S2WMAMBA: A WAVELET-ASSISTED MAMBA-BASED DUAL-BRANCH NETWORK FOR PANSHARPENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pansharpening fuses a high-resolution panchromatic (PAN) image with a low-resolution multispectral (MS) image to produce a high-resolution multispectral (HRMS) image. A key difficulty is that jointly processing PAN and MS features often entangles spatial detail enhancement with spectral fidelity. To address this feature entanglement, we propose S2WMamba, a framework that disentangles modality-specific frequency information to guide cross-modal fusion. Concretely, unlike global frequency transforms, a localized 2D Haar DWT is applied to PAN features to isolate spatial edges and textures. Concurrently, a novel channel-wise 1D Haar DWT treats each pixel's spectrum as a 1D signal, re-parameterizing it as a shared spectral base plus band-specific variations; operating in this basis limits spectral distortion. The resulting Spatial branch injects wavelet-extracted spatial details into MS features, while the Spectral branch refines PAN features using spectra from the DWT1D process. To fuse these decoupled sub-bands, the two branches exchange information via Mamba-based cross-modulation, which models long-range dependencies across the sub-bands with linear complexity. On WV3, GF2, and QB datasets, S2WMamba matches or surpasses strong baselines (FusionMamba, CANNet, U2Net, PanNet), improving reduced-resolution PSNR by up to 0.23 dB and reaching a Hybrid Quality with No Reference (HQNR) of 0.956 on full-resolution WV3. Ablations show that both branches are necessary and that parallel dual-branch fusion outperforms sequential variants.

1 INTRODUCTION

High-resolution multispectral (HRMS) images are important for applications such as environmental monitoring and urban planning. Because of sensor limitations, satellites usually acquire low-resolution multispectral (LRMS) images with rich spectral information and high-resolution panchromatic (PAN) images with fine spatial details. Pansharpening fuses them into an HRMS image with high fidelity in both domains.

Existing pansharpening methods mainly fall into traditional and deep-learning-based categories. Traditional methods—Component Substitution (CS), Multi-Resolution Analysis (MRA) and Variational Optimization (VO) (Vivone et al., 2015)—often struggle to balance spatial detail injection and spectral consistency. Deep models improve on them, but convolutions are limited by local receptive fields and Transformers incur quadratic complexity.

Recent visual state space models, such as VMamba (Liu et al., 2024) and Mamba2D (Baty et al., 2024), provide global receptive fields with linear-time complexity; Pan-Mamba (He et al., 2025) and FusionMamba (Peng et al., 2024) bring them into pansharpening, and recent studies have also explored quaternion interaction and diffusion-based modeling (Li et al., 2025; Xing et al., 2024; Zhong et al., 2024). Nevertheless, with few exceptions these methods recombine the modalities in a shared feature space, and any spatial/spectral split is not anchored to an explicit frequency decomposition; two issues therefore remain insufficiently addressed: (1) **Feature Entanglement**: jointly modeling PAN and LRMS features often trades spatial enhancement against spectral fidelity; and (2) **Inadequate Frequency Fusion**: simple fusion operators cannot effectively weight and integrate frequency-specific cues.

Wavelet transforms (Mallat, 1989) offer localized multi-scale representations and are therefore well suited to this problem. Recent frequency-aware restoration and wavelet-fusion models likewise

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

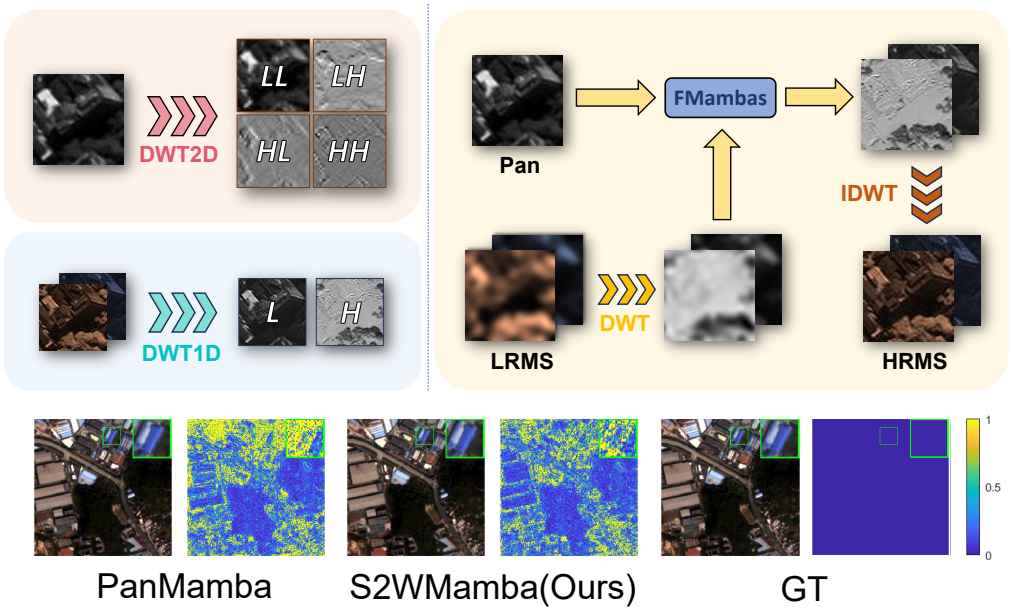

Figure 1: Overview of the proposed S2WMamba framework and visual comparison with state-of-the-art methods. The model utilizes modality-specific 2D and 1D Haar DWT to disentangle spatial-spectral features. These features are fused via FMamba modules and reconstructed through IDWT, achieving superior reconstruction quality.

highlight the value of selective frequency modulation and structured subband processing (Cui et al., 2025; Yang et al., 2024; Cui et al., 2024). However, existing wavelet-based methods often rely on shallow fusion and do not explicitly account for the distinct physical roles of spatial and spectral information.

To address these issues, we propose **S2WMamba** (Spatial–Spectral Wavelet Mamba), a pansharpening framework built on modality-specific wavelet disentanglement and efficient State Space Models. A 2D Haar DWT is applied to PAN features and a channel-wise 1D Haar DWT to the MS spectrum; a dual-branch cascade then injects spatial detail into MS features and spectral cues into PAN features, with the two branches interacting through Mamba-based cross-modulation (FMamba) at linear complexity. Fig. 1 illustrates the framework and compares it visually with competing methods.

In summary, our main contributions are:

- We propose **S2WMamba**, a pansharpening framework that uses modality-specific wavelet disentanglement (2D DWT for spatial structure and 1D DWT for spectral structure) to mitigate feature entanglement.
- We introduce a **channel-wise 1D Haar DWT** for remote sensing imagery to preserve spectral consistency and provide a perspective for related multimodal tasks.
- We adopt a bespoke **FMamba module**, adapted from the cross-modal selective-scan operator of FusionMamba (Peng et al., 2024), to model long-range interactions among decoupled wavelet sub-bands, improving frequency fusion with linear complexity.
- We obtain results that match or surpass strong baselines on WV3, GF2, and QB under reduced- and full-resolution settings, at a competitive parameter budget.

