ADAPTIVE WAVELET ATTENTION-GUIDED TENSOR COMPLETION WITH FUSED LOW-RANKNESS AND SMOOTHNESS FOR HYPERSPECTRAL IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank tensor completion is widely used to recover missing data in hyperspectral images. However, existing methods often rely on fixed low-rank models and hand-crafted priors, which may fail to preserve spatial and spectral structures when large amounts of data are missing. To address this issue, we propose Adaptive Wavelet Attention-guided Smooth Tensor Completion (AWSTC), an adaptive wavelet attention-guided deep unfolding framework for recovering missing entries in hyperspectral image tensors by jointly exploiting low-rankness and smoothness. AWSTC models tensor completion as a low-rank factorization problem with fused smoothness regularization. Its adaptive wavelet attention module learns the importance of different channels from wavelet subbands to guide low-rank recovery. The resulting optimization algorithm is unfolded into a multi-stage network that combines model-based updates with learnable residual dense blocks. Experiments on three hyperspectral image datasets show that AWSTC performs well across various missing rates, producing clearer images with fewer stripe artifacts.

1 INTRODUCTION

Hyperspectral images (HSIs) record scenes across many spectral bands and provide rich information for remote sensing Goetz (2009); Ghamisi et al. (2017). Sensor failures and limited sampling can leave entries missing, making HSI completion necessary Zeng et al. (2020); Wang et al. (2021). Since spatial and spectral correlations should be modeled jointly, an HSI is naturally represented as a third-order tensor.

HSIs exhibit strong correlations across bands, while neighboring pixels are generally smooth. These properties motivate low-rankness and smoothness as complementary priors. Low-rank tensor completion (TC) exploits spectral and spatial redundancy Liu et al. (2012); Zhou et al. (2017); Zhang et al. (2020); Kong et al. (2023), but may lose local details when many entries are missing. Combining low-rank models with total variation (TV) helps preserve these details Wang et al. (2017); Fan et al. (2018); Zeng et al. (2020). As shown in Figure 1, the concentrated gradient distributions and decaying singular-value spectra motivate our use of low-rank factorization on directional gradients. However, fixed regularizers may not adapt well to different HSI structures.

Deep learning can learn spatial–spectral features for TC Chai et al. (2021); Zhao et al. (2020), but purely learned recovery can be difficult to interpret and may depend heavily on the training data. Deep unfolding combines learning with an optimization model by turning its iterations into trainable stages Dutta et al. (2024). Existing methods have unfolded low-rank TC Mai et al. (2022) and incorporated attention into tensor recovery Mai et al. (2024).

Motivated by these observations, we propose AWSTC, an adaptive wavelet attention-guided deep unfolding framework that jointly exploits low-rankness and smoothness for HSI tensor completion. The model couples low-rank factorization with directional gradients to capture correlated structures. Its adaptive wavelet attention (AWA) module uses a learnable wavelet-like transform to extract low-frequency structure and directional high-frequency details. These features provide channel-wise weights for the gradient-domain low-rank fitting term. As illustrated in Figure 2, AWSTC unfolds

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.

the optimization algorithm into K stages, combining model-based updates with learnable residual dense blocks to refine local details. The AWA weights are shared across stages to guide reconstruction.

The main contributions of this work are summarized as follows:

(1) We introduce an adaptive wavelet attention (AWA) module that jointly captures low-frequency structural information and directional high-frequency details. The resulting channel-wise weights adaptively guide low-rank tensor recovery by emphasizing informative spatial–spectral features.

(2) We integrate fused smoothness regularization into a deep unfolding framework for low-rank tensor completion. This design combines interpretable model-based updates with learnable proximal mappings, improving spatial–spectral coherence and suppressing reconstruction artifacts.

(3) We conduct experiments on multiple HSI datasets under a wide range of missing rates. The results demonstrate that AWSTC consistently achieves competitive or superior performance compared with representative baselines.

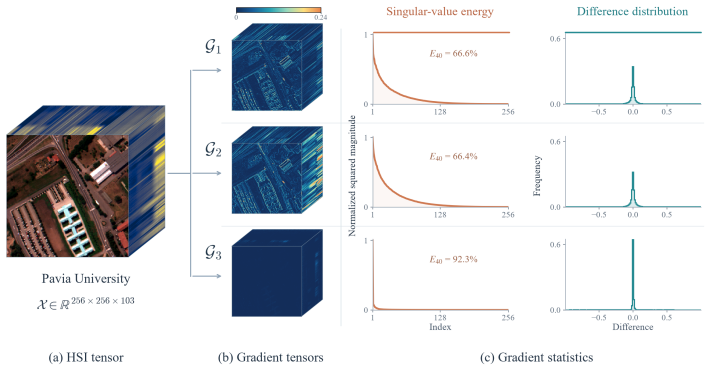


Figure 1: Structural properties of a Pavia University crop.

2 Preliminaries

In this paper, scalars, vectors, matrices, and tensors are denoted by lowercase letters (e.g., x), bold lowercase letters (e.g., $\mathbf{x}$), uppercase letters (e.g., X or $\mathbf{X}$), and calligraphic uppercase letters (e.g., $\mathcal{X}$), respectively. Basic tensor concepts and operations are introduced in Appendix A (Kilmer & Martin, 2011; Lu et al., 2020).

Definition 2.1 (Gradient tensor (Wang et al., 2023)) *For $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the mode-$k$ gradient tensor is*

$$\mathcal{G}_k := \nabla_k(\mathcal{X}) = \mathcal{X} \times_k \mathbf{D}_{I_k}, \qquad k = 1, 2, 3, \tag{1}$$

where $\mathbf{D}_{I_k}$ is the periodic finite-difference matrix and $\times_k$ is the mode-k product. The boundary convention is detailed in Appendix A.

Definition 2.2 (Tensor correlated total variation (Wang et al., 2023)) *For a nonempty set of modes $\Gamma \subseteq \{1, 2, 3\}$, tensor correlated total variation is*

$$\|\mathcal{X}\|_{\text{t-CTV}} := \frac{1}{\gamma} \sum_{k \in \Gamma} \|\mathcal{G}_k\|_{\circledast}, \qquad \gamma = |\Gamma|. \tag{2}$$

3 Proposed Method

We begin with the basic robust tensor completion model:

$$\min_{\mathcal{X}, \mathcal{E}} \operatorname{rank}(\mathcal{X}) + \|\mathcal{E}\|_0 \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathcal{X} + \mathcal{E}) = \mathcal{P}_\Omega(\mathcal{D}), \tag{3}$$

where $\mathcal{D}, \mathcal{X}, \mathcal{E} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ denote the incomplete observations, the tensor to be recovered, and sparse errors, respectively. Here, $\operatorname{rank}(\cdot)$ denotes a tensor rank measure, $\|\cdot\|_0$ counts nonzero entries, and $\mathcal{P}_\Omega$ retains observed entries indexed by Ω and zeros the rest.

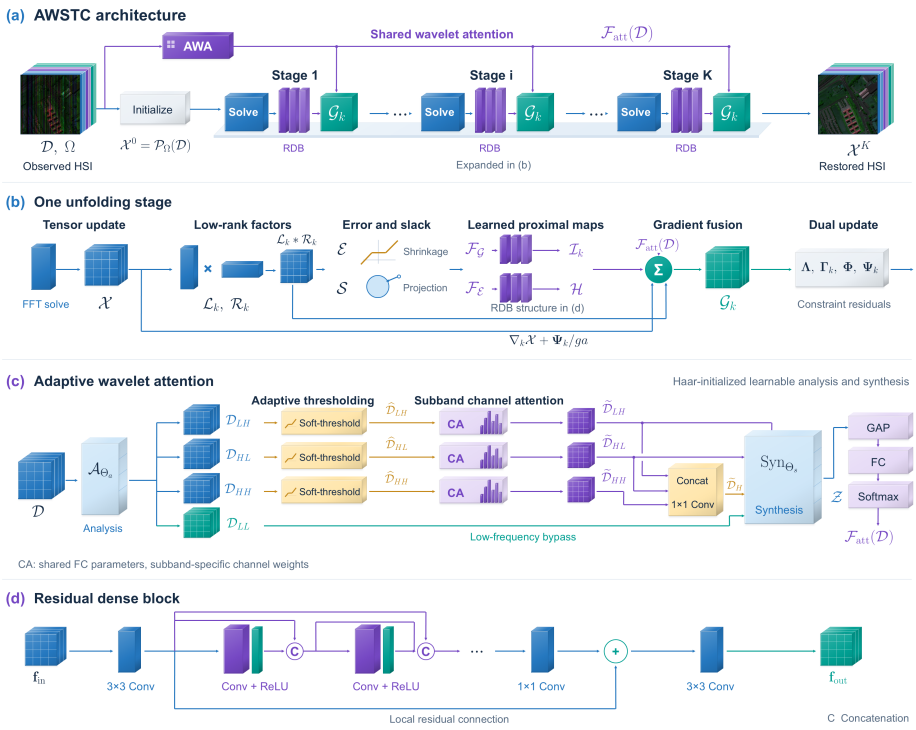


Figure 2: Overview of the proposed AWSTC for hyperspectral image restoration.

