Laboring on less labors: RPCA Paradigm for Pan-sharpening

Honghui Xu[1,2], Chuangjie Fang[1], Yibin Wang[3], Jie Wu[1], Jianwei Zheng[1*]

[1]Zhejiang University of Technology [2]Taizhou University [3]Fudan University

xhh9609@gmail.com, fangcj@zjut.edu.cn, yibinwang1121@163.com, {wuj,zjw}@zjut.edu.cn

Abstract

*Deep unfolding network (DUN) based pansharpening has shed new light on high-resolution/spectrum image acquisition, serving as a computational alternative to physical devices. While with both merits of deep feature learning and acceptable interpretability enjoyed, current pansharpening necessitates substantial effort in approximating the degradation matrices along the spatial and spectral dimensions, yet with performance hardly guaranteed within the complex scenarios. Moreover, as a key step during DUN update, current solutions rely solely on black-box networks to learn the data-driven priors, which further results in laborious architecture crafting and compromised interpretability. To counteract the dilemmas, we propose a new solution, namely **R**PCA-based **U**nfolding **N**etwork (RUN), which shrinks the original two degradations to only one. Specifically, grounded in the significant sparsity of spatial offset components, i.e., the difference between upsampled image and the desired target, we shift the original pansharpening issue into a novel Robust Principal Component Analysis (RPCA)-based paradigm. On that basis, the tricky approximation to the spatial degradation matrix as well as its transposed counterpart is naturally avoided. Specific for the prior learning step of RPCA unfolding, an efficient Nonlinear transformation-based Tensor Nuclear Norm (NTNN) is meticulously engineered, in which the computationally intensive Singular Value Decomposition is avoided with the aid of depthwise convolutions. More importantly, NTNN plays a plug-and-play role and can be easily embedded into Transformer/CNN architectures for the learning of both global and local features. Experimental results on multiple remote datasets demonstrate the superiority of the proposal over previous SOTA methods.*

1. Introduction

Due to the rapidly evolving remote sensing (RS) technology, a multitude of satellites outfitted with optical sensing apparatus, including QuickBird, GaoFen-2 (GF2), WorldView-2 (WV2), and WorldView-3 (WV3), have been deployed, facilitating a wide array of application scenarios. Nevertheless, optical imagery often encounters the challenge of balancing spatial and spectral resolution-a dilemma rooted in the inherent limitations of imaging sensors. This fundamental compromise restricts the ability to simultaneously enrich spatial and spectral resolutions in image acquisition [12, 18, 23]. Practically, the spatial size of panchromatic (PAN) images often surpasses that of low-resolution multispectral images (LRMS). In the meantime, an LRMS image is rich in spectral information, in stark contrast to PANs that involve only a single channel. Against this backdrop, **Pansharpening** emerges as an alternative solution to expensive hardware devices, which hammers at synthesizing a high spatial resolution multispectral (HRMS) image given a PAN/LRMS pair, paving the way for imagery that captures the best of both domains.

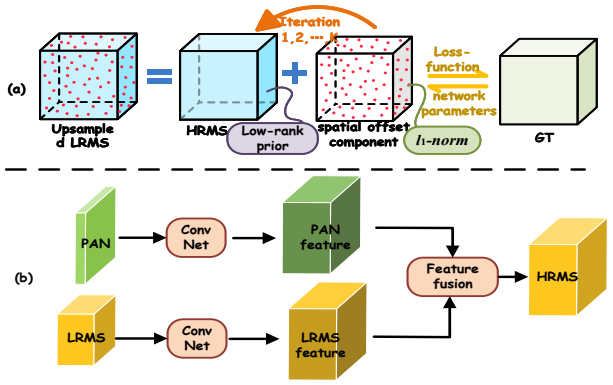


Figure 1. The proposed RPCA Paradigm for pansharpening (a) *versus* the traditional pansharpening framework (b).

Contemporary pansharpening methodologies can be systematically classified into two principal groups: model-based schemes and deep learning (DL)-based techniques. Model-based schemes [2, 24, 29, 31] typically enjoy a high population due to solid theoretical foundation. Technically, this branch often formulates a model incorporating two key elements, *i.e.*, spatial-spectral fidelity and prior property ex-

*Corresponding author.

tractor, which are then refined through a specific optimization algorithm. Recent studies have successfully employed massive prior assumptions, including spectral correlation [40], local spatial smoothness [35], and global low rankness [26, 28], to earn performance improvements. However, although underpinned by solid theoretical support and coherent physical interpretation, peak performance is limited due to the singleton usage of a PAN/LRMS pair during model update. In addition, the necessary setting of several manual hyperparameters incurs a labor-intensive model selection, leaving a notable hurdle in real-world deployment.

In a data-driven and end-to-end fashion, DL-based solutions directly establish the mapping relationship for LRMS-PAN fusion, following the general framework illustrated in the lower segment of Fig. 1. Convolution neural networks (CNNs) [3–5, 8, 19, 36], as the cornerstone of DL techniques, enjoy a high-performance guarantee owing to the endowed ability in local feature learning. Transformer, based on the self-attention mechanism [6, 20, 37], hammers at capturing long-range interdependencies and is recently ushering in a new wave of advancements. Generally, the attractive performance of DL lies mainly in the strong capability to explore implicit features. However, the black-box pattern also leaves poor interpretability. In practice, how to achieve a well balance between accuracy and interpretability presents a non-trivial problem.

More recently, massive DUNs emerged to bridge the gap between model-based and learning-based methods, with the naive expectation of both high performance and high interpretability. The core idea of DUN lies in unfolding the iterations of an inference algorithm into an end-to-end and data-driven pattern. Technically, the DUN architecture decouples the original problem into multiple sub-problems, allowing for the individual introduction of handcrafted priors and data-driven knowledge, *e.g.*, CNN/Transformer-based priors. Generally, this treatment not only retains solid support from optimization theory, but also achieves superior performance. However, specific to pansharpening, there are two pressing issues within the current DUN framework. 1) The degradation matrices along the spatial and spectral dimensions, as well as their transposed counterparts, shall be adaptively learned during the iterative update. Accordingly, the individual design of four network architectures results in an extremely laborious investigation. This issue is particularly pronounced in the context of aerial remote sensing and satellite imagery, where images often present large scenes with massive objects of various scales. 2) For the key step of prior extraction, the typically handcrafted treatment often suffers from heavy computation overhead, *e.g.*, from SVD operation. In contrast, the pure employment of CNN/Transformer or their combination would inevitably drag the unfolding back to the black-box pattern.

