (Govindarajan et al., 2026) normalizes keys onto a sphere while retaining query control over attention concentration. NaLaFormer (Meng et al., 2026) combines query-norm-aware feature maps with a nonnegative geometric similarity. However, preserving query magnitude or normalizing keys does not by itself explicitly characterize relative amplitude allocation within a response pair. Product-based scores assign identical values to pairs with equal amplitude products and directional agreement, while key normalization removes candidate-magnitude information. RPSA therefore incorporates relative amplitude allocation into the affinity, retaining candidate-strength information while allowing the target amplitude to regulate candidate weights.

3 Method

3.1 Problem Formulation and Overall Architecture

Let $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ denote the desired high-resolution hyperspectral image (HR-HSI), $\boldsymbol{Y} \in \mathbb{R}^{h\times w\times C}$ the observed low-resolution hyperspectral image (LR-HSI), and $\boldsymbol{Z} \in \mathbb{R}^{H\times W\times C_m}$ the registered high-resolution multispectral image (HR-MSI). Here $H = sh$, $W = sw$, s is the spatial scale factor, and $C_m < C$. We express the observations as

$$\boldsymbol{Y} = \mathcal{D}_s(\boldsymbol{X}) + \boldsymbol{n}_y, \qquad \boldsymbol{Z} = \mathcal{R}(\boldsymbol{X}) + \boldsymbol{n}_z, \tag{1}$$

where $\mathcal{D}_s$ denotes spatial degradation, $\mathcal{R}$ denotes the sensor spectral response, and $\boldsymbol{n}_y, \boldsymbol{n}_z$ account for noise and model mismatch.

As shown in Fig. 2, RPSA-Net first encodes $\boldsymbol{Y}$ to initialize a low-resolution hyperspectral feature state, while $\boldsymbol{Z}$ supplies high-resolution spatial guidance. At each high-resolution stage, the state is lifted to the HR-MSI grid, combined with $\boldsymbol{Z}$, and refined by an RPSA fusion block that combines attention-based aggregation with local convolutional mixing and residual connections. Low-resolution feedback then downsamples the refined features and merges them with the preceding state through a low-resolution RPSA fusion block. After three high-resolution stages and two low-resolution feedback steps, the final features predict a hyperspectral residual that is added to bicubic upsampling of $\boldsymbol{Y}$ to obtain $\widehat{\boldsymbol{X}}$.

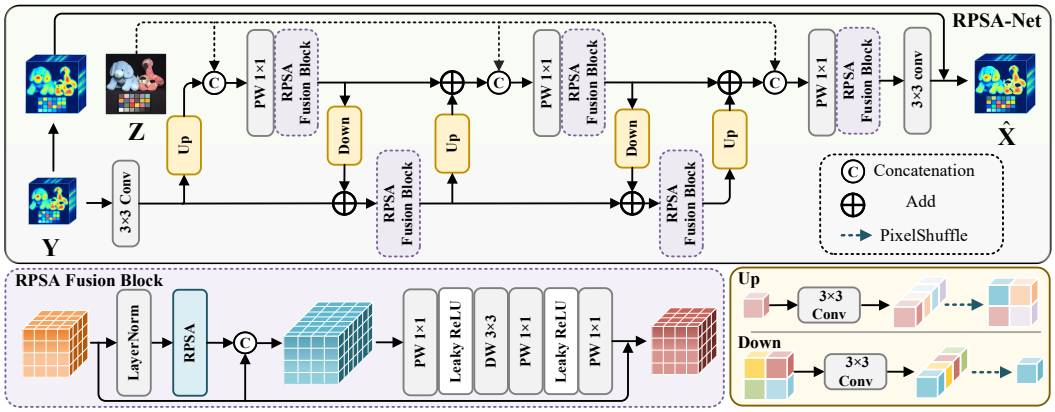


Figure 2: Overview of the RPSA-Net architecture.

3.2 Response-Pair Spherical Geometry

To evaluate neighboring evidence during fusion, we consider candidate strength in relation to the target response. A dot product depends on the two amplitudes only through their product, allowing a larger candidate amplitude to offset a smaller target amplitude without changing the score at a fixed relative orientation. We therefore construct a geometric representation that separates overall strength from amplitude allocation while retaining relative orientation. Figure 3 relates this geometry to the RPSA computation.

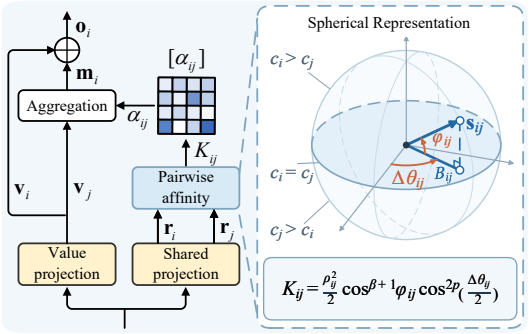


Figure 3: RPSA architecture.

Let $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^{C_f}$ denote the target and candidate features. We map them into a common two-dimensional response plane using a shared projection $\boldsymbol{W}_r \in \mathbb{R}^{2\times C_f}$ to compare their amplitudes and orientations, while a separate value projection carries the content for aggregation. Each projected response has the polar form

$$\boldsymbol{r}_i = \boldsymbol{W}_r \boldsymbol{x}_i = \bar{\rho}_i \begin{bmatrix} \cos\theta_i \\ \sin\theta_i \end{bmatrix}, \qquad \bar{\rho}_i = \|\boldsymbol{r}_i\|_2, \tag{2}$$

where $\bar{\rho}_i$ measures response amplitude and θ_i records orientation in the learned plane. At low response amplitudes, small perturbations can cause large changes in orientation. We therefore attenuate weak responses before constructing the pair geometry and define the effective amplitude as

$$c_i = \bar{\rho}_i \frac{\bar{\rho}_i^2}{\bar{\rho}_i^2 + \tau}, \qquad \tau > 0, \tag{3}$$

where τ controls the attenuation of weak responses.