We then introduce a smoothness prior and use an ℓ_1 penalty for sparse errors, yielding

$$\min \frac{1}{\gamma} \sum_{k\in\Gamma} \|\mathcal{G}_k\|_{\circledast} + \lambda\|\mathcal{E}\|_1 \quad \text{s.t.} \quad \mathcal{G}_k = \nabla_k(\mathcal{X}), \quad \mathcal{P}_\Omega(\mathcal{X}+\mathcal{E}) = \mathcal{P}_\Omega(\mathcal{D}) \tag{4}$$

where $\mathcal{G}_k = \nabla_k(\mathcal{X})$ are gradient tensors, Γ contains the selected modes, and $\lambda > 0$ balances the sparse-error penalty.

Then we use tensor factorization to model low-rank structure. For each $k \in \Gamma$, we approximate $\mathcal{G}_k$ by $\mathcal{L}_k * \mathcal{R}_k$, where $\mathcal{L}_k \in \mathbb{R}^{I_1\times r\times I_3}$ and $\mathcal{R}_k \in \mathbb{R}^{r\times I_2\times I_3}$. The fixed rank parameter $1 \le r \le \min(I_1, I_2)$ limits the tubal rank of this product to at most r. Replacing the nuclear-norm term with the squared fitting error gives the following alternative model:

$$\min \frac{1}{2\gamma} \sum_{k\in\Gamma} \|\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k\|_F^2 + \lambda\|\mathcal{E}\|_1 \quad \text{s.t.} \quad \mathcal{G}_k = \nabla_k(\mathcal{X}), \quad \mathcal{P}_\Omega(\mathcal{X}+\mathcal{E}) = \mathcal{P}_\Omega(\mathcal{D}) \tag{5}$$

We then add attention weights $\mathcal{F}_{\text{att}}(\mathcal{D})$ to adjust the contribution of each channel:

$$\min \frac{1}{2\gamma} \sum_{k\in\Gamma} \|\sqrt{\mathcal{F}_{\text{att}}(\mathcal{D})} \odot (\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\|_F^2 + \lambda\|\mathcal{E}\|_1 \text{ s.t. } \mathcal{G}_k = \nabla_k(\mathcal{X}), \mathcal{P}_\Omega(\mathcal{X}+\mathcal{E}) = \mathcal{P}_\Omega(\mathcal{D}) \tag{6}$$

3.1 Adaptive Wavelet Attention

AWA produces channel weights $\mathcal{F}_{\text{att}}(\mathcal{D})$ shared across unfolding stages (Figure 2(c)). Its analysis and synthesis filters are initialized with Haar bases and then learned without an orthogonality or perfect-reconstruction constraint. With $h_0 = [1, 1]/\sqrt{2}$ and $h_1 = [-1, 1]/\sqrt{2}$, the initial two-dimensional filters are

$$\mathbf{K}_{LL} = h_0^\top h_0, \quad \mathbf{K}_{LH} = h_0^\top h_1, \quad \mathbf{K}_{HL} = h_1^\top h_0, \quad \mathbf{K}_{HH} = h_1^\top h_1. \tag{7}$$

Learnable analysis $\mathcal{A}_{\Theta_a}$ decomposes each channel into a low-frequency component and three directional high-frequency components:

$$(\mathcal{D}_{LH}, \mathcal{D}_{HL}, \mathcal{D}_{HH}, \mathcal{D}_{LL}) = \mathcal{A}_{\Theta_a}(\mathcal{D}), \tag{8}$$

For $s \in \{LH, HL, HH\}$, a learned threshold ratio and spatial global average pooling (GAP) suppress small coefficients:

$$\tau_s = \sigma(\boldsymbol{\theta}_s) \odot \mathrm{GAP}(|\mathcal{D}_s|), \qquad \widehat{\mathcal{D}}_s = \mathrm{sign}(\mathcal{D}_s) \odot \mathrm{ReLU}(|\mathcal{D}_s| - \tau_s), \tag{9}$$

Here σ is the sigmoid and $\boldsymbol{\theta}_s$ is learned separately for each subband and channel. The shared fully connected mapping ϕ then computes subband-specific channel weights:

$$\mathbf{a}_s = \mathrm{Softmax}\left(\phi(\mathrm{GAP}(\widehat{\mathcal{D}}_s))\right), \qquad \widetilde{\mathcal{D}}_s = \mathbf{a}_s \odot \widehat{\mathcal{D}}_s, \quad s \in \{LH, HL, HH\}, \tag{10}$$

The attended high-frequency subbands are fused by a 1×1 convolution:

$$\widetilde{\mathcal{D}}_H = \mathrm{Conv}_{1\times1}\left([\widetilde{\mathcal{D}}_{LH}, \widetilde{\mathcal{D}}_{HL}, \widetilde{\mathcal{D}}_{HH}]\right). \tag{11}$$

The fused feature $\widetilde{\mathcal{D}}_H$ replaces the third high-frequency input of the learnable grouped-transposed-convolution synthesis; it is not interpreted as a directional HH band. Synthesis gives

$$\mathcal{Z} = \mathrm{Syn}_{\Theta_s}\left(\widetilde{\mathcal{D}}_{LH}, \widetilde{\mathcal{D}}_{HL}, \widetilde{\mathcal{D}}_H, \mathcal{D}_{LL}\right), \tag{12}$$

Finally, GAP, ϕ, and Softmax produce the channel weights

$$\mathcal{F}_{\mathrm{att}}(\mathcal{D}) = \mathrm{Softmax}\left(\phi(\mathrm{GAP}(\mathcal{Z}))\right). \tag{13}$$

These positive relative weights have size $C \times 1 \times 1$ and are broadcast spatially in the low-rank fitting term. A larger value penalizes the corresponding channel residual more strongly.

3.2 Optimization

After obtaining $\mathcal{F}_{\mathrm{att}}(\mathcal{D})$, we solve the model by the augmented Lagrangian method. We add learned regularizers $g_{\mathcal{G}}$ and $h_{\mathcal{E}}$ for the gradient and sparse error tensors, and use a slack tensor $\mathcal{S}$ to account for bounded noise on observed entries:

$$\begin{aligned} \min\ & \frac{1}{2\gamma}\sum_{k\in\Gamma}\left\|\sqrt{\mathcal{F}_{\mathrm{att}}(\mathcal{D})} \odot (\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\right\|_F^2 + \lambda\|\mathcal{E}\|_1 + \sum_{k\in\Gamma} g_{\mathcal{G}}(\mathcal{G}_k) + h_{\mathcal{E}}(\mathcal{E}) \\ \text{s.t. } & \mathcal{G}_k = \nabla_k(\mathcal{X}),\ \mathcal{X} + \mathcal{E} + \mathcal{S} = \mathcal{P}_\Omega(\mathcal{D}),\ \|\mathcal{P}_\Omega(\mathcal{S})\|_F \le \delta. \end{aligned} \tag{14}$$

We introduce $\mathcal{I}_k = \mathcal{G}_k$ and $\mathcal{H} = \mathcal{E}$ to separate the learned regularizers. Omitting constant terms, the augmented Lagrangian is

$$\begin{aligned} \mathcal{L} = & \frac{1}{2\gamma}\sum_{k\in\Gamma}\left\|\sqrt{\mathcal{F}_{\mathrm{att}}(\mathcal{D})} \odot (\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\right\|_F^2 + \lambda\|\mathcal{E}\|_1 + \sum_{k\in\Gamma} g_{\mathcal{G}}(\mathcal{I}_k) + h_{\mathcal{E}}(\mathcal{H}) \\ & + \frac{\mu}{2}\left\|\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S} + \tfrac{\mathbf{\Lambda}}{\mu}\right\|_F^2 + \frac{\alpha}{2}\sum_{k\in\Gamma}\left\|\mathcal{G}_k - \mathcal{I}_k + \tfrac{\mathbf{\Gamma}_k}{\alpha}\right\|_F^2 \\ & + \frac{\beta}{2}\left\|\mathcal{E} - \mathcal{H} + \tfrac{\mathbf{\Phi}}{\beta}\right\|_F^2 + \frac{\rho}{2}\sum_{k\in\Gamma}\left\|\nabla_k(\mathcal{X}) - \mathcal{G}_k + \tfrac{\mathbf{\Psi}_k}{\rho}\right\|_F^2. \end{aligned} \tag{15}$$

Here $\mathbf{\Lambda}$, $\mathbf{\Gamma}_k$, $\mathbf{\Phi}$, and $\mathbf{\Psi}_k$ are multipliers, and μ, α, β, and ρ are positive penalty parameters. We update each variable with the others fixed and omit iteration indices. Detailed derivations are given in Appendix B.

(1) Updating $\mathcal{X}$

Setting the derivative with respect to $\mathcal{X}$ to zero gives

$$\left(\mu\mathbf{1} + \rho\sum_{k\in\Gamma}\nabla_k^T\nabla_k\right)\mathcal{X} = \mu\left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right) + \rho\sum_{k\in\Gamma}\nabla_k^T\left(\mathcal{G}_k - \frac{\mathbf{\Psi}_k}{\rho}\right). \tag{16}$$

Here $\mathbf{1}$ is the identity operator and ∇_k^T is the adjoint of ∇_k. Under periodic boundaries, this linear system is diagonalized by FFTs along the modes in Γ.

