CORRECTING SPECTRA OUTSIDE THE BACKBONE: A MODEL-AGNOSTIC RECTIFIER FOR HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral image super-resolution (HSI-SR) aims to recover spatial detail while preserving the spectral shape on which quantitative analysis relies. Recent HSI-SR methods, from repurposed RGB super-resolution backbones to dedicated spectral-spatial architectures, have greatly improved spatial reconstruction. However, overlooking the compact spectral structure of hyperspectral data leaves residual spectral errors, while binding the spectral treatment to each architecture forces it to be rebuilt for every new backbone. Yet the low-dimensional structure of spectra belongs to the data, not to any backbone. Backbones differ in the errors they leave, but not in the structure of the true spectra. One rectifier design can therefore serve any backbone. Building on this insight, we propose the **S**pectral **R**ectification **S**uper-**R**esolution Network (***SR*^{2}*-Net***), a model-agnostic rectifier that needs nothing from the backbone but its output, leaves its internal architecture untouched, and is trained per backbone. *SR*^{2}*-Net* follows an *enhance-then-rectify* pipeline in which Hierarchical Spectral-Spatial Synergy Attention (**H-S**^{3}**A**) reinforces cross-band interactions, while Mode-Constrained Rectification (**MCR**) confines the correction to a learned compact spectral subspace. A degradation-consistency constraint further ties the output to the observed low-resolution input. Experiments with five backbones spanning CNN, Transformer, and diffusion families show that one fixed configuration improves spectral fidelity in every reported setting while preserving or improving spatial quality. Averaged over thirty in-domain settings, *SR*^{2}*-Net* removes 22.6% of the residual spectral error at a backbone-independent cost of 0.048M parameters.

1 INTRODUCTION

A hyperspectral image (HSI) records tens to hundreds of contiguous spectral bands at each pixel, and this dense spectrum is what makes it useful from remote sensing to biomedical imaging (Aburaed et al., 2023; Gao et al., 2025; Wang et al., 2021; Yang et al., 2025). Sensors trade spatial resolution for spectral resolution, so a hyperspectral cube is usually spectrally rich but spatially coarse, and Hyperspectral Image Super-resolution (HSI-SR) fills in the missing spatial detail. Unlike in three-channel natural-image SR (Liang et al., 2021; Chen et al., 2023b), the recovered bands must stay radiometrically faithful, since once the spectral shape or its absorption features drift, the quantitative analysis that motivates hyperspectral imaging no longer holds (Kotwal et al., 2025).

Current HSI-SR methods follow two directions. One repurposes mature RGB-SR backbones such as RCAN (Zhang et al., 2018b) and SwinIR (Liang et al., 2021), which recover spatial texture well but treat the bands as ordinary channels. The other builds HSI-specific architectures such as SSPSR (Jiang et al., 2020) and ESSAformer (Zhang et al., 2023) with dedicated spectral-spatial interaction mechanisms (Wu et al., 2024; Zhang et al., 2025a). Although both recover spatial detail well, the spectra they reconstruct still carry residual error, drifting from the ground-truth shape even where the images in Figure 1 look spatially plausible (He et al., 2024). The root lies in a premise the two directions share, namely that spectral fidelity is a capability every architecture has to build in for itself. That premise leads to two challenges.

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.

1) Overlooking the compact spectral structure of hyperspectral data leaves the reconstructed spectral shape unconstrained. Neighboring bands sample a continuous reflectance response, so a scene's spectra span only a few principal directions. Few methods turn this regularity into a constraint. RGB-derived backbones are tuned for spatial texture and impose nothing on the multi-band axis, whereas most HSI-specific models, despite their dedicated interaction mechanisms, still learn cross-band relations implicitly, so nothing in the reconstruction states what a valid spectrum looks like. The spectral axis is then corrected only as a by-product of minimizing a pixel-wise loss, and band-to-band error survives. ***2) Binding the spectral treatment to a specific architecture forces it to be rebuilt for every new backbone.*** Whatever spectral treatment an architecture does have, including the explicit low-dimensional priors of EigenSR (Su et al., 2025) and abundance-domain reconstruction (Hu et al., 2024), is embedded in its own layers or solving domain, so the mechanism built for one cannot be reused by the next. Adopting a stronger spatial backbone means solving spectral fidelity again from scratch, and the effort never accumulates across the field.

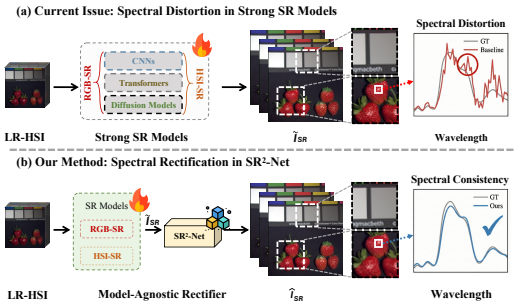

Figure 1: Motivation. (a) Strong SR models, general-purpose and HSI-specific alike, introduce spectral distortion during reconstruction; (b) the model-agnostic *SR*^{2}*-Net* rectifier restores spectral consistency.

Both challenges follow from the same premise, and that premise need not hold. The low-dimensional structure of spectra (Chen et al., 2024) belongs to the data, not to any backbone. Backbones differ in the spectral errors they leave behind, but not in the structure of the true spectra. The missing constraint therefore belongs in a module of its own, stated explicitly rather than learned implicitly and placed outside rather than rebuilt inside each backbone, so that one rectifier design serves any backbone.

In this paper, we propose the **S**pectral **R**ectification **S**uper-**R**esolution Network (*SR*^{2}*-Net*), a model-agnostic rectifier that needs nothing from the backbone but its output: it leaves the internal architecture untouched and is trained per backbone, even on a frozen one. The second challenge is answered by construction, since the same fixed design and configuration attach to CNN, Transformer, and diffusion backbones alike with no redesign. The first challenge is answered by a two-stage *enhance-then-rectify* strategy, whose stages supply the two sides of the missing constraint. Hierarchical Spectral-Spatial Synergy Attention (H-S^{3}A) first makes cross-band interaction explicit, strengthening the dependencies that the backbone learns only weakly. Mode-Constrained Rectification (MCR) then confines the correction to a few learned spectral modes, so that it reshapes a spectrum along those directions rather than band by band. A degradation-consistency constraint keeps that freedom anchored, requiring the reconstructed high-resolution (HR) image, once degraded, to reproduce the low-resolution (LR) input. Our main contributions are as follows:

- We trace the residual spectral error of HSI-SR to the premise that spectral fidelity must be built into each architecture, and recast spectral correction as a backbone-independent problem grounded in the low-dimensional spectral structure of the data.
- We realize this as *SR*^{2}*-Net*, an *enhance-then-rectify* rectifier in which H-S^{3}A makes cross-band interaction explicit, a deliberately simple MCR bottleneck confines the correction to a learned compact spectral subspace, and an auxiliary degradation-consistency constraint keeps the output faithful to the LR measurement, all without modifying the backbone's internal architecture.
- We validate *SR*^{2}*-Net* on five CNN, Transformer, and diffusion backbones, where one fixed configuration improves spectral fidelity in every reported setting and removes 22.6% of the in-domain residual spectral error on average for 0.048M extra parameters. With the backbone frozen, it still recovers 76% of the jointly trained gain, and controlled comparisons show that the gain is not a capacity effect, since widening the MCR bottleneck to the band count gives part of it back.

2 Related Work

2.1 Image and Hyperspectral Super-Resolution

Deep SR architectures have steadily improved spatial detail recovery, and strong backbones are now available across families, including CNN-based RCAN (Zhang et al., 2018b), Transformer-based SwinIR (Liang et al., 2021) and HAT (Chen et al., 2023b), the state-space MambaIR (Guo et al., 2024), and diffusion-based ResShift (Yue et al., 2023) and DiffBIR (Lin et al., 2024). Hyperspectral image super-resolution (HSI-SR) inherits these designs but must also preserve band-to-band continuity, a requirement it shares with recovering a full spectral cube from a single RGB image (Mei et al., 2022; Wang et al., 2024; 2025). A first group treats spectral bands as channels and trains general-purpose backbones on hyperspectral cubes, inheriting strong spatial reconstruction but constraining the spectral axis only implicitly. A second, larger group designs spectral-spatial coupled architectures, exemplified by SSPSR (Jiang et al., 2020) and ESSAformer (Zhang et al., 2023) and by remote sensing work that jointly reconstructs spatial and spectral content (Zhang et al., 2021; Li & Gu, 2022; Zhu et al., 2022; Chen et al., 2023a; Zhang et al., 2025b; Dong et al., 2021), while self-supervised and unsupervised schemes (Hong et al., 2023; Wu et al., 2024) and diffusion priors (Miao et al., 2023; Pang et al., 2024) reduce reliance on paired data. Across these directions the spectral treatment remains architecture-specific, and residual spectral errors persist.

2.2 Spectral Fidelity and Model-Agnostic Refinement

A complementary direction targets spectral fidelity directly rather than through architecture alone. One group enforces spectral-consistency or degradation-consistency objectives so that reconstructions stay compatible with the observed LR input (Wang et al., 2022a; Dong et al., 2021). A second group makes that structure explicit inside the reconstruction, solving in the abundance domain (Hu et al., 2024), bridging a pre-trained RGB backbone through eigenimages (Su et al., 2025), or learning a low-rank prior (Wang et al., 2022b), each tied to one architecture. Classical post-hoc corrections such as iterative back-projection are hand-crafted, and a learned refiner such as the two-stage DualSR (Li et al., 2022) sharpens spectra with a spectral-angle-constrained back-projection co-designed with its own coarse network. Model-agnostic spectral correction has also been studied for RGB-to-spectral reconstruction, such as the analytic projection of Lin & Finlayson (2020), which needs no training but does no spatial reasoning. In contrast to these co-designed or fixed corrections, *SR*^{2}*-Net* is one learned rectifier design that attaches unchanged to CNN-, Transformer-, and diffusion-based backbones for spatial HSI-SR.

3 Method

3.1 Overview

SR^2*-Net* places the spectral correction after the backbone rather than inside it, as shown in Fig. 2. Following an enhance-then-rectify design, an existing SR backbone first recovers a high-resolution HSI from the low-resolution input, and the rectifier then corrects that reconstruction along spectral correlations. It operates at the output resolution and uses local spatial context to guide the correction rather than performing a second super-resolution.

Notation and pipeline Given a low-resolution input $I_{\mathrm{LR}} \in \mathbb{R}^{h\times w\times S}$ with S ordered spectral bands, an SR backbone $f_{\mathrm{SR}}(\cdot)$ first produces an initial super-resolved reconstruction $\tilde{I}_{\mathrm{SR}} = f_{\mathrm{SR}}(I_{\mathrm{LR}}) \in \mathbb{R}^{H\times W\times S}$. For an upscaling factor α, its spatial dimensions satisfy $H = \alpha h$ and $W = \alpha w$.

The rectifier refines $\tilde{I}_{\mathrm{SR}}$ in two stages. First, a Hierarchical Spectral-Spatial Synergy Attention (H-S^{3}A) module enhances cross-band interactions via spectral grouping, trilateral synergy attention (TSA), and adjacent spectra shuffle. Next, a Mode-Constrained Rectification (MCR) module confines the correction to a learned compact spectral subspace, suppressing residual spectral errors. Finally, a reconstruction loss and an auxiliary degradation-consistency constraint jointly train the network, tying the output back to the observed LR input.

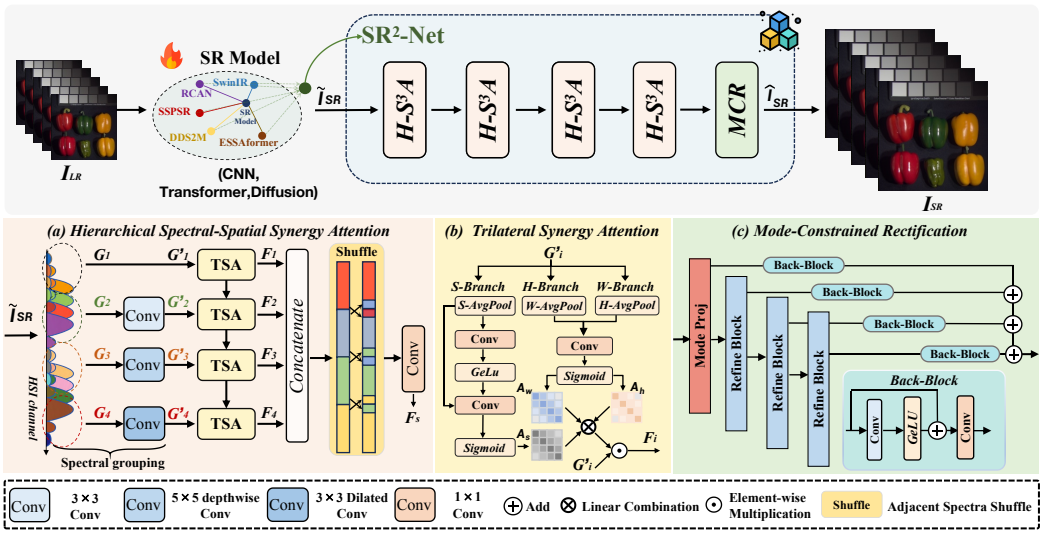

Figure 2: Overview of *SR²-Net*. Given the LR input I_{LR}, an SR backbone produces an initial reconstruction $\tilde{I}_{\mathrm{SR}}$, which *SR²-Net* rectifies into the output $\hat{I}_{\mathrm{SR}}$. (a) H-S³A enhances cross-band interaction via spectral grouping and adjacent shuffle; (b) TSA fuses spectral, height, and width views to recalibrate features; (c) MCR, drawn with N stages (N=1 in our configuration), projects features onto a compact spectral subspace and aggregates the back-projected corrections.