For a pair of nonzero responses (i, j), let $\Delta\theta_{ij} \in [-\pi, \pi]$ denote $\theta_i - \theta_j$ modulo 2π. We represent the combined effective strength by a radius and encode amplitude allocation and relative orientation as latitude and longitude on a unit sphere, with

$$\rho_{ij} = \sqrt{c_i^2 + c_j^2}, \qquad \boldsymbol{s}_{ij} = \frac{1}{\rho_{ij}^2} \begin{bmatrix} 2c_i c_j \cos\Delta\theta_{ij} \\ 2c_i c_j \sin\Delta\theta_{ij} \\ c_i^2 - c_j^2 \end{bmatrix}. \tag{4}$$

Since $4c_i^2 c_j^2 + (c_i^2 - c_j^2)^2 = (c_i^2 + c_j^2)^2$, $\boldsymbol{s}_{ij}$ lies on the unit sphere. Its longitude $\Delta\theta_{ij}$ records relative orientation, while its latitude $\varphi_{ij} = \arcsin((\boldsymbol{s}_{ij})_3)$ encodes the allocation of squared amplitude

between the two responses. The equator corresponds to equal effective amplitudes, with positive and negative latitudes indicating stronger target and candidate responses, respectively. The radius and latitude jointly determine both effective amplitudes through

$$c_i^2 = \frac{\rho_{ij}^2}{2}(1 + \sin\varphi_{ij}), \qquad c_j^2 = \frac{\rho_{ij}^2}{2}(1 - \sin\varphi_{ij}). \tag{5}$$

These identities follow by substituting $\rho_{ij}^2 = c_i^2 + c_j^2$ and $\sin\varphi_{ij} = (c_i^2 - c_j^2)/\rho_{ij}^2$. The radius and spherical point therefore retain overall strength, amplitude allocation, and relative orientation. Numerical stabilization, including zero-response handling, is described in Appendix B.1.

Common scaling of the effective amplitudes changes only the radius, whereas amplifying either amplitude alone also changes latitude. The magnitude of the equatorial projection, $B_{ij} = \cos\varphi_{ij} = 2c_ic_j/(c_i^2 + c_j^2)$, gives a scale-invariant measure of amplitude balance that equals one for equal amplitudes and decreases with imbalance.

3.3 Response-Pair Spherical Attention

We use this balance measure to modulate the effective-amplitude product, preserving sensitivity to joint amplification while limiting the affinity gain from amplifying a candidate alone. Let $C_{ij} = c_ic_j$ and $P_{ij} = \cos^{2p}(\Delta\theta_{ij}/2)$, where $p \geq 1$ controls angular selectivity. For a balance exponent $\beta \geq 0$, define

$$K_{ij} = C_{ij}B_{ij}^{\beta}P_{ij} = \frac{\rho_{ij}^2}{2}\cos^{\beta+1}\varphi_{ij}\cos^{2p}\left(\frac{\Delta\theta_{ij}}{2}\right). \tag{6}$$

Since $C_{ij} = (\rho_{ij}^2/2)\cos\varphi_{ij}$, the amplitude product couples overall effective strength with amplitude balance. The additional factor B_{ij}^{β} strengthens balance selectivity for $\beta > 0$, while P_{ij} favors alignment and vanishes for opposed responses. This construction yields a symmetric, nonnegative affinity with $K_{ij} = K_{ji} \geq 0$.

We adopt $\beta = 1$, for which the amplitude term becomes the harmonic mean of the squared effective amplitudes. Holding c_i and the relative orientation fixed gives

$$K_{ij} = \frac{2c_i^2c_j^2}{c_i^2 + c_j^2}P_{ij}, \qquad \lim_{c_j\to\infty} K_{ij} = 2c_i^2P_{ij}. \tag{7}$$

For $P_{ij} > 0$, this limit is finite and nonzero, with its value set by the target response and directional agreement. Choosing $\beta < 1$ allows unbounded growth under the same amplification, while $\beta > 1$ eventually drives the affinity to zero.

To characterize the affinity's response across amplitude ratios, we compute its logarithmic sensitivity to c_j for $P_{ij} > 0$ as

$$\frac{\partial\log K_{ij}}{\partial\log c_j} = 2 - \frac{2c_j^2}{c_i^2 + c_j^2} = \frac{2c_i^2}{c_i^2 + c_j^2} = 1 + \sin\varphi_{ij}. \tag{8}$$

The latitude expression shows that relative sensitivity equals one for balanced responses and tends to zero as the candidate becomes much stronger than the target. Joint amplification preserves latitude and scales the affinity quadratically.

Let $\Omega(i)$ denote the aggregation neighborhood of target i, including i itself. We normalize the pairwise affinities K_{ij} over this neighborhood to obtain attention weights α_{ij} that determine each candidate's relative contribution. These weights aggregate the separately projected values $\boldsymbol{v}_j$ into a local message $\boldsymbol{m}_i$:

$$\alpha_{ij} = \frac{K_{ij}}{\sum_{k\in\Omega(i)} K_{ik}}, \qquad \boldsymbol{m}_i = \sum_{j\in\Omega(i)} \alpha_{ij}\boldsymbol{v}_j. \tag{9}$$

For $c_i > 0$, the self-affinity $K_{ii} = c_i^2$ ensures a positive normalization denominator. Multiplying all effective amplitudes in $\Omega(i)$ by $\lambda > 0$ scales each affinity by λ^2, so the normalized weights are unchanged.

To determine how the target amplitude influences candidate competition, we examine the weight ratio between two candidates.

Proposition 1 (Target-dependent candidate allocation)**.** *Let $\beta = 1$ and let $j, k \in \Omega(i) \setminus \{i\}$ be distinct candidates. Hold their amplitudes $c_j, c_k > 0$ and all response directions fixed, with $P_{ij}, P_{ik} > 0$. For $c_i > 0$, their weight ratio is*

$$R_{jk}(c_i) = \frac{\alpha_{ij}}{\alpha_{ik}} = \frac{c_j^2}{c_k^2} \frac{c_i^2 + c_k^2}{c_i^2 + c_j^2} \frac{P_{ij}}{P_{ik}}. \tag{10}$$

If $c_j > c_k$, this ratio increases strictly from P_{ij}/P_{ik} as $c_i \to 0^+$ to $(c_j/c_k)^2 P_{ij}/P_{ik}$ as $c_i \to \infty$.

Proof. The common row denominator cancels in α_{ij}/α_{ik}. Substituting Eq. 7 gives Eq. 10. For $c_j > c_k$, differentiation gives

$$\frac{\mathrm{d} \log R_{jk}}{\mathrm{d} c_i} = \frac{2c_i(c_j^2 - c_k^2)}{(c_i^2 + c_j^2)(c_i^2 + c_k^2)} > 0. \tag{11}$$

The two limits follow directly from Eq. 10. □

Proposition 1 shows that the target amplitude controls how strongly candidate amplitude differences affect their weight ratio. The ratio depends only on directional agreement in the weak-target limit and reflects more of the candidates' amplitude contrast as the target strengthens. This dependence on the target amplitude disappears when $\beta = 0$, because c_i cancels from the ratio. In HMF, the evolving target response thus provides a local reference for allocating neighboring spatial cues during hyperspectral reconstruction. Appendix A.1 extends the candidate-ratio analysis to general β, while Appendix A.2 derives bounds on nonself attention weights and analyzes message sensitivity to single-candidate amplification.

