# ENTROPY-GUIDED HETEROSCEDASTICITY MODELING FOR FAST ZERO-SHOT REMOTE SENSING IMAGE FUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hyperspectral pansharpening (HyPan) and hyperspectral-multispectral image fusion (HMIF) reconstruct a high-resolution hyperspectral image (HR-HSI) from a low-resolution hyperspectral image (LR-HSI) and a high-spatial-resolution panchromatic or multispectral observation (PAN/MSI), a fundamental and highly ill-posed inverse problem in remote sensing. We address the zero-shot setting, in which the result must be obtained from the input observation pair alone without any external training data, and focus on optimization-based zero-shot methods, which optimize an untrained network directly on the input observations and stay faithful to the physical observation model. Uniformly weighting their spatial and spectral fidelity terms is, however, problematic: the spatial term is mainly informative at edges and fine structures, while the flat background that dominates remote sensing scenes dilutes these sparse constraints, and the spectral term is dominated by low-frequency content, so the two objectives drive the network in conflicting directions and the optimization settles at a sub-optimal compromise. We propose a fast zero-shot optimization framework with entropy-guided heteroscedasticity modeling: an information gate computed from the local Shannon entropy down-weights uninformative regions and strengthens genuine high-frequency structure, aligning the two objectives, and a two-stage schedule first anchors the spectral manifold and then refines spatial details. On HyPan and HMIF benchmarks, our method improves spectral fidelity and spatial sharpness over state-of-the-art methods, reconstructing an image in about 17 seconds without any external training data.

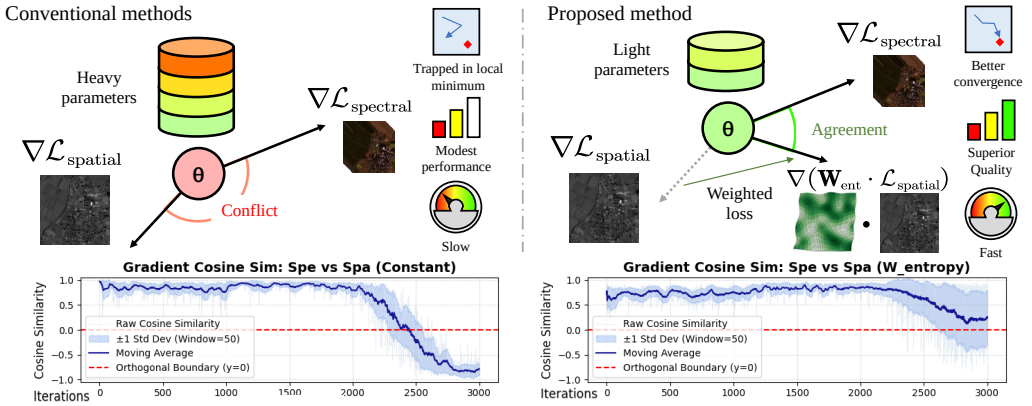



Figure 1: Conceptual comparison and empirical gradient analysis. **Top:** under conflicting spatial and spectral gradients, a heavily parameterized network (left) settles into a sub-optimal compromise, while weighting the spatial loss with an entropy prior (right) aligns the gradients and lets a lightweight network converge to a high-quality solution. **Bottom:** measured cosine similarity between the spatial and spectral loss gradients during optimization with constant weighting (left) and entropy-guided weighting (right); the gate keeps the two objectives in agreement throughout the optimization.

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 INTRODUCTION

Hyperspectral pansharpening (HyPan) and hyperspectral-multispectral image fusion (HMIF) are remote sensing fusion tasks that reconstruct a high-resolution hyperspectral image (HR-HSI) from a low-spatial-resolution hyperspectral image (LR-HSI) and a high-spatial-resolution panchromatic or multispectral observation (PAN/MSI). Because a single sensor cannot simultaneously achieve high spectral resolution and high spatial resolution, such fusion is a fundamental and highly ill-posed inverse problem, and accurate reconstruction matters for downstream earth-observation tasks that require physical and radiometric consistency Kwarteng & Chavez (1989); Mallat (2002). This paper studies the zero-shot setting, in which the result must be obtained from the input observation pair alone without any external training data.

Deep learning methods are now widely used for this task and can model highly nonlinear transformations, while variational optimization (VO) methods remain an active line of research Tian et al. (2021); Dian et al. (2019); deep models, however, typically require large paired datasets, which are scarce for hyperspectral data Sidorov & Hardeberg (2019); Li et al. (2025) and costly to collect in practice. Most current methods, including zero-shot learning (ZSL) approaches, follow Wald's protocol to construct pseudo-paired data by blurring and downsampling the input observations Dian et al. (2023); Cao et al. (2024). A model trained this way operates in the downsampled space, and applying it to full-resolution images leaves a representational gap: it tends to overfit the training scale, fails to extrapolate high-frequency structure, and produces blurred textures at inference time Wang et al. (2024); Lu et al. (2025). Pre-trained generative priors such as diffusion models offer an alternative route to inject spatial details Rui et al. (2024); Pang et al. (2024), yet their iterative sampling is slow and the domain gap between pre-training data and the fusion task can induce spectral hallucinations Zhao et al. (2026b); Xiao et al. (2025); Zhao et al. (2026a); Wei et al. (2025).

Within this zero-shot setting, optimization-based zero-shot methods are appealing: an untrained network is optimized directly on the test observations, so no external data or pre-training are needed and the solution stays faithful to the physical observation model. This line of work originates from Lempitsky et al. Lempitsky et al. (2018) and has been extended to hyperspectral restoration and super-resolution Sidorov & Hardeberg (2019); Li et al. (2025); Jiang et al. (2022), including recent zero-shot fusion frameworks such as CLoRF Wang et al. (2025). When such methods are applied to fusion, however, the spatial and spectral fidelity terms can interfere with each other, a difficulty reported in recent zero-shot fusion studies Liu et al. (2026); Nie et al. (2024) and analogous to the gradient interference studied in multi-task learning Yu et al. (2020). The two terms are unbalanced in what they can constrain. The spectral term acts on every degraded pixel and is dominated by low-frequency content, whereas the spatial term is mainly informative at edges and fine textures and contributes little in locally constant regions Vivone et al. (2014). Because such regions dominate remote sensing scenes, a uniform scale parameter, i.e., equal weight on every spatial residual, lets their many uninformative pixels dilute the sparse structural gradients; the spatial term then cannot anchor high-frequency structure, and jointly with the low-frequency spectral term the optimization drifts toward over-smoothed solutions. Existing zero-shot methods already exhibit this difficulty: CS2DIPs Fang et al. (2024) notes that optimizing on the degraded image alone can hardly capture dense spatial and spectral characteristics jointly and that the results turn out sub-optimal, and our own measurements confirm it (bottom of Figure 1), where the cosine similarity between the two gradients turns negative in the later optimization stages. Although the two gradient directions can cancel so that the optimization still converges, the resulting compromise is sub-optimal: the network spends its capacity on the redundant smooth regions instead of on the sparse informative structures, so genuine edges remain under-sharpened. The compromise thus neither preserves spectral fidelity in flat areas nor sharpens genuine edges; the problem is not the absence of convergence but the quality of the reached solution.

To solve this problem, we propose a fast zero-shot optimization framework with entropy-guided heteroscedasticity modeling. Re-deriving the fusion objective from a maximum a posteriori (MAP) perspective shows that the reliability of the spatial guidance varies across pixels. Using the local Shannon entropy of the spatial observation as a proxy for the structural signal-to-noise ratio (SNR), we build an information gate that down-weights the loss in low-entropy regions, where the local spatial constraint is redundant and carries little structural information, and up-weights it along genuine structural boundaries. The gate reroutes the fitting capacity of the network from the redundant smooth background to genuine high-frequency structure, so that the informative spatial gradients are no longer

diluted and no longer oppose the spectral ones (bottom of Figure 1). In addition, the optimization is decoupled into two stages that first anchor the spectral manifold and then refine spatial details, which avoids the spectral bias that typically limits zero-shot methods.