3.2 H-S³A: Cross-Band Interaction Enhancement

The first side of the missing constraint is that backbones learn cross-band relations only implicitly. H-S³A models them explicitly while keeping the features spatially resolved. As shown in Fig. 2(a), H-S³A consists of three components: spectral grouping, trilateral synergy attention, and adjacent spectra shuffle.

Spectral grouping Since the band count S need not be a multiple of four, we first zero-pad the spectra of $\tilde{I}_{\mathrm{SR}} \in \mathbb{R}^{H\times W\times S}$ to $S' = 4\lceil S/4\rceil$ bands; for the 31-band benchmarks this gives $S' = 32$. We then evenly split the padded spectra into four contiguous groups $\{G_1, G_2, G_3, G_4\}$ along the band dimension, with $G_i \in \mathbb{R}^{H\times W\times S'/4}$. The padded band sits at the tail of the last group and is discarded after the final back-projection, so the rectifier preserves the original number of bands. The four groups use complementary per-group stems: G_1 keeps an identity path, while G_2, G_3, and G_4 are processed by a 3×3 convolution, a 5×5 depthwise convolution, and a 3×3 dilated convolution, respectively. This diversifies the effective receptive field across groups at low cost, producing $\{G'_1, G'_2, G'_3, G'_4\}$ that are fed into TSA.

Trilateral Synergy Attention We apply TSA to each group feature $G'_i \in \mathbb{R}^{H\times W\times S'/4}$ to jointly capture spectral and spatial dependencies. TSA first summarizes G'_i into three complementary descriptors by average pooling along one axis at a time:

$$A_i^h = \sigma_g(\mathrm{AvgPool}_w\, G'_i), \quad A_i^w = \sigma_g(\mathrm{AvgPool}_h\, G'_i),$$
$$A_i^s = \sigma_s(\mathrm{AvgPool}_s\, G'_i), \tag{1}$$

where the subscript of AvgPool is the axis averaged over. Pooling over width gives the height descriptor $A_i^h \in \mathbb{R}^{H\times 1\times S'/4}$ and pooling over height gives $A_i^w \in \mathbb{R}^{1\times W\times S'/4}$; both retain per-band coordinate context. Pooling over the spectral axis gives $A_i^s \in \mathbb{R}^{H\times W\times 1}$, which aggregates the bands into a single spatial gate. The spectral branch $\sigma_s(\cdot)$ applies a convolution, a GeLU activation, a second convolution, and a sigmoid, while the height and width descriptors share a lighter gate $\sigma_g(\cdot)$ of a single convolution followed by a sigmoid, so every descriptor lies in $[0, 1]$. We then broadcast the three descriptors to $\mathbb{R}^{H\times W\times S'/4}$ and fuse them to recalibrate the group feature:

$$F_i = G'_i \odot \left(\lambda_h A_i^h + \lambda_w A_i^w + \lambda_s A_i^s\right), \tag{2}$$

where $\lambda_h, \lambda_w, \lambda_s$ are learnable scalars that weight the three broadcast descriptors and $\odot$ is element-wise multiplication. Applying TSA to the four groups yields $\{F_1, F_2, F_3, F_4\}$.

Adjacent Spectra Shuffle Group-wise processing keeps the cost low but leaves the groups isolated. An adjacent spectra shuffle, inspired by channel shuffle (Zhang et al., 2018a), restores exchange between them. Specifically, we concatenate the TSA-enhanced groups along the spectral dimension, $F = [F_1, F_2, F_3, F_4] \in \mathbb{R}^{H\times W\times S'}$, where $[\cdot]$ denotes concatenation. The shuffle swaps the first channel of each group with the last channel of the preceding group, a fixed permutation designed to couple adjacent groups while preserving spectral locality. The zero-padded channel sits at the tail of the last group and is excluded from this boundary swap. A 1×1 convolution then mixes the shuffled bands, completing one H-S^{3}A block. Four such blocks are applied in sequence, each wrapped in a residual connection, and the output $F_s \in \mathbb{R}^{H\times W\times S'}$ of the last one is fed into the MCR for further rectification.

3.3 MCR: Mode-Constrained Rectification

Hyperspectral spectra concentrate near a low-dimensional subspace (Chen et al., 2024), spanned by the few principal directions that PCA exploits to compress an HSI. Mode-Constrained Rectification (MCR) supplies the other side of the missing constraint by constraining the correction itself rather than the loss. It models the correction the way hyperspectral data are modeled, as $r = 8$ learned spectral modes shared by all pixels, each pixel with its own coefficients (Sec. 4.5). Every stage shares one bias-free back-projection $g_m(\cdot)$ whose r output directions are these modes, so the aggregated correction never leaves their span and reshapes a spectrum along them rather than band by band.

As illustrated in Fig. 2(c), MCR first applies a Mode Projection $P(\cdot)$ that maps the input feature $F_s \in \mathbb{R}^{H\times W\times S'}$ to a reduced dimension r:

$$m_0 = P(F_s) \in \mathbb{R}^{H\times W\times r}. \tag{3}$$

MCR then applies N refinement stages (N=1 in our configuration), where the i-th stage updates the embedding through a Refine Block $f_m(\cdot)$:

$$m_i = f_m(m_{i-1}), \quad i = 1, 2, \ldots, N. \tag{4}$$

For each m_i, a Back-Block $g_m(\cdot)$ maps the embedding back to the padded S'-band space; summed over all stages, these terms form a correction that is added to the backbone's reconstruction once the zero-padded band is discarded:

$$\hat{I}_{\mathrm{SR}} = \tilde{I}_{\mathrm{SR}} + \Pi_S\Big(\sum_{i=0}^{N} g_m(m_i)\Big), \tag{5}$$

where $\Pi_S(\cdot)$ discards the zero-padded band. The $i = 0$ term back-projects the un-refined embedding, and each refinement stage adds a further term to the sum.

MCR is deliberately simple: three small convolution blocks. The Mode Projection $P(\cdot)$ is a 1×1 convolution that reduces the channel dimension from S' to r, followed by a GeLU activation, and the Refine Block $f_m(\cdot)$ shares this structure at dimension r. The Back-Block $g_m(\cdot)$ applies a 3×3 convolution with a GeLU activation and a residual connection at rank r, followed by a 1×1 convolution that maps r back to S'.

3.4 Degradation-Consistency Loss

The degradation-consistency loss anchors the rectified output to the measurement, acting as a soft data-fidelity term for HSI-SR. The predicted HR image $\hat{I}_{\mathrm{SR}}$ is degraded back to the LR space by bicubic downsampling and compared with the input I_{LR}:

$$\mathcal{L}_{\mathrm{deg}} = \frac{1}{hwS}\left\|\mathcal{D}\left(\hat{I}_{\mathrm{SR}}\right) - I_{\mathrm{LR}}\right\|_1, \tag{6}$$

where $\mathcal{D}(\cdot)$ denotes the degradation model used to generate the LR input. We retain the standard mean ℓ_1 reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ of the SR model, giving the total loss $\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{deg}}\mathcal{L}_{\mathrm{deg}}$, where λ_{rec} and λ_{deg} balance the two terms.

Table 1: In-domain results on ARAD-1K and CAVE at $\times 2$, $\times 4$, and $\times 8$, with training and test sets drawn from disjoint splits of each dataset. Every backbone is evaluated as Base and after attaching *SR*^{2}*-Net* (+Ours, in bold), with one fixed rectifier configuration throughout. mPSNR and mSSIM are spatial, mSAM and CC spectral, and ↑/↓ mark the better direction.

Dataset	Scale	Metric	CNN-based				Transformer-based				Diffusion-based	
			RCAN		SSPSR		SwinIR		ESSAformer		DDS2M	
			Base	**+Ours**	Base	**+Ours**	Base	**+Ours**	Base	**+Ours**	Base	**+Ours**
ARAD-1K	$\times 2$	mPSNR↑	48.7390	**50.2735**	46.0882	**48.5005**	50.4662	**51.6551**	49.4500	**49.9900**	36.5792	**37.2283**
		mSAM↓	1.2071	**0.7849**	0.9906	**0.6658**	0.7727	**0.4842**	0.6090	**0.5080**	7.1986	**4.7788**
		mSSIM↑	0.9936	**0.9964**	0.9893	**0.9920**	0.9972	**0.9975**	0.9937	**0.9942**	0.9089	**0.9277**
		CC↑	0.9978	**0.9990**	0.9960	**0.9981**	0.9990	**0.9993**	0.9986	**0.9987**	0.9723	**0.9799**
	$\times 4$	mPSNR↑	38.3786	**40.0972**	38.8792	**40.1004**	39.5717	**40.9720**	39.9537	**40.7200**	30.8476	**33.2613**
		mSAM↓	1.9009	**1.3010**	1.6076	**1.3128**	1.3950	**1.2819**	1.3670	**1.2300**	7.0404	**5.8757**
		mSSIM↑	0.9470	**0.9719**	0.9463	**0.9543**	0.9709	**0.9752**	0.9592	**0.9599**	0.7984	**0.8671**
		CC↑	0.9812	**0.9916**	0.9858	**0.9891**	0.9902	**0.9924**	0.9903	**0.9904**	0.9421	**0.9549**
	$\times 8$	mPSNR↑	33.3062	**34.3745**	34.1149	**34.9975**	33.4185	**34.8318**	34.6531	**35.0546**	27.7449	**29.5794**
		mSAM↓	3.0518	**2.1900**	2.4223	**2.0352**	2.4682	**2.1164**	2.3531	**2.0140**	11.5589	**9.2067**
		mSSIM↑	0.8891	**0.9029**	0.8724	**0.8850**	0.8869	**0.9103**	0.8882	**0.8889**	0.6060	**0.7993**
		CC↑	0.9628	**0.9693**	0.9622	**0.9681**	0.9551	**0.9720**	0.9691	**0.9698**	0.8861	**0.9019**
CAVE	$\times 2$	mPSNR↑	41.0944	**43.4074**	39.4003	**40.4833**	40.3123	**41.4861**	39.4700	**40.7900**	35.5928	**36.5567**
		mSAM↓	5.7619	**4.4010**	4.2518	**3.7893**	0.8214	**0.6275**	4.0820	**3.3450**	11.0914	**8.1802**
		mSSIM↑	0.9838	**0.9897**	0.9711	**0.9830**	0.9754	**0.9877**	0.9704	**0.9747**	0.9053	**0.9357**
		CC↑	0.9946	**0.9965**	0.9907	**0.9943**	0.9807	**0.9905**	0.9950	**0.9960**	0.9882	**0.9908**
	$\times 4$	mPSNR↑	35.9929	**37.4461**	35.2386	**35.8929**	37.2675	**38.1303**	34.1100	**35.3300**	29.8048	**31.4389**
		mSAM↓	7.6445	**5.4742**	4.8735	**4.4121**	0.9015	**0.7083**	5.6070	**4.5410**	13.7359	**10.0409**
		mSSIM↑	0.9536	**0.9643**	0.9384	**0.9427**	0.9492	**0.9577**	0.9267	**0.9367**	0.7493	**0.8875**
		CC↑	0.9825	**0.9856**	0.9795	**0.9807**	0.9315	**0.9478**	0.9846	**0.9875**	0.9699	**0.9766**
	$\times 8$	mPSNR↑	32.0839	**33.2764**	31.5901	**32.0282**	32.8788	**34.2636**	30.0600	**30.8700**	27.1577	**28.1137**
		mSAM↓	10.1641	**6.8252**	7.6479	**6.1349**	1.5060	**0.8553**	7.8170	**6.0120**	16.6298	**13.7381**
		mSSIM↑	0.8923	**0.9104**	0.8781	**0.8828**	0.9125	**0.9321**	0.8530	**0.8664**	0.6298	**0.7948**
		CC↑	0.9616	**0.9654**	0.9577	**0.9599**	0.9185	**0.9233**	0.9637	**0.9683**	0.9186	**0.9389**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and degradation We evaluate on three widely used 31-band hyperspectral datasets: ARAD-1K (Arad et al., 2022) with the standard $900/50$ train/test split, CAVE (Park et al., 2007) with $16/16$ scenes, and 18 ICVL (Arad & Ben-Shahar, 2016) HSIs reserved for cross-domain evaluation only. LR inputs are synthesized by bicubic downsampling at each scale factor, and robustness to test-time degradation mismatch is examined separately without changing this training setup.

