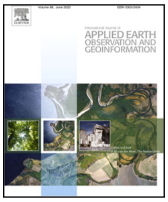
Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Reverse degradation for remote sensing pan-sharpening

Jiang He, Xiao Xiang Zhu *

Chair of Data Science in Earth Observation, Technical University of Munich, 80333, Munich, Germany
Munich Center for Machine Learning, 80333, Munich, Germany

ARTICLE INFO

Keywords:
Pan-sharpening
Degradation model
Unsupervised fusion
Model interpretability
Self-learning
Contrastive learning

ABSTRACT

Accurate pan-sharpening of multispectral images is essential for high-resolution remote sensing, yet supervised methods are limited by the need for paired training data and poor generalization. Existing unsupervised approaches often neglect the physical consistency between degradation and fusion and lack sufficient constraints, resulting in suboptimal performance in complex scenarios. We propose RevFus, a novel two-stage pan-sharpening framework. In the first stage, an invertible neural network models the degradation process and reverses it for fusion with cycle-consistency self-learning, ensuring a physically grounded mapping. In the second stage, structural detail compensation and spatial–spectral contrastive learning alleviate detail loss and enhance spectral–spatial fidelity. To further understand the network's decision-making, we design a quantitative and systematic measure of model interpretability, the Interpretability Efficacy Coefficient (IEC). IEC integrates multiple statistics derived from SHapley Additive exPlanations (SHAP) values into a single unified score and try to evaluate how effectively a model balances spatial detail enhancement with spectral preservation. Experiments on three datasets demonstrate that RevFus outperforms state-of-the-art unsupervised and traditional methods, delivering superior spectral fidelity, enhanced spatial detail, and high model interpretability, thereby validating the effectiveness of the interpretable deep learning framework for robust, high-quality pan-sharpening.

1. Introduction

Pan-sharpening is an efficient remote sensing image processing technique that fuses spectral information from different data sources to obtain high spatial resolution multispectral images. It mainly combines the high spatial resolution (HR) details of panchromatic (PAN) images with the rich spectral information of low spatial resolution (LR) multispectral (MS) images. This technique provides higher-quality data for numerous remote sensing applications (Liu et al., 2025), including land cover classification (Yang et al., 2025), urban monitoring (Chen et al., 2025), agricultural assessment (Xia et al., 2023), and environmental management (Eugenio et al., 2014).

According to information and feature processing, traditional pan-sharpening methods can be categorized to three groups: Component substitution-based methods (Carper et al., 1990; Gillespie et al., 1987; Kwarteng and Chavez, 1989; Laben and Brower, 2000), Multi-resolution analysis-based methods (Burt and Adelson, 1987; Nason and Silverman, 1995; Starck et al., 2002; Do and Vetterli, 2005) and Hybrid methods (González-Audícana et al., 2004; Otazu et al., 2005; Javan et al., 2021). CS-based methods are considered to provide rich spatial details, although spectral fidelity may be compromised. Meanwhile, MRA-based methods are considered to better preserve spectral information, but their spatial details are generally less sharp.

While these methods offer high computational efficiency, they often struggle to balance spatial detail preservation with spectral fidelity in complex scenarios, frequently resulting in significant spectral distortions. To address these limitations, more advanced techniques have emerged, leveraging sparse representation and optimization-based frameworks. Representative approaches include P+ XS (Ballester et al., 2006), Total Variation (TV) regularization (Palsson et al., 2013), the $\ell_{1/2}$ gradient prior (Zeng et al., 2016), filter estimation techniques (Vivone et al., 2014), and local gradient constraint-based fusion (Fu et al., 2019). Although these methods offer enhanced modeling flexibility and stronger representational capabilities, they typically depend on strict assumptions and pre-specified degradation models, which can restrict their general applicability and necessitate careful adjustment of hyper-parameters. As a result, their applicability to multi-sensor environments and large-scale remote sensing tasks remains constrained.

* Corresponding author at: Chair of Data Science in Earth Observation, Technical University of Munich, 80333, Munich, Germany
E-mail address: xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.jag.2026.105085
Received 10 November 2025; Received in revised form 20 December 2025; Accepted 1 January 2026

In recent years, with the rapid advancement of deep learning, directly learning the mapping between low- and high-spatial resolution information has emerged as a prominent approach for pan-sharpening. In particular, supervised convolutional neural network (CNN)-based methods have demonstrated remarkable performance in this task (Zhong et al., 2016; Rao et al., 2017; Wu et al., 2023, 2025). For instance, Masi et al. (2016) formulates pan-sharpening as an image super-resolution problem and addresses it using a three-layer CNN (PNN). To construct deeper networks and improve learning capacity, residual learning was introduced by Shao and Cai (2018), while Wei et al. (2017) proposed global residual skip connections to better maintain spatial details. Building on these strategies, Yang et al. (2017) incorporated high-pass filtering to facilitate the extraction of high-frequency components. To further boost CNN modeling capabilities, a range of techniques has been explored, including adaptive weighting schemes (Liu et al., 2020a), pyramid network architectures (Zhang et al., 2019), gradient priors (Zhang and Ma, 2021), two-stream networks (Liu et al., 2020c), deep unrolling approaches (He et al., 2022), generative adversarial networks (GANs) (Liu et al., 2020b; Gastineau et al., 2021), and diffusion-based models (Zhong et al., 2024). Through end-to-end training, these networks can directly learn fusion mappings from paired LR and HR images, achieving substantially higher performance than traditional methods (Deng et al., 2022). Nevertheless, such approaches are highly dependent on high-quality real-world training data, which are often unavailable in practical scenarios.

To alleviate the dependency on ground truth data, a range of unsupervised pan-sharpening methods have been proposed (Li et al., 2021; Zhou et al., 2020). These methods typically simulate the degradation process or incorporate adversarial mechanisms to learn the fusion mapping, thereby offering greater practicality. Nevertheless, existing unsupervised approaches generally overlook the physical consistency between the degradation and fusion processes. Most of them optimize solely based on reconstruction errors, which limits the model's ability to accurately capture the true mapping between LR and HR domains, leading to poor performance in complex scenarios.

To address the limitations of existing unsupervised pan-sharpening methods, this paper presents a novel framework, **Rev**erse degradation for **Fus**ion (RevFus). RevFus is built upon an Invertible Neural Network comprising multiple Invertible Spatial–Spectral Degradation Blocks (ISSDB), enabling bidirectional mapping between high-quality MS images and degraded spectral and spatial domains. Rather than directly learning a fusion mapping, we first explicitly model the degradation process and then reverse it to perform fusion, which is termed degradation-to-fusion learning. This approach thus provides a more interpretable and physically grounded fusion process compared to conventional methods. To further enhance training stability and fidelity in degradation-to-fusion learning, we introduce a cycle-consistency self-learning (CCSL) objective, which enforces consistency between the degradation and fusion processes.

Although degradation-to-fusion learning enhance the stability and fidelity of unsupervised pan-sharpening, a spatial detail compensation stage is further designed in this paper to enhance the spatial details of the initial fused image. Building upon the initial fusion results, this stage aligns with the PAN image through the degradation in the spectral branch, and subsequently reprojects the missing spatial details into the MS domain using spectral transformation. Moreover, to enhance the collaborative modeling of spatial and spectral information, we introduce a spatial–spectral contrastive learning objective. Unlike traditional contrastive learning approaches that focus on feature alignment within a single modality (Khosla et al., 2020), our method considers improvement in both two domains, thereby enabling higher-quality fusion feature alignment.

While these strategies improve the visual quality and fidelity of the fused images, deep learning-based pan-sharpening models still suffer from a lack of interpretability. Their internal feature extraction and fusion mechanisms remain largely hidden, making it difficult to understand how the network balances spatial detail enhancement with spectral preservation. To address this, we propose a quantitative and systematic measure of model interpretability, the Interpretability Efficacy Coefficient (IEC). The IEC integrates multiple statistics derived from SHapley Additive exPlanations (SHAP) values, such as mean influence, standard deviation, top-3 feature focus, and influence entropy, into a single unified score. By combining concepts from information theory, statistical mechanics, and signal processing, IEC provides a quantitative measure of model interpretability.

The main contributions of this work are summarized as follows:

- We propose a novel unsupervised pan-sharpening framework based on invertible networks and cycle-consistency self-learning, termed degradation-to-fusion learning, which first models the degradation and then reverses it to achieve image fusion.
- A structural detail compensation stage is designed to further enhance spatial detail representation.
- A spatial–spectral contrastive learning objective is proposed to improve alignment of spatial and spectral features with high-quality references, ensuring effective integration of spatial and spectral information.
- A composite metric called interpretability efficacy coefficient is proposed to quantitatively measure the model interpretability of deep learning-based pan-sharpening models.

The remainder of this paper is organized as follows. Section 2 reviews the related work. The framework and methodological details are presented in Section 3. Section 4 reports experimental results on three datasets to validate the effectiveness of the proposed model and analyzes the proposed IEC. Finally, Section 5 concludes the paper.

2. Related works

2.1. Unsupervised pan-sharpening

Unsupervised pan-sharpening has attracted considerable attention, with numerous methods proposed to address the challenge of fusing high-resolution panchromatic (PAN) images and low-resolution multispectral (MS) images without relying on ground truth data. A prominent line of research explores the application of generative adversarial networks (GANs) in this area (Li et al., 2021; Zhou et al., 2020). For instance, Ma et al. (2020) introduced a dual-discriminator framework designed to enforce spatial and spectral fidelity separately. Ozcelik et al. (2020) leveraged reduced-resolution MS images processed via grayscale conversion and spatial downsampling to train their GAN. Zhou et al. (2021) developed a two-stream generator combined with dual discriminators and incorporated a cycle-consistency-based hybrid loss to enhance reconstruction quality (Zhou et al., 2022). Xu et al. (2023) proposed treating pan-sharpening as a two-stage fusion process by introducing an intermediate resolution scale to bridge the gap between PAN and MS images. Although GAN-based approaches are effective in scenarios lacking ground truth, they tend to generate new image content rather than strictly fuse existing spectral and spatial information, which can lead to undesirable artifacts.

Beyond GANs, other studies focus on refining loss functions and imposing stronger constraints to improve fusion quality (Liu et al., 2023; He et al., 2023). Luo et al. (2020) designed a novel loss formulation utilizing input MS and PAN images to simultaneously enhance spatial details and maintain spectral consistency. Ciotola et al. (2022a) further proposed a target-adaptive operating mode to adapt the network behavior dynamically. Qu et al. (2020) introduced a self-attention mechanism combined with sparse constraints and detail reconstruction loss, achieving improved unsupervised pan-sharpening performance. Ni et al. (2022) modeled degradation processes through multiple CNN blocks, strengthening the fusion constraints.

Additionally, some approaches integrate prior domain knowledge into the unsupervised learning framework. Seo et al. (2020) combined unsupervised learning with implicit registration techniques to better align PAN and MS images. Wang et al. (2022) employed meta-learning strategies to adapt supervised models for unsupervised scenarios. Uezato et al. (2020) introduced a guided deep decoder network aimed at refining features between PAN and multispectral domains, further improving fusion outcomes.

2.2. Contrastive learning

Contrastive learning has become one of the most influential self-supervised learning frameworks in recent years. Its fundamental goal is to learn feature representations by pulling together semantically similar samples (positive pairs) while pushing apart dissimilar ones (negative pairs) in the embedding space (Jaiswal et al., 2020). Early contrastive methods demonstrated remarkable performance in computer vision by leveraging large batch sizes and memory banks to construct effective positive and negative pairs. SimCLR (Chen et al., 2020) emphasized the importance of strong data augmentations and large batch sizes, whereas MoCo (He et al., 2020) introduced a momentum encoder to maintain a dynamic dictionary of negative samples, enabling effective training with smaller batches.

Beyond these, newer frameworks have relaxed the reliance on negative pairs. Methods like BYOL (Grill et al., 2020) and SimSiam (Chen and He, 2021) successfully learned meaningful representations without explicit negatives by using asymmetric network architectures and stop-gradient techniques to prevent collapse. Furthermore, advancements include the incorporation of hard negative mining (Kalantidis et al., 2020), multi-view contrastive learning (Tian et al., 2020), and multi-modal contrastive frameworks (Radford et al., 2021).

Contrastive learning has been widely adopted across various domains beyond natural images, including medical imaging, speech processing, and remote sensing. In particular, its ability to learn from unlabeled data makes it well-suited for fields where annotation is expensive or impractical. The surveyed literature also highlights theoretical insights into the relationships between contrastive loss functions, mutual information estimation, and downstream task performance (Tschannen et al., 2019).

2.3. Model interpretability

Growing requirements for trust, transparency, and controllability in satellite imagery processing have driven significant interest in model interpretability. These approaches are broadly categorized into four groups: Feature Attribution, Model Distillation, Intrinsic Interpretability, and Contrastive Examples (Höhl et al., 2024).

Feature Attribution methods include backpropagation-based and perturbation-based approaches, both aiming to reveal which input features most influence model predictions. Backpropagation methods exploit model internals, such as gradient maps or Integrated Gradients (Sundararajan et al., 2017), while Grad-CAM produces class-specific heatmaps for CNNs (Selvaraju et al., 2017). In contrast, perturbation methods modify inputs to measure changes in predictions (Fisher et al., 2019). The key difference lies in their reliance on the model: backpropagation uses internal information, whereas perturbation treats the model as a black box.