3.4 Training Objective

We penalize bandwise reconstruction errors with an ℓ_1 term and encourage agreement in spectral direction with a cosine-distance term. The overall training objective is

$$\mathcal{L}_1 = \frac{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_1}{HWC}, \quad \mathcal{L}_{\mathrm{spec}} = \frac{1}{HW} \sum_{i=1}^{HW} \left(1 - \frac{\widehat{\boldsymbol{x}}_i^\top \boldsymbol{x}_i}{\|\widehat{\boldsymbol{x}}_i\|_2 \|\boldsymbol{x}_i\|_2}\right), \quad \mathcal{L} = \mathcal{L}_1 + \lambda_{\mathrm{spec}} \mathcal{L}_{\mathrm{spec}}. \tag{12}$$

Here $\widehat{\boldsymbol{x}}_i, \boldsymbol{x}_i \in \mathbb{R}^C$ are the reconstructed and reference spectra at pixel i, respectively, and λ_{spec} controls the relative contribution of the spectral term.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate RPSA-Net on CAVE (Yasuma et al., 2010), Chikusei (Yokoya & Iwasaki, 2016), and Pavia Centre (Gamba, 2004) under simulated degradations and conduct real-world experiments on WHU-MHF (Zhu et al., 2024). CAVE contains 32 scenes, each of size 512×512 pixels, with 20 used for training and 12 for testing. For Chikusei, we extract 16 nonoverlapping 512×512 tiles from a 2048×2048 crop, using 12 for training and the remaining four for testing. For Pavia Centre, we train on the left 1024×459 image strip and test on four nonoverlapping 256×256 tiles from the right strip. For WHU-MHF, reduced-resolution samples from six ZY-1-02D scene pairs are used for training, while patches from a seventh scene pair are used for visual comparisons.

Data preparation. For the simulated benchmarks, we generate paired LR-HSI and HR-MSI observations from reference HR-HSIs, which serve as ground truth for reconstruction. LR-HSIs are obtained by applying Gaussian blurring with an $s \times s$ kernel ($\sigma = 2.0$) to the reference images, followed by spatial downsampling by a factor of $s \in \{4, 8\}$. HR-MSIs are synthesized by applying dataset-specific spectral response matrices to the same reference images.

Table 1: Quantitative comparison on CAVE, Chikusei, and Pavia Centre at $\times 4$ and $\times 8$. Bold and underlined values indicate the best and second-best results, respectively.

Scale	Method	CAVE				Chikusei				Pavia Centre			
		PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
$\times 4$	MSST	50.52	0.9964	2.107	0.917	48.51	0.9795	1.526	2.010	48.73	0.9846	3.106	1.641
	DSPNet	51.89	0.9970	1.953	0.781	49.23	0.9778	1.497	1.930	49.12	0.9851	3.029	1.604
	DCINN	52.94	0.9974	1.700	0.693	51.44	0.9862	1.263	1.688	49.70	0.9861	2.941	1.556
	DCT	52.52	0.9973	1.831	0.724	51.74	0.9860	1.256	1.688	49.95	0.9863	2.934	1.544
	CYformer	52.84	0.9974	1.799	0.698	51.81	0.9861	1.249	1.697	50.05	0.9864	2.923	1.538
	DPFormer	52.52	0.9972	1.834	0.726	52.06	0.9865	1.234	1.664	49.77	0.9862	2.934	1.575
	TRT	53.02	0.9975	1.707	0.682	51.98	0.9865	1.233	1.667	49.98	0.9852	3.008	1.630
	RPSA-Net	**53.56**	**0.9977**	**1.584**	**0.638**	**52.34**	**0.9874**	**1.201**	**1.638**	**50.25**	**0.9866**	**2.893**	**1.524**
$\times 8$	MSST	48.76	0.9954	2.465	0.564	46.76	0.9702	1.854	1.189	48.26	0.9842	3.172	0.843
	DSPNet	50.18	0.9960	2.209	0.493	46.63	0.9663	1.916	1.179	48.56	0.9847	3.112	0.828
	DCINN	50.47	0.9963	2.115	0.471	48.34	0.9764	1.629	1.042	48.89	0.9853	3.055	0.811
	DCT	50.54	0.9963	2.089	0.471	48.86	0.9765	1.591	1.031	49.20	0.9855	3.020	0.807
	CYformer	50.80	0.9963	2.054	0.459	48.91	0.9768	1.616	1.046	49.38	0.9856	3.032	0.798
	DPFormer	50.39	0.9962	2.133	0.475	48.73	0.9766	1.626	1.035	48.96	0.9854	3.048	0.810
	TRT	51.02	0.9965	2.008	0.442	49.27	0.9755	1.590	1.033	49.37	0.9856	3.017	0.801
	RPSA-Net	**51.41**	**0.9966**	**1.921**	**0.431**	**49.52**	**0.9791**	**1.538**	**1.007**	**49.49**	**0.9857**	**3.006**	**0.797**

Table 2: Computational efficiency on the Chikusei $\times 4$ test set. Time denotes end-to-end inference over the test set. Bold and underlined values indicate the best and second-best results, respectively.

Metric ↓	MSST	DSPNet	DCINN	DCT	CYformer	DPFormer	TRT	RPSA-Net
Params. (M)	100.59	6.13	5.64	8.44	**1.73**	5.12	9.79	4.28
MACs (G)	11,948	**442**	503	4,997	462	691	2,587	838.66
Time (s)	2.10	2.35	7.84	17.66	4.00	3.26	4.86	**1.16**
Mem. (GiB)	5.07	4.53	4.15	7.85	**3.50**	6.04	10.23	5.48

Implementation details. The main reconstruction experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. We train RPSA-Net for 600 epochs using Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and a batch size of 32. The learning rate remains fixed at 10^{-4} throughout training. We use a feature width of $C_f = 192$ and $H_a = 4$ attention heads. We set the balance exponent to $\beta = 1$, the directional exponent to $p = 2$, the window size to $w_r = 8$, and the spectral loss weight to $\lambda_{\text{spec}} = 0.05$. We evaluate reconstruction quality using PSNR, SSIM, SAM, and ERGAS, and computational cost using parameter count (Params.), multiply-accumulate operations (MACs), inference time, and peak GPU memory (Mem.).

4.2 Comparison with State-of-the-Art Methods

We benchmark RPSA-Net against seven representative HMF methods: MSST (Jia et al., 2023), DSPNet (Sun et al., 2023), DCT (Ma et al., 2024), DCINN (Wang et al., 2024), CYformer (Chen et al., 2025), DPFormer (Liu et al., 2026), and TRT (Xu et al., 2026b).