Evaluation metrics We measure spatial reconstruction with mean PSNR (mPSNR) and mean SSIM (mSSIM) (Wang et al., 2004), and spectral fidelity with mean spectral angle mapper (mSAM, in degrees) (Kruse et al., 1993) and the mean spectral correlation coefficient (CC) (Loncan et al., 2015). Most tables report mPSNR, mSSIM, and mSAM, with formal definitions in Appendix A.

Compared methods To verify model-agnostic generality, we attach *SR*^{2}*-Net* to five representative baseline models: CNN-based RCAN (Zhang et al., 2018b) and SSPSR (Jiang et al., 2020), Transformer-based SwinIR (Liang et al., 2021) and ESSAformer (Zhang et al., 2023), and diffusion-based DDS2M (Miao et al., 2023).

For each SR backbone $\mathcal{B}$, we evaluate the baseline $\mathcal{B}$ and the rectified version $\mathcal{B} + \mathrm{SR}^2$-Net under the same settings. *SR*^{2}*-Net* is therefore evaluated as an architecture-preserving refinement. Each of the four feed-forward backbones is trained jointly with the rectifier (frozen-backbone results are in Sec. 4.5). Their Base models are trained with the reconstruction loss only, whereas the degradation-consistency constraint applies to the rectified models. The diffusion-based DDS2M

Table 2: Cross-domain results, with every backbone trained on ARAD-1K and tested on CAVE and ICVL without fine-tuning. Each backbone is evaluated as Base and after attaching *SR*^{2}*-Net* (+Ours, in bold) under the identical protocol.

Method	Variant	×2				×4			
		CAVE		ICVL		CAVE		ICVL	
		mPSNR↑	mSAM↓	mPSNR↑	mSAM↓	mPSNR↑	mSAM↓	mPSNR↑	mSAM↓
RCAN	Base	42.7269	5.0004	49.6448	0.7389	35.6214	6.2153	39.9389	1.0969
	+Ours	**44.2866**	**4.0978**	**51.8506**	**0.4682**	**36.3326**	**5.3594**	**40.9148**	**0.7897**
SSPSR	Base	41.4071	4.0419	48.4758	0.8140	36.7963	4.9284	43.9812	1.0271
	+Ours	**43.0076**	**2.9243**	**50.7217**	**0.5264**	**37.5721**	**3.9189**	**45.1945**	**0.6698**
SwinIR	Base	40.7157	4.4366	47.9367	0.8260	34.7725	4.8551	41.0743	1.0552
	+Ours	**42.3415**	**2.6231**	**50.0704**	**0.4709**	**35.7256**	**3.8360**	**42.6530**	**0.6682**
ESSAformer	Base	42.1763	3.3575	49.7895	0.6404	37.0305	4.8258	44.3869	0.9349
	+Ours	**43.4978**	**2.5611**	**51.3981**	**0.4625**	**37.6350**	**3.5556**	**45.7012**	**0.6079**

instead reconstructs each scene by self-supervised per-image inference; the rectifier is trained on its outputs for the training scenes under the same losses and evaluated on the held-out test scenes.

Implementation details All networks are implemented in PyTorch on an NVIDIA RTX 3090 GPU and trained using AdamW, with the learning rate decayed from 1×10^{-4} to 1×10^{-5} by cosine annealing over 1000 epochs. We optimize the mean ℓ_1 reconstruction loss together with the degradation-consistency constraint, using $\lambda_{\text{rec}} = 1.0$ and $\lambda_{\text{deg}} = 0.1$, and *SR*^{2}*-Net* uses 4 spectral groups, 4 H-S^{3}A blocks, and 1 MCR block with bottleneck rank $r = 8$. Full hyperparameters are listed in Appendix B.

4.2 Quantitative Results and Analysis

In-domain performance Training and testing within each of ARAD-1K and CAVE, Tab. 1 covers $\times 2$, $\times 4$, and $\times 8$. Spectral fidelity improves in all thirty configurations, with mSAM falling by 1.1203 on average and by at least 0.1010. The gain is not bought with spatial quality: mPSNR rises by between 0.4015 and 2.4137 dB, and mSSIM and CC never decrease.

How much the rectifier corrects follows how much spectral error the backbone leaves. The largest gains occur for the diffusion-based DDS2M, whose mSAM drops by 2.4198 on ARAD-1K at $\times 2$ and by 3.6950 on CAVE at $\times 4$, whereas the spectrally strong ESSAformer leaves the least to correct (0.1010 on ARAD-1K at $\times 2$). Over the thirty configurations the reduction tracks the Base mSAM with a correlation of 0.93, and one fixed configuration removes 22.6% of that error on average, with per-backbone means between 16.9% and 29.9%. The pattern persists as the task gets harder, since the reduction at $\times 8$ exceeds the one at $\times 2$ for all eight feed-forward combinations of backbone and dataset.

Cross-domain generalization Trained on the 900 ARAD-1K training images and tested on all CAVE scenes and all ICVL HSIs without fine-tuning, Tab. 2 covers $\times 2$ and $\times 4$. Models generalize better under domain shift after attaching *SR*^{2}*-Net*, with mSAM reductions in every reported configuration while preserving spatial metrics. For instance, SwinIR on CAVE at $\times 2$ improves mPSNR from 40.7157 to 42.3415 and reduces mSAM from 4.4366 to 2.6231. DDS2M is reported in the in-domain setting only, as its diffusion inference is far costlier than the other backbones. Full results including mSSIM are in Appendix E.

Overhead and robustness *SR*^{2}*-Net* adds a backbone-independent overhead of $+0.048$M parameters in the 31-band setting, and its FLOPs overhead depends only on the output spatial size, amounting to $+7.7\%$ at $\times 4$ and $+22.9\%$ at $\times 8$ for the lightest backbone SwinIR and proportionally less for heavier ones. It also remains effective under mild test-time mismatch. Keeping SwinIR trained under standard bicubic degradation, mSAM improves over Base in all nine test-time settings and scales (with the B1 blur at $\times 4$ it drops from 1.4760 to 1.2709), with full results in Appendix D.

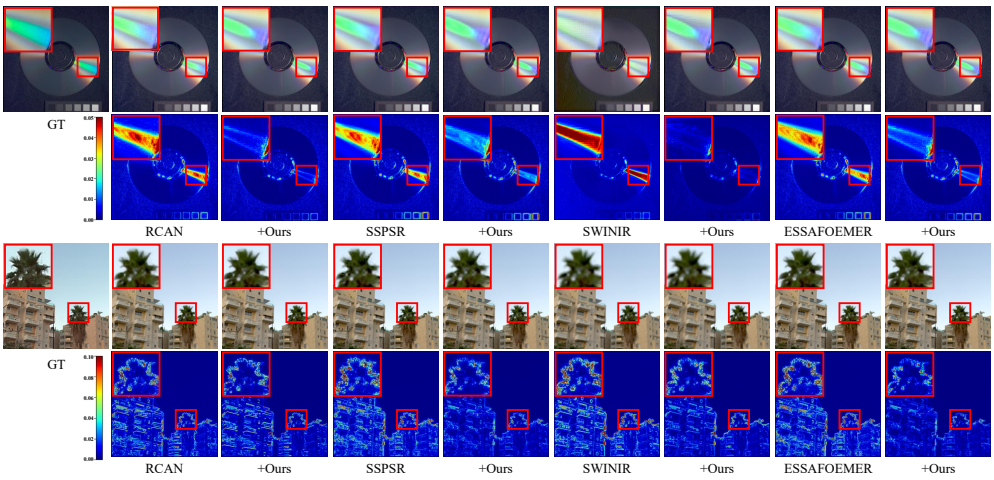

Figure 3: Error maps at $\times 4$ (in-domain, CAVE and ARAD-1K). Pseudo-RGB uses bands 25/15/5 as R–G–B, and band-averaged absolute error maps (darker is better) pair each +Ours result with its Base. Red boxes mark where *SR*^{2}*-Net* sharpens edges and fine structures; the cross-domain ICVL comparison and more galleries are in Appendix I.

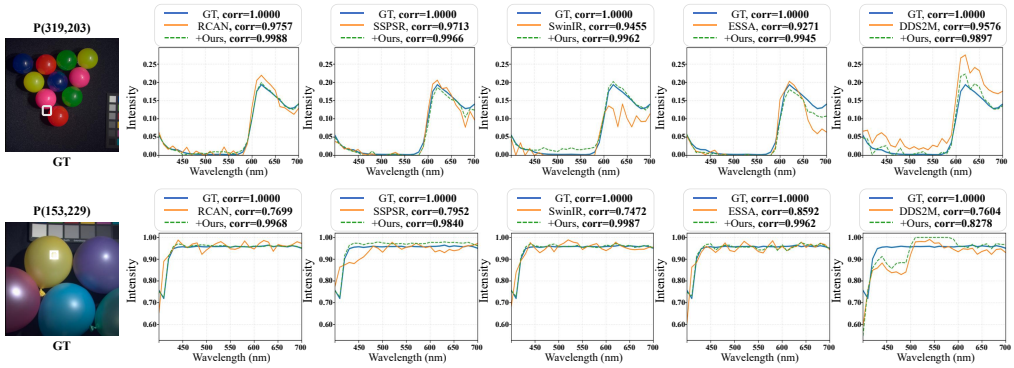

Figure 4: Per-pixel spectra at $\times 4$ on CAVE across CNN-, Transformer-, and diffusion-based backbones. For two representative pixels (marked in the pseudo-RGB images), the baseline spectra oscillate away from the ground truth while *SR*^{2}*-Net* follows it; additional scenes and datasets are in Appendix J.

4.3 Visualization

In Fig. 3, baseline SR models produce localized errors near object boundaries and thin structures, visible as halos. After attaching *SR*^{2}*-Net* these errors are visibly reduced while fine textures and sharp edges are preserved; per-scene galleries are in Appendix I.

Fig. 4 visualizes per-pixel spectra under $\times 4$ upsampling at CAVE locations across CNN-, Transformer-, and diffusion-based backbones. The baseline spectra oscillate away from the ground-truth trend, while *SR*^{2}*-Net* yields smoother profiles that follow it on these pixels. Beyond these examples, the effect is dataset-wide. On the 50 ARAD-1K test scenes *SR*^{2}*-Net* lowers the residual spectral error on all 50 and improves per-scene mSAM on all 50 (one-sided Wilcoxon signed-rank, $p < 10^{-6}$), and the residual analysis is in Appendix H.

4.4 Ablation Study

Component ablation All ablations use ARAD-1K at $\times 4$ with SwinIR. Tab. 3 evaluates the two-stage design of *SR*^{2}*-Net*. Each stage moves the axis it was designed for. In isolation, H-S^{3}A contributes most of the spatial gain (mPSNR 40.7059 against 39.5717) but little of the spectral one,

Table 3: Component ablation on ARAD-1K at $\times 4$ with SwinIR. Best in **bold**, second underlined.

Method		Metric		
H-S^{3}A	MCR	mPSNR↑	mSSIM↑	mSAM↓
×	×	39.5717	0.9709	1.3950
✓	×	40.7059	0.9735	1.3476
×	✓	40.2550	0.9717	1.3173
✓	✓	**40.9720**	**0.9752**	**1.2819**

Table 4: Loss-weight ablation on ARAD-1K at $\times 4$ with SwinIR. Best in **bold**, second underlined.

λ_{rec}	λ_{deg}	mPSNR↑	mSSIM↑	mSAM↓
1.0	0.0	40.6920	0.9740	1.3262
1.0	0.1	**40.9720**	**0.9752**	**1.2819**
1.0	0.2	40.9225	0.9751	1.2833
1.0	0.5	40.7184	0.9745	1.3174
1.0	1.0	40.6217	0.9739	1.3305

Table 5: Comparison to simpler rectifiers (ARAD-1K, $\times 4$, SwinIR). SG/PCA/IBP are fixed post-hoc rectifiers; "$+\mathcal{L}_{\text{deg}}$ only" adds the loss to the Base with no rectifier, and "Generic head" is a same-budget ($+0.048$M) residual head. +MCR and *SR*^{2}*-Net* are the jointly trained MCR-only and full models. Best in **bold**, second underlined.

Method	mPSNR↑	mSAM↓	mSSIM↑	CC↑
Base (SwinIR)	39.5717	1.3950	0.9709	0.9902
+ SG	38.7834	2.4512	0.9698	0.9889
+ PCA	39.5729	1.4028	0.9709	0.9902
+ IBP	39.5782	1.3693	0.9709	0.9902
+ $\mathcal{L}_{\text{deg}}$ only	39.7097	1.3692	0.9711	0.9905
+ Generic head	39.9815	1.3648	0.9712	0.9907
+ MCR	40.2550	1.3173	0.9717	0.9912
+ *SR*^{2}*-Net*	**40.9720**	**1.2819**	**0.9752**	**0.9924**

whereas MCR gives the larger spectral gain (mSAM 1.3173 against 1.3950) at roughly half the spatial improvement. The two are complementary rather than additive, and the full model is the best on every metric, reaching 40.9720 mPSNR and 1.2819 mSAM. Removing only the adjacent spectra shuffle, with the 1×1 mixing kept, raises mSAM to 1.3102 (Appendix G).