Building on feature-level explanations, model distillation techniques seek to approximate complex models with interpretable surrogates. Local distillation methods focus on individual predictions, using approaches such as LIME (Cheng et al., 2022) or Deep SHapley Additive exPlanations (SHAP) (Temenos et al., 2023). In contrast, model translation creates global surrogates, often with decision trees or rule-based models, to provide overall interpretability across the dataset (Augasta and Kathirvalavakumar, 2012). Together, these methods complement feature attribution by offering explanations at both local and global levels.

Intrinsic Interpretability takes a different approach by designing models or analyzing latent spaces to be inherently understandable. Interpretable-by-design models allow direct inspection of decision rules or coefficients (Guo et al., 2023). Latent-space methods interpret hidden activations using attention mechanisms (Khan et al., 2024) or concept-based approaches, providing insight into deep model representations. Joint training strategies, such as concept bottleneck models or ProtoPNet (Barnes et al., 2022), integrate auxiliary tasks to align latent representations with human-understandable concepts, bridging local and global interpretability.

Contrastive Examples explain model behavior by comparison with alternative instances. Counterfactual methods identify minimal changes required to alter predictions (Dantas et al., 2023), while example-based approaches retrieve similar historical instances (Ishikawa et al., 2023). The distinction is that counterfactuals generate artificial cases, whereas example-based methods rely on real instances. By providing intuitive comparisons, these approaches complement other interpretability methods and offer human-aligned reasoning.

3. Methodology

In this study, we propose a unsupervised fusion method called RevFus, which consists of two main stages: degradation-to-fusion learning and structural detail compensation. In the first stage, the original MSI is degraded into a LR MSI and a LR PAN image, which are then processed in reverse by RevFus to reconstruct the estimation of original MSI, with CCSL enforcing reconstruction fidelity. Thus, the information in the original data is exploited to initialize the network of Revfus. In the second stage, the parameters of RevFus are frozen. The initial fusion output is degraded into the PAN spectral domain in alignment-based detail reprojection, and missing spatial details are compensated using spectral alignment. Finally, S^2 contrastive learning is proposed to improve both spatial and spectral fidelity. Details are shown in Fig. 1. Furthermore, to measure the model interpretability comprehensively, a new index named IEC is proposed.

3.1. Problem formulations

Let $X \in \mathbb{R}^{W\times H\times C}$ denote the ideal HRMS image, where C represents the band number, and W and H are the spatial dimensions. The PAN image is denoted as $P \in \mathbb{R}^{W\times H\times 1}$, which shares the same spatial resolution as X but contains only a single spectral channel. The corresponding LRMS image is denoted by $M \in \mathbb{R}^{w\times h\times C}$, where $W/w = H/h = r$, with r being the scale factor between the HR and LR images. Typically, it is 4 in pan-sharpening tasks.

In real-world remote sensing scenarios, due to the inherent limitations of imaging sensors, the observed data consists of the spatially degraded multispectral image M and the spectrally degraded PAN image P, instead of the ideal HRMS image X. These degradation processes can be mathematically formulated as:

$$P = X\Phi \tag{1}$$

$$M = DX \tag{2}$$

where $\Phi \in \mathbb{R}^{C\times 1}$ is the spectral response functions of the PAN sensor, and $D \in \mathbb{R}^{(w\times h)\times(W\times H)}$ is the spatial degradation operator applied to X to generate the LRMS image.

Based on the formulations in Eqs. (1) and (2), the pan-sharpening task can be regarded as a typical inverse problem: reconstructing the HR multispectral image X from its spatially and spectrally degraded counterparts M and P. In deep learning-based approaches, this inverse mapping is generally learned through a model $\mathscr{F}(\cdot)$, which directly estimates the HRMS output from the degraded inputs:

$$\hat{X} = \mathscr{F}(M, P) \tag{3}$$

where $\hat{X}$ denotes the reconstructed HR multispectral image.

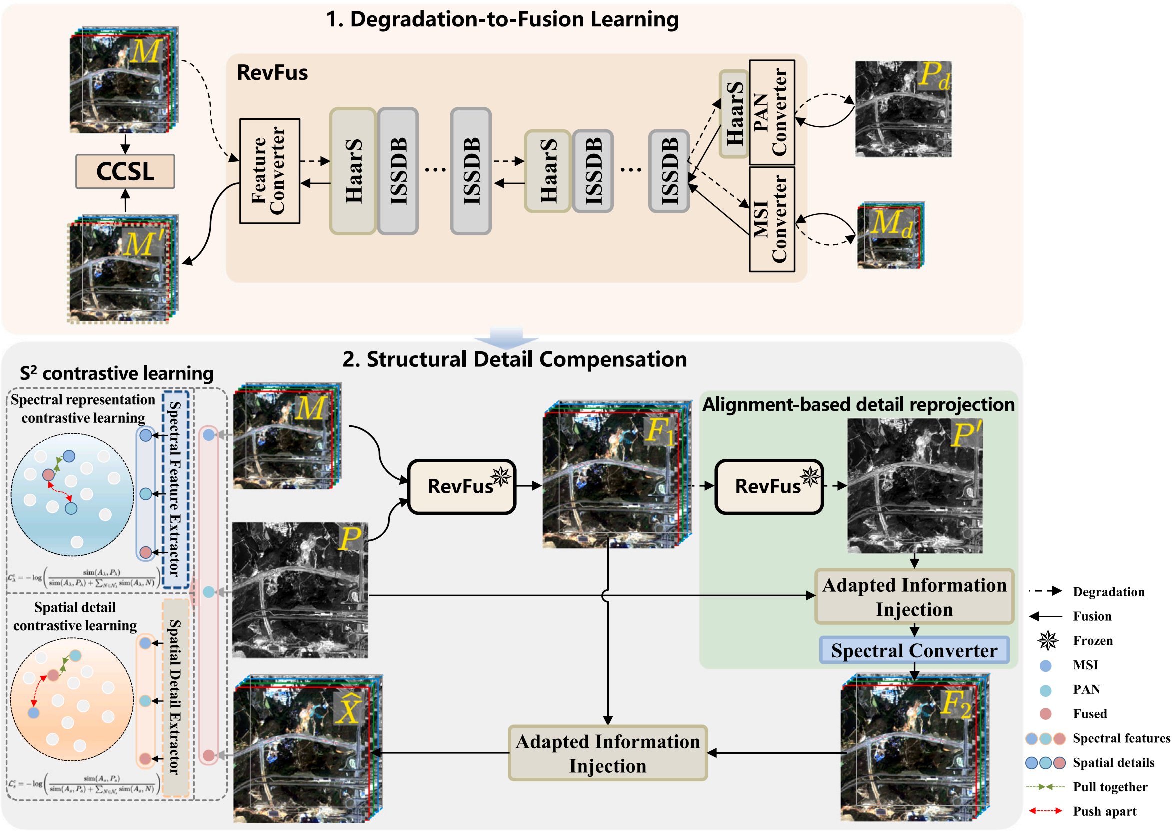

1. Degradation-to-Fusion Learning
M
CCSL
M'
RevFus
Feature Converter
HaarS
ISSDB
...
ISSDB
HaarS
ISSDB
...
ISSDB
HaarS
PAN Converter
MSI Converter
P_d
M_d
2. Structural Detail Compensation
S² contrastive learning
Spectral representation contrastive learning
Spectral Feature Extractor
$\mathcal{L}_\lambda^c = -\log\left(\frac{\mathrm{sim}(A_\lambda, P_\lambda)}{\mathrm{sim}(A_\lambda, P_\lambda) + \sum_{N\in\mathcal{N}_\lambda}\mathrm{sim}(A_\lambda, N)}\right)$
Spatial detail contrastive learning
Spatial Detail Extractor
$\mathcal{L}_s^c = -\log\left(\frac{\mathrm{sim}(A_s, P_s)}{\mathrm{sim}(A_s, P_s) + \sum_{N\in\mathcal{N}_s}\mathrm{sim}(A_s, N)}\right)$
M
P
RevFus
F_1
Alignment-based detail reprojection
RevFus
P'
Adapted Information Injection
Spectral Converter
F_2
Adapted Information Injection
X̂
Degradation
Fusion
Frozen
MSI
PAN
Fused
Spectral features
Spatial details
Pull together
Push apart

Fig. 1. Overview of the proposed RevFus framework.

3.2. *Degradation-to-fusion learning*

Most existing unsupervised fusion frameworks treat the pan-sharpening problem as a generation task with prior constraints and address it using generative algorithms. The advantage of this approach lies in leveraging large-scale existing samples to build a knowledge base, thereby enhancing the model's generalization capability in complex scenarios. However, this strategy overlooks that pan-sharpening fundamentally represents the inverse problem of the degradation process in remote sensing imaging. Consequently, generative algorithms may introduce artificial spectral information or spatial details, which significantly undermines the reliability of remote sensing images as physically meaningful data sources.

In this work, we exploit the unique bidirectional inference capability of invertible neural networks to propose a method grounded in the degradation mechanism of the imaging process. Our approach simultaneously learns both spatial and spectral degradation processes using the existing multispectral data itself. During this stage, spatial and spectral degradations are learned at a coarser scale, which reduces memory consumption and accelerates training. After the RevFus model $f(\cdot)$ is trained to convergence, we obtain two types of degraded observations derived from the original multispectral image M. Specifically, the model estimates a spatially degraded M_d^s, as well as a spectrally degraded version M_d^λ, where M_d^λ preserves the same spatial dimensions as the original image M:

$$\hat{M}_d^s, \hat{M}_d^\lambda = f(M) \tag{4}$$

Leveraging the near-lossless reversibility of invertible neural networks, when the multispectral image M and the panchromatic image P are fed backward into the model, the fused result F_1 can be obtained.

$$F_1 = f^{-1}(M, P) \tag{5}$$

where F_1 denotes the initial fused results and $f^{-1}(\cdot)$ represents the reverse process of RevFus. To enable the model to achieve reversibility while learning the degradation process, we not only propose invertible spatial–spectral degradation blocks and invertible Haar wavelet sampler, but also introduce cycle-consistency self-learning.

3.2.1. *Network architecture of RevFus*

Considering that directly performing a one-step up- and down-sampling would introduce a large information gap, complicating spatial–spectral alignment and fusion, the proposed RevFus adopts a multi-scale architecture. At each scale, the data is up- or down-sampled by a ratio of two. Feature extraction at different scales incorporates multiple invertible spatial–spectral degradation blocks, as illustrated in the first step of Fig. 1.

Most existing methods typically rely on simple interpolation or deconvolution to change image scales. However, these operations are generally non-invertible, and some, such as pixelshuffle, achieve only approximate invertibility while still incurring information loss. Inspired by traditional multi-resolution analysis-based pan-sharpening methods, RevFus employs an invertible wavelet-based approach, termed the Haar wavelet sampler, to achieve lossless up- and down-sampling. The Haar wavelet transform, one of the simplest and most classical discrete wavelet transforms, decomposes an image into four approximation components with different frequencies (LL, LH, HL, HH) using simple addition and subtraction. Since the transformation relies solely on linear and invertible operations, it is inherently lossless and can perfectly reconstruct the original data via the inverse transform (King and Wang, 2001).

This multi-resolution decomposition can be represented as a set of fixed 2 × 2 convolution kernels that decompose the input image into a low-frequency component and three high-frequency components.

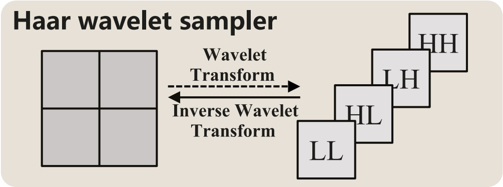


Fig. 2. Haar wavelet sampler with down- and up-sampling processes.

Specifically, the Haar wavelet decomposition kernels are defined as:

$$K_{LL} = \frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \quad K_{LH} = \frac{1}{2}\begin{bmatrix}-1 & -1\\ 1 & 1\end{bmatrix}, \tag{6}$$

$$K_{HL} = \frac{1}{2}\begin{bmatrix}-1 & 1\\ -1 & 1\end{bmatrix}, \quad K_{HH} = \frac{1}{2}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}. \tag{7}$$

Since these operations correspond to orthogonal transformations, they are fully invertible. The inverse transform simply convolves the sub-bands with the corresponding kernels and sums them to recover the original image. This convolutional implementation of the wavelet sampler can be integrated into CNNs, enabling RevFus to achieve truly lossless and invertible information propagation. Compared to traditional interpolation, Haar wavelet up- and down-sampling not only guarantees invertibility but also explicitly separates structural and detailed features, providing cleaner and more complementary representations for subsequent spatial–spectral fusion. The down- and up-sampling process of Haar wavelet sampler are shown in Fig. 2. Cause the typical resolution ratio in pan-sharpening is four, RevFus employs two cascaded Haar wavelet samplers to progressively match resolutions.

In RevFus, feature extraction and fusion at each scale are composed of multiple invertible spatial–spectral degradation blocks, as shown in Fig. 3. The key concept is to divide the input feature evenly along the channel dimension, resulting in two parts, (x_1, x_2), and perform a sequence of invertible transformations to enable bidirectional information flow and reconstruction. The forward propagation is formulated as:

$$\begin{aligned} y_1 &= x_1 \odot \exp\big(\text{Scaling}(E(x_2))\big) + F(x_2), \\ y_2 &= x_2 \odot \exp\big(\text{Scaling}(G(y_1))\big) + H(y_1) \end{aligned} \tag{8}$$

where $\text{Scaling}(x) = 2\cdot\sigma(x) - 1$ denotes the scaling operation, which maps the output of the *Sigmoid* function $\sigma(\cdot)$ from the range $[0, 1]$ to $[-1, 1]$. E, F, G, H are nonlinear mappings implemented using DenseBlocks ().

