SEPARATE AND VARY TOGETHER: CROSS-SENSOR PANSHARPENING WITH REVERSIBLE SPECTRAL TRANSPORT AND PAIRED SENSOR TRANSFORMATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pansharpening fuses a low-resolution multispectral (MS) image with a panchromatic (PAN) image. Existing cross-sensor methods rely on per-sensor training, multi-sensor data, or target-time adaptation, which are costly and require target labels or weight updates. We study a restrictive setting: train once on a source sensor and deploy the same weights on unseen sensors. The challenge is that different sensors observe different band counts, centers, and spectral responses, and the PAN–MS relation varies. A fixed input–output mapping learned on one sensor is therefore tied to that sensor's observation process and difficult to transfer. We instead learn how the output should change when the sensing process changes, formulated as approximate equivariance under sensor transformations: fusion should commute with a sampled transformation applied jointly to the source MS observations and their HRMS targets. Since the transformation is sampled from source data and paired targets are exact, this behavior is supervised without target data, enabling source-only training. We propose EquiPan, which promotes this approximate equivariance through two complementary mechanisms. A Spectral–Contrast Transformation (SCT) first samples virtual observation models from source data and applies each transformation consistently to the MS observations and their HRMS target, providing exact supervision of how the fusion output should change with the sensor. A Reversible Spectral Transport (RST) then projects the resulting heterogeneous observations into a shared latent space and maps predicted residuals back to their native bands through an analytic left inverse on the observed spectral degrees of freedom. A parameter-free PAN mean alignment (PMA) operation removes branch-dependent global offsets while preserving the measured PAN contrast. Across seven backbones and three unseen sensors, EquiPan improves zero-shot transfer over traditional and interpolation-only baselines.

1 INTRODUCTION

Pansharpening transfers fine spatial detail from a high-resolution panchromatic (PAN) image to a low-resolution multispectral (MS) image of the same scene, reconstructing a high-resolution multispectral (HRMS) image without corrupting the band-specific spectral measurements of MS (Vivone et al., 2015; 2021). Deep networks learn this transfer from paired LRMS, PAN, and HRMS examples (Masi et al., 2016; Yang et al., 2017). Because these pairs are sensor-specific, a network trained on one sensor is tied to its observation process and generally requires retraining for another.

This motivates cross-sensor generalization, where a model trained on a source sensor is deployed on unseen sensors. Existing cross-sensor methods rely on multi-sensor data, input-distribution alignment, or target-time adaptation (Wang et al., 2026a; Cui et al., 2025; Xin et al., 2026; Xia et al., 2026). These approaches require additional sensor diversity, target-side processing, or deployment-time optimization. We instead study a novel setting: train once on a source sensor and deploy the same weights on unseen sensors. A central challenge is that the target sensing process is unknown during source training. Table 1 illustrates representative differences among the sensors used in this study.

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.

Table 1: Representative differences among the four sensors used in this study. PAN support and native nadir ground sample distances (GSDs) follow sensor documentation (DigitalGlobe, 2014; 2016; 2003; Chen et al., 2022).

Sensor	Role	MS bands	MS centers supplied to RST (nm)	PAN support (nm)	PAN/MS GSD (m)
WV3	Source	8	426/481/547/605/661/724/832/948	450–800	0.31/1.24
WV2	Unseen target	8	426/481/547/605/661/724/832/948	450–800	0.46/1.85
QuickBird	Unseen target	4	485/560/660/830	450–900	0.61/2.44
GF2	Unseen target	4	485/555/660/830	450–900	0.80/3.20

To address this challenge, we require the fusion operator to be approximately equivariant to changes in the sensing process: when the sensing process changes, the fused output should change accordingly. This turns the unknown target mapping into a source-side constraint. Since such changes can be simulated from source data and applied consistently to the source observations and their HRMS targets, the required output change is supervised without target labels. Equivariance thus provides a source-only training principle for fixed-weight cross-sensor deployment.

Based on this principle, we propose EquiPan, a backbone-compatible framework that realizes sensor equivariance through two mechanisms operating at different stages. At the data level, a Spectral–Contrast Transformation (SCT) samples virtual observation models from source data and applies each sampled transformation jointly to the MS observations and their HRMS target, providing exact supervision without target labels. These virtual views vary in band configuration and spectral contrast. At the model level, a Reversible Spectral Transport (RST) maps both measured and virtual observations into a shared latent space that a fixed-channel backbone can process, and maps the predicted residual back to the corresponding native bands through an analytic left inverse on the observed spectral degrees of freedom. A parameter-free PAN mean alignment (PMA) matches the global PAN mean to the MS view of each branch while preserving the measured PAN contrast. No further PAN transformation is applied because the target PAN response is unavailable during source training.

We instantiate EquiPan with seven architecturally different fusion backbones and transfer models trained on WV3 to unseen WV2, QuickBird, and GaoFen-2 data. Evaluation covers reduced-resolution PSNR, ERGAS, and SAM against an MS reference, and full-resolution HQNR with its spectral and spatial distortion terms. In summary, our contributions are:

- We formulate single-source cross-sensor pansharpening as approximate equivariance under sensor transformations, where the required output change is defined by the transformation and can be supervised on source data.
- We propose EquiPan, in which SCT samples virtual observation models with exact paired targets, RST provides a reversible latent interface for their heterogeneous spectral configurations, and PMA removes branch-dependent global offsets without learned parameters.
- Experiments with seven backbones show that EquiPan, trained on one source sensor and deployed on unseen sensors without target training or weight updates, improves over traditional and interpolation-only baselines at reduced resolution and, in most cases, at full resolution.

2 Related Work

From spatial to sensor-transformation equivariance. Equivariance is widely used in image classification and detection to make features respond predictably to translations, rotations, and reflections (Cohen & Welling, 2016; Weiler & Cesa, 2019; Han et al., 2021). Sharing a transformation rule across such appearances increases weight sharing and reduces the need to relearn every transformed case. Beyond geometric symmetry, Reichert et al. (2024) explored sensor equivariance under changes in camera projection for semantic segmentation, while Wang et al. (2026b) coupled reconstruction with a specified degradation process in mosaiced hyperspectral–PAN fusion. We use approximate equivariance to require that fusion commutes with sensor transformations: applying a sampled transformation to the source observations before fusion should agree with applying the

same transformation to the fused output. Unlike methods that encode a predefined geometric symmetry or imaging degradation, EquiPan is trained on exactly paired source-derived transformations to promote this cross-space fusion consistency, without target-side adaptation.