(2) Updating $\mathcal{L}_k$ and $\mathcal{R}_k$

The factors are updated alternately in the Fourier domain. For the i-th frontal slice,

$$\widehat{\mathcal{L}}_k^{(i)} = \widehat{\mathcal{G}}_k^{(i)} \left(\widehat{\mathcal{R}}_k^{(i)}\right)^H \left(\widehat{\mathcal{R}}_k^{(i)} \left(\widehat{\mathcal{R}}_k^{(i)}\right)^H\right)^\dagger, \quad \widehat{\mathcal{R}}_k^{(i)} = \left(\left(\widehat{\mathcal{L}}_k^{(i)}\right)^H \widehat{\mathcal{L}}_k^{(i)}\right)^\dagger \left(\widehat{\mathcal{L}}_k^{(i)}\right)^H \widehat{\mathcal{G}}_k^{(i)}. \tag{17}$$

Here the hat denotes an FFT along the third mode, H denotes conjugate transpose, and $\dagger$ denotes the pseudoinverse. Inverse FFT recovers $\mathcal{L}_k$ and $\mathcal{R}_k$.

(3) Updating $\mathcal{E}$

Combining the two quadratic terms gives the soft-thresholding update

$$\mathcal{E} = \mathcal{T}_{\frac{\lambda}{\mu+\beta}} \left(\frac{\mu \left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right) + \beta \left(\mathcal{H} - \frac{\mathbf{\Phi}}{\beta}\right)}{\mu + \beta} \right), \tag{18}$$

where $\mathcal{T}_\tau$ is the elementwise soft-thresholding operator (Hale et al., 2008).

(4) Updating $\mathcal{S}$

With $\Psi_\mathcal{S} = \mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} + \mathbf{\Lambda}/\mu$, projection onto $\|\mathcal{P}_\Omega(\mathcal{S})\|_F \le \delta$ gives

$$\mathcal{S} = \mathcal{P}_{\Omega^C}(\Psi_\mathcal{S}) + \min\left\{1, \frac{\delta}{\|\mathcal{P}_\Omega(\Psi_\mathcal{S})\|_F}\right\} \mathcal{P}_\Omega(\Psi_\mathcal{S}). \tag{19}$$

The observed component is defined as zero when its norm is zero.

(5) Updating $\mathcal{I}_k$ and $\mathcal{H}$

Their subproblems correspond to proximal mappings of $g_\mathcal{G}/\alpha$ and $h_\mathcal{E}/\beta$ (Parikh & Boyd, 2014). We replace these mappings with two CNN modules:

$$\mathcal{I}_k = \mathcal{F}_\mathcal{G}\left(\mathcal{G}_k + \frac{\mathbf{\Gamma}_k}{\alpha}; \Theta_\mathcal{G}\right), \qquad \mathcal{H} = \mathcal{F}_\mathcal{E}\left(\mathcal{E} + \frac{\mathbf{\Phi}}{\beta}; \Theta_\mathcal{E}\right), \tag{20}$$

where $\Theta_\mathcal{G}$ and $\Theta_\mathcal{E}$ are learnable parameters.

(6) Updating $\mathcal{G}_k$

Setting the derivative with respect to $\mathcal{G}_k$ to zero yields

$$\mathcal{G}_k = \frac{\frac{1}{\gamma}\mathcal{F}_{\text{att}}(\mathcal{D}) \odot (\mathcal{L}_k * \mathcal{R}_k) + \alpha\left(\mathcal{I}_k - \frac{\mathbf{\Gamma}_k}{\alpha}\right) + \rho\left(\nabla_k(\mathcal{X}) + \frac{\mathbf{\Psi}_k}{\rho}\right)}{\frac{1}{\gamma}\mathcal{F}_{\text{att}}(\mathcal{D}) + \alpha + \rho}. \tag{21}$$

The division is elementwise. Larger attention weights give more importance to the low-rank factor product.

(7) Updating Lagrange multipliers

Finally, the multipliers are updated using the constraint residuals:

$$\begin{aligned} &\mathbf{\Lambda} = \mathbf{\Lambda} + \mu\left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S}\right), \quad && \mathbf{\Gamma}_k = \mathbf{\Gamma}_k + \alpha(\mathcal{G}_k - \mathcal{I}_k), \\ &\mathbf{\Phi} = \mathbf{\Phi} + \beta(\mathcal{E} - \mathcal{H}), && \mathbf{\Psi}_k = \mathbf{\Psi}_k + \rho(\nabla_k(\mathcal{X}) - \mathcal{G}_k). \end{aligned} \tag{22}$$

3.3 Deep Unfolded Network

Each of the K stages follows the prescribed updates above, with the AWA weights shared across stages (Figure 2). The reconstruction is initialized with $\mathcal{P}_\Omega(\mathcal{D})$, the sparse error, slack, auxiliaries and multipliers with zeros, and the factors with small values. We set $r = 10$. The two learned mappings in the auxiliary-variable updates are residual dense blocks with eight and six convolutional layers. The penalty parameters, attention module, and learned mappings are trained by backpropagation; the output of stage K is the restored HSI. These blocks are learned maps and are not claimed to be exact proximal operators of explicit regularizers.

4 EXPERIMENTS

4.1 SETUP

We conduct experiments on three widely used hyperspectral image (HSI) datasets, including Indian Pines, Salinas, and Pavia University.

Indian Pines. The Indian Pines dataset was collected by the AVIRIS sensor and contains 145×145 pixels with 200 spectral bands after preprocessing. In our experiments, a $128 \times 128 \times 200$ subcube is used for evaluating.

Salinas. The Salinas dataset was also acquired by the AVIRIS sensor. It consists of 512×217 pixels and 224 spectral bands. A $192 \times 192 \times 224$ subcube is selected for evaluating.

Pavia University. The Pavia University dataset was acquired by the ROSIS sensor over the University of Pavia, Italy. It contains 610×340 pixels and 103 spectral bands. Following the experimental setting, the upper-left $256 \times 256 \times 103$ subcube is used for evaluating.

For each dataset, the HSI data are first normalized to the range $[0, 1]$. To simulate HSI degradation, we generate stripe-corrupted observations under three missing rates, namely 0.5, 0.7, and 0.9. The binary mask is used to indicate the locations of observed and missing entries. Training samples are obtained by randomly cropping spatial patches of size 32×32 for Indian Pines and 64×64 for Salinas and Pavia University from the original HSIs, and the same degradation setting is applied to generate the corresponding corrupted inputs. Random rotations and flips are further used for data augmentation.

All experiments are conducted on a computing instance equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of graphics memory, an AMD EPYC 9354 processor, and 60.1 GB of RAM. The degraded HSI tensor $\mathcal{D}$ and the binary observation mask $\mathcal{M}_\Omega$ are provided to the proposed deep unfolding network, where $\mathcal{P}_\Omega(\mathcal{D}) = \mathcal{M}_\Omega \odot \mathcal{D}$ denotes the masked observation tensor. The network configuration follows Section 3. The training objective is defined as the ℓ_1 loss between the restored HSI and the corresponding ground truth. The network is optimized using Adam with an initial learning rate of 3×10^{-5} and trained for 60 epochs.

Baselines are SNN (Liu et al., 2012), TNN (Qin et al., 2022), TVWTR (Wang et al., 2021), SPC_TV (Yokota et al., 2016), AGTC (Mai et al., 2024), KBR (Xie et al., 2017), TNN_TV (Qiu et al., 2021), and SNN_TV (Li et al., 2017), evaluated under the same degradation settings. Metrics are PSNR, SSIM (Wang et al., 2004), FSIM (Zhang et al., 2011), and ERGAS (Wald, 2000); higher is better except for ERGAS.

4.2 RESTORATION RESULTS

Tables 1–3 report the complete PSNR, SSIM, FSIM, and ERGAS comparisons. AWSTC performs strongly across datasets and missing rates. At 90% missingness, its PSNR is 41.29, 47.10, and 36.06 dB on Indian Pines, Salinas, and Pavia University, respectively. On Pavia University at this rate, AGTC has the higher SSIM (0.97 versus 0.96), while AWSTC has the best PSNR, FSIM, and ERGAS.

Table 1: Quantitative results on the Indian Pines dataset under different Missing Rate settings. The best results are highlighted in bold.

Method	Missing Rate = 0.5				Missing Rate = 0.7				Missing Rate = 0.9			
	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS
Original	15.51	0.10	0.50	605.00	14.08	0.07	0.50	713.77	13.00	0.03	0.57	805.55
SNN	32.16	0.86	0.88	91.64	31.19	0.82	0.84	107.52	24.01	0.54	0.74	275.26
TNN	29.20	0.80	0.86	128.04	27.81	0.73	0.81	153.58	24.17	0.50	0.70	245.41
TVWTR	33.13	0.84	0.87	87.35	33.14	0.82	0.85	92.72	32.22	0.75	0.81	111.33
SPC_TV	34.71	0.89	0.92	68.41	35.20	0.89	0.91	66.70	36.38	0.85	0.89	68.33
AGTC	39.39	0.94	0.87	50.80	38.25	0.91	0.85	60.06	37.12	0.88	0.80	71.02
KBR	42.68	0.95	0.97	51.33	31.74	0.71	0.83	175.64	20.86	0.43	0.68	397.35
TNN_TV	42.99	0.94	0.95	39.05	39.65	0.89	0.90	57.38	36.05	0.83	0.82	83.31
SNN_TV	44.99	0.95	0.96	34.22	41.19	0.90	0.91	51.42	36.84	0.83	0.83	82.32
AWSTC	**52.92**	**0.99**	**0.99**	**17.03**	**47.75**	**0.98**	**0.97**	**24.47**	**41.29**	**0.95**	**0.92**	**40.04**

Table 2: Quantitative results on the Salinas dataset under different Missing Rate settings. The best results are highlighted in bold.