Since the scaling operations are exponential, the Jacobian determinant of the transformation can be expressed as:

$$\det J = \exp\Big(\sum \text{Scaling}(E(x_2)) + \sum \text{Scaling}(G(y_1))\Big) > 0 \tag{9}$$

which is strictly positive, ensuring the transformation is mathematically invertible. Leveraging this property, the transformation can be exactly reversed as:

$$\begin{aligned} x_2 &= (y_2 - H(y_1)) / \exp\big(\text{scaling}(G(y_1))\big), \\ x_1 &= (y_1 - F(x_2)) / \exp\big(\text{scaling}(E(x_2))\big) \end{aligned} \tag{10}$$

This design ensures that each block preserves information in both forward and backward directions, minimizing redundancy loss in spatial–spectral fusion and enhancing the invertibility and stability of the reconstructed images.

3.2.2. Optimization objective

In the Degradation-to-Fusion Learning stage, our goal is to accurately model the image degradation process and, combined with the nearly lossless inverse propagation of invertible neural networks, achieve a natural transition from degradation learning to fusion learning. In this stage, how to use the information in the original MS image to learn the spatial and spectral degradation is crucial. Therefore, we propose a coarse-scale self-supervised learning (CSSL) strategy that uses the features of the original MS and PAN images at the coarse scale to simulate the degradation process, thereby using the original MS information as the target to accurately learn the spatial–spectral degradation. The corresponding loss function is defined as follows:

$$\mathcal{L}_{cs} = \beta_\lambda \mathcal{L}_\lambda + \beta_s \mathcal{L}_s \tag{11}$$

$$\mathcal{L}_\lambda = \|\hat{M}_d^\lambda - P_d^s\|_1 \tag{12}$$

$$\mathcal{L}_s = \|\hat{M}_d^s - M_d^s\|_1 \tag{13}$$

Here, $\mathcal{L}cs$ represents the coarse-scale self-supervised loss, which is composed of a spectral consistency term $\mathcal{L}_\lambda$ and a spatial consistency term $\mathcal{L}_s$. The spectral component enforces that the predicted degraded image $\hat{M}_d^\lambda$ remains consistent with the spectrally-degraded MS image M_d^λ, as well as with the spatially-degraded PAN image P_d^s. The spatial component, on the other hand, ensures that the spatial details of the predicted degraded MS image $\hat{M}_d^s$ are well preserved with respect to the reference degraded MS image M_d^s. The two terms are weighted by coefficients $\beta\lambda$ and β_s to control their relative influence.

To further improve the model fidelity of the proposed RevFus, we introduce a cycle-consistency self-learning strategy. Specifically, in the forward degradation process, RevFus takes the original MS image as input and generates the degraded MS image and the PAN image. In the backward fusion process, the network reconstructs a HR fused MS image from the LR MS image with PAN data. The cycle-consistency constraint feeds the forward-degraded outputs back into the network in the backward direction to produce an estimate of the original MS image, enforcing consistency between the reconstructed and original MS images. This design encourages the network to learn a nearly lossless mapping between the degradation and fusion processes, thereby enhancing spectral and spatial fidelity, stabilizing the training procedure, and effectively reducing artifacts in the reconstructed HR MS images.

The corresponding cycle-consistency loss is formulated as:

$$\mathcal{L}_{cc} = \|f^{-1}(f(M)) - M\|_1 \tag{14}$$

where $f(\cdot)$ denotes the forward degradation process, $f^{-1}(\cdot)$ denotes the backward fusion process, and $\|\cdot\|_1$ measures the L_1 distance between the reconstructed and original MS images. This loss explicitly enforces that the network's forward and backward mappings are consistent.

Overall, the optimization objective in the Degradation-to-Fusion Learning stage consists of two components: a coarse-scale self-supervised loss and a cycle-consistency self-learning constraint. These two terms are balanced using weighting coefficients α and β:

$$\mathcal{L}_{D2F} = \alpha\mathcal{L}_{cc} + \beta\mathcal{L}_{cs} \tag{15}$$

3.3. Structural detail compensation

Compared with supervised pan-sharpening methods, the main challenge of unsupervised pan-sharpening lies in the absence of high-resolution reference images to constrain the model learning. Without sufficient supervision, the fused results often suffer from insufficient spatial details. Although the proposed degradation-to-fusion learning effectively models both the real degradation process and its inverse fusion, the lack of HR multispectral ground truth still limits the recovery of fine spatial structures.

To address this challenge, this paper proposes a further structural detail compensation stage based on spectral alignment to restore the spatial details lost during the Degradation-to-Fusion Learning stage. Specifically, after training RevFus in degradation-to-fusion learning stage and freezing its parameters, the MS and PAN images are first fed into the fusion path of RevFus to generate an initial fused result F_1. Then, by applying the forward degradation process of RevFus to F_1, P' is obtained, which carries the spatial information embedded in F_1 due to its spectral degradation origin. Next, a spectral alignment-based

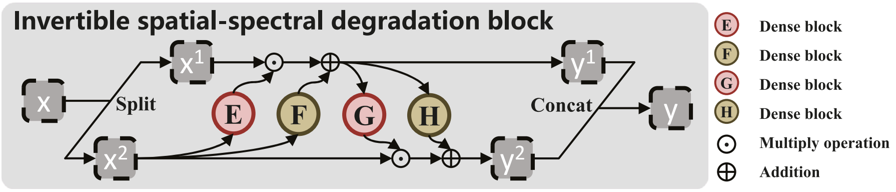


Fig. 3. The degradation process of the proposed invertible spatial–spectral degradation block.

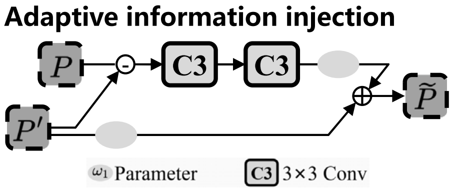


Fig. 4. Adaptive information injection module for alignment-based detail reprojection.

detail reprojection mechanism is applied to recover the spatial details absent in F_1 but present in the original PAN, yielding F_2. Finally, an adaptive information injection module integrates F_1 and F_2 to obtain the final fused result $\hat{X}$. Meanwhile a spatial–spectral contrastive learning scheme is adopted to further enhance spatial–spectral consistency and detail preservation.

3.3.1. Alignment-based detail reprojection

In practice, the input PAN and MS images are not only different in spatial resolution but also in spectral characteristics, making it particularly challenging to simultaneously address both types of discrepancies. To more effectively recover the missing details in F_1, it is first transformed into PAN domain, while also accounting for the intrinsic spectral relationship between the two sensors.

In RevFus, the spatial–spectral degradation process has already been modeled during the degradation-to-fusion learning stage. Thus, its forward degradation path can be exploited to project F_1 into the PAN spectral space:

$$P' = f^{-1}(F_1) \tag{16}$$

Although P' has been spectrally aligned, it still lacks the fine spatial details contained in the original PAN image P. To address this, both P' and P are input into an Adaptive Information Injection (AII) module (see Fig. 4), which adaptively transfers the missing spatial information from P to P', resulting in an enhanced representation $\widetilde{P}$:

$$\widetilde{P} = \mathrm{AII}\left(P', P\right) = \omega_1 \cdot \mathrm{Conv}_2\left(\mathrm{Conv}_1(P' - P)\right) + \omega_2 \cdot P' \tag{17}$$

Here, ω_1 and ω_2 are learnable weights, while Conv_1 and Conv_2 are 3×3 convolution layers. These weights balance the contribution of the original aligned features and the enhanced difference information to construct $\widetilde{P}$.

Through this reprojection mechanism, P' is effectively supplemented with the spatial structures guided by the original PAN image. The enhanced representation $\widetilde{P}$ is then mapped back into the MS spectral domain using a spectral converter $\mathcal{F}sc$, yielding a spectrally consistent yet spatially enriched representation F_2:

$$F_2 = \mathcal{F}sc\left(\widetilde{P}\right) \tag{18}$$

Finally, after obtaining the spatially enhanced representation F_2, another AII module is employed to adaptively integrate F_1 and F_2, leading to the final fused output:

$$\hat{X} = \mathrm{AII}\left(F_1, F_2\right) \tag{19}$$

The final fused output $\hat{X}$ thus preserves spectral consistency while restoring high-quality spatial details.

3.3.2. Spatial–spectral contrastive learning

To enhance the representational capacity of the fusion network, we propose a spatial–spectral contrastive learning (S^{2}CL) strategy that jointly constrains the spectral consistency and spatial structure preservation during training. As shown in Fig. 1, the framework integrates both spectral representation contrast and spatial detail contrast under a unified optimization objective, guided by a consistency constraint.

Given a fused image $\hat{X}$, original MS image M, and PAN image P, we first extract spectral representations using a spectral encoder:

$$E_\lambda^A = \mathcal{E}_\lambda(\hat{X}), \quad E_\lambda^P = \mathcal{E}_\lambda(M), \quad E_\lambda^N = \mathcal{E}_\lambda(P) \tag{20}$$

where $\mathcal{E}_\lambda(\cdot)$ denotes the spectral feature extractor consists of two large-kernel convolutional layers and a global average pooling layer, yielding a 128-dimensional spectral feature vector. To enforce spectral fidelity, the spectral feature contrastive loss is defined as:

$$\mathcal{L}_\lambda^c = -\log\left(\frac{\mathrm{sim}(E_\lambda^A, E_\lambda^P)}{\mathrm{sim}(E_\lambda^A, E_\lambda^P) + \sum_{E^N \in \mathcal{N}_\lambda} \mathrm{sim}(E_\lambda^A, E^N)}\right) \tag{21}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity, and $\mathcal{N}_\lambda$ represents the set of negative PAN spectral samples.

At the spatial level, we extract structural details from $\hat{X}$, M, and P using a Sobel operator, obtaining:

$$E_s^A = \nabla \hat{X}, \quad E_s^P = \nabla P, \quad E_s^N = \nabla M \tag{22}$$

The fused image spatial feature E_s^A is encouraged to align with the HR PAN image while remaining distinct from the LR MS image. The spatial detail contrastive loss is formulated as:

$$\mathcal{L}_s^c = -\log\left(\frac{\mathrm{sim}(E_s^A, E_s^P)}{\mathrm{sim}(E_s^A, E_s^P) + \sum_{E^N \in \mathcal{N}_s} \mathrm{sim}(E_s^A, E^N)}\right) \tag{23}$$

To further ensure global spectral consistency, a reconstruction-based consistency loss is introduced:

$$\mathcal{L}_{\mathrm{con}} = \| D_{spa}(\hat{X}) - M \|_1 \tag{24}$$

where $D_{spa}(\cdot)$ denotes the spatial degradation. A cosine annealing strategy is adopted to dynamically balance spectral and spatial constraints during training:

$$\omega_\lambda = \frac{1}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \tag{25}$$

where t is the current training iteration, and T is the total schedule length. This design emphasizes spectral learning in early stages while gradually balancing spatial and spectral objectives.

Finally, the overall training objective in this stage integrates all three components:

$$\mathcal{L}_{SDC} = \mathcal{L}_{\mathrm{con}} + \varepsilon\left[(1 - \omega_\lambda)\mathcal{L}_\lambda^c + \omega_\lambda \tau \mathcal{L}_s^c\right] \tag{26}$$

where ε and τ serve as scaling factors to unify the magnitude of different loss terms, ensuring they contribute effectively in the same optimization framework. This formulation enables the fused image to achieve a desirable trade-off between spectral fidelity and spatial detail preservation.

3.4. Interpretability efficacy coefficient

In pan-sharpening, deep learning models have achieved remarkable progress in improving both spatial and spectral quality. However, their inherent black-box nature makes it difficult to clearly understand how features are extracted and fused within the network. The balance between spatial detail enhancement and spectral preservation is often implicit and not directly observable. This limitation reduces the trustworthiness, transparency, and controllability of such models, and may further affect the reliability of fused outputs in downstream applications. Therefore, incorporating interpretability analysis is essential for evaluating and achieving explainable artificial intelligence-based fusion methods.

SHAP, one of the most classical and widely used model interpretability methods, is a game theory-based feature attribution approach that quantifies the influence of different features on the result. It can effectively describe the local contribution of each input spectral bands to the model output in pan-sharpening. However, existing studies still lack a global, quantitative, and comparable measure that can consistently evaluate interpretability across models.

Friedman and Popescu (2008) suggested that an effective interpretable model should strike a balance between conciseness and feature importance. Motivated by this principle, this study builds upon the SHAP framework and introduces a new composite interpretability metric, interpretability efficacy coefficient. Drawing inspiration from the concept of system efficiency in energy theory, the IEC integrates multiple statistical measures derived from SHAP values into a single, unified score. By combining principles from information theory, statistical mechanics, and signal processing, the proposed IEC aims to quantify the efficiency of interpretability in pan-sharpening models. In other words, it evaluates the sufficiency, stability, and conciseness of the model's interpretability throughout the fusion process.

3.4.1. Fundamental statistics of SHAP matrix

Since SHAP values are local explanation indicators and can only be compared in terms of relative magnitude within the same model and sample, we normalize them to obtain a unified representation. In this work, given a model with N input spectral bands and M output bands, we denote the normalized SHAP matrix as $\tilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M\times N}$, where each element $\tilde{\phi}_{j,i}$ measures the absolute contribution of the ith input band to the jth fused output band. The matrix $\tilde{\boldsymbol{\Phi}}$ is obtained by applying a global normalization:

$$\tilde{\phi}_{j,i} = \frac{|\phi_{j,i}|}{\max|\boldsymbol{\Phi}|}, \tag{27}$$

where $\max|\boldsymbol{\Phi}|$ denotes the maximum of all sampled SHAP values. Based on $\tilde{\boldsymbol{\Phi}}$, we extract four fundamental metrics that capture different aspects of model interpretability for each output band.