Degradation-consistency loss Tab. 4 varies λ_{deg}. Removing the constraint ($\lambda_{\text{deg}} = 0$) raises mSAM from 1.2819 to 1.3262, still below the Base (1.3950). We select $\lambda_{\text{deg}} = 0.1$ once on this configuration and fix it everywhere else; $\lambda_{\text{deg}} = 0.2$ is within one seed standard deviation of it.

4.5 Isolating the Source of the Improvement

We ask whether the gain comes from a cheaper alternative (Tab. 5). Among the fixed post-hoc rectifiers, Savitzky–Golay smoothing (SG) degrades every metric (mSAM 2.4512), while iterative back-projection (IBP) and a PCA projection of the same rank as MCR barely change mSAM (1.3693 and 1.4028). The degradation-consistency loss alone (1.3692) and a same-budget generic residual head (1.3648) stay far above MCR (1.3173) and the full *SR*^{2}*-Net* (1.2819); the generic head shares the full model's $+0.048$M budget, whereas MCR alone uses less. Nor is the gain a capacity effect: narrowing the bottleneck from $r = 31$ to $r = 8$ lowers mSAM from 1.331 to 1.2819 while removing parameters, and $r = 1$ is worse again (1.349), so the rank acts as a two-sided constraint. The rectifier also works with the four backbones frozen at their Base weights: mSAM falls in all sixteen ARAD-1K and CAVE configurations at $\times 2$ and $\times 4$, recovering 76% of the jointly trained gain on average (Tab. A11). The ranking from Base through the cheap controls to MCR and the full model is stable across three seeds, and the full model also leads the fixed rectifiers and MCR at $\times 2$ and $\times 8$; implementation, seed stability, and the structural sweeps are in Appendices F and G.

5 Conclusion

We presented *SR*^{2}*-Net*, a model-agnostic spectral rectifier for hyperspectral image super-resolution. Following an enhance-then-rectify design, it attaches to an SR backbone's output without modifying its internal architecture: H-S^{3}A strengthens cross-band interactions, MCR confines the correction to a compact spectral subspace, and a degradation-consistency constraint ties the output to the observed LR input. Across five backbones spanning CNN, Transformer, and diffusion families, it improves spectral fidelity in every reported setting while preserving or improving spatial quality, in domain, under cross-domain transfer (four of five backbones), and under mild test-time mismatch, at a constant $+0.048$M overhead.

AI USE STATEMENT

We have not used generative AI tools for any task that requires disclosure under the ICLR 2027 AI Policy for Authors: no such tool was used to develop the method or its conceptual framework, to formulate hypotheses or claims, to design the experiments, to implement the method, to generate or process data, to translate, or to interpret the results. We used generative AI tools only to edit and rephrase the text of this paper for readability. We have reviewed all AI-assisted edits, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only the public ARAD-1K, CAVE, and ICVL hyperspectral benchmarks under their original terms of use and involves no new data collection or human subjects. We are not aware of ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

All experiments use the public ARAD-1K, CAVE, and ICVL datasets with the splits and bicubic degradation described in Section 4.1. Appendix A defines every metric and the border cropping applied before evaluation. Appendix B lists the training hyperparameters and the full rectifier configuration (Tables A1 and A2), shared across backbones and scales except for the scale-dependent patch size. Appendix C specifies the test-time degradation protocol, including the per-image seeding of the added noise, and Appendix F details the fixed and learned control rectifiers and reports seed stability over three seeds. The code will be released upon publication.

REFERENCES

Nour Aburaed, Mohammed Q Alkhatib, Stephen Marshall, Jaime Zabalza, and Hussain Al Ahmad. A review of spatial enhancement of hyperspectral remote sensing imaging techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2275–2300, 2023.

Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural RGB images. In *European conference on computer vision*, pp. 19–34, 2016.

Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Luc Van Gool, Shuai Liu, Yongqiang Li, Chaoyu Feng, Lei Lei, Jiaojiao Li, Songcheng Du, Chaoxiong Wu, Yihong Leng, Rui Song, Mingwei Zhang, Chongxing Song, Shuyi Zhao, Zhiqiang Lang, Wei Wei, Lei Zhang, Renwei Dian, Tianci Shan, Anjing Guo, Chengguo Feng, Jinyang Liu, Mirko Agarla, Simone Bianco, Marco Buzzelli, Luigi Celona, Raimondo Schettini, Jiang He, Yi Xiao, Jiajun Xiao, Qiangqiang Yuan, Jie Li, Liangpei Zhang, Taesung Kwon, Dohoon Ryu, Hyokyoung Bae, Hao-Hsiang Yang, Hua-En Chang, Zhi-Kai Huang, Wei-Ting Chen, Sy-Yen Kuo, Junyu Chen, Haiwei Li, Song Liu, Sabarinathan, K Uma, B Sathya Bama, and S. Mohamed Mansoor Roomi. NTIRE 2022 spectral recovery challenge and data set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 863–881, June 2022.

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

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377, 2023b.

Yong Chen, Jinshan Zeng, Wei He, Xi-Le Zhao, Tai-Xiang Jiang, and Qing Huang. Fast large-scale hyperspectral image denoising via noniterative low-rank subspace representation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 30: 5754–5768, 2021.

Feng Gao, Xuepeng Jin, Xiaowei Zhou, Junyu Dong, and Qian Du. MSFMamba: Multi-scale feature fusion state space model for multi-source remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025.

Hang Guo, Jinmin Li, Tao Dai, Zhi Ouyang, Xidong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision (ECCV)*, pp. 222–241, 2024.

Chengxun He, Yang Xu, Zebin Wu, and Zhihui Wei. Connecting low-level and high-level visions: A joint optimization for hyperspectral image super-resolution and target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Danfeng Hong, Jing Yao, Chenyu Li, Deyu Meng, Naoto Yokoya, and Jocelyn Chanussot. Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

Qian Hu, Xinya Wang, Junjun Jiang, Xiao-Ping Zhang, and Jiayi Ma. Exploring the spectral prior for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 33:5260–5272, 2024.

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.

Alankar Kotwal, Vishwanath Saragadam, Joshua D Bernstock, Alfredo Sandoval, Ashok Veeraraghavan, and Pablo A Valdés. Hyperspectral imaging in neurosurgery: a review of systems, computational methods, and clinical applications. *Journal of Biomedical Optics*, 30(2):023512, 2025.

Fred A Kruse, Adam B Lefkoff, Joseph W Boardman, Kathleen B Heidebrecht, AT Shapiro, PJ Barloon, and Alexander FH Goetz. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3): 145–163, 1993.

Qiang Li, Yuan Yuan, Xiuping Jia, and Qi Wang. Dual-stage approach toward hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 31:7252–7263, 2022.

Tianshuai Li and Yanfeng Gu. Progressive spatial-spectral joint network for hyperspectral image reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1833–1844, 2021.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 430–448, 2024.

Yi-Tun Lin and Graham D. Finlayson. Physically plausible spectral reconstruction from RGB images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 532–533, 2020.

Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015.

Shaohui Mei, Yunhao Geng, Junhui Hou, and Qian Du. Learning hyperspectral images from RGB images via a coarse-to-fine CNN. *Science China Information Sciences*, 65(5):152102, 2022.

Yuchun Miao, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. DDS2M: Self-supervised denoising diffusion spatio-spectral model for hyperspectral image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12086–12096, 2023.

Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. HIR-Diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3005–3014, 2024.

Jong-Il Park, Moon-Hyun Lee, Michael D Grossberg, and Shree K Nayar. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

Xi Su, Xiangfei Shen, Mingyang Wan, Jing Nie, Lihui Chen, Haijun Liu, and Xichuan Zhou. EigenSR: Eigenimage-bridged pre-trained RGB learners for single hyperspectral image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7033–7041, 2025.

Chunying Wang, Baohua Liu, Lipeng Liu, Yanjun Zhu, Jialin Hou, Ping Liu, and Xiang Li. A review of deep learning used in the hyperspectral image analysis for agriculture. *Artificial Intelligence Review*, 54(7):5205–5253, 2021.

Nan Wang, Shaohui Mei, Yifan Zhang, Mingyang Ma, and Xiangqing Zhang. Hyperspectral image reconstruction from RGB input through highlighting intrinsic properties. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Nan Wang, Shaohui Mei, Yi Wang, Yifan Zhang, and Duo Zhan. WHANet: Wavelet-based hybrid asymmetric network for spectral super-resolution from RGB inputs. *IEEE Transactions on Multimedia*, 27:414–428, 2025.

Xinya Wang, Jiayi Ma, and Junjun Jiang. Hyperspectral image super-resolution via recurrent feedback embedding and spatial-spectral consistency regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022a.

Xiuheng Wang, Jie Chen, Qi Wei, and Cédric Richard. Hyperspectral image super-resolution via deep prior regularization with parameter estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1708–1723, 2022b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Huajing Wu, Suqin Wu, Kefei Zhang, Xuexi Liu, Shuangshuang Shi, and Chaofa Bian. Unsupervised blind spectral–spatial cross-super-resolution network for HSI and MSI fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Chen Yang, Zhiming Guo, Douglas Fernandes Barbin, Zhiqiang Dai, Nicholas Watson, Megan Povey, and Xiaobo Zou. Hyperspectral imaging and deep learning for quality and safety inspection of fruits and vegetables: A review. *Journal of Agricultural and Food Chemistry*, 73(17):10019–10035, 2025.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 13294–13306, 2023.

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.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018a.

Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. SSR-NET: spatial-spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2021.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018b.

Zhanxu Zhang, Linzi Yang, Guanglian Zhang, Jiangwei Deng, Lifeng Bian, and Chen Yang. CASS-Net: Cross-attention enhanced spectral-spatial interaction network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:11716–11730, 2025a.

Zhen Zhang, Yemao Qi, Qingwang Wang, Bo-Hui Tang, Yabin Hu, Lehao Huang, and Chenxi Liu. Frequency-enhanced spatial-spectral network for hyperspectral imagery reconstruction from multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025b.

Dehui Zhu, Bo Du, Yanni Dong, and Liangpei Zhang. Spatial-spectral joint reconstruction with interband correlation for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

APPENDIX

This appendix provides the formal metric definitions, full configurations and protocols, complete quantitative results, and additional qualitative galleries complementing the main paper. The quantitative sections (Sections A–H) are self-contained and precede the qualitative galleries (Sections I and J).

A EVALUATION METRICS

Let $X \in \mathbb{R}^{H\times W\times S}$ denote the ground-truth HSI and $\hat{X}$ its reconstruction, with S spectral bands. We use two spatial-fidelity metrics (mPSNR and mSSIM) and two spectral-fidelity metrics (mSAM and CC). Spatial metrics are averaged over the S bands, whereas spectral metrics are averaged over pixels and images. Unless noted otherwise, LR inputs are produced by bicubic downsampling of the HR cube, all cubes are scaled to $[0, 1]$, and each metric is computed after cropping a border of s pixels (s the scale factor) to remove boundary artifacts.

Mean PSNR (mPSNR) For each band b, the peak signal-to-noise ratio is

$$\mathrm{PSNR}_b = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\frac{1}{HW}\sum_p \left(X_{b,p} - \hat{X}_{b,p}\right)^2}\right), \tag{A1}$$

where the sum runs over the HW spatial positions p and MAX is the dynamic range of the data. mPSNR averages PSNR_b over all bands; higher is better.

Mean SSIM (mSSIM) For a reference band x and its reconstruction y, the structural similarity is

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{A2}$$

where μ, σ^2, and σ_{xy} are the local means, variances, and covariance, and C_1, C_2 are stabilizing constants. mSSIM averages the per-band SSIM; higher is better.

Mean SAM (mSAM) The spectral angle mapper measures the angle between the reference and reconstructed spectra at each pixel p,

$$\mathrm{SAM}_p = \arccos\left(\frac{\langle x_p, \hat{x}_p\rangle}{\|x_p\|_2\,\|\hat{x}_p\|_2}\right), \tag{A3}$$

where $x_p, \hat{x}_p \in \mathbb{R}^S$ are the spectra at pixel p. mSAM averages SAM_p over all pixels and images and is reported in degrees; lower is better. Because it compares the direction of the spectral vector rather than its magnitude, mSAM is the primary indicator of spectral fidelity in this paper.

Spectral correlation coefficient (CC) CC is the Pearson correlation between the reference and reconstructed spectra at each pixel,

$$\mathrm{CC}_p = \frac{\sum_s (x_{p,s} - \bar{x}_p)(\hat{x}_{p,s} - \bar{\hat{x}}_p)}{\sqrt{\sum_s (x_{p,s} - \bar{x}_p)^2}\,\sqrt{\sum_s (\hat{x}_{p,s} - \bar{\hat{x}}_p)^2}}, \tag{A4}$$

where $\bar{x}_p$ and $\bar{\hat{x}}_p$ are the per-pixel spectral means. CC is averaged over pixels and images; higher is better.

