RECURRENT OPERATOR TRANSPORT WITH SPECTRAL ADAPTATION FOR UNREGISTERED HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unregistered hyperspectral image (HSI) super-resolution transfers high-resolution spatial detail across misaligned sensors while preserving spectral fidelity. Existing methods typically transfer detail through cross-modal feature fusion, leaving the spatial reconstruction process implicit in the interaction between reference and HSI features. In this paper, we propose Recurrent Operator Transport with Spectral Adaptation (**ROTSA**), a framework that explicitly learns spatial reconstruction operators from the reference and transfers them to HSI restoration. Specifically, a self-reconstructing reference branch models cross-scale spatial transformations using neighborhood weights and an affine detail state, capturing local propagation relationships and missing high-frequency detail. Structural correspondence then transports these operators to the HSI grid, where they recurrently guide spatial restoration. Besides, to adapt the transferred correction to scene-dependent spectral characteristics, we further introduce adaptive spectral capacity, which represents the correction in a complete spectral basis estimated from the low-resolution HSI and continuously adjusts its contribution across the basis. This preserves the reconstruction supported by the observed HSI while allowing reference guidance to adapt to the spectral content of each scene. Experiments on simulated and real datasets demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Hyperspectral images (HSIs) capture rich spectral information for material analysis and scene understanding (Im & Jensen, 2008; Pan et al., 2003), but their spatial resolution is often limited by the trade-off between spectral and spatial sampling. Single-image HSI super-resolution reconstructs spatial detail from learned image priors (Liang et al., 2023b; Yang et al., 2025; 2026), while fusion-based methods further exploit a high-resolution RGB or multispectral image to provide complementary spatial information (Liu et al., 2025a;b; Song et al., 2026). However, the HSI and reference are often acquired by different sensors and are not spatially registered in practical scenarios.

Recently, unregistered HSI super-resolution (SR) methods (Zheng et al., 2022; Lai et al., 2024; Zhang et al., 2025; 2026) have explored how to leverage mismatched high-resolution (HR) details to enhance low-resolution (LR) HSI via correspondence estimation and feature fusion. Despite recent progress, differences in viewpoint and sensor response complicate reference guidance, making it challenging to transfer useful spatial representations across misaligned sensors and use them effectively for HSI super-resolution.

Existing methods typically inject reference features into HSI representations through cross-modal fusion, leaving spatial reconstruction implicit in the interaction between modalities (Lai et al., 2024; Zhang et al., 2025; 2026). Although correspondence determines where reference information is sampled, the fusion module must still learn how to use spatial structure alongside sensor-dependent appearance. Under misregistration, displaced reference boundaries further complicate this process. The correction responses in Fig. 1 therefore motivate explicitly learning transferable spatial relationships, beyond estimating correspondence alone. Moreover, the usefulness of transferred detail depends on the observed spectra. The scene-dependent spectral ranks in Fig. 2 suggest that one fixed subspace may not suit different inputs. Rank alone also leaves correction strength undeter-

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.

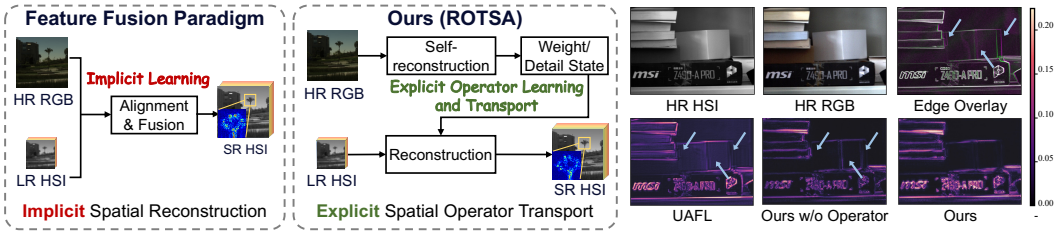

Figure 1: Feature fusion versus explicit spatial operator transport. Left: ROTSA learns reconstruction operators from the reference before applying them to HSI features. Right: the edge overlay highlights misalignment between RGB (magenta) and HSI (green) boundaries. The lower panels show the total output-domain response induced by the reference, reporting the RMS response across HSI bands. Operator transport reduces responses around displaced reference structures, supporting more selective spatial guidance (see Appendix B.3).

mined, while hard truncation can discard components supported by the HSI. This motivates adapting reference corrections across spectral modes while retaining the complete HSI estimate.

In this paper, we propose a Recurrent Operator Transport with Spectral Adaptation (**ROTSA**) framework. Specifically, coarse and fine-scale references provide a reconstruction target without cross-sensor misalignment. The reference self-reconstruction branch learns neighborhood weights for local propagation and an affine state for detail missing at coarse resolution. Structural correspondence transfers both fields to HSI coordinates, where the weights operate on HSI features and an HSI-conditioned gate controls the added detail. A shared recurrent module applies these operations across scales while maintaining a separate estimate from the LR HSI.

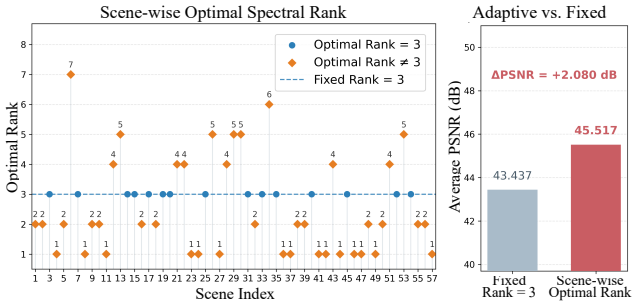

Figure 2: Scene-wise optimal spectral rank versus fixed rank. Following UAFL (Zhang et al., 2026), combining LR-HSI bases with HR-HSI coefficients shows that the preferred rank varies across scenes on REAL (Lai et al., 2024), motivating adaptation to scene-specific spectral characteristics.

We further adapt the transferred correction in the spectral domain by estimating a complete basis from global and local variation in the LR-HSI. Scene and correction statistics determine continuous weights over this basis. Weighting the difference between guided and HSI predictions adapts reference detail to scene spectra while retaining the underlying HSI estimate. Extensive experiments on simulated and real datasets demonstrate that ROTSA outperforms SOTA approaches, improving both spatial reconstruction and spectral fidelity. Our contributions are summarized as follows:

- We propose the ROTSA framework for unregistered HSI super-resolution that combines spatial operator transport with adaptive spectral capacity, enabling reliable reference guidance while preserving spectral fidelity.
- We develop recurrent spatial operator transport, which learns cross-scale spatial transformations with self-reconstructing operators and maps them to the HSI grid via structural correspondence, enabling effective spatial guidance under misregistration.
- We introduce adaptive spectral capacity, which continuously adjusts transferred corrections over a complete spectral basis estimated from the low-resolution HSI, enabling scene-adaptive guidance while preserving spectral fidelity.

2 Related Work

Registered HSI fusion. Registered HSI fusion combines complementary spatial and spectral information under an alignment assumption. Early approaches formulate reconstruction through obser-

vation models and hand-crafted priors, including coupled factorization and subspace regularization (Akhtar et al., 2014; 2015; Kawakami et al., 2011; Lanaras et al., 2015; Yokoya et al., 2011; Simoes et al., 2014). Learning-based methods further explore convolution and attention (Fu et al., 2019; Zhang et al., 2024; Liu et al., 2025a;b; Song et al., 2026). Despite their different reconstruction strategies, these methods rely on aligned spatial correspondence when integrating reference detail, an assumption that is difficult to maintain in real scenarios.

Unregistered HSI fusion. Unregistered methods incorporate alignment into reconstruction through joint registration, progressive matching, or deformable aggregation (Fu et al., 2020; Nie et al., 2020; Qu et al., 2022; Zheng et al., 2022; Zhou et al., 2019; Ying et al., 2022). MoE-PNP (Qu et al., 2024) jointly models registration and fusion, SSC-HSR (Zhang et al., 2025) refines their interaction through deformable aggregation, and UAFL (Zhang et al., 2026) restores abundances while retaining SVD-derived endmembers. These methods improve unregistered fusion, but leave spatial reconstruction largely implicit in cross-modal interaction. ROTSA instead learns explicit reconstruction operators from the reference and transfers them to HSI features.

Spatially adaptive upsampling. Adaptive upsampling uses image content to guide local reconstruction. Joint bilateral upsampling derives edge-aware weights from an HR guide (Kopf et al., 2007), while CARAFE and dynamic upsampling filters predict local kernels for feature reassembly and detail recovery (Wang et al., 2019; Jo et al., 2018). DKN learns spatially varying sampling locations and kernel weights for joint image filtering (Kim et al., 2021). Upsample Anything combines spatial and range cues through anisotropic Gaussian kernels (Seo et al., 2026). Inspired by these content-adaptive reconstruction schemes, ROTSA learns neighborhood weights and an affine detail state through reference self-reconstruction, then transports both fields via structural correspondence to guide recurrent HSI restoration under misregistration.

Spectral adaptation. Inter-band correlation motivates subspace-based fusion and adaptive dimension selection (Yokoya et al., 2011; Simoes et al., 2014; Rasti et al., 2015), with compact representations also obtained through singular-value thresholding and graph structure (Gavish & Donoho, 2014; Jiang et al., 2013). HySime selects a subspace by minimizing estimation error (Bioucas-Dias & Nascimento, 2008), while UAFL reduces reconstruction dimensionality through spectral unmixing (Zhang et al., 2026). However, dimension selection does not specify the correction strength for each component. ROTSA adjusts these strengths across a complete spectral basis.

3 METHODOLOGY

3.1 MOTIVATION AND OVERVIEW

Effective unregistered HSI super-resolution (SR) requires deciding both *what spatial information to transfer* and *how to incorporate it into the HSI reconstruction*. As shown in Fig. 1, reference correction responses differ around displaced reference boundaries, motivating us to learn transferable spatial relationships through reference self-reconstruction and align the resulting operators for application to HSI features. Meanwhile, the scene-dependent ranks in Fig. 2 indicate that the spectral contribution of reference detail should vary with the observed scene. We therefore retain a complete spectral basis and continuously adjust the transferred correction across its modes.

The proposed framework is illustrated in Fig. 3. Let $\mathbf{L} \in \mathbb{R}^{C\times h\times w}$ denote the low-resolution (LR) HSI with C spectral bands and $\mathbf{R}_S \in \mathbb{R}^{3\times H\times W}$ the unregistered high-resolution (HR) RGB reference. We obtain a base reconstruction $\mathbf{B}$ by Bicubic interpolation of $\mathbf{L}$ to the target resolution. The model predicts an HSI residual $\mathbf{R}_{\mathrm{H}}$ and a reference-guided residual $\mathbf{R}_{\mathrm{G}}$, both at this resolution. Their difference $\Delta\mathbf{R} = \mathbf{R}_{\mathrm{G}} - \mathbf{R}_{\mathrm{H}}$ represents the proposed reference correction. Adaptive spectral capacity modulates this correction to obtain $\Delta\mathbf{R}_a$, which is added to the retained HSI estimate:

$$\widehat{\mathbf{Y}} = \mathbf{B} + \mathbf{R}_{\mathrm{H}} + \Delta\mathbf{R}_a. \tag{1}$$

Matrix products act on the spectral vector of each pixel. The reconstruction thus retains the HSI estimate $\mathbf{B} + \mathbf{R}_{\mathrm{H}}$ while regulating the additional reference contribution.

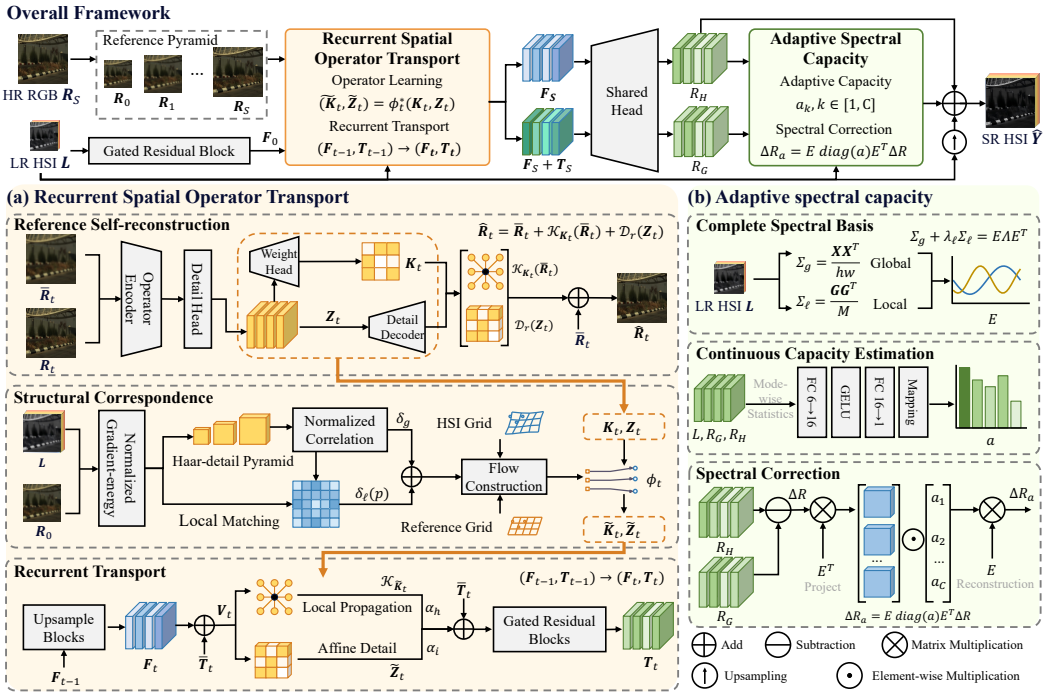

Figure 3: Overview of our proposed framework. (a) Recurrent spatial operator transport learns neighborhood weights and an affine detail state through reference self-reconstruction, aligns these fields through structural correspondence, and recurrently applies them to HSI features. (b) Adaptive spectral capacity estimates continuous mode-wise weights to regulate the transferred correction in a complete spectral basis.