Mean Influence: This metric quantifies the average contribution strength of input bands to each output, reflecting the overall utilization of spectral information:

$$R^j = \frac{1}{N}\sum_{i=1}^{N}\tilde{\phi}_{j,i}, \qquad R = \frac{1}{M}\sum_{j=1}^{M}R^j. \tag{28}$$

where R^j denotes the overall contribution to the jth output band. A higher value indicates stronger overall feature utilization (Friedman and Popescu, 2008).

Standard Deviation of Influence: It measures the variation of band contributions within each output, indicating the stability and balance of model attention across inputs:

$$S^j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\phi}_{j,i} - R^j\right)^2}, \qquad S = \frac{1}{M}\sum_{j=1}^{M}S^j. \tag{29}$$

A lower S indicates that the model distributes attention more evenly across input bands, leading to more robust interpretability, whereas a higher S suggests excessive sensitivity to certain bands (Ghorbani et al., 2019).

Top-3 Focus: It reflects the concentration degree of model attention by computing the proportion of the top-3 largest contributions relative to the total influence:

$$F^j = \frac{\sum_{a=1}^{3}\tilde{\phi}_j^a}{\sum_{i=1}^{N}\tilde{\phi}_{j,i}}, \qquad F = \frac{1}{M}\sum_{j=1}^{M}F^j, \tag{30}$$

where $\tilde{\phi}_j^a$ denotes the ath largest contribution for the jth output channel. A higher F indicates that only a few key spectral bands dominate the fusion decision, which often aligns with physical priors in remote sensing.

Influence Entropy: From an information-theoretic perspective, this metric measures the dispersion of contributions across input bands using Shannon entropy (Shannon, 1948):

$$p_{j,i} = \frac{\tilde{\phi}_{j,i}}{\sum_{i=1}^{N}\tilde{\phi}_{j,i}}, \qquad E^j = -\sum_{i=1}^{N}p_i^j\log(p_i^j), \qquad E = \frac{1}{M}\sum_{j=1}^{M}\frac{E^j}{\log N}. \tag{31}$$

where $p_{j,i}$ denotes row-wise normalization of the Shapley matrix, E^j is the Shannon entropy of jth output channel. A higher E indicates an overly uniform contribution distribution, suggesting that the model's decision logic is diffuse and lacks clear focus.

3.4.2. The proposed IEC

Different statistical metrics focus on different aspects of model interpretability. Relying on any single metric alone may result in a partial assessment, failing to capture the overall behavior of the model. Therefore, in this work, we propose a novel comprehensive metric, the IEC, which consolidates four fundamental statistics and can comprehensively evaluate the strength, uncertainty, and focus of model interpretability. The formulation of IEC is defined as:

$$\eta = \frac{R\cdot F}{S\cdot E}\cdot e^{-|1-F|} \tag{32}$$

where the product $R\cdot F$ captures the Effective Explanatory Energy, reflecting the overall strength and focus of the model's explanations. The term $S\cdot E$ quantifies the Explanatory Uncertainty Dissipation, indicating variability and dispersion in feature contributions. Finally, $e^{-|1-F|}$ serves as a Decision Conciseness Correction, quantifying the concentration of the model's explanations.

Effective Explanatory Energy: This term quantifies the "effective energy" provided by the explanation. The mean influence R reflects the overall contribution of features, while the focus F reflects the simplicity of the decision logic. Their product indicates that an efficient explanation must exhibit both high impact and high focus. As we all know, in pan-sharpening tasks, the fused results are typically expected to be highly correlated with the spectral information of the corresponding original bands and the spatial details provided by the PAN band. This corresponds to a model that fully exploits band information (high R) while focusing its decision on a few physically meaningful key bands (high F), consistent with multispectral imaging priors (Friedman and Popescu, 2008).

Explanatory Uncertainty Dissipation: This component quantifies the "energy loss" or noise inherent in the explanation. By placing S and E in the denominator, the metric penalizes explanations that are either unstable or ambiguous. *(1) Statistical Noise S:* A large S indicates substantial variation in Shapley values across spectral bands, implying low stability in feature attribution (Ghorbani et al., 2019). In pan-sharpening, high S implies minor fluctuations in non-key bands disproportionately affect the fusion result. *(2) Information Noise E:* A high entropy E reflects a diffuse and unfocused decision logic with lower informational clarity (Shannon, 1948; Kay, 1993). In the context of pan-sharpening, this suggests that the model treats all bands nearly equally, failing to highlight the core spectral bands that are most critical for producing high-quality fusion results.

Decision Conciseness Correction: Inspired by the Boltzmann distribution in statistical mechanics, $|1-F|$ is defined as the "energy of decision dispersion" (Chandler, 1987). When F approaches 1, the decision becomes highly concentrated, corresponding to a stable, low-energy state in which the correction term $e^{-|1-F|}$ behaves almost linearly. As F decreases, the decision grows increasingly dispersed, leading to an unstable, high-energy state where the penalty $e^{-|1-F|}$ escalates exponentially. The term $e^{-|1-F|}$ thus encourages concise and transparent explanations that align with physical priors, ultimately enhancing both interpretability and generalizability (Rasmussen and Ghahramani, 2001).

Overall, the IEC constructs a multi-factor cooperative evaluation framework. It combines "Effective Explanatory Energy" and "Explanatory Uncertainty Dissipation" as the core efficacy foundation, further enhanced by the non-linear "Decision Conciseness Correction" to incentivize simplicity. This composite structure integrates principles from information theory, statistical mechanics, and signal processing with the physical characteristics of pan-sharpening, elevating interpretability evaluation from abstract attribute description to a quantitative measure of model interpretability efficiency. It provides a powerful tool for quantifying decision logic simplicity and robustness in pan-sharpening models.

4. Experimental results

This section presents experiments designed to evaluate RevFus. We include both reduced- and full-resolution experiments on three data sets, as well as an ablation study and further discussions.

4.1. Experimental setting

4.1.1. Data sets

This study employs three satellite data sets for both reduced- and full-resolution experiments: QuickBird (QB), Gaofen-2 (GF2), and WorldView-2 (WV2).

QB captures a PAN image along with four multispectral (MS) channels covering the visible to near-infrared spectrum (Red, Green, Blue, and NIR). The resolutions are 0.61 m and 2.44 m, respectively. For our experiments, QB images are drawn from Shenzhen, China, representing an urban environment dominated by buildings and roads.

GF2 provides images through PAN and four MS channels (Blue, Green, Red, NIR), with resolutions of 0.81 m and 3.24 m, respectively. The GF2 imagery used in this work comes from Nanning, China, including diverse terrain such as vegetation, land, and water bodies.

WV2 acquires one PAN channel with eight MS channels, including standard Blue, Green, Red, plus Coastal-Blue, Yellow, Red-Edge, NIR1, and NIR2. Its spatial resolutions are 0.46 m and 1.85 m. The selected WV2 data covers San Francisco, USA, featuring a mixture of urban structures, hills, and vegetation.

Collectively, these data sets offer a range of landscapes and acquisition conditions, providing a robust basis for evaluating the performance of the proposed model. All results presented are derived from these selected satellite images.

4.1.2. Comparison methods

In this study, we compare our RevFus with seven traditional algorithms and six state-of-the-art deep learning-based methods, including BDSD (Garzelli et al., 2007), Adaptive Gram–Schmidt transformation (GSA) (Aiazzi et al., 2007), Adaptive Component Substitution with Partial Replacement (PRACS) (Choi et al., 2010), Modulation Transfer Functions-Generalized Laplacian Pyramid-High-Pass Modulation (MTF-GLP-HPM) (Aiazzi et al., 2006), ATrous WaveleT (ATWT)-M3 (Ranchin and Wald, 2000), AWLP (Otazu et al., 2005), TV (Palsson et al., 2013), Z-PNN (Ciotola et al., 2022b), LDPNet (Ni et al., 2022), ZSPan (Cao et al., 2024), PGMAN (Zhou et al., 2021), TFResNet (Liu et al., 2020c), PLRDiff (Rui et al., 2024), and UCL (Xiao et al., 2026).

Traditional algorithms and five of deep learning-based methods can quickly achieve unsupervised pan-sharpening without training, while TFResNet are data-driven. We deploy it with the proposed CCSL.

4.1.3. Quantitative metrics

Two types of experiments are conducted in this study: reduced-resolution testing following Wald's protocol (Wald et al., 1997), and full-resolution testing.

In reduced-resolution evaluation, five quantitative metrics are employed to assess pan-sharpening performance in both spatial and spectral domains: *Correlation Coefficient* (CC), *mean Peak Signal-to-Noise Ratio* (mPSNR, in dB), *mean Structural SIMilarity* (mSSIM) (Wang et al., 2004), *Spectral Angle Mapper* (SAM, in degrees) (Kruse et al., 1993), and *Erreur Relative Globale Adimensionnelle de Synthèse* (ERGAS). Higher values of CC, mPSNR, and mSSIM indicate better image quality, whereas lower values of RMSE, SAM, and ERGAS correspond to reduced distortion.

For full-resolution evaluation, four metrics are considered: *Quality with No Reference* (QNR) (Alparone et al., 2008), *High-Quality QNR* (HQNR) (Aiazzi et al., 2014), and the two components of HQNR, namely the spectral distortion D_λ^K and the spatial distortion D_s. Better fusion results are reflected by lower distortion indices, which lead to higher QNR or HQNR values.

4.1.4. Implementation details

The proposed approach consists of two stages. In RevFus, the feature dimension is 64, and a scale of two indicates the employment of two Haar wavelet samplers. In the degradation-fusion learning stage, the weighting coefficients are set as $\alpha = 0.33$ and $\beta = 0.67$, with $\beta_\lambda = \beta_s = 0.5$. In the structural detail compensation stage, the parameters ϵ and τ are set to 0.1 and 0.001, respectively. Both stages are optimized using the Adamax algorithm with an initial learning rate of 0.001 under a OneCycleLR schedule, which dynamically adjusts the learning rate to improve convergence and overall performance. All CNN-based models are implemented in PyTorch and trained on a Linux workstation equipped with 1 TB of RAM and an NVIDIA A40 GPU. In contrast, conventional methods are executed in MATLAB on an Intel Core i7-1355U CPU (1.70 GHz). Results for all traditional methods were generated using the pan-sharpening MATLAB toolbox developed by Vivone et al. (2020), while results for deep learning-based methods were obtained using the implementations provided by their respective authors.

4.2. Results on QuickBird dataset

Table 1 presents the quantitative results obtained on the QuickBird dataset. Additionally, Fig. 5 illustrates the visual performance in false-color composite format for the reduced-resolution experiments.

4.2.1. Reduced-resolution experiments

As shown in Table 1, among the traditional pan-sharpening methods, PRACS achieves the best overall performance, followed by TV, which slightly surpasses PRACS in SAM. BDSD performs the worst in both spectral fidelity and spatial preservation. Deep learning-based approaches generally outperform conventional algorithms. For instance, TFResNet improves mPSNR by approximately 1.8–2 dB over traditional methods and reduces SAM, highlighting the benefits of learning spatial and spectral features. In contrast, generative-based methods, including GAN- and diffusion model-based approaches, tend to exhibit lower performance across these metrics. Our proposed RevFus achieves the best results in all evaluated indices, attaining the highest spatial evaluation indexes, as well as the lowest SAM and ERGAS. This marked improvement can be attributed to the degradation-to-fusion learning strategy and the cycle-consistency self-learning, which provide physically grounded constraints and enable more effective spectral–spatial feature alignment.