B NETWORK CONFIGURATION AND TRAINING DETAILS

Table A1 summarizes the training hyperparameters, and Table A2 lists the configuration of the two rectification stages. Unless stated otherwise, all settings are shared across backbones and scales; only the patch size depends on the scale factor. These values match those used to produce every result reported in the main paper.

Table A1: Training hyperparameters, shared across all backbones.

Setting	Value
Framework	PyTorch
Hardware	NVIDIA RTX 3090 GPU
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.99$)
Learning rate	$1 \times 10^{-4} \rightarrow 1 \times 10^{-5}$ (cosine)
Epochs	1000
Batch size	8
LR patch ($\times 2/\times 4/\times 8$)	64^2 / 32^2 / 32^2
HR patch ($\times 2/\times 4/\times 8$)	128^2 / 128^2 / 256^2
Reconstruction loss	mean ℓ_1, $\lambda_{\text{rec}} = 1.0$
Degradation loss	mean ℓ_1, $\lambda_{\text{deg}} = 0.1$

Table A2: Configuration of the *SR*^{2}*-Net* rectifier, where S is the number of spectral bands, $S' = 4\lceil S/4 \rceil$ the zero-padded band count, and r the MCR bottleneck rank.

Component	Configuration
Spectral groups	4 contiguous groups of $\lceil S/4 \rceil$ bands (bands zero-padded to a multiple of four; padding removed after back-projection)
H-S^{3}A blocks	4, applied in sequence with independent parameters and a residual connection around each: $F^{(l)} = F^{(l-1)} + \text{Block}_l(F^{(l-1)})$, where $F^{(0)}$ is the zero-padded input and $F^{(4)} = F_s$ enters MCR
▷ per-group stems	identity / 3×3 / 5×5 depthwise / 3×3 dilated (one per group); 1×1 mixing after the shuffle
▷ TSA views	height, width, spectral (h, w, s)
MCR blocks (refinement stages N)	1; ablations that drop MCR remove the module entirely (projection and back-projection included), so the rectified output is $\Pi_S(F_s)$, giving the H-S^{3}A-only variant of the component ablation
▷ projection $P(\cdot)$	1×1 conv, $S' \rightarrow r$, GeLU
▷ refine $f_m(\cdot)$	1×1 conv at rank r, GeLU
▷ back-projection $g_m(\cdot)$	residual 3×3 conv at rank r, then 1×1, $r \rightarrow S'$
Bottleneck rank r	8
Added parameters	$+0.048$ M (independent of the backbone)

C TEST-TIME DEGRADATION PROTOCOL

In the robustness study, every model is trained once under the standard bicubic degradation and then evaluated without retraining, changing only the test-time degradation. Table A3 lists the nine settings, grouped into kernel, blur, noise, and combined mismatch. Blur is applied in the HR space before downsampling; noise is added independently in the LR space after downsampling, using a deterministic per-image seed so that the perturbation is reproducible.

Table A3: Test-time degradation settings. Each is applied only at test time to a model trained under standard bicubic downsampling.

Tag	Type	Setting
K1	Kernel	area-based downsampling
K2	Kernel	alternative bicubic implementation
B1	Blur	Gaussian $\sigma = 0.6$, kernel 5×5
B2	Blur	Gaussian $\sigma = 1.0$, kernel 7×7
N1	Noise	Gaussian $\sigma = 1/255$
N2	Noise	Gaussian $\sigma = 2/255$
N3	Noise	Gaussian $\sigma = 5/255$
J1	Blur + Noise	$\sigma = 0.6$, $n = 2$
J2	Blur + Noise	$\sigma = 1.0$, $n = 5$

D FULL ROBUSTNESS RESULTS

This section reports the complete robustness study summarized in the main paper, under the nine test-time degradation settings and the training protocol defined in Section C. Table A4 lists the full results on ARAD-1K for SwinIR and SwinIR + *SR*^{2}*-Net* at the $\times 2$, $\times 4$, and $\times 8$ scales.

Observation Across all nine mismatch settings and scales, *SR*^{2}*-Net* reduces mSAM in every setting (Table A4), while preserving (and frequently improving) spatial metrics. The rectified model stays ahead of the Base in every setting, while the margin varies across conditions.

E FULL CROSS-DOMAIN RESULTS

Table A5 reports the complete cross-domain results at the $\times 2$ and $\times 4$ scales (stacked as two blocks), including the mSSIM columns omitted from the main paper for brevity. DDS2M is reported in the in-domain setting only, as its diffusion inference is far more costly than the other backbones. The four models in the table are trained on the 900 training images of ARAD-1K and directly evaluated on all CAVE scenes and all ICVL HSIs without fine-tuning. As in the main paper, attaching *SR*^{2}*-Net* yields mSAM reductions in every reported configuration on both target domains while preserving (and frequently improving) the spatial metrics.

F RECTIFIER CONTROLS AND SEED STABILITY: IMPLEMENTATION AND FULL RESULTS

The main paper reports the $\times 4$ subset of the comparison against simple spectral rectifiers. Here we provide the full three-scale comparison and the implementation details of the fixed post-hoc rectifiers. All three baselines operate on the output of a fixed SwinIR model and do not modify the SR backbone weights, whereas *SR*^{2}*-Net* is trained end-to-end as a model-agnostic rectification module.

Fixed rectifiers **Savitzky–Golay smoothing (SG)** applies a 1D Savitzky–Golay filter along the spectral dimension for each pixel independently, with window length $w = 7$ and polynomial order

Table A4: Robustness to mild test-time degradation mismatch on ARAD-1K (SwinIR vs. SwinIR + *SR*^{2}*-Net*). We consider nine test-time settings grouped as kernel mismatch (K1: area downsampling; K2: an alternative bicubic implementation), blur mismatch (B1/B2: HR-space Gaussian blur with $\sigma \in \{0.6, 1.0\}$ using 5×5 and 7×7 kernels before bicubic downsampling), noise mismatch (N1/N2/N3: i.i.d. Gaussian noise added in LR after downsampling with $\sigma = n/255$ for $n \in \{1, 2, 5\}$), and combined blur+noise mismatch (J1/J2: $(\sigma, n) \in \{(0.6, 2), (1.0, 5)\}$). Gain denotes the change over the baseline (+Ours − Base); since lower mSAM is better, a negative mSAM Gain indicates improvement.

Scale	Metric	Var.	Kernel		Blur		Noise			Blur+Noise	
			K1	K2	B1	B2	N1	N2	N3	J1	J2
×2	mPSNR↑	Base	46.9930	45.2782	42.6891	39.8390	44.1997	42.8779	39.0326	41.2689	36.9781
		+Ours	**47.3542**	**48.6912**	**46.3376**	**41.2105**	**46.1393**	**42.9718**	**39.2598**	**41.9803**	**37.0884**
		Gain	+0.3612	+3.4130	+3.6485	+1.3715	+1.9396	+0.0939	+0.2272	+0.7114	+0.1103
	mSSIM↑	Base	0.9945	0.9917	0.9859	0.9742	0.9879	0.9806	0.9361	0.9754	0.9200
		+Ours	**0.9949**	**0.9968**	**0.9934**	**0.9810**	**0.9892**	**0.9807**	**0.9408**	**0.9766**	**0.9270**
		Gain	+0.0004	+0.0051	+0.0075	+0.0068	+0.0013	+0.0001	+0.0047	+0.0012	+0.0070
	mSAM↓	Base	0.6412	0.7029	0.8262	1.0349	2.2169	3.8313	7.6378	3.8589	7.6421
		+Ours	**0.5743**	**0.5793**	**0.6685**	**0.9354**	**2.0748**	**3.7932**	**6.3217**	**3.7346**	**7.2463**
		Gain	-0.0669	-0.1236	-0.1577	-0.0995	-0.1421	-0.0381	-1.3161	-0.1243	-0.3958
×4	mPSNR↑	Base	38.0659	37.3518	37.0133	36.4237	36.9821	36.6436	35.2832	36.3791	34.7529
		+Ours	**38.7683**	**40.8013**	**40.4649**	**39.2507**	**39.2291**	**38.0369**	**35.9106**	**37.9000**	**35.3443**
		Gain	+0.7024	+3.4495	+3.4516	+2.8270	+2.2470	+1.3933	+0.6274	+1.5209	+0.5914
	mSSIM↑	Base	0.9596	0.9531	0.9498	0.9437	0.9464	0.9395	0.9024	0.9367	0.8948
		+Ours	**0.9670**	**0.9748**	**0.9731**	**0.9672**	**0.9618**	**0.9464**	**0.9086**	**0.9452**	**0.8970**
		Gain	+0.0074	+0.0217	+0.0233	+0.0235	+0.0154	+0.0069	+0.0062	+0.0085	+0.0022
	mSAM↓	Base	1.3722	1.4583	1.4760	1.5187	2.5611	3.8740	7.0666	3.8760	7.0590
		+Ours	**1.3204**	**1.2944**	**1.2709**	**1.2949**	**2.2280**	**3.2619**	**6.8138**	**3.2247**	**6.8102**
		Gain	-0.0518	-0.1639	-0.2051	-0.2238	-0.3331	-0.6121	-0.2528	-0.6513	-0.2488
×8	mPSNR↑	Base	32.6296	32.3557	32.3143	32.2368	32.0617	31.8439	31.1646	31.7823	31.0966
		+Ours	**33.3445**	**34.8300**	**34.8294**	**34.7736**	**34.1853**	**33.6229**	**31.9019**	**33.6025**	**31.9108**
		Gain	+0.7149	+2.4743	+2.5151	+2.5368	+2.1236	+1.7790	+0.7373	+1.8202	+0.8142
	mSSIM↑	Base	0.8769	0.8698	0.8689	0.8672	0.8608	0.8526	0.8230	0.8518	0.8211
		+Ours	**0.8968**	**0.9102**	**0.9098**	**0.9087**	**0.8953**	**0.8796**	**0.8289**	**0.8790**	**0.8265**
		Gain	+0.0199	+0.0404	+0.0409	+0.0415	+0.0345	+0.0270	+0.0059	+0.0272	+0.0054
	mSAM↓	Base	2.3049	2.4180	2.4190	2.4220	3.0980	4.0295	6.5479	4.0262	6.5525
		+Ours	**2.2712**	**2.1173**	**2.1052**	**2.0901**	**2.8654**	**3.5249**	**6.3644**	**3.5022**	**5.9158**
		Gain	-0.0337	-0.3007	-0.3138	-0.3319	-0.2326	-0.5046	-0.1835	-0.5240	-0.6367

Table A5: Full cross-domain results at the $\times 2$ and $\times 4$ scales (trained on ARAD-1K, tested on CAVE and ICVL without fine-tuning), grouped by backbone and scale. Higher mPSNR and mSSIM and lower mSAM are better; the Gain row is reported as (+Ours − Base), so a negative mSAM gain denotes improved spectral fidelity. Base denotes each backbone trained on ARAD-1K and directly evaluated on CAVE and ICVL, and +Ours denotes the same backbone augmented with *SR*^{2}*-Net* under the identical protocol.