Our contributions are summarized as follows:

- We analyze the gradient conflict of zero-shot fusion under the uniform scale assumption, and formulate spatial heteroscedasticity through an entropy-based prior computed from the observation, deriving an information gate that removes the conflict and keeps the spatial fidelity focused on the genuinely informative structures.
- We design a decoupled two-stage optimization strategy that anchors the spectral manifold before spatial refinement, avoiding the spectral bias and the sub-optimal compromise typical of zero-shot methods.
- On several benchmark datasets, the proposed method improves spectral fidelity and spatial sharpness over state-of-the-art approaches with a lightweight network and an execution time of about 17 seconds per image.

## 2 Related Work

### 2.1 Image Fusion and Learning-Based Methods

Early remote sensing image fusion methods fall into three families: component substitution Kwarteng & Chavez (1989), multi-resolution analysis Mallat (2002), and variational optimization Tian et al. (2021); Dian et al. (2019). The first two are computationally cheap but rely on simple fusion rules, while variational methods enforce physical consistency through alternating optimization but depend on hand-crafted regularizers. Deep learning models substantially improve structural expressiveness Deng et al. (2020); Tang et al. (2022). Since ground-truth high-resolution hyperspectral images are unavailable, zero-shot learning (ZSL) methods construct pseudo-paired data by spatial downsampling Dian et al. (2023); Cao et al. (2024); Wang et al. (2024). This creates a representational gap Lu et al. (2025): networks trained only in the low-frequency downsampled domain generalize poorly to full resolution, showing scale-specific overfitting and reduced spatial sharpness.

Pre-trained diffusion models have also been used as generative priors for image restoration: remote-sensing adaptations such as PLRDiff Rui et al. (2024) and HIR-Diff Pang et al. (2024) recover missing spatial frequencies with pre-trained diffusion backbones, but the iterative reverse sampling is computationally expensive and the domain gap between natural pre-training images and remote-sensing spectra can induce localized spectral hallucinations Zhao et al. (2026b); Xiao et al. (2025); Zhao et al. (2026a); Wei et al. (2025). By contrast, the methods considered next optimize the observations directly.

### 2.2 Zero-Shot Optimization

In this paradigm, an untrained convolutional network is optimized on the degraded image itself, so that its structure acts as an implicit regularizer that favors smooth, natural images over noise, and a randomly initialized network can solve inverse problems without any training Lempitsky et al. (2018). This idea has been extended to hyperspectral data for denoising, inpainting, and super-resolution Sidorov & Hardeberg (2019), for multispectral super-resolution with enhanced architectures Li et al. (2025), and across multiple scales for unsupervised denoising Jiang et al. (2022). In fusion tasks, zero-shot optimization frameworks built on this paradigm, such as CLoRF Wang et al. (2025), integrate implicit neural representations with self-supervised low-rank factorization and require no external data, and CS2DIPs Fang et al. (2024) couples spatial and spectral branches for unsupervised hyperspectral super-resolution.

Because these methods remain faithful to the physical observation model, they avoid both the representational gap of downsampled training and the domain gap of pre-trained priors. Applied to fusion, however, they must jointly satisfy the spatial and spectral observation constraints, and the two corresponding loss terms can interfere with each other, a phenomenon analogous to the gradient interference studied in multi-task learning Yu et al. (2020) and reported in recent zero-shot fusion studies Liu et al. (2026); Nie et al. (2024). The two terms are unbalanced in their information

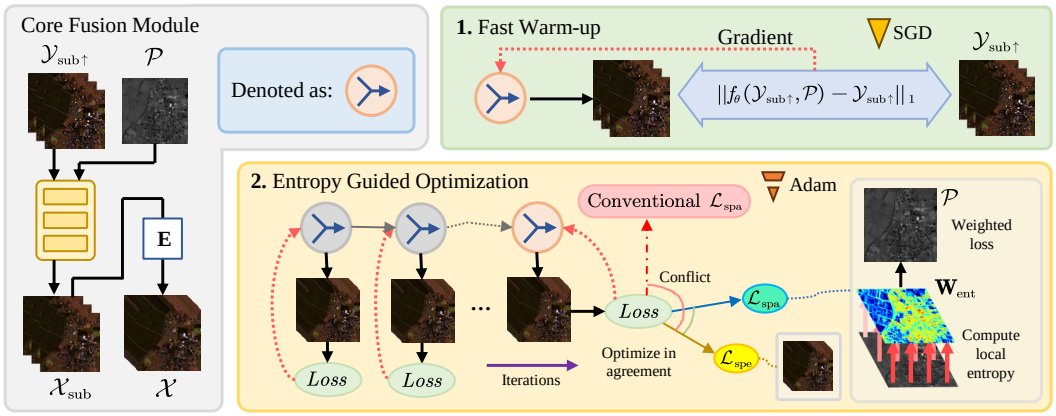



Figure 2: Overview of the proposed zero-shot fusion framework. **Left:** the core fusion module ($f_\theta$) takes the upsampled spectral observation ($\mathcal{Y}_{\text{sub}\uparrow}$) and the spatial observation ($\mathcal{P}$), projects them onto the subspace $\mathbf{E}$ to reconstruct $\mathcal{X}$, and is shown as a single icon. **Top right:** the Stage I warm-up (SGD) anchors the spectral subspace. **Bottom right:** the Stage II refinement (Adam), where the entropy map ($\mathbf{W}_{\text{ent}}$) computed from $\mathcal{P}$ weights the spatial loss; this gate removes the gradient conflict of a uniformly weighted spatial loss (red dashed) and aligns $\mathcal{L}_{\text{spa}}$ with $\mathcal{L}_{\text{spe}}$.

content: the spatial term is mainly informative at edges and fine textures Vivone et al. (2014), and uniform weighting dilutes these sparse constraints against the structure-less smooth background, while the spectral term is dominated by low-frequency content, so the two objectives pull the network in conflicting directions. Although the two gradient directions can cancel so that the optimization still settles at a compromise point, this equilibrium is sub-optimal: the network wastes its capacity on the redundant smooth regions instead of sharpening genuine edges. In contrast, the proposed framework weights the spatial loss with an entropy-guided information gate computed from the observation and decouples the optimization into two stages, which removes this conflict.

## 3 Method

This section describes the proposed fast zero-shot image fusion framework, whose overall pipeline is shown in Figure 2. We first formulate fusion as a MAP estimation problem with spatially varying reliability of the spatial observation, and then introduce the subspace projection, the lightweight network, the entropy-based information gate, and the decoupled optimization strategy.

### 3.1 Problem Formulation

Both HyPan and HMIF aim to reconstruct an ideal high-resolution hyperspectral image (HR-HSI) $\mathcal{X} \in \mathbb{R}^{H\times W\times C_{hsi}}$ from a low-resolution hyperspectral image (LR-HSI) $\mathcal{Y} \in \mathbb{R}^{H_v\times W_v\times C_{hsi}}$ and a high-spatial-resolution PAN/MSI observation $\mathcal{P} \in \mathbb{R}^{H\times W\times C_{msi}}$. We adopt a channels-last layout, so that $\mathcal{X}(i,j) \in \mathbb{R}^{1\times C_{hsi}}$ denotes the spectral vector at pixel $(i,j)$. This inverse problem can be formulated as a MAP estimation:

$$\mathcal{X}^* = \arg\max_{\mathcal{X}} P(\mathcal{X}|\mathcal{Y},\mathcal{P}) = \arg\max_{\mathcal{X}} \frac{P(\mathcal{Y},\mathcal{P}|\mathcal{X})P(\mathcal{X})}{P(\mathcal{Y},\mathcal{P})}. \tag{1}$$

Because the spectral sensor ($\mathcal{Y}$) and the spatial sensor ($\mathcal{P}$) are physically independent imaging systems, the joint likelihood factorizes under the conditional independence assumption, i.e., $P(\mathcal{Y},\mathcal{P}|\mathcal{X}) = P(\mathcal{Y}|\mathcal{X})P(\mathcal{P}|\mathcal{X})$. Taking the negative log-likelihood turns the MAP estimate into the minimization of the energy functional:

$$\mathcal{L}(\mathcal{X}) = -\log P(\mathcal{Y}|\mathcal{X}) - \log P(\mathcal{P}|\mathcal{X}) - \log P(\mathcal{X}). \tag{2}$$