Since the key distinction of remote sensing images from conventional RGB images lies in the near-infrared (NIR) band, we adopt false-color synthesis on the QuickBird dataset to better highlight the NIR information. This allows us to more clearly analyze the performance of different algorithms in preserving spectral fidelity within this

Table 1
Quantitative assessment on the QuickBird data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution testing					Full-resolution testing			
	CC	mPSNR	mSSIM	SAM	ERGAS	QNR	D_s	D_λ^K	HQNR
BDSD	0.9294	30.4495	0.8291	4.3523	3.6824	0.9190	0.0150	0.3438	0.6463
GSA	0.9373	31.4089	0.8234	3.9263	3.3650	0.8160	0.0690	0.2290	0.7177
PRACS	0.9528	33.6325	0.8627	3.2344	2.5214	0.9484	0.0293	0.2239	0.7534
MTF-GLP-HPM	0.9414	31.7537	0.8485	3.4660	3.2187	0.8151	0.0645	**0.1269**	0.8168
ATWT-M3	0.9393	32.3005	0.8318	3.7516	2.9624	0.9301	0.0351	0.3106	0.6652
AWLP	0.9323	31.0635	0.8278	3.3574	3.4181	0.7903	0.0570	0.2986	0.6614
TV	0.9497	32.7300	0.8574	3.2342	2.8122	0.9450	0.0415	0.3664	0.6073
Z-PNN	0.9104	29.4261	0.7882	5.3894	4.4199	0.8160	0.0317	0.1589	0.8144
LDPNet	0.9293	31.7835	0.8101	3.7093	3.0926	0.9565	0.0313	0.2517	0.7249
ZSPan	0.9170	30.2634	0.7863	4.9296	3.7621	0.8665	0.0580	0.1580	0.7932
PGMAN	0.9224	28.5246	0.7779	4.2196	4.5097	0.9576	0.0412	0.2789	0.6914
TFResNet	0.9382	32.4348	0.8562	3.8137	2.8966	0.9028	0.0287	0.3173	0.6631
PLRDiff	0.8124	27.4977	0.6527	9.2623	5.3881	0.8086	0.1457	0.2622	0.6303
UCL	0.9385	31.7408	0.8507	3.3000	3.0796	0.8924	0.0300	0.2265	0.7503
RevFus	**0.9614**	**34.3730**	**0.9084**	**3.0439**	**2.3173**	**0.9776**	**0.0117**	0.1379	**0.8520**

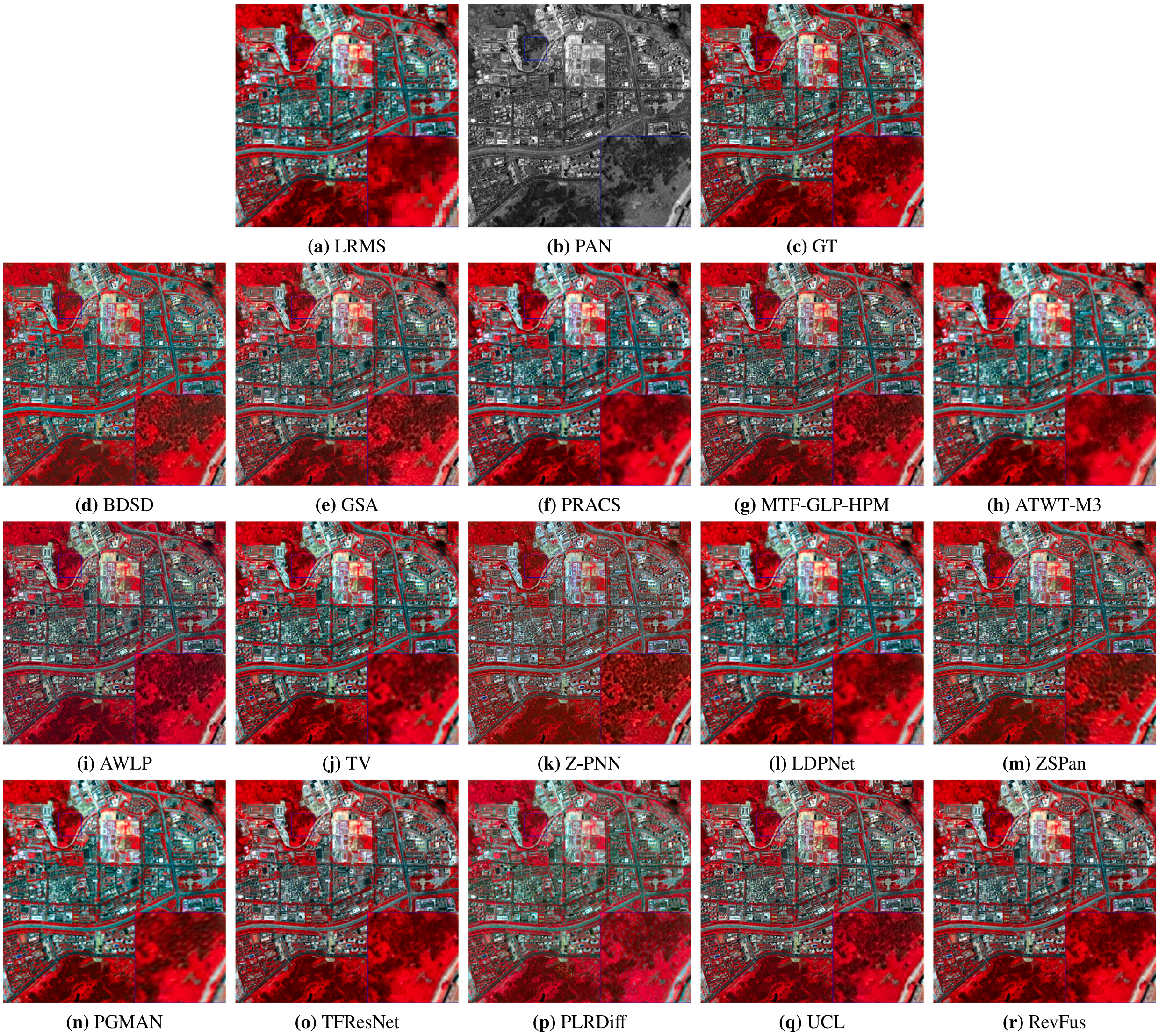

(a) LRMS (b) PAN (c) GT
(d) BDSD (e) GSA (f) PRACS (g) MTF-GLP-HPM (h) ATWT-M3
(i) AWLP (j) TV (k) Z-PNN (l) LDPNet (m) ZSPan
(n) PGMAN (o) TFResNet (p) PLRDiff (q) UCL (r) RevFus

Fig. 5. Visual comparison of pan-sharpening methods on the QuickBird dataset in the reduced-resolution testing scenario. The results are displayed in false-color, where bands 4, 3, and 2 correspond to the RGB channels.

critical band. As illustrated in the enlarged vegetation regions of Fig. 5, traditional methods such as GSA and AWLP often either over-sharpen or blur spectral details. PRACS and TV preserve spectral fidelity but exhibit minor blurring or smoothing. Z-PNN produces overly sharp textures, making vegetation details appear unnatural, while PLRDiff introduces noticeable noise and artifacts. LDPNet renders the vegetation regions overly smooth, losing fine texture details. In contrast, ZSPan and our proposed RevFus maintain texture patterns that closely

Table 2
Quantitative assessment on the Gaofen-2 data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution testing					Full-resolution testing			
	CC	mPSNR	mSSIM	SAM	ERGAS	QNR	D_s	D_λ^K	HQNR
BDSD	0.9048	31.3159	0.8376	4.5906	5.1625	0.9199	0.0555	0.0435	0.9034
GSA	0.9307	31.9031	0.8569	3.8487	4.7055	0.8554	0.0594	0.0330	0.9095
PRACS	0.9499	36.0971	0.9061	3.6958	3.5148	0.9193	**0.0177**	0.0478	0.9353
MTF-GLP-HPM	0.9337	34.2323	0.8915	3.6741	4.4738	0.7983	0.0739	0.0235	0.9043
ATWT-M3	0.9364	34.5333	0.8851	4.1665	4.0644	0.8910	0.0576	0.0538	0.8917
AWLP	0.9163	32.1412	0.8507	3.5407	4.6949	0.7962	0.0715	0.0247	0.9055
TV	**0.9635**	**37.4370**	0.9329	3.3583	**2.9843**	0.9265	0.0588	0.0384	0.9051
Z-PNN	0.8818	29.8225	0.7929	7.7254	9.0632	0.7966	0.0622	0.0476	0.8931
LDPNet	0.9246	33.9958	0.8721	4.2671	3.9661	0.9041	0.0819	0.0503	0.8719
ZSPan	0.8856	31.6318	0.8049	6.3007	4.9554	0.8010	0.1310	0.0404	0.8338
PGMAN	0.9253	33.9528	0.8686	3.8240	3.9588	0.8767	0.1206	0.0509	0.8347
TFResNet	0.9008	32.9904	0.8672	4.6878	4.7426	0.8822	0.0454	0.0573	0.8999
PLRDiff	0.9180	29.6733	0.8440	12.8467	7.5355	0.8777	0.0756	0.0475	0.8805
UCL	0.9261	33.5803	0.8730	3.8389	4.2803	0.9046	0.0521	0.0289	0.9205
RevFus	0.9602	36.5191	**0.9379**	**3.2124**	3.0838	**0.9604**	0.0269	**0.0231**	**0.9506**

resemble the ground truth, with RevFus achieving the most accurate balance between spatial detail and spectral fidelity, demonstrating the effectiveness of the proposed structural detail compensation in preserving fine structural features.

4.2.2. Full-resolution experiments

As reported in the last four columns of Table 1, among traditional model-based methods, PRACS and TV achieve relatively better performance, with QNR values exceeding 0.94. In contrast, GSA exhibits the highest D_s, indicating inaccurate extraction of spatial details, while MTF-GLP-HPM attains the lowest spectral distortion (D_λ^K). This can be attributed to the presence of small, densely clustered buildings in the QuickBird images, which complicates the extraction of spatial features.

Among deep learning-based methods, LDPNet, PGMAN, and ZSPan improve over traditional algorithms but still exhibit certain limitations in balancing spatial and spectral information. Specifically, LDPNet shows slightly higher spectral distortion in some regions, and PGMAN fails to achieve the lowest spatial distortion. ZSPan, benefiting from better spectral preservation, attains a relatively high HQNR. In contrast, our proposed RevFus consistently outperforms all competing methods, achieving the highest QNR, the lowest spatial distortion, and competitive spectral distortion. These improvements can be attributed to the degradation-to-fusion learning strategy, which first models the degradation process to achieve the fusion, as well as the introduction of alignment-based spatial detail reprojection, effectively improving spatial detail representation.

4.3. Results on Gaofen-2 dataset

Table 2 reports the quantitative results on the Gaofen-2 dataset, where the best performance is shown in bold and the second best are underlined. Moreover, Fig. 6 displays the visual results in true-color synthesis for full-resolution testing.

4.3.1. Reduced-resolution experiments

As shown in Table 2, traditional methods demonstrate competitive performance. Among them, TV achieves the best overall results, yielding the highest CC, mPSNR, and the lowest ERGAS, but this comes at the cost of high computational complexity, which limits its practicality in large-scale applications. PRACS follows closely with strong performance in all metrics, particularly achieving a low SAM. However, methods such as Z-PNN and PLRDiff suffer from large spectral distortions, reflected by significantly higher SAM and ERGAS values. Deep learning-based approaches like LDPNet and PGMAN outperform most traditional algorithms in certain metrics but still show limitations in maintaining a balance between spatial and spectral fidelity. By comparison, the proposed RevFus achieves consistently superior results across almost all metrics, attaining the highest mSSIM and the lowest SAM, while also delivering second-best CC, mPSNR, and ERGAS. More importantly, RevFus provides a more robust balance between spatial and spectral fidelity, which can be attributed to the accurate modeling of the degradation process in the first stage and the spatial–spectral contrastive learning in the second stage, jointly ensuring reliable and physically consistent fusion.

4.3.2. Full-resolution experiments

In the full-resolution evaluation, RevFus further demonstrates clear advantages. It achieves the highest QNR and HQNR, along with the lowest spectral distortion D_λ^K. Moreover, it maintains a competitive spatial distortion D_s, ranking second among all methods. Although PRACS achieves the lowest D_s, showing the advance of CS-based methods in spatial detail injection. Similarly, MTF-GLP-HPM yields the second-best D_λ^K but suffers from lower QNR. These results indicate that RevFus is more capable of jointly optimizing spatial and spectral fidelity. The superior performance of RevFus can be attributed to its invertible degradation modeling, which enforces physical consistency between fusion and degradation.

The visual results for full-resolution testing on the Gaofen-2 dataset further complement the quantitative assessment. Among traditional methods, AWLP appears visually superior with well-preserved textures, whereas ATWT-M3 suffers from noticeable blurring, and TV shows insufficient fine details. For deep learning-based methods, Z-PNN and TFResNet achieve a relatively good balance between spatial and spectral fidelity, while LDPNet exhibits loss of fine details. ZSPan and PGMAN produce somewhat blurred results, and PLRDiff introduces severe spectral distortions. In comparison, the proposed RevFus effectively preserves both spatial details and spectral fidelity, delivering fusion results that are visually more consistent with the ground truth. This demonstrates that the combination of degradation-to-fusion learning and spatial–spectral contrastive learning enables robust and physically consistent image fusion.

4.4. Results on WorldView-2 dataset

Table 3 reports the quantitative results on the WorldView-2 dataset, where the best performance is shown in bold and the second best are underlined. Moreover, Fig. 7 displays the visual results in true-color synthesis for reduced-resolution testing. Compared with the previous two sensors, which provide four-band multispectral data, WorldView-2 captures eight spectral bands, significantly increasing the fusion difficulty and resulting in noticeable decreases across all quantitative metrics.

