# TRT: Harnessing Tensor Ring Transformer for Hyperspectral Image Super-Resolution

**Honghui Xu[1,2], Junwei Zhu[2], Yubin Gu[3], Yueqian Quan[2], Chuangjie Fang[2], Hong Qiu[4], Jianwei Zheng[2, 5*]**

[1]College of Artificial Intelligence, Taizhou University, Taizhou 318000, Zhejiang, China
[2]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[3] School of Informatics, Xiamen University, Xiamen 361005, China
[4]College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China
[5]Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou 310023
xhh9609@gmail.com, {xhh, jwzhu, fangcj, zjw}@zjut.edu.cn, guyubin@stu.xmu.edu.cn, qiuhong@zwu.edu.cn

## Abstract

Deep unfolding networks (DUNs) have recently emerged as a promising approach for hyperspectral image super-resolution (HSISR) by combining the benefits of nonlinear deep learning architectures with interpretable optimization techniques. Despite their advantages, current DUNs face significant challenges, particularly in approximating degradation matrices across both spatial and spectral dimensions, which results in complex and cumbersome model construction. By analyzing the difference between the upsampled low-resolution hyperspectral images (LRHS) and the true target image, we observed that the residual image exhibits strong sparsity, akin to noise. Leveraging this insight, we reformulate the HSISR problem as a robust principal component analysis (RPCA)-based denoising task, effectively eliminating the need for the complex approximation of spatial degradation matrix and its transpose. In addition, we introduce a Tensor Ring Transformer based on multilinear products as the prior term, wherein tokens are mapped to a tensor ring factor domain and the traditional dot product is replaced with a multilinear tensor ring product. This significantly reduces the computational complexity of the Transformer model, from $\mathcal{O}(N^2d)$ to $\mathcal{O}(Nr^2)$, where the tensor ring rank $r$ is significantly smaller than $d$, while maintaining the expressive power. The proposed Tensor Ring Transformer integrates both Softmax and linear attention mechanisms, striking a balance between interpretability—characteristic of model-based approaches—and the efficiency inherent in deep learning techniques. Experimental results across multiple remote sensing datasets demonstrate the superiority of the designed Tensor Ring Transformer, achieving substantial improvements in image quality and computational efficiency compared to current state-of-the-art methods.

## Introduction

Hyperspectral imaging (HSI) is a sophisticated technique that acquires images over numerous subtle narrow spectral bands ranging from