Because real acquisitions may contain mild noise or outliers, we model the observation residuals with zero-mean Laplace distributions Fu et al. (2006); Lanaras et al. (2015). Unlike the Gaussian prior, which penalizes outliers aggressively, the Laplace prior tolerates sparse deviations and yields

an $\ell_1$-norm objective. The per-pixel scale parameters of this model additionally control the emphasis of each residual, which lets us account for the spatially varying information content of the spatial observation. Expanding the negative log-likelihood terms, we obtain:

$$\mathcal{L}(\mathcal{X}) = \frac{1}{b_{\text{spe}}}\|(\mathbf{B} \otimes \mathcal{X}) \downarrow - \mathcal{Y}\|_1 + \sum_{i,j} \frac{1}{b_{\text{spa}}(i,j)}\|\mathcal{X}(i,j)\mathbf{R} - \mathcal{P}(i,j)\|_1 + \lambda\Phi(\mathcal{X}), \tag{3}$$

where $\mathbf{B}$ is the spatial blur kernel, $\otimes$ the spatial convolution applied to each band, $\downarrow$ the downsampling operator, and $\mathbf{R} \in \mathbb{R}^{C_{hsi} \times C_{msi}}$ the spectral response matrix. The scale parameter $b_{\text{spa}}(i,j)$ varies across pixels; its inverse controls the emphasis of the spatial residual at pixel $(i,j)$. The term $\lambda\Phi(\mathcal{X})$ is a prior on the target image.

### 3.2 Dimension-Agnostic Subspace Projection

Hyperspectral images contain strong spectral redundancy, so their spectral vectors lie close to a low-dimensional subspace. To handle sensors with different numbers of bands without modifying the network, we project the target $\mathcal{X}$ onto a common low-dimensional subspace:

$$\mathcal{X} = \mathcal{X}_{\text{sub}} \times_3 \mathbf{E}, \tag{4}$$

where $\times_3$ denotes the mode-3 product, i.e., the mode-3 matrix unfolding of the tensor multiplied by $\mathbf{E}$: $(\mathcal{X}_{\text{sub}} \times_3 \mathbf{E})(i,j,c) = \sum_k \mathcal{X}_{\text{sub}}(i,j,k)\mathbf{E}(k,c)$ Sidiropoulos et al. (2017). Here $\mathbf{E} \in \mathbb{R}^{K \times C_{hsi}}$ is a precomputed projection basis ($K \ll C_{hsi}$) and $\mathcal{X}_{\text{sub}} \in \mathbb{R}^{H \times W \times K}$ is the subspace coefficient map. In practice, $\mathbf{E}$ is estimated by least squares on a uniformly selected $K$-band subset $\mathcal{Y}_{\text{sub}\uparrow} \in \mathbb{R}^{H \times W \times K}$ of the bicubically upsampled LR-HSI $\mathcal{Y}_\uparrow \in \mathbb{R}^{H \times W \times C_{hsi}}$:

$$\mathbf{E} = \arg\min_{\mathbf{M}} \|\mathcal{Y}_{\text{sub}\uparrow} \times_3 \mathbf{M} - \mathcal{Y}_\uparrow\|_F^2. \tag{5}$$

Optimizing over the coefficient map $\mathcal{X}_{\text{sub}}$ instead of the full tensor $\mathcal{X}$ makes the framework independent of the band count of the sensor.

### 3.3 Lightweight Architecture

Optimizing the coefficient map directly is prone to noise overfitting, so we parameterize it with a lightweight network $f_\theta$, expressed as $\mathcal{X}_{\text{sub}} = f_\theta(\mathcal{Y}_{\text{sub}\uparrow}, \mathcal{P})$. Following this zero-shot paradigm, optimizing the network parameters $\theta$ rather than the pixel values provides an implicit structural regularizer, which plays the role of the prior $\lambda\Phi(\mathcal{X})$ in Eq. (3). In our implementation, $f_\theta$ is a four-layer convolutional neural network (CNN) with a feature extraction head, a residual body with local shortcut connections, and a projection tail with a sigmoid activation that keeps the output in the physically valid range $[0,1]$.

### 3.4 SNR-Driven Heteroscedasticity Modeling via Local Entropy

Conventional formulations assign a single scale parameter to every spatial residual ($b_{\text{spa}}(i,j) = \text{const}$), which forces the network to fit microscopic fluctuations everywhere. In homogeneous regions the true structural signal has no spatial gradient, so any minute fluctuation of the residual is dominated by sensor noise or interpolation artifacts; the local structural SNR drops toward zero, and the spatial guidance from such regions is unreliable ($b_{\text{spa}}(i,j) \to \infty$, i.e., a vanishing weight). In structural edge areas, by contrast, the true signal overwhelms the fluctuations, the local SNR is high, and the residual carries reliable guidance ($b_{\text{spa}}(i,j) \to 0$, i.e., a large weight). A uniform scale parameter therefore spends a large share of the spatial fidelity on uninformative smooth regions while under-using the sparse informative edges.

To estimate this spatially varying emphasis without external data, we use the local Shannon entropy of the spatial observation $\mathcal{P}$ as a proxy for structural informativeness. For a local window $\Omega$ centered at pixel $(i,j)$, the entropy is:

$$H(i,j) = -\sum_k p_k \log p_k, \tag{6}$$

where $p_k$ is the probability of intensity level $k$ within $\Omega$. This choice follows directly from the information-theoretic meaning of entropy: in a smooth homogeneous region the local histogram

concentrates on a few intensity levels and $H$ approaches zero, which corresponds to low structural confidence (large $b_{\text{spa}}$); near an edge the histogram spreads out and $H$ is high, which corresponds to high confidence. The entropy is therefore a data-intrinsic estimate of the spatially varying weight of each spatial residual. Unlike stochastic uncertainty estimates such as Monte Carlo dropout, it requires no extra forward passes, which is essential for a fast zero-shot framework. After min-max normalization to obtain $H_{\text{norm}}(i,j) \in [0,1]$, we approximate the inverse scale parameter by:

$$\frac{1}{b_{\text{spa}}(i,j)} \approx \eta \cdot \left[H_{\text{norm}}(i,j)\right]^2. \tag{7}$$

The quadratic exponent acts as a continuous soft-threshold on the confidence map: because $H_{\text{norm}}(i,j) \in [0,1]$, squaring further suppresses low- and mid-entropy regions, including weak interpolation artifacts and slightly textured background, while leaving the constraints of strong structural boundaries nearly unchanged. This non-linear contrast enhancement concentrates the effective spatial constraints on genuine high-frequency structures. Substituting Eq. (7) into Eq. (3) yields the weighted spatial loss:

$$\mathcal{L}_{\text{spa}} = \sum_{i,j} \mathbf{W}_{\text{ent}}(i,j) \cdot \|\mathcal{X}(i,j)\mathbf{R} - \mathcal{P}(i,j)\|_1, \tag{8}$$

with $\mathbf{W}_{\text{ent}}(i,j) = \left[H_{\text{norm}}(i,j)\right]^2$. The gate down-weights uninformative gradients in low-entropy smooth regions and amplifies the informative constraints along structural boundaries.

### 3.5 Two-Stage Decoupled Optimization

Networks exhibit spectral bias: low-frequency components are fitted quickly, while high-frequency details converge much more slowly. To prevent the optimization from stalling in low-frequency solutions or poor local minima, we decouple spectral anchoring from spatial refinement into two stages.