Method	Scale	Variant	CAVE			ICVL		
			mPSNR↑	mSSIM↑	mSAM↓	mPSNR↑	mSSIM↑	mSAM↓
RCAN	×2	Base	42.7269	0.9866	5.0004	49.6448	0.9976	0.7389
		+Ours	**44.2866**	**0.9886**	**4.0978**	**51.8506**	**0.9979**	**0.4682**
		Gain	+1.5597	+0.0020	-0.9026	+2.2058	+0.0003	-0.2707
	×4	Base	35.6214	0.9537	6.2153	39.9389	0.9707	1.0969
		+Ours	**36.3326**	**0.9576**	**5.3594**	**40.9148**	**0.9723**	**0.7897**
		Gain	+0.7112	+0.0039	-0.8559	+0.9759	+0.0016	-0.3072
SSPSR	×2	Base	41.4071	0.9816	4.0419	48.4758	0.9966	0.8140
		+Ours	**43.0076**	**0.9845**	**2.9243**	**50.7217**	**0.9972**	**0.5264**
		Gain	+1.6005	+0.0029	-1.1176	+2.2459	+0.0006	-0.2876
	×4	Base	36.7963	0.9521	4.9284	43.9812	0.9851	1.0271
		+Ours	**37.5721**	**0.9557**	**3.9189**	**45.1945**	**0.9856**	**0.6698**
		Gain	+0.7758	+0.0036	-1.0095	+1.2133	+0.0005	-0.3573
SwinIR	×2	Base	40.7157	0.9839	4.4366	47.9367	0.9964	0.8260
		+Ours	**42.3415**	**0.9877**	**2.6231**	**50.0704**	**0.9971**	**0.4709**
		Gain	+1.6258	+0.0038	-1.8135	+2.1337	+0.0007	-0.3551
	×4	Base	34.7725	0.9313	4.8551	41.0743	0.9708	1.0552
		+Ours	**35.7256**	**0.9577**	**3.8360**	**42.6530**	**0.9799**	**0.6682**
		Gain	+0.9531	+0.0264	-1.0191	+1.5787	+0.0091	-0.3870
ESSAformer	×2	Base	42.1763	0.9835	3.3575	49.7895	0.9973	0.6404
		+Ours	**43.4978**	**0.9855**	**2.5611**	**51.3981**	**0.9977**	**0.4625**
		Gain	+1.3215	+0.0020	-0.7964	+1.6086	+0.0004	-0.1779
	×4	Base	37.0305	0.9556	4.8258	44.3869	0.9856	0.9349
		+Ours	**37.6350**	**0.9579**	**3.5556**	**45.7012**	**0.9866**	**0.6079**
		Gain	+0.6045	+0.0023	-1.2702	+1.3143	+0.0010	-0.3270

$p = 3$; the spectrum is symmetrically reflection-padded before filtering and cropped back to avoid boundary bias. **PCA projection** fits a dataset-level PCA subspace from 20,000 pixel spectra randomly sampled from the ARAD-1K training split and keeps it fixed at test time; with rank $k = 8$, each estimated spectrum is projected onto the top-k principal components and reconstructed back to the original spectral space. **Iterative back-projection (IBP)** refines the SwinIR output by enforcing consistency with the observed LR image under the assumed degradation, performing $T = 10$ steps with step size $\eta = 1.0$,

$$I_{\mathrm{SR}}^{(t+1)} = I_{\mathrm{SR}}^{(t)} + \eta \cdot \mathcal{U}\left(I_{\mathrm{LR}} - \mathcal{D}\left(I_{\mathrm{SR}}^{(t)}\right)\right), \tag{A5}$$

where $\mathcal{D}$ is the same bicubic downsampling operator used in training and $\mathcal{U}$ is its bicubic upsampling counterpart, applied independently per band to keep the procedure lightweight.

Learned controls Two learned baselines isolate the source of the gain, both trained jointly with SwinIR under the identical loss, cosine schedule, 1000 epochs, and seed as *SR*^{2}*-Net* at $\times 4$ on ARAD-1K. $\mathcal{L}_{\mathrm{deg}}$ **only** adds the degradation-consistency constraint to the Base with no rectifier module (no extra parameters). **Generic head** is a same-budget rectifier that replaces H-S^{3}A and MCR by a plain residual block of three 3×3 convolutions ($31 \to 48 \to 48 \to 31$ channels with GeLU activations) and a global residual connection, adding $+0.048$M parameters to match *SR*^{2}*-Net*; unlike MCR it has no low-dimensional spectral bottleneck. These two learned controls are evaluated at $\times 4$ only.

Seed stability To confirm the attribution is robust, we retrain the five SwinIR/ARAD-1K variants with three seeds ($\{42, 123, 2024\}$) and report the mean $\pm$ std in Table A7. The across-seed standard deviation is small (≤ 0.012 mSAM), and the gap between the cheap controls (1.362 and 1.371 mSAM) and MCR / *SR*^{2}*-Net* (1.319 and 1.280) ranges from 0.043 to 0.091, at least $3.5\times$ that deviation, so the ranking, and the conclusion that the gain comes from the architecture, is stable across seeds.

Table A6: Spectral-rectifier comparison on ARAD-1K with SwinIR. SG, PCA, and IBP are fixed post-hoc methods; MCR and *SR*^{2}*-Net* are jointly trained. Best in **bold**, second best underlined.

Method	Scale	mPSNR↑	mSAM↓	mSSIM↑	CC↑
Base (SwinIR)	×2	50.4662	0.7727	0.9972	0.9990
	×4	39.5717	1.3950	0.9709	0.9902
	×8	33.4185	2.4682	0.8869	0.9551
Base + SG	×2	46.1893	1.9059	0.9959	0.9975
	×4	38.7834	2.4512	0.9698	0.9889
	×8	33.2419	3.2638	0.8859	0.9541
Base + PCA	×2	50.7837	0.6188	0.9972	0.9990
	×4	39.5729	1.4028	0.9709	0.9902
	×8	33.4198	2.4601	0.8869	0.9552
Base + IBP	×2	51.0962	0.5211	0.9973	0.9992
	×4	39.5782	1.3693	0.9709	0.9902
	×8	33.4041	2.4365	0.8865	0.9548
Base + MCR	×2	51.2531	0.5182	0.9973	0.9992
	×4	40.2550	1.3173	0.9717	0.9912
	×8	33.9241	2.2159	0.8947	0.9675
Base + Ours	×2	**51.6551**	**0.4842**	**0.9975**	**0.9993**
	×4	**40.9720**	**1.2819**	**0.9752**	**0.9924**
	×8	**34.8318**	**2.1164**	**0.9103**	**0.9720**

G Structural Hyperparameters of the Rectifier

We vary the three structural hyperparameters of *SR*^{2}*-Net* one at a time on ARAD-1K at $\times 4$ with SwinIR, keeping every other setting at its default ($r = 8$, 4 H-S^{3}A blocks, $N = 1$ MCR block, 4 spectral groups, $\lambda_{\mathrm{deg}} = 0.1$). Table A9 reports the sweeps. The default row of each sweep is the

Table A7: Seed stability on ARAD-1K ($\times 4$, SwinIR): mean $\pm$ std over three seeds $\{42, 123, 2024\}$. The architecture gap (cheap controls versus MCR/*SR*^{2}*-Net*) is 0.043 to 0.091 mSAM, at least $3.5\times$ the across-seed std.

Variant (SwinIR, $\times 4$)	mPSNR↑	mSAM↓	mSSIM↑
Base	$39.58_{\pm.03}$	$1.397_{\pm.011}$	$0.9710_{\pm.0005}$
+ $\mathcal{L}_{\text{deg}}$ only	$39.70_{\pm.03}$	$1.371_{\pm.010}$	$0.9710_{\pm.0007}$
+ Generic head	$39.99_{\pm.04}$	$1.362_{\pm.012}$	$0.9713_{\pm.0006}$
+ MCR only	$40.25_{\pm.03}$	$1.319_{\pm.010}$	$0.9716_{\pm.0006}$
+ *SR*^{2}*-Net*	$\mathbf{40.96}_{\pm.03}$	$\mathbf{1.280}_{\pm.009}$	$\mathbf{0.9753}_{\pm.0006}$

Table A8: Complexity of *SR*^{2}*-Net* using SwinIR on ARAD-1K at $\times 4$. Parameters and FLOPs are computed for the 32^2 LR training patch (128^2 HR); the rectifier adds 0.048M parameters and 1.482G FLOPs.

Scale	Metric	Base	+Ours	Δ
$\times 4$	Params (M)	8.228	**8.276**	+0.048
	FLOPs (G)	19.302	**20.785**	+1.482

full model of the main paper, and the 0-block and w/o-MCR rows are its single-stage ablations; these rows quote the main paper's values. The w/o-MCR row removes the module entirely, projection and back-projection included; it is not obtained by truncating the correction sum at $i = 0$, which would retain the term $g_m(m_0)$.

The bottleneck rank behaves as a two-sided constraint. Setting $r = 1$ raises mSAM to 1.349, and widening the projection to the band count ($r = 31$, one below its $S' = 32$ input) raises it to 1.331, against 1.2819 at the default; both margins exceed the across-seed standard deviation of Section F (≤ 0.012 mSAM) by more than fourfold. The $r = 31$ variant has more parameters than the default, so the loss is not a capacity effect. Between $r = 4$ and $r = 16$ mSAM varies by 0.006, which is below that standard deviation, so the width of the bottleneck is not critical in this range.

Both depth sweeps flatten at or before the default. A single H-S^{3}A block leaves mSAM at 1.322, close to the 1.3173 obtained with none; two and four blocks reach 1.298 and 1.2819, and eight give 1.288. For MCR, $N = 2$ is nominally the best setting (1.278 against 1.2819 at $N = 1$), while $N = 4$ and $N = 8$ fall back to 1.289 and 1.294; every $N \geq 1$ lies within 0.013 of the default, and we use $N = 1$, the cheapest of them.

Disabling the adjacent spectra shuffle while keeping the 1×1 mixing leaves the parameter count unchanged but raises mSAM from 1.2819 to 1.3102, more than twice the across-seed standard deviation, while mPSNR moves only from 40.9720 to 40.8871. Compared with the MCR-only variant (1.3173), the shuffle thus carries most of the spectral contribution of the H-S^{3}A stage, whereas the spatial gain largely persists without it. Because the fixed permutation is immediately followed by a dense 1×1 convolution, it does not enlarge the function class; its benefit is an optimization effect of the fixed adjacent-group coupling.

H Dataset-Level Spectral-Error and Significance Analysis

To complement the per-pixel spectral profiles in the main paper, we quantify the spectral correction over the full ARAD-1K test set (50 scenes) at $\times 4$ with SwinIR. For each pixel we form the residual spectrum $e = \hat{I}_{\text{SR}} - I_{\text{GT}}$ and report its mean absolute value (MAE), its mean absolute first spectral difference $|\Delta e|$, and its mean absolute second spectral difference $|\Delta^2 e|$ (a proxy for the band-to-band roughness of the error). As summarized in Table A10, *SR*^{2}*-Net* lowers all three statistics: the error magnitude on every scene, and its first- and second-difference variation on 44 and 42 of 50, respectively; per-scene mSAM improves on all 50 scenes. A one-sided Wilcoxon signed-rank test finds every difference significant. This shows at the dataset level that the improvement is systematic rather than driven by a few scenes. The three statistics fall at different rates: $|\Delta^2 e|$ decreases in absolute terms together with the error magnitude, but relative to the (smaller) residual it increases.

Table A9: Structural hyperparameter sweeps on ARAD-1K ($\times 4$, SwinIR). Each block varies one hyperparameter of the rectifier while every other setting keeps the default of the main paper ($r = 8$, four H-S^{3}A blocks, $N = 1$ MCR block, adjacent spectra shuffle on); * marks that default, so all * rows are the same full model. Best in **bold**, second best underlined within each block; the across-seed standard deviation is ≤ 0.012 mSAM (Table A7).

Sweep	Setting	mPSNR↑	mSSIM↑	mSAM↓
MCR rank r	1	40.44	0.9726	1.349
	2	40.67	0.9738	1.326
	4	40.91	0.9748	1.288
	8*	**40.97**	**0.9752**	**1.282**
	16	40.92	0.9745	1.284
	31	40.82	0.9741	1.331
H-S^{3}A blocks	0 (MCR only)	40.26	0.9717	1.317
	1	40.58	0.9727	1.322
	2	40.84	0.9743	1.298
	4*	**40.97**	**0.9752**	**1.282**
	8	40.93	0.9747	1.288

Sweep	Setting	mPSNR↑	mSSIM↑	mSAM↓
MCR blocks N	w/o (H-S^{3}A only)	40.71	0.9735	1.348
	1*	40.97	0.9752	1.282
	2	**40.99**	**0.9754**	**1.278**
	4	40.97	0.9753	1.289
	8	40.91	0.9746	1.294
Spectra shuffle	with*	**40.97**	**0.9752**	**1.282**
	w/o	40.89	0.9743	1.310

Table A10: Dataset-level residual spectral-error analysis over 50 ARAD-1K test scenes (SwinIR, $\times 4$). "Improved" counts scenes on which *SR^2-Net* lowers the statistic; p is from a one-sided Wilcoxon signed-rank test. Per-scene mSAM improves on all 50 scenes ($p < 10^{-6}$).

Residual statistic	Base	+ *SR^2-Net*	Improved	p
Spectral MAE	0.00783	**0.00534**	50/50	$< 10^{-6}$
Gradient $\lvert\Delta e\rvert$	0.000838	**0.000781**	44/50	$< 10^{-3}$
Curvature $\lvert\Delta^2 e\rvert$	0.000697	**0.000618**	42/50	$< 10^{-3}$

Table A11: Frozen-backbone results. Each backbone is held at its Base weights and only *SR^2-Net* is trained on top of it, so the backbone is used exactly as it was originally trained. mSAM falls in all sixteen configurations, recovering 76% of the jointly trained gain reported in the main paper on average, with per-configuration recovery ranging from 71% to 87%. Across the remaining metrics the single exception is ESSAformer's mSSIM on ARAD-1K at $\times 2$, which changes by -0.0001.