Cross-sensor generalization. Existing methods place this adaptation at different stages. Wang et al. (2026a) proposed FoundPS, a dedicated foundation model for arbitrary-band fusion trained on multi-satellite data. Cui et al. (2025) introduced UniPAN, which leaves the fusion architecture unchanged but maps training and test observations toward a common marginal distribution. Xin et al. (2026) introduced ERFT, which learns an additional feature-tailoring module from target observations. By contrast, Xia et al. (2026) proposed SWIFT, which selects informative target samples and updates only sensor-sensitive source weights. These routes rely respectively on multi-sensor training, test-distribution alignment, or target-time learning. EquiPan instead studies single-source training with fixed-weight deployment and no learned target calibration.

3 METHOD

In this section, we present EquiPan for fixed-weight cross-sensor pansharpening under single-source training. We first formulate cross-sensor generalization as approximate equivariance under sensor transformations, which turns the unknown target mapping into a source-side constraint and enables supervision without target-side labels. We then realize this principle through two complementary mechanisms: the data-level Spectral–Contrast Transformation (SCT), and the model-level Reversible Spectral Transport (RST) together with parameter-free PAN Mean Alignment (PMA). Figure 1 summarizes how these components compose during training and fixed-weight deployment.

3.1 CROSS-SENSOR GENERALIZATION AS APPROXIMATE EQUIVARIANCE

Let $Y^\star$ denote a latent high-resolution spectral scene, and let $q \in \{s, t\}$ index the source or target sensor. The sensor-specific HRMS target, LRMS observation, and PAN observation are

$$Y_q = \mathcal{O}_q^{\mathrm{MS}}(Y^\star), \qquad M_q = \mathcal{S}_q(Y_q), \qquad P_q = \mathcal{O}_q^{\mathrm{PAN}}(Y^\star), \tag{1}$$

where $\mathcal{O}_q^{\mathrm{MS}}$ and $\mathcal{O}_q^{\mathrm{PAN}}$ denote the spectral observation operators, and $\mathcal{S}_q$ denotes the spatial degradation that produces LRMS. Let λ_q and C_q denote the MS band centers and band count, respectively, and let L_q be M_q upsampled to the PAN grid. Source training provides

$$\mathcal{D}_s = \{(M_s^{(i)}, P_s^{(i)}, Y_s^{(i)}, \lambda_s)\}_{i=1}^{N_s}, \tag{2}$$

whereas deployment provides only (M_t, P_t, λ_t), without Y_t or target-side weight updates. We seek a fixed predictor $\Phi_{\theta,\lambda_q}(M_q, L_q, P_q)$ that returns an HRMS estimate in the native C_q-band configuration. The objective is to learn a single parameter set θ from $\mathcal{D}_s$ such that

$$\Phi_{\theta,\lambda_t}(M_t, L_t, P_t) \approx Y_t. \tag{3}$$

for an unseen target sensor.

To characterize the MS-side variation relevant to this objective, consider the same scene observed under the source and target MS configurations. We assume that a spectral transformation $\mathcal{T}_{s\to t}$ approximately relates the corresponding quantities:

$$M_t \approx \mathcal{T}_{s\to t}(M_s), \qquad L_t \approx \mathcal{T}_{s\to t}(L_s), \qquad Y_t \approx \mathcal{T}_{s\to t}(Y_s). \tag{4}$$

This assumption defines the MS-side sensor variation we aim to handle; it does not assert that $\mathcal{T}_{s\to t}$ maps P_s to P_t. Because the target transformation is unavailable during source training, we approximate its relevant family using source-derived transformations. For a source sample, we suppress the sample index and write $(M, L, P, Y, \lambda) = (M_s^{(i)}, L_s^{(i)}, P_s^{(i)}, Y_s^{(i)}, \lambda_s)$. A sampled transformation $\mathcal{T}_g$ produces

$$M_g = \mathcal{T}_g(M), \qquad L_g = \mathcal{T}_g(L), \qquad Y_g = \mathcal{T}_g(Y), \tag{5}$$

together with transformed band metadata λ_g. Approximate equivariance requires

$$\Phi_{\theta,\lambda_g}(M_g, L_g, P) \approx \mathcal{T}_g(\Phi_{\theta,\lambda}(M, L, P)). \tag{6}$$

Paired supervision of the two paths encourages the predictor to respond consistently to sampled MS-side changes while holding the measured PAN image fixed. Source reconstruction and this approximate equivariance can support fixed-weight transfer when the sampled transformations adequately cover the target MS-side variation and the remaining PAN discrepancy is sufficiently limited.

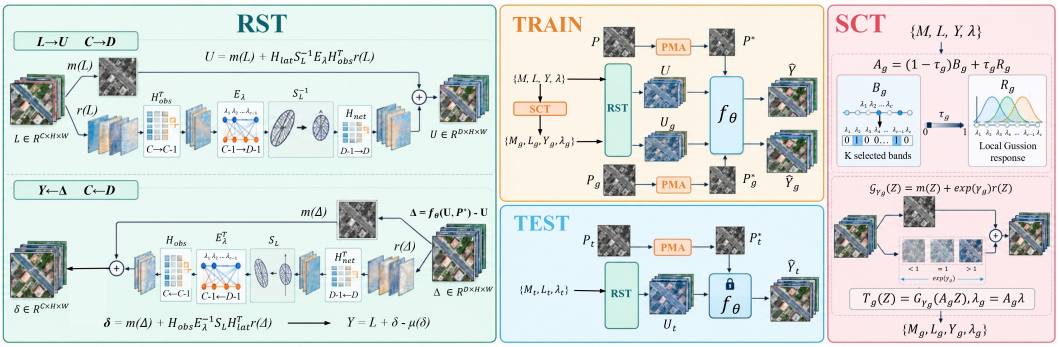

Figure 1: Overview of EquiPan. During training, **SCT** first constructs an exactly paired virtual MS view from the source tuple. The source and virtual observations then pass through **RST** into a shared latent interface and are processed by the same backbone, while **PMA** aligns the measured PAN image with the MS view of each branch. At deployment, **SCT** is removed, whereas **RST** and **PMA** are retained with fixed source-trained weights.

3.2 Spectral–Contrast Transformation for Sampling Virtual Observation Models

The first mechanism operates at the data level. Because the target observation model $\mathcal{T}_{s\to t}$ is unavailable during source training, we cannot directly construct training pairs from an unseen sensor. SCT instead samples virtual observation models from the source data and applies each sampled transformation consistently to the MS observations and their HRMS target. This produces exactly paired source-derived views that approximate changes in the sensing process and define the transformations under which the fusion model is trained to behave equivariantly. An observation model is determined by two complementary factors: how the source bands are combined into virtual bands, and how strongly the resulting bands differ from one another.