**Stage I: Identity Warm-up.** In the first stage, the network is isolated from spatial details and trained to reproduce the upsampled spectral base. We optimize the parameters $\theta$ for $T_{warm}$ iterations with SGD at a learning rate of 1.0 under the global spectral alignment loss:

$$\mathcal{L}_{warm}(\theta) = \|f_\theta(\mathcal{Y}_{\text{sub}\uparrow}, \mathcal{P}) - \mathcal{Y}_{\text{sub}\uparrow}\|_1. \tag{9}$$

The use of SGD is motivated by the optimization literature rather than engineering convenience: SGD is known to converge to flat regions of the loss landscape and to favor smooth, low-frequency solutions, which acts as an implicit low-pass regularizer Zhou et al. (2020). Stage I therefore anchors the spectral subspace quickly, before the refinement stage can fit spurious high-frequency content.

**Stage II: Spatial Refinement.** Starting from the warm-up weights, we switch to the Adam optimizer with a cosine annealing learning-rate schedule and minimize the joint objective:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{\text{spe}} + \beta\mathcal{L}_{\text{spa}}, \tag{10}$$

where $\mathcal{L}_{\text{spe}} = \|(\mathbf{B} \otimes \mathcal{X}) \downarrow - \mathcal{Y}\|_1$. In this stage the entropy gate retains only a sparse set of structural constraints, and the per-coordinate adaptive step sizes of Adam are well suited to such sparse, steep high-frequency gradients, making the refinement phase more efficient than isotropic SGD updates Wilson et al. (2017). The decoupling avoids the gradient conflict and lets the network reach a better Pareto balance between spectral and spatial fidelity.

## 4 Experiments

We evaluate the proposed framework on the HyPan and HMIF tasks. We report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and spectral angle mapper (SAM); relative dimensionless global error in synthesis (ERGAS), root mean square error (RMSE), and the end-to-end runtime are reported in the appendix (Tables 4 and 5). All methods are compared under the same zero-shot protocol, and the hyperparameters of the competing methods are tuned for a fair comparison.

**Implementation details.** The proposed network is a four-layer CNN with 64 hidden channels. The optimization uses 1,000 SGD warm-up iterations at a learning rate of 1.0, followed by 3,000 Adam

Table 1: Quantitative results on the HyPan task (Chikusei, Pavia, and Houston datasets). Bold: best; underline: second best. Time(s) is the end-to-end zero-shot runtime per image; ERGAS and RMSE are reported in Table 4 in the appendix.

| Method | Chikusei | | | | Pavia | | | | Houston | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | SAM↓ | Time(s)↓ | PSNR↑ | SSIM↑ | SAM↓ | Time(s)↓ | PSNR↑ | SSIM↑ | SAM↓ | Time(s)↓ |
| CNMF | 30.032 | 0.844 | 4.705 | **2.62** | 27.757 | 0.841 | 5.956 | **2.24** | 34.743 | 0.858 | 6.409 | **2.82** |
| LTTR | 31.304 | 0.830 | 4.562 | 99.01 | 32.091 | 0.888 | 5.312 | 68.94 | 36.973 | 0.919 | 5.927 | 113.34 |
| GTNN | 31.853 | 0.848 | 4.469 | 67.33 | 31.779 | 0.878 | 5.575 | 65.24 | <u>38.204</u> | 0.921 | 5.869 | 67.44 |
| CTDF | 29.662 | 0.838 | 4.614 | <u>8.23</u> | 28.456 | 0.835 | 6.566 | <u>7.95</u> | 35.760 | 0.898 | 6.647 | <u>8.97</u> |
| ZSL | <u>32.162</u> | <u>0.886</u> | <u>4.378</u> | 281.31 | 29.973 | 0.853 | 7.312 | 246.80 | 37.355 | 0.919 | 6.900 | 289.50 |
| CS2DIPs | 26.837 | 0.706 | 7.456 | 565.91 | 28.076 | 0.797 | 8.709 | 440.65 | 33.361 | 0.852 | 9.229 | 541.04 |
| PLRDiff | 29.567 | 0.802 | 5.116 | 94.01 | 33.341 | 0.905 | 5.073 | 94.01 | 36.914 | <u>0.924</u> | <u>5.608</u> | 93.90 |
| DM-ZS | 31.994 | 0.858 | 4.536 | 2181.79 | <u>33.838</u> | <u>0.909</u> | <u>4.824</u> | 1809.35 | 36.313 | 0.908 | 6.024 | 2447.88 |
| CLoRF | 31.143 | 0.847 | 4.869 | 891.88 | 31.609 | 0.861 | 5.618 | 777.75 | 34.416 | 0.863 | 8.880 | 925.38 |
| **Proposed** | **33.671** | **0.905** | **3.659** | 16.84 | **34.498** | **0.931** | **4.137** | 16.17 | **39.253** | **0.941** | **5.106** | 17.21 |

iterations with cosine annealing (initial learning rate $3 \times 10^{-3}$). All experiments are conducted on a single NVIDIA RTX 3080 GPU. An architecture study and the remaining ablations are given in Section 5.

### 4.1 Hyperspectral Pansharpening

We evaluate on the Chikusei, Pavia, and Houston datasets, each containing an HR-HSI of $256 \times 256$ pixels with 128, 93, and 144 bands, respectively. The LR-HSI is constructed by applying a $5 \times 5$ Gaussian blur kernel ($\sigma = 3$) followed by $4\times$ spatial downsampling; the PAN image is the spectral average of all bands. We compare against CNMF Yokoya et al. (2011), LTTR Dian et al. (2019), GTNN Dian et al. (2024), CTDF Xu et al. (2024), ZSL Dian et al. (2023), CS2DIPs Fang et al. (2024), PLRDiff Rui et al. (2024), DM-ZS Xiao et al. (2025), and CLoRF Wang et al. (2025).

Table 1 summarizes the results, including the per-image runtime. The proposed method achieves the best PSNR, SSIM, and SAM on all three datasets. Our method converges in about 16–17 seconds per image, two orders of magnitude faster than the diffusion-based methods and comparable to the fastest classical ones. Qualitative results are shown in Figure 4 (Appendix B);

### 4.2 Hyperspectral and Multispectral Image Fusion

We extend the evaluation to the HMIF task on the CAVE (44 image pairs) and Harvard (20 image pairs) datasets, using the same competing methods as in the HyPan experiments. Following the standard zero-shot protocol, the LR-HSI is produced with a $5 \times 5$ Gaussian kernel ($\sigma = 3$) and $4\times$ downsampling, and the high-resolution multispectral image (HR-MSI) is generated through the spectral response functions of the sensor.

Table 2 reports the results. On Harvard, the proposed method achieves the best score on all three metrics. On CAVE, it obtains the highest PSNR and the lowest RMSE (reported in the appendix), while its SAM and ERGAS trail the best competitor. This is expected for a purely optimization-based method: without downsampled training data or diffusion priors, the reconstruction is driven entirely by the observation losses, which yields a balanced trade-off between spectral fidelity and spatial sharpness. The end-to-end runtime is about 22 seconds per image (Table 2). Qualitative results are shown in Figure 5 (Appendix B); consistent with the HyPan experiments, the residuals are small in high-frequency regions and no spurious high-frequency artifacts are introduced.

## 5 Ablation Studies

We analyze the main components of the proposed framework: the entropy-based weighting, the decoupled warm-up, and the resulting optimization dynamics, together with a spectral and spatial frequency analysis of the reconstructions. A study of the network architecture is deferred to Appendix C to respect the page limit.