		RCAN		SSPSR		SwinIR		ESSAformer	
Setting	Metric	Base	Frozen	Base	Frozen	Base	Frozen	Base	Frozen
ARAD-1K $\times 2$	mPSNR↑	48.74	49.16	46.09	46.49	50.47	50.64	49.45	49.61
	mSSIM↑	0.9936	0.9948	0.9893	0.9894	0.9972	0.9974	0.9937	0.9936
	mSAM↓	1.2071	0.9043	0.9906	0.7541	0.7727	0.5642	0.6090	0.5321
	CC↑	0.9978	0.9984	0.9960	0.9979	0.9990	0.9991	0.9986	0.9989
ARAD-1K $\times 4$	mPSNR↑	38.38	38.65	38.88	39.05	39.57	39.76	39.95	40.08
	mSSIM↑	0.9470	0.9536	0.9463	0.9483	0.9709	0.9725	0.9592	0.9595
	mSAM↓	1.9009	1.4622	1.6076	1.3965	1.3950	1.3142	1.3670	1.2682
	CC↑	0.9812	0.9895	0.9858	0.9875	0.9902	0.9914	0.9903	0.9906
CAVE $\times 2$	mPSNR↑	41.09	41.72	39.40	39.71	40.31	40.68	39.47	39.86
	mSSIM↑	0.9838	0.9863	0.9711	0.9748	0.9754	0.9786	0.9704	0.9708
	mSAM↓	5.7619	4.7216	4.2518	3.8642	0.8214	0.6723	4.0820	3.4925
	CC↑	0.9946	0.9965	0.9907	0.9942	0.9807	0.9882	0.9950	0.9954
CAVE $\times 4$	mPSNR↑	35.99	36.42	35.24	35.41	37.27	37.51	34.11	34.48
	mSSIM↑	0.9536	0.9572	0.9384	0.9401	0.9492	0.9527	0.9267	0.9274
	mSAM↓	7.6445	6.0418	4.8735	4.4726	0.9015	0.7582	5.6070	4.7213
	CC↑	0.9825	0.9853	0.9795	0.9808	0.9315	0.9461	0.9846	0.9873

I ADDITIONAL ERROR-MAP GALLERIES

Figures A1–A3 provide per-scene error-map galleries at $\times 4$: on ARAD-1K and CAVE under the in-domain setting, and on ICVL under the cross-domain setting, where models are trained on ARAD-1K and tested on ICVL without fine-tuning. For each scene we show pseudo-RGB renderings and the band-averaged absolute error maps of different SR backbones, with each rectified result (+Ours) placed next to its baseline. On the unseen ICVL data, the boundary-related errors are reduced in the same way as in the in-domain comparisons of the main paper.

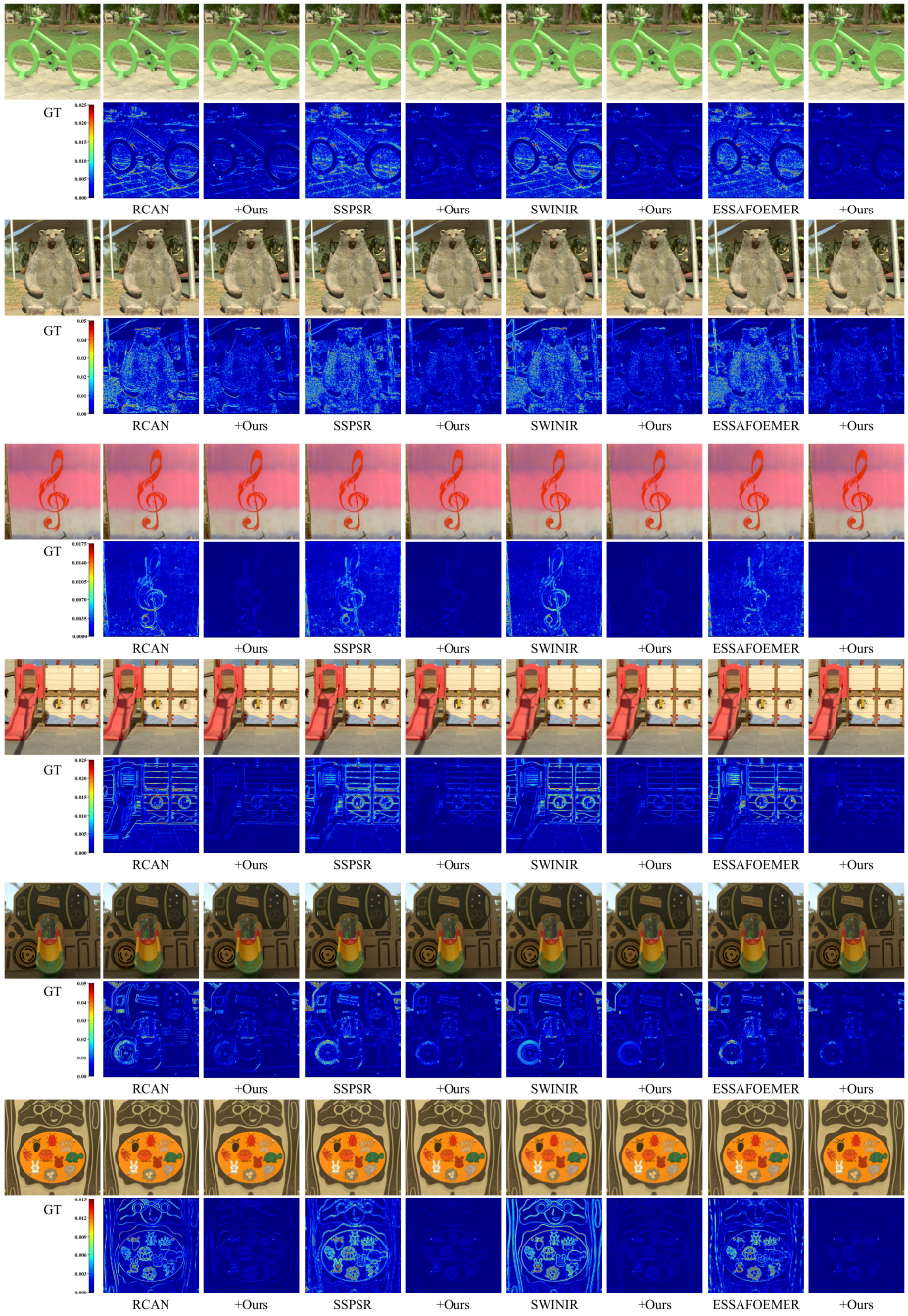

Figure A1: Error-map gallery on ARAD-1K at $\times 4$ (in-domain).

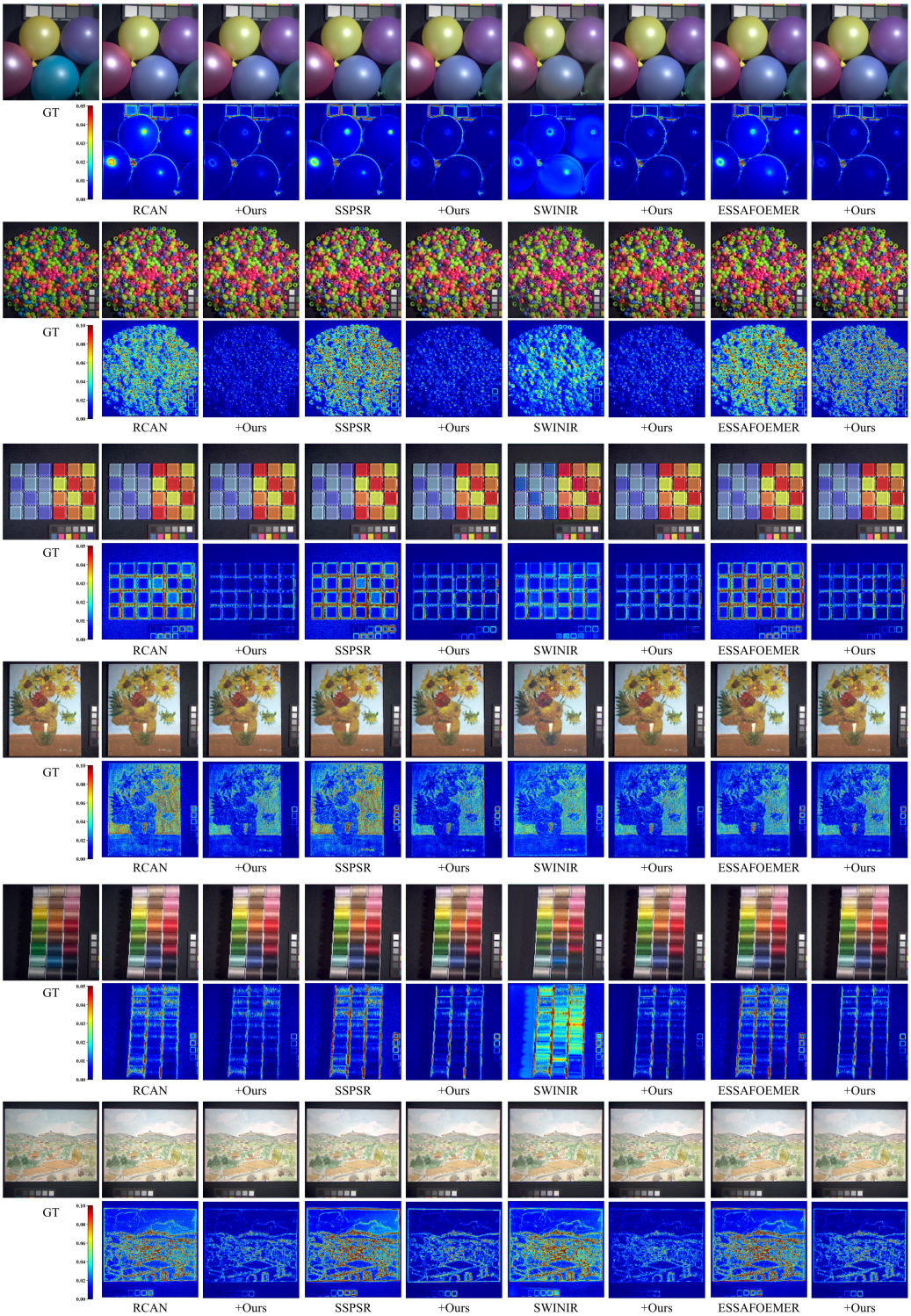

Figure A2: Error-map gallery on CAVE at $\times 4$ (in-domain). The visualization protocol is identical to Figure A1.

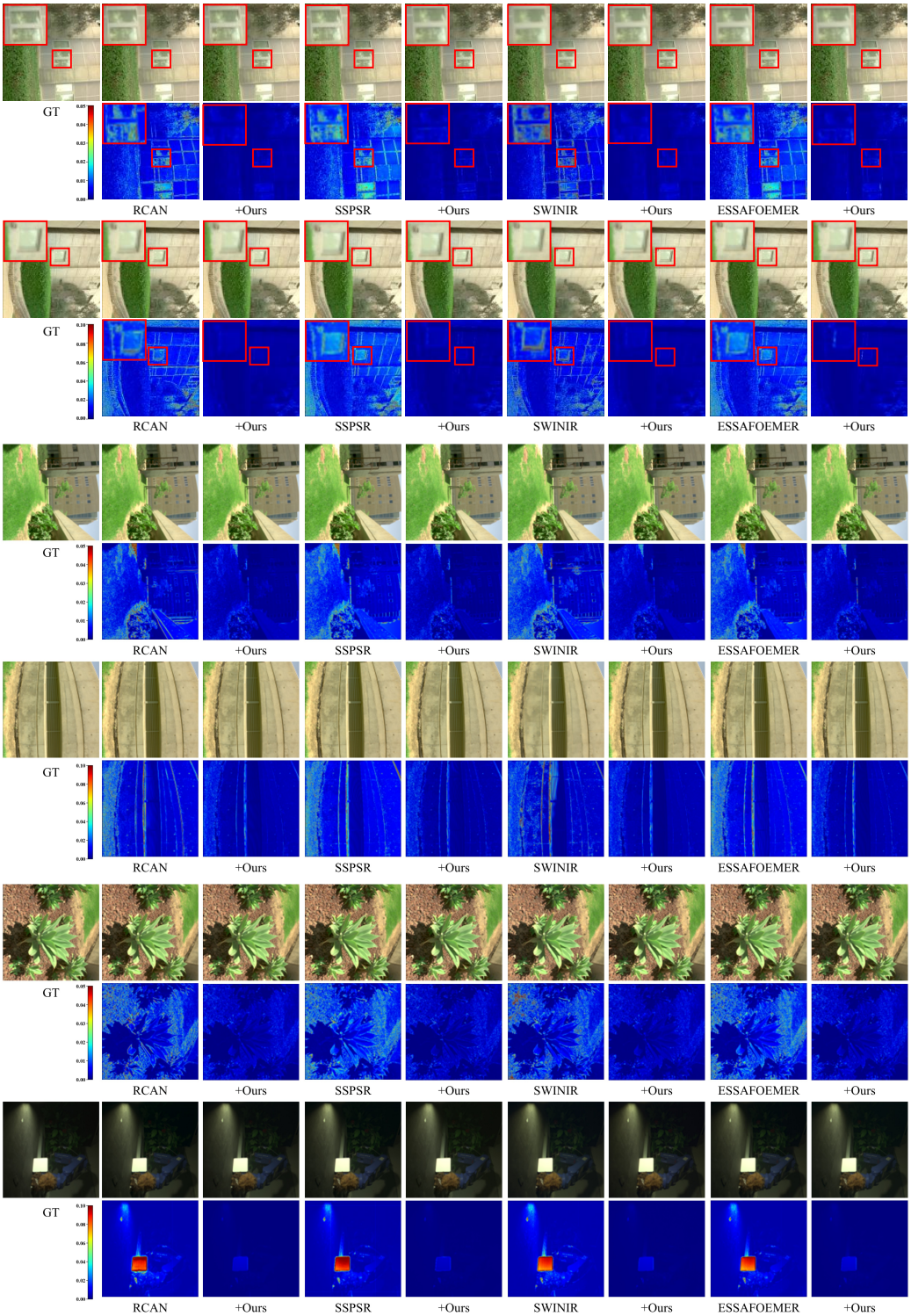

Figure A3: Error-map gallery on ICVL at $\times 4$ under the cross-domain setting (trained on ARAD-1K, tested on ICVL without fine-tuning). The visualization protocol is identical to Figure A1.