Method	Missing Rate = 0.5				Missing Rate = 0.7				Missing Rate = 0.9			
	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS
Original	28.22	0.44	0.68	403.53	26.78	0.36	0.67	476.12	25.68	0.26	0.69	539.61
SNN	43.20	0.93	0.93	62.52	40.56	0.89	0.89	87.55	32.94	0.73	0.78	264.18
TNN	38.42	0.89	0.92	139.19	35.07	0.81	0.88	232.37	29.15	0.60	0.76	503.16
TVWTR	39.23	0.88	0.88	102.15	38.05	0.86	0.86	114.36	36.14	0.85	0.84	134.19
SPC_TV	46.66	0.96	0.97	37.24	46.52	0.97	0.96	36.76	41.79	0.92	0.91	69.77
AGTC	42.93	0.97	0.92	53.81	42.19	0.97	0.90	59.22	39.66	0.94	0.83	80.49
KBR	51.61	0.95	0.97	85.37	44.60	0.84	0.90	238.80	24.43	0.46	0.71	1035.84
TNN_TV	47.74	0.98	0.97	34.18	44.44	0.95	0.93	50.77	39.94	0.91	0.86	89.28
SNN_TV	49.02	0.98	0.97	30.22	45.77	0.96	0.94	44.07	41.21	0.92	0.88	79.79
AWSTC	**59.02**	**1.00**	**0.99**	**11.08**	**51.25**	**0.99**	**0.98**	**22.50**	**47.10**	**0.98**	**0.96**	**35.15**

Table 3: Quantitative results on the Pavia University dataset under different Missing Rate settings. The best results are highlighted in bold.

Method	Missing Rate = 0.5				Missing Rate = 0.7				Missing Rate = 0.9			
	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS	PSNR	SSIM	FSIM	ERGAS
Original	16.92	0.25	0.66	555.26	15.51	0.15	0.57	652.53	14.42	0.05	0.50	739.41
SNN	27.11	0.79	0.87	171.82	24.03	0.66	0.77	247.77	19.12	0.42	0.65	432.46
TNN	31.38	0.92	0.96	110.52	33.49	0.93	0.96	103.29	27.44	0.75	0.88	185.64
TVWTR	24.72	0.66	0.79	227.71	23.32	0.57	0.73	268.22	21.74	0.49	0.65	323.40
SPC_TV	23.33	0.53	0.79	266.84	21.86	0.43	0.74	315.45	20.77	0.37	0.70	357.32
AGTC	42.74	0.99	0.99	32.07	34.31	0.97	0.97	91.02	32.01	**0.97**	0.97	118.63
KBR	39.11	0.98	0.99	44.39	42.33	0.99	0.99	32.74	30.51	0.87	0.93	127.84
TNN_TV	29.39	0.86	0.89	132.59	25.82	0.73	0.78	201.50	22.56	0.54	0.63	293.35
SNN_TV	29.96	0.86	0.90	123.96	26.30	0.74	0.80	190.58	22.46	0.55	0.64	297.95
AWSTC	**46.16**	**0.99**	**1.00**	**22.74**	**43.26**	**0.99**	**0.99**	**30.84**	**36.06**	0.96	**0.98**	**64.75**

4.3 VISUAL COMPARISON

Figure 3 shows the restored images; corresponding error maps are in Appendix C. With fewer observations, conventional low-rank methods often blur boundaries, while several competing methods retain visible stripes or color distortions. TV-based variants improve spatial continuity but can still lose local details.

AWSTC better preserves land-cover boundaries in Indian Pines, continuous field regions in Salinas, and urban structures in Pavia University. These comparisons are consistent with the quantitative results, while the error maps in Figure 6 provide a complementary view of residual reconstruction errors. All visual comparisons use the same degradation settings as the quantitative evaluation.

4.4 ABLATIONS

Table 4 compares attention modules on Indian Pines with 50% missingness while keeping the remaining components fixed. AWA increases PSNR from 45.75 dB without attention to 52.92 dB and outperforms AHDRNet, FFA, SimAM, and CBAM in all four reported metrics. Figure 4 compares AWA with CBAM across all three datasets and missing rates. On Salinas, AWA gains 3.92, 4.95, and 1.69 dB at missing rates 0.5, 0.7, and 0.9, respectively.

Table 4: Effect of different attention modules on the Indian Pines dataset with a Missing Rate of 0.5. The best results are highlighted in bold.

Attention Module	PSNR	SSIM	FSIM	ERGAS
None	45.75	0.96	0.95	27.56
AHDRNet (Yan et al., 2019)	47.10	0.97	0.96	26.76
FFA (Qin et al., 2020)	47.02	0.97	0.96	28.32
SimAM (Yang et al., 2021)	50.96	0.98	0.98	19.31
CBAM (Woo et al., 2018)	49.27	0.98	0.98	20.10
AWA	**52.92**	**0.99**	**0.99**	**17.03**

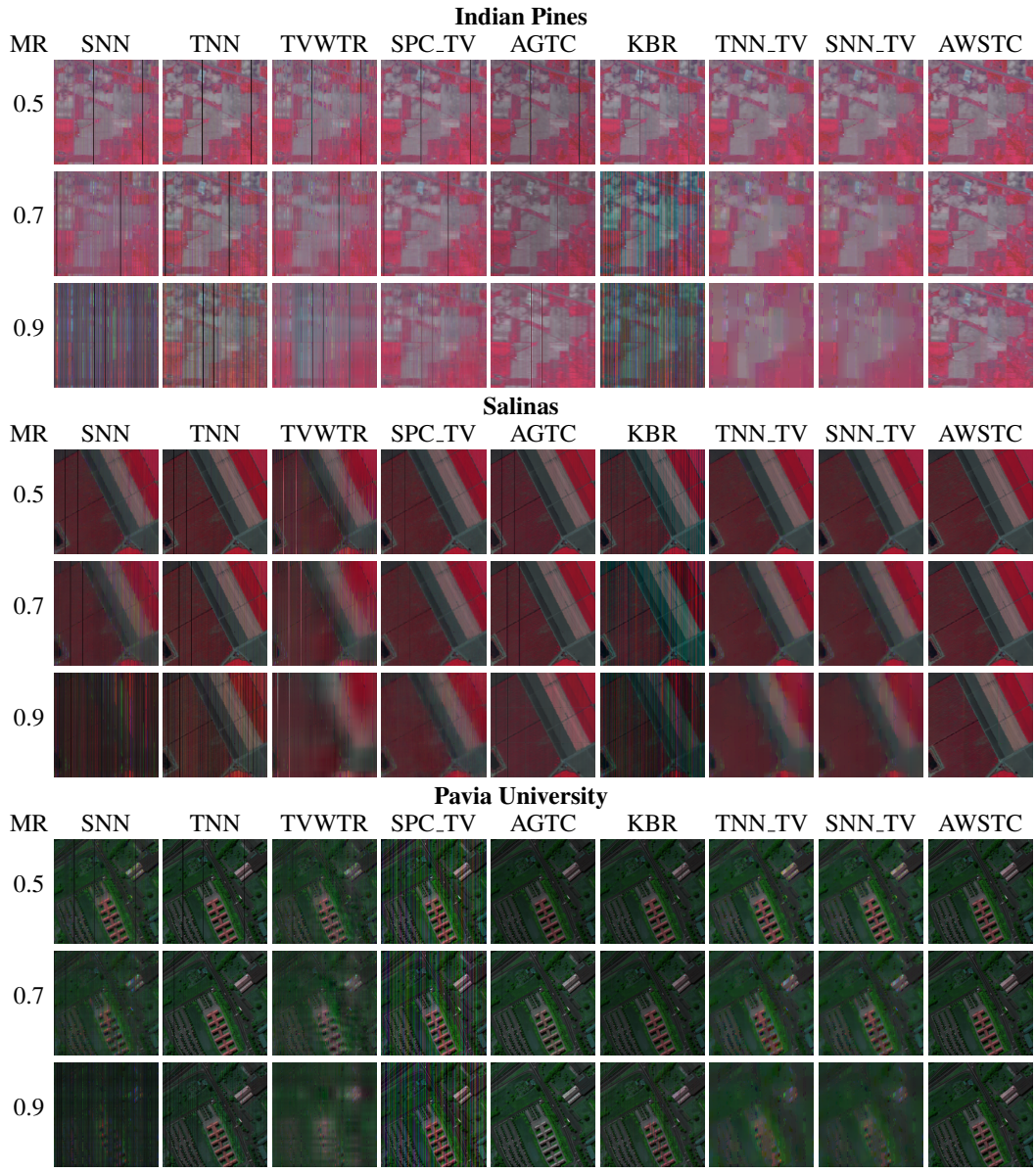


Figure 3: Visual comparison of different methods on the Indian Pines, Salinas, and Pavia University datasets under Missing Rate settings of 0.5, 0.7, and 0.9.