Table 2: Quantitative results on the HMIF task (CAVE and Harvard datasets). Bold: best; underline: second best. Time(s) is the end-to-end zero-shot runtime per image; ERGAS and RMSE are reported in Table 5 in the appendix.

| Method | CAVE (44 pairs) | | | | Harvard (20 pairs) | | | |
|---|---|---|---|---|---|---|---|---|
| | CAVE Dataset | | | | Harvard Dataset | | | |
| | PSNR↑ ± std | SSIM↑ ± std | SAM↓ ± std | Time(s)↓ | PSNR↑ ± std | SSIM↑ ± std | SAM↓ ± std | Time(s)↓ |
| CNMF | 40.431 ± 5.945 | 0.970 ± 0.024 | 5.048 ± 2.187 | **3.82** | 46.499 ± 5.203 | 0.962 ± 0.010 | 8.110 ± 1.922 | **4.50** |
| LTTR | 46.356 ± 3.834 | 0.987 ± 0.006 | 3.314 ± 1.202 | 27.31 | 47.977 ± 4.057 | 0.964 ± 0.010 | 8.058 ± 1.865 | 24.83 |
| GTNN | 46.162 ± 4.061 | 0.986 ± 0.006 | 3.544 ± 1.237 | 69.59 | 47.926 ± 4.037 | 0.964 ± 0.010 | 8.095 ± 1.848 | 63.23 |
| CTDF | 40.883 ± 6.197 | 0.947 ± 0.112 | 7.699 ± 7.236 | 5.50 | 44.582 ± 4.697 | 0.944 ± 0.037 | 9.966 ± 4.746 | 5.22 |
| ZSL | 48.077 ± 4.983 | **0.993 ± 0.005** | **3.074 ± 1.121** | 205.03 | 47.699 ± 4.390 | 0.964 ± 0.010 | 8.158 ± 1.873 | 203.49 |
| CS2DIPs | 39.776 ± 5.444 | 0.970 ± 0.019 | 7.998 ± 8.276 | 101.88 | 44.668 ± 4.831 | 0.956 ± 0.012 | 8.568 ± 1.787 | 106.53 |
| PLRDiff | 46.817 ± 5.644 | 0.968 ± 0.076 | 6.610 ± 10.799 | 89.67 | 45.138 ± 2.943 | 0.936 ± 0.049 | 12.346 ± 9.539 | 90.43 |
| DM-ZS | 48.769 ± 4.218 | 0.990 ± 0.006 | 3.937 ± 1.471 | 967.74 | 45.831 ± 3.930 | 0.959 ± 0.012 | 8.237 ± 1.788 | 974.64 |
| CLoRF | 38.036 ± 4.038 | 0.922 ± 0.038 | 13.469 ± 7.262 | 581.40 | 41.193 ± 3.583 | 0.926 ± 0.018 | 10.541 ± 2.905 | 578.47 |
| **Proposed** | **49.223 ± 4.515** | 0.987 ± 0.041 | 4.594 ± 11.065 | 22.96 | **48.689 ± 4.041** | **0.966 ± 0.009** | **8.007 ± 1.880** | 22.76 |

Table 3: Ablation of the spatial weighting mechanism on three datasets. Constant: tuned scalar weight; Dropout: dropout-based uncertainty map; Proposed: entropy-guided gate. Bold: best.

| | Chikusei | | Pavia | | Houston | |
|---|---|---|---|---|---|---|
| Mechanism | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| Constant Weighting | 33.5059 | 0.9021 | 34.3089 | 0.9282 | 38.9631 | 0.9383 |
| Dropout | 33.6232 | 0.9040 | **34.5524** | **0.9319** | 39.2308 | **0.9415** |
| **Proposed (Entropy)** | **33.6705** | **0.9049** | 34.4983 | 0.9313 | **39.2533** | 0.9414 |

### 5.1 Impact of Spatial Uncertainty Weighting

We first evaluate the spatial weighting mechanism on the three HyPan datasets, replacing the entropy weight with a tuned constant scalar. As Table 3 shows, removing the spatial adaptability lowers the reconstruction fidelity on Chikusei and Houston, although the constant-weight baseline remains competitive; the lightweight CNN and the two-stage initialization thus already provide a strong foundation, and the entropy map adds less than one second of computation.

Following Kendall & Gal (2017); Gal & Ghahramani (2016), we also test Monte Carlo (MC) dropout, which runs $T = 20$ stochastic forward passes after a constant-weight optimization to build a pixel-wise uncertainty map. It reaches accuracy close to the entropy gate, but requires the initial optimization plus twenty forward passes, more than doubling the runtime, whereas the entropy map provides the same guidance directly from the observation.

### 5.2 Impact of the Decoupled Warm-up Initialization

To evaluate the two-stage schedule, we track PSNR and SSIM against wall-clock time with and without the warm-up; the curves are reported in Figure 6 (Appendix D). Direct joint optimization stays at low values for a long time, because the network must fit low-frequency spectral content and high-frequency spatial details simultaneously and the spectral bias of convolutional architectures drives it toward a poor local minimum. With the warm-up, SGD anchors the spectral subspace first; after the switch to Adam, the network focuses on spatial structure and reaches higher final PSNR and SSIM.

### 5.3 Gradient Conflict Mitigation via Entropy Weighting

To examine the effect of the gate on the optimization dynamics, we track the cosine similarity between the gradients of the spatial and spectral losses during the optimization (bottom of Figure 1). With constant weighting (right), the gradients are roughly aligned at the beginning but collapse into strong conflicts later, with the similarity turning negative; this means that the spatial constraint degrades the established spectral fidelity. This baseline still converges in terms of the total loss: the two gradient directions cancel and the parameters settle at a compromise point. The trajectory shows why this equilibrium is unsatisfactory: the spatial constraint actively destroys the spectral fidelity established in the warm-up, so the compromise sacrifices one objective instead of balancing the two. With

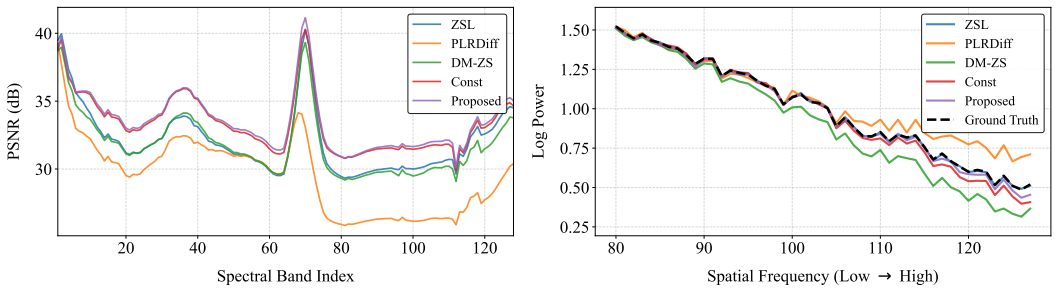



Figure 3: Evaluation on the Chikusei dataset from spectral and spatial perspectives. **Left:** band-wise PSNR across all channels. **Right:** 1D radial power spectrum in the high-frequency range (spatial frequency index $> 80$).

the entropy-guided weighting (left), the moving average (window size 10) of the cosine similarity stays positive throughout the optimization, because the gate concentrates the spatial gradients on informative high-entropy regions.

### 5.4 Spectral and Spatial Frequency Analysis

We deconstruct the reconstruction quality on Chikusei from two perspectives (Figure 3). From the spectral perspective (left), the proposed method maintains a consistent advantage across the whole spectral range: DM-ZS shows steep per-band drops in specific channels, consistent with spectral hallucination caused by its domain shift, and ZSL suffers from scale-specific overfitting, whereas our optimization keeps every band close to the fidelity constraint. From the spatial perspective (right), we compare the radial power spectrum in the high-frequency range, where the methods differ; in the low-frequency range the spectra nearly coincide. ZSL closely matches the high-frequency energy of the ground truth yet has a much lower PSNR, which shows that matching high-frequency energy statistically is not sufficient: the details must also be spatially aligned. Our method does not inject external high-frequency statistics, and the entropy gate suppresses gradients in homogeneous regions, so the reconstructed high-frequency content remains aligned with the true structure.

## 6 Conclusion and Limitations

We presented a fast zero-shot optimization framework for hyperspectral and multispectral image fusion, addressing the gradient conflict and the sub-optimal compromise equilibrium of conventional optimization-based zero-shot methods under the uniform scale assumption. The method extracts a structural heteroscedasticity map from the PAN/MSI observation using local Shannon entropy and uses it as an information gate that weights the spatial loss to remove the gradient conflict between the spatial and spectral objectives; the subspace projection and the decoupled two-stage optimization (an SGD warm-up followed by Adam) anchor the spectral subspace and reduce the spectral bias of zero-shot methods. On the HyPan and HMIF benchmarks, the method achieves state-of-the-art accuracy with an end-to-end runtime of about 17 seconds per image and no external training data.