Quantitative comparison. Table 1 reports the reconstruction results on CAVE, Chikusei, and Pavia Centre at $\times 4$ and $\times 8$. RPSA-Net outperforms all competing methods across all four metrics in all six settings, showing consistent advantages across datasets and upsampling factors. At $\times 8$, RPSA-Net improves PSNR over the runner-up by 0.39, 0.25, and 0.11 dB on CAVE, Chikusei, and Pavia Centre, respectively. In terms of spectral fidelity, it reduces SAM by 0.087, 0.052, and 0.011 relative to the second-best results on the three datasets, respectively. These improvements in both spatial reconstruction and spectral fidelity suggest that response-pair modeling helps recover spatial detail while preserving spectral information.

Table 2 further compares computational cost on Chikusei at $\times 4$. RPSA-Net uses 4.28M parameters, the second-lowest count among the compared methods, and achieves the shortest inference time of 1.16 s over the entire test set. These results show that its reconstruction gains are achieved with a relatively compact model and low inference latency. While some baselines require fewer MACs

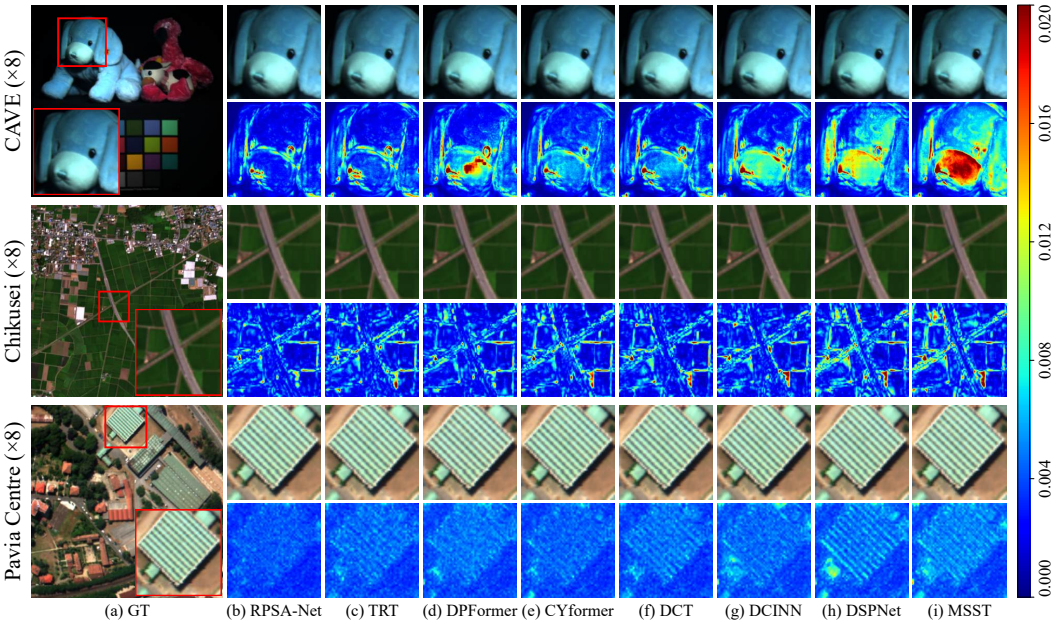


Figure 4: Visual comparisons on CAVE, Chikusei, and Pavia Centre at $\times 8$. False-color results and absolute-error maps are shown for each method, with enlarged views of the highlighted regions.

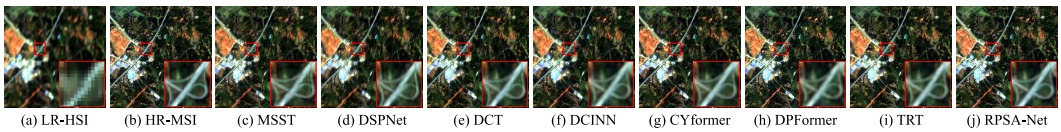


Figure 5: Visual comparisons on the WHU-MHF dataset.

or less memory, RPSA-Net combines more accurate reconstruction with faster inference, offering a favorable trade-off between reconstruction quality and computational efficiency.

Qualitative comparison. Figure 4 complements the quantitative results with false-color reconstructions and absolute-error maps on CAVE, Chikusei, and Pavia Centre at $\times 8$. On CAVE, RPSA-Net better preserves the contours and facial details of the stuffed toy, retaining both its shape and finer features. On Chikusei, it recovers clearer boundaries along the intersecting roads while better preserving the surrounding field textures. On Pavia Centre, it more faithfully reproduces the regular stripe pattern on the building roof, maintaining the continuity of these repetitive structures. The corresponding error maps show smaller residuals in the highlighted regions, supporting these visual observations and indicating improved reconstruction of both object boundaries and local textures.

4.3 Ablation Studies and Analysis

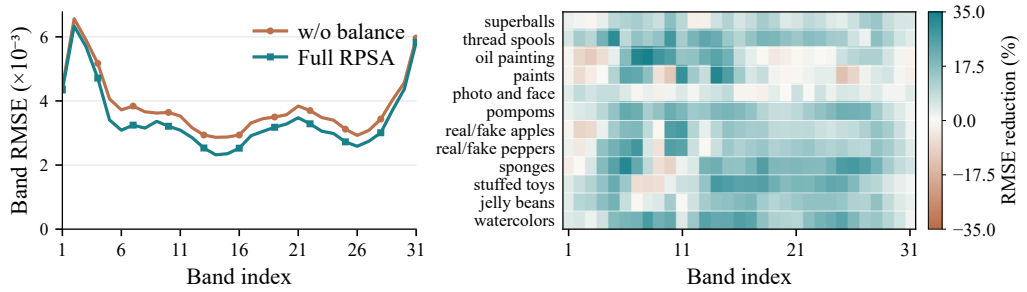
Ablation study. Table 3 evaluates the contributions of response strength and amplitude balance on Chikusei $\times 4$. All attention variants share the backbone, response projection, value pathway, windows, and refinement schedule. Local fusion serves as a baseline without attention-based aggregation, while standard dot-product attention computes scaled query–key dot products and normalizes the resulting scores with softmax. Compared with the direction-only affinity $K_{ij} = P_{ij}$, incorporating the effective-amplitude product through $K_{ij} = C_{ij}P_{ij}$ increases PSNR from 52.10 to 52.17 dB and reduces SAM from 1.235 to 1.229. Adding the balance factor to obtain $K_{ij} = C_{ij}B_{ij}P_{ij}$ further improves all four metrics, raising PSNR to 52.34 dB and lowering SAM to 1.201. Full RPSA achieves the best results across all four metrics, with PSNR gains of 0.54 and 0.31 dB over local fusion and dot-product attention, respectively. These results support the use of response strength and amplitude balance alongside directional agreement to improve reconstruction quality.