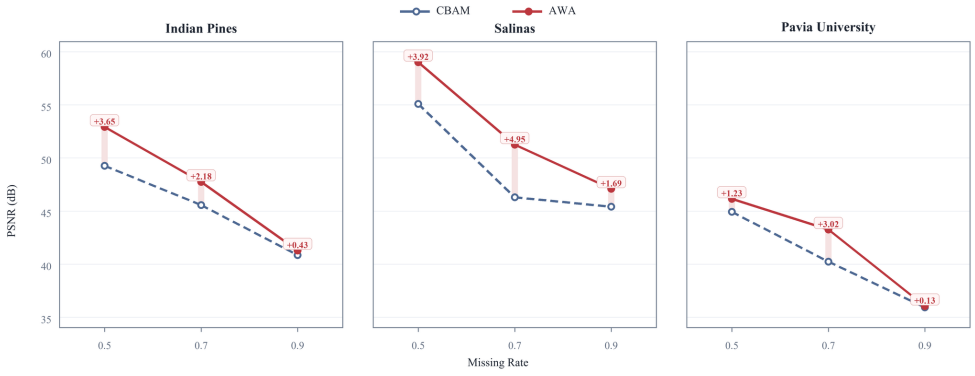


Figure 4: PSNR comparison between CBAM and the proposed AWA module on the Indian Pines, Salinas, and Pavia University datasets under different Missing Rate settings.

Table 5: Ablation of AWA components on Indian Pines at a Missing Rate of 50%. PSNR drops are measured relative to the complete AWA model.

Metric	Fixed Haar filters	No thresholding	No subband attention	No high-frequency fusion	Complete AWA
PSNR	52.09	52.52	51.75	52.66	**52.92**
Drop	0.83	0.40	1.17	0.26	—

Table 5 isolates four AWA components under the same setting. Fixing analysis/synthesis filters to Haar, removing thresholding, removing subband attention (while retaining final channel attention), and removing high-frequency fusion reduce PSNR by 0.83, 0.40, 1.17, and 0.26 dB, respectively. These results support each component, with the largest drop from removing subband attention. The following depth study examines the choice of K.

4.5 Unfolding depth

We vary K on Indian Pines with 50% missingness while keeping the other settings unchanged (Figure 5). Increasing K from 9 to 11 raises PSNR from 43.09 to 52.92 dB and lowers ERGAS from 29.00 to 17.03. The slight decline at $K = 12$ supports choosing $K = 11$ among the tested configurations; more stages do not necessarily improve restoration.

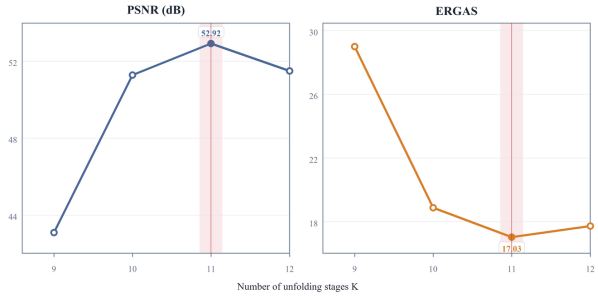


Figure 5: Effect of the number of unfolding stages K on PSNR and ERGAS for the Indian Pines dataset with a Missing Rate of 0.5.

5 Conclusion

In this paper, we proposed AWSTC, an adaptive wavelet attention-guided deep unfolding method for hyperspectral image tensor completion. The proposed method integrates low-rank tensor factorization with fused smoothness regularization. In addition, the adaptive wavelet attention module extracts informative frequency-aware features from wavelet subbands and uses them to guide the low-rank recovery process. By unfolding the derived optimization algorithm into a learnable network, AWSTC keeps the interpretability of model-based tensor recovery while improving the flexibility of data-driven restoration.

Experiments on the Indian Pines, Salinas, and Pavia University datasets show that AWSTC achieves strong restoration performance under different missing rates. Compared with representative tensor completion methods, the proposed method obtains better quantitative results in most settings and produces clearer visual structures with fewer stripe artifacts. The error maps further show that AWSTC can effectively reduce reconstruction errors, especially when the missing rate is high. Moreover, the discussion on different attention modules and unfolding stages verifies the effectiveness of the proposed AWA module and supports the adopted network configuration. In future work, we will consider extending the proposed framework to more general high-order tensor recovery tasks and more complex real-world degradation models.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, adapt the LaTeX template, identify baseline methods for comparison, and assist in checking the correctness of our algorithm implementation. All AI-assisted work was reviewed by the authors, and the algorithm code and experimental results were manually checked. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We will release the source code upon request from the reviewers.

REFERENCES

Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016. doi: 10.1017/S096249291600009X.

Sayantan Dutta, Adrian Basarab, Bertrand Georgeot, and Denis Kouamé. Diva: Deep unfolded network from quantum interactive patches for image restoration. *Pattern Recognition*, 155:110676, 2024.

Haiyan Fan, Chang Li, Yulan Guo, Gangyao Kuang, and Jiayi Ma. Spatial–spectral total variation regularized low-rank tensor decomposition for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 56(10):6196–6213, 2018. doi: 10.1109/TGRS.2018.2833473.

Pedram Ghamisi, Naoto Yokoya, Jun Li, Wenzhi Liao, Sicong Liu, Javier Plaza, Behnood Rasti, and Antonio Plaza. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):37–78, 2017. doi: 10.1109/MGRS.2017.2762087.

Alexander F. H. Goetz. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote Sensing of Environment*, 113:S5–S16, 2009. doi: 10.1016/j.rse.2007.12.014.

Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

Misha E Kilmer and Carla D Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011.

Weichao Kong, Feng Zhang, Wenjin Qin, and Jianjun Wang. Low-tubal-rank tensor recovery with multilayer subspace prior learning. *Pattern Recognition*, 140:109545, 2023.

Xutao Li, Yunming Ye, and Xiaofei Xu. Low-rank tensor completion with total variation for visual data inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2012.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):925–938, 2020. doi: 10.1109/TPAMI.2019.2891760.

Truong Thanh Nhat Mai, Edmund Y Lam, and Chul Lee. Deep unrolled low-rank tensor completion for high dynamic range imaging. *IEEE Transactions on Image Processing*, 31:5774–5787, 2022.

Truong Thanh Nhat Mai, Edmund Y Lam, and Chul Lee. Attention-guided low-rank tensor completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9818–9833, 2024. doi: 10.1109/TPAMI.2024.3429498.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Wenjin Qin, Hailin Wang, Feng Zhang, Jianjun Wang, Xin Luo, and Tingwen Huang. Low-rank high-order tensor completion with applications in visual data. *IEEE Transactions on Image Processing*, 31:2433–2448, 2022.

Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11908–11915, 2020.

Duo Qiu, Minru Bai, Michael K Ng, and Xiongjun Zhang. Robust low-rank tensor completion via transformed tensor nuclear norm with total variation regularization. *Neurocomputing*, 435: 197–215, 2021.

Lucien Wald. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pp. 99–103. SEE/URISCA, 2000.

Hailin Wang, Jiangjun Peng, Wenjin Qin, Jianjun Wang, and Deyu Meng. Guaranteed tensor recovery fused low-rankness and smoothness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10990–11007, 2023.

Minghua Wang, Qiang Wang, Jocelyn Chanussot, and Danfeng Hong. Total variation regularized weighted tensor ring decomposition for missing data recovery in high-dimensional optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

Yao Wang, Jiangjun Peng, Qian Zhao, Deyu Meng, Yee Leung, and Xi-Le Zhao. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *arXiv preprint arXiv:1707.02477*, 2017. URL `https://arxiv.org/abs/1707.02477`.

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.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1888–1902, 2017.

Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1751–1760, 2019.

Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11863–11874, 2021.

Tatsuya Yokota, Qibin Zhao, and Andrzej Cichocki. Smooth parafac decomposition for tensor completion. *IEEE Transactions on Signal Processing*, 64(20):5423–5436, 2016.

Haijin Zeng, Xiaozhen Xie, Haojie Cui, Hanping Yin, and Jifeng Ning. Hyperspectral image restoration via global l 1-2 spatial–spectral total variation regularized local low-rank tensor recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3309–3325, 2020.

Feng Zhang, Jianjun Wang, Wendong Wang, and Chen Xu. Low-tubal-rank plus sparse tensor recovery with prior subspace information. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3492–3507, 2020.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Xi-Le Zhao, Wen-Hao Xu, Tai-Xiang Jiang, Yao Wang, and Michael K. Ng. Deep plug-and-play prior for low-rank tensor completion. *Neurocomputing*, 400:137–149, 2020. doi: 10.1016/j.neucom.2020.03.018.

Pan Zhou, Canyi Lu, Zhouchen Lin, and Chao Zhang. Tensor factorization for low-rank tensor completion. *IEEE Transactions on Image Processing*, 27(3):1152–1163, 2017.

A NOTATIONS AND PRELIMINARIES

For a third-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, its (i_1, i_2, i_3)-th entry is denoted by $\mathcal{X}(i_1, i_2, i_3)$, and its i_3-th frontal slice is denoted by $\mathcal{X}_{i_3} = \mathcal{X}(:, :, i_3)$.