3.2 Recurrent Spatial Operator Transport

Reference pyramid. We construct a coarse-to-fine reference pyramid $\{\mathbf{R}_0, \ldots, \mathbf{R}_S\}$ to learn spatial self-reconstruction. The coarsest reference matches the LR-HSI resolution, while intermediate levels are sampled directly from the HR reference at successively halved resolutions $(H/2^j, W/2^j)$:

$$\mathbf{R}_0 = \downarrow_{h,w} (\boldsymbol{k}_\theta * \mathbf{R}_S), \qquad \mathbf{R}_t = \mathcal{B}_t(\mathbf{R}_S), \quad 1 \le t < S, \tag{2}$$

where $\boldsymbol{k}_\theta$ is a learned nonnegative and normalized kernel, and $\downarrow_{h,w}$ subsamples the filtered image to (h, w). The operator $\mathcal{B}_t$ denotes Bicubic resampling to the size (H_t, W_t) of level t. These reference pairs provide self-reconstruction targets.

Reference self-reconstruction. For each transition, let $\overline{\mathbf{R}}_t = \mathcal{B}_t(\mathbf{R}_{t-1})$. The Operator Encoder and Detail Head in Fig. 3(a) extract a normalized state $\mathbf{Z}_t$ from the coarse–fine pair, and the Weight Head predicts neighborhood weights $\mathbf{K}_t$. For a multichannel signal $\mathbf{V}$, these weights define:

$$\mathcal{K}_{\mathbf{K}_t}(\mathbf{V})(\boldsymbol{p}) = \sum_{\boldsymbol{q} \in \mathcal{N}_8(\boldsymbol{p})} \kappa_{t,\boldsymbol{p},\boldsymbol{q}} \big[\mathbf{V}(\boldsymbol{q}) - \mathbf{V}(\boldsymbol{p})\big], \qquad \sum_{\boldsymbol{q}} |\kappa_{t,\boldsymbol{p},\boldsymbol{q}}| \le \gamma < 1, \tag{3}$$

where $\mathcal{N}_8(\boldsymbol{p})$ contains the eight neighbors of location $\boldsymbol{p}$, and $\kappa_{t,\boldsymbol{p},\boldsymbol{q}}$ is a signed coefficient shared across channels. Signed weights permit smoothing and sharpening, with their total magnitude bounded by γ.

Since reweighting cannot recover variation from a constant neighborhood, the Detail Decoder $\mathcal{D}_\mathrm{r}$ supplies an additive high-frequency residual:

$$\widehat{\mathbf{R}}_t = \overline{\mathbf{R}}_t + \mathcal{K}_{\mathbf{K}_t}(\overline{\mathbf{R}}_t) + \mathcal{D}_\mathrm{r}(\mathbf{Z}_t). \tag{4}$$

Supervision against $\mathbf{R}_t$ trains local propagation and missing-detail reconstruction jointly. The shared learner operates at every scale, transferring $\mathbf{K}_t$ and the latent detail state $\mathbf{Z}_t$.

Structural correspondence. We estimate correspondence between $\mathbf{L}$ and $\mathbf{R}_0$ using normalized gradient-energy representations. First, the global translation $\boldsymbol{\delta}_{\mathrm{g}}$ is estimated by maximizing overlap-normalized correlation and searching at a multi-scale pyramid between the HSI and reference Haar-detail energy maps. Then, the local residual displacement $\boldsymbol{\delta}_\ell(\boldsymbol{p})$ is estimated as a softmax-weighted average of candidate offsets, using similarities between normalized HSI descriptors and globally warped reference descriptors within a local neighborhood. The weighted offsets yield a subpixel residual displacement, giving $\boldsymbol{d}(\boldsymbol{p}) = \boldsymbol{\delta}_{\mathrm{g}} + \boldsymbol{\delta}_\ell(\boldsymbol{p})$.

Flow Construction bilinearly resizes $\boldsymbol{d}$ to each stage and scales its horizontal and vertical components by W_t/w and H_t/h, respectively. The resulting sampling map $\boldsymbol{\phi}_t$ transports the reference fields to the HSI grid:

$$(\widetilde{\mathbf{K}}_t, \widetilde{\mathbf{Z}}_t) = (\mathbf{K}_t \circ \boldsymbol{\phi}_t, \mathbf{Z}_t \circ \boldsymbol{\phi}_t), \tag{5}$$

where $\circ$ denotes bilinear sampling at the reference locations specified by $\boldsymbol{\phi}_t$.

Recurrent transport. We embed $\mathbf{L}$ into HSI features $\mathbf{F}_0$ and initialize the accumulated reference correction as $\mathbf{T}_0 = 0$. At each stage, the Upsample Blocks refine the HSI features while resizing the preceding correction:

$$\mathbf{F}_t = \mathcal{H}(\mathcal{B}_t(\mathbf{F}_{t-1})), \qquad \overline{\mathbf{T}}_t = \mathcal{U}_t(\mathbf{T}_{t-1}), \qquad \mathbf{V}_t = \mathbf{F}_t + \overline{\mathbf{T}}_t, \tag{6}$$

where $\mathcal{H}$ is the shared HSI refinement module and $\mathcal{U}_t$ denotes bilinear resampling to (H_t, W_t). The transported fields refine $\mathbf{V}_t$ through local propagation and gated detail injection:

$$\begin{aligned} \mathcal{M}_t(\mathbf{V}_t) &= \mathcal{P}_{\mathrm{out}}\left[\mathcal{K}_{\widetilde{\mathbf{K}}_t}(\mathcal{P}_{\mathrm{in}}(\mathbf{V}_t))\right], \\ \mathcal{J}_t(\mathbf{V}_t) &= \mathcal{P}_Z(\widetilde{\mathbf{Z}}_t) \odot \tanh \mathcal{G}(\mathbf{V}_t, \boldsymbol{\Gamma}_t), \end{aligned} \tag{7}$$

where $\mathcal{P}_{\mathrm{in}}$ and $\mathcal{P}_{\mathrm{out}}$ project features into and out of the propagation space, and $\mathcal{P}_Z$ maps the detail state to the HSI feature dimension. The gate $\mathcal{G}$ uses $\mathbf{V}_t$ and scale encoding $\boldsymbol{\Gamma}_t$ to control the projected detail. $\odot$ denotes element-wise multiplication. The Gated Residual Blocks $\mathcal{Q}$ update the accumulated correction:

$$\mathbf{T}_t = \mathcal{Q}\big(\overline{\mathbf{T}}_t + \alpha_h \mathcal{M}_t(\mathbf{V}_t) + \alpha_i \mathcal{J}_t(\mathbf{V}_t)\big), \tag{8}$$

where α_h, α_i are learned bounded gains and $\mathcal{Q}$ is shared across stages. The independent recurrence of $\mathbf{F}_t$ retains an estimate from the HSI alone. At the finest level, the Shared Head g produces:

$$\mathbf{R}_{\mathrm{H}} = g(\mathbf{F}_S), \qquad \mathbf{R}_{\mathrm{G}} = g(\mathbf{F}_S + \mathbf{T}_S). \tag{9}$$

Both residuals contain C bands, and their difference is the proposed reference correction (more implementation details in Appendix A).

3.3 Adaptive spectral capacity

Complete spectral basis. We regulate the proposed correction using global and local spectral statistics of $\mathbf{L}$. Let $\boldsymbol{X} \in \mathbb{R}^{C \times hw}$ contain spatially centered LR spectra, and let $\boldsymbol{G} \in \mathbb{R}^{C \times M}$ collect differences between adjacent spectra, where M is the number of horizontal and vertical pixel pairs. The complete spectral basis can be obtained by:

$$\boldsymbol{\Sigma}_{\mathrm{g}} = \frac{\boldsymbol{X}\boldsymbol{X}^\top}{hw}, \qquad \boldsymbol{\Sigma}_\ell = \frac{\boldsymbol{G}\boldsymbol{G}^\top}{M}, \qquad \boldsymbol{\Sigma}_{\mathrm{g}} + \lambda_\ell \boldsymbol{\Sigma}_\ell = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^\top, \tag{10}$$

where λ_ℓ balances local variation, $\boldsymbol{\Lambda}$ contains the eigenvalues, and $\boldsymbol{E} \in \mathbb{R}^{C \times C}$ contains the eigenvectors in decreasing eigenvalue order.

Continuous capacity estimation. We extract mode-wise statistics to describe the observed scene and the proposed reconstruction. Projecting the scene covariances onto each basis vector and normalizing across modes gives p_k^{g} and p_k^{ℓ}, which measure global spectral variation and local spectral differences, respectively. We also project the guided residual and reference correction, $\mathbf{A}_{\mathrm{G}} = \boldsymbol{E}^\top \mathbf{R}_{\mathrm{G}}$ and $\mathbf{A}_\Delta = \boldsymbol{E}^\top \Delta\mathbf{R}$. Their spatial mean-square energies, normalized across modes, yield q_k^{G} and q_k^{Δ}, describing the reconstruction residual and the additional change proposed by reference guidance.

For each mode, we assemble a descriptor $\boldsymbol{z}_k \in \mathbb{R}^6$ containing these four log-energy statistics, a normalized version of the remaining-scene-energy prior π_k, and the normalized mode index $(k-1)/(C-1)$. A shared MLP h_ψ predicts an adjustment to the prior:

$$a_k = \sigma(\pi_k + \xi \tanh h_\psi(\boldsymbol{z}_k)), \tag{11}$$

Table 1: Quantitative comparison on ICVL at three scales. PSNR is measured in dB. Best and second-best results are bold and underlined. FLOPs are measured for 256×256 outputs at $\times 4$.

Method	Venue	$\times 4$			$\times 8$			$\times 16$			Params	FLOPs
		PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	(M)	(G)
Bicubic	–	35.33	0.960	0.028	30.20	0.914	0.043	26.69	0.871	0.064	–	–
SSPSR	TCI'20	40.19	0.982	0.033	33.65	0.944	0.045	28.96	0.902	<u>0.058</u>	20.82	180.40
MCNet	RS'20	39.90	0.980	0.033	33.36	0.941	0.048	28.75	0.899	0.071	2.20	280.30
BiQRNN	JSTARS'21	37.85	0.974	0.039	32.77	0.937	0.054	28.21	0.892	0.077	1.30	468.70
ESSA	ICCV'23	38.41	0.977	0.057	32.75	0.936	0.076	28.51	0.893	0.087	11.10	78.14
NonReg	TGRS'22	33.47	0.949	0.131	30.66	0.926	0.202	27.06	0.879	0.294	2.00	161.61
MoE-PNP	TNNLS'24	33.97	0.957	0.074	31.92	0.931	0.086	29.61	0.908	0.080	44.67	3040.81
HSIFN	TNNLS'24	41.14	<u>0.987</u>	0.041	35.18	<u>0.963</u>	0.063	29.81	0.910	0.082	21.01	594.10
LRTN	IJCV'25	37.58	0.975	0.061	32.74	0.936	0.070	29.25	0.902	0.082	3.79	33.15
SRLF	CVPR'25	38.75	0.977	0.041	32.92	0.932	0.069	29.06	0.889	0.101	1.33	20.31
SSC-HSR	IJCV'25	41.38	<u>0.987</u>	0.031	34.74	0.954	0.042	30.77	0.917	0.071	11.01	165.68
SSDAN	TGRS'26	41.30	0.984	0.029	34.67	0.950	0.042	30.45	0.915	0.062	6.77	337.27
UAFL	CVPR'26	<u>41.84</u>	0.985	<u>0.025</u>	<u>35.78</u>	<u>0.963</u>	<u>0.039</u>	<u>31.35</u>	<u>0.922</u>	0.059	5.94	96.17
ROTSA	Ours	**44.70**	**0.993**	**0.023**	**37.92**	**0.977**	**0.037**	**32.05**	**0.938**	**0.052**	1.39	101.94

where σ is the sigmoid function, and ξ bounds the learned adjustment. Each a_k is shared across spatial positions. Descriptor detail appear in Appendix C.

Spectral correction. We project the reference correction $\Delta\mathbf{R}$ onto the complete basis $\boldsymbol{E}$, scale each coefficient map by its estimated weight, and reconstruct the modulated correction:

$$\Delta\mathbf{R}_a = \boldsymbol{E}\,\mathrm{diag}(\boldsymbol{a})\boldsymbol{E}^\top\Delta\mathbf{R}, \tag{12}$$

where $\boldsymbol{a} = [a_1, \ldots, a_C]^\top$ contains the correction weights. Adding $\Delta\mathbf{R}_a$ to the HSI estimate $\mathbf{B}+\mathbf{R}_\mathrm{H}$ yields the final reconstruction in Eq. (1). Unit weights recover the guided prediction, while smaller weights reduce its contribution in the corresponding modes. This adapts reference guidance to the scene without truncating the underlying HSI estimate.