Table 3: Affinity ablation on Chikusei $\times 4$. Bold and underlined values indicate the best and second-best results, respectively.

Variant	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
Local fusion	51.80	0.9861	1.276	1.707
Dot-product	52.03	0.9866	1.247	1.684
Direction only	52.10	0.9865	1.235	1.673
w/o balance	<u>52.17</u>	<u>0.9870</u>	<u>1.229</u>	<u>1.662</u>
Full RPSA	**52.34**	**0.9874**	**1.201**	**1.638**

Table 4: Effect of β and p on Chikusei $\times 4$. Bold and underlined values indicate the best and second-best results, respectively.

β	p	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
0	2	52.17	0.9870	1.229	1.662
0.5	2	52.29	<u>0.9871</u>	1.210	1.647
1	1	52.20	0.9868	1.222	1.659
1	2	**52.34**	**0.9874**	**1.201**	**1.638**
1	3	<u>52.30</u>	0.9870	<u>1.207</u>	<u>1.644</u>
2	2	52.22	0.9869	1.219	1.654



(a) Spectral error (lower is better)

(b) Error reduction by scene and band

Figure 6: Spectral reconstruction errors of Full RPSA and its variant without amplitude balance across all 12 CAVE test scenes ($\times 4$). (a) Band-wise RMSE averaged across scenes. (b) Relative RMSE reduction, where teal and brown indicate lower and higher errors for Full RPSA, respectively.

Parameter sensitivity. Table 4 examines the sensitivity to the balance exponent β and the directional exponent p on Chikusei $\times 4$, with all other settings fixed. With $p = 2$, increasing β from 0 to 1 improves PSNR from 52.17 to 52.34 dB and reduces SAM from 1.229 to 1.201, whereas further increasing β to 2 degrades both metrics. This trend favors a moderate degree of amplitude balance modulation. With $\beta = 1$, increasing p from 1 to 2 improves both metrics, while $p = 3$ yields slightly lower performance, suggesting that stronger angular selectivity does not necessarily improve reconstruction. We therefore choose $(\beta, p) = (1, 2)$ as the default setting.

Amplitude balance comparison. Figure 6 compares the spectral reconstruction errors of Full RPSA and its variant without amplitude balance on CAVE $\times 4$. As shown in Fig. 6(a), both variants exhibit similar overall error profiles, while Full RPSA achieves lower band-wise RMSE across the spectral range. The heatmap in Fig. 6(b) complements these scene-averaged curves by showing error reductions across most individual scene-band combinations. The observed improvements are consistent with the motivation for amplitude balance, which evaluates candidate response strength relative to the target when regulating neighboring contributions for spectral reconstruction.

5 Conclusion

We presented response-pair spherical attention for hyperspectral and multispectral image fusion. By combining overall effective amplitude, relative amplitude allocation, and directional agreement in its affinity, RPSA lets the target response regulate neighboring feature contributions while limiting the gain in unnormalized affinity from amplifying a candidate alone. We incorporated RPSA into RPSA-Net, which alternates high-resolution refinement with low-resolution feedback to reconstruct the HR-HSI from LR-HSI and HR-MSI observations. Experiments on three simulated benchmarks and a real-world dataset demonstrate state-of-the-art spectral fidelity and spatial reconstruction.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising the manuscript, including improving sentence structure, correcting grammar, and refining wording for clarity and readability.

REPRODUCIBILITY STATEMENT

The proposed method is described in Section 3, with data preparation and experimental settings in Section 4.1. Appendices A and B provide the theoretical derivations and numerical implementation details, respectively.

REFERENCES

Shi Chen, Lefei Zhang, and Liangpei Zhang. Cyclic cross-modality interaction for hyperspectral and multispectral image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (1):741–753, 2025.

Qihang Fan, Huaibo Huang, Yuang Ai, and Ran He. Rectifying magnitude neglect in linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21505–21514, 2025. URL https://arxiv.org/abs/2507.00698.

Chuangjie Fang, Yan Li, Junwei Zhu, Hong Qiu, Honghui Xu, and Jianwei Zheng. Subspace-frequency regularization for hyperspectral image super-resolution. *Knowledge-Based Systems*, 339, 2026. doi: 10.1016/j.knosys.2026.115620. Art. no. 115620.

Jian Fang, Jingxiang Yang, Abdolraheem Khader, and Liang Xiao. MIMO-SST: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. doi: 10.1109/TGRS.2024.3361553. Art. no. 5510020.

Paolo Gamba. A collection of data for urban area characterization. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 1, pp. 69–72, 2004.

Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. QUEST: A robust attention formulation using query-modulated spherical attention. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HkztQWZfl2.

Shuying Huang, Mingyang Ren, Yong Yang, Xiaozheng Wang, and Yingzhi Wei. MFTN: A multi-scale feature transfer network based on IMatchFormer for hyperspectral image super-resolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 20063–20072, 2024.

Sen Jia, Zhichao Min, and Xiyou Fu. Multiscale spatial-spectral transformer network for hyperspectral and multispectral image fusion. *Information Fusion*, 96:117–129, 2023.

Baisong Li, Xingwang Wang, and Haixiao Xu. PIF-Net: Ill-posed prior guided multispectral and hyperspectral image fusion via invertible Mamba and fusion-aware LoRA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5964–5972, 2026a. doi: 10.1609/aaai.v40i8.37519. URL https://ojs.aaai.org/index.php/AAAI/article/view/37519.

Jiaxin Li, Ke Zheng, Lianru Gao, Li Ni, Min Huang, and Jocelyn Chanussot. Model-informed multistage unsupervised network for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. doi: 10.1109/TGRS.2024.3391014. Art. no. 5516117.

Jiaxin Li, Ke Zheng, Lianru Gao, Zhu Han, Zhi Li, and Jocelyn Chanussot. Enhanced deep image prior for unsupervised hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025. doi: 10.1109/TGRS.2025.3531646. Art. no. 5504218.

Wei Li, Jieyuan Pei, Junnan Xu, Xuanfeng Ding, Junwei Zhu, Wanjun Chen, and Jianwei Zheng. Solving spatial-spectral fusion with latent spectral operators. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026b. URL https://weili419.github.io/latent-spectral-operators/.

Yan Li, Chuangjie Fang, Zhijie Chen, Yiqun Meng, Honghui Xu, Hao Wang, and Yi-Peng Liu. Block term decomposition-guided frequency Mamba modulation for hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026c. doi: 10.1109/TGRS.2026.3699818. Art. no. 5516414.