To remedy the current issues, we propose in this work a new pansharpening paradigm with two innovations, *i.e.*, degradation shrinkage and joint explicit-implicit priors. On the one hand, following a deep contrast between the updating flow of target variable and the sparsification of spatial offset components, we conjecture that the original pansharpening issue can be naturally transformed into a RPCA-based paradigm. Along this line, the original approximation to two degradation matrices can be shrunk to one only. More importantly, the more complex matrix along the spatial dimension is removed, leaving a simpler degradation with a much smaller size. On the other hand, specific to the low-rank constraint in RPCA model, most hand-crafted treatments struggle to match the complexity of remote sensing imagery, while data-driven alternatives require large-scale samples to ensure acceptable performance. Therefore, we propose a Nonlinear transformation-based Tensor Nuclear Norm (NTNN), which further possesses two distinct features. First, two dual transformations are employed to substitute the traditional Fourier/adaptive counterparts, in which an explicit threshold-solving operator is also considered based on the vanilla convolution layer, capturing the low-rank prior with reduced computational overhead. Second, NTNN can be embedded into current Transformer/CNN modules in a plug-and-play fashion, thus capturing holistic features with little effort additionally made. The final RPCA model is unfolded in an end-to-end manner through an alternating direction method of multipliers (ADMM) framework, enjoying both solid mathematical properties and high-performance guarantee. To sum up, the major contributions are summarized as follows.

1. With strong sparsity of spatial offset component tapped, we omit the approximation of spatial degradation, leaving the adaptation only to spectral response with a much simpler architecture. On that basis, a RPCA paradigm is innovated for pansharpening, which can be easily solved by most off-the-shelf first-order optimizers.
2. Specific to the low-rank constraint in RPCA, a dual Nonlinear transformation-based Tensor Nuclear Norm is crafted, in which the explicit Singular Value Thresholding (SVT) operation is upgraded as a nonlinear CNN-based treatment, enjoying jointly a lightweight computation and a solid mathematical foundation.
3. NTNN can be readily embedded in most CNN and Transformer modules, leading to enriched features with both explicit and implicit priors. With these components equipped within an ADMM solver, the final RPCA unfolding solution for pansharpening is engineered, which requires less labor in degradation simulations and enjoys the more holistic feature information.

2. Methodology

In this section, we initially provide a concise overview of typical pansharpening architecture, followed by a compre-

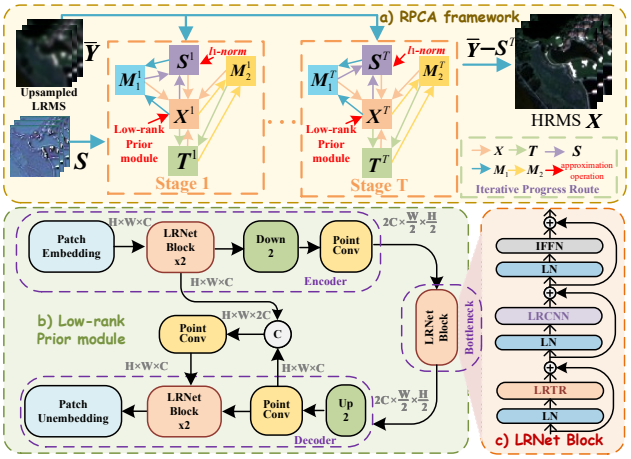


Figure 2. Illustration of the proposed RUN. Each stage lies a RPCA solver equipped with a prior extractor LRNet.

hensive exposition of the evolutionary path leading to the RPCA Paradigm and the concerned solution. The overall RUN network is given in Fig. 2.

2.1. Original Problem Formulation

Given two images from the same scene, *i.e.*, LRMS $\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$ and PAN $\mathcal{Z} \in \mathbb{R}^{H\times W\times 1}$, pansharpening aims to estimate HRMS $\mathcal{X} \in \mathbb{R}^{H\times W\times S}$ with high spatial size and deep spectral channels. Here, W/w ($W > w$), H/h ($H > h$), and S denote width, height, and band size, respectively. In general, the explicit relationship between LRMS, PAN, and HRMS can be expressed as follows [11].

$$\boldsymbol{Y}_{(3)} = \boldsymbol{X}_{(3)}\boldsymbol{B} + \boldsymbol{E}_{\boldsymbol{Y}}, \boldsymbol{Z}_{(3)} = \boldsymbol{R}\boldsymbol{X}_{(3)} + \boldsymbol{E}_{\boldsymbol{Z}} \quad (1)$$

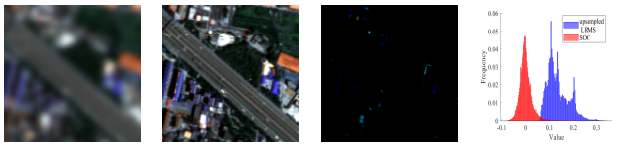
where $\boldsymbol{X}_{(3)} \in \mathbb{R}^{S\times HW}$, $\boldsymbol{Y}_{(3)} \in \mathbb{R}^{S\times hw}$, and $\boldsymbol{Z}_{(3)} \in \mathbb{R}^{1\times HW}$ are the mode-3 matricization of HRMS, LRMS, and PAN, respectively. $\boldsymbol{B} \in \mathbb{R}^{HW\times hw}$ and $\boldsymbol{R} \in \mathbb{R}^{1\times S}$ sequentially denote the spatial and spectral degradation matrices. $\boldsymbol{E}_{\boldsymbol{Y}}$ and $\boldsymbol{E}_{\boldsymbol{Z}}$ are the noise in the observations. To estimate $\boldsymbol{X}_{(3)}$, certain constraints are required to characterize the intrinsic properties of target images. Hence, pansharpening can be formulated with the energy function as

$$\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{R}\boldsymbol{X}\|_F^2 + \lambda J_1(\boldsymbol{X}) \quad (2)$$

where λ is the penalty parameter and J_1 is the regularizer. To optimize (2), it is known that laborious efforts are needed to adapt the spatial/spectral degradation matrices ($\boldsymbol{B}$ and $\boldsymbol{R}$) and their transposition counterparts ($\boldsymbol{B}^T$ and $\boldsymbol{R}^T$).

2.2. RPCA Paradigm for Pansharpening

Note the adaptation of massive matrices would not only incur laborious work, but also suffer from the risk of performance drop due to the suboptimal simulations. To alleviate this issue, we attempt to remove one of the degradations



(a) ULRMS (b) HRHS (c) SOC (d) Histograms

Figure 3. Sparse SOCs and the concerned histograms.

with the RPCA paradigm, given in the upper part of Fig. 2. As can be easily observed, the size of spatial matrix $\boldsymbol{B}$ is much larger than that of spectral matrix $\boldsymbol{R}$. Therefore, it is preferable to remove $\boldsymbol{B}$, leaving $\boldsymbol{R}$ with limited elements to be simulated. Upon investigation of upsampled LRMS $\bar{\boldsymbol{Y}}$, which plays the initial step in most pansharpening solutions, an intriguing property can be easily dug. That is, the disparity between $\bar{\boldsymbol{Y}}$ and $\boldsymbol{X}$ is significantly sparse regardless of variations in the dataset, as shown in Fig. 3. That is to say, the spatial offset component (SOC) $\boldsymbol{S} = \bar{\boldsymbol{Y}} - \boldsymbol{X}$ would remove a significant portion of low- and mid-frequency information, retaining only high-frequency details such as edges. The histogram in Fig. 3(d) corroborates this assumption. As shown, the search for a significantly sparse matrix $\boldsymbol{S}$ by removing the massive low-frequency components fits perfectly the flow of updating $\boldsymbol{X}$ to be the ground-truth HRMS. On that basis, we impose the sparsity penalty l_1 norm SOC $\boldsymbol{S}$, rather than using meticulously designed modules to approximate the spatial degradations $\boldsymbol{B}$ and $\boldsymbol{B}^T$. Eq. (2) can then be reformulated as follows.