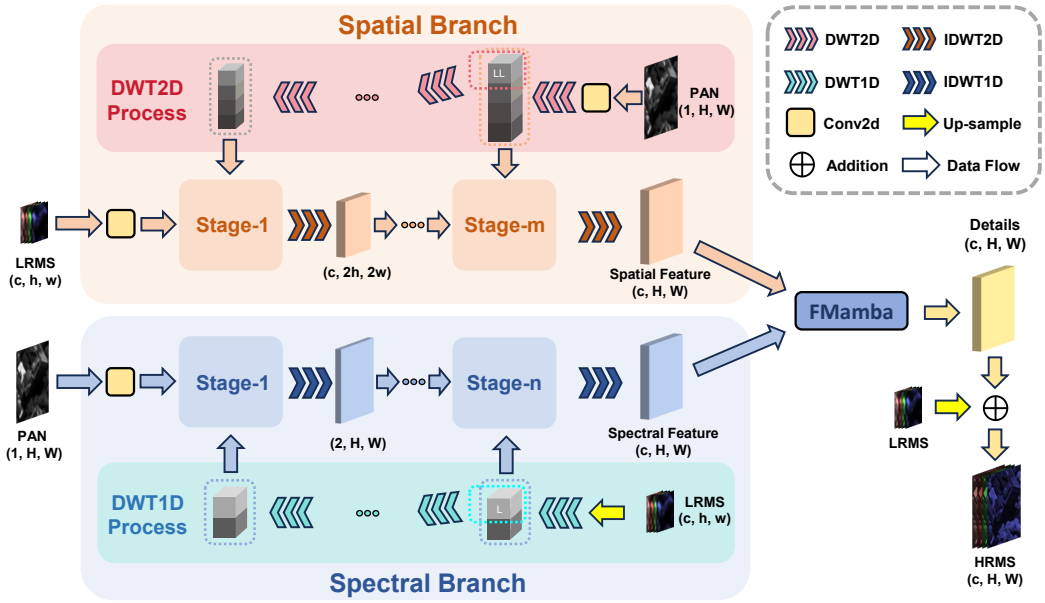

Figure 2: The overall workflow of our S2WMamba. Our network consists of two main components: a Spatial Branch (top), built on the 2D Haar DWT of the PAN image, and a Spectral Branch (bottom), built on the channel-wise 1D Haar DWT of the MS image. The two branches exchange information through FMamba at every stage, and their outputs are merged to produce the HRMS image.

2 Proposed Method

2.1 Overall Architecture

The main challenge in pansharpening is feature entanglement between spatial detail enhancement and spectral fidelity. To address this, we propose **S2WMamba**, shown in Fig. 2, which combines modality-specific wavelet disentanglement with efficient sequence modeling for controlled frequency-guided fusion.

Let the high-resolution panchromatic image be $\mathbf{P} \in \mathbb{R}^{B\times 1\times H\times W}$ and the low-resolution multispectral image be $\mathbf{M} \in \mathbb{R}^{B\times C\times \frac{H}{r}\times \frac{W}{r}}$, where B is the batch size, C is the number of spectral bands, and r is the resolution ratio (typically $r = 4$). We first bilinearly upsample $\mathbf{M}$ to obtain $\mathbf{M}_{up} \in \mathbb{R}^{B\times C\times H\times W}$.

S2WMamba consists of two phases: modality-specific wavelet disentanglement, followed by dual-branch cascaded fusion whose outputs are aggregated by a global FMamba module. The network is trained end-to-end with global residual learning:

$$\hat{\mathbf{Y}} = \mathbf{M}_{up} + \mathcal{F}_{net}(\mathbf{P}, \mathbf{M}) \tag{1}$$

where $\mathcal{F}_{net}(\cdot)$ extracts and injects the required spatial and spectral components.

2.2 Modality-Specific Wavelet Disentanglement

Unlike global transforms such as DCT or FFT, the Discrete Wavelet Transform (DWT) preserves both frequency and localization. We therefore design two wavelet processes matched to the roles of PAN and MS data.

2.2.1 2D Spatial Haar DWT for Structural Extraction

PAN mainly provides textures, edges, and structural boundaries, so we use a 2D Haar DWT to extract them. Let $f_L = \frac{1}{\sqrt{2}}[1, 1]$ and $f_H = \frac{1}{\sqrt{2}}[-1, 1]$ denote the 1D low-pass and high-pass Haar filters.

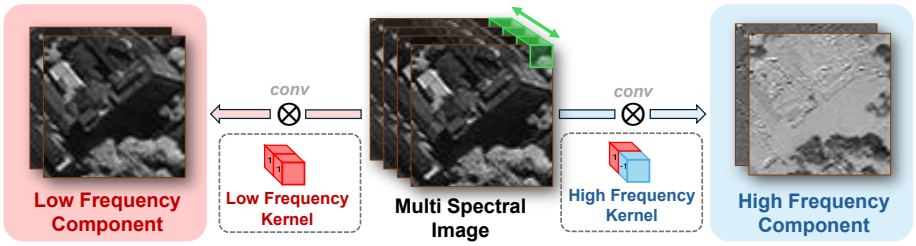

Figure 3: The core DWT1D strategy: low- and high-frequency kernels are applied along the channel axis, splitting the spectrum into low- and high-frequency components.

The 2D filters are defined by tensor products:

$$\mathbf{W}_{LL} = f_L \otimes f_L, \quad \mathbf{W}_{LH} = f_L \otimes f_H, \quad \mathbf{W}_{HL} = f_H \otimes f_L, \quad \mathbf{W}_{HH} = f_H \otimes f_H \tag{2}$$

Applying them to the convolved PAN feature $\mathbf{F}_P$ with stride 2 gives

$$\mathcal{P}^{(1)} = \{\mathbf{P}^{(1)}_{LL}, \mathbf{P}^{(1)}_{LH}, \mathbf{P}^{(1)}_{HL}, \mathbf{P}^{(1)}_{HH}\} = \mathrm{DWT}_{2D}(\mathbf{F}_P) \tag{3}$$

where $\mathbf{P}^{(1)}_{LL}$ is the coarse component and the other three sub-bands encode horizontal, vertical, and diagonal details. We recursively decompose the low-frequency component to obtain a multi-scale hierarchy:

$$\mathcal{P}^{(i+1)} = \{\mathbf{P}^{(i+1)}_{LL}, \mathbf{P}^{(i+1)}_{LH}, \mathbf{P}^{(i+1)}_{HL}, \mathbf{P}^{(i+1)}_{HH}\} = \mathrm{DWT}_{2D}(\mathbf{P}^{(i)}_{LL}), \quad i \in \{1, 2, 3...\} \tag{4}$$

2.2.2 1D Spectral Haar DWT for Spectral Consistency

We further introduce a channel-wise 1D Haar DWT for the MS image. Although spectral bands do not form a spatial grid, adjacent bands remain strongly correlated, so each pixel spectrum can be treated as a 1D signal.

Using stride-2 1D convolutions with kernels $f_L = \frac{1}{\sqrt{2}}[1, 1]$ and $f_H = \frac{1}{\sqrt{2}}[-1, 1]$ along the channel dimension, we obtain

$$\mathcal{S}^{(1)} = \{\mathbf{S}_L, \mathbf{S}_H\} = \mathrm{DWT}_{1D}(\mathbf{M}_{up}) \tag{5}$$

which re-expresses each pixel spectrum as a shared spectral baseline ($\mathbf{S}_L$) plus band-specific variations ($\mathbf{S}_H$), as shown in Fig. 3. Because the Haar transform is orthonormal, this step re-parameterizes the spectrum rather than separating it lossily: it changes the basis in which the following stages operate, so that the shared envelope and the band-specific variation can be weighted independently instead of one being discarded. As in the spatial branch, the decomposition is applied hierarchically:

$$\mathcal{S}^{(i+1)} = \{\mathbf{S}_L, \mathbf{S}_H\} = \mathrm{DWT}_{1D}(\mathbf{S}^{(i)}_L), \quad i \in \{1, 2, 3...\} \tag{6}$$

This separation reduces spectral distortion during spatial-detail injection.