Shuaiqi Liu, Ruixia Cai, Huanru Yue, Bing Li, Yudong Zhang, and Siyuan Liu. Dual-phase transformer with spectral-spatial synergy for hyperspectral image fusion. *Pattern Recognition*, 175: 113116, 2026.

Yuanye Liu, Renwei Dian, and Shutao Li. Low-rank transformer for high-resolution hyperspectral computational imaging. *International Journal of Computer Vision*, 133(2):809–824, 2025a. doi: 10.1007/s11263-024-02203-7.

Yuanye Liu, Jinyang Liu, Renwei Dian, and Shutao Li. A selective re-learning mechanism for hyperspectral fusion imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7437–7446, 2025b. URL https://github.com/YuanyeLiu/SRLF-Net.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022. URL https://arxiv.org/abs/2111.09883.

Qing Ma, Junjun Jiang, Xianming Liu, and Jiayi Ma. Reciprocal transformer for hyperspectral and multispectral image fusion. *Information Fusion*, 104:102148, 2024.

Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. PolaFormer: Polarity-aware linear attention for vision transformers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kN6MFmKUSK.

Weikang Meng, Yadan Luo, Liangyu Huo, Yingjian Li, Yaowei Wang, Xin Li, and Zheng Zhang. Norm×Direction: Restoring the missing query norm in vision linear attention. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.21137.

Yucheng Sun, Han Xu, Yong Ma, Minghui Wu, Xiaoguang Mei, Jun Huang, and Jiayi Ma. Dual spatial-spectral pyramid network with transformer for hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wu Wang, Liang-Jian Deng, Ran Ran, and Gemine Vivone. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4):1029–1054, 2024.

Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, Liang-Jian Deng, Jocelyn Chanussot, and Gemine Vivone. Fully-connected transformer for multi-source image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2071–2088, 2025. doi: 10.1109/TPAMI.2024.3523364.

Honghui Xu, Chuangjie Fang, Yiqun Meng, Jiawei Jiang, Sixian Chan, Shiqing Zhang, and Jianwei Zheng. TPTransformer: Tensor-tensor product transformer for hyperspectral image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 1670–1679, 2026a.

Honghui Xu, Junwei Zhu, Yubin Gu, Yueqian Quan, Chuangjie Fang, Hong Qiu, and Jianwei Zheng. TRT: Harnessing tensor ring transformer for hyperspectral image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11232–11240, 2026b.

Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5(5):5, 2016.

Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie Guo, Xinbo Gao, and Jing Zhang. ESSAformer: Efficient transformer for hyperspectral image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23073–23084, 2023. URL https://arxiv.org/abs/2307.14010.

Mingjin Zhang, Zhongkai Yang, and Fei Gao. CO^2IF: Language-bridging hyperspectral-multispectral image fusion with coordinated and cross-modal optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12654–12662, 2026. doi: 10.1609/aaai.v40i15.38261.

Jian Zhu, He Wang, Yang Xu, Zebin Wu, and Zhihui Wei. Self-learning hyperspectral and multispectral image fusion via adaptive residual guided subspace diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17862–17871, 2025.

Zengliang Zhu, Xinyu Wang, Guanzhong Li, and Yanfei Zhong. A self-supervised spaceborne multispectral and hyperspectral image fusion unrolling network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

A Theoretical Analysis

This appendix further analyzes how effective amplitude allocation affects candidate competition and the resulting attention message. We use the ideal affinity and normalization in Eqs. 6 and 9, with $\beta \geq 0$ and $p \geq 1$. Throughout, we assume $c_i > 0$, unit response directions, and a finite neighborhood $\Omega(i)$ containing the target. The effective amplitudes defined in Eq. 3 are varied directly while all response directions remain fixed. The value vectors are additionally held fixed when analyzing message sensitivity. The corresponding stabilized implementation, including zero-response handling, is specified in Appendix B.1.

A.1 Amplitude Allocation and Candidate Competition

To compare response pairs with the same amplitude product and relative orientation, fix $C = c_i c_j > 0$ and $P = P_{ij} > 0$. Writing the degree of imbalance as $q = |\log(c_i/c_j)|$ gives

$$B_{ij} = \frac{2}{c_i/c_j + c_j/c_i} = \operatorname{sech} q, \qquad K(q) = CP \operatorname{sech}^{\beta} q. \tag{13}$$

Since $\mathrm{d}K/\mathrm{d}q = -\beta K(q) \tanh q$, the affinity strictly decreases with imbalance for $\beta > 0$ and reaches its maximum CP at equal amplitudes. At $\beta = 0$, it is constant across these equal-product pairs. Exchanging the target and candidate preserves q, so the balance term distinguishes the degree of imbalance symmetrically.

We next fix the candidate amplitudes and all response directions, and examine how varying the target amplitude changes the candidates' relative weights. To extend Proposition 1 to general β, consider two distinct nonself candidates $j, k \in \Omega(i) \setminus \{i\}$ with fixed positive amplitudes $c_j, c_k > 0$ and $P_{ij}, P_{ik} > 0$. Substituting $B_{ij} = 2c_i c_j/(c_i^2 + c_j^2)$ into Eq. 6 and cancelling the common row denominator gives

$$R_{jk}(c_i) = \frac{\alpha_{ij}}{\alpha_{ik}} = \left(\frac{c_j}{c_k}\right)^{1+\beta} \left(\frac{c_i^2 + c_k^2}{c_i^2 + c_j^2}\right)^{\beta} \frac{P_{ij}}{P_{ik}}. \tag{14}$$

Differentiating with respect to the target amplitude yields

$$\frac{\mathrm{d} \log R_{jk}}{\mathrm{d}c_i} = \frac{2\beta c_i (c_j^2 - c_k^2)}{(c_i^2 + c_k^2)(c_i^2 + c_j^2)}. \tag{15}$$

For $\beta > 0$ and $c_j > c_k$, increasing the target amplitude therefore shifts the relative allocation toward the stronger candidate. The limiting ratios are

$$\begin{aligned} \lim_{c_i \to 0^+} R_{jk}(c_i) &= \left(\frac{c_j}{c_k}\right)^{1-\beta} \frac{P_{ij}}{P_{ik}}, \\ \lim_{c_i \to \infty} R_{jk}(c_i) &= \left(\frac{c_j}{c_k}\right)^{1+\beta} \frac{P_{ij}}{P_{ik}}. \end{aligned} \tag{16}$$

At $\beta = 1$, these expressions recover the directional ratio and amplitude-weighted ratio in Proposition 1. At $\beta = 0$, target amplitude cancels from the ratio. Individual weights still compete with the self-affinity $K_{ii} = c_i^2$, which dominates the row as $c_i \to \infty$ when the candidate amplitudes remain fixed.