3.4 Learning objective

Given the HR-HSI target $\mathbf{Y}$, we supervise the final reconstruction, the retained HSI estimate, and the reference scale transitions:

$$\mathcal{L} = \|\widehat{\mathbf{Y}} - \mathbf{Y}\|_1 + \lambda_{\mathrm{sam}}\mathcal{L}_{\mathrm{SAM}}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_h\|\mathbf{B} + \mathbf{R}_\mathrm{H} - \mathbf{Y}\|_1 + \frac{\lambda_r}{S}\sum_{t=1}^{S}\|\widehat{\mathbf{R}}_t - \mathbf{R}_t\|_1, \tag{13}$$

where $\|\cdot\|_1$ denotes mean absolute error and $\mathcal{L}_{\mathrm{SAM}}$ is the mean spectral-angle loss. The coefficients λ_{sam}, λ_h, and λ_r balance spectral fidelity, auxiliary HSI reconstruction, and reference self-reconstruction, respectively. Reference supervision is averaged over the S transitions.

4 Experiments

4.1 Experimental setting

Datasets. We evaluate on ICVL (Arad & Ben-Shahar, 2016) and REAL (Lai et al., 2024) under the unregistered reference-guided protocol of UAFL (Zhang et al., 2026). ICVL uses synthesized RGB references with simulated viewpoint changes, whereas REAL contains captured, unregistered pairs. We reserve 20 ICVL scenes and 10 REAL scenes for testing and use the remainder for training. Further details appear in the Appendix E.1.

Implementation details. Following Zhang et al. (2025), we generate LR-HSI inputs by applying an 8×8 Gaussian kernel with $\sigma = 3$ and subsequent downsampling. ROTSA uses 64-channel HSI features and a 32-channel reference detail state, with reconstruction and transport parameters shared across recurrent stages. For each scale ($\times 4$, $\times 8$, and $\times 16$), we train a separate model for 100 epochs on ICVL and 300 epochs on REAL. Training minimizes Eq. (13) using AdamW (Loshchilov & Hutter, 2017), with a learning rate of 10^{-5}, weight decay of 5×10^{-5}, and batch size of 1. All experiments use PyTorch (Paszke et al., 2019) on a single NVIDIA RTX 4090 GPU.

Table 2: Quantitative comparison on REAL at three scales. PSNR is measured in dB. Best and second-best results are bold and underlined.

Method	Venue	×4			×8			×16		
		PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Bicubic	–	34.01	0.941	0.042	28.43	0.870	0.058	24.84	0.815	0.086
SSPSR	TCI'20	39.04	0.976	0.040	31.13	0.906	0.069	27.20	0.851	0.077
MCNet	RS'20	39.07	0.974	0.038	31.31	0.904	0.063	27.04	0.845	0.088
BiQRNN	JSTARS'21	37.80	0.969	0.044	31.66	0.908	0.060	26.53	0.832	0.111
ESSA	ICCV'23	39.94	0.982	0.040	31.85	0.913	0.064	26.70	0.918	0.111
NonReg	TGRS'22	31.60	0.951	0.081	30.00	0.923	0.167	29.09	0.904	0.584
MoE-PNP	TNNLS'24	32.57	0.951	0.101	31.12	0.946	0.119	29.91	0.934	0.121
HSIFN	TNNLS'24	40.15	0.982	0.037	34.39	0.953	0.052	30.07	0.939	0.100
LRTN	IJCV'25	38.62	0.978	0.054	34.79	0.957	0.066	31.41	0.937	0.081
SRLF	CVPR'25	39.06	0.978	0.043	34.63	0.953	0.064	31.12	0.929	0.091
SSC-HSR	IJCV'25	41.16	0.986	0.036	36.19	0.969	0.052	31.91	0.940	0.084
SSDAN	TGRS'26	40.93	0.984	0.036	35.84	0.961	0.049	31.81	0.941	0.070
UAFL	CVPR'26	<u>42.05</u>	<u>0.988</u>	<u>0.033</u>	<u>37.23</u>	<u>0.972</u>	<u>0.046</u>	<u>32.28</u>	<u>0.942</u>	<u>0.065</u>
ROTSA	Ours	**42.63**	**0.989**	**0.030**	**37.75**	**0.973**	**0.042**	**33.68**	**0.949**	**0.057**

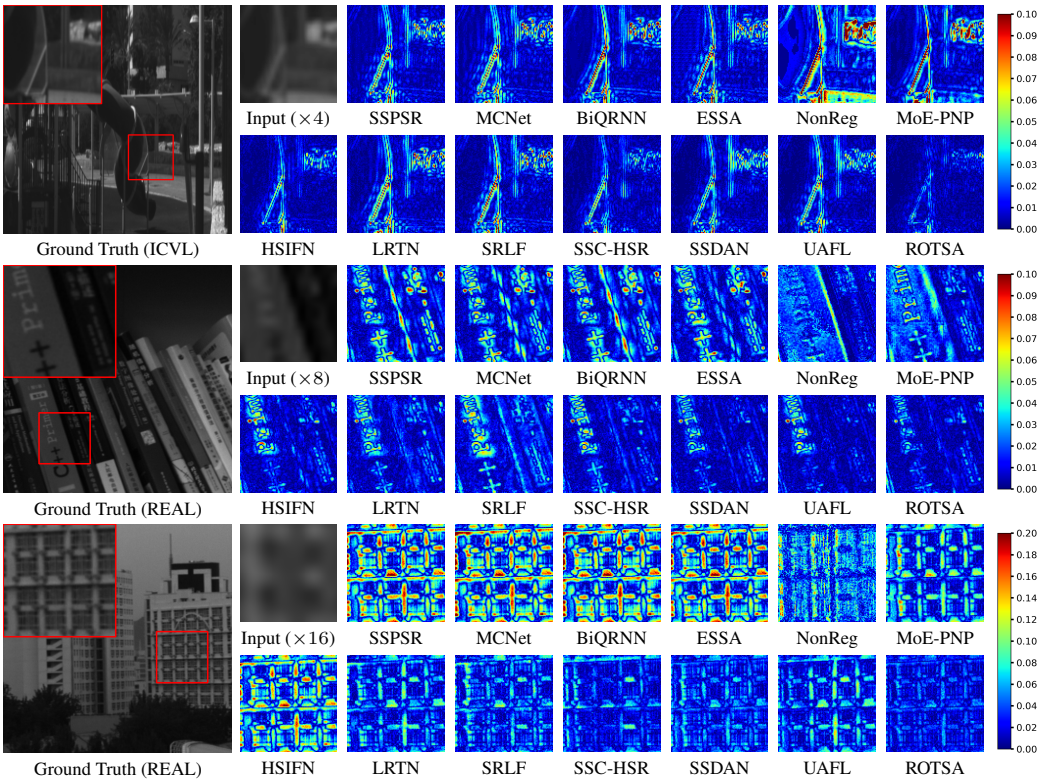

Figure 4: Visual comparison on ICVL ×4 (top), REAL ×8 (middle), and REAL ×16 (bottom). Error maps share a color scale within each group.

Compared methods and metrics. We compare ROTSA with SOTA methods, including single HSI SR methods SSPSR (Jiang et al., 2020), MCNet (Li et al., 2020), BiQRNN (Fu et al., 2021), and ESSA (Zhang et al., 2023); and fusion-based methods NonReg (Zheng et al., 2022), MoE-PNP (Qu et al., 2024), HSIFN (Lai et al., 2024), LRTN (Liu et al., 2025a), SRLF (Liu et al., 2025b), SSC-HSR (Zhang et al., 2025), SSDAN (Song et al., 2026), and UAFL (Zhang et al., 2026). Reconstruction quality is evaluated using peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and spectral angle mapper (SAM, in radians). Higher PSNR and SSIM and lower SAM indicate better performance.

4.2 Comparison with existing methods

Quantitative comparison. As shown in Tables 1 and 2, ROTSA achieves the best PSNR, SSIM, and SAM at all three scales on both datasets. On ICVL, it exceeds UAFL by 2.86, 2.14, and 0.70 dB at $\times 4$, $\times 8$, and $\times 16$, respectively. These gains are accompanied by improved SSIM and SAM, indicating that spatial detail recovery does not come at the expense of spectral fidelity. The advantage also holds on REAL, where the HSI and reference are captured by separate sensors. At $\times 16$, ROTSA improves PSNR over UAFL by 1.40 dB and reduces SAM from 0.065 to 0.057. Its consistent performance across simulated and real misregistration supports the effectiveness of spatial operator transport with adaptive spectral capacity correction. ROTSA achieves these improvements with 1.39 M parameters, 76.6% fewer than UAFL, at a comparable computational cost. Relative to SSDAN, it reduces parameters and FLOPs by 79.5% and 69.8%, respectively, demonstrating a favorable balance between reconstruction quality and model complexity.

Visual comparison. Fig. 4 compares reconstruction errors in a spectral band for each example. On ICVL, ROTSA reduces the concentrated errors along the narrow curved boundary that remain visible in competing methods. On REAL, it produces lower errors around oblique structures at $\times 8$ and repeated window boundaries at $\times 16$. These examples illustrate more accurate recovery of fine structures across different resolution gaps, complementing the quantitative improvements in spatial and spectral fidelity.

4.3 Ablation studies

Framework design. As shown in Fig. 5, we examine the roles of reference guidance, recurrent spatial operator transport, and adaptive spectral capacity through three ablation variants: removing the reference (w/o Ref), replacing recurrent spatial operator transport with direct feature fusion (w/o RSOT), and disabling adaptive spectral capacity (w/o ASC). Without the reference, PSNR drops from 37.75 to 32.47 dB, confirming the importance of reference guidance. Direct fusion reduces PSNR by 1.13 dB and increases SAM from 0.042 to 0.046, supporting the use of learned spatial operators to transfer reference detail. The performance of w/o ASC (direct guided residual without spectral routing) drops, indicating that spectral modulation further improves the use of transferred information. The full model achieves the best results across all three metrics, demonstrating the joint benefit of spatial operator transport and adaptive spectral capacity.

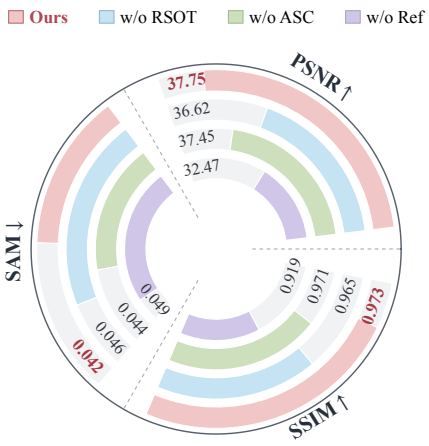

Figure 5: Ablation of main component.

Spatial operator design. Table 3 examines the operator representation and its supervision. The variants w/o neighborhood weights and w/o affine state disable local propagation and additive detail, respectively. Removing the affine state causes the largest PSNR decrease, from 37.75 to 36.47 dB, highlighting the importance of supplying detail beyond local reweighting. Removing neighborhood weights gives 37.61 dB, supporting the additional contribution of local propagation. The variant w/o self-rec. loss removes only the reference reconstruction loss while retaining the reference branch. Its lower PSNR of 37.66 dB and higher SAM of 0.043 indicate that reference supervision helps constrain the learned spatial transformation. Without explicit alignment, ROTSA still performs close to the full model, indicating that its reconstruction capability does not rely solely on displacement compensation (additional comparisons in Appendix B.1).

Table 3: Ablation of recurrent spatial operator transport on REAL $\times 8$.

Variant	PSNR↑	SSIM↑	SAM↓
w/o neighborhood weights	37.61	<u>0.972</u>	**0.042**
w/o affine state	36.47	0.965	0.045
w/o alignment	37.63	0.971	<u>0.043</u>
w/o self-rec. loss	<u>37.66</u>	<u>0.972</u>	<u>0.043</u>
ROTSA	**37.75**	**0.973**	**0.042**

Structural correspondence. We compare separately trained models on ICVL $\times 8$ under clean (aligned), slight (default), and large misregistration in Table 4. The large setting increases the rigid, parallax, and nonrigid transformations. The variant w/o alignment transfers operators without dis-

Table 4: Alignment comparison on ICVL $\times 8$ under different misregistration levels. Models are trained separately for each setting.

Method	Clean			Slight			Large		
	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
SSC-HSR	41.51	0.990	0.042	34.74	0.954	0.042	<u>34.99</u>	**0.960**	0.043
UAFL	44.30	<u>0.993</u>	<u>0.036</u>	35.78	0.963	0.039	34.32	0.947	<u>0.037</u>
ROTSA w/o alignment	**48.89**	**0.997**	**0.024**	<u>37.15</u>	<u>0.973</u>	<u>0.038</u>	34.78	<u>0.951</u>	**0.036**
ROTSA	<u>48.78</u>	**0.997**	**0.024**	**37.92**	**0.977**	**0.037**	**35.71**	**0.960**	**0.036**

Table 5: Ablation of adaptive spectral capacity on REAL $\times 8$.

Variant	PSNR↑	SSIM↑	SAM↓
Fixed rank = 3	37.56	0.972	0.043
Full capacity	37.45	0.971	0.044
ROTSA	**37.75**	**0.973**	**0.042**

Table 6: Ablation of reference pyramid construction on REAL $\times 8$.