2.3 Dual-Branch Cascaded Fusion

After disentanglement, S2WMamba fuses the two modalities with a dual-branch architecture in which each branch carries the shallow features of one modality—$\mathbf{F}_P$ for PAN and $\mathbf{F}_M$ for MS—and is refined by complementary frequency cues from the other. Within a stage both operands are first mapped to a common channel width d by a 3×3 convolution.

2.3.1 Spectral Branch: Spectral Guidance

The Spectral Branch carries the PAN-derived features $\mathbf{F}_P$ and refines them with the spectral structure of the MS image. Denoting the level-l 1D spectral sub-bands by $\mathcal{S}^{(l)} = \{\mathbf{S}^{(l)}_L, \mathbf{S}^{(l)}_H\}$, stage i cross-modulates the PAN representation through FMamba:

$$\mathbf{F}^{(i)}_{spe} = \mathrm{FMamba}(\mathbf{F}^{(L_{Spe}-i+1)}_P \mid \mathcal{S}^{(L_{Spe}-i+1)}) \tag{7}$$

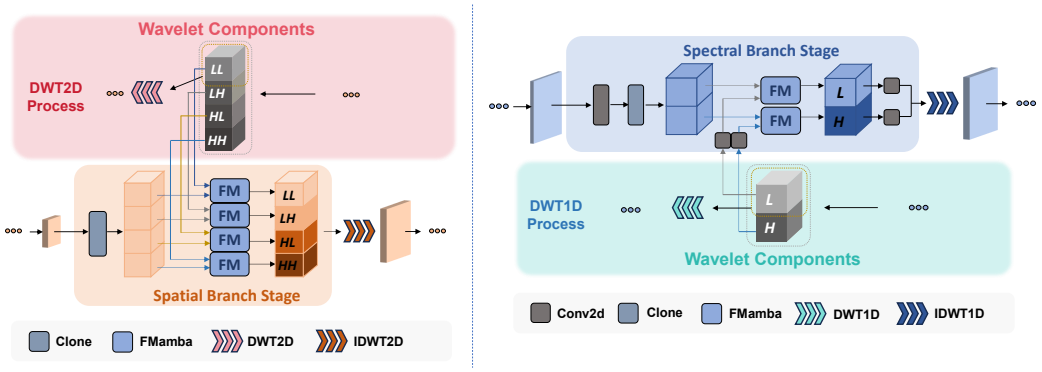

Figure 4: The internals of one stage in each branch, expanding the *Stage-i* blocks of Fig. 2. Left: in the Spatial Branch, the MS-derived features are the backbone and are modulated at each sub-band position by the four 2D Haar sub-bands of the PAN image, which are then fused back by IDWT2D. Right: in the Spectral Branch, the PAN-derived features are the backbone and are modulated by the channel-wise 1D Haar sub-bands of the MS image, which are fused back by IDWT1D. Tables 4 and 5 give the tensor shape at every level.

Cascading from the coarsest level upwards and reconstructing each stage by IDWT1D,

$$\mathbf{P}_{rec}^{(i-1)} = \mathrm{IDWT}_{1D}(\mathbf{F}_{spe}^{(i)}) \tag{8}$$

lets the per-pixel spectra of the MS image guide the refinement of the PAN features, yielding the branch output $\mathbf{F}_{spe}$.

2.3.2 Spatial Branch: Spatial Detail Injection

Symmetrically, the Spatial Branch carries the MS-derived features $\mathbf{F}_M$ and injects the structural detail captured by the 2D Haar decomposition of the PAN image:

$$\mathbf{F}_{spa}^{(i)} = \mathrm{FMamba}(\mathbf{F}_{M}^{(L_{Spa}-i+1)} \mid \mathcal{P}^{(L_{Spa}-i+1)}) \tag{9}$$

The enhanced features are fused back by IDWT2D:

$$\mathbf{M}_{rec}^{(i-1)} = \mathrm{IDWT}_{2D}(\mathbf{F}_{spa}^{(i)}) \tag{10}$$

producing the branch output $\mathbf{F}_{spa}$. Both cascades therefore run from the coarsest level of their pyramid to the finest.

Fig. 4 details a single stage of each branch and Algorithm 1 summarises the complete pipeline; Appendix H lists the tensor shapes at every level, with Table 4 covering the spatial cascade and Table 5 the spectral one. The two cascades need not have equal depth: the spatial cascade runs for $L_{Spa} = \log_2(r)$ stages and the spectral one for $L_{Spe} = \log_2(C)$ stages (WV3: 2 and 3; GF2/QB: 2 and 2). Finally, $\mathbf{F}_{spa}$ and $\mathbf{F}_{spe}$ are merged by a global FMamba module to produce the residual $\mathcal{F}_{net}(\mathbf{P}, \mathbf{M})$.

2.4 Cross-Modal Dynamical System via Parameterized State Spaces

To facilitate the state evolutions formulated above, FMamba must efficiently integrate the driving priors into the base representation. Standard self-attention mechanisms suffer from quadratic computational complexity $\mathcal{O}(N^2)$ with respect to spatial resolution $N = H \times W$, making them prohibitive for high-resolution pansharpening tasks. Instead, we build on the cross-modal selective-scan block of FusionMamba (Peng et al., 2024), re-derived here as a parameter-dependent continuous dynamical system.

For a continuous-time state space formulation, the sequence modeling maps an input continuous state $x(t)$ to an output $y(t)$ via a hidden latent state $h(t)$:

$$\frac{d}{dt}h(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) \tag{11}$$

Algorithm 1 S2WMamba Inference Pipeline

Require: PAN $\mathbf{P}$, LRMS $\mathbf{M}$, spatial stages L_{Spa}, spectral stages L_{Spe}
1: Extract shallow features: $\mathbf{F}_P \leftarrow f_P(\mathbf{P})$, $\mathbf{F}_M \leftarrow f_M(\mathbf{M})$
2: Build 2D wavelet pyramid from $\mathbf{F}_P$ and 1D spectral pyramid from $\mathbf{F}_M$
3: Initialize branch states: $\mathbf{H}^{(0)} \leftarrow \mathbf{F}_M$, $\mathbf{V}^{(0)} \leftarrow \mathbf{F}_P$
4: **for** $i = 1$ to L_{Spa} **do**
5: $\quad \mathbf{H}^{(i)} \leftarrow \text{FMamba}(\mathbf{H}^{(i-1)} \mid \mathcal{P}^{(L_{Spa}-i+1)})$
6: $\quad$ Reconstruct intermediate state by IDWT-2D
7: **end for**
8: **for** $j = 1$ to L_{Spe} **do**
9: $\quad \mathbf{V}^{(j)} \leftarrow \text{FMamba}(\mathbf{V}^{(j-1)} \mid \mathcal{S}^{(L_{Spe}-j+1)})$
10: $\quad$ Reconstruct intermediate state by IDWT-1D
11: **end for**
12: Fuse final states and predict residual $\mathbf{R}$
13: **return** $\hat{\mathbf{Y}} = \text{Upsample}(\mathbf{M}) + \mathbf{R}$

In traditional State Space Models (SSMs), the matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are statically learned. In our cross-modal context, we let the target wavelet sub-band $\mathbf{X}_{\text{cond}}$ (e.g., spectral base $\mathcal{S}$ or spatial details $\mathcal{P}$) parameterize the input-dependent terms of the state-space recursion of the primary feature $\mathbf{X}_{\text{base}}$. Precisely, the state-transition matrix $\mathbf{A}$ remains a static learned parameter; $\mathbf{B}$, $\mathbf{C}$ and the timescale Δ are generated from $\mathbf{X}_{\text{cond}}$:

$$\frac{d}{dt}h_{\text{base}}(t) = \mathbf{A}h_{\text{base}}(t) + \mathbf{B}(\mathbf{X}_{\text{cond}})x_{\text{base}}(t) \tag{12}$$

The state evolution of one modality is therefore driven by the frequency content of the other through these input-dependent terms.