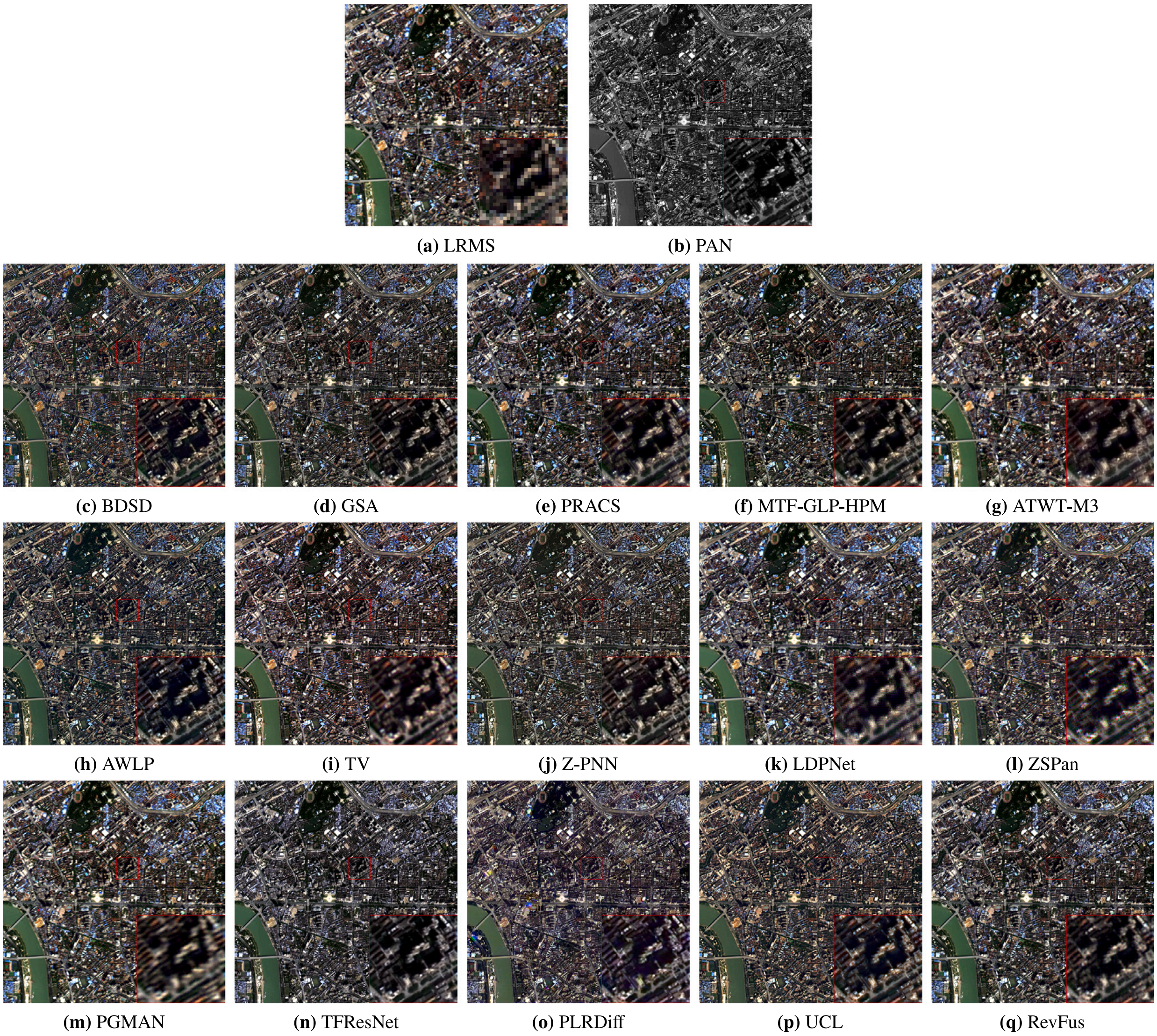


Fig. 6. Visual comparison of pan-sharpening methods on the Gaofen-2 dataset in the full-resolution testing scenario. The results are displayed in true-color, where bands 3, 2, and 1 correspond to the RGB channels.

Table 3
Quantitative assessment on the WorldView-2 data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution testing					Full-resolution testing			
	CC	mPSNR	mSSIM	SAM	ERGAS	QNR	D_s	D_λ^K	HQNR
BDSD	0.6851	21.4516	0.4801	8.7772	12.2564	0.9598	0.0400	0.3332	0.6402
GSA	0.7633	22.6819	0.5366	7.5127	10.6750	0.9077	0.0700	0.2640	0.6845
PRACS	0.7819	23.5347	0.5606	7.1914	9.6707	0.9429	0.0345	0.2449	0.7291
MTF-GLP-HPM	0.8095	24.0484	0.5949	**6.6441**	9.1836	0.9513	0.0309	0.1325	0.8407
ATWT-M3	0.8141	24.3457	0.5643	6.9154	8.8416	0.9071	0.0729	0.2826	0.6651
AWLP	0.8004	23.7039	0.5784	7.3005	9.4939	0.9467	0.0282	0.1418	0.8340
TV	0.8004	23.4319	0.5745	8.6758	9.7802	0.9614	0.0286	0.2450	0.7334
Z-PNN	0.7549	21.4874	0.5308	8.9331	11.3725	0.9197	0.0578	0.1921	0.7612
LDPNet	0.7940	23.8314	0.5362	9.8029	9.1734	0.9230	0.0313	0.2131	0.7622
ZSPan	0.8496	23.7649	0.6824	8.7633	9.2171	0.8362	0.1404	0.1401	0.7392
PGMAN	0.8016	23.9945	0.5621	7.3515	8.9129	0.8822	0.0954	0.2409	0.6866
TFResNet	0.8835	25.9612	0.7527	7.9309	7.4842	0.8679	0.0647	0.4644	0.5009
PLRDiff	0.7345	22.3739	0.5139	7.4734	10.5139	0.9033	0.0667	0.2554	0.6949
UCL	0.7743	22.8183	0.5620	6.9051	10.2032	0.9077	0.0657	0.1661	0.7791
RevFus	**0.9002**	**26.8413**	**0.7600**	6.7277	**6.7089**	**0.9711**	**0.0216**	**0.0958**	**0.8846**

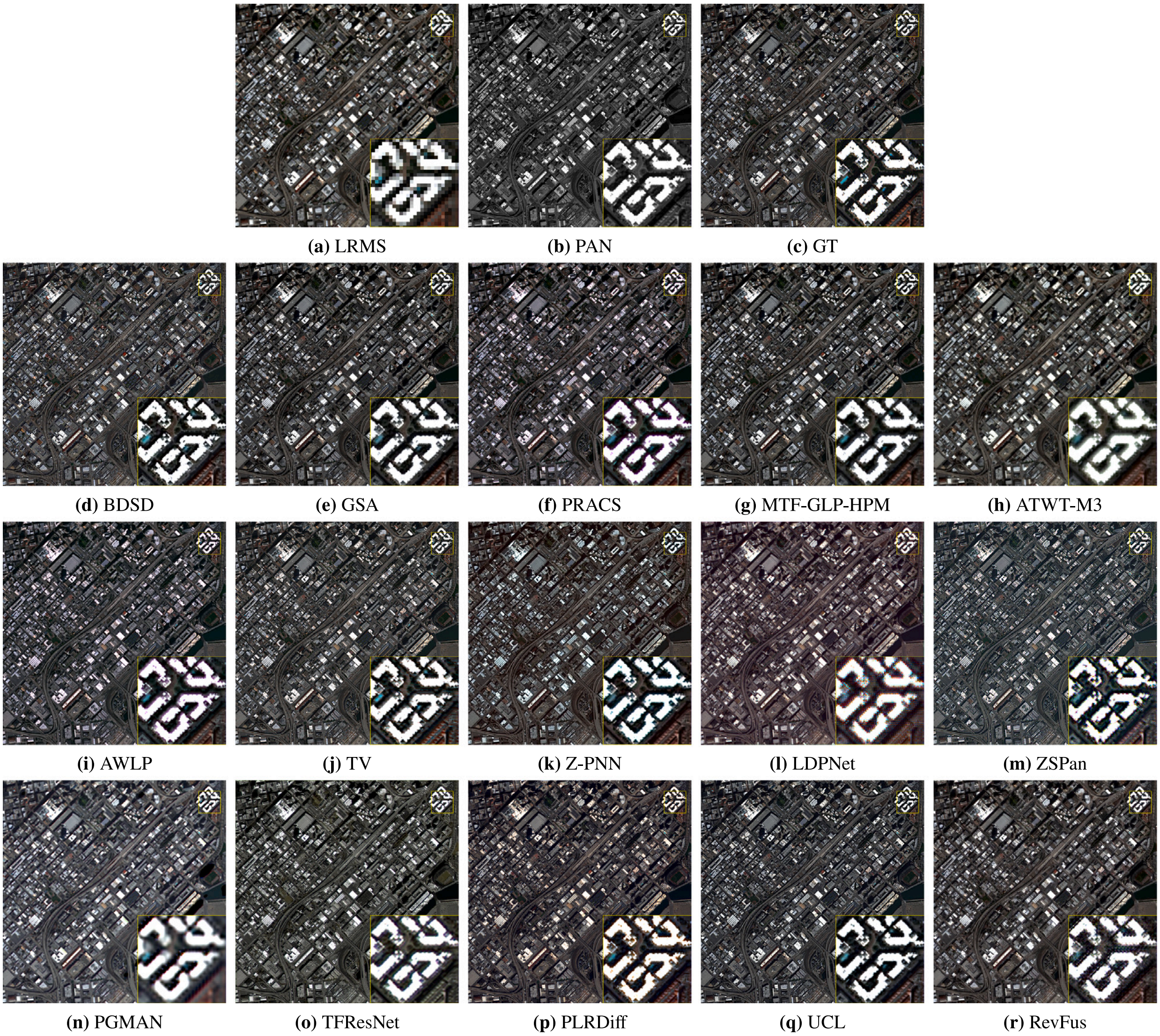

(a) LRMS (b) PAN (c) GT
(d) BDSD (e) GSA (f) PRACS (g) MTF-GLP-HPM (h) ATWT-M3
(i) AWLP (j) TV (k) Z-PNN (l) LDPNet (m) ZSPan
(n) PGMAN (o) TFResNet (p) PLRDiff (q) UCL (r) RevFus

Fig. 7. Visual comparison of pan-sharpening methods on the WorldView-2 dataset in the reduced-resolution testing scenario. The results are displayed in true-color, where bands 5, 3, and 2 correspond to the RGB channels.

4.4.1. Reduced-resolution experiments

Among traditional algorithms, MTF-GLP-HPM and ATWT-M3 achieve competitive results. PRACS also demonstrates robust performance with a relatively low SAM and ERGAS. In contrast, BDSD and GSA exhibit substantially lower correlation and higher spectral errors compared to the best traditional methods. Deep learning-based approaches generally improve performance over traditional algorithms, with TFResNet increasing correlation and mSSIM by approximately 10%–15% relative to the best traditional baseline. However, challenges remain for some learning-based models. ZSPan over-sharpens textures, leading to a marked rise in SAM, whereas LDPNet demonstrates only moderate spectral degradation. By comparison, RevFus consistently outperforms all competing methods, achieving the best CC, mPSNR, mSSIM, and ERGAS, enhancing correlation and mPSNR by roughly 5%–10% and reducing SAM by around 10% relative to the second-best approach. These results demonstrate the effectiveness of the proposed framework in maintaining a stable trade-off between spectral fidelity and spatial detail preservation.

The visual results for reduced-resolution testing on the WorldView-2 dataset further illustrate the quantitative findings. Among traditional methods, BDSD is strongly influenced by the PAN image, resulting in noticeable spectral distortions, while MTF-GLP-HPM appears visually optimal with well-preserved textures. ATWT-M3, however, introduces visible artifacts that slightly degrade image quality. For deep learning-based methods, Z-PNN and PLRDiff demonstrate competitive performance in both spatial detail preservation and spectral fidelity. Nevertheless, our proposed RevFus produces fusion results that are most consistent with the ground truth, achieving the highest alignment of spatial structures and spectral information.

4.4.2. Full-resolution experiments

In full-resolution evaluation, traditional methods often exhibit an imbalance between spatial and spectral preservation, with some achieving low spatial distortion at the expense of higher spectral errors, and vice versa. Deep learning-based models improve one aspect but still compromise the other, resulting in limited HQNR gains. RevFus, however, reduces both spatial and spectral distortions by roughly 15%–25% compared with the best competing methods, yielding a substantially higher overall quality index.

Table 4
Ablation study on WorldView-2 dataset. The inclusion of INN, HaarS, CCSL, SDC, and S^{2}CL modules is indicated by ✓. CC, PSNR, SSIM, SAM, and ERGAS are reported. The rows in between show the relative change (%) compared with the previous configuration.

Methods	INN	HaarS	CCSL	SDC	S^{2}CL	CC	PSNR	SSIM	SAM	ERGAS
Z-PNN	✗	✗	✗	✗	✗	0.7549	21.4874	0.5308	8.9331	11.3725
Percentage change	–					*+5.13%*	*+11.36%*	*+2.81%*	*−10.97%*	*−18.92%*
CSSL w/o HaarS	✓	✗	✗	✗	✗	0.7936	23.9283	0.5457	7.9533	9.2207
Percentage change	–					*+9.16%*	*+6.42%*	*+19.55%*	*+20.79%*	*−18.14%*
D2FL w/o CCSL	✓	✓	✗	✗	✗	0.8663	25.4644	0.6524	9.6071	7.5485
Percentage change	–					*+0.53%*	*+1.99%*	*+5.73%*	*−26.38%*	*−7.49%*
RevFus w/o SDC	✓	✓	✓	✗	✗	0.8709	25.9705	0.6898	7.0730	6.9831
Percentage change	–					*+3.28%*	*+3.16%*	*+9.70%*	*−1.55%*	*−2.06%*
RevFus w/o S^{2}CL	✓	✓	✓	✓	✗	0.8995	26.7922	0.7567	6.9637	6.8389
Percentage change	–					*+0.08%*	*+0.18%*	*+0.44%*	*−3.39%*	*−1.90%*
RevFus	✓	✓	✓	✓	✓	0.9002	26.8413	0.7600	6.7277	6.7089

Table 5
Computational efficiency comparison of the proposed RevFus among deep learning-based methods based on WorldView-2 dataset.

Method	Params (M)	FLOPs (G)	Training time (s)	Inference time (s)	Total time (s)
Z-PNN	0.081	20.105	20.196	0.024	20.220
LDPNet	0.131	23.138	1543.968	0.046	1544.014
ZSPan	0.079	40.158	256.635	0.008	256.642
PGMAN	3.932	73.742	91.324	0.014	91.338
TFResNet	2.366	111.319	15.575	0.013	15.588
PLRDiff	391.048	2868.580	–	221.199	221.199
UCL	3.452	76.633	40.111	174.332	214.443
RevFus	72.104	1908.9	183.901	0.273	184.174

4.5. Ablation study

To investigate the contribution of each key module in the proposed RevFus framework, we conducted an ablation study on the WorldView-2 dataset, as summarized in Table 4. The involved modules are INN, Haar wavelet sampler, the cycle-consistency self-learning, structural detail compensation, and spatial–spectral contrastive learning.