$$\min_{\boldsymbol{X},\boldsymbol{S}} \alpha\|\boldsymbol{S}\|_1 + J_1(\boldsymbol{X}) \text{ s.t. } \bar{\boldsymbol{Y}} = \boldsymbol{X} + \boldsymbol{S} \quad (3)$$

where α is the penalty parameter. Typically, low rankness is the most frequently used J_1 regularizer, such as tensor nuclear norm (TNN), hence Eq. (3) fits perfectly to the well-known RPCA model. Note the size of $\boldsymbol{R}$ is much smaller than $\boldsymbol{B}$. That is, in stark contrast to the spatial matrix with complex textures and structures, the simulation of spectral degradation is surely easier. Hence, we introduce a spectral regularization term based on Eq. (3), thus constructing the RPCA paradigm for pansharpening.

$$\min_{\boldsymbol{X},\boldsymbol{S}} \alpha\|\boldsymbol{S}\|_1 + J_1(\boldsymbol{X}) + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{R}\boldsymbol{X}\|_F^2 \text{ s.t. } \bar{\boldsymbol{Y}} = \boldsymbol{X} + \boldsymbol{S} \quad (4)$$

Most first-order optimization algorithms are readily available to solve model (4). We simply employ typical ADMM scheme, yet put more effort in network architecture during unfolding stages. With an auxiliary variable $\boldsymbol{T} = \boldsymbol{X}$, Eq. (4) can be reformed into an unconstrained one.

$$\begin{aligned} L_{\mu_i}(\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}_1, \boldsymbol{M}_2) &= \alpha\|\boldsymbol{S}\|_1 + J_1(\boldsymbol{T}) \\ &+ \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{R}\boldsymbol{X}\|_F^2 + \frac{\mu_1}{2}\|\bar{\boldsymbol{Y}} - (\boldsymbol{X} + \boldsymbol{S}) + \frac{\boldsymbol{M}_1}{\mu_1}\|_F^2 \\ &+ \frac{\mu_2}{2}\|\boldsymbol{T} - \boldsymbol{X} + \frac{\boldsymbol{M}_2}{\mu_2}\|_F^2 \end{aligned} \quad (5)$$

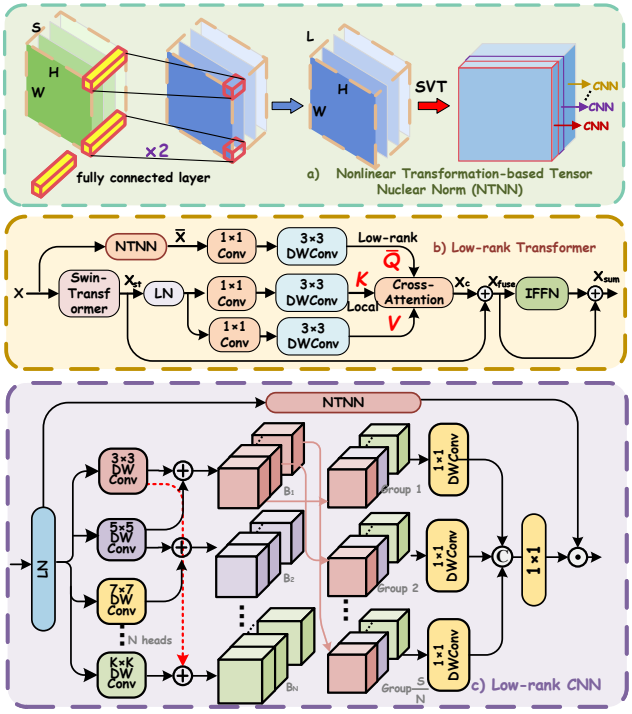


Figure 4. Illustration of the proposed prior extractor. (a) NTNN. (b) Low-rank Transformer. (c) Low-rank CNN.

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are the Lagrange multipliers. Based on this formulation, we unroll the iterative pansharpening algorithm into a series of blocks, forming the RPCA unfolding network. Note RPCA serves as the backbone, where the sparsity of the spatial offset component allows that the spatial fidelity term in Eq.(5) is decomposed into five subproblems. The interpolated LRMS image is used as input, with other variables initialized to zero, and propagated through K unfolded blocks, each corresponding to one iteration of Algorithm 2 (Appendix). $\boldsymbol{S}$ and $\boldsymbol{M}$ are updated via thresholding; $\boldsymbol{X}$ via point-wise convolutions approximating $\boldsymbol{R}$ and $\boldsymbol{R}^\top$; and $\boldsymbol{T}$ via the proposed Low-Rank Embedding using Transformer or CNN layers. The final output is trained end-to-end with an ℓ_1 loss.

3. Low-rank prior extractor

Within the RPCA-based paradigm of (5), a well-behaved prior extractor is a crucial decision-maker for pansharpening performance. Instead of most hand-crafted priors, CNNs and Transformers have emerged as the most sought-after choices owing to the powerful feature extraction. However, both of these two focus on learning implicit priors in a data-driven manner. Without the consideration of any explicit priors, high performance cannot be guaranteed due to the purely black-box nature. To strive for joint implicit-explicit usage, we propose a low-rank network (LRNet), which can be easily embedded into the off-the-shelf Transformers and CNNs, as shown in Fig. 4.

3.1. NTNN Module

Upon reviewing the TNN framework, it is typically represented by matrices of linear transformations, such as DFT [27], DCT [14], or data-adaptive matrices [30], which can be generally expresses as

$$\|\mathcal{X}\|_{\mathrm{TNN}} \triangleq \frac{1}{S}\sum_{k=1}^{S}\|\boldsymbol{Z}_k\|_*, \tag{6}$$

where $\mathcal{X} = \mathcal{Z} \times_3 \boldsymbol{F}^\top$, $\boldsymbol{F}$ is a linear transformation, and $\boldsymbol{Z}_k$ is the k-th frontal slice of $\mathcal{Z}$. Along this line, Luo et al. [16, 17] proposed self-supervised nonlinear transform-based TNN (STNN) to model the complex and nonlinear nature of real-world data. However, a single transformation may still be inadequate to accommodate the complexity of the data structure. Hence, we elaborate a dual-enhanced TNN, where data undergo two transformations for more enriched features. Specifically, we use two-stage 1-D convolution layers and a one-time activation function *ReLU* along mode-3 as the temporal linear transform. Here, the two linear layers convey information about the data in different states, leading to a more accurate transformation domain. The concerned *ReLU* serves to ensure that the input values are restricted to non-negative range. The expression concerned is written as

$$\mathcal{F}(\mathcal{X}) = \mathrm{ReLU}((\mathcal{X} \times_3 \boldsymbol{W}) \times_3 \bar{\boldsymbol{W}}) \tag{7}$$

where $\boldsymbol{W} \in \mathbb{R}^{S\times L}$ and $\bar{\boldsymbol{W}} \in \mathbb{R}^{L\times L}$ are parameters in linear layers. A larger L would enrich a well-rounded $\mathcal{F}: \mathbb{R}^{H\times W\times S} \to \mathbb{R}^{H\times W\times L}$ to achieve a better low-rank approximation. On that basis, the overall NTNN is given as

$$\|\mathcal{X}\|_{\mathrm{NTNN}} \triangleq \frac{1}{L}\sum_{k=1}^{L}\|\mathcal{F}(\mathcal{X})^{(k)}\|_*, \tag{8}$$