Using a zero-order hold discretization rule with a timescale parameter Δ, the continuous parameters are converted to their discrete counterparts $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$, allowing efficient recurrent computation:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t \tag{13}$$

Within our architecture, the FMamba module instantiates two parallel sequence-to-sequence mappings. Let $\mathbf{X}_{\text{a}}$ and $\mathbf{X}_{\text{b}}$ denote the flattened input features. They undergo layer normalization (LN) and a projection mapping ($\mathcal{F}_{\text{proj}}$) with a residual connection:

$$\hat{\mathbf{X}}_{\text{a}} = \mathbf{X}_{\text{a}} + \mathcal{F}_{\text{proj}}(\text{LN}(\mathbf{X}_{\text{a}})), \quad \hat{\mathbf{X}}_{\text{b}} = \mathbf{X}_{\text{b}} + \mathcal{F}_{\text{proj}}(\text{LN}(\mathbf{X}_{\text{b}})) \tag{14}$$

Subsequently, the dual branches cross-modulate each other via the CrossMamba mapping given as Algorithm 2 in the appendix. Applying this to our dual modalities yields the spatial-guided and spectral-guided representations:

$$\mathbf{Y}_{\text{a}} = \text{CrossMamba}(\hat{\mathbf{X}}_{\text{a}} \mid \hat{\mathbf{X}}_{\text{b}}), \quad \mathbf{Y}_{\text{b}} = \text{CrossMamba}(\hat{\mathbf{X}}_{\text{b}} \mid \hat{\mathbf{X}}_{\text{a}}) \tag{15}$$

The two modulated features are then aggregated with a global skip connection:

$$\mathbf{Y}_{\text{out}} = \mathbf{Y}_{\text{a}} + \mathbf{Y}_{\text{b}} + \mathbf{X}_{\text{a}} + \mathbf{X}_{\text{b}} \tag{16}$$

This yields linear-time $\mathcal{O}(N)$ global modulation, linking the wavelet priors to the constraints of the dynamical system.

2.5 Loss Function

The entire network is trained end-to-end by minimizing the Mean Absolute Error (ℓ_1 loss) between the predicted HRMS and the ground truth (GT); ℓ_1 is widely preferred over ℓ_2 in low-level vision because it yields less blurry reconstructions (Zhao et al., 2017):

$$\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K}\left\|\hat{\mathbf{Y}}^{\{i\}} - I^{\{i\}}\right\|_1 \tag{17}$$

where K is the number of training samples, $\hat{\mathbf{Y}}^{\{i\}}$ is the network's output, and $I^{\{i\}}$ is the corresponding ground truth image.

3 EXPERIMENTS

3.1 SETTINGS

We evaluate S2WMamba on WorldView-3 (WV3), GaoFen-2 (GF2), and QuickBird (QB). Following Wald's protocol (Wald et al., 1997), we generate reduced-resolution PAN/LRMS/GT training pairs; for WV3, the sizes are 64×64 (PAN), 16×16×8 (LRMS), and 64×64×8 (GT). Data and preprocessing follow PanCollection (Deng et al., 2022). We compare with three traditional methods (MTF-GLP-FS (Vivone et al., 2018), BDSD-PC (Vivone, 2019), TV (Palsson et al., 2014)) and eight deep baselines (PNN (Masi et al., 2016), PanNet (Yang et al., 2017), DiCNN (He et al., 2019), FusionNet (Deng et al., 2021), Pan-Mamba (He et al., 2025), CANNet (Duan et al., 2024), U2Net (Peng et al., 2023), and FusionMamba (Peng et al., 2024)).

The model is trained in PyTorch with AdamW (Loshchilov & Hutter, 2019) at learning rate 4×10^{-4}, decayed by 0.7 every 100 epochs for 500 epochs, with batch size 32. We report SAM (Yuhas et al., 1992), ERGAS (Wald, 2002), and Q4/Q8 (Garzelli & Nencini, 2009) at reduced resolution, and HQNR, D_s, and D_λ (Arienzo et al., 2022) at full resolution.

3.2 ASSESSMENTS

3.2.1 COMPARISON TO EXISTING STATE OF THE ART METHODS.

S2WMamba differs from prior work in three ways: it keeps *linear* sequence complexity relative to Transformer-based PanFormer (Zhou et al., 2022); it decomposes each modality into wavelet sub-bands before any cross-modal interaction, so the FMamba operator we adopt from Fusion-Mamba (Peng et al., 2024) acts on decoupled frequency bands rather than on the shared features used by Pan-Mamba (He et al., 2025); and it fuses sub-bands in a dual-branch cascade rather than a content-adaptive non-local operator (Duan et al., 2024).

3.2.2 REDUCED AND FULL RESOLUTION ASSESSMENTS.

Table 1 reports the quantitative results. S2WMamba attains the best reduced-resolution PSNR on all three sensors, though the margins differ in size: 0.231 dB over the second-best method on GF2 and 0.045 dB on QB, against only 0.017 dB on WV3, where the strongest methods are effectively tied. It also attains the highest full-resolution HQNR on all three sensors, though on WV3 the margin over the strongest baseline is within the per-image spread; Fig. 5 shows the darkest residual maps. On WV3 the model uses 0.63M parameters, between FusionMamba (0.73M) and Pan-Mamba (0.48M). Appendix G gives a spatially resolved view of full-resolution quality through local HQNR maps and restates the reduced-resolution PSNR column as a bar chart.

3.3 ABLATION ANALYSIS: VALIDATING THE DYNAMICAL HYPOTHESES

We validate the design of S2WMamba through ablations on the dynamical constraints and frequency disentanglement of Section 2.4; results are listed in Table 2.

3.3.1 NECESSITY OF SYMMETRICAL FREQUENCY CONSTRAINTS.

Removing the Spectral Branch (**SpeO**) costs 0.622 dB PSNR, indicating that spatial enhancement without spectral guidance introduces color distortion; removing the Spatial Branch (**SpaO**) costs 0.657 dB, showing that spectral modeling alone cannot recover fine textures; the two branches are therefore complementary and both necessary.

3.3.2 PARALLEL VS. SEQUENTIAL FUSION.

The parallel design outperforms both sequential variants: Spectral→Spatial (**SeqB1**) costs 0.198 dB and Spatial→Spectral (**SeqB2**) costs 0.286 dB, since sequential fusion introduces bottlenecks and error accumulation.

Table 1: Quantitative results on the WV3, GF2 and QB datasets at both reduced and full resolutions. (Best: **bold**, second-best: underline.)