Coarsest	Intermediate	PSNR↑	SSIM↑	SAM↓
Learned	Recursive	37.28	0.970	0.043
Learned	Area	37.66	0.972	**0.042**
Learned	Bicubic (ROTSA)	**37.75**	**0.973**	**0.042**

placement compensation. Structural correspondence improves PSNR by 0.77 and 0.93 dB under slight and large misregistration, respectively. Under large misregistration, ROTSA exceeds UAFL and SSC-HSR by 1.39 and 0.72 dB. On clean inputs, disabling alignment gives a marginally higher PSNR with unchanged SSIM and SAM. These results indicate that correspondence is particularly useful when spatial mismatch must be resolved, allowing reference operators to guide reconstruction at the appropriate HSI locations (more evaluation in Appendix B.2).

Adaptive spectral capacity. We examine how spectral capacity affects reference correction through the three variants in Table 5. Fixed rank = 3 applies reference corrections only to the first three spectral modes and retains the HSI prediction elsewhere. Full capacity sets all mode weights to one, accepting the complete correction, whereas ROTSA adaptively adjusts the weights across all modes. ROTSA improves PSNR over these alternatives by 0.19 and 0.3 dB, respectively, with better SSIM and SAM. These results support adapting correction strength rather than using a fixed interaction subspace or uniformly accepting reference corrections (more ablation in Appendix D).

Reference pyramid construction. Keeping the coarsest reference learned, we vary the construction of intermediate pyramid levels in Table 6. Recursive constructs intermediate levels through successive convolutions, whereas Area and Bicubic directly resize the HR reference to each intermediate resolution. Bicubic achieves 37.75 dB, compared with 37.66 dB for Area and 37.28 dB for Recursive. The similar performance of Bicubic and Area suggests that operator learning accommodates different intermediate scale constructions. The lower score of Recursive indicates that learned intermediate levels do not necessarily provide better reconstruction targets. Further degradation and loss ablations appear in Appendix F.

5 Conclusion

We propose ROTSA for unregistered HSI super-resolution, combining recurrent spatial operator transport with adaptive spectral capacity. Through reference self-reconstruction, ROTSA learns neighborhood weights and an affine detail state that jointly model local propagation and detail missing from the coarse scale. Structural correspondence aligns these fields with the HSI grid for recurrent restoration, separating operator learning from application across modalities. This design gives reference guidance an explicit spatial reconstruction role. Adaptive spectral capacity regulates the transferred correction across a complete spectral basis according to scene content and prediction statistics, while preserving the reconstruction supported by the HSI input. Experiments on ICVL and REAL demonstrate consistent improvements in spatial reconstruction and spectral fidelity across multiple upsampling scales. Ablation studies support the contributions of the spatial and spectral components, while recurrent parameter sharing keeps the model compact. Together, these findings highlight the value of learning transferable spatial operators and adapting their spectral contribution to HSI super-resolution.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting, language revision, and formatting. The authors reviewed and revised the AI-assisted material and take full responsibility for the final manuscript, including its scientific claims, experimental results, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

We document the method formulation and training objective in Section 3, and report dataset protocols, evaluation metrics, and training settings in Section 4. Additional architectural and implementation details are provided in the appendix. The appendix further describes ablation configurations and motivation analysis to clarify how the reported comparisons were obtained. Together, these descriptions support implementation of the method and reproduction of the experimental evaluation.

REFERENCES

Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Sparse spatio-spectral representation for hyperspectral image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 63–78, 2014.

Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Bayesian sparse representation for hyperspectral image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3631–3640, 2015.

Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *Proceedings of the European Conference on Computer Vision*, pp. 19–34, 2016.

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

Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11661–11670, 2019.

Ying Fu, Yongrong Zheng, Lin Zhang, Yinqiang Zheng, and Hua Huang. Simultaneous hyperspectral image super-resolution and geometric alignment with a hybrid camera system. *Neurocomputing*, 384:282–294, 2020.

Ying Fu, Zhiyuan Liang, and Shaodi You. Bidirectional 3d quasi-recurrent neural network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2674–2688, 2021.

Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

Jungho Im and John R Jensen. Hyperspectral remote sensing of vegetation. *Geography Compass*, 2 (6):1943–1961, 2008.

Bo Jiang, Chris Ding, Bio Luo, and Jin Tang. Graph-laplacian pca: Closed-form solution and robustness. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3492–3498, 2013.

Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6:1082–1096, 2020.

Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, 2018.

Rei Kawakami, Yasuyuki Matsushita, John Wright, Moshe Ben-Ezra, Yu-Wing Tai, and Katsushi Ikeuchi. High-resolution hyperspectral imaging via matrix factorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2329–2336, 2011.

Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021.

Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics*, 26(3):96–es, 2007.

Zeqiang Lai, Ying Fu, and Jun Zhang. Hyperspectral image super resolution with real unaligned rgb guidance. *IEEE Transactions on Neural Networks and Learning Systems*, 1(1):1–13, 2024.

Charis Lanaras, Emmanuel Baltsavias, and Konrad Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3586–3594, 2015.

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

Yingping Liang, Jiaming Liu, Debing Zhang, and Ying Fu. Mpi-flow: Learning realistic optical flow with multiplane images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13857–13868, October 2023a.

Zhiyuan Liang, Shuai Wang, Tao Zhang, and Ying Fu. Blind super-resolution of single remotely sensed hyperspectral image. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023b.

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.

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. IEEE, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiangtao Nie, Lei Zhang, Wei Wei, Chen Ding, and Yanning Zhang. Unsupervised deep hyperspectral super-resolution with unregistered images. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2020.

Zhihong Pan, Glenn Healey, Manish Prasad, and Bruce Tromberg. Face recognition in hyperspectral images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1552–1560, 2003.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Proceedings of the Advances in Neural Information Processing Systems*, 32, 2019.

Jiahui Qu, Jizhou Cui, Wenqian Dong, Qian Du, Xiaoyang Wu, Song Xiao, and Yunsong Li. A principle design of registration-fusion consistency: Toward interpretable deep unregistered hyperspectral image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5): 9648–9662, 2024.

Ying Qu, Hairong Qi, Chiman Kwan, Naoto Yokoya, and Jocelyn Chanussot. Unsupervised and unregistered hyperspectral image super-resolution with mutual dirichlet-net. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

Behnood Rasti, Magnus O Ulfarsson, and Johannes R Sveinsson. Hyperspectral subspace identification using sure. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2481–2485, 2015.

Minseok Seo, Mark Hamilton, and Changick Kim. Upsample anything: A simple and hard to beat baseline for feature upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29707–29716, 2026.

Miguel Simoes, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3373–3388, 2014.

Qiya Song, Shuo Guo, Tao Yang, Bin Sun, Renwei Dian, and Shutao Li. S2-differential feature awareness network for hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3007–3016, 2019.

Pei Yang, Yong Ma, Xiaoguang Mei, Qihai Chen, Minghui Wu, and Jiayi Ma. Deep blind super-resolution for hyperspectral images. *Pattern Recognition*, 157:110916, 2025.

Zhixiong Yang, Jingyuan Xia, Shengxi Li, Lingyu Zheng, Shuanghui Zhang, Li Liu, Yaowen Fu, and Yongxiang Liu. Band-kernel stochastic learning for unsupervised blind hyperspectral image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jiacheng Ying, Hui-Liang Shen, and Si-Yuan Cao. Unaligned hyperspectral image fusion via registration and interpolation modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022.

Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537, 2011.

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.

Tao Zhang, Ying Fu, Liwei Huang, Siyuan Li, Shaodi You, and Chenggang Yan. Rgb-guided hyperspectral image super-resolution with deep progressive learning. *CAAI Transactions on Intelligence Technology*, 9(3):679–694, 2024.

Yingkai Zhang, Zeqiang Lai, Tao Zhang, Ying Fu, and Chenghu Zhou. Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance. *International Journal of Computer Vision*, 133(9):6590–6610, 2025.

Yingkai Zhang, Tao Zhang, Jing Nie, and Ying Fu. Enhancing unregistered hyperspectral image super-resolution via unmixing-based abundance fusion learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41573–41583, 2026.

Ke Zheng, Lianru Gao, Danfeng Hong, Bing Zhang, and Jocelyn Chanussot. Nonregsrnet: A non-rigid registration hyperspectral super-resolution network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

Yuan Zhou, Anand Rangarajan, and Paul D Gader. An integrated approach to registration and fusion of hyperspectral and multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3020–3033, 2019.

APPENDIX

We provide implementation details, mathematical analysis, and additional experiments that complement the main paper. It is organized as follows:

- Section A describes the spatial operator, correspondence estimation, and local reconstruction properties.
- Section B evaluates operator replacements, alignment across train–test conditions, and output-domain reference responses.
- Section C details the spectral basis, continuous capacity weights, and reconstruction properties.
- Section D studies controller inputs, fixed interaction ranks, and scene-wise spectral variation.
- Section E specifies the datasets, training protocol, and evaluation metrics.
- Section F examines kernel support, training objectives, and reference pyramid construction.
- Section G analyzes recurrent computation and model complexity.
- Section H presents additional qualitative comparisons on REAL and ICVL.

A SPATIAL OPERATOR TRANSPORT: IMPLEMENTATION

This section specifies the reference learner, structural correspondence, recurrent application, and local properties of the reconstruction operator.

A.1 REFERENCE CONSTRUCTION AND SHARED ARCHITECTURE

The HSI feature width is $d = 64$, and the reference state has 32 channels. The input embedding uses a 3×3 convolution and two gated residual blocks. Each shared $2\times$ HSI restoration stage contains Bicubic upsampling, a convolution, and five gated residual blocks. The reference learner uses three such blocks, and operator refinement uses two. The shared output head contains one gated residual block and a 3×3 convolution to C bands. The number of recurrent stages is $S = \log_2 s$; parameters are shared across stages, while the benchmark trains separate models for each overall scale.

Each gated residual block uses a 3×3 convolution to form content and gate branches, multiplies the GELU-activated content by a sigmoid gate, and projects the result with a second 3×3 convolution. A learned $\tanh$-bounded gain scales this update before it is added to the block input.

The reference kernel uses nonnegative, normalized coefficients with point symmetry. For free coefficients $\mathbf{\Omega}$:

$$\boldsymbol{k}_\theta = \mathrm{softmax}\left(\frac{\mathbf{\Omega} + \mathrm{rot}_{180}(\mathbf{\Omega})}{2}\right), \tag{14}$$

where normalization covers all kernel entries. A shared 8×8 kernel is applied independently to each reference channel, followed by stride-s sampling. Intermediate reference levels are direct Bicubic resamplings of the observed HR reference. The kernel is optimized without weight decay. The kernel learns the coarse reference used to supervise cross-scale reconstruction.

A.2 LEARNING AND TRANSPORTING THE SPATIAL OPERATOR

The reference learner receives the finer image, its interpolated coarse counterpart, their difference, a high-pass reference image, and scale coordinates. Writing $[\cdot]$ for channel concatenation, its state and raw weight prediction are:

$$\mathbf{Z}_t = \mathrm{GN}\big(\mathcal{E}_{\mathrm{r}}([\mathbf{R}_t, \overline{\mathbf{R}}_t, \mathbf{R}_t - \overline{\mathbf{R}}_t, \mathrm{HP}_5(\mathbf{R}_t), \mathbf{\Gamma}(2)])\big), \qquad \mathbf{u}_t = \mathcal{W}(\mathbf{Z}_t). \tag{15}$$

Here $\mathcal{E}_{\mathrm{r}}$ is the shared convolutional encoder, GN uses one group, and HP_m subtracts an $m \times m$ local average. The weight head $\mathcal{W}$ is a 1×1 convolution. Both weight prediction and detail decoding use the normalized state. The eight weights correspond to the immediate horizontal, vertical, and diagonal neighbors. Values outside the image use replicated boundary pixels. The coefficient constraint

in Eq. (3) is implemented as:

$$\mathbf{K}_t = \frac{\gamma \tanh \mathbf{u}_t}{\max\Big(1, \sum_{j=1}^{8} |\tanh[\mathbf{u}_t]_j|\Big)}, \qquad \gamma = 0.85, \tag{16}$$

with the denominator evaluated at each pixel. The detail decoder uses two 3×3 convolutions with GELU and a final HP_3 filter. Reference self-reconstruction is supervised by the coarse–fine reference pair at each level. The latent $\mathbf{Z}_t$ and the eight-channel $\mathbf{K}_t$ are transported; the decoded RGB residual is used only for reference reconstruction. All stages are trained jointly with hyperspectral reconstruction.