**Limitations.** Because the method is iterative, very high-resolution inputs require considerable memory and time, limiting real-time use; the current shallow CNN also bounds the achievable accuracy, and stronger architectures may improve both.

### AI use statement

In this work, we used generative AI tools for the following tasks with recommended disclosure: language polishing, i.e., editing the manuscript to improve readability and clarity, and checking the formatting and internal consistency of the manuscript, such as notation usage and reference formatting. We have not used generative AI tools for any task with required disclosure: we did not use AI to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to propose or refine hypotheses, to design or provide

feedback on the research methodology or experiments, to implement methods, to perform translation, to clean or reformat datasets, or to conduct qualitative or thematic data analysis or interpret results. All scientific content, technical decisions, and interpretations in this paper were made by the authors. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including its text, claims, and artifacts.

## REFERENCES


Qi Cao, Liang-Jian Deng, Wu Wang, Junming Hou, and Gemine Vivone. Zero-shot semi-supervised learning for pansharpening. *Inf. Fusion*, 101:102001, 2024.

L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 59(8):6995–7010, 2020.

R. Dian, A. Guo, and S. Li. Zero-shot hyperspectral sharpening. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12650–12666, 2023.

Renwei Dian, Shutao Li, and Leyuan Fang. Learning a low tensor-train rank representation for hyperspectral image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(9):2672–2683, 2019.

Renwei Dian, Yuanye Liu, and Shutao Li. Hyperspectral image fusion via a novel generalized tensor nuclear norm regularization. *IEEE Trans. Neural Netw. Learn. Syst.*, 36(4):7437–7448, 2024.

Yuan Fang, Yipeng Liu, Chong-Yung Chi, Zhen Long, and Ce Zhu. CS2DIPs: Unsupervised hsi super-resolution using coupled spatial and spectral dips. *IEEE Trans. Image Process.*, 33: 3090–3101, 2024. doi: 10.1109/TIP.2024.3390582.

Haoying Fu, Michael K Ng, Mila Nikolova, and Jesse L Barlow. Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. *SIAM J. Sci. Comput.*, 27(6):1881–1902, 2006.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hao Jiang, Qing Zhang, Yongwei Nie, Lei Zhu, and Wei-Shi Zheng. Learning multi-scale deep image prior for high-quality unsupervised image denoising. *Comput. Graph. Forum*, 41(7):323–334, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.*, 30, 2017.

P Kwarteng and A Chavez. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogramm. Eng. Remote Sens*, 55(1):339–348, 1989.

Charis Lanaras, Emmanuel Baltsavias, and Konrad Schindler. Hyperspectral super-resolution by coupled spectral unmixing. In *Int. Conf. Comput. Vis.*, pp. 3586–3594, 2015.

Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. Deep image prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9446–9454, 2018.

Jiaxin Li, Ke Zheng, Lianru Gao, Zhu Han, Zhi Li, and Jocelyn Chanussot. Enhanced deep image prior for unsupervised hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.*, 63:1–18, 2025.

Ye Liu, Wubin Gao, Youchen Chen, Fan Lyu, and Xing Wang. Low-level continual test-time adaptation for image restoration. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9032–9036. IEEE, 2026.

Hailiang Lu, Mercedes E Paoletti, Juan M Haut, Sergio Moreno-Álvarez, Guangsheng Chen, and Weipeng Jing. A novel spectral-spatial attention network for zero-shot pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 18:19805–19820, 2025.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 2002.

Yongwei Nie, Mingxian Fan, Chengjiang Long, Qing Zhang, Jian Zhu, and Xuemiao Xu. Incorporating test-time optimization into training with dual networks for human mesh recovery. *Adv. Neural Inf. Process. Syst.*, 37:11634–11663, 2024.

L. Pang, X. Rui, L. Cui, H. Wang, D. Meng, and X. Cao. HIR-Diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 3005–3014, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

X. Rui, X. Cao, L. Pang, Z. Zhu, Z. Yue, and D. Meng. Unsupervised hyperspectral pansharpening via low-rank diffusion model. *Inf. Fusion*, 107:102325, 2024.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.*, 65(13):3551–3582, 2017.

Oleksii Sidorov and Jon Yngve Hardeberg. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *Int. Conf. Comput. Vis. Workshops*, pp. 3844–3851, 2019.

Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.*, 31:5134–5149, 2022.

X. Tian, Y. Chen, C. Yang, and J. Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16, 2021.

Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Trans. Geosci. Remote Sens.*, 53(5):2565–2586, 2014.

Hebaixu Wang, Hao Zhang, Xin Tian, and Jiayi Ma. Zero-sharpen: A universal pansharpening method across satellites for reducing scale-variance gap via zero-shot variation. *Inf. Fusion*, 101: 102003, 2024.

Ting Wang, Zipei Yan, Jizhou Li, Xile Zhao, Chao Wang, and Michael Ng. Hyperspectral and multispectral image fusion with arbitrary resolution through self-supervised representations. *Int. J. Comput. Vis.*, pp. 1–21, 2025.

Zhe-Min Wei, Heng-Chao Li, Yu-Bang Zheng, Jian-Li Wang, and Qian Du. Uladiff: Unmixing-guided learnable abundance-latent diffusion for hyperspectral image denoising. *IEEE Trans. Geosci. Remote Sens.*, 63:1–15, 2025.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

J.-L. Xiao, T.-Z. Huang, L.-J. Deng, G. Lin, Z. Cao, C. Li, and Q. Zhao. Hyperspectral pansharpening via diffusion models with iteratively zero-shot guidance. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 12669–12678, 2025.

Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, Jin-Liang Xiao, Clifford Broni-Bediako, Junshi Xia, and Naoto Yokoya. A coupled tensor double-factor method for hyperspectral and multispectral image fusion. *IEEE Trans. Geosci. Remote Sens.*, 62:1–17, 2024.

N. Yokoya, T. Yairi, and A. Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Trans. Geosci. Remote Sens.*, 50(2):528–537, 2011.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Adv. Neural Inform. Process. Syst.*, 33:5824–5836, 2020.

Jiaxuan Zhao, Lingling Li, Licheng Jiao, Life Mengru Ma, Xu Liu, Chao Wang, Fang Liu, Wenping Ma, and Shuyuan Yang. Modality-prior aware progressive diffusion pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 64:1–15, 2026a.

Zuopeng Zhao, Ying Liu, Xiaoyu Li, Su Luo, Lu Li, and Wenwen Liu. Anyband-diff: A unified remote sensing image generation and band repair framework with spectral priors. *arXiv preprint arXiv:2605.14341*, 2026b.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Adv. Neural Inform. Process. Syst.*, 33:21285–21296, 2020.

## A Additional Quantitative Results

Tables 4 and 5 report the complete quantitative results for the two tasks, including ERGAS, RMSE, and the end-to-end runtime. Bold: best; underline: second best. Time(s) is the end-to-end zero-shot execution time per image.

## B Qualitative Comparison on HyPan and HMIF

Figure 4 and Figure 5 show the qualitative comparison on the HyPan and HMIF tasks.

For the HyPan, the entropy map computed from the PAN image highlights the regions of fine structure, where the competing methods leave the largest errors. By amplifying the spatial loss in these high-entropy regions and attenuating it in smooth ones, the network reconstructs sharp boundaries while homogeneous areas stay clean and consistent, which yields the smallest residual errors among all methods.

## C Impact of the Network Architecture

The network architecture acts as an implicit prior in such zero-shot optimization, so we compare five configurations on the CAVE dataset. The proposed shallow CNN uses four convolutional layers with 64 hidden channels and follows the two-stage schedule (1,000 SGD iterations at learning rate $1.0$, then 3,000 Adam iterations with cosine annealing starting at $3 \times 10^{-3}$). The alternatives are: (1) a U-Net Ronneberger et al. (2015), a deep encoder-decoder with skip connections; (2) a ResNet He et al. (2016) with five residual blocks; (3) a two-layer CNN with 16 hidden channels (Tiny CNN); and (4) direct optimization of the pixel tensor (Direct Tensor). The Adam learning rate is reduced to $1 \times 10^{-3}$ for the U-Net, the ResNet, and the Direct Tensor, whose loss landscapes are more difficult to optimize, and raised to $1 \times 10^{-2}$ for the Tiny CNN; other hyperparameters are kept fixed.