Methods	WV3 (Reduced-resolution)				WV3 (Full-resolution)		
	PSNR↑	SAM↓	ERGAS↓	Q8↑	D_λ ↓	D_s ↓	HQNR↑
MTF-GLP-FS	32.963	5.316	4.700	0.833	0.020	0.063	0.919
BDSD-PC	32.970	5.428	4.697	0.829	0.063	0.073	0.870
TV	32.381	5.692	4.855	0.795	0.023	0.039	0.938
PNN	37.313	3.677	2.681	0.893	0.021	0.043	0.937
PanNet	37.346	3.613	2.664	0.891	**0.017**	0.047	0.937
DiCNN	37.390	3.592	2.672	0.900	0.036	0.046	0.920
FusionNet	38.047	3.324	2.465	0.904	0.024	0.036	0.941
Pan-Mamba	39.012	2.913	2.184	0.920	0.018	0.031	0.952
CANNet	39.003	2.941	2.174	0.920	0.020	0.030	0.951
U2Net	39.117	2.888	2.150	0.920	0.018	0.031	0.951
FusionMamba	39.374	2.844	2.092	0.922	0.019	0.027	0.955
S2WMamba (Ours)	**39.391**	**2.825**	**2.087**	**0.923**	0.024	**0.021**	**0.956**
Methods	**GF2 (Reduced-resolution)**				**GF2 (Full-resolution)**		
	PSNR↑	SAM↓	ERGAS↓	Q4↑	D_λ ↓	D_s ↓	HQNR↑
MTF-GLP-FS	41.565	1.655	1.589	0.897	0.035	0.143	0.828
BDSD-PC	41.205	1.681	1.667	0.892	0.076	0.155	0.781
TV	41.262	1.911	1.737	0.907	0.055	0.112	0.839
PNN	45.096	1.048	1.057	0.960	0.032	0.094	0.877
PanNet	46.268	0.997	0.919	0.967	0.018	0.080	0.904
DiCNN	44.931	1.053	1.081	0.959	0.037	0.099	0.868
FusionNet	45.663	0.974	0.988	0.964	0.035	0.101	0.867
Pan-Mamba	48.931	0.743	0.684	0.982	0.023	0.057	0.921
CANNet	49.520	0.707	0.630	0.983	0.019	0.063	0.919
U2Net	49.404	0.714	0.632	0.982	0.024	0.051	0.927
FusionMamba	49.678	0.705	0.615	**0.984**	0.017	0.030	0.954
S2WMamba (Ours)	**49.909**	**0.676**	**0.599**	**0.984**	**0.016**	**0.028**	**0.957**
Methods	**QB (Reduced-resolution)**				**QB (Full-resolution)**		
	PSNR↑	SAM↓	ERGAS↓	Q4↑	D_λ ↓	D_s ↓	HQNR↑
MTF-GLP-FS	32.709	7.792	7.373	0.835	0.047	0.150	0.811
BDSD-PC	32.550	8.085	7.513	0.831	0.198	0.164	0.672
TV	32.136	7.510	7.690	0.821	0.055	0.101	0.850
PNN	36.942	5.181	4.468	0.918	0.058	0.062	0.884
PanNet	34.678	5.767	5.859	0.885	0.043	0.114	0.849
DiCNN	35.781	5.367	5.133	0.904	0.095	0.107	0.809
FusionNet	37.540	4.904	4.156	0.925	0.057	0.052	0.894
Pan-Mamba	37.356	4.625	4.277	0.929	0.049	0.044	0.910
CANNet	38.488	4.496	3.698	0.937	0.037	0.050	0.915
U2Net	38.065	4.642	3.987	0.931	0.059	0.026	0.916
FusionMamba	37.986	4.610	4.054	0.930	0.057	0.040	0.906
S2WMamba (Ours)	**38.533**	**4.445**	**3.679**	**0.938**	**0.035**	**0.024**	**0.942**

3.3.3 Effectiveness of Mamba-based Cross-Modulation.

Replacing FMamba with concatenation followed by a residual block (**CRM**) costs 0.521 dB, and cross-attention at a similar parameter budget (**ARM**) also underperforms, suggesting that Mamba's conditional state transition integrates the wavelet priors more effectively.

4 Conclusion

This paper presented S2WMamba, a pansharpening framework that resolves spatial-spectral feature entanglement through modality-specific wavelet disentanglement. A 2D Haar DWT decomposes PAN-derived features into coarse and directional sub-bands, while a channel-wise 1D Haar DWT re-parameterizes each pixel spectrum as a shared spectral base plus band-specific residuals. A dual-branch cascade recombines these representations through FMamba at every stage, a condition-parameterized state-space operator with linear cost. The ablations confirm that both halves are necessary—removing either branch costs over 0.6 dB and replacing FMamba with concatenation

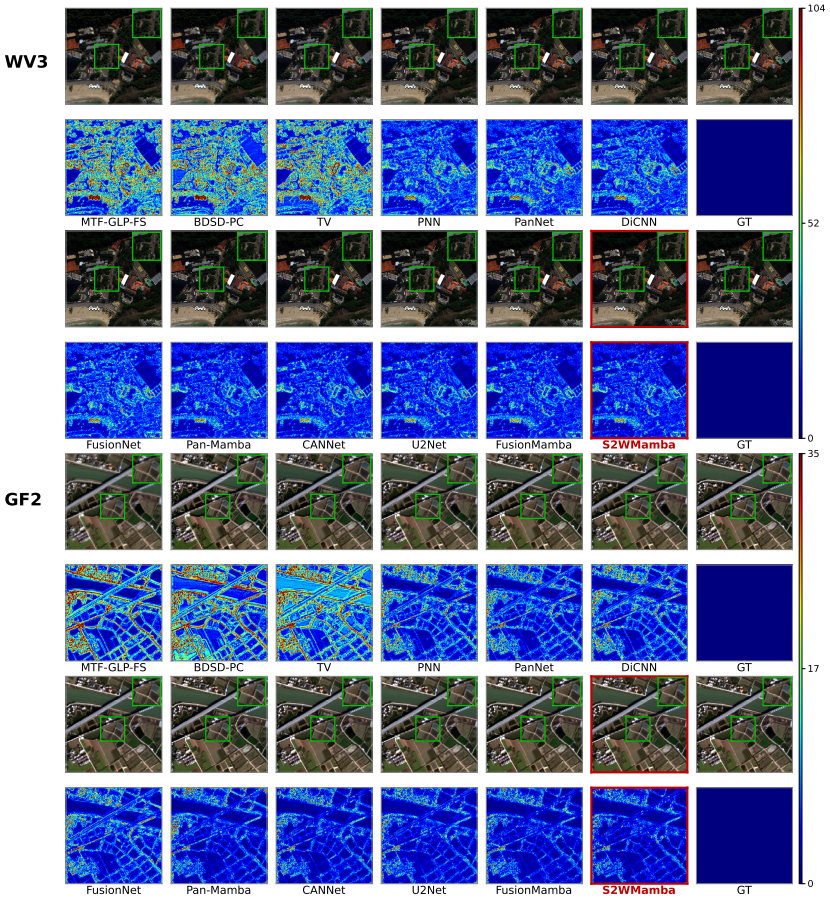

Figure 5: Visual results and residuals on the WV3 (top) and GF2 (bottom) reduced-resolution test set. Each pair shows a fused product with a magnified inset, and below it the band-averaged absolute residual $|\hat{M} - I|$ against the ground truth, on a per-sensor colour scale shared by all methods. S2WMamba is outlined in red.

Table 2: Ablation on key components and strategy, WV3 reduced-resolution dataset. **Orig** is the complete model; the remaining rows are the ablated variants defined in the text.

Ablation	PSNR↑	SAM↓	ERGAS↓	Q8↑
SpeO	38.769	3.100	2.252	0.916
SpaO	38.734	3.010	2.259	0.916
SeqB1	39.193	2.876	2.136	0.921
SeqB2	39.105	2.969	2.163	0.918
CRM	38.870	2.986	2.244	0.919
ARM	39.295	2.897	2.105	0.922
Orig	**39.391**	**2.825**	**2.087**	**0.923**

a further 0.521 dB—and Fig. 5 shows the residual maps of the proposed cascade to be the least structured of those compared.

LIMITATIONS

Our comparisons use a single random seed, so margins over the strongest baselines carry no claim of statistical significance; further limitations, including the band-count assumption of the 1D DWT and generalization to unseen sensors, are in Appendix I.