Structural correspondence. Both inputs are converted to spatially normalized gradient-energy maps for matching. The Haar-detail representation uses fixed horizontal and vertical adjacent differences, pooled across channels into an energy map and normalized spatially. Repeated 2×2 average pooling forms a coarse-to-fine pyramid. Overlap-normalized correlation selects a global translation at the coarsest level, followed by local displacement searches at finer levels. A shared structural encoder followed by separate query/key projections forms unit-normalized descriptors $\boldsymbol{q}(\boldsymbol{p})$ and $\boldsymbol{k}(\boldsymbol{p})$. After globally warping the reference descriptors, let $\mathcal{N} = \{-2, \dots, 2\}^2$. Local matching computes:

$$w_{\boldsymbol{p},\boldsymbol{o}} = \frac{\exp(8\boldsymbol{q}(\boldsymbol{p})^\top \boldsymbol{k}(\boldsymbol{p}+\boldsymbol{o}))}{\sum_{\boldsymbol{o}' \in \mathcal{N}} \exp(8\boldsymbol{q}(\boldsymbol{p})^\top \boldsymbol{k}(\boldsymbol{p}+\boldsymbol{o}'))}, \qquad \boldsymbol{\delta}_\ell(\boldsymbol{p}) = \sum_{\boldsymbol{o} \in \mathcal{N}} w_{\boldsymbol{p},\boldsymbol{o}} \boldsymbol{o}. \tag{17}$$

The expected offset provides a subpixel residual displacement within the local search window. The main configuration estimates correspondence once on the LR grid and rescales its displacements for finer stages. At stage t, the sampling map is:

$$\boldsymbol{\phi}_t(\boldsymbol{p}) = \boldsymbol{p} + \boldsymbol{D}_t[\mathcal{U}_t(\boldsymbol{d})](\boldsymbol{p}), \qquad \boldsymbol{D}_t = \mathrm{diag}(W_t/w, H_t/h), \tag{18}$$

where $\boldsymbol{d} = \boldsymbol{\delta}_{\mathrm{g}} + \boldsymbol{\delta}_\ell$ and $\mathcal{U}_t$ bilinearly resizes the LR displacement. The operator fields are sampled bilinearly at $\boldsymbol{\phi}_t(\boldsymbol{p})$. Sampling uses pixel-center coordinates and border replication. Displacements are measured in LR pixels before being multiplied by 2^t at stage t. The transported coefficients act on the eight neighbors of the current HSI grid.

The spatial gate uses the current HSI features and two scale coordinates, $\log_2(s_t)/4$ and $1/s_t$, where $s_t = 2^t$ is the current magnification relative to the LR grid. The matching representation is spatially normalized before correspondence estimation.

Spatial application and notation. The neighborhood operator projects HSI features from 64 to 32 channels and then projects the result back to 64 channels, using 1×1 convolutions. The state projection $\mathcal{P}_Z$ uses a 1×1 convolution, GELU, and a 3×3 convolution to produce 64 channels. The gate $\mathcal{G}$ concatenates $\mathbf{V}_t$ with $\mathbf{\Gamma}_t$, then applies two 3×3 convolutions separated by GELU. The gains are $\alpha_h = \tanh \beta_h$ and $\alpha_i = \tanh \beta_i$, where the learned scalars β_h, β_i are initialized to 0.15. HSI features use Bicubic upsampling; the accumulated update uses bilinear upsampling. The neighborhood operation and additive state form the affine reconstruction operator; feature-dependent gating and residual refinement adapt its application to the HSI features.

The shared output head produces $\mathbf{R}_{\mathrm{H}} = g(\mathbf{F}_S)$ and $\mathbf{R}_{\mathrm{G}} = g(\mathbf{F}_S + \mathbf{T}_S)$. The final reconstruction uses $\mathbf{B} + \mathbf{R}_{\mathrm{H}}$ and adds a spectrally weighted correction. Projection and synthesis mix spectral coordinates, whereas capacity weights scale individual modes uniformly over spatial positions.

A.3 Properties of the reconstruction operator

The nonnegative reference construction kernel and the signed reconstruction weights serve different purposes. The former creates coarse observations; the latter acts on differences between neighboring values. The properties below hold with the neighborhood weights fixed.

Proposition 1 (Constant preservation and channel compatibility)**.** *For a spatially constant vector $\boldsymbol{b}$ and a pointwise linear channel map $\boldsymbol{P}$:*

$$\mathcal{K}_{\mathbf{K}}(\mathbf{V} + \boldsymbol{b}) = \mathcal{K}_{\mathbf{K}}(\mathbf{V}), \qquad \mathcal{K}_{\mathbf{K}}(\boldsymbol{P}\mathbf{V}) = \boldsymbol{P}\mathcal{K}_{\mathbf{K}}(\mathbf{V}). \tag{19}$$

Proof. The constant cancels in each neighbor difference. Channel-shared scalar coefficients commute with $\boldsymbol{P}$, so applying $\boldsymbol{P}$ before or after the weighted sum gives the same result. □

The reconstruction $\mathbf{V} + \mathcal{K}_{\mathbf{K}}(\mathbf{V})$ therefore preserves spatially constant inputs, and the same scalar neighborhood coefficients can operate across different channel spaces. When the coarse neighborhood is constant, the propagation term vanishes and the additive state supplies missing variation. This motivates learning local propagation and detail reconstruction together.

Proposition 2 (Local update bound)**.** *Let* $\mathrm{osc}_{\boldsymbol{p}}(\mathbf{V}) = \max_{\boldsymbol{q}\in\mathcal{N}_8(\boldsymbol{p})} \|\mathbf{V}(\boldsymbol{q}) - \mathbf{V}(\boldsymbol{p})\|_2$. *Under* $\sum_{\boldsymbol{q}} |\kappa_{\boldsymbol{p},\boldsymbol{q}}| \le \gamma$*:*

$$\|\mathcal{K}_{\mathbf{K}}(\mathbf{V})(\boldsymbol{p})\|_2 \le \gamma\, \mathrm{osc}_{\boldsymbol{p}}(\mathbf{V}). \tag{20}$$

Proof. Apply the triangle inequality and bound each neighbor difference by $\mathrm{osc}_{\boldsymbol{p}}(\mathbf{V})$. □

Bilinear field sampling preserves the coefficient constraint because it forms convex combinations:

$$\left\| \sum_{\boldsymbol{r}} b_{\boldsymbol{p},\boldsymbol{r}} \boldsymbol{\kappa}(\boldsymbol{r}) \right\|_1 \le \sum_{\boldsymbol{r}} b_{\boldsymbol{p},\boldsymbol{r}} \|\boldsymbol{\kappa}(\boldsymbol{r})\|_1 \le \gamma, \quad b_{\boldsymbol{p},\boldsymbol{r}} \ge 0, \quad \sum_{\boldsymbol{r}} b_{\boldsymbol{p},\boldsymbol{r}} = 1. \tag{21}$$

Border replication preserves this argument at boundaries. Thus transport retains the local update bound while allowing signed coefficients to represent both smoothing and sharpening.

A.4 Reference-independent HSI path

In Eq. (6), $\mathbf{F}_0$ depends only on $\mathbf{L}$, and each $\mathbf{F}_t$ is computed from the preceding HSI features. For fixed parameters, $\mathbf{R}_{\mathrm{H}} = g(\mathbf{F}_S)$ therefore provides a reference-independent estimate. The reference acts through the accumulated correction $\mathbf{T}_S$ and the guided residual $\mathbf{R}_{\mathrm{G}}$. This separation retains the HSI estimate while allowing the transferred spatial information to refine it.

B Spatial operator transport: experiments

B.1 Operator components and replacement controls

The component ablations in Table 3 retain the HSI backbone, spectral controller, recurrent stages, and training protocol. The *w/o neighborhood weights* variant removes local propagation; *w/o affine state* removes additive detail injection; *w/o alignment* applies the learned fields at the corresponding grid coordinates without displacement compensation; and *w/o self-reconstruction* sets the reference reconstruction loss weight to zero while retaining the reference branch. Relative to 37.75 dB for ROTSA on REAL $\times 8$, these variants give 37.61, 36.47, 37.63, and 37.66 dB, respectively. The largest decrease, 1.28 dB, occurs without the affine state, supporting explicit reconstruction of detail missing from the coarse observations. The 0.12 dB difference without alignment also shows that the spatial reconstruction operator remains useful before displacement compensation.

Table 7 compares alternative local reconstruction operators within the same framework. *CARAFE* replaces the proposed weight-and-state operator with content-aware feature reassembly (Wang et al., 2019). *DKN* replaces it with deformable kernel filtering, which predicts sampling offsets and local weights (Kim et al., 2021). For a controlled comparison, we replace only the reference-operator predictor and its matched application rule: CARAFE predicts softmax-normalized 3×3 fixed-neighborhood kernels from the coarse reference, whereas DKN predicts zero-sum 3×3 weights and fractional offsets from the HR reference and its bicubic coarse-scale baseline; both reconstruct the same next reference-pyramid level under the identical self-reconstruction loss, while the HSI restorer, reference pyramid, alignment and pullback, recurrent schedule, spectral routing, and all loss weights remain unchanged. These are operator replacements within ROTSA, trained and evaluated under the same REAL $\times 8$ protocol.

ROTSA exceeds CARAFE and DKN by 4.27 and 1.25 dB, respectively. It also achieves higher SSIM (0.973 versus 0.939 and 0.965) and lower SAM (0.042 versus 0.049 and 0.046). Together with the component ablations, this comparison supports the combination of reference self-reconstruction, local propagation, and additive detail for unregistered HSI reconstruction.

Table 7: Spatial operator replacement on REAL $\times 8$. Each alternative replaces both the neighborhood-weight and affine-state paths; the rest of the framework is unchanged.

Spatial operator	PSNR↑	SSIM↑	SAM↓
CARAFE	33.48	0.939	0.049
DKN	36.50	0.965	0.046
ROTSA	**37.75**	**0.973**	**0.042**

B.2 Alignment across training and test conditions

We evaluate whether a model trained with slight misregistration can also use clean and more strongly misregistered references. For each method, the same slight-trained model is tested on both clean and large-misregistration inputs. Separately trained clean and large models provide matched-condition comparisons. All experiments use ICVL $\times 8$, the same scene split, and the same test condition within each block of Table 8.

ROTSA w/o alignment disables displacement compensation for both transported fields while preserving the operator learner, HSI backbone, spectral controller, and training protocol. *ROTSA* retains structural correspondence. *UAFL* and *SSC-HSR* are evaluated as complete comparison methods under the same train–test conditions. Clean inputs use aligned references. Slight references follow the camera-pose perturbation protocol (Liang et al., 2023a; Zhang et al., 2025), whereas large references use sequential global, perspective, and elastic transformations with the parameters specified in Appendix E.1.

Table 8: Alignment under matched and changed train–test conditions on ICVL $\times 8$. Each slight-trained model is evaluated on both test conditions. Best results are bold within each train–test block.

Train	Test	Method	PSNR↑	SSIM↑	SAM↓
Clean	Clean	ROTSA	48.78	**0.997**	**0.024**
		ROTSA w/o alignment	**48.89**	**0.997**	**0.024**
		UAFL	44.30	0.993	0.036
		SSC-HSR	41.51	0.990	0.042
Slight	Clean	ROTSA	**41.12**	**0.989**	**0.034**
		ROTSA w/o alignment	39.99	0.986	0.037
		UAFL	38.03	0.976	0.039
		SSC-HSR	35.93	0.968	0.045
Large	Large	ROTSA	**35.71**	**0.960**	**0.036**
		ROTSA w/o alignment	34.78	0.951	**0.036**
		UAFL	34.32	0.947	0.037
		SSC-HSR	34.99	**0.960**	0.043
Slight	Large	ROTSA	**34.74**	**0.951**	**0.037**
		ROTSA w/o alignment	34.38	0.948	**0.037**
		UAFL	33.89	0.944	0.038
		SSC-HSR	33.34	0.938	0.044

With trained on slight misregisteration, ROTSA reaches 41.12 dB on clean inputs and 34.74 dB under large misregistration. Alignment improves PSNR over the corresponding no-alignment variant by 1.13 and 0.36 dB. Under the same fixed training condition, ROTSA exceeds UAFL by 3.09 and 0.85 dB, and SSC-HSR by 5.19 and 1.40 dB, respectively. It achieves the best PSNR and SSIM in both tests, with the best or tied-best SAM.

The matched-condition models reach 48.78 dB on clean inputs and 35.71 dB under large misregistration. Relative to these results, the slight-trained ROTSA is lower by 7.66 and 0.97 dB, respectively. The matched clean comparison gives a 0.11 dB advantage to the no-alignment variant with identical SSIM and SAM, whereas matched large training benefits from alignment by 0.93 dB. These results distinguish the benefit of training on the target condition from the contribution of correspon-

dence: correspondence improves transfer under the changed test conditions, while operator-based reconstruction remains effective in the no-alignment control.

B.3 Output-domain reference responses and reconstruction errors

We compare UAFL, direct feature fusion, and ROTSA on the same example to examine how reference guidance affects the HSI output around displaced structures. Direct fusion replaces the spatial operator with feature fusion while retaining the HSI restoration backbone, spectral controller, and training protocol. UAFL provides a comparison with an existing fusion architecture.

Total reference-induced response. Let $f_m(\mathbf{L}, \mathbf{R})$ denote the final HSI output of method m. With model parameters and the LR-HSI fixed, we measure the response to the reference by:

$$\mathbf{D}^{(m)} = f_m(\mathbf{L}, \mathbf{R}) - f_m(\mathbf{L}, \mathbf{0}), \qquad M^{(m)}(\boldsymbol{p}) = \left[\frac{1}{C} \sum_{c=1}^{C} D_c^{(m)}(\boldsymbol{p})^2 + \epsilon \right]^{1/2}. \tag{22}$$

Here $\mathbf{0}$ replaces the reference input in a complete second forward pass and $\epsilon = 10^{-12}$. Both predictions include all reconstruction stages and the method's final output mapping. Fig. 1 displays the cross-band RMS maps $M^{(m)}$ with one shared color scale. UAFL and direct fusion show stronger responses around displaced reference structures, while ROTSA concentrates its response more closely on the HSI structure. The final-output errors below assess the reconstruction quality associated with these response patterns.