From the table, compared with Z-PNN, the INN-based model exhibits better spatial fidelity and spectral preservation. Since D2FL w/o HaarS corresponds to an INN-based model without Haar samplers or CCSL, the comparison with D2FL w/o CCSL highlights the effect of introducing Haar samplers. We observe that spatial fidelity is further improved, while SAM also increases, which can be attributed to the lack of strong spectral consistency constraints. Comparing D2FL w/o CCSL with RevFus w/o SDC, the inclusion of CCSL moderately improves the metrics, particularly reducing SAM by over 26%, highlighting its importance in enforcing cycle-consistency to stabilize unsupervised training and maintain spectral fidelity. Adding SDC (RevFus w/o S^{2}CL) further boosts performance, with mPSNR and mSSIM increasing by 3.2% and 9.7%, respectively, indicating that structural detail compensation effectively enhances spatial representation. Finally, incorporating S^{2}CL in the full RevFus model achieves the best results, yielding additional gains of 0.2–0.4% in mPSNR and mSSIM and reducing SAM by 3.4%, demonstrating that spatial–spectral contrastive learning effectively aligns high-quality spatial and spectral features and contributes to robust fusion of spatial details and spectral fidelity.

Overall, these results confirm that the combination of degradation-to-fusion learning, structural detail compensation, and the proposed learning strategies is essential for achieving superior pan-sharpening performance.

Table 6
Quantitative comparison of interpretability metrics across seven deep learning-based pan-sharpening models on the WorldView-2 dataset. R, S, F, and E denote the Mean Influence, Standard Deviation, Top-3 Focus, and Influence Entropy, respectively, while η represents the proposed IEC. Higher η values indicate stronger, more stable, and more focused interpretability performance. Superscript numbers indicate per-metric ranks (1 = best). Cells highlighted in red, green, and blue mark the top-1, top-2 and top-3 performances per metric, respectively.

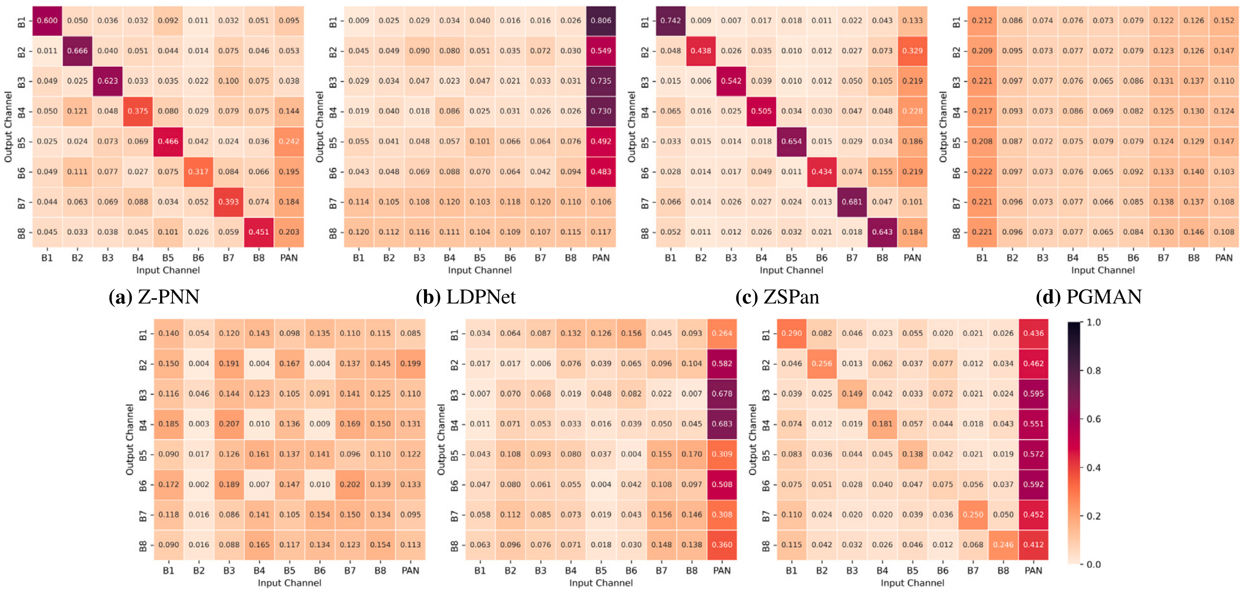
Method	R	S	F	E	η
Z-PNN	0.2062^{4}	0.2516^{5}	0.7302^{3}	0.7460^{4}	0.6123
LDPNet	0.0464^{7}	0.1399^{1}	0.6607^{5}	0.7153^{3}	0.2184
ZSPan	0.1793^{5}	0.2820^{7}	0.8542^{1}	0.6013^{1}	<u>0.7806</u>
PGMAN	0.4126^{2}	0.1770^{2}	0.4879^{6}	0.9668^{7}	0.7050
PLRDiff	0.5766^{1}	0.2751^{6}	0.4807^{7}	0.9224^{6}	0.6500
TFResNet	0.1693^{6}	0.2274^{4}	0.6964^{4}	0.7569^{5}	0.5057
RevFus	0.2126^{3}	0.1989^{3}	0.7931^{2}	0.6986^{2}	**0.9868**

4.6. Computational efficiency

Table 5 presents a comparison of the computational efficiency of RevFus with other deep learning algorithms on WorldView-2 dataset, measuring model parameters, Floating Point Operations (FLOPs), and runtime. It is important to note that PLRDiff is a diffusion-based model, which requires 1000 steps for inference. For fairness, the reported FLOPs and runtime are calculated based on the whole steps.

As shown in the table, due to the use of INN, the proposed RevFus has a relatively larger model size. However, compared with most self-supervised fusion methods, it achieves a shorter inference time, albeit at the cost of a longer training time. For frameworks that require iterative inference, such as PLRDiff and UCL, although the training time is relatively short, their inference time is significantly longer.

Overall, considering the total computational time, the efficiency of the proposed RevFus is acceptable, especially when taking its superior unsupervised performance into account. Meanwhile, computational efficiency continues to be a limitation of this work. Moving forward, a major research direction is to realize efficient unsupervised fusion framework through more lightweight INN modules.



(a) Z-PNN **(b)** LDPNet **(c)** ZSPan **(d)** PGMAN

(e) PLRDiff **(f)** TFResNet **(g)** RevFus

Fig. 8. Row-normalized SHAP matrices derived from seven deep learning models on the WorldView-2 dataset, visualized as heatmaps. The color intensity reflects the relative contribution of each spectral band to the model's output. The horizontal axis denotes the input spectral bands, while the vertical axis represents the output fused bands.

4.7. *Model interpretability evaluation*

4.7.1. *Quantitative comparison*

Table 6 compares interpretability across seven deep learning–based pan-sharpening models on the WorldView-2 dataset using four Shapley-derived metrics and the proposed IEC η, which unifies them into a single measure of interpretability efficiency.

RevFus achieves the highest η, ranking top-3 in all four metrics, indicating strong, stable, and focused explanatory patterns, reflecting a well-structured fusion mechanism with minimal uncertainty. ZSPan attains the second-highest η, benefiting from high focus F and low uncertainty E, though its higher variance S suggests less stable interpretability across outputs. PGMAN shows competitive η with high mean influence R and stability S, but low top-3 focus F indicates a more distributed contribution pattern. PLRDiff achieves the highest R, yet low F and high E reveal scattered, less interpretable contributions, while Z-PNN and TFResNet demonstrate moderate interpretability. LDPNet records the lowest η, with minimal influence R and poor focus F, reflecting weak and diffuse attribution despite low variance S.

In summary, models with balanced high influence, low uncertainty, and concentrated focus tend to achieve superior interpretability efficiency, such as RevFus and ZSPan. Moreover, the proposed η metric effectively captures these multidimensional traits, offering a unified and physically meaningful criterion for evaluating interpretability across deep pan-sharpening models.

4.7.2. *Contribution analysis*

To better understand the feature utilization patterns of different pan-sharpening models, we visualized row-normalized SHAP matrices, as shown in Fig. 8. Row normalization ensures that, for each output spectral band, the contributions of all input channels sum to one, allowing for a clear comparison of the relative importance of each input channel. This approach highlights which input channels are most influential in predicting each output band, making the models' decision logic more interpretable and facilitating task-specific analysis.

The heatmaps reveal distinct patterns that reflect both the models' internal strategies and the intrinsic characteristics of the pan-sharpening task. For Z-PNN, ZSPan, and RevFus, the contributions are highly concentrated: for each output band, the corresponding original spectral channel and the PAN band consistently exhibit the highest attribution. This observation aligns with the pan-sharpening principle, where the reconstruction of each output band should primarily rely on its original multispectral information complemented by high-resolution spatial details from PAN. Notably, RevFus places an even greater emphasis on the PAN band compared to Z-PNN and ZSPan, which likely explains its superior performance in spatial detail reconstruction and its higher interpretability.

In contrast, TFResNet and LDPNet predominantly rely on the PAN band across nearly all output bands, with weaker contributions from the original spectral channels, indicating an emphasis on spatial information while the utilization of spectral information remains further improvement. PGMAN exhibits a systematic bias in which input channels B1, B7, B8, and PAN contribute more strongly across all output bands, regardless of the target band, suggesting that the network architecture favors certain spectral features and thus reduces the alignment between input and output channels expected in pan-sharpening. PLRDiff displays scattered and less structured contributions, with no clear correspondence between output bands and specific input channels, implying a diffuse integration of information from all inputs and weaker interpretive clarity.

Overall, the heatmaps reveal that models such as RevFus, Z-PNN, and ZSPan effectively combine the corresponding spectral band with PAN information, producing clear and focused interpretability. These visual observations complement the quantitative η metric, providing task-specific insight into how each model balances spectral fidelity and spatial detail in pan-sharpening.

5. Conclusions

This work presents RevFus, a novel unsupervised pan-sharpening framework that effectively integrates physically grounded modeling with advanced learning strategies. By explicitly modeling the degradation process through an invertible neural network and enforcing cycle-consistency, the proposed method ensures reliable and interpretable fusion. The introduction of structural detail compensation and spatial–spectral contrastive learning further strengthens spatial detail preservation and spectral fidelity. Moreover, combining four SHAP-derived metrics, IEC is proposed to evaluate the model interpretability comprehensively. Extensive reduced-resolution and full-resolution experiments on QuickBird, Gaofen-2, and WorldView-2 datasets demonstrate that RevFus consistently surpasses state-of-the-art unsupervised and traditional methods with high model interpretability. The improvements highlight the importance of combining degradation-to-fusion learning with targeted constraints for robust and high-quality pansharpening. These findings suggest that frameworks motivated by physical processes are effective in addressing the limitations of unsupervised approaches in complex scenarios. Future work will investigate the extension of this framework to multi-temporal and hyperspectral data, aiming to further enhance generalization and applicability in diverse remote sensing tasks.

CRediT authorship contribution statement

Jiang He: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xiao Xiang Zhu:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the German Research Foundation (DFG GZ: ZH 498/18-1; Project number: 519016653) and by the Munich Center for Machine Learning, Germany.

Data availability

Data will be made available on request.

References

Aiazzi, B., Alparone, L., Baronti, S., Carlà, R., Garzelli, A., Santurri, L., 2014. Full-scale assessment of pansharpening methods and data products. In: Image Signal Process. Remote Sens. XX. vol. 9244, SPIE, 924402.

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M., 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. Photogramm. Eng. Remote Sens. 72 (5), 591–596.

Aiazzi, B., Baronti, S., Selva, M., 2007. Improving component substitution pansharpening through multivariate regression of MS + Pan data. IEEE Trans. Geosci. Remote Sens. 45 (10), 3230–3239.

Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M., 2008. Multispectral and panchromatic data fusion assessment without reference. Photogramm. Eng. Remote Sens. 74 (2), 193–200.

Augasta, M.G., Kathirvalavakumar, T., 2012. Reverse engineering the neural networks for rule extraction in classification problems. Neural Process. Lett. 35 (2), 131–150.

Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B., 2006. A variational model for P+XS image fusion. Int. J. Comput. Vis. 69 (1), 43–58.

Barnes, E.A., Barnes, R.J., Martin, Z.K., Rader, J.K., 2022. This looks like that there: Interpretable neural networks for image tasks when location matters. Artif. Intell. Earth Syst. 1 (3), e220001.

Burt, P.J., Adelson, E.H., 1987. The Laplacian pyramid as a compact image code. In: Readings in Computer Vision. Elsevier, pp. 671–679.

Cao, Q., Deng, L.-J., Wang, W., Hou, J., Vivone, G., 2024. Zero-shot semi-supervised learning for pansharpening. Inf. Fusion 101, 102001.

Carper, W., Lillesand, T., Kiefer, R., 1990. The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. 56 (4), 459–467.

Chandler, D., 1987. Introduction to Modern Statistical Mechanics. Oxford University Press.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proc. IEEECVF Conf. Comput. Vis. Pattern Recognit.. pp. 15750–15758.

Chen, T., Kornblith, S., et al., 2020. A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn..

Chen, K., Wang, Y., Huang, C., Wang, J., Li, S.L., Guan, H., Ma, L., 2025. GreenNet: A dual-encoder network for urban green space classification using high-resolution remotely sensed images. Int. J. Appl. Earth Obs. Geoinf. 142, 104709.