Traditionally, SVT is a widely used approach for solving TNN-constrained problems. However, existing TNNs omit this step due to its high computational cost, which may compromise the enforcement of absolute low-rankness. To enjoy the advantages of SVT yet within a lightweight computation, we elaborate a joint Depthwise Convolution and SoftShrink framework as a substitute. The red operations in Algorithm 1 indicate the proposed NTSVT variant. As shown in Algorithm 1 and Fig. 4(a), traditional SVT requires slice-wise SVD computations and an entry-wise shrinkage operation. Instead, we reformulate this process using a Depthwise Convolution + SoftShrink function, which together perform local spectral decomposition and adaptive singular value thresholding. Specifically, depth-wise convolution serves as a channel-wise feature filter, extracting dominant structures, while SoftShrink directly en-

forces adaptive spectral thresholding, analogous to the singular value shrinkage in SVT. By leveraging 3×3 kernel, the computational complexity is reduced from $O(n^4)$ in standard SVT to $O(9n^3)$. More importantly, this nonlinear SVT retains the principles of the original one while ensuring structural consistency with low-rank modeling, as evidenced in Theorem 1.

Algorithm 1 Traditional SVT $\rightarrow$ NTSVT

Input: Input feature $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{H \times W \times S}$.
$\mathcal{X} = \boldsymbol{F}(\mathcal{X}) \rightarrow \mathcal{F}(\mathcal{X}) = \text{ReLU}((\mathcal{X} \times_3 \boldsymbol{W}) \times_3 \bar{\boldsymbol{W}})$
for $i = 1$ to $S \rightarrow L$ **do**
  $[\mathcal{U}^i, \mathcal{S}^i, \mathcal{V}^i] = \text{SVD}(\mathcal{X}^i) \rightarrow \mathcal{Q}^i = DW(\mathcal{F}(\mathcal{X})^i)$
  $\mathcal{S}^i = \mathcal{D}_\tau(\mathcal{S}^i) \rightarrow \mathcal{Q}^i = SoftShrink(\mathcal{Q}^i)$
  $\mathcal{P}^i = \mathcal{U}^i \mathcal{S}^i \mathcal{V}^i$
end for
Output: $\mathcal{P} \rightarrow \mathcal{Q}$

Theorem 1 *Let $\mathcal{T} \in \mathbb{R}^{H \times W \times S}$ be a third-order tensor. Define the Nonlinear Tensor Singular Value Thresholding (NTSVT) operation as: $\mathcal{T}^\star = \mathcal{H}_\tau \circ \mathcal{F}_\theta(\mathcal{T})$, where $\mathcal{F}_\theta$ is a depthwise separable convolution applied as: $\mathcal{F}_\theta(\mathcal{T}) = \mathcal{T} \times_1 \boldsymbol{K}_h \times_2 \boldsymbol{K}_w$, with convolutional kernels $\boldsymbol{K}_h, \boldsymbol{K}_w \in \mathbb{R}^{3\times 3}$, and $\mathcal{H}_\tau$ is a SoftShrink function performing element-wise thresholding. Since depthwise convolution approximates a localized spectral decomposition, the overall transformation is equivalent to applying SVT in a localized manner: $\mathcal{T}^\star = \mathcal{U}\tilde{\boldsymbol{\Sigma}}\mathcal{V}^\top, \quad \tilde{\sigma}_i = \max(\sigma_i - \tau, 0)$, where $\mathcal{U}, \mathcal{V}$ are local orthonormal matrices, and $\tilde{\boldsymbol{\Sigma}}$ is the thresholded singular value matrix. Moreover, NTSVT preserves or reduces the rank, satisfying:*

$$\text{rank}(\mathcal{T}^\star) \leq \text{rank}(\mathcal{T}). \tag{9}$$

The proof of the theorem is provided in the Appendix.

Remark: *The proposed NTNN framework strictly adheres to the fundamental principles of TNN while modifying only the transformation process, ensuring that it naturally inherits the low-rank guarantee. The low-rank property is empirically confirmed by the rapid singular value decay shown in the log-domain visualization of Fig. 5, where NTNN exhibits a steeper decay than TNN and STNN, demonstrating its superior ability to suppress high-rank noise and model low-rank structures effectively. Additionally, the integration of NTNN into Low-Rank Transformer and CNN architectures is systematically detailed in Algorithms 2 and 3, demonstrating its adaptability across different network structures.*

3.2. Low-rank Embedding on Transformer/CNN

The limited receptive field is the congenital disadvantage of CNNs. While self-attention in Transfomers enjoys a high reputation in long-range dependency, the mostly window-based treatment again incurs a certain loss of the contextual information. The proposed NTNN, armed with NTSVT, presents an excellent solution to this problem, which is born with global low-rank modeling capability. Therefore, we amalgamate the information derived from nonlinear TNN with the local characteristics inherent to CNNs and the attention weights of Transformers, engineering low-rank Transformer/CNN networks as shown in Fig. 4(b) and (c).

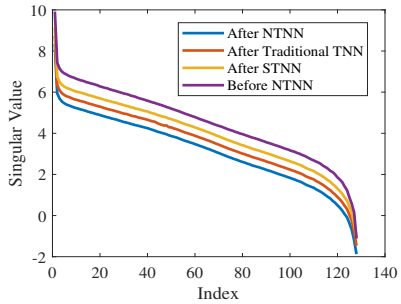


Figure 5. Variations of singular values by using NTNN.

Algorithm 2 Low-rank Transformer

Input: The input feature $\mathcal{X} \in \mathbb{R}^{H \times W \times S}$
Output: The summed feature $\boldsymbol{X}_{\text{sum}}$
$\bar{\mathcal{X}} = \text{NTSVT}(\boldsymbol{\mathcal{X}})$
$(\bar{\boldsymbol{Q}}, \sim, \sim) = \text{Attention}(\bar{\mathcal{X}})$
$(\sim, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{X}_{st} = \text{Swin-Attention}(\mathcal{X})$
$\boldsymbol{X}_c = \text{Cross-Attention}(\bar{\boldsymbol{Q}}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\bar{\boldsymbol{Q}}\boldsymbol{K}^T\right)\boldsymbol{V}$
$\boldsymbol{X}_{\text{fuse}} = \boldsymbol{X}_{st} + \boldsymbol{X}_c$
$\boldsymbol{X}_{\text{sum}} = \text{IFFN}(\text{LN}(\boldsymbol{X}_{\text{fuse}})) + \boldsymbol{X}_{\text{fuse}}$

3.2.1. Low-rank Transformer (LRTR)

In traditional Transformer, the query, key, and value all originate from the same data source. Instead, we employ the output from NTNN to serve as the query, while adopting the output from a Swim-Transformer as the key and value, which achieves explicit-implicit prior fusion in a cross-attention manner. Specifically, given the input features $\mathcal{X} \in \mathbb{R}^{H \times W \times S}$, NTNN provides a globally low-rank approximation as $\bar{\mathcal{X}} = \|\mathcal{X}\|_{\text{NTNN}}$, in which the solution $\bar{\mathcal{X}} \in \mathbb{R}^{H \times W \times L}$ can be easily achieved via Algorithm 1.