As Table 6 shows, a deeper network does not necessarily help in this zero-shot setting. The Direct Tensor is the fastest but the least accurate: optimizing 786.43 K parameters in the unconstrained pixel space lacks the implicit low-pass filtering that convolutional structure provides. The deep networks (U-Net, 458.38 K parameters; ResNet, 27.59 G FLOPs) also underperform: without training data, their excess capacity overfits the noisy components of the single degraded observation, and the optimization ends in a poor local minimum. The Tiny CNN is fast and stable, but its 3.92 K parameters cannot represent the full spectral-spatial mapping. The proposed shallow CNN offers enough capacity for high-frequency details while keeping the loss landscape easy to optimize; with the SGD warm-up it reaches the best fidelity in a reasonable time.

Table 4: Complete quantitative results on the HyPan task (Chikusei, Pavia, and Houston datasets), including ERGAS, RMSE, and runtime.

| **Chikusei Dataset** | | | | | | |
|---|---|---|---|---|---|---|
| Method | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | RMSE↓ | Time(s)↓ |
| CNMF | 30.032 | 0.844 | 4.705 | 6.601 | 0.0343 | **2.62** |
| LTTR | 31.304 | 0.830 | 4.562 | 6.631 | 0.0294 | 99.01 |
| GTNN | 31.853 | 0.848 | 4.469 | 5.762 | 0.0281 | 67.33 |
| CTDF | 29.662 | 0.838 | 4.614 | 6.731 | 0.0361 | 8.23 |
| ZSL | 32.162 | 0.886 | 4.378 | 5.684 | 0.0264 | 281.31 |
| CS2DIPs | 26.837 | 0.706 | 7.456 | 10.776 | 0.0487 | 565.91 |
| PLRDiff | 29.567 | 0.802 | 5.116 | 6.732 | 0.0366 | 94.01 |
| DM-ZS | 31.994 | 0.858 | 4.536 | 5.739 | 0.0271 | 2181.79 |
| CLoRF | 31.143 | 0.847 | 4.869 | 6.763 | 0.0290 | 891.88 |
| **Proposed** | **33.671** | **0.905** | **3.659** | **4.842** | **0.0221** | 16.84 |
| **Pavia Dataset** | | | | | | |
| Method | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | RMSE↓ | Time(s)↓ |
| CNMF | 27.757 | 0.841 | 5.956 | 6.511 | 0.0427 | **2.24** |
| LTTR | 32.091 | 0.888 | 5.312 | 3.629 | 0.0277 | 68.94 |
| GTNN | 31.779 | 0.878 | 5.575 | 3.758 | 0.0287 | 65.24 |
| CTDF | 28.456 | 0.835 | 6.566 | 5.559 | 0.0398 | 7.95 |
| ZSL | 29.973 | 0.853 | 7.312 | 4.742 | 0.0352 | 246.80 |
| CS2DIPs | 28.076 | 0.797 | 8.709 | 5.922 | 0.0434 | 440.65 |
| PLRDiff | 33.341 | 0.905 | 5.073 | 3.227 | 0.0231 | 94.01 |
| DM-ZS | 33.838 | 0.909 | 4.824 | 2.998 | 0.0220 | 1809.35 |
| CLoRF | 31.609 | 0.861 | 5.618 | 3.810 | 0.0285 | 777.75 |
| **Proposed** | **34.498** | **0.931** | **4.137** | **2.760** | **0.0203** | 16.17 |
| **Houston Dataset** | | | | | | |
| Method | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | RMSE↓ | Time(s)↓ |
| CNMF | 34.743 | 0.858 | 6.409 | 5.388 | 0.0218 | **2.82** |
| LTTR | 36.973 | 0.919 | 5.927 | 6.622 | 0.0151 | 113.34 |
| GTNN | 38.204 | 0.921 | 5.869 | 3.651 | 0.0147 | 67.44 |
| CTDF | 35.760 | 0.898 | 6.647 | 4.889 | 0.0193 | 8.97 |
| ZSL | 37.355 | 0.919 | 6.900 | 4.040 | 0.0156 | 289.50 |
| CS2DIPs | 33.361 | 0.852 | 9.229 | 9.524 | 0.0225 | 541.04 |
| PLRDiff | 36.914 | 0.924 | 5.608 | 4.367 | 0.0158 | 93.90 |
| DM-ZS | 36.313 | 0.908 | 6.024 | 4.482 | 0.0176 | 2447.88 |
| CLoRF | 34.416 | 0.863 | 8.880 | 5.363 | 0.0171 | 925.38 |
| **Proposed** | **39.253** | **0.941** | **5.106** | **3.261** | **0.0123** | 17.21 |
| Ideal value | $+\infty$ | 1 | 0 | 0 | 0 | 0 |

## D Additional Convergence and Qualitative Results

**Convergence of the two-stage schedule.** Figure 6 tracks PSNR (left) and SSIM (right) against wall-clock time on CAVE with and without the warm-up. Direct joint optimization from a random initialization (blue dashed) improves both metrics slowly and stays at low values for a long time, because the network must fit low-frequency spectral content and high-frequency spatial details simultaneously, and the spectral bias of convolutional architectures drives it toward a poor local minimum. With the warm-up (solid green), SGD anchors the spectral subspace first, without interference from the spatial refinement constraint; after the switch to Adam (solid red), the network focuses on spatial structure. The decoupled schedule thus raises PSNR and SSIM quickly and reaches higher final values than direct joint optimization.

Table 5: Complete quantitative results on the HMIF task (CAVE and Harvard datasets), including ERGAS, RMSE, and runtime.