Edge regions. For an image $\mathbf{X}$ with C_X channels, spatially standardize each channel to obtain $\widetilde{\mathbf{X}}$ and compute:

$$G_{\mathbf{X}}(\boldsymbol{p}) = \left[\frac{1}{C_X} \sum_{c=1}^{C_X} \big((\partial_x \widetilde{X}_c(\boldsymbol{p}))^2 + (\partial_y \widetilde{X}_c(\boldsymbol{p}))^2 \big) + \epsilon \right]^{1/2}. \tag{23}$$

The derivatives use finite differences. Pixels at or above each image's 88th gradient percentile define the HSI and reference edge masks, Ω_{H} and Ω_{R}. The main figure shows reference edges in magenta and HSI edges in green. With $\mathcal{D}_2$ denoting dilation by two pixels, the reference-only and flat regions are:

$$\Omega_{\mathrm{ref}} = \Omega_{\mathrm{R}} \setminus \mathcal{D}_2(\Omega_{\mathrm{H}}), \qquad \Omega_{\mathrm{flat}} = \big[\mathcal{D}_2(\Omega_{\mathrm{R}} \cup \Omega_{\mathrm{H}}) \big]^c. \tag{24}$$

These common evaluation masks distinguish HSI boundaries, displaced reference boundaries, and areas away from either set of edges.

Spatial selectivity of the response. To quantify variation around different structures, we compute the spatial gradient magnitude of the output-domain response:

$$Q^{(m)}(\boldsymbol{p}) = \left[\frac{1}{C} \sum_{c=1}^{C} \big((\partial_x D_c^{(m)}(\boldsymbol{p}))^2 + (\partial_y D_c^{(m)}(\boldsymbol{p}))^2 \big) + \epsilon \right]^{1/2},$$

$$q_\Omega^{(m)} = \frac{1}{|\Omega|} \sum_{\boldsymbol{p} \in \Omega} Q^{(m)}(\boldsymbol{p}), \qquad r_\Omega^{(m)} = \frac{q_\Omega^{(m)}}{\max(q_{\Omega_{\mathrm{H}}}^{(m)}, \epsilon)}. \tag{25}$$

The response gradients use the original output scale. Normalizing by each method's HSI-edge response sets its HSI-edge bar to one. Smaller reference-only and flat-region bars indicate a stronger concentration of response variation around HSI structure. The selectivity annotation reports $q_{\Omega_{\mathrm{H}}}^{(m)} / \max(q_{\Omega_{\mathrm{ref}}}^{(m)}, \epsilon)$. In Fig. 6, ROTSA has lower relative responses in reference-only and flat regions, consistent with more selective spatial guidance.

Final reconstruction error. For each method, the error map and regional mean absolute error are:

$$E^{(m)}(\boldsymbol{p}) = \frac{1}{C} \sum_{c=1}^{C} |\widehat{Y}_c^{(m)}(\boldsymbol{p}) - Y_c(\boldsymbol{p})|, \qquad \mathrm{MAE}_\Omega^{(m)} = \frac{1}{|\Omega|} \sum_{\boldsymbol{p} \in \Omega} E^{(m)}(\boldsymbol{p}). \tag{26}$$

Fig. 6 reports the reconstructions and errors for the same example. The HSI is displayed with bands 21, 11, and 3 as pseudo-RGB channels. All reconstruction panels share the ground-truth 1st–99th

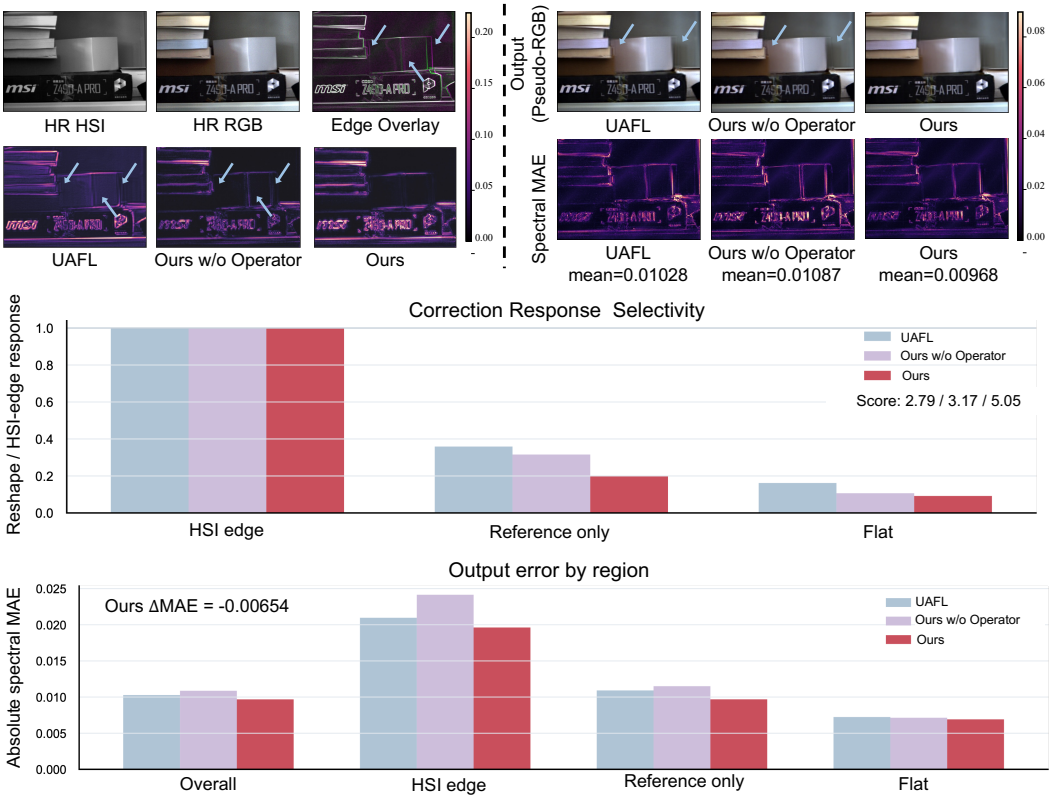

Figure 6: Spatial analysis for the example in Fig. 1. Top: final reconstructions and cross-band MAE maps. Bottom: normalized output-domain reference-response variation and regional reconstruction error. ROTSA reduces relative responses around displaced reference structures and gives lower reconstruction errors.

percentile display ranges, and the error maps share a color scale. The mean MAE is 0.01028 for UAFL, 0.01087 for direct fusion, and 0.00968 for ROTSA, corresponding to reductions of 5.8% and 10.9% relative to the two comparisons. ROTSA also gives lower error in the displayed edge and flat regions. The base-to-final annotation measures the MAE change from $\mathbf{B} + \mathbf{R}_{\mathrm{H}}$ to the final ROTSA output. The response visualization and output-error measurements together support selective use of reference structure for reconstruction under misregistration.

C Spectral Adaptation: Formulation

C.1 Complete Basis and Observable Descriptors

The LR-HSI is centered by subtracting each band's spatial mean. Horizontal and vertical differences between adjacent spectra form the local samples in Eq. (10), with $M = h(w-1) + (h-1)w$ and $\lambda_\ell = 0.5$. Its eigenvectors are ordered by decreasing eigenvalue to form the complete orthonormal basis $\boldsymbol{E}$. Each mode represents a linear combination of spectral bands.

Let $\epsilon_k^{\mathrm{g}} = \max(\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{\mathrm{g}} \boldsymbol{e}_k, 0)$ and $\epsilon_k^{\ell} = \max(\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{\ell} \boldsymbol{e}_k, 0)$, where the clamps remove negative round-off. With $\mathcal{N}(\boldsymbol{v}) = \boldsymbol{v} / \max(\sum_j v_j, 10^{-12})$, the scene descriptors are:

$$\boldsymbol{p}^{\mathrm{g}} = \mathcal{N}(\boldsymbol{\epsilon}^{\mathrm{g}}), \qquad \boldsymbol{p}^{\ell} = \mathcal{N}(\boldsymbol{\epsilon}^{\ell}), \qquad \boldsymbol{p} = \mathcal{N}(\boldsymbol{\epsilon}^{\mathrm{g}} + \lambda_\ell \boldsymbol{\epsilon}^{\ell}). \tag{27}$$

The combined distribution normalizes the sum of the unnormalized energies. The residual and proposed correction are projected as:

$$\mathbf{A}_{\mathrm{G}} = \boldsymbol{E}^\top \mathbf{R}_{\mathrm{G}}, \qquad \mathbf{A}_{\Delta} = \boldsymbol{E}^\top (\mathbf{R}_{\mathrm{G}} - \mathbf{R}_{\mathrm{H}}). \tag{28}$$

Their mode-wise spatial energies give two further descriptors:

$$e_k^{\mathrm{G}} = \frac{\|[\mathbf{A}_{\mathrm{G}}]_k\|_F^2}{HW}, \qquad e_k^{\Delta} = \frac{\|[\mathbf{A}_{\Delta}]_k\|_F^2}{HW}, \qquad \boldsymbol{q}^{\mathrm{G}} = \mathcal{N}(\boldsymbol{e}^{\mathrm{G}}), \quad \boldsymbol{q}^{\Delta} = \mathcal{N}(\boldsymbol{e}^{\Delta}). \tag{29}$$

We compute these energies from the uncentered coefficient maps and normalize each energy vector independently across modes. Together, the four descriptors characterize the observed spectral content and the proposed reference correction.

C.2 Prior and continuous weights

For mode k, the remaining scene energy is $u_k = \max(1 - \sum_{j<k} p_j, 10^{-8})$. It determines the prior score:

$$\pi_k = \mathrm{clip}\left(\frac{\log u_k - \log \delta_0}{\tau}, -\chi, \chi\right), \qquad \delta_0 = 10^{-4}, \quad \tau = 0.5, \quad \chi = 4. \tag{30}$$

For normalized energy x, write $\ell(x) = \mathrm{clip}(\log \max(x, 10^{-8}), -18, 0)/18$. The shared controller receives:

$$\boldsymbol{z}_k = \big[\ell(p_k^{\mathrm{g}}), \ell(p_k^{\ell}), \ell(q_k^{\mathrm{G}}), \ell(q_k^{\Delta}), \pi_k/\chi, \nu_k\big], \qquad \nu_k = \frac{k-1}{C-1}. \tag{31}$$

Its two-layer MLP and weight mapping are:

$$\begin{aligned} s_k &= \boldsymbol{W}_2\, \mathrm{GELU}(\boldsymbol{W}_1 \boldsymbol{z}_k + \boldsymbol{b}_1) + b_2, \\ a_k &= \sigma(\pi_k + \xi \tanh s_k), \qquad \xi = 6. \end{aligned} \tag{32}$$

Here $\boldsymbol{W}_1 \in \mathbb{R}^{16\times 6}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{1\times 16}$. The final MLP layer is initialized to zero, giving $a_k = \sigma(\pi_k)$ initially. Learned weights need not be monotonic in mode index. Each weight is shared across the spatial positions of its coefficient map.

The basis and descriptors are computed with gradients detached. The controller receives supervision through the final reconstruction, and gradients reach both residual predictions through their contributions to that reconstruction.

C.3 Complete-basis reconstruction and correction energy

The weighted correction in Eq. (1) is:

$$\Delta\mathbf{R}_a = \boldsymbol{E}\,\mathrm{diag}(\boldsymbol{a})\boldsymbol{E}^{\top}\Delta\mathbf{R}. \tag{33}$$

With $\mathbf{A}_{\mathrm{H}} = \boldsymbol{E}^{\top}\mathbf{R}_{\mathrm{H}}$, completeness of the basis gives:

$$\widehat{\mathbf{Y}} = \mathbf{B} + \boldsymbol{E}\mathbf{A}_{\mathrm{out}}, \qquad [\mathbf{A}_{\mathrm{out}}]_k = (1 - a_k)[\mathbf{A}_{\mathrm{H}}]_k + a_k[\mathbf{A}_{\mathrm{G}}]_k. \tag{34}$$

Thus each spectral coordinate interpolates between the HSI and guided residuals. The complete basis retains the HSI contribution across all modes while controlling the strength of the additional reference correction.

For fixed input and weights, orthogonality yields:

$$\|\boldsymbol{E}\,\mathrm{diag}(\boldsymbol{a})\boldsymbol{E}^{\top}\Delta\mathbf{R}\|_F^2 = \sum_{k=1}^{C} a_k^2 \|[\mathbf{A}_{\Delta}]_k\|_F^2 \le \|\Delta\mathbf{R}\|_F^2. \tag{35}$$

The spectral controller therefore redistributes the reference contribution across modes while keeping its total energy bounded by that of the proposed correction.

Changing an eigenvector's sign changes its coefficient signs but preserves the squared-energy descriptors. For a diagonal sign matrix $\boldsymbol{S}$, $(\boldsymbol{E}\boldsymbol{S})\,\mathrm{diag}(\boldsymbol{a})(\boldsymbol{E}\boldsymbol{S})^{\top} = \boldsymbol{E}\,\mathrm{diag}(\boldsymbol{a})\boldsymbol{E}^{\top}$. The weighted reconstruction is therefore invariant to eigenvector sign choices.