To bring the superiority of joint explicit-implicit priors into full play, a hybrid operation is designed, as shown in Fig. 4(b). To be specific, a 3×3 depth-wise layer followed by a 1×1 convolution is imposed to ensure a low-rank query $\bar{\boldsymbol{Q}} \in \mathbb{R}^{L \times (HW)}$, i.e., $\bar{\boldsymbol{Q}} = \text{Conv}_{1\times1}(\text{DW}_{3\times3}(\mathcal{F}(\mathcal{X})))$. The same operation as well as an additional LayerNorm layer is then employed on Swin-Transformer output $\boldsymbol{X}_{st}$, generating $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{L \times (HW)}$. Afterward, the cross-attention can

be learned as

$$\boldsymbol{X}_c = \text{Cross-Attention}(\bar{\boldsymbol{Q}}, \boldsymbol{K}, \boldsymbol{V}) = Softmax(\bar{\boldsymbol{Q}}\boldsymbol{K}^T)\boldsymbol{V} \tag{10}$$

Subsequently, $\boldsymbol{X}_{st}$ and $\boldsymbol{X}_c$ are fused through a residual connection, getting a higher-quality output $\boldsymbol{X}_{\text{fuse}}$. Finally, we feed $\boldsymbol{X}_{\text{fuse}}$ to IFFN [25] to further aggregate the features. Note that $\mathcal{F}(\mathcal{X})$ is generated by NTNN, which enforces a low-rank representation. The sequential depthwise and 1×1 convolutions won't increase rank. As a result, $\bar{\boldsymbol{Q}}$ remains its low-rankness. Based on the multiplication property, we have $\text{rank}(\bar{\boldsymbol{Q}}\boldsymbol{K}^T) \leq \min(\text{rank}(\bar{\boldsymbol{Q}}), \text{rank}(\boldsymbol{K}))$. Given that $\bar{\boldsymbol{Q}}$ is inherently low-rank, the term $\bar{\boldsymbol{Q}}\boldsymbol{K}^T$ inherits this nature. Furthermore, the softmax operation does not increase rank, ensuring that $\boldsymbol{X}_c$ remains low-rank throughout the transformation process.

3.2.2. Low-rank CNN (LRCNN)

Although the output of LRTR is already filled with features as global as possible, it is believed that local information is also not negligible in complex remote sensing images. To that end, we instinctively pivot towards leveraging the capabilities of CNNs, as depicted in Fig. 4(c). To be concrete, multi-scale information flow with kernel size varied is realized. The input feature $\mathcal{X}$ is initially partitioned into N equal heads along the channel dimension, followed by depth-wise separable convolutions with kernel sizes starting from 3×3 and gradually increasing by 2 per head. Afterward, we broadcast information from larger spatial dimensions to the output features of the preceding ones to alleviate the lack of global modeling capability in CNNs.[1], i.e.,

$$\begin{aligned}
\mathcal{A}_1 &= DW_{3\times3}(\mathcal{X}_1) + DW_{5\times5}(\mathcal{X}_2) \\
\mathcal{A}_2 &= DW_{5\times5}(\mathcal{X}_2) + DW_{7\times7}(\mathcal{X}_3) \\
&\vdots \\
\mathcal{A}_N &= DW_{(2N+1)\times(2N+1)}(\mathcal{X}_N) + DW_{3\times3}(\mathcal{X}_1)
\end{aligned} \tag{11}$$

Subsequently, we merge the broadcasted multiple heads and modulate the multi-scale output by multiplying it with the output of NTNN, i.e.,$\mathcal{X}_{\text{final}} = \text{NTSVT}(\boldsymbol{\mathcal{X}}) \odot (\text{Concat}(\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_N))$, where $\odot$ is the element-wise multiplication. Here, NTNN functions as a special feature modulator to provide CNNs with global low rankness, which cannot be achieved solely through multi-scale interactions. Given the modulated $\text{NTSVT}(\boldsymbol{\mathcal{X}})$, the applied element-wise multiplication further yields

$$\text{rank}(\mathcal{X}_{\text{final}}) \leq \min(\text{rank}(\text{NTSVT}(\boldsymbol{\mathcal{X}})), \text{rank}(\text{Concat}(\mathcal{A}_i))). \tag{12}$$

Note the spectral-wise concatenation has no impact on spatial rank, then the final output $\mathcal{X}_{\text{final}}$ enjoys low-rankness.

[1]To form a complete aggregation chain, we broadcast the 3×3 outputs to the largest spatial features.

3.2.3. Hybrid integration

As depicted in Fig. 2(b) and (c), the overall hybrid integration (HI) holds a U-shaped architecture, which is primarily composed of several LRNet blocks. Three components, *i.e.*, LRTR, LRCNN, and IFFN, occupy the main positions. The U-shaped network follows an "embedding-encoder-bottleneck-decoder-unembedding" flow to extract refined feature representations. To commence, a patch embedding layer is utilized to segment the intermediate image $\boldsymbol{T}^t$ into non-overlapping patch tokens. Following this, two LRNet blocks for each encoder and decoder, as well as a single LRNet block for the bottleneck, collaborate to explore local-global properties. In particular, an additional connection unit is facilitated to change the channel size via downsampling/upsampling operations and point convolution. Finally, we obtain the approximate solution $\boldsymbol{T}^{t+1}$ via a patch unembedding layer.