| **CAVE Dataset (44 Image Pairs)** | | | | | | |
|---|---|---|---|---|---|---|
| Method | PSNR↑ ± std | SSIM↑ ± std | SAM↓ ± std | ERGAS↓ ± std | RMSE↓ ± std | Time(s)↓ |
| CNMF | 40.431 ± 5.945 | 0.970 ± 0.024 | 5.048 ± 2.187 | 4.463 ± 3.093 | 0.0126 ± 0.0072 | **3.82** |
| LTTR | 46.356 ± 3.834 | 0.987 ± 0.006 | 3.314 ± 1.202 | 2.132 ± 1.143 | 0.0064 ± 0.0022 | 27.31 |
| GTNN | 46.162 ± 4.061 | 0.986 ± 0.006 | 3.544 ± 1.237 | 2.166 ± 1.186 | 0.0065 ± 0.0025 | 69.59 |
| CTDF | 40.883 ± 6.197 | 0.947 ± 0.112 | 7.699 ± 7.236 | 4.890 ± 7.213 | 0.0146 ± 0.0220 | 5.50 |
| ZSL | 48.077 ± 4.983 | **0.993 ± 0.005** | **3.074 ± 1.121** | 2.060 ± 1.454 | 0.0057 ± 0.0042 | 205.03 |
| CS2DIPs | 39.776 ± 5.444 | 0.970 ± 0.019 | 7.998 ± 8.276 | 4.466 ± 2.251 | 0.0145 ± 0.0084 | 101.88 |
| PLRDiff | 46.817 ± 5.644 | 0.968 ± 0.076 | 6.610 ± 10.799 | 4.685 ± 8.267 | 0.0100 ± 0.0150 | 89.67 |
| DM-ZS | 48.769 ± 4.218 | 0.990 ± 0.006 | 3.937 ± 1.471 | **1.719 ± 1.058** | 0.0047 ± 0.0024 | 967.74 |
| CLoRF | 38.036 ± 4.038 | 0.922 ± 0.038 | 13.469 ± 7.262 | 5.773 ± 3.316 | 0.0160 ± 0.0081 | 581.40 |
| **Proposed** | **49.223 ± 4.515** | 0.987 ± 0.041 | 4.594 ± 11.065 | 2.134 ± 4.321 | **0.0045 ± 0.0027** | 22.96 |
| **Harvard Dataset (20 Image Pairs)** | | | | | | |
| Method | PSNR↑ ± std | SSIM↑ ± std | SAM↓ ± std | ERGAS↓ ± std | RMSE↓ ± std | Time(s)↓ |
| CNMF | 46.499 ± 5.203 | 0.962 ± 0.010 | 8.110 ± 1.922 | 14.143 ± 4.402 | 0.0221 ± 0.0155 | **4.50** |
| LTTR | 47.977 ± 4.057 | 0.964 ± 0.010 | 8.058 ± 1.865 | 14.144 ± 4.429 | 0.0219 ± 0.0154 | 24.83 |
| GTNN | 47.926 ± 4.037 | 0.964 ± 0.010 | 8.095 ± 1.848 | 14.152 ± 4.428 | 0.0219 ± 0.0154 | 63.23 |
| CTDF | 44.582 ± 4.697 | 0.944 ± 0.037 | 9.966 ± 4.746 | 16.439 ± 7.906 | 0.0234 ± 0.0145 | 5.22 |
| ZSL | 47.699 ± 4.390 | 0.964 ± 0.010 | 8.158 ± 1.873 | 14.138 ± 4.416 | 0.0218 ± 0.0155 | 203.49 |
| CS2DIPs | 44.668 ± 4.831 | 0.956 ± 0.012 | 8.568 ± 1.787 | 14.328 ± 4.443 | 0.0234 ± 0.0152 | 106.53 |
| PLRDiff | 45.138 ± 2.943 | 0.936 ± 0.049 | 12.346 ± 9.539 | 15.702 ± 5.991 | 0.0262 ± 0.0127 | 90.43 |
| DM-ZS | 45.831 ± 3.930 | 0.959 ± 0.012 | 8.237 ± 1.788 | 14.249 ± 4.313 | 0.0223 ± 0.0149 | 974.64 |
| CLoRF | 41.193 ± 3.583 | 0.926 ± 0.018 | 10.541 ± 2.905 | 16.568 ± 5.276 | 0.0246 ± 0.0143 | 578.47 |
| **Proposed** | **48.689 ± 4.041** | **0.966 ± 0.009** | **8.007 ± 1.880** | **14.116 ± 4.435** | **0.0216 ± 0.0155** | 22.76 |
| Ideal value | $+\infty$ | 1 | 0 | 0 | 0 | 0 |

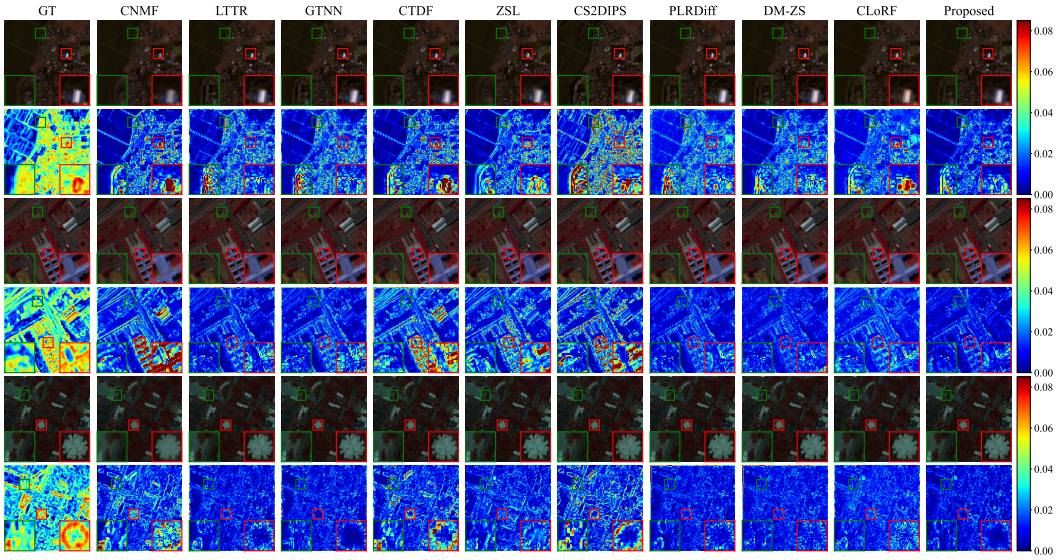



Figure 4: Qualitative comparison on the HyPan task. **Top:** reconstructed fused images. **Bottom:** the image directly below the ground truth (GT) shows the spatial entropy map computed from the PAN image, which serves as the guidance prior; the remaining images are the error residual maps of the competing methods.

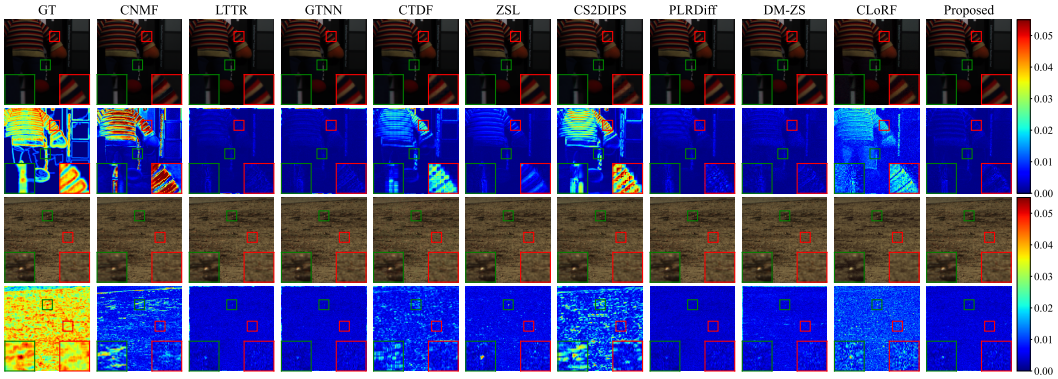



Figure 5: Qualitative comparison on the HMIF task with the CAVE (top two rows) and Harvard (bottom two rows) datasets. For each dataset, the fused images are shown with magnified structural details. The image directly below the ground truth (GT) shows the spatial entropy map, and the remaining images are the error residual maps of the competing methods.

Table 6: Architecture ablation on the CAVE dataset ($256 \times 256$ spatial resolution).

| Architecture | Params (K) | FLOPs (G) | PSNR↑ | SSIM↑ | Time(s)↓ |
|---|---|---|---|---|---|
| Direct Tensor | 786.43 | 0.0008 | 38.254 | 0.9861 | **5.61** |
| Tiny CNN | 3.92 | 0.2548 | 46.294 | 0.9923 | 6.21 |
| Complex U-Net | 458.38 | 10.1481 | 40.366 | 0.9942 | 25.97 |
| Deep ResNet | 421.84 | 27.5943 | 42.382 | 0.9949 | 69.81 |
| **CNN (Proposed)** | 155.02 | 10.1460 | **51.576** | **0.9984** | 23.50 |

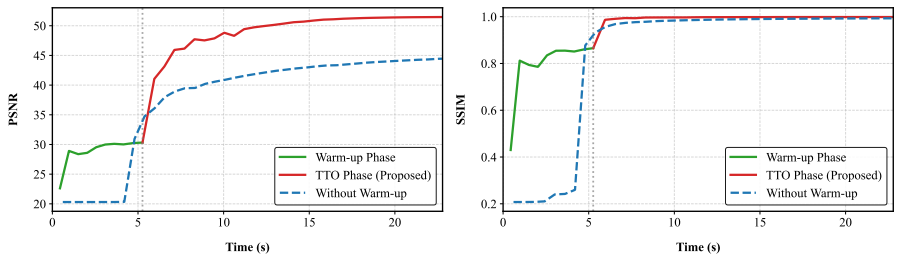



Figure 6: Convergence on the CAVE dataset in PSNR (left) and SSIM (right) against wall-clock time. The decoupled schedule (green: warm-up; red: refinement) improves both metrics quickly, while direct joint optimization (blue dashed) stays at low values for a long time; the final PSNR and SSIM are also higher.