Cheng, S., Cheng, L., Qin, S., Zhang, L., Liu, P., Liu, L., Xu, Z., Wang, Q., 2022. Improved understanding of how catchment properties control hydrological partitioning through machine learning. Water Resour. Res. 58 (4), e2021WR031412.

Choi, J., Yu, K., Kim, Y., 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE Trans. Geosci. Remote Sens. 49 (1), 295–309.

Ciotola, M., Vitale, S., Mazza, A., Poggi, G., Scarpa, G., 2022a. Pansharpening by convolutional neural networks in the full resolution framework. IEEE Trans. Geosci. Remote Sens..

Ciotola, M., Vitale, S., Mazza, A., Poggi, G., Scarpa, G., 2022b. Pansharpening by convolutional neural networks in the full resolution framework. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Dantas, C.F., Drumond, T.F., Marcos, D., Ienco, D., 2023. Counterfactual explanations for remote sensing time series data: An application to land cover classification. In: Jt. Eur. Conf. Mach. Learn. Knowl. Discov. Databases. Springer, pp. 20–36.

Deng, L.-J., Vivone, G., Paoletti, M.E., Plaza, A., 2022. Machine learning in pansharpening: A benchmark, from shallow to deep networks. IEEE Geosci. Remote. Sens. Mag. 10 (3), 279–315. http://dx.doi.org/10.1109/MGRS.2022.3187652.

Do, M.N., Vetterli, M., 2005. The contourlet transform: An efficient directional multiresolution image representation. IEEE Trans. Image Process. 14 (12), 2091–2106.

Eugenio, F., Martin, J., Marcello, J., Fraile-Nuez, E., 2014. Environmental monitoring of El Hierro Island submarine volcano, by combining low and high resolution satellite imagery. Int. J. Appl. Earth Obs. Geoinf. 29, 53–66.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20 (177), 1–81.

Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles.

Fu, X., Lin, Z., Huang, Y., Ding, X., 2019. A variational pan-sharpening with local gradient constraints. In: Proc. IEEECVF Conf. Comput. Vis. Pattern Recognit.. pp. 10265–10274.

Garzelli, A., Nencini, F., Capobianco, L., 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. IEEE Trans. Geosci. Remote Sens. 46 (1), 228–236.

Gastineau, A., Aujol, J.-F., Berthoumieu, Y., Germain, C., 2021. Generative adversarial network for pansharpening with spectral and spatial discriminators. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Ghorbani, A., Abid, A., Zou, J., 2019. Interpretation of neural networks is fragile. In: Proc. AAAI Conf. Artif. Intell.. vol. 33, pp. 3681–3688.

Gillespie, A.R., Kahle, A.B., Walker, R.E., 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. Remote Sens. Environ. 22 (3), 343–365.

González-Audícana, M., Saleta, J.L., Catalán, R.G., García, R., 2004. Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Trans. Geosci. Remote Sens. 42 (6), 1291–1299.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

Guo, X., Hou, B., Yang, C., Ma, S., Ren, B., Wang, S., Jiao, L., 2023. Visual explanations with detailed spatial information for remote sensing image classification via channel saliency. Int. J. Appl. Earth Obs. Geoinf. 118, 103244.

He, K., Fan, H., et al., 2020. Momentum contrast for unsupervised visual representation learning. In: Proc. IEEECVF Conf. Comput. Vis. Pattern Recognit..

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144.

He, J., Yuan, Q., Li, J., Zhang, L., 2022. A knowledge optimization-driven network with normalizer-free group ResNet prior for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Höhl, A., Obadic, I., Fernández-Torres, M.-Á., Najjar, H., Oliveira, D.A.B., Akata, Z., Dengel, A., Zhu, X.X., 2024. Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing. IEEE Geosci. Remote. Sens. Mag. 12 (4), 261–304. http://dx.doi.org/10.1109/MGRS.2024.3467001.

Ishikawa, S.-n., Todo, M., Taki, M., Uchiyama, Y., Matsunaga, K., Lin, P., Ogihara, T., Yasui, M., 2023. Example-based explainable AI and its application for remote sensing image classification. Int. J. Appl. Earth Obs. Geoinf. 118, 103215.

Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F., 2020. A survey on contrastive self-supervised learning. Technologies 9 (1), 2.

Javan, F.D., Samadzadegan, F., Mehravar, S., Toosi, A., Khatami, R., Stein, A., 2021. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. ISPRS J. Photogramm. Remote Sens. 171, 101–117.

Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D., 2020. Hard negative mixing for contrastive learning. Adv. Neural Inf. Process. Syst. 33, 21798–21809.

Kay, S.M., 1993. Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice-Hall, Inc..

Khan, M., Hanan, A., Kenzhebay, M., Gazzea, M., Arghandeh, R., 2024. Transformer-based land use and land cover classification with explainability using satellite imagery. Sci. Rep. 14 (1), 16744.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Adv. Neural Inf. Process. Syst. 33, 18661–18673.

King, R., Wang, J., 2001. A wavelet based algorithm for pan sharpening landsat 7 imagery. In: Proc. IEEE Int. Geosci. Remote Sens. Symp.. IGARSS, vol. 2, pp. 849–851, vol.2.

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., Goetz, A.F.H., 1993. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44 (2–3), 145–163.

Kwarteng, P., Chavez, A., 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. Photogramm. Eng. Remote Sens. 55 (1), 339–348.

Laben, C.A., Brower, B.V., 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening.

Li, J., Sun, W., Jiang, M., Yuan, Q., 2021. Self-supervised pansharpening based on a cycle-consistent generative adversarial network. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Liu, J., Feng, Y., Zhou, C., Zhang, C., 2020a. Pwnet: An adaptive weight network for the fusion of panchromatic and multispectral images. Remote. Sens. 12 (17), 2804.

Liu, X., Liu, Q., Wang, Y., 2020c. Remote sensing image fusion based on two-stream fusion network. Inf. Fusion 55, 1–15.

Liu, Q., Meng, X., Shao, F., Li, S., 2023. Supervised-unsupervised combined deep convolutional neural networks for high-fidelity pansharpening. Inf. Fusion 89, 292–304.

Liu, Z., Zheng, K., Song, Y., Zhang, J., 2025. Daily high-resolution PM2. 5 mapping using spatiotemporal CNN-transformer-KAN model. Int. J. Appl. Earth Obs. Geoinf. 144, 104900.

Liu, Q., Zhou, H., Xu, Q., Liu, X., Wang, Y., 2020b. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 59 (12), 10227–10242.

Luo, S., Zhou, S., Feng, Y., Xie, J., 2020. Pansharpening via unsupervised convolutional neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 4295–4310.

Ma, J., Yu, W., Chen, C., Liang, P., Guo, X., Jiang, J., 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. Inf. Fusion 62, 110–120.

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by convolutional neural networks. Remote. Sens. 8 (7), 594.

Nason, G.P., Silverman, B.W., 1995. The stationary wavelet transform and some statistical applications. In: Wavelets and Statistics. Springer, pp. 281–299.

Ni, J., Shao, Z., Zhang, Z., Hou, M., Zhou, J., Fang, L., Zhang, Y., 2022. LDP-Net: An unsupervised pansharpening network based on learnable degradation processes. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 5468–5479.

Otazu, X., González-Audícana, M., Fors, O., Núñez, J., 2005. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods. IEEE Trans. Geosci. Remote Sens. 43 (10), 2376–2385.

Ozcelik, F., Alganci, U., Sertel, E., Unal, G., 2020. Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs. IEEE Trans. Geosci. Remote Sens. 59 (4), 3486–3501.

Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., 2013. A new pansharpening algorithm based on total variation. IEEE Geosci. Remote. Sens. Lett. 11 (1), 318–322.

Qu, Y., Baghbaderani, R.K., Qi, H., Kwan, C., 2020. Unsupervised pansharpening based on self-attention mechanism. IEEE Trans. Geosci. Remote Sens. 59 (4), 3192–3208.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn.. PmLR, pp. 8748–8763.

Ranchin, T., Wald, L., 2000. Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. Photogramm. Eng. Remote Sens. 66 (1), 49–61.

Rao, Y., He, L., Zhu, J., 2017. A residual convolutional neural network for pan-shaprening. In: 2017 Int. Workshop Remote Sens. Intell. Process.. RSIP, IEEE, pp. 1–4.

Rasmussen, C.E., Ghahramani, Z., 2001. Occam's razor. In: Adv. Neural Inf. Process. Syst.. vol. 13, pp. 294–300.

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

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 618–626.

Seo, S., Choi, J.-S., Lee, J., Kim, H.-H., Seo, D., Jeong, J., Kim, M., 2020. UPSNet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images. IEEE Access 8, 201199–201217.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 11 (5), 1656–1669.

Starck, J.-L., Candès, E.J., Donoho, D.L., 2002. The curvelet transform for image denoising. IEEE Trans. Image Process. 11 (6), 670–684.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: Int. Conf. Mach. Learn.. PMLR, pp. 3319–3328.

Temenos, A., Temenos, N., Kaselimi, M., Doulamis, A., Doulamis, N., 2023. Interpretable deep learning framework for land use and land cover classification in remote sensing using SHAP. IEEE Geosci. Remote. Sens. Lett. 20, 1–5.

Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding. In: Comput. Vision–ECCV 2020 16th Eur. Conf. Glasg. UK August 23–28 2020 Proc. Part XI 16. Springer, pp. 776–794.

Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M., 2019. On mutual information maximization for representation learning. arXiv prepr. arXiv:190713625. arXiv:1907.13625.

Uezato, T., Hong, D., Yokoya, N., He, W., 2020. Guided deep decoder: Unsupervised image pair fusion. In: 16th Eur. Conf. Comput. Vis.. ECCV 2020, vol. 12351, Springer, pp. 87–102.

Vivone, G., Dalla Mura, M., Garzelli, A., Restaino, R., Scarpa, G., Ulfarsson, M.O., Alparone, L., Chanussot, J., 2020. A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. IEEE Geosci. Remote. Sens. Mag. 9 (1), 53–81.

Vivone, G., Simões, M., Dalla Mura, M., Restaino, R., Bioucas-Dias, J.M., Licciardi, G.A., Chanussot, J., 2014. Pansharpening based on semiblind deconvolution. IEEE Trans. Geosci. Remote Sens. 53 (4), 1997–2010.

Wald, L., Ranchin, T., Mangolini, M., 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogramm. Eng. Remote Sens. 63 (6), 691–699.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, D., Zhang, P., Bai, Y., Li, Y., 2022. MetaPan: Unsupervised adaptation with meta-learning for multispectral pansharpening. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Wei, Y., Yuan, Q., Shen, H., Zhang, L., 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci. Remote. Sens. Lett. 14 (10), 1795–1799.

Wu, X., Cao, Z.-H., Huang, T.-Z., Deng, L.-J., Chanussot, J., Vivone, G., 2025. Fully-connected transformer for multi-source image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 47 (3), 2071–2088.

Wu, Z.-C., Huang, T.-Z., Deng, L.-J., Vivone, G., 2023. A framelet sparse reconstruction method for pansharpening with guaranteed convergence. Inverse Probl. Imaging 17 (6), 1277–1300.

Xia, J., Yokoya, N., Adriano, B., Kanemoto, K., 2023. National high-resolution cropland classification of Japan with agricultural census information and multi-temporal multi-modality datasets. Int. J. Appl. Earth Obs. Geoinf. 117, 103193.

Xiao, J.-L., Huang, T.-Z., Deng, L.-J., Jiang, H., Zhao, Q., Vivone, G., 2026. Unsupervised coefficient learning framework for variational pansharpening. Inf. Fusion (ISSN: 1566-2535) 127, 103790.

Xu, Q., Li, Y., Nie, J., Liu, Q., Guo, M., 2023. UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained Generative Adversarial Network. Inf. Fusion 91, 31–46.

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J., 2017. PanNet: A deep network architecture for pan-sharpening. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 5449–5457.

Yang, H., Jiang, Z., Zhang, Y., Wu, Y., Luo, H., Zhang, P., Wang, B., 2025. A high-resolution remote sensing land use/land cover classification method based on multi-level features adaptation of segment anything model. Int. J. Appl. Earth Obs. Geoinf. 141, 104659.

Zeng, D., Hu, Y., Huang, Y., Xu, Z., Ding, X., 2016. Pan-sharpening with structural consistency and l1/2 gradient prior. Remote. Sens. Lett. 7 (12), 1170–1179.

Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019. Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans. Geosci. Remote Sens. 57 (8), 5549–5563.

Zhang, H., Ma, J., 2021. GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening. ISPRS J. Photogramm. Remote Sens. 172, 223–239.

Zhong, Y., Wu, X., Cao, Z., Dou, H.-X., Deng, L.-J., 2024. SSDiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. In: Annu. Conf. Neural Inf. Process. Syst.. NeurIPS.

Zhong, J., Yang, B., Huang, G., Zhong, F., Chen, Z., 2016. Remote sensing image fusion with convolutional neural network. Sens. Imaging 17 (1), 1–16.

Zhou, H., Liu, Q., Wang, Y., 2021. Pgman: An unsupervised generative multiadversarial network for pansharpening. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 6316–6327.

Zhou, H., Liu, Q., Weng, D., Wang, Y., 2022. Unsupervised cycle-consistent generative adversarial networks for Pan sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Zhou, C., Zhang, J., Liu, J., Zhang, C., Fei, R., Xu, S., 2020. PercepPan: Towards unsupervised pan-sharpening based on perceptual loss. Remote. Sens. 12 (14), 2318.