4. Experiments

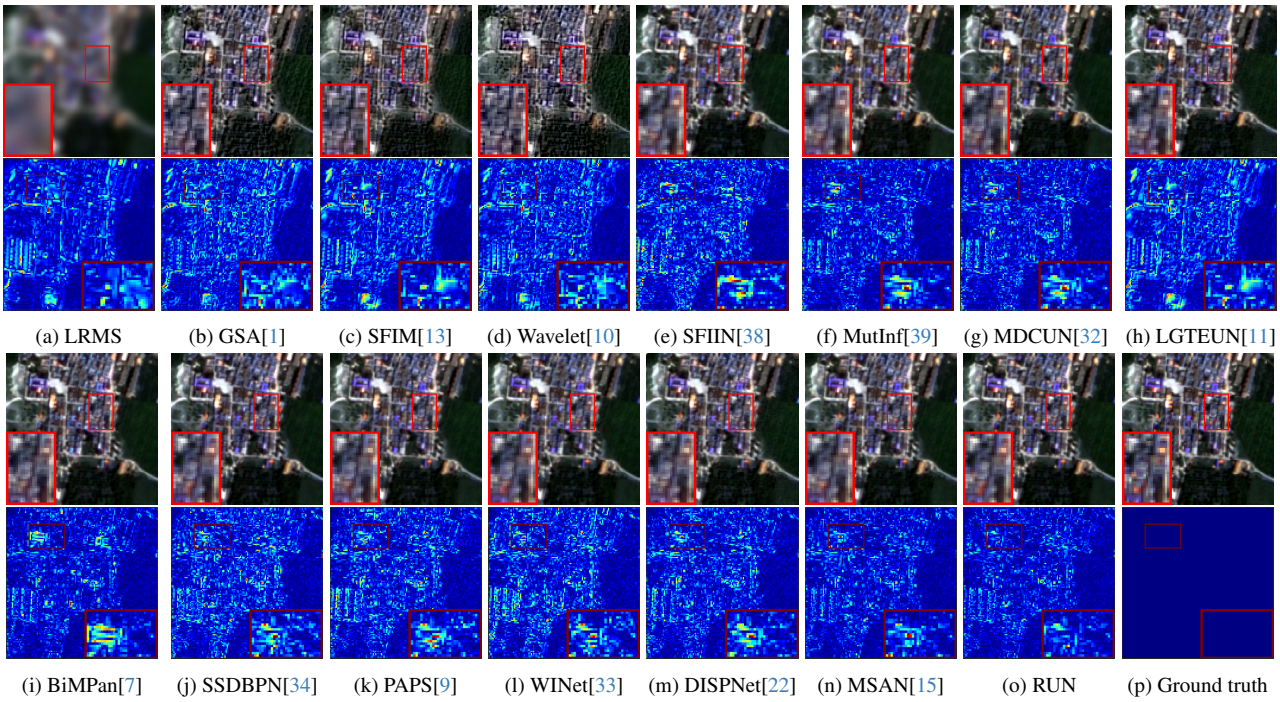
For the evaluation of RUN, we conduct comprehensive experiments on acknowledged remote sensing images obtained from three satellites, including GF2, WV2[2], and WV3 [11]. Following the Wald protocol [21], downsampling is used to generate the concerned low-resolution dataset for each satellite sensor. The three datasets are further split into non-overlapping subsets for training (approximately 1000 LRMS/PAN/GT image pairs) and testing (140 LRMS/PAN/GT image pairs). Due to page limitation, more contents such as metrics, competitors, and real-world experiments are presented in the supplementary material.

4.1. Comparisons with State-of-the-arts

Table 1 presents the numerical results of all competing methods. Evidently, DL-based methods enjoy substantial superiority over model-based ones in all metrics. Notably, our model consistently achieves top rankings in all cases. From a more intuitive perspective, specific to the PSNR values, our proposal outperforms the second-best by 0.0493 dB, 0.127 dB, and 0.899 dB on WV3, WV2, and GF2 datasets, respectively. This exceptional performance demonstrates that RPCA framework matches desirably to the degradations of the pansharpening issue. Furthermore, the designed low-rank module possesses powerful capabilities for both global and local feature acquisition.

For visual comparison, a recovered example from GF-2 and the corresponding error maps are shown in Fig. 6. At first glance, model-based methods often exhibit vagueness and fail to achieve a clear target, while DL-based models offer relatively satisfactory results with some loss of texture details and several unexpected artifacts. In the enlarged

[2]https://github.com/lms-07/LGTEUN.



(a) LRMS (b) GSA[1] (c) SFIM[13] (d) Wavelet[10] (e) SFIIN[38] (f) MutInf[39] (g) MDCUN[32] (h) LGTEUN[11]

(i) BiMPan[7] (j) SSDBPN[34] (k) PAPS[9] (l) WINet[33] (m) DISPNet[22] (n) MSAN[15] (o) RUN (p) Ground truth

Figure 6. Visual comparisons on typical GF-2 data.

Table 1. Quantitative comparison of different methods on WV-3, WV-2, and GF-2 satellite data sets. The best values and the second best values are respectively highlighted by red and blue colors.

Method	WorldView-3					WorldView-2					GaoFen-2				
	PSNR↑	SSIM↑	Q8↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	Q4↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	Q4↑	SAM↓	ERGAS↓
GSA_{07}	22.5164	0.6343	0.5742	6.3369	7.8267	33.5975	0.8899	0.5681	3.2830	2.5402	36.0557	0.8838	0.5517	3.6727	3.5758
$SFIM_{00}$	21.4154	0.5415	0.4525	6.5718	8.8553	32.6334	0.8728	0.5159	3.4205	3.1919	34.7715	0.8572	0.4584	3.7643	4.2073
$Wavelet_{09}$	21.4464	0.5656	0.5271	8.6116	9.1545	32.1992	0.8500	0.4577	3.6555	3.3799	33.9208	0.8197	0.4033	3.9821	4.6445
$SFIIN_{22}$	31.6587	0.9492	0.9435	3.7357	2.8016	41.9489	0.9752	0.8108	1.3121	1.0084	44.7248	0.9802	0.8721	1.6043	1.3361
$MutInf_{22}$	31.8298	0.9523	0.9469	3.6440	2.7526	41.9522	0.9760	0.8258	1.3006	1.0153	44.8305	0.9800	0.8836	1.5871	1.3394
$MDCUN_{22}$	31.2978	0.9429	0.9363	3.7873	2.9295	42.3351	0.9772	0.8370	1.2376	0.9638	45.5677	0.9825	0.8915	1.4439	1.2249
$LGTEUN_{23}$	32.2188	0.9545	0.9494	3.4664	2.6286	42.6240	0.9782	0.8401	1.2089	0.9282	45.8364	0.9840	0.8973	1.4152	1.1824
$BiMPan_{23}$	32.2850	0.9565	0.9518	3.4549	2.5821	42.2729	0.9771	0.8380	1.2433	0.9649	45.5398	0.9822	0.8928	1.4324	1.2357
$SSDBPN_{23}$	31.0621	0.9434	0.9379	3.9649	3.0059	41.7904	0.9743	0.8243	1.3350	1.0317	45.5734	0.9825	0.8916	1.4725	1.2248
$PAPS_{24}$	32.2540	0.9550	0.9497	3.4606	2.6127	41.6371	0.9746	0.8241	1.3407	1.0316	45.2308	0.9815	0.8836	1.5126	1.2710
$WINet_{24}$	32.5091	0.9532	0.9509	3.4518	2.5830	42.3578	0.9731	0.8375	1.2893	0.9765	45.9842	0.9825	0.8889	1.3581	1.2147
$DISPNet_{24}$	32.5571	0.9568	0.9521	3.4406	2.5798	42.4620	0.9748	0.8411	1.2147	1.0880	45.9184	0.9801	0.8845	1.3808	1.2154
$MSAN_{25}$	32.5018	0.9535	0.9511	3.4512	2.6053	42.4589	0.9739	0.8320	1.2178	0.9875	46.0123	0.9850	0.8988	1.3986	1.1510
RUN	32.6064	0.9575	0.9530	3.4377	2.5610	42.7509	0.9791	0.8423	1.1860	0.9139	46.9117	0.9872	0.9122	1.2660	1.0456

thumbnails, all competitors struggle to effectively reconstruct the roof section, as indicated by the red points in the pseudocolored images. Besides, various degrees of dimness happen that disable the ability to showcase more details. Comparatively, RUN shows sharper results in these areas, with fewer artifacts on the error map, especially in the zoom-in region. The advantage can be attributed to the LRNet module, which excels in extracting both implicit and explicit features. Fig. 7 further visualizes the overall performance in terms of PSNR, FLOPS, and Params. Evidently, the other competitors lag behind our method by a large margin. Yet with the parameters and FLOPs lying only in the above-average place, we believe this level of overhead is acceptable taking the efficacy into account.