For $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the entrywise ℓ_1 norm and Frobenius norm of $\mathcal{A}$ are defined as

$$\|\mathcal{A}\|_1 = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} |\mathcal{A}(i_1, i_2, i_3)|, \qquad \|\mathcal{A}\|_F = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} |\mathcal{A}(i_1, i_2, i_3)|^2}. \tag{A.1}$$

The corresponding inner product is defined for $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ by

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \mathcal{A}(i_1, i_2, i_3) \mathcal{B}(i_1, i_2, i_3). \tag{A.2}$$

We next introduce the t-product, which extends matrix multiplication to third-order tensors and is used in the low-rank factorization below.

Definition A.1 (T-product Kilmer & Martin (2011)) *Let $\mathcal{X} \in \mathbb{R}^{I_1 \times k \times I_3}$ and $\mathcal{Y} \in \mathbb{R}^{k \times I_2 \times I_3}$ be two third-order tensors. The t-product of $\mathcal{X}$ and $\mathcal{Y}$ produces a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, defined by*

$$\mathcal{T} = \mathcal{X} * \mathcal{Y} = \mathrm{fold}(\mathrm{bcirc}(\mathcal{X}) \cdot \mathrm{unfold}(\mathcal{Y})), \tag{A.3}$$

where $\mathrm{bcirc}(\mathcal{X}) \in \mathbb{R}^{I_1 I_3 \times k I_3}$ is the block circulant matrix formed by arranging the frontal slices of $\mathcal{X}$ in a circulant pattern:

$$\mathrm{bcirc}(\mathcal{X}) = \begin{bmatrix} \mathcal{X}_1 & \mathcal{X}_{I_3} & \cdots & \mathcal{X}_2 \\ \mathcal{X}_2 & \mathcal{X}_1 & \cdots & \mathcal{X}_3 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{X}_{I_3} & \mathcal{X}_{I_3-1} & \cdots & \mathcal{X}_1 \end{bmatrix}. \tag{A.4}$$

The operator $\mathrm{unfold}(\mathcal{Y}) \in \mathbb{R}^{k I_3 \times I_2}$ stacks the frontal slices of $\mathcal{Y}$ vertically:

$$\mathrm{unfold}(\mathcal{Y}) = \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \\ \vdots \\ \mathcal{Y}_{I_3} \end{bmatrix}. \tag{A.5}$$

The operator fold *is the inverse of* unfold*; here, it maps the matrix product in $\mathbb{R}^{I_1 I_3 \times I_2}$ to a tensor in $\mathbb{R}^{I_1 \times I_2 \times I_3}$.*

To introduce the t-SVD, we define tensor transpose and orthogonality Kilmer & Martin (2011). For $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the tensor transpose $\mathcal{X}^\top \in \mathbb{R}^{I_2 \times I_1 \times I_3}$ is formed by transposing each frontal slice and reversing the order of slices $2, \ldots, I_3$. The identity tensor $\mathcal{I} \in \mathbb{R}^{n \times n \times I_3}$ has the identity matrix as its first frontal slice and zero matrices as the remaining slices. A tensor $\mathcal{Q} \in \mathbb{R}^{n \times n \times I_3}$ is orthogonal if $\mathcal{Q}^\top * \mathcal{Q} = \mathcal{Q} * \mathcal{Q}^\top = \mathcal{I}$. A tensor is called *f-diagonal* if each frontal slice is a diagonal matrix, which may be rectangular.

Using these definitions, the matrix SVD extends to the following tensor SVD (t-SVD).

Theorem A.1 (t-SVD Kilmer & Martin (2011)) *Third-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ can be written as*

$$\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^\top, \tag{A.6}$$

where $\mathcal{U} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ and $\mathcal{V} \in \mathbb{R}^{I_2 \times I_2 \times I_3}$ are orthogonal tensors, and $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is an f-diagonal tensor.

Based on this framework, the tensor nuclear norm is used to promote low-rank structure. We use the definition of Lu et al. (2020) below.

Definition A.2 (Tensor nuclear norm Lu et al. (2020)) *For $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the tensor nuclear norm induced by the t-product is defined as*

$$\|\mathcal{X}\|_{\circledast} := \frac{1}{I_3} \sum_{j=1}^{I_3} \sum_{i=1}^{m} \sigma_{ij}, \tag{A.7}$$

where $m = \min(I_1, I_2)$, and σ_{ij} denotes the i-th singular value of the j-th frontal slice of $\overline{\mathcal{X}}$. Here, $\overline{\mathcal{X}} = \mathrm{fft}(\mathcal{X}, [], 3) \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ is obtained using the unnormalized discrete Fourier transform along the third mode. Equivalently, $\|\mathcal{X}\|_{\circledast} = I_3^{-1} \sum_{j=1}^{I_3} \|\overline{\mathcal{X}}_j\|_*$, where $\|\cdot\|_*$ denotes the matrix nuclear norm.

In addition to low-rank structure, we consider local smoothness through differences between neighboring entries. These differences form the gradient tensors defined below, with periodic boundary conditions Wang et al. (2023).

Definition A.3 (Gradient tensor Wang et al. (2023)) *Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a third-order tensor. The gradient tensor of $\mathcal{X}$ along the k-th mode is defined as*

$$\mathcal{G}_k := \nabla_k(\mathcal{X}) = \mathcal{X} \times_k \mathbf{D}_{I_k}, \qquad k = 1, 2, 3, \tag{A.8}$$

where $\mathbf{D}_{I_k} \in \mathbb{R}^{I_k \times I_k}$ is, for $I_k \geq 2$, the circulant difference matrix whose first row is $(-1, 1, 0, \ldots, 0)$ and whose successive rows are circular shifts to the right. The operator $\times_k$ denotes the mode-k product between a tensor and a matrix. Specifically, $\mathcal{X} \times_k \mathbf{D}_{I_k}$ applies the finite-difference operation to all mode-k fibers of $\mathcal{X}$. Thus, for a mode-k fiber $\mathbf{x}$, $(\mathbf{D}_{I_k}\mathbf{x})_i = x_{i+1} - x_i$, with $x_{I_k+1} = x_1$. For $I_k = 1$, we set $\mathbf{D}_{I_k} = 0$.

Total variation (TV) measures local changes using these gradients. Let $\Gamma \subseteq \{1, 2, 3\}$ be the nonempty set of selected modes. The standard anisotropic and isotropic TV are defined as Chambolle & Pock (2016)

$$\|\mathcal{X}\|_{\text{TV-1}} := \sum_{k \in \Gamma} \|\mathcal{G}_k\|_1,$$

$$\|\mathcal{X}\|_{\text{TV-2}} := \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \left(\sum_{k \in \Gamma} |\mathcal{G}_k(i_1, i_2, i_3)|^2 \right)^{1/2},$$

where $\mathcal{G}_k = \nabla_k(\mathcal{X})$. Here, isotropic TV combines the gradient directions at each entry; it differs from the sum of Frobenius norms called TV-2 in Wang et al. (2023). For HSIs, $\Gamma = \{1, 2, 3\}$ includes both spatial and spectral directions.

To also use the low-rank structure of gradient tensors, Wang et al. (2023) introduced tensor correlated total variation (t-CTV). It combines low-rankness and smoothness by applying the tensor nuclear norm to the gradients.

Definition A.4 (Tensor correlated total variation Wang et al. (2023)) *Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a third-order tensor, and let $\Gamma \subseteq \{1, 2, 3\}$ be a nonempty set of selected modes. With $\mathcal{G}_k = \nabla_k(\mathcal{X})$, the third-order Fourier-based form of t-CTV is defined as*

$$\|\mathcal{X}\|_{\text{t-CTV}} := \frac{1}{\gamma} \sum_{k \in \Gamma} \|\mathcal{G}_k\|_{\circledast}, \tag{A.9}$$

where $\|\cdot\|_{\circledast}$ is the tensor nuclear norm defined above and $\gamma = |\Gamma|$ is the number of selected modes.

TV and t-CTV are seminorms: both are zero for tensors that are constant along the selected modes, even when the tensors are nonzero.