C.4 FIXED-RANK AND FULL-CAPACITY CONTROLS

Fixed interaction rank K replaces the learned weights by $m_k = \mathbb{I}[k \le K]$:

$$\mathbf{R}_{\text{fixed}} = \mathbf{R}_{\text{H}} + \boldsymbol{E}\,\text{diag}(\boldsymbol{m})\boldsymbol{E}^{\top}(\mathbf{R}_{\text{G}} - \mathbf{R}_{\text{H}}). \tag{36}$$

The first K modes use the guided residual and the remaining modes retain the HSI residual; all variants retain the full-band Bicubic base. Full capacity sets every weight to one without adaptive reweighting. These controls keep the restoration backbone, spatial operator, and training protocol unchanged and replace only the spectral interaction rule. In Table 5, adaptive capacity improves PSNR over fixed rank three and full capacity by 0.19 and 0.30 dB, respectively.

C.5 MODE-WISE RECONSTRUCTION ERROR

For fixed residual predictions and basis, let $\boldsymbol{t}_k = [\boldsymbol{E}^{\top}(\mathbf{Y} - \mathbf{B} - \mathbf{R}_{\text{H}})]_k$ denote the missing target component and $\boldsymbol{d}_k = [\boldsymbol{E}^{\top}\Delta\mathbf{R}]_k$ the proposed correction. Applying a weight a gives:

$$\mathcal{E}_k(a) = \|\boldsymbol{t}_k - a\boldsymbol{d}_k\|_F^2 = \|\boldsymbol{t}_k\|_F^2 - 2a\langle \boldsymbol{t}_k, \boldsymbol{d}_k\rangle + a^2\|\boldsymbol{d}_k\|_F^2. \tag{37}$$

For $\boldsymbol{d}_k \neq 0$, minimizing over $[0, 1]$ yields:

$$a_k^{\star} = \text{clip}\left(\frac{\langle \boldsymbol{t}_k, \boldsymbol{d}_k\rangle}{\|\boldsymbol{d}_k\|_F^2}, 0, 1\right). \tag{38}$$

The expression relates an effective correction strength to both its magnitude and its agreement with the missing signal. It motivates combining scene and residual information in the learned controller. The target-dependent optimum provides an analytical reference. ROTSA estimates its weights from the observable descriptors in Eq. (31).

D SPECTRAL ADAPTATION: EXPERIMENTS

D.1 SCENE AND RESIDUAL EVIDENCE FOR CAPACITY ALLOCATION

We compare two simplified controllers in Table 9. Each replaces only capacity allocation; the HSI backbone, spatial operator, correspondence module, losses, and remaining training settings are kept unchanged.

Global scalar. This variant pools the mode descriptors and prior with the normalized scene energies, predicts one input-dependent weight, and broadcasts it across all modes:

$$\overline{\boldsymbol{z}} = \sum_k p_k \boldsymbol{z}_k, \qquad \overline{\pi} = \sum_k p_k \pi_k, \qquad a_k = \sigma(\overline{\pi} + \xi \tanh h_{\psi}(\overline{\boldsymbol{z}}))\,. \tag{39}$$

It retains scene-dependent strength adjustment while applying a common strength to every spectral coordinate.

Prior only. This variant retains the global–local basis and combined scene energies but removes the learned adjustment: $a_k = \sigma(\pi_k)$.

Table 9: Capacity-controller ablation on REAL $\times 8$. Only the spectral allocation rule or its inputs are changed; all remaining components are fixed.

Variant	PSNR↑	SSIM↑	SAM↓
Global scalar	37.55	0.971	0.044
Prior only	37.60	0.972	**0.042**
ROTSA	**37.75**	**0.973**	**0.042**

The full controller reaches 37.75 dB, exceeding global scalar and prior only by 0.20 and 0.15 dB, respectively. It also gives the highest SSIM and the lowest reported SAM, tied with prior only. The scalar comparison supports adapting strengths across modes. The additional improvement to 37.75 dB shows the value of describing the correction currently proposed by the reference branch.

D.2 SENSITIVITY TO A FIXED INTERACTION RANK

Table 10 varies the retained interaction rank K in equation 36 while keeping the backbone, spatial operator, training objective, and evaluation protocol unchanged. Fixed ranks two through five allow the guided residual in the corresponding leading modes and retain the HSI residual in the others. ROTSA instead applies learned continuous weights across the complete basis.

Table 10: Fixed interaction ranks on REAL $\times 8$. The unselected modes retain the HSI prediction.

Variant	PSNR↑	SSIM↑	SAM↓
Fixed rank = 2	37.38	0.971	0.044
Fixed rank = 3	37.56	0.972	0.043
Fixed rank = 4	37.59	0.972	0.043
Fixed rank = 5	37.53	0.972	0.043
ROTSA	**37.75**	**0.973**	**0.042**

The highest PSNR among the fixed ranks is 37.59 dB at rank four; rank five gives 37.53 dB. Adaptive capacity improves on the best tested fixed rank by 0.16 dB, with higher SSIM and lower SAM. Together with Table 9, these results support adapting both the distribution and strength of reference corrections to the input.

D.3 SCENE-WISE RANK ANALYSIS

Basis–coefficient reconstruction. To examine scene variation in spectral dimensionality, we follow the SVD factor representation in UAFL (Zhang et al., 2026). Reshape the LR and HR observations as $\boldsymbol{L} \in \mathbb{R}^{C\times hw}$ and $\boldsymbol{Y} \in \mathbb{R}^{C\times HW}$ and decompose them independently:

$$\boldsymbol{L} = \boldsymbol{U}_{\mathrm{L}}\boldsymbol{S}_{\mathrm{L}}\boldsymbol{V}_{\mathrm{L}}^{\top}, \qquad \boldsymbol{Y} = \boldsymbol{U}_{\mathrm{Y}}\boldsymbol{S}_{\mathrm{Y}}\boldsymbol{V}_{\mathrm{Y}}^{\top}. \tag{40}$$

The SVD is applied to the uncentered HSI, with components ordered by decreasing singular value. The LR decomposition supplies a spectral basis, while the HR decomposition supplies spatial coefficient maps. Retaining the first k components gives:

$$\boldsymbol{A}_{\mathrm{Y},k} = \boldsymbol{U}_{\mathrm{Y},k}^{\top}\boldsymbol{Y}, \qquad \widehat{\boldsymbol{Y}}_k^{\mathrm{diag}} = \boldsymbol{U}_{\mathrm{L},k}\boldsymbol{A}_{\mathrm{Y},k}. \tag{41}$$

The HR coefficients provide an oracle analysis of LR–HR factor pairing. Successively retaining the leading components yields a PSNR curve for each scene. Its maximizing component count defines:

$$\begin{aligned} k_i^{\star} &= \arg\max_k \mathrm{PSNR}(\widehat{\boldsymbol{Y}}_{i,k}^{\mathrm{diag}}, \boldsymbol{Y}_i), \\ \Delta_{\mathrm{rank}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{PSNR}(\widehat{\boldsymbol{Y}}_{i,k_i^{\star}}^{\mathrm{diag}}, \boldsymbol{Y}_i) - \mathrm{PSNR}(\widehat{\boldsymbol{Y}}_{i,3}^{\mathrm{diag}}, \boldsymbol{Y}_i)\right]. \end{aligned} \tag{42}$$

Here k counts the leading paired components, with consistent ordering and sign conventions for the LR and HR decompositions. Both the HR coefficient construction and the scene-wise rank selection use the target for this analysis.

REAL and ICVL observations. The main-text analysis uses 57 REAL scenes at $\times 8$. The selected rank is three in 15 scenes and differs from three in 42, with selected values from one to seven. Fixed rank three gives 43.437 dB, whereas scene-wise selection gives 45.517 dB, a 2.080 dB difference. The additional ICVL experiment in Fig. 7 uses 20 scenes, also at $\times 8$. The selected counts are two for one scene, three for six, four for nine, and five for four. Hence 14 of 20 scenes favor a count other than three. The corresponding means are 53.054 and 54.393 dB, a 1.339 dB difference. Thus the selected rank differs from three in 73.7% of REAL scenes and 70.0% of ICVL scenes. The scene-dependent distributions motivate adapting spectral interactions to each input.

Connection to adaptive capacity. ROTSA implements input-dependent spectral correction through continuous weights over a complete LR-derived basis. Its controller combines scene statistics with predicted residuals, and its performance is evaluated by the trained controls in Tables 9 and 10.

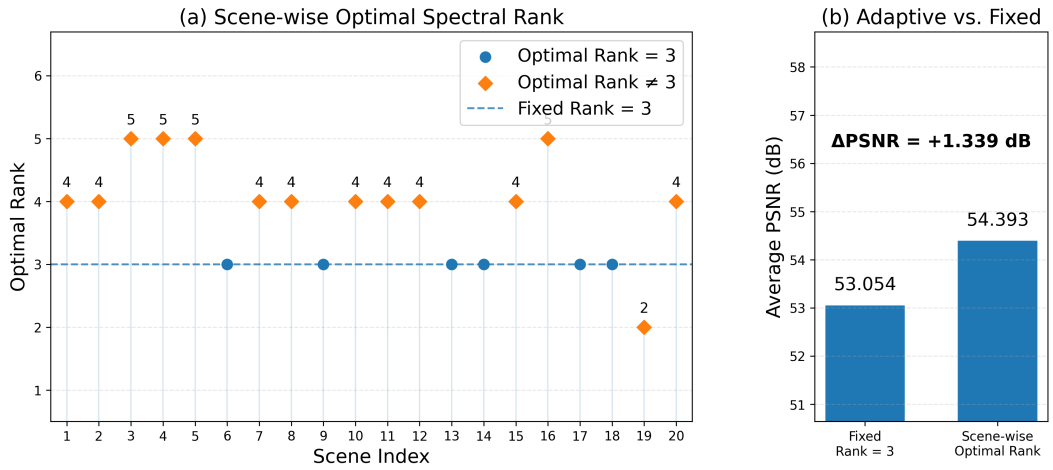

Figure 7: Scene-wise rank analysis on 20 ICVL scenes at $\times 8$, using LR basis vectors and HR coefficient maps. Target-based rank selection improves the mean PSNR from 53.054 to 54.393 dB.

E Datasets, Training, and Evaluation

E.1 Datasets and Observation Protocol

ICVL: simulated unregistered guidance. ICVL provides natural-scene hyperspectral images acquired at 1392×1300 spatial resolution. We use the distributed 31-band representation covering 400–700 nm at 10 nm intervals (Arad & Ben-Shahar, 2016). Following the UAFL benchmark (Zhang et al., 2026), images are center-cropped to 512×512, RGB references are synthesized with the Nikon D700 spectral response, and their spatial correspondence is controlled by the settings below. The benchmark partition reserves 20 scenes for testing and uses the remaining scenes for training; the split is made before any crop extraction.

ICVL misregistration settings. The *clean* setting uses aligned RGB references. The *slight* setting follows the previous camera-pose-based generation protocol, perturbing the virtual-camera pose to produce a shifted reference viewpoint (Liang et al., 2023a; Zhang et al., 2025). The *large* setting applies three transformations sequentially to the HR reference: a global geometric perturbation with a rotation of $3.0°$, a scale factor of 1.05, and a translation of $(5, -3)$ pixels; a perspective transform with strength 0.02 to simulate parallax; and an elastic deformation with $\alpha = 15$ and $\sigma = 3$ for local non-rigid distortion. These settings are used consistently across methods in the matched-condition and cross-condition comparisons.

REAL: acquired cross-view misregistration. REAL contains 57 indoor and outdoor scene pairs acquired with a dual-camera system using SOC710-VP hyperspectral cameras (Lai et al., 2024). We use the processed HSI–RGB pairs following UAFL, with 47 scenes for training and 10 for testing. The target and reference views exhibit acquired misregistration, including viewpoint differences and parallax. The LR-HSI is generated from the target using the spatial degradation below.

Degradation and normalization. The data loader normalizes each HSI cube to $[0, 1]$ using a single minimum and maximum across its bands and pixels, then applies the same spatial degradation independently to every band:

$$\mathbf{L} = \downarrow_s (\boldsymbol{k}_{\mathrm{HSI}} * \mathbf{Y}), \qquad \boldsymbol{k}_{\mathrm{HSI}}(i,j) = \frac{\exp[-(i^2+j^2)/(2\sigma^2)]}{\sum_{u,v}\exp[-(u^2+v^2)/(2\sigma^2)]}. \tag{43}$$

The benchmark uses an 8×8 support with $\sigma = 3$, centered at half-pixel coordinates, and retains every s-th pixel for $s \in \{4, 8, 16\}$. RGB values are scaled by 255. The bicubic base is derived from $\mathbf{L}$. The reference construction kernel in Eq. (2) is learned jointly with reconstruction.

E.2 Training protocol and controlled comparisons

In addition to the training protocol of the main paper, the coefficients λ_{sam}, λ_h, and λ_r are set to 0.01, 0.1, and 0.05 in our experiments, respectively. The learned reference kernel is excluded from weight decay. All ablations retain the same data split, observation generation, and training protocol. Apart from the component specified by each ablation, the architecture is kept unchanged. For alignment transfer, the slight-trained model is reused across test conditions as described in Appendix B.2. For each ablation, the named component or rule is changed while the remaining architecture and training protocol are kept unchanged.