Sampling virtual band configurations. To expose the model to sensors with unseen band counts, SCT samples a virtual band count $K \in \{4, 8\}$ with equal probability. Exact selection uses all WV3 bands for $K = 8$; for $K = 4$, it uniformly selects one band from each adjacent pair. For anchor i_k, let g_k be its smaller adjacent wavelength gap. The Gaussian center is sampled within $\lambda_{i_k} \pm 0.35g_k$, and its FWHM is sampled uniformly from $[0.65g_k, 1.35g_k]$. The exact and Gaussian endpoints are combined by a scalar τ_g into a non-negative row-stochastic matrix $A_g \in \mathbb{R}^{K\times C_s}$:

$$A_g = (1-\tau_g)B_g + \tau_g R_g, \qquad A_g \geq 0, \qquad A_g\mathbf{1} = \mathbf{1}, \tag{7}$$

where B_g and R_g are the selection and Gaussian matrices. We set $\tau_g = 0$ or 1 with probability 0.25 each and otherwise sample $\tau_g \sim \mathcal{U}[0, 1]$; the virtual centers are $\lambda_g = A_g\lambda$.

Sampling virtual contrast gains. The matrix A_g changes how the source bands are combined but does not change the amplitude of the band-to-band differences. We sample $\gamma_g \sim \mathcal{U}[-\log 2, \log 2]$, giving $\exp(\gamma_g) \in [0.5, 2]$. For a Q-band tensor Z, define its per-pixel mean $\bar{Z} = Q^{-1}\mathbf{1}_Q^\top Z$ and chroma residual $r_Q(Z) = Z - \mathbf{1}_Q\bar{Z}$. SCT preserves the mean and scales only this residual:

$$\mathcal{G}_{\gamma_g}(Z) = \mathbf{1}_Q\bar{Z} + \exp(\gamma_g)r_Q(Z). \tag{8}$$

Because the gain is positive, the transformation is bijective at each pixel: the virtual observation retains the same pixel-level information as the source while rescaling its band-to-band differences. This one-to-one correspondence keeps the transformed target Y_g a valid observation of the same scene. The complete sampled transformation applies band mixing followed by contrast scaling:

$$\mathcal{T}_g(Z) = \mathcal{G}_{\gamma_g}(A_g Z), \qquad \lambda_g = A_g\lambda. \tag{9}$$

Only A_g determines the virtual band centers; the contrast gain does not alter λ_g.

Exactly paired transformed view. The same sampled transformation is applied to the upsampled MS image and the HRMS target:

$$L_g = \mathcal{T}_g(L), \qquad Y_g = \mathcal{T}_g(Y). \tag{10}$$

If the backbone consumes the low-resolution MS image as an auxiliary statistic, the same transformation is applied to M as well. Because $\mathcal{T}_g$ acts only along the spectral dimension, the transformed input and target remain exactly aligned in space. The paired target Y_g therefore turns the abstract equivariance requirement into supervised learning.

3.3 Reversible Spectral Transport as Pre- and Post-Processing

SCT produces observations with varying band counts, center wavelengths, and spectral contrasts, which a fixed-channel backbone cannot process directly. RST therefore maps native-band residuals to a shared latent representation before fusion and returns predicted residuals to the observed bands afterward. It aligns spectral geometry with a fixed Helmert basis and normalizes residual statistics with an invertible Cholesky transform. Reversibility here denotes a left inverse on the observed spectral subspace, not a bijection between the complete C- and D-channel spaces.

Spectral-geometry alignment by orthogonal transformation. This step addresses two sensor discrepancies that SCT samples but RST must resolve: different band counts and center wavelengths, and different luminance–chroma compositions of the MS observations, as evidenced in component-substitution pansharpening (Aiazzi et al., 2007; Vivone et al., 2015). Both are handled by a single spectral transformation.

The first issue is that the observed MS bands and the latent spectral grid live on different spectral axes. A natural bridge is a wavelength interpolation matrix $W_\lambda \in \mathbb{R}^{D\times C}$ that maps the C observed band centers λ to the fixed latent grid of dimension D. Applying W_λ alone, however, is not sufficient, because it acts on the full observed spectrum and does not preserve spectral geometry: it mixes luminance and chroma, and the resulting map is not orthogonal.

The second issue is therefore addressed first, at the input. Let $L \in \mathbb{R}^{C\times N}$, where columns index pixels, and let $\mathbf{1}_C \in \mathbb{R}^C$ be the all-ones vector. Each observed pixel is decomposed into a per-pixel common intensity and a chroma residual:

$$\bar{L} = \frac{1}{C}\mathbf{1}_C^\top L, \qquad r_C(L) = L - \mathbf{1}_C\bar{L}, \qquad L = \mathbf{1}_C\bar{L} + r_C(L), \tag{11}$$

so that $\mathbf{1}_C^\top r_C(L) = 0$. The common intensity $\bar{L}$ is passed through unchanged because it should not be rotated by spectral resampling. Only $r_C(L)$ is adapted to the latent spectral axes.

To make this adaptation well-posed, we need to isolate the chroma subspace on both the observed and the latent sides. This is the role of the two Helmert bases $H_{\text{obs}} \in \mathbb{R}^{C\times(C-1)}$ and $H_{\text{lat}} \in \mathbb{R}^{D\times(D-1)}$, which are fixed orthonormal bases of the respective zero-mean subspaces. They are not an additional design choice but the algebraic tool that lets W_λ act on chroma alone. With them, the chroma-to-chroma wavelength map becomes

$$H_{\text{lat}}^\top W_\lambda H_{\text{obs}} \in \mathbb{R}^{(D-1)\times(C-1)}, \tag{12}$$

which no longer mixes luminance into the transformation. This map is still non-orthogonal because W_λ is an interpolation operator. We therefore orthogonalize only the chroma-to-chroma map:

$$E_\lambda = \operatorname{orth}\left(H_{\text{lat}}^\top W_\lambda H_{\text{obs}}\right), \tag{13}$$

which yields $E_\lambda^\top E_\lambda = I_{C-1}$ whenever the oriented matrix has full column rank. Hence E_λ is an isometric embedding of the observed chroma subspace into the latent chroma subspace, and its left inverse is $E_\lambda^\top$. The common intensity $\bar{L}$ never enters this construction and is therefore preserved exactly. In this way, a single transformation resolves both band-count/center-wavelength mismatch and luminance–chroma mismatch, which is exactly the sensor variation that SCT simulates during training.