B Detailed Optimization and Variable Updates

After obtaining the adaptive attention weight $\mathcal{F}_{\text{att}}(\mathcal{D})$, we solve the proposed tensor completion model by the augmented Lagrangian method. The model uses both low-rank information and smoothness information. To better fit real data, we also use two learned regularizers, $g_{\mathcal{G}}(\cdot)$ and $h_{\mathcal{E}}(\cdot)$, for the gradient-related tensor $\mathcal{G}_k$ and the sparse error tensor $\mathcal{E}$, respectively. Since the observed tensor may contain missing entries and bounded noise, we assume $\mathcal{P}_\Omega(\mathcal{D}) = \mathcal{P}_\Omega(\mathcal{X}+\mathcal{E}+\mathcal{N})$, where $\mathcal{N}$ is a noise tensor with a noise level $\|\mathcal{P}_\Omega(\mathcal{N})\|_F \leq \delta$ for some $\delta > 0$. A slack tensor $\mathcal{S}$ is introduced to model the bounded noise on the observed entries. The resulting model is written as

$$\begin{aligned}\min_{\mathcal{L}_k,\mathcal{R}_k,\mathcal{X},\mathcal{E},\mathcal{S},\{\mathcal{G}_k\}_{k\in\Gamma}} \frac{1}{\gamma}\sum_{k\in\Gamma}\frac{1}{2}\left\|\sqrt{\mathcal{F}_{\text{att}}(\mathcal{D})}\odot(\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\right\|_F^2 + \lambda\|\mathcal{E}\|_1 + \sum_{k\in\Gamma} g_{\mathcal{G}}(\mathcal{G}_k) + h_{\mathcal{E}}(\mathcal{E}) \\ \text{s.t. } \mathcal{G}_k = \nabla_k(\mathcal{X}), \quad \mathcal{X}+\mathcal{E}+\mathcal{S} = \mathcal{P}_\Omega(\mathcal{D}), \quad \|\mathcal{P}_\Omega(\mathcal{S})\|_F \leq \delta.\end{aligned} \tag{B.1}$$

To decouple the learned regularization terms, we introduce two auxiliary variables $\mathcal{I}_k$ and $\mathcal{H}$, where $\mathcal{I}_k = \mathcal{G}_k$ and $\mathcal{H} = \mathcal{E}$. The corresponding positive penalty parameters are μ, α, β, and ρ. Omitting terms that are constant with respect to the primal variables, the augmented Lagrangian function is written as

$$\begin{aligned}\mathcal{L} =& \frac{1}{\gamma}\sum_{k\in\Gamma}\frac{1}{2}\left\|\sqrt{\mathcal{F}_{\text{att}}(\mathcal{D})}\odot(\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\right\|_F^2 + \lambda\|\mathcal{E}\|_1 + \sum_{k\in\Gamma} g_{\mathcal{G}}(\mathcal{I}_k) + h_{\mathcal{E}}(\mathcal{H}) \\ &+ \frac{\mu}{2}\left\|\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right\|_F^2 + \frac{\alpha}{2}\sum_{k\in\Gamma}\left\|\mathcal{G}_k - \mathcal{I}_k + \frac{\mathbf{\Gamma}_k}{\alpha}\right\|_F^2 \\ &+ \frac{\beta}{2}\left\|\mathcal{E} - \mathcal{H} + \frac{\mathbf{\Phi}}{\beta}\right\|_F^2 + \frac{\rho}{2}\sum_{k\in\Gamma}\left\|\nabla_k(\mathcal{X}) - \mathcal{G}_k + \frac{\mathbf{\Psi}_k}{\rho}\right\|_F^2.\end{aligned} \tag{B.2}$$

We update the variables one by one as follows.

(1) Updating $\mathcal{X}$

When the other variables are fixed, the subproblem for $\mathcal{X}$ is

$$\min_{\mathcal{X}} \frac{\mu}{2}\left\|\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right\|_F^2 + \frac{\rho}{2}\sum_{k\in\Gamma}\left\|\nabla_k(\mathcal{X}) - \mathcal{G}_k + \frac{\mathbf{\Psi}_k}{\rho}\right\|_F^2. \tag{B.3}$$

This is a quadratic least-squares problem. By setting the derivative with respect to $\mathcal{X}$ to zero, we obtain

$$\left(\mu\mathbf{1} + \rho\sum_{k\in\Gamma}\nabla_k^T\nabla_k\right)\mathcal{X} = \mu\left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right) + \rho\sum_{k\in\Gamma}\nabla_k^T\left(\mathcal{G}_k - \frac{\mathbf{\Psi}_k}{\rho}\right). \tag{B.4}$$

Thus, the update of $\mathcal{X}$ is

$$\begin{aligned}\mathcal{X} = \left(\mu\mathbf{1} + \rho\sum_{k\in\Gamma}\nabla_k^T\nabla_k\right)^{-1}\Bigg[&\mu\left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right) \\ &+ \rho\sum_{k\in\Gamma}\nabla_k^T\left(\mathcal{G}_k - \frac{\mathbf{\Psi}_k}{\rho}\right)\Bigg].\end{aligned} \tag{B.5}$$

(2) Updating $\mathcal{L}_k$ and $\mathcal{R}_k$

With $\mathcal{G}_k$ fixed, the factor tensors are updated by solving

$$\min_{\mathcal{L}_k,\mathcal{R}_k} \frac{1}{2}\|\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k\|_F^2. \tag{B.6}$$

Based on the t-product, this problem can be solved slice by slice in the Fourier domain. For the i-th frontal slice, we have

$$\min_{\widehat{\mathcal{L}}_k^{(i)}} \frac{1}{2}\left\|\widehat{\mathcal{L}}_k^{(i)}\widehat{\mathcal{R}}_k^{(i)} - \widehat{\mathcal{G}}_k^{(i)}\right\|_F^2, \tag{B.7}$$

The solution is

$$\widehat{\mathcal{L}}_k^{(i)} = \widehat{\mathcal{G}}_k^{(i)} \left(\widehat{\mathcal{R}}_k^{(i)}\right)^H \left(\widehat{\mathcal{R}}_k^{(i)} \left(\widehat{\mathcal{R}}_k^{(i)}\right)^H\right)^\dagger. \tag{B.8}$$

Similarly, when $\widehat{\mathcal{L}}_k^{(i)}$ is fixed, $\widehat{\mathcal{R}}_k^{(i)}$ is obtained from

$$\min_{\widehat{\mathcal{R}}_k^{(i)}} \frac{1}{2} \left\| \widehat{\mathcal{L}}_k^{(i)} \widehat{\mathcal{R}}_k^{(i)} - \widehat{\mathcal{G}}_k^{(i)} \right\|_F^2, \tag{B.9}$$

which gives

$$\widehat{\mathcal{R}}_k^{(i)} = \left(\left(\widehat{\mathcal{L}}_k^{(i)}\right)^H \widehat{\mathcal{L}}_k^{(i)}\right)^\dagger \left(\widehat{\mathcal{L}}_k^{(i)}\right)^H \widehat{\mathcal{G}}_k^{(i)}. \tag{B.10}$$

Finally, the tensors in the original domain are recovered by inverse FFT:

$$\mathcal{L}_k = \mathcal{F}_{FFT}^{-1}\{\widehat{\mathcal{L}}_k\}, \qquad \mathcal{R}_k = \mathcal{F}_{FFT}^{-1}\{\widehat{\mathcal{R}}_k\}. \tag{B.11}$$

(3) Updating $\mathcal{E}$

The subproblem for $\mathcal{E}$ is

$$\min_{\mathcal{E}} \lambda\|\mathcal{E}\|_1 + \frac{\mu}{2} \left\| \mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu} \right\|_F^2 + \frac{\beta}{2} \left\| \mathcal{E} - \mathcal{H} + \frac{\mathbf{\Phi}}{\beta} \right\|_F^2. \tag{B.12}$$

After combining the two quadratic terms, this problem becomes a standard ℓ_1-norm proximal problem. Therefore, $\mathcal{E}$ is updated by soft-thresholding:

$$\mathcal{E} = \mathcal{T}_{\frac{\lambda}{\mu+\beta}} \left(\frac{\mu\left(\mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{S} + \frac{\mathbf{\Lambda}}{\mu}\right) + \beta\left(\mathcal{H} - \frac{\mathbf{\Phi}}{\beta}\right)}{\mu + \beta} \right), \tag{B.13}$$

where $\mathcal{T}_\tau(\mathcal{A})$ denotes the element-wise soft-thresholding operator (Hale et al., 2008) with parameter $\tau > 0$.

(4) Updating $\mathcal{S}$

For the slack tensor $\mathcal{S}$, we first define

$$\Psi_\mathcal{S} = \mathcal{P}_\Omega(\mathcal{D}) - \mathcal{X} - \mathcal{E} + \frac{\mathbf{\Lambda}}{\mu}. \tag{B.14}$$

Then $\mathcal{S}$ is obtained by projecting $\Psi_\mathcal{S}$ onto the set $\|\mathcal{P}_\Omega(\mathcal{S})\|_F \le \delta$. This gives

$$\mathcal{S} = \mathcal{P}_{\Omega^C}(\Psi_\mathcal{S}) + \min\left\{1, \frac{\delta}{\|\mathcal{P}_\Omega(\Psi_\mathcal{S})\|_F}\right\} \mathcal{P}_\Omega(\Psi_\mathcal{S}). \tag{B.15}$$

When $\mathcal{P}_\Omega(\Psi_\mathcal{S}) = 0$, the observed component of this projection is defined to be zero by continuous extension.

(5) Updating $\mathcal{I}_k$ and $\mathcal{H}$

The auxiliary variables $\mathcal{I}_k$ and $\mathcal{H}$ are used for the learned regularizers. Their subproblems are

$$\mathcal{I}_k = \arg\min_{\mathcal{I}_k} g_\mathcal{G}(\mathcal{I}_k) + \frac{\alpha}{2} \left\| \mathcal{G}_k - \mathcal{I}_k + \frac{\mathbf{\Gamma}_k}{\alpha} \right\|_F^2, \tag{B.16}$$

and

$$\mathcal{H} = \arg\min_{\mathcal{H}} h_\mathcal{E}(\mathcal{H}) + \frac{\beta}{2} \left\| \mathcal{E} - \mathcal{H} + \frac{\mathbf{\Phi}}{\beta} \right\|_F^2. \tag{B.17}$$

Thus, they can be written as proximal mappings:

$$\mathcal{I}_k = \mathrm{prox}_{\frac{1}{\alpha} g_\mathcal{G}} \left(\mathcal{G}_k + \frac{\mathbf{\Gamma}_k}{\alpha} \right), \qquad \mathcal{H} = \mathrm{prox}_{\frac{1}{\beta} h_\mathcal{E}} \left(\mathcal{E} + \frac{\mathbf{\Phi}}{\beta} \right), \tag{B.18}$$

where $\mathrm{prox}_{g_\mathcal{G}}(\cdot)$ and $\mathrm{prox}_{h_\mathcal{E}}(\cdot)$ are the proximal operators (Parikh & Boyd, 2014) corresponding to regularization functions $g_\mathcal{G}$ and $h_\mathcal{E}$, respectively.