The change in relative allocation can also reverse candidate ordering. For $\beta = 1$, write $a = c_i$, $b = c_j > c = c_k$, and $h = P_{ij}/P_{ik}$. The ratio increases continuously from h to $(b^2/c^2)h$, so a unique crossing occurs exactly when $c^2/b^2 < h < 1$. Solving $R_{jk}(a_*) = 1$ gives

$$a_*^2 = \frac{b^2 c^2 (1 - h)}{h b^2 - c^2}. \tag{17}$$

Below this threshold, the better-aligned, weaker candidate receives more weight. Above it, the stronger candidate receives more weight. Thus the target amplitude determines the tradeoff between candidate strength and directional agreement even when both candidate responses are fixed.

A.2 One-Sided Amplification and Message Sensitivity

We next fix the target amplitude $a = c_i > 0$ and vary the amplitude $t = c_j$ of one nonself candidate, keeping all directions and other candidate amplitudes fixed. For $P = P_{ij} > 0$, Eq. 6 gives

$$K_{ij}(t) \sim 2^{\beta} a^{1+\beta} P\, t^{1-\beta} \qquad \text{as } t \to \infty. \tag{18}$$

The exponent $\beta = 1$ therefore gives a finite, nonzero limit, whereas $\beta < 1$ permits unbounded growth and $\beta > 1$ drives the affinity to zero. We now examine how the adopted exponent transfers this saturation to normalized weights and messages.

For $\beta = 1$, let $f(t) = 2a^2t^2P/(a^2 + t^2)$ and $U = 2a^2P$, as in Eq. 7. The remaining affinity mass is $A_i = \sum_{k\in\Omega(i)\setminus\{j\}} K_{ik} \geq K_{ii} = a^2$. Since $f(t) \leq U$ and $x/(A_i + x)$ is increasing for $x \geq 0$,

$$\alpha_{ij}(t) = \frac{f(t)}{A_i + f(t)} \leq \frac{U}{A_i + U} \leq \frac{2P}{1 + 2P}. \tag{19}$$

The first bound is the exact limit as $t \to \infty$. Full alignment gives the upper bound $2/3$, approached when the self term is the only other positive affinity. Lower directional agreement or additional positive affinity tightens the bound. The effect of numerical stabilization on this ideal bound is analyzed in Appendix B.2. For comparison, with $P > 0$, the normalized weight tends to one when $\beta < 1$ and to zero when $\beta > 1$. When $P = 0$, the candidate has zero affinity and weight for every amplitude.

To trace the effect on the aggregated message, also hold all value vectors fixed and define

$$\boldsymbol{b}_i = \sum_{k\in\Omega(i)\setminus\{j\}} K_{ik}\boldsymbol{v}_k, \qquad \boldsymbol{m}_i(t) = \frac{\boldsymbol{b}_i + f(t)\boldsymbol{v}_j}{A_i + f(t)}. \tag{20}$$

Differentiating the affinity and normalized candidate weight gives

$$f'(t) = \frac{4a^4tP}{(a^2 + t^2)^2}, \qquad \frac{\mathrm{d}\alpha_{ij}}{\mathrm{d}t} = \frac{A_i f'(t)}{(A_i + f(t))^2}. \tag{21}$$

For $P > 0$, the candidate weight increases with amplitude, and its derivative is $\mathcal{O}(t^{-3})$ as $t \to \infty$. Applying the quotient rule to Eq. 20 then gives

$$\frac{\mathrm{d}\boldsymbol{m}_i}{\mathrm{d}t} = \frac{f'(t)}{A_i + f(t)} \left(\boldsymbol{v}_j - \boldsymbol{m}_i(t)\right). \tag{22}$$

Let $D_i = \max_{k\in\Omega(i)} \|\boldsymbol{v}_j - \boldsymbol{v}_k\|_2$. The message is a convex combination of these fixed values, so $\|\boldsymbol{v}_j - \boldsymbol{m}_i(t)\|_2 \leq D_i$. Using $A_i + f(t) \geq a^2$ gives

$$\left\| \frac{\mathrm{d}\boldsymbol{m}_i}{\mathrm{d}t} \right\|_2 \leq \frac{4a^2tP}{(a^2 + t^2)^2} D_i = \mathcal{O}(t^{-3}). \tag{23}$$

The message approaches $(\boldsymbol{b}_i + U\boldsymbol{v}_j)/(A_i + U)$ with diminishing sensitivity to further candidate amplification. This extends the pairwise sensitivity in Eq. 8 to the influence transmitted through normalized weighting, with the value content held fixed.

B Numerical Implementation and Reproducibility

We first specify the stabilized operator and its treatment of small and zero responses, then derive the resulting weight bounds and describe the fusion block and window evaluation. Each head computes the stabilized operator independently. We omit the head index in its formulas.

B.1 Stabilized Affinities and Zero Responses

Let $\boldsymbol{r}_i = \boldsymbol{W}_r\boldsymbol{x}_i$ and $e_i = \|\boldsymbol{r}_i\|_2^2$. We use $\epsilon = 10^{-6}$ throughout the operator and a learnable attenuation parameter $\widehat{\tau}$ for each head. The same safeguard ϵ is used in the response calculation, balance denominator, and row normalization. The stabilized quantities are computed as

$$\begin{aligned} \bar{\rho}_i^{\epsilon} &= \sqrt{\max(e_i, \epsilon)}, \quad \widetilde{\boldsymbol{n}}_i = \frac{\boldsymbol{r}_i}{\bar{\rho}_i^{\epsilon}}, \\ \tau &= \mathrm{softplus}(\widehat{\tau}), \qquad g_i = \frac{e_i}{e_i + \tau + \epsilon}, \quad c_i = \bar{\rho}_i^{\epsilon} g_i. \end{aligned} \tag{24}$$

Here c_i is the stabilized effective amplitude. The direction has unit norm when $e_i \geq \epsilon$ and squared norm e_i/ϵ below this threshold, so the norm floor also attenuates low-energy directions.