Residual-statistics alignment by Cholesky normalization. Even after geometry alignment, MS observations can retain different first- and second-order residual statistics. RST therefore applies invertible Cholesky whitening to map the geometry-aligned residuals to a common standardized distribution. Let

$$\tilde{r} = E_\lambda H_{\text{obs}}^\top r_C(L) \tag{14}$$

be the geometry-aligned chroma residual. From the residuals $\{\tilde{r}_i\}_{i=1}^N$ of the current MS image, RST estimates a detached center b_L and a stabilized Cholesky factor S_L as

$$b_L = \frac{1}{N}\sum_{i=1}^{N}\tilde{r}_i, \qquad \Sigma = \frac{1}{N}\sum_{i=1}^{N}(\tilde{r}_i - b_L)(\tilde{r}_i - b_L)^\top, \tag{15}$$

$$\Sigma \leftarrow \frac{\Sigma}{\mathrm{tr}(\Sigma)}, \qquad \Sigma \leftarrow \frac{1}{2}\Sigma + \frac{1}{2}I, \qquad \Sigma = S_L S_L^\top. \tag{16}$$

The normalized residual is then

$$\hat{r} = S_L^{-1}(\tilde{r} - b_L), \tag{17}$$

with inverse $\tilde{r} = S_L\hat{r} + b_L$. The latent input is

$$U = \mathbf{1}_D\bar{L} + H_{\mathrm{lat}}\hat{r}, \tag{18}$$

where $\mathbf{1}_D\bar{L} \in \mathbb{R}^{D\times N}$ broadcasts the unchanged common intensity to the latent channels. The second term applies geometry alignment and statistics normalization to the chroma residual. SCT is applied first to both MS observations and HRMS targets; RST then applies geometry alignment followed by statistics normalization.

Residual prediction and inverse transformation. The latent tensor U is only a network interface. The backbone predicts a residual relative to U, and RST applies the inverse of the input transformation to return it to the native bands:

$$\begin{aligned}
\Delta &= f_\theta(U, P^*) - U,\\
\bar{\Delta} &= \frac{1}{D}\mathbf{1}_D^\top\Delta, \qquad r_D(\Delta) = \Delta - \mathbf{1}_D\bar{\Delta},\\
\delta &= \mathbf{1}_C\bar{\Delta} + H_{\mathrm{obs}}E_\lambda^\top S_L H_{\mathrm{lat}}^\top r_D(\Delta),\\
\widehat{Y} &= L + \delta - \mu(\delta).
\end{aligned} \tag{19}$$

The term $\mathbf{1}_C\bar{\Delta}$ broadcasts the latent common component to the native bands. The residual component is recolored by S_L, mapped back through $E_\lambda^\top$, and synthesized by H_{obs}. Here $\mu(\delta)$ is the samplewise global mean broadcast to all entries, so its subtraction removes only the global common-mode shift. Hence zero residual returns L exactly, and the analytic inverse is a left inverse on the observed spectral degrees of freedom; RST is not a bijection between the complete C- and D-channel spaces.

PAN and MS are acquired through different spectral responses, so their global means can differ by an additive offset. For the source and virtual branches, PMA forms $P^* = P - \mu(P) + \mu(L)$ and $P_g^* = P - \mu(P) + \mu(L_g)$, respectively. Both preserve the measured spatial contrast $P - \mu(P)$ and change only the global offset. No further PAN transformation is applied because the target PAN response is unavailable during source training.

3.4 Joint Supervision from Source and Virtual Observations

After SCT constructs the paired virtual observation and RST maps each native MS view through the shared latent interface, the source and virtual branches are processed by the same predictor with shared backbone parameters θ. As shown in Figure 1, their predictions are

$$\widehat{Y} = \Phi_{\theta,\lambda}(M, L, P), \qquad \widehat{Y}_g = \Phi_{\theta,\lambda_g}(M_g, L_g, P). \tag{20}$$

Training both branches with the same θ indirectly promotes the approximate equivariance in equation 6: the source term anchors reconstruction on the measured source sensor, whereas the virtual term teaches the shared weights how the correct output changes with band configuration and spectral contrast. Training minimizes equally weighted reconstruction losses against the two exact targets:

$$\mathcal{L}_\theta = \frac{1}{2}\left\|\widehat{Y} - Y\right\|_1 + \frac{1}{2}\left\|\widehat{Y}_g - Y_g\right\|_1. \tag{21}$$

During deployment, SCT is removed, while RST and PMA are retained because they are deterministic functions of the target observation. RST and PMA estimate their detached states from the complete target scene and reuse them for all inference tiles; the learned parameters θ remain fixed.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We train all learned variants exclusively on WorldView-3 (WV3). Evaluation on WV3 measures source-sensor performance, whereas WorldView-2 (WV2), QuickBird (QB), and Gaofen-2 (GF2) test zero-shot transfer to unseen sensors. Reduced-resolution (RR) evaluation follows the protocol of Wald et al. (1997) and reports PSNR (dB), ERGAS, and SAM (degrees) against an MS reference. At full resolution (FR), where no reference HRMS image is available, we report HQNR and its spectral and spatial distortion terms, D_λ and D_s (Khan et al., 2009; Vivone et al., 2019).

Benchmark comparison. Tables 2 and 3 separate the methods into four groups according to their training and evaluation roles. Group I comprises eight established non-learning methods from the component-substitution and multiresolution-analysis families (Vivone et al., 2015; 2021). Group II contains three supervised deep baselines (Yang et al., 2017; Yuan et al., 2018; He et al., 2019). Their QB and GF2 results use models trained on the corresponding sensors, while their WV2 results use WV3-trained models in a zero-shot setting.

Groups III and IV form the controlled comparison. Group III trains each backbone on WV3 with source-only supervision and handles different band configurations by wavelength interpolation. Group IV equips the same backbone with EquiPan and uses paired source–virtual supervision generated entirely from WV3. Both groups use the same source sensor and exclude target-domain training data, learned target calibration, and deployment-time weight updates. The paired rows therefore measure the effect of the EquiPan training and interface design for each backbone.