REPRODUCIBILITY STATEMENT

All experiments use the PanCollection benchmark (Deng et al., 2022) under Wald's protocol (Wald et al., 1997). For reduced-resolution training and evaluation, the reference HRMS is low-pass filtered and decimated to build the LRMS/PAN inputs; for WV3 a network input consists of a $16\times16\times8$ LRMS patch and a 64×64 PAN patch with a $64\times64\times8$ HRMS target, and the analogue holds for GF2 and QB. We evaluate on the standard 20-sample test split of each sensor and report means over those 20 samples. Assessment follows the public PanCollection toolbox: reduced resolution metrics are computed after cropping a 30-pixel border (PSNR band-averaged; SAM averaging per-pixel spectral angles; ERGAS aggregating per-band errors; Q4/Q8 jointly over bands), while full-resolution metrics (D_λ, D_s, HQNR) use 32-pixel blocks without a border crop. The compared methods were run under this identical protocol using publicly released implementations—the parameter-free classical methods with their authors' recommended settings, and the deep baselines with their official released weights—and are assessed with the same public toolbox and the same settings.

Models are trained with AdamW (Loshchilov & Hutter, 2019) (weight decay 0), initial learning rate 4×10^{-4} decayed by a factor of 0.7 every 100 epochs, and batch size 32, on a single NVIDIA RTX 4090 (24 GB). All reported results come from a single training run per sensor under a fixed random seed; we do not report variance across seeds. Code and trained weights will be released upon acceptance; the script used to rebuild and verify the bibliography is released with them.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: implementing parts of the model code. We have not used generative AI tools for proposing or refining the research hypotheses, for designing or refining the experimental methodology, or for interpreting the experimental results. We have also not used them to generate synthetic datasets or to prove mathematical claims, and these tasks are not applicable to this work.

Additionally, we used generative AI tools for the following tasks with recommended disclosure: drafting and editing portions of the manuscript text, identifying related literature, and formatting bibliographic entries.

We have reviewed and take responsibility for all AI-assisted work. In particular, every reference in this paper was independently verified against the publisher's own record or against a standard catalogue record (DOI, Crossref, DBLP): each entry was checked to exist, to carry the stated author list, and to appear in the stated venue, and any entry that could not be verified in this way was removed rather than retained. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Alberto Arienzo, Gemine Vivone, Andrea Garzelli, Luciano Alparone, and Jocelyn Chanussot. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):168–201, 2022. doi: 10.1109/mgrs.2022.3170092.

Enis Baty, Alejandro Hernández Díaz, Rebecca Davidson, Chris Bridges, and Simon Hadfield. Mamba2d: A natively multi-dimensional state-space model for vision tasks. *arXiv preprint arXiv:2412.16146v3*, 2024.

Joseph W. Boardman and Fred A. Kruse. Automated spectral analysis: A geological example using aviris data, north grapevine mountains, nevada. In *Proceedings of the Tenth Thematic Conference on Geologic Remote Sensing*, pp. 407–418, San Antonio, Texas, USA, 1994.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1093–1108, 2024. doi: 10.1109/tpami.2023.3330416.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, 2025. arXiv:2403.14614.

Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2021. doi: 10.1109/tgrs.2020.3031366.

Liang-jian Deng, Gemine Vivone, Mercedes E. Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022. doi: 10.1109/mgrs.2022.3187652.

Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27738–27747. IEEE, 2024. doi: 10.1109/cvpr52733.2024.02620.

A. Garzelli and F. Nencini. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):662–665, 2009. doi: 10.1109/lgrs.2009.2022650.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019. doi: 10.1109/jstars.2019.2898574.

Xuanhua He, Ke Cao, Jie Zhang, Keyu Yan, Yingying Wang, Rui Li, Chengjun Xie, Danfeng Hong, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115:102779, 2025. doi: 10.1016/j.inffus.2024.102779.

Dong Li, Chunhui Luo, Yuanfei Bao, Gang Yang, Jie Xiao, Xueyang Fu, and Zheng-Jun Zha. Enhanced pansharpening via quaternion spatial-spectral interactions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10908–10918. IEEE, 2025. doi: 10.1109/iccv51701.2025.01015.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems 37*, pp. 103031–103063. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3273.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. arXiv:1711.05101.

S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. doi: 10.3390/rs8070594.

Frosti Palsson, Johannes R. Sveinsson, and Magnus O. Ulfarsson. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters*, 11(1):318–322, 2014. doi: 10.1109/lgrs.2013.2257669.

Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3219–3227. ACM, 2023. doi: 10.1145/3581783.3612084.

Siran Peng, Xiangyu Zhu, Haoyu Deng, Liang-Jian Deng, and Zhen Lei. Fusionmamba: Efficient remote sensing image fusion with state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/tgrs.2024.3496073.

Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6421–6433, 2019. doi: 10.1109/tgrs.2019.2906073.

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 Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015. doi: 10.1109/tgrs.2014.2361734.

Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing*, 27(7):3418–3431, 2018. doi: 10.1109/tip.2018.2819501.

Lucien Wald. *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Les Presses de l'École des Mines, Paris, France, 2002.

Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63(6):691–699, 1997.

Yinghui Xing, Litao Qu, Shizhou Zhang, Kai Zhang, Yanning Zhang, and Lorenzo Bruzzone. Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model. *IEEE Transactions on Image Processing*, 33:5496–5509, 2024. doi: 10.1109/tip.2024.3461476.

Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1753–1761. IEEE, 2017. doi: 10.1109/iccv.2017.193.

Yunsong Yang, Genji Yuan, and Jinjiang Li. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. doi: 10.1109/tgrs.2024.3427370.

Roberta H. Yuhas, Alexander F. H. Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Summaries of the 3rd Annual JPL Airborne Geoscience Workshop*, pp. 147–149, Pasadena, CA, USA, 1992.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi: 10.1109/tci.2016.2644865.

Yu Zhong, Xiao Wu, Zihan Cao, Hong-Xia Dou, and Liang-Jian Deng. Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. In *Advances in Neural Information Processing Systems 37*, pp. 77962–77986. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2478.

Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pan-sharpening. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2022. doi: 10.1109/icme52920.2022.9859770.

A Notations and Problem Re-Statement

Given PAN image $\mathbf{P} \in \mathbb{R}^{1\times rH\times rW}$ and LRMS image $\mathbf{M} \in \mathbb{R}^{c\times H\times W}$, pansharpening seeks an HRMS estimate $\hat{\mathbf{Y}} \in \mathbb{R}^{c\times rH\times rW}$. Following Wald's protocol (Wald et al., 1997), an ideal solution should preserve spectral consistency with $\mathbf{M}$ after downsampling while inheriting high-frequency geometry from $\mathbf{P}$.

S2WMamba factorizes this into two coupled latent trajectories:

$$\mathcal{T}_{spa}:\ (\mathbf{M}, \mathcal{W}_{2D}(\mathbf{P})) \mapsto \mathbf{F}_{spa}, \qquad \mathcal{T}_{spe}:\ (\mathbf{P}, \mathcal{W}_{1D}(\mathbf{M})) \mapsto \mathbf{F}_{spe}, \tag{18}$$

where $\mathcal{W}_{2D}$ is 2D Haar DWT and $\mathcal{W}_{1D}$ is channel-wise 1D Haar DWT.

B Why Modality-Specific DWT is Well-Posed

B.1 Spatial branch as constrained detail injection

The PAN decomposition produces $\{\mathbf{P}^l_{LL}, \mathbf{P}^l_{LH}, \mathbf{P}^l_{HL}, \mathbf{P}^l_{HH}\}_{l=1}^{L_{Spa}}$. In each stage, only sub-band-conditioned updates are allowed:

$$\mathbf{H}^{(i)} = \mathbf{H}^{(i-1)} + \Phi_i\Big(\mathbf{H}^{(i-1)}; \mathbf{P}^{(L_{Spa}-i+1)}_{LL,LH,HL,HH}\Big). \tag{19}$$

This form provides an inductive bias: the structural update is conditioned on wavelet detail carriers rather than left free-form.