4.2. Analysis and Discussion

Effect of unfolding stages. To investigate the performance fluctuation along with the increasing iterations, the results of RUN adapting to different stages, ranging from 1 to 5, are reported in Fig. 8(a). Due to the powerful learning capabilities of both implicit prior extractor and explicit low-rankness, only a very small number of iterations are required to reach and stabilize the peak performance, *e.g.*, 3 or 4. As evidenced in Fig. 8(b), more iterations naturally incur heavier computational overhead. Hence, $T = 4$ is recommended as the default setting, with possible attempts

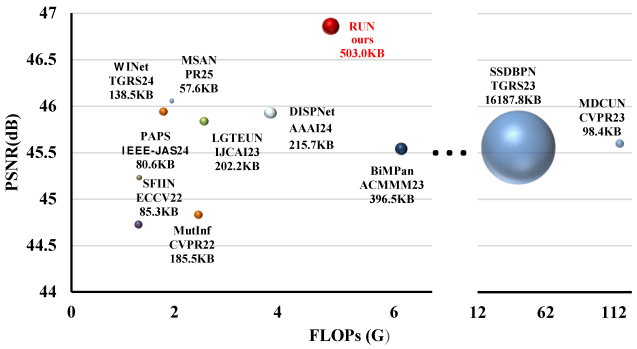


Figure 7. Comparisons of FLOPs (X-axis), PSNR (Y-axis), and parameters (Params, circle radius) on GF2 dataset.

Table 2. Ablation of different CNN components on GF2.

Multiscale	Broadcast	NTNN	PSNR	Params (M)	FLOPs (G)
✔			46.4248	0.4630	3.36
✔	✔		46.6487	0.4742	3.51
✔	✔	✔	46.9117	0.5030	3.60

to go deeper for better efficacy.

Effect of sparse SOC. Recall that the employment of RPCA model leads to avoidance of spatial degradation. Here, we adopt one-round (OR) and two-rounds (TR) 3×3 convolutions respectively to simulate matrices $\boldsymbol{B}$ and $\boldsymbol{B}^T$, and compare them with the finally used sparse SOC treatment. The experiments are carried out on GF2 dataset with 4×4 downsampling factor, whose results are given in Fig. 9(a). Evidently, our proposal consistently occupies the summit point with a higher median value. The improvements achieved can be attributed to the elimination of complex spatial degradation and the reservation of relatively simpler/smaller spectral matrix, which prevents simulation loss to a deep extent. Moreover, the continuous sparsification of spatial offset component also leads to better adaptation.

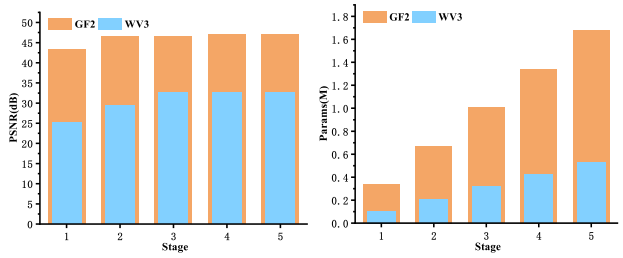


Figure 8. Efficacy and efficiency of RUN with different numbers of stages on GF2 and WV3 satellite data sets.

Effect of NTNN module. To validate the behavior of NTNN, Fig. 9(b), Fig. 10, and Table 2 offer the ablation results. Three observations can be drawn. First, in comparison to typical self-attention and Swin-Transformer,

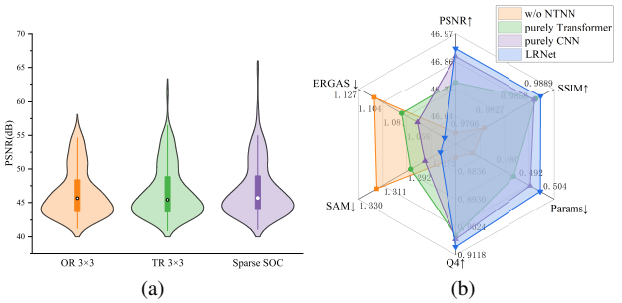


Figure 9. Ablations of sparse SOC (a) and the low-rank component (b) in terms of representative metrics.

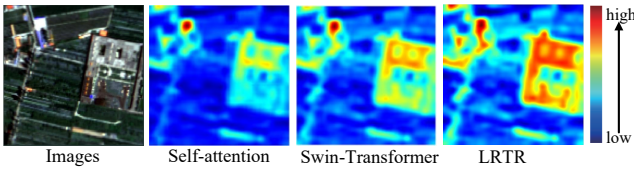


Figure 10. Different heat maps of various attentions.

NTNN delivers superiority in capturing salient features, as shown with higher red luminance in Fig. 10 that highlights saliency. Second, through the complementation across scale-varied receptive fields and the imposition of global low rankness, NTNN enjoys a lead of 0.48dB PSNR over the pure CNN, with only 0.04M parameters added. Third, the radar chart in Fig. 9(b) demonstrates that NTNN, when integrated with CNN and Transformer, achieves outstanding pansharpening performance, again verifying the benefit of merging implicit-explicit knowledge.

5. Conclusion

To alleviate the notorious issue that solving pansharpening should meticulously design four learning modules, we propose a new paradigm, namely RPCA-based unfolding solver, characterized by reduced labor intensity. Concretely, it is derived that the sparsification of spatial offset component fits concordantly with the update of target variable. Following this, the simulation of spatial degradation matrices can be omitted, replaced by sparsity modeling of the difference between HRMS and upsampled LRMS. In addition, for the prior extraction within the RPCA framework, the newly designed low-rank Transformer performs deep cross-complementarity of explicit-implicit features without compromising efficiency and flexibility. The concerned low-rank CNN provides an enhanced ability to capture spatial-spectral correlations through multi-scale spatial cross-collaboration and global low-rankness. Across three satellite datasets, our proposal outperforms all competitors, even with less molded degradation matrices.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62276232 and 62406285, and the Key Program of Natural Science Foundation of Zhejiang Province under Grant LZ24F030012.