4.2 RESULTS AND ANALYSIS

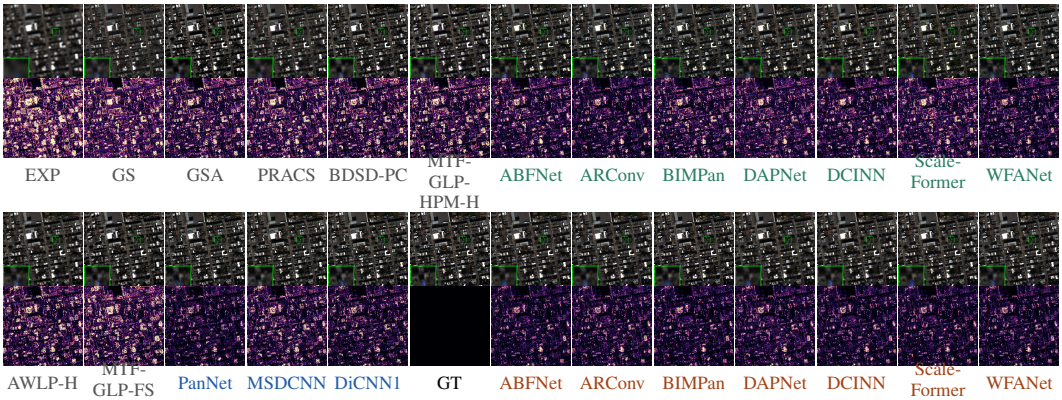

Figure 2: Qualitative RR comparison on QB. Each entry pairs an RGB reconstruction with its four-band mean absolute error. Labels denote traditional (gray), target-trained deep (blue), interpolation-only (green), and EquiPan (orange) methods; paired green and orange columns share backbones.

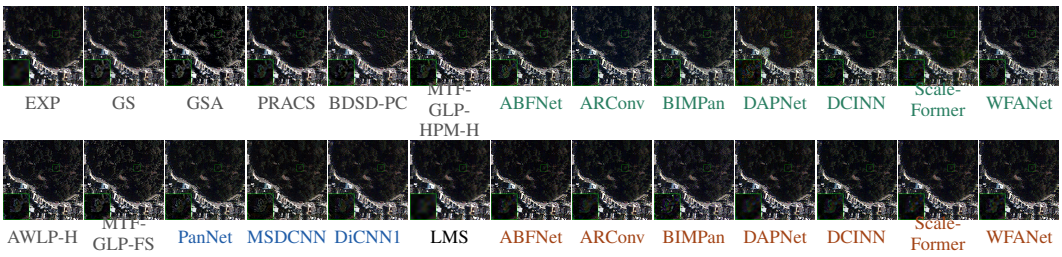

Figure 3: Qualitative comparison on QB full-resolution data.

We apply both configurations to seven diverse backbones: ABFNet (Wang et al., 2026c), ARConv (Wang et al., 2025), BIMPan (Hou et al., 2023), DAPNet (Wang & Ma, 2025), DCINN (Wang

Table 2: Reduced-resolution comparison. Bold and underlined values denote the best and second-best results in each column across all methods, respectively.

	WV3			QB			GF2			WV2 zero-shot		
Method	PSNR↑	ERGAS↓	SAM↓	PSNR↑	ERGAS↓	SAM↓	PSNR↑	ERGAS↓	SAM↓	PSNR↑	ERGAS↓	SAM↓
I. Traditional methods												
EXP (Vivone et al., 2021)	27.07	7.09	5.91	28.39	11.57	8.16	38.15	2.89	2.48	25.14	6.80	6.59
GS (Laben & Brower, 2000)	29.57	5.31	5.74	30.28	9.21	8.19	38.22	2.97	2.98	27.18	5.26	6.67
GSA (Aiazzi et al., 2007)	30.22	4.95	6.29	31.84	7.62	7.75	40.17	2.30	2.48	28.31	4.53	6.49
PRACS (Choi et al., 2011)	30.15	4.92	5.70	31.05	8.41	7.99	40.43	2.20	2.33	26.98	5.37	6.38
BDSD-PC (Vivone, 2019)	31.03	4.51	5.66	32.09	7.44	7.70	40.99	2.09	2.34	28.82	4.27	6.22
MTF-GLP-FS (Vivone et al., 2018)	30.57	4.76	5.70	31.50	7.93	7.83	40.01	2.31	2.45	27.88	4.80	6.41
MTF-GLP-HPM-H (Lolli et al., 2017)	29.79	5.21	5.34	31.39	7.96	6.97	40.89	2.08	2.25	27.94	4.81	5.99
AWLP-H (Vivone et al., 2019)	31.19	4.45	5.23	32.18	7.32	7.03	41.56	1.93	2.06	28.53	4.50	5.97
II. Deep baselines												
PanNet (Yang et al., 2017)	34.43	3.01	4.15	35.07	5.27	5.36	**45.60**	**1.22**	**1.40**	29.11	4.22	5.49
MSDCNN (Yuan et al., 2018)	29.67	5.25	6.48	33.57	6.26	6.54	42.58	1.73	2.00	27.23	5.25	7.43
DiCNN1 (He et al., 2019)	35.33	2.73	3.95	**35.36**	5.09	**5.20**	45.37	1.27	1.43	26.78	5.63	6.28
III. Interpolation only												
ABFNet (Wang et al., 2026c)	**37.54**	**2.16**	**3.19**	33.84	6.17	6.76	34.62	3.89	4.11	28.86	4.40	5.27
ARConv (Wang et al., 2025)	36.89	2.30	3.31	34.61	5.49	6.17	38.00	2.78	2.81	28.79	4.42	5.45
BIMPan (Hou et al., 2023)	36.91	2.31	3.35	33.38	6.44	6.88	32.91	4.79	5.07	29.59	4.03	5.43
DAPNet (Wang & Ma, 2025)	37.13	2.26	3.30	31.72	16.19	7.52	35.33	3.53	3.89	29.00	4.33	5.69
DCINN (Wang et al., 2024)	37.19	2.25	3.31	33.94	6.01	6.70	35.68	3.71	3.52	27.93	4.94	5.82
ScaleFormer (Cao et al., 2026)	37.36	2.20	3.24	32.90	6.80	7.74	34.40	4.01	3.71	28.23	4.77	5.89
WFANet (Huang et al., 2025)	37.03	2.28	3.35	34.00	5.94	6.70	34.51	3.93	4.67	28.69	4.55	5.79
IV. EquiPan												
ABFNet (Wang et al., 2026c)	37.32	2.21	3.26	35.10	**5.02**	5.99	44.00	1.50	1.61	31.85	3.10	4.60
ARConv (Wang et al., 2025)	36.45	2.42	3.54	34.55	5.54	6.45	42.38	1.78	1.98	30.73	3.48	4.92
BIMPan (Hou et al., 2023)	36.38	2.45	3.55	34.39	5.69	6.00	40.26	2.21	2.28	30.61	3.52	4.91
DAPNet (Wang & Ma, 2025)	36.70	2.38	3.42	33.98	5.88	6.45	42.35	1.74	1.79	31.25	3.29	4.81
DCINN (Wang et al., 2024)	36.75	2.36	3.45	34.87	5.32	6.20	43.55	1.62	1.70	31.34	3.22	4.76
ScaleFormer (Cao et al., 2026)	37.19	2.24	3.29	34.78	5.89	5.91	42.61	1.69	1.72	**32.00**	**2.95**	**4.49**
WFANet (Huang et al., 2025)	36.73	2.35	3.45	34.83	5.36	6.45	43.05	1.60	1.75	31.23	3.24	4.87

et al., 2024), ScaleFormer (Cao et al., 2026), and WFANet (Huang et al., 2025). Their varied fusion designs test the architectural portability of EquiPan.