B.2 Spectral branch as low/high-band manifold correction

For each pixel location, 1D Haar analysis over channels yields low-frequency envelope and inter-band residual terms. Let $\mathbf{s} \in \mathbb{R}^c$ be a spectral vector. The decomposition recursively satisfies

$$\mathbf{s} = \mathcal{W}^{-1}_{1D}\Big(\mathbf{s}_L^{(L_{Spe})}, \mathbf{s}_H^{(L_{Spe})}, \ldots, \mathbf{s}_H^{(1)}\Big). \tag{20}$$

Using these components as controls provides an inductive bias towards physically plausible cross-band relationships, consistent with spectral-mixing insights in remote sensing (Boardman & Kruse, 1994; Deng et al., 2022).

C Conditional State-Space Interpretation of FMamba

FMamba follows selective SSM principles (Gu & Dao, 2024; Liu et al., 2024). For base sequence $\mathbf{x}_t$ and condition sequence $\mathbf{z}_t$, the input-dependent parameters $\mathbf{B}$, $\mathbf{C}$ and Δ are generated from the condition, while $\mathbf{A}$ stays a static learned parameter, exactly as in Sec. 2.4:

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}(\mathbf{z}_t)\mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}(\mathbf{z}_t)\mathbf{h}_t. \tag{21}$$

Compared with static convolutions or static attention kernels, this mechanism can be interpreted as *input-dependent linear dynamical evolution* with linear-time scan complexity.

Proposition 1. *Assume stage operators are Lipschitz: $\|\Phi_i(\mathbf{u}) - \Phi_i(\mathbf{v})\| \le \kappa_i\|\mathbf{u}-\mathbf{v}\|$ and residual scaling $\alpha_i \in (0,1]$. If $\prod_i(1+\alpha_i\kappa_i)$ is bounded, then perturbation propagation from inputs to branch outputs is bounded.*

Proof. Let ϵ_0 be an input perturbation and ϵ_i its image after stage i. The Lipschitz assumption together with the residual scaling gives $\|\epsilon_i\| \le (1+\alpha_i\kappa_i)\|\epsilon_{i-1}\|$ for every i. Applying this inequality recursively over the L stages yields

$$\|\epsilon_L\| \le \Big(\prod_{i=1}^{L}(1+\alpha_i\kappa_i)\Big)\|\epsilon_0\|, \tag{22}$$

so the perturbation at the branch output is bounded whenever the product is. □

This proposition is not a training guarantee; it only provides a stability-oriented interpretation that supports the staged residual design.

Algorithm 2 CrossMamba Block. Here d is the channel width of the flattened feature, n the SSM state dimension, and HW the number of spatial positions.

Inputs: $\mathbf{x}_{\mathrm{a}}, \mathbf{x}_{\mathrm{b}} : (HW, d)$
Output: $\mathbf{y}_{\mathrm{a}} : (HW, d)$
$\mathbf{A} : (d, n) \leftarrow \mathbf{Parameter}_{\mathrm{A}}$
/* $\mathbf{A}$ represents d sets of structured $n \times n$ matrices */
$\mathbf{B} : (HW, n) \leftarrow \mathbf{Linear}_{\mathrm{B}}(\mathbf{x}_{\mathrm{b}})$
$\mathbf{C} : (HW, n) \leftarrow \mathbf{Linear}_{\mathrm{C}}(\mathbf{x}_{\mathrm{b}})$
$\mathbf{\Delta} : (HW, d) \leftarrow \log(1 + \exp(\mathbf{Linear}_{\Delta}(\mathbf{x}_{\mathrm{b}}) + \mathbf{Parameter}_{\Delta}))$
/* $\mathbf{Parameter}_{\Delta}$ is a bias vector with a size of d */
$\bar{\mathbf{A}} : (HW, d, n) \leftarrow \exp(\mathbf{\Delta} \otimes \mathbf{A})$
$\bar{\mathbf{B}} : (HW, d, n) \leftarrow \mathbf{\Delta} \otimes \mathbf{B}$
$\mathbf{y}_{\mathrm{a}} \leftarrow \mathrm{SSM}(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{C})(\mathbf{x}_{\mathrm{a}})$
/* SSM represents Eq. (13) implemented using selective scan */
return $\mathbf{y}_{\mathrm{a}}$

Table 3: Symbol-level complexity comparison (per feature map with $N{=}HW$, channel width D (distinct from the band count C), convolution kernel size k, SSM state dimension n).

Module type	Time complexity	Memory trend
Self-attention fusion	$\mathcal{O}(N^2 D)$	quadratic in N
Convolutional fusion	$\mathcal{O}(k^2 N D^2)$	linear in N
FMamba-style selective scan	$\mathcal{O}(ND)$ to $\mathcal{O}(NDn)$	linear in N

D Expanded Procedure

Algorithm 2 details the CrossMamba block that FMamba instantiates internally at every stage; the end-to-end S2WMamba pipeline is given as Algorithm 1 in the main text.

E Implementation Clarifications

Table 3 is a theoretical comparison to explain scaling behavior; it does not constitute a runtime benchmark claim. In practice, constants depend on implementation details, operator fusion, and hardware libraries.

Our training objective is a single ℓ_1 reconstruction term (Eq. (17)); we add no perceptual or structural-fidelity term. Importantly, our supplementary analysis does not alter any setting, split, or metric definition reported in the main manuscript.

F Discussion

The central message is that S2WMamba is not merely stacking wavelet and Mamba blocks. Its key principle is *bidirectional constrained evolution*: PAN high-frequency components regulate spatial-detail transfer, while MS spectral wavelet components regulate cross-band fidelity. This dual constraint helps bridge traditional physics-informed intuition and modern sequence-based representation learning. We hope these derivations make the design choices easier to reproduce and audit.

G Visualization Analysis

Figure 6 restates the reduced-resolution PSNR column of Table 1 for every compared method. Figure 7 gives a spatially resolved view of full-resolution quality, where each pixel reports the local balance between spectral and spatial fidelity rather than a single dataset-level score.

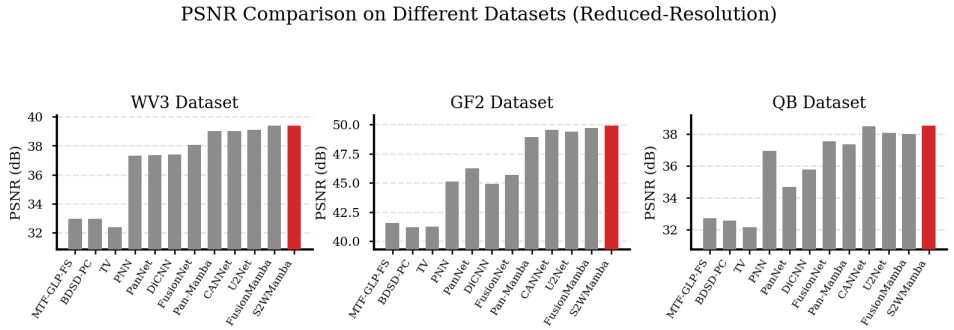

Figure 6: Reduced-resolution PSNR (dB) of every compared method on the WV3, GF2 and QB datasets; the red bar is S2WMamba.