J ADDITIONAL SPECTRAL-CURVE GALLERIES

Figures A4–A6 provide additional per-pixel spectral-curve galleries on CAVE, ICVL, and ARAD-1K at $\times 4$, spanning CNN-, Transformer-, and diffusion-based backbones (including DDS2M). For each marked pixel, *SR*^{2}*-Net* (green dashed) follows the ground-truth spectrum (blue) more closely than the baseline (orange).

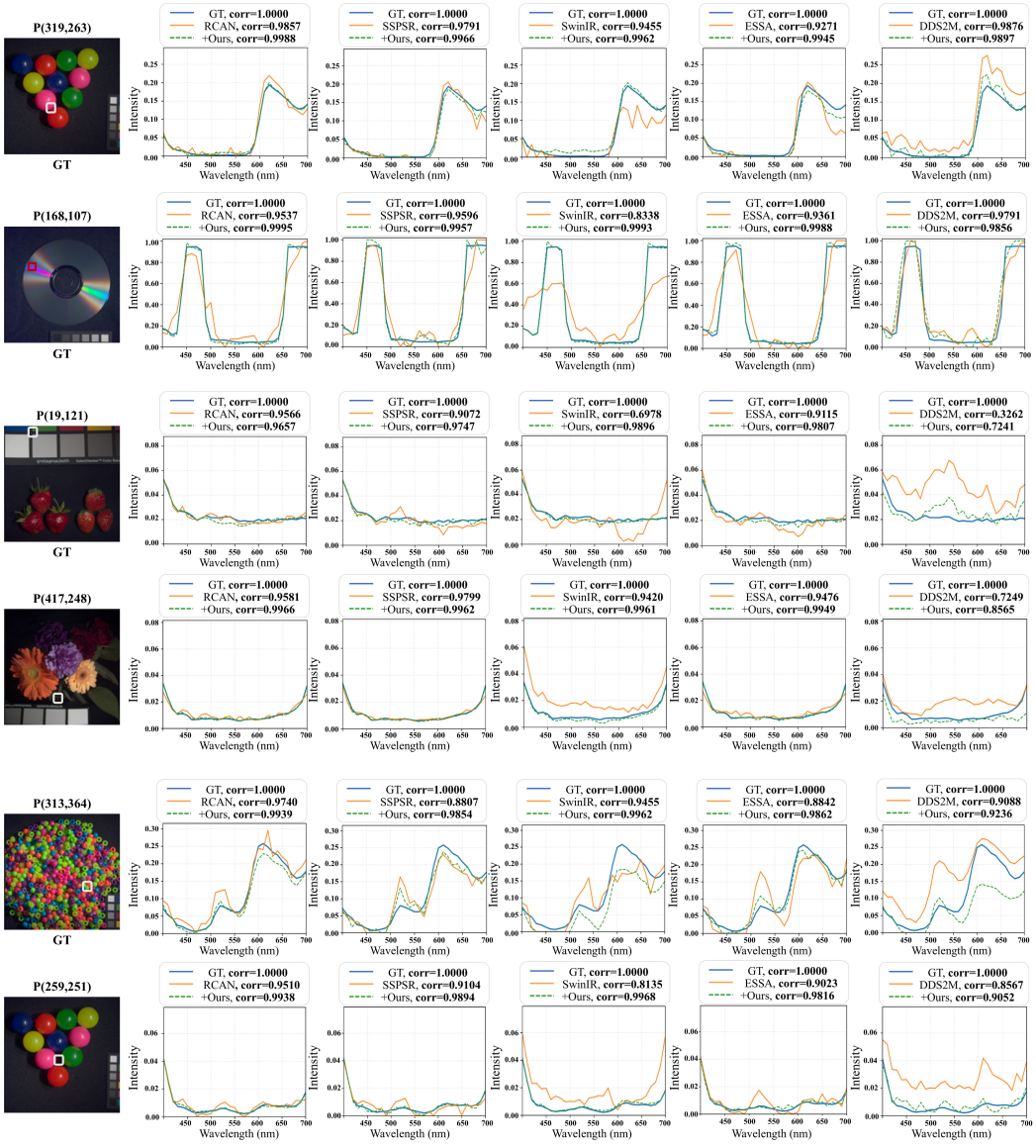

Figure A4: Additional per-pixel spectral-curve gallery on CAVE at $\times 4$.

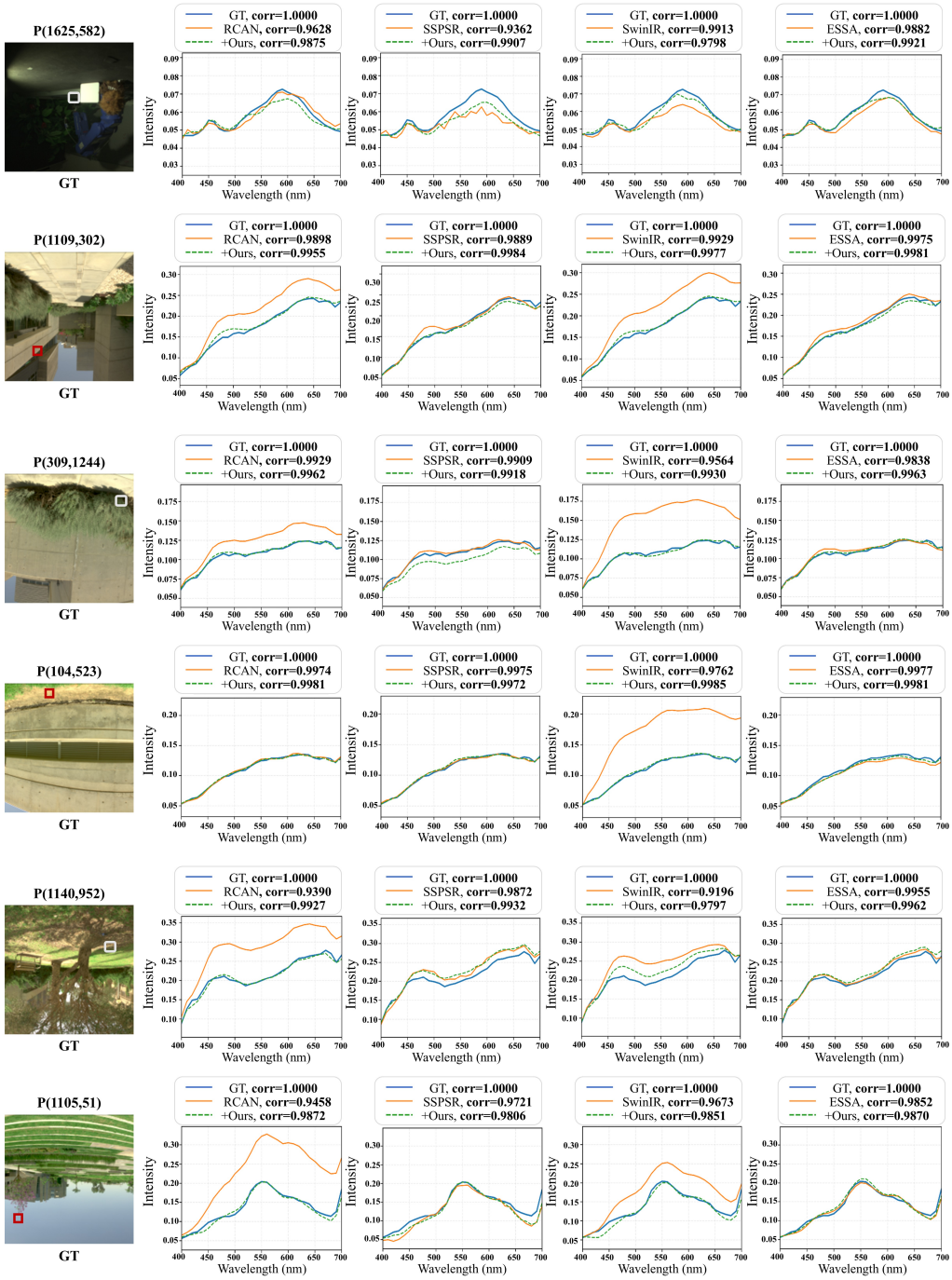

Figure A5: Additional per-pixel spectral-curve gallery on ICVL at $\times 4$ under the cross-domain setting.

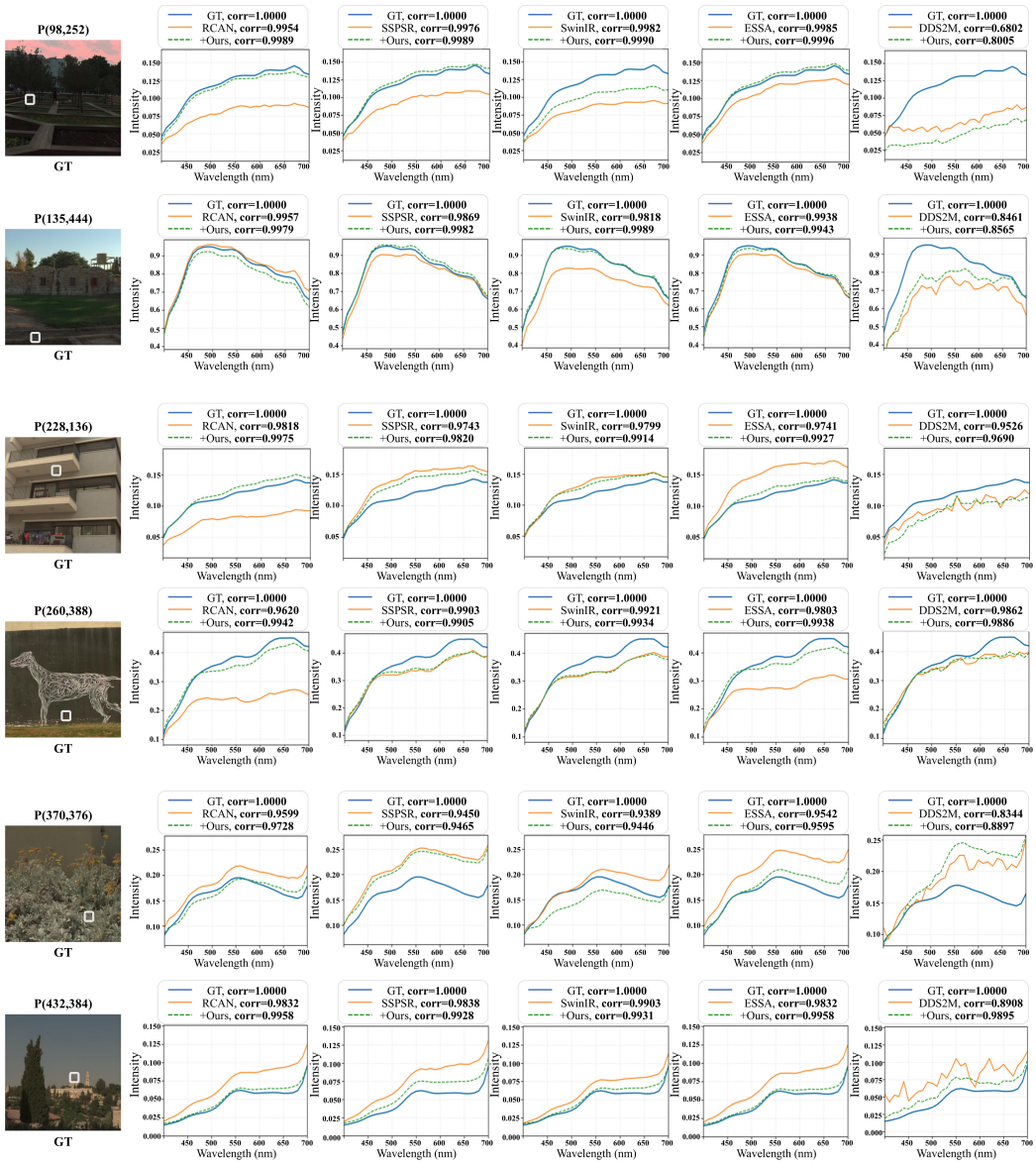

Figure A6: Additional per-pixel spectral-curve gallery on ARAD-1K at $\times 4$.