Reduced-resolution results. Table 2 shows that EquiPan improves RR PSNR in 20 of 21 backbone–target comparisons: all seven on GF2 by 4.38–9.38 dB and WV2 by 1.02–3.77 dB, and six on QB by 0.83–2.26 dB. WV3 decreases by 0.17–0.53 dB. The strongest EquiPan model ranks first overall on WV2 but remains below the sensor-trained deep baselines on QB and GF2.

Full-resolution results. Table 3 shows that EquiPan improves HQNR in 19 of 21 backbone–target comparisons: five on QB and all seven on GF2 and WV2, by 0.0167–0.0989, 0.0521–0.2752, and 0.0116–0.1145, respectively. It attains the highest HQNR on QB (0.9201) and GF2 (0.9250), and the second-highest on WV2 (0.9389); WV3 changes are smaller and mixed (-0.0010 to 0.0139).

Qualitative comparison. Figure 2 shows stronger building and road residuals for traditional methods, whereas deep baselines and EquiPan recover sharper boundaries; paired columns show lower broad spectral errors for most EquiPan variants. In FR, several variants sharpen roads and vegetation without conspicuous color shifts (Figure 3). Without an HRMS reference, FR images support only visual PAN–MS consistency comparisons.

Component analysis. Table 4 evaluates RST, SCT, and PMA with ABFNet under a shared protocol. No EquiPan disables all three, while each remaining variant removes one component. Relative to No EquiPan, the complete framework improves RR PSNR by 1.26/9.38/2.99 dB and HQNR by 0.0533/0.2626/0.0569 on QB/GF2/WV2, respectively. On WV3, PSNR decreases by 0.22 dB while HQNR rises by 0.0030. Removing RST most reduces GF2/WV2 RR PSNR, SCT removal lowers RR PSNR on every unseen sensor, and PMA removal most reduces QB HQNR. Full EquiPan gives the best HQNR throughout and GF2 PSNR; w/o RST and w/o PMA slightly lead QB and WV2 PSNR, respectively.

Table 3: Full-resolution comparison. Bold and underlined values denote the best and second-best unrounded results in each column across all methods, respectively.

	WV3			QB			GF2			WV2 zero-shot		
Method	HQNR ↑	D_λ ↓	D_s ↓	HQNR ↑	D_λ ↓	D_s ↓	HQNR ↑	D_λ ↓	D_s ↓	HQNR ↑	D_λ ↓	D_s ↓
I. Traditional methods												
EXP (Vivone et al., 2021)	0.8998	0.025	0.077	0.7842	0.052	0.172	0.8852	0.021	0.096	0.8984	0.045	0.060
GS (Laben & Brower, 2000)	0.8469	0.097	0.064	0.6718	0.213	0.150	0.6155	0.221	0.212	0.7740	0.148	0.092
GSA (Aiazzi et al., 2007)	0.8582	0.075	0.075	0.6402	0.212	0.191	0.6965	0.126	0.203	0.8224	0.097	0.090
PRACS (Choi et al., 2011)	0.9168	0.045	0.041	0.7742	0.119	0.123	0.8109	0.073	0.126	0.9144	0.056	0.031
BDSD-PC (Vivone, 2019)	0.8777	0.065	0.062	0.7135	0.200	0.110	0.7673	0.087	0.160	0.8064	0.158	0.043
MTF-GLP-FS (Vivone et al., 2018)	0.9059	0.048	0.049	0.7933	0.085	0.135	0.7859	0.080	0.146	0.8765	0.058	0.070
MTF-GLP-HPM-H (Lolli et al., 2017)	0.8937	0.059	0.051	0.8178	0.094	0.100	0.8500	0.051	0.105	0.8728	0.061	0.071
AWLP-H (Vivone et al., 2019)	0.9270	0.019	0.055	0.8244	0.050	0.134	0.8678	0.024	0.111	0.8939	0.034	0.075
II. Deep baselines												
PanNet (Yang et al., 2017)	0.9481	**0.017**	0.036	0.8925	0.058	0.054	0.9228	**0.019**	0.060	**0.9461**	0.031	**0.024**
MSDCNN (Yuan et al., 2018)	0.8962	0.050	0.057	0.8341	0.083	0.093	0.8352	0.053	0.119	0.8560	0.101	0.049
DiCNN1 (He et al., 2019)	0.9156	0.048	0.038	0.8368	0.099	0.073	0.7718	0.100	0.143	0.7070	0.203	0.113
III. Interpolation only												
ABFNet (Wang et al., 2026c)	0.9498	0.024	0.027	0.8588	0.077	0.073	0.6586	0.259	0.111	0.8772	0.077	0.050
ARConv (Wang et al., 2025)	0.9528	0.023	0.025	0.8971	0.058	0.051	0.8671	0.069	0.069	0.9215	0.051	0.029
BIMPan (Hou et al., 2023)	0.9407	0.028	0.032	0.8503	0.090	0.068	0.7985	0.149	0.062	0.9036	0.063	0.036
DAPNet (Wang & Ma, 2025)	0.9470	0.026	0.028	0.7694	0.129	0.117	0.8367	0.108	0.062	0.8912	0.051	0.060
DCINN (Wang et al., 2024)	0.9520	0.022	0.026	0.8813	0.070	0.053	0.8022	0.095	0.114	0.9020	0.047	0.053
ScaleFormer (Cao et al., 2026)	0.9424	0.032	0.027	0.8214	0.142	0.046	0.6129	0.298	0.128	0.8204	0.107	0.081
WFANet (Huang et al., 2025)	0.9505	0.027	0.024	0.8500	0.072	0.086	0.8533	0.099	0.054	0.9180	0.048	0.036
IV. EquiPan												
ABFNet (Wang et al., 2026c)	0.9528	0.020	0.028	0.9121	0.045	0.046	0.9212	0.035	0.045	0.9341	**0.026**	0.041
ARConv (Wang et al., 2025)	0.9518	0.022	0.027	**0.9201**	0.045	**0.038**	0.9192	0.042	**0.040**	0.9368	0.032	0.032
BIMPan (Hou et al., 2023)	0.9515	0.025	0.024	0.8990	0.063	0.042	0.8805	0.037	0.086	0.9371	0.040	0.025
DAPNet (Wang & Ma, 2025)	0.9463	0.022	0.032	0.8683	0.044	0.092	0.9021	0.035	0.065	0.9268	0.027	0.047
DCINN (Wang et al., 2024)	**0.9611**	0.020	**0.019**	0.8649	0.042	0.098	0.8991	0.033	0.071	0.9389	0.027	0.035
ScaleFormer (Cao et al., 2026)	0.9563	0.025	0.019	0.8381	0.066	0.102	0.8881	0.043	0.072	0.9349	0.038	0.028
WFANet (Huang et al., 2025)	0.9555	0.021	0.024	0.8212	**0.041**	0.143	**0.9250**	0.031	0.045	0.9296	0.029	0.042

Table 4: Component ablation with ABFNet trained on WV3. bold marks the column best.