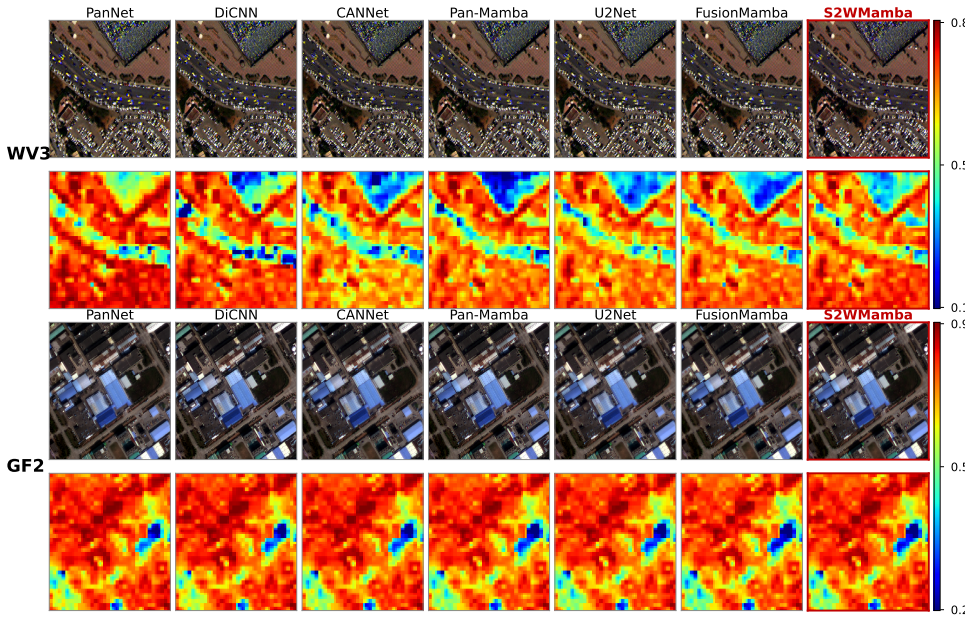

Figure 7: Local HQNR on full-resolution WV3 (top) and GF2 (bottom) for several deep pansharpening methods and our S2WMamba. Each pixel is $(1 - D_\lambda)(1 - D_s)$ computed over a 32×32 sliding patch, and the colour scale is shared by all methods within a sensor row. Warmer colours denote a better local balance between spatial detail and spectral fidelity. S2WMamba is outlined in red.

H Stage-wise Tensor Shapes

Tables 4 and 5 list the tensor shape at every level of the two branch cascades described in Section 2.

I Limitations

Beyond the single-seed protocol noted in the main text, our study has three further limitations. First, the channel-wise 1D DWT assumes that adjacent spectral bands are strongly correlated and terminates cleanly only when the band count is a power of two; other band counts, or bands not ordered by spectral proximity, would require a padding or reordering strategy that we do not evaluate. Second, our evaluation covers three sensors at a fixed ratio $r{=}4$, and generalization to unseen sensors remains untested. Finally, we report no dedicated failure-case analysis: strongly textured scenes remain the regime in which residual spectral distortion is most visible.

Table 4: Spatial Branch: stage-wise tensors (all batches omitted for brevity), r=4 (for WV3, GF2 and QB). The branch carries the MS-derived backbone $\mathbf{F}_M$; the four 2D Haar sub-bands of the convolved PAN feature supply the conditioning input at every stage.

r=4 (WV3, GF2, QB): 2 Stages	
Backbone (MS)	$\mathbf{F}_M \in \mathbb{R}^{C\times H\times W}$
Conditioning (convolved PAN)	$\mathbf{F}_P \in \mathbb{R}^{C\times H\times W}$
Level-1 DWT2D	$[\mathrm{LL}_1, \mathrm{LH}_1, \mathrm{HL}_1, \mathrm{HH}_1] \in \mathbb{R}^{4C\times \frac{H}{2}\times \frac{W}{2}}$
Level-2 DWT2D	$[\mathrm{LL}_2, \mathrm{LH}_2, \mathrm{HL}_2, \mathrm{HH}_2] \in \mathbb{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$
FMamba (Spatial Branch Stage 1)	$[\mathrm{F}_{LL_2}, \mathrm{F}_{LH_2}, \mathrm{F}_{HL_2}, \mathrm{F}_{HH_2}] \in \mathbb{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$
IDWT2D (Spatial Branch Stage 1)	$M_1 \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$
FMamba (Spatial Branch Stage 2)	$[\mathrm{F}_{LL_1}, \mathrm{F}_{LH_1}, \mathrm{F}_{HL_1}, \mathrm{F}_{HH_1}] \in \mathbb{R}^{4C\times \frac{H}{2}\times \frac{W}{2}}$
IDWT2D (Spatial Branch Stage 2)	$M_2 \in \mathbb{R}^{C\times H\times W}$
Reduce to C	$\mathbf{F}_{spa} \in \mathbb{R}^{C\times H\times W}$

Table 5: Spectral Branch: channel-wise DWT1D on C bands (non-interleaved $[L, H]$ layout). The branch carries the PAN-derived backbone $\mathbf{F}_P$; the 1D spectral sub-bands of the MS image supply the conditioning input at every stage.

C=8 (WV3 for example): 3 Stages	
Backbone (PAN)	$\mathbf{F}_P \in \mathbb{R}^{C\times H\times W}$
Level-3 DWT1D: L_3, H_3	$\mathbb{R}^{1\times H\times W}$
IDWT1D (Spectral Branch Stage 1)	$\mathbb{R}^{2\times H\times W}$
Level-2 DWT1D: L_2, H_2	$\mathbb{R}^{2\times H\times W}$
IDWT1D (Spectral Branch Stage 2)	$\mathbb{R}^{4\times H\times W}$
Level-1 DWT1D: L_1, H_1	$\mathbb{R}^{4\times H\times W}$
IDWT1D (Spectral Branch Stage 3)	$\mathbb{R}^{8\times H\times W}$
C=4 (GF2/QB for example): 2 Stages	
Level-2 DWT1D: L_2, H_2	$\mathbb{R}^{1\times H\times W}$
IDWT1D (Spectral Branch Stage 1)	$\mathbb{R}^{2\times H\times W}$
Level-1 DWT1D: L_1, H_1	$\mathbb{R}^{2\times H\times W}$
IDWT1D (Spectral Branch Stage 2)	$\mathbb{R}^{4\times H\times W}$

J Practical Reproducibility Notes

For reproducibility, we recommend keeping three implementation invariants unchanged when reusing the framework on new sensors. First, keep stage depth tied to sensor geometry ($L_{Spa} = \log_2 r$) and spectral dimensionality ($L_{Spe} = \log_2 C$ when feasible), rather than treating both as unconstrained hyperparameters. Second, preserve the branch semantics: the PAN-conditioned path should remain the dominant source of spatial edges, while the MS-conditioned path should remain the dominant source of spectral priors. Third, if replacing FMamba with another linear-time operator, ensure the operator is still condition-parameterized; otherwise, the model degenerates to feature mixing without explicit cross-modal dynamical control. These constraints are lightweight but important for maintaining consistency with the theoretical formulation and with the training behavior reported in the main paper.

K Conclusion

This supplementary material explains S2WMamba from three perspectives: frequency-domain decomposition, conditional state-space fusion, and implementation-level reproducibility. The added derivations and algorithmic description clarify that the method is driven by explicit cross-modal constraints rather than unconstrained feature mixing. Specifically, modality-specific wavelet priors guide spatial detail transfer and spectral consistency preservation, while FMamba enables linear-time conditioned interaction. Together with the HQNR visualization analysis, these results explain the

framework's effectiveness on full-resolution remote-sensing scenes and make the method easier to understand, reproduce, and extend in future pansharpening research.