Writing $\mathrm{clip}_{[a,b]}$ for clipping to $[a, b]$, the implemented factors and weights are

$$\begin{aligned} d_{ij} &= \mathrm{clip}_{[-1,1]}(\widetilde{\boldsymbol{n}}_i^\top \widetilde{\boldsymbol{n}}_j), \quad & \widetilde{P}_{ij} &= \left[\mathrm{clip}_{[0,1]}\left(\frac{1+d_{ij}}{2}\right)\right]^p, \\ \widetilde{B}_{ij} &= \frac{2c_i c_j}{c_i^2 + c_j^2 + \epsilon}, & \widetilde{K}_{ij} &= c_i c_j \widetilde{B}_{ij}^{\beta} \widetilde{P}_{ij}, \\ R_i &= \sum_{k \in \Omega(i)} \widetilde{K}_{ik}, & \widetilde{\alpha}_{ij} &= \frac{\widetilde{K}_{ij}}{\max(R_i, \epsilon)}. \end{aligned} \tag{25}$$

Clipping preserves the valid ranges of the directional factors. At $\beta = 0$, the balance factor is set to one, including when $\widetilde{B}_{ij} = 0$. The amplitude product makes every interaction with a zero response vanish. A zero target response therefore gives a zero affinity row and a zero attention message. More generally,

$$\sum_{j \in \Omega(i)} \widetilde{\alpha}_{ij} = \frac{R_i}{\max(R_i, \epsilon)} \leq 1. \tag{26}$$

Rows with $R_i \geq \epsilon$ sum to one, while smaller row sums attenuate the message proportionally. The value residual in Eq. 30 remains active in both cases. The stabilized direction norms also enter $\widetilde{P}_{ij}$, giving the self-directional factor

$$\widetilde{P}_{ii} = \left(\frac{1 + e_i/\max(e_i, \epsilon)}{2}\right)^p. \tag{27}$$

B.2 Effect of Numerical Stabilization on Weight Bounds

The numerical safeguards in Appendix B.1 change the self-affinity. We therefore adapt the ideal candidate-weight bound from Appendix A.2 (Eq. 19) to the implemented operator. For $\beta = 1$, define

$$H_i = \widetilde{K}_{ii} = \frac{2c_i^4}{2c_i^2 + \epsilon} \widetilde{P}_{ii}, \qquad U_{ij} = 2c_i^2 \widetilde{P}_{ij}. \tag{28}$$

For a nonself candidate j and a neighborhood containing i, nonnegativity gives $R_i \geq H_i + \widetilde{K}_{ij}$ and $\widetilde{K}_{ij} = U_{ij} c_j^2/(c_i^2 + c_j^2 + \epsilon) \leq U_{ij}$. Since $t/\max(H_i + t, \epsilon)$ is nondecreasing for $t \geq 0$, these inequalities yield

$$\widetilde{\alpha}_{ij} \leq \frac{\widetilde{K}_{ij}}{\max(H_i + \widetilde{K}_{ij}, \epsilon)} \leq \frac{U_{ij}}{\max(H_i + U_{ij}, \epsilon)}. \tag{29}$$

The bound includes $c_i = 0$ and approaches the ideal expression when the stabilizers are negligible and the directions have unit norm. Near the numerical scale, the actual self-affinity remains essential. For example, let $e_i \geq \epsilon$, $c_i^2 = \epsilon$, and retain only the target and one aligned candidate. Then $H_i = 2\epsilon/3$ and $\widetilde{K}_{ij} \to 2\epsilon$ as the candidate amplitude grows, so its weight approaches $3/4$. Equation 29 covers this regime, where the ideal aligned bound $2/3$ no longer applies.

B.3 Fusion Block and Window Evaluation

Fusion block. For an input map $\boldsymbol{F} \in \mathbb{R}^{C_f \times H_f \times W_f}$, spatial tokenization $\mathcal{Q}$ followed by LayerNorm supplies the features $\boldsymbol{x}_i$. A bias-free linear layer produces $2H_a$ response coordinates, and a separate linear layer with bias produces C_f value coordinates. Splitting these outputs into H_a heads gives each head a two-dimensional response and a value vector $\boldsymbol{v}_i^h \in \mathbb{R}^{d_h}$, where $d_h = C_f/H_a$. Each head computes a message $\widetilde{\boldsymbol{m}}_i^h$ using the stabilized weights in Eq. 25, adds it to the value stream with a learned scale γ_h, and contributes to the output projection

$$\widetilde{\boldsymbol{v}}_i^h = \boldsymbol{v}_i^h + \gamma_h \widetilde{\boldsymbol{m}}_i^h, \qquad \boldsymbol{o}_i = \boldsymbol{W}_o \,\mathrm{Concat}_h\big(\widetilde{\boldsymbol{v}}_i^h\big) + \boldsymbol{b}_o. \tag{30}$$

Let $\boldsymbol{O}$ collect the output tokens and set $\mathcal{A}(\boldsymbol{F}) = \mathcal{Q}^{-1}(\boldsymbol{O})$ to restore the spatial layout. Spatial dropout and a convolutional mixer Φ then combine this output with the input through

$$\mathcal{G}(\boldsymbol{F}) = \boldsymbol{F} + \Phi([\mathrm{Drop}(\mathcal{A}(\boldsymbol{F})), \boldsymbol{F}])\,. \tag{31}$$

Brackets denote channel concatenation. The mixer Φ uses a 1×1 convolution at width $2C_f$, a 3×3 depthwise convolution followed by a 1×1 pointwise convolution at the same width, and a final 1×1 projection to C_f. The first convolution and the depthwise–pointwise pair are each followed by LeakyReLU and spatial dropout.

Window evaluation and complexity. To evaluate the messages, we reshape the directions, effective amplitudes, and values into spatial maps and pad them for partitioning into $w_r \times w_r$ windows. For each target i, the positions in its assigned window form the neighborhood $\Omega(i)$. A window offset δ is implemented by zero padding without cyclic wrapping. With $a = \delta \bmod w_r$, the top and left padding widths are a, and the bottom and right widths are

$$b = (w_r - (H_f + a) \bmod w_r) \bmod w_r, \qquad r = (w_r - (W_f + a) \bmod w_r) \bmod w_r. \tag{32}$$

The padded dimensions are $H_p = H_f + a + b$ and $W_p = W_f + a + r$. Padding is applied to the computed directions, amplitudes, and values, so padded candidates have exactly zero amplitude and contribute no affinity. Each window includes its diagonal and requires no separate attention mask. The implementation multiplies the affinity matrix by the values, then divides each message row by $\max(R_i, \epsilon)$. Reversing the partition and cropping to rows $[a, a + H_f)$ and columns $[a, a + W_f)$ restores the original domain before the residual and projection in Eq. 30.

With $N_p = H_p W_p$ padded positions and w_r^2 positions per window, affinity evaluation costs $\mathcal{O}(H_a N_p w_r^2)$ and value aggregation costs $\mathcal{O}(H_a N_p w_r^2 d_h)$. The resulting total time and affinity storage are

$$T_{\mathrm{window}} = \mathcal{O}(H_a N_p w_r^2 d_h), \qquad M_{\mathrm{window}} = \mathcal{O}(H_a N_p w_r^2). \tag{33}$$

These costs exclude feature projections and convolutional mixing.