				WV3		QB		GF2		WV2 zero-shot	
Variant	RST	SCT	PMA	PSNR ↑	HQNR ↑	PSNR ↑	HQNR ↑	PSNR ↑	HQNR ↑	PSNR ↑	HQNR ↑
No EquiPan	No	No	No	**37.54**	0.9498	33.84	0.8588	34.62	0.6586	28.86	0.8772
w/o RST	No	Yes	Yes	37.42	0.9513	**35.16**	0.8989	36.08	0.8688	29.46	0.9299
w/o SCT	Yes	No	Yes	37.45	0.9508	33.30	0.8811	40.02	0.9073	29.69	0.9228
w/o PMA	Yes	Yes	No	37.34	0.9516	34.27	0.8332	40.30	0.9169	**31.99**	0.9239
Full EquiPan	Yes	Yes	Yes	37.32	**0.9528**	35.10	**0.9121**	**44.00**	**0.9212**	31.85	**0.9341**

5 CONCLUSION

We presented EquiPan for fixed-weight cross-sensor pansharpening under single-source training. By treating changes in the sensing process as approximate equivariance, EquiPan separates cross-sensor adaptation from the fusion backbone itself. SCT exposes the model to exactly paired source-derived sensor variations, RST provides a reversible latent interface for the resulting heterogeneous spectral observations, and PMA reduces global PAN–MS offsets without altering the measured PAN contrast. Experiments across seven backbones and three unseen sensors showed that this design generally improved zero-shot transfer over interpolation-only counterparts and achieved competitive performance against representative baselines.

AI USE STATEMENT

We used AI to assist with manuscript drafting, refinement of the conceptual explanation and mathematical descriptions of the framework, feedback on ablation design and experimental protocols. We also used it to check related literature and manuscript structure, improve readability, format references and tables, and edit result-summary code. The authors retain responsibility for reviewing and verifying all AI-assisted content and for the final text, claims, references, code, and numerical reports.

REFERENCES

Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of MS+Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. doi: 10.1109/TGRS.2007.901007.

Ke Cao, Xuanhua He, Xueheng Li, Lingting Zhu, Yingying Wang, Ao Ma, Zhanjie Zhang, Man Zhou, Chengjun Xie, and Jie Zhang. Cross-scale pansharpening via ScaleFormer and the PanScale benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13211–13221, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cao_Cross-Scale_Pansharpening_via_ScaleFormer_and_the_PanScale_Benchmark_CVPR_2026_paper.html.

Liangfu Chen, Husi Letu, Meng Fan, Huazhe Shang, Jinhua Tao, Laixiong Wu, Ying Zhang, Chao Yu, Jianbin Gu, Ning Zhang, Jin Hong, Zhongting Wang, and Tianyu Zhang. An introduction to the chinese high-resolution earth observation system: Gaofen-1–7 civilian satellites. *Journal of Remote Sensing*, 2022:9769536, 2022. doi: 10.34133/2022/9769536.

Jaewan Choi, Kiyun Yu, and Yongil Kim. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):295–309, 2011. doi: 10.1109/TGRS.2010.2051674.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Yongchuan Cui, Peng Liu, and Hui Zhang. Enpowering your pansharpening models with generalizability: Unified distribution is all you need. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11850–11860, 2025.

DigitalGlobe. QuickBird imagery products: Product guide. Technical Report Revision 4.0, DigitalGlobe, 2003. URL https://engineering.purdue.edu/~bethel/qb_pg.pdf.

DigitalGlobe. WorldView-3 data sheet. Technical Report DS-WV3 05/14, DigitalGlobe, 2014. URL https://content.satimagingcorp.com.s3.amazonaws.com/static/satellite-sensor-specification/WorldView-3-PDF-Download.pdf.

DigitalGlobe. WorldView-2 data sheet. Technical Report DS-WV2 02/16, DigitalGlobe, 2016. URL https://www.nv5geospatialsoftware.com/Portals/0/pdfs/WorldView2-DS-WV2-rev2.pdf.

Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. ReDet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2786–2795, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Han_ReDet_A_Rotation-Equivariant_Detector_for_Aerial_Object_Detection_CVPR_2021_paper.html.

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.

Junming Hou, Qi Cao, Ran Ran, Che Liu, Junling Li, and Liang-Jian Deng. Bidomain modeling paradigm for pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 347–357, 2023. doi: 10.1145/3581783.3612188.

Jie Huang, Rui Huang, Jinghao Xu, Siran Peng, Yule Duan, and Liang-Jian Deng. Wavelet-assisted multi-frequency attention network for pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3662–3670, 2025. doi: 10.1609/aaai.v39i4.32381.

Muhammad Murtaza Khan, Luciano Alparone, and Jocelyn Chanussot. Pansharpening quality assessment using the modulation transfer functions of instruments. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3880–3891, 2009. doi: 10.1109/TGRS.2009.2029094.

Craig A. Laben and Bernard V. Brower. Process for enhancing the spatial resolution of multispectral imagery using Pan-Sharpening. U.S. Patent 6,011,875, January 2000. URL https://patents.google.com/patent/US6011875A/en. Issued January 4, 2000.

Simone Lolli, Luciano Alparone, Andrea Garzelli, and Gemine Vivone. Haze correction for contrast-based multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2255–2259, 2017. doi: 10.1109/LGRS.2017.2761021.

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

Hannes Reichert, Manuel Hetzel, Andreas Hubert, Konrad Doll, and Bernhard Sick. Sensor equivariance: A framework for semantic segmentation with diverse camera models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1254–1261, 2024. doi: 10.1109/CVPRW63382.2024.00132. URL https://openaccess.thecvf.com/content/CVPR2024W/OmniCV2024/html/Reichert_Sensor_Equivariance_A_Framework_for_Semantic_Segmentation_with_Diverse_Camera_CVPRW_2024_paper.html.