E.3 Evaluation metrics and aggregation

Let $\widehat{\mathbf{Y}}, \mathbf{Y} \in \mathbb{R}^{C\times H\times W}$ be an output and its target. Metrics are first evaluated for each image, then averaged across evaluation images. PSNR and SSIM are averaged over bands; SAM compares the full spectral vectors at each pixel. Visualization contrast adjustments are not used when computing these metrics.

Peak signal-to-noise ratio. With unit intensity range, the reported band-averaged PSNR is:

$$\mathrm{MSE}_c = \frac{1}{HW}\sum_{\boldsymbol{p}} (\widehat{Y}_c(\boldsymbol{p}) - Y_c(\boldsymbol{p}))^2, \qquad \mathrm{PSNR} = \frac{1}{C}\sum_{c=1}^{C} 10\log_{10}\frac{1}{\mathrm{MSE}_c}. \tag{44}$$

We average the band-wise PSNR values to measure radiometric reconstruction accuracy.

Structural similarity. SSIM is evaluated separately for each spectral band using the benchmark evaluator, then averaged across bands and evaluation images. It measures spatial structure and complements the full-spectrum comparison in SAM. Metric inputs use the benchmark intensity normalization, without visualization contrast adjustments.

Spectral angle mapper. Let $\widehat{\boldsymbol{y}}(\boldsymbol{p})$ and $\boldsymbol{y}(\boldsymbol{p})$ denote the C-dimensional spectra. The evaluation routine computes:

$$\mathrm{SAM} = \frac{1}{HW}\sum_{\boldsymbol{p}} \arccos\left[\mathrm{clip}\left(\frac{\widehat{\boldsymbol{y}}(\boldsymbol{p})^\top \boldsymbol{y}(\boldsymbol{p}) + \epsilon}{(\|\widehat{\boldsymbol{y}}(\boldsymbol{p})\|_2 + \epsilon)(\|\boldsymbol{y}(\boldsymbol{p})\|_2 + \epsilon)}, -1, 1\right)\right], \qquad \epsilon = 10^{-8}. \tag{45}$$

We report radians. Smaller angles indicate better spectral agreement; away from numerical regularization, the angle is unchanged by a positive rescaling of either spectrum.

F Additional controlled ablations

F.1 Reference-kernel support

Table 11 changes only the spatial support of the learned coarse-reference kernel; the remaining architecture and training protocol are fixed. The 6×6, 8×8, 10×10, and 12×12 variants therefore examine the extent of reference aggregation used to construct the coarse observation.

Table 11: Learned reference-kernel support on REAL $\times 8$.

Kernel support	PSNR↑	SSIM↑	SAM↓
6×6	37.67	0.972	0.043
8×8 (ROTSA)	**37.75**	**0.973**	**0.042**
10×10	37.69	0.972	**0.042**
12×12	37.67	**0.973**	**0.042**

PSNR ranges from 37.67 to 37.75 dB, a spread of 0.08 dB, while SAM ranges from 0.042 to 0.043. The 8×8 support gives the highest PSNR and ties for the best SSIM and SAM, providing an effective reference construction within the tested range.

F.2 TRAINING OBJECTIVE ABLATION

Training objective. The reconstruction terms in Eq. (13) are means over bands and pixels; the reference term is additionally averaged over stages. Spectral-angle supervision averages the angle between predicted and target spectra:

$$\mathcal{L}_{\mathrm{SAM}}(\mathbf{U},\mathbf{V}) = \frac{1}{HW}\sum_{\boldsymbol{p}} \arccos\left(\mathrm{clip}\left[\frac{\langle \mathbf{U}(\boldsymbol{p}),\mathbf{V}(\boldsymbol{p})\rangle}{\max(\|\mathbf{U}(\boldsymbol{p})\|_2\|\mathbf{V}(\boldsymbol{p})\|_2,\epsilon)}, -1+\eta, 1-\eta\right]\right). \tag{46}$$

For numerical stability, $\epsilon = 10^{-8}$ and $\eta = 10^{-6}$; numerically identical spectra are assigned zero loss. The auxiliary HSI reconstruction is $\mathbf{B} + \mathbf{R}_{\mathrm{H}}$; removing its supervision does not remove the HSI branch. Reference self-reconstruction is supervised by the reference pyramid, complementing the HR-HSI supervision of the restoration output.

Definition of the controls. Removing reference self-reconstruction sets $\lambda_r = 0$; removing SAM sets $\lambda_{\mathrm{sam}} = 0$. The *w/o HSI expert supervision* variant sets $\lambda_h = 0$ but retains the branch and its role in the final prediction. Reconstruction only retains $\|\widehat{\mathbf{Y}} - \mathbf{Y}\|_1$. All variants retain the same architecture and optimization settings and change only the specified loss terms.

Table 12: Loss ablation on REAL $\times 8$. Architecture and optimization settings are unchanged.

Training objective	PSNR↑	SSIM↑	SAM↓
w/o self-reconstruction	37.66	0.972	0.043
w/o SAM	37.56	0.971	0.043
w/o HSI expert supervision	37.65	0.972	**0.042**
Reconstruction only	37.61	0.971	0.043
ROTSA	**37.75**	**0.973**	**0.042**

Effect of supervision. The complete objective gives 37.75 dB, 0.973 SSIM, and 0.042 SAM. Removing self-reconstruction, SAM, or HSI auxiliary supervision reduces PSNR by 0.09, 0.19, and 0.10 dB, respectively. Removing SAM also lowers SSIM to 0.971 and increases SAM to 0.043. Reconstruction only gives 37.61 dB with the same SSIM and SAM. The combined objective gives the highest reported PSNR in this comparison. The complementary reconstruction and spectral objectives improve the final result jointly.

F.3 REFERENCE PYRAMID CONSTRUCTION

Table 13 varies the construction of coarse and intermediate reference levels while retaining the operator learner, HSI restoration backbone, correspondence module, spectral controller, and training protocol. The observed HR reference remains the finest level in every variant.

Area/Area and *Bicubic/Bicubic* directly resize the HR reference to all required resolutions using the respective interpolation rule. The three *learned-convolution* variants use the learned kernel for the coarsest reference and differ in their intermediate targets: *recursive convolution* constructs these levels through successive convolutions, whereas *Area* and *Bicubic* resize the HR reference directly. *Matched Gaussian/Bicubic* uses the known HSI degradation to construct the coarsest reference, and *Matched Gaussian/Matched Gaussian* uses the corresponding known degradation at every coarse and intermediate level.

Learning the coarsest reference with Bicubic intermediate levels reaches 37.75 dB, improving on Area/Area and Bicubic/Bicubic by 0.50 and 0.39 dB. With the learned coarse reference fixed, direct Area and Bicubic resampling give 37.66 and 37.75 dB, compared with 37.28 dB for recursive convolution. These results favor learning the coarse reference while forming intermediate self-reconstruction targets directly from the observed HR image.

The fully matched construction reaches 37.77 dB, a 0.02 dB difference from ROTSA, with the same SAM of 0.042. ROTSA gives higher SSIM, 0.973 versus 0.972. Matching only the coarsest level gives 37.65 dB and the lowest SAM, 0.041. The comparison shows that the learned construction attains similar reconstruction quality to the fully matched reference pyramid while estimating its coarse reference directly from the RGB observation.

Table 13: Reference pyramid construction on REAL $\times 8$. The remaining model and training settings are unchanged.

Coarsest reference	Intermediate levels	PSNR↑	SSIM↑	SAM↓
Area	Area	37.25	0.971	0.043
Bicubic	Bicubic	37.36	0.971	0.042
Learned convolution	Recursive convolution	37.28	0.970	0.043
Learned convolution	Area	37.66	0.972	0.042
Learned convolution	Bicubic (ROTSA)	37.75	**0.973**	0.042
Matched Gaussian	Bicubic	37.65	0.972	**0.041**
Matched Gaussian	Matched Gaussian	**37.77**	0.972	0.042

G Computational analysis

Shared recurrent parameters keep the model size independent of the number of stages, although computation changes with the number and resolution of those stages. At a fixed HR output size and $S = \log_2 s$, the total number of spatial positions processed across stages is

$$\sum_{t=1}^{S} (2^t h)(2^t w) = \frac{4}{3} HW(1 - 4^{-S}) < \frac{4}{3} HW. \tag{47}$$

Thus progressively finer processing has a geometric area sum rather than applying every recurrent stage at the final resolution. The eight-neighbor propagation uses $O(8d' H_t W_t)$ operations for projected width d'; the surrounding learned convolutions add their usual channel-dependent cost. Constructing the spectral statistics costs $O(C^2 hw)$, eigendecomposition costs $O(C^3)$, and full-resolution projection and synthesis cost $O(C^2 HW)$. The controller itself acts on only C six-dimensional vectors per image. The comparison in Table 1 measures FLOPs at 256×256 output resolution and $\times 4$. ROTSA uses 1.39 M parameters and 101.94 G FLOPs. Relative to UAFL (5.94 M, 96.17 G), it reduces parameters by 76.6% with 6.0% more FLOPs. Relative to SSDAN (6.77 M, 337.27 G), it reduces parameters and FLOPs by 79.5% and 69.8%, respectively.

H Additional qualitative comparisons

Fig. 8 and Fig. 9 provide additional comparisons on REAL $\times 4$ and ICVL $\times 8$/$\times 16$. Each group shows the target image with a marked crop, the interpolated input, and fixed-band reconstruction-error maps for the compared methods. The methods use the same test example, crop, spectral band, and color scale within each group, following the main-text evaluation protocol.

On REAL $\times 4$, the marked region contains repeated square patches and their boundaries. The ROTSA error map shows weaker boundary-shaped responses than the comparison methods, indicating more accurate reconstruction of this repeated structure. On ICVL $\times 8$, the palm canopy contains narrow leaves and a thin trunk; ROTSA reduces concentrated errors around these details. At $\times 16$, the marked region contains rock boundaries and surrounding texture. ROTSA shows lower error around several of these structures, complementing its quantitative advantage at the larger upsampling factor.

The REAL $\times 4$ and ICVL $\times 8$ error maps use a 0–0.1 color range, and ICVL $\times 16$ uses 0–0.2. Within each group, weaker responses indicate smaller errors at the displayed band. These spatial comparisons complement the full-HSI PSNR, SSIM, and SAM results in Tables 1 and 2.

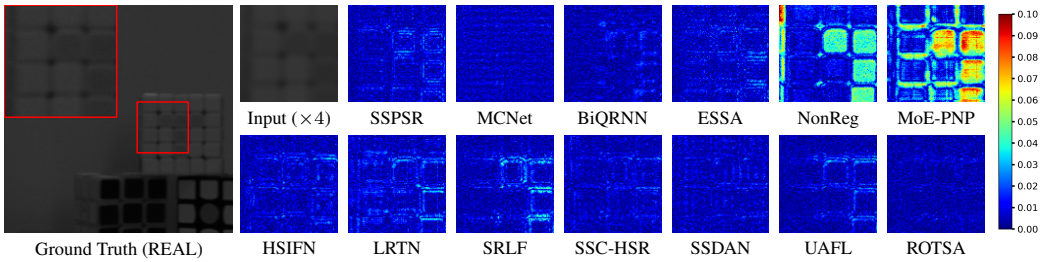

Figure 8: Additional qualitative comparison on REAL $\times 4$. The target image marks the evaluated crop; the other panels show the input and fixed-band reconstruction errors. All methods share the same crop, band, and 0–0.1 error scale.

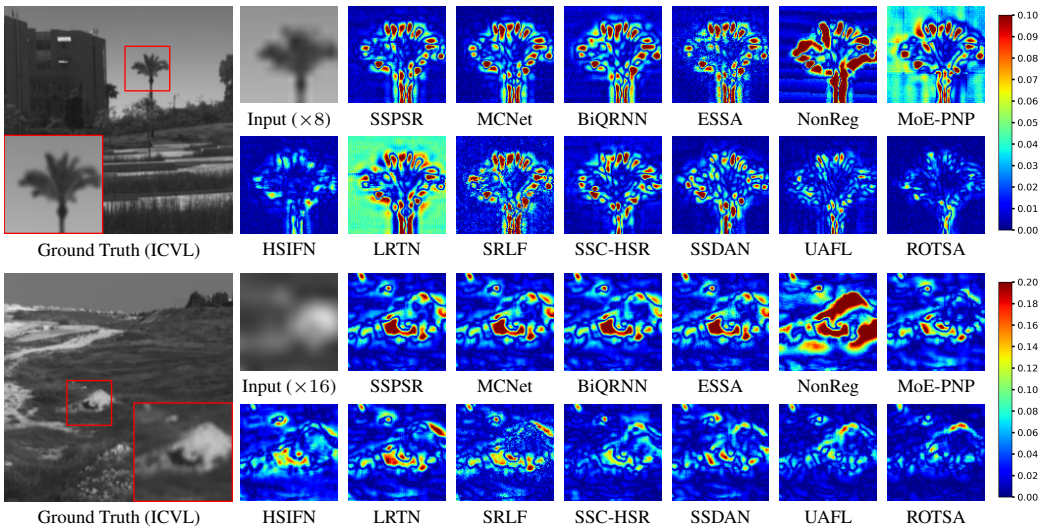

Figure 9: Additional qualitative comparisons on ICVL at $\times 8$ (top) and $\times 16$ (bottom). The error scales are 0–0.1 and 0–0.2, respectively, and are shared by all methods within each group. ROTSA reduces concentrated errors around fine structures in the marked crops.