In traditional model-based methods, the regularization functions are usually manually designed. However, such fixed priors may not fully describe complex image structures. To improve the representation ability of the model, we use CNNs to learn the proximal mappings of $g_{\mathcal{G}}(\cdot)$ and $h_{\mathcal{E}}(\cdot)$. Specifically, the proximal operators are replaced by two CNN modules, denoted as $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{E}}$. Therefore, the updates of $\mathcal{I}_k$ and $\mathcal{H}$ can be rewritten as

$$\mathcal{I}_k = \mathcal{F}_{\mathcal{G}}\left(\mathcal{G}_k + \frac{\mathbf{\Gamma}_k}{\alpha}; \Theta_{\mathcal{G}}\right), \qquad \mathcal{H} \quad = \mathcal{F}_{\mathcal{E}}\left(\mathcal{E} + \frac{\mathbf{\Phi}}{\beta}; \Theta_{\mathcal{E}}\right), \tag{B.19}$$

where $\Theta_{\mathcal{G}}$ and $\Theta_{\mathcal{E}}$ are the learnable parameters of $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{E}}$, respectively.

(6) Updating $\mathcal{G}_k$

The update of $\mathcal{G}_k$ is the main step where the proposed adaptive wavelet attention is used. When the other variables are fixed, the subproblem is

$$\begin{aligned}\min_{\mathcal{G}_k} \frac{1}{2\gamma}\left\|\sqrt{\mathcal{F}_{\mathrm{att}}(\mathcal{D})} \odot (\mathcal{L}_k * \mathcal{R}_k - \mathcal{G}_k)\right\|_F^2 + \frac{\alpha}{2}\left\|\mathcal{G}_k - \mathcal{I}_k + \frac{\mathbf{\Gamma}_k}{\alpha}\right\|_F^2 \\ + \frac{\rho}{2}\left\|\nabla_k(\mathcal{X}) - \mathcal{G}_k + \frac{\mathbf{\Psi}_k}{\rho}\right\|_F^2.\end{aligned} \tag{B.20}$$

Since all terms are quadratic with respect to $\mathcal{G}_k$, we set the derivative to zero and obtain

$$\left(\frac{1}{\gamma}\mathcal{F}_{\mathrm{att}}(\mathcal{D}) + \alpha + \rho\right)\mathcal{G}_k = \frac{1}{\gamma}\mathcal{F}_{\mathrm{att}}(\mathcal{D}) \odot (\mathcal{L}_k * \mathcal{R}_k) + \alpha\left(\mathcal{I}_k - \frac{\mathbf{\Gamma}_k}{\alpha}\right) + \rho\left(\nabla_k(\mathcal{X}) + \frac{\mathbf{\Psi}_k}{\rho}\right). \tag{B.21}$$

Therefore,

$$\mathcal{G}_k = \frac{\frac{1}{\gamma}\mathcal{F}_{\mathrm{att}}(\mathcal{D}) \odot (\mathcal{L}_k * \mathcal{R}_k) + \alpha\left(\mathcal{I}_k - \frac{\mathbf{\Gamma}_k}{\alpha}\right) + \rho\left(\nabla_k(\mathcal{X}) + \frac{\mathbf{\Psi}_k}{\rho}\right)}{\frac{1}{\gamma}\mathcal{F}_{\mathrm{att}}(\mathcal{D}) + \alpha + \rho}. \tag{B.22}$$

This update shows how $\mathcal{F}_{\mathrm{att}}(\mathcal{D})$ affects the low-rank term. If a component has a larger AWA weight, $\mathcal{G}_k$ is pulled more strongly toward $\mathcal{L}_k * \mathcal{R}_k$. If the weight is smaller, the update depends more on the smoothness term and the learned proximal prior.

Multiplier updates. After the primal variables are updated, the multipliers follow the four constraint residuals:

$$\begin{aligned}\mathbf{\Lambda} &= \mathbf{\Lambda} + \mu\left(\mathcal{P}_{\Omega}(\mathcal{D}) - \mathcal{X} - \mathcal{E} - \mathcal{S}\right), \\ \mathbf{\Gamma}_k &= \mathbf{\Gamma}_k + \alpha(\mathcal{G}_k - \mathcal{I}_k), \\ \mathbf{\Phi} &= \mathbf{\Phi} + \beta(\mathcal{E} - \mathcal{H}), \\ \mathbf{\Psi}_k &= \mathbf{\Psi}_k + \rho(\nabla_k(\mathcal{X}) - \mathcal{G}_k).\end{aligned} \tag{B.23}$$

C Reconstruction Error Maps

Figure 6 complements the restored images in Figure 3 of the main text, retaining all datasets and missing rates. Darker regions indicate smaller reconstruction errors.

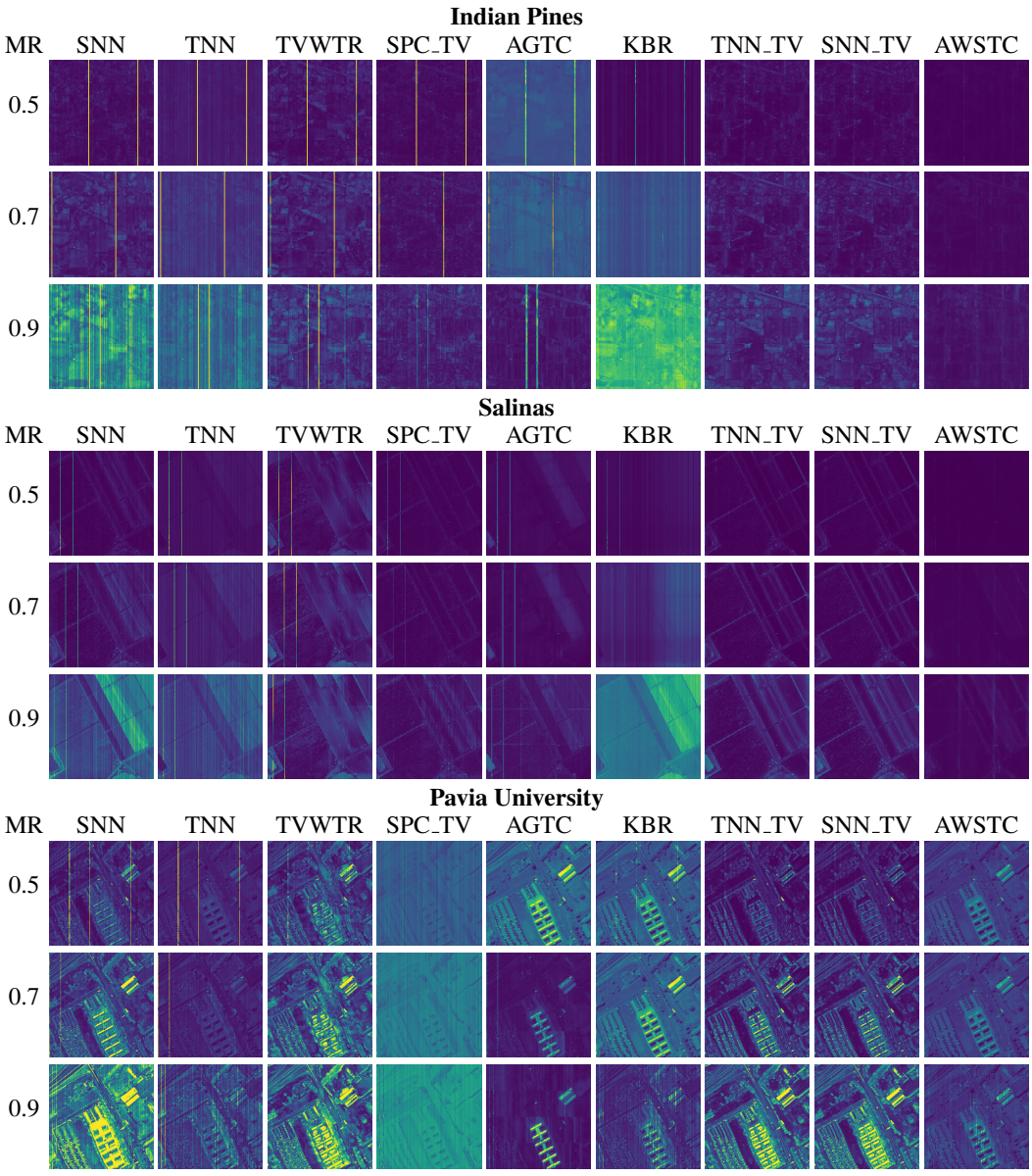


Figure 6: Error maps of different methods on the Indian Pines, Salinas, and Pavia University datasets under Missing Rate settings of 0.5, 0.7, and 0.9. Darker colors indicate smaller reconstruction errors.