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.

Gemine Vivone, Luciano Alparone, Andrea Garzelli, and Simone Lolli. Fast reproducible pansharpening based on instrument and acquisition modeling: AWLP revisited. *Remote Sensing*, 11(19):2315, 2019. doi: 10.3390/rs11192315.

Gemine Vivone, Mauro Dalla Mura, Andrea Garzelli, Rocco Restaino, Giuseppe Scarpa, Magnus Orn Ulfarsson, Luciano Alparone, and Jocelyn Chanussot. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):53–81, 2021. doi: 10.1109/MGRS.2020.3019315.

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

Hebaixu Wang and Jiayi Ma. Deep adaptive unfolded network via spatial morphology stripping and spectral filtration for pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10730–10740, 2025.

Hebaixu Wang, Jing Zhang, Haonan Guo, Di Wang, Jiayi Ma, Bo Du, and Liangpei Zhang. Universal pansharpening foundation model. *arXiv preprint arXiv:2603.03831*, 2026a. URL https://arxiv.org/abs/2603.03831.

Nan Wang, Anjing Guo, Renwei Dian, and Shutao Li. Equivariant high-resolution hyperspectral imaging via mosaiced and PAN image fusion. *IEEE Transactions on Image Processing*, 35:1246–1260, 2026b. doi: 10.1109/TIP.2026.3657219.

Wu Wang, Liang-Jian Deng, Ran Ran, and Gemine Vivone. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4):1029–1054, 2024. doi: 10.1007/s11263-023-01924-5.

Xueyang Wang, Zhixin Zheng, Jiandong Shao, Yule Duan, and Liang-Jian Deng. Adaptive rectangular convolution for remote sensing pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17872–17881, 2025.

Zhenggang Wang, Wu Wang, Huazhe Liang, and Tai-Xiang Jiang. Butterworth as attention: Anisotropic spectral gating for pansharpening. In *International Conference on Machine Learning*, 2026c. URL https://it.swufe.edu.cn/info/1143/21728.htm. Accepted paper; proceedings details not yet verified.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html.

Zeyu Xia, Chenxi Sun, Tianyu Xin, Yubo Zeng, Haoyu Chen, and Liang-Jian Deng. SWIFT: A general sensitive weight identification framework for fast sensor-transfer pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10960–10968, 2026. doi: 10.1609/aaai.v40i13.38073.

Tianyu Xin, Jin-Liang Xiao, Zeyu Xia, Shan Yin, and Liang-Jian Deng. Training and inference within 1 second—tackle cross-sensor degradation of real-world pansharpening with efficient residual feature tailoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11141–11149, 2026. doi: 10.1609/aaai.v40i13.38093.

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

Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018. doi: 10.1109/JSTARS.2018.2794888.

A LEFT-INVERSE PROPERTY OF RST

For fixed E_λ, S_L, and b_L, the RST input and return operations can be written as

$$u = F_{\text{in}} l + b, \tag{22}$$

$$F_{\text{in}} = \frac{\mathbf{1}_D \mathbf{1}_C^\top}{C} + H_{\text{lat}} S_L^{-1} E_\lambda H_{\text{obs}}^\top, \tag{23}$$

$$b = -H_{\text{lat}} S_L^{-1} b_L, \tag{24}$$

$$F_{\text{out}} = \frac{\mathbf{1}_C \mathbf{1}_D^\top}{D} + H_{\text{obs}} E_\lambda^\top S_L H_{\text{lat}}^\top. \tag{25}$$

The Helmert bases satisfy $H^\top H = I$, $H^\top \mathbf{1} = 0$, and $HH^\top = I - \mathbf{1}\mathbf{1}^\top/C$ in the corresponding channel space. Together with $E_\lambda^\top E_\lambda = I_{C-1}$, these identities eliminate the common-mode/chromatic cross terms and give

$$F_{\text{out}} F_{\text{in}} = \frac{\mathbf{1}_C \mathbf{1}_C^\top}{C} + H_{\text{obs}} E_\lambda^\top S_L H_{\text{lat}}^\top H_{\text{lat}} S_L^{-1} E_\lambda H_{\text{obs}}^\top \tag{26}$$

$$= \frac{\mathbf{1}_C \mathbf{1}_C^\top}{C} + H_{\text{obs}} H_{\text{obs}}^\top = I_C. \tag{27}$$

Hence $F_{\text{out}}(u - b) = l$. RST is therefore injective on the observed spectral subspace. When $C < D$, its image has dimension C and does not cover the complete latent space, which is why the construction is a left inverse rather than a global bijection.

B IMPLEMENTATION DETAILS

For virtual center ν_k and width σ_k, the smooth spectral-response endpoint is

$$(R_g)_{k,c} = \frac{d_c \exp[-(\lambda_c - \nu_k)^2/(2\sigma_k^2)]}{\sum_{j=1}^{C_s} d_j \exp[-(\lambda_j - \nu_k)^2/(2\sigma_k^2)]}. \tag{28}$$

$\sigma_k = \text{FWHM}_k/(2\sqrt{2\log 2})$, and d_c denotes the midpoint interval width, with each outer boundary extended by half the nearest wavelength gap. Table 5 collects settings not specified in the main text.

Table 5: Training and EquiPan-specific implementation settings.

Setting	Value
Optimizer	Adam with the original backbone parameter grouping
Training and selection	360 epochs; batch size 32; validation every 5 epochs; checkpoint selected by WV3 validation PSNR
Initial learning rate	ABFNet, BIMPan, DAPNet: 10^{-3}; ARConv: $6{\times}10^{-4}$; DCINN: $8{\times}10^{-4}$; ScaleFormer: 1.5×10^{-3}; WFANet: 9×10^{-4}
Learning-rate schedule	MultiStepLR at $[100, 200, 300, 340]$, decay factor 0.5
Latent spectral grid	$D = 8$, using the WV3 band centers
RST covariance	Trace normalization, 0.5 shrinkage toward identity, Cholesky factorization

Bands are sorted by center wavelength before RST and restored afterward. Deterministic QR signs fix the orientation of E_λ, while rank-deficient embeddings are rejected. RST statistics and factorization use float64; the resulting state and network inputs use float32. Negligible residual energy or failed factorization sets $S_L = I_{D-1}$ while retaining residual centering.

Scope. Virtual responses cannot recover spectra outside the observed WV3 bands, and band centers do not encode complete response functions. SCT does not model spatial blur, registration error, or sensor noise. Deployment requires $C \leq 8$, a full-rank embedding, and retraining each backbone with EquiPan; RST's left inverse alone does not imply cross-sensor accuracy.