References

[1] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. 7

[2] Cong Bai, Hongkai Li, Jinglin Zhang, Ling Huang, and Lu Zhang. Unsupervised adversarial instance-level image retrieval. *IEEE Transactions on Multimedia*, 23:2199–2207, 2021. 1

[3] Yubin Gu, Honghui Xu, Yueqian Quan, Wanjun Chen, and Jianwei Zheng. Orsi salient object detection via bidimensional attention and full-stage semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2

[4] Yubin Gu, Siting Chen, Xiaoshuai Sun, Jiayi Ji, Yiyi Zhou, and Rongrong Ji. Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration. *Pattern Recognition*, 164:111478, 2025.

[5] Yubin Gu, Yuan Meng, Siting Chen, Jiayi Ji, Xiaoshuai Sun, Weijian Ruan, and Rongrong Ji. Sfir: Optimizing spatial and frequency domains for image restoration. *Pattern Recognition*, page 112188, 2025. 2

[6] Yubin Gu, Yuan Meng, Jiayi Ji, and Xiaoshuai Sun. Acl: Activating capability of linear attention for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17913–17923, 2025. 2

[7] Junming Hou, Qi Cao, Ran Ran, Che Liu, Junling Li, and Liang-jian Deng. Bidomain modeling paradigm for pansharpening. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 347–357, 2023. 7

[8] Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting. *Nature Communications*, 16(1): 5923, 2025. 2

[9] Yanan Jia, Qiming Hu, Renwei Dian, Jiayi Ma, and Xiaojie Guo. Paps: Progressive attention-based pan-sharpening. *IEEE/CAA Journal of Automatica Sinica*, 11(2):391–404, 2024. 7

[10] Roger L King and Jianwen Wang. A wavelet based algorithm for pan sharpening landsat 7 imagery. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 849–851. IEEE, 2001. 7

[11] Mingsong Li, Yikun Liu, Tao Xiao, Yuwen Huang, and Gongping Yang. Local-global transformer enhanced unfolding network for pan-sharpening. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1071–1079, 2023. 3, 6, 7

[12] Huangxing Lin, Yuhang Dong, Xinghao Ding, Tianpeng Liu, and Yongxiang Liu. Unsupervised pan-sharpening via mutually guided detail restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3386–3394, 2024. 1

[13] JG Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472, 2000. 7

[14] Canyi Lu. Transforms based tensor robust pca: Corrupted low-rank tensors recovery via convex optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1145–1152, 2021. 4

[15] Hangyuan Lu, Yong Yang, Shuying Huang, Rixian Liu, and Huimin Guo. Msan: Multiscale self-attention network for pansharpening. *Pattern Recognition*, 162:111441, 2025. 7

[16] Yisi Luo, Xi-Le Zhao, Deyu Meng, and Tai-Xiang Jiang. Hlrtf: Hierarchical low-rank tensor factorization for inverse problems in multi-dimensional imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19303–19312, 2022. 4

[17] Yi-Si Luo, Xi-Le Zhao, Tai-Xiang Jiang, Yi Chang, Michael K Ng, and Chao Li. Self-supervised nonlinear transform-based tensor nuclear norm for multi-dimensional image recovery. *IEEE Transactions on Image Processing*, 31:3793–3808, 2022. 4

[18] Qing Ma, Jiancheng Pan, and Cong Bai. Direction-oriented visual–semantic embedding model for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 1

[19] Yueqian Quan, Honghui Xu, Renfang Wang, Qiu Guan, and Jianwei Zheng. Orsi salient object detection via progressive semantic flow and uncertainty-aware refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 2

[20] Zhao Su, Yong Yang, Shuying Huang, Weiguo Wan, Wei Tu, Hangyuan Lu, and Changjie Chen. Ctcp: Cross transformer and cnn for pansharpening. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 3003–3011, 2023. 2

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

[22] Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, and Jiayi Ma. Deep unfolded network with intrinsic supervision for pan-sharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5419–5426, 2024. 7

[23] Jiaming Wang, Tao Lu, Xiao Huang, Ruiqian Zhang, and Xiaoxiao Feng. Pan-sharpening via conditional invertible neural network. *Information Fusion*, 101:101980, 2024. 1

[24] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Jie Huang, Jocelyn Chanussot, and Gemine Vivone. Lrtcfpan: Low-rank tensor completion based framework for pansharpening. *IEEE Transactions on Image Processing*, 32:1640–1655, 2023. 1

[25] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Chia-Wen Lin, and Liangpei Zhang. Ttst: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33:738–752, 2024. 6

[26] Honghui Xu, Mengjie Qin, Shengyong Chen, Yuhui Zheng, and Jianwei Zheng. Hyperspectral-multispectral image fusion via tensor ring and subspace decompositions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8823–8837, 2021. 2

[27] Honghui Xu, Jianwei Zheng, Xiaomin Yao, Yuchao Feng, and Shengyong Chen. Fast tensor nuclear norm for structured low-rank visual inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):538–552, 2022. 4

[28] Honghui Xu, Jiawei Jiang, Yuchao Feng, Yiting Jin, and Jianwei Zheng. Tensor completion via hybrid shallow-and-deep priors. *Applied Intelligence*, 53(13):17093–17114, 2023. 2

[29] Honghui Xu, Chuangjie Fang, Yilin Ge, Yubin Gu, and Jianwei Zheng. Cascade-transform-based tensor nuclear norm for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[30] Honghui Xu, Chuangjie Fang, Renfang Wang, Shengyong Chen, and Jianwei Zheng. Dual-enhanced high-order self-learning tensor singular value decomposition for robust principal component analysis. *IEEE Transactions on Artificial Intelligence*, 5(7):3564–3578, 2024. 4

[31] Honghui Xu, Yueqian Quan, Mengjie Qin, Yibin Wang, Chuangjie Fang, Yan Li, and Jianwei Zheng. Nonlinear learnable triple-domain transform tensor nuclear norm for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025. 1

[32] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1788–1797, 2022. 7

[33] Jie Zhang, Xuanhua He, Keyu Yan, Ke Cao, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Pan-sharpening with wavelet-enhanced high-frequency information. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 7

[34] Kai Zhang, Anfei Wang, Feng Zhang, Wenbo Wan, Jiande Sun, and Lorenzo Bruzzone. Spatial-spectral dual back-projection network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 7

[35] Yaqun Zhang, Zhichang Guo, Yao Li, and Boying Wu. Structure tensor-driven block-based adaptive variational pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. 2

[36] Xin Zhao, Jiayi Guo, Yueting Zhang, and Yirong Wu. Asymmetric bidirectional fusion network for remote sensing pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 2

[37] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pan-sharpening. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[38] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 274–291. Springer, 2022. 7

[39] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1798–1808, 2022. 7

[40] Man Zhou, Keyu Yan, Xueyang Fu, Aiping Liu, and Chengjun Xie. Pan-guided band-aware multi-spectral feature enhancement for pan-sharpening. *IEEE Transactions on Computational Imaging*, 9:238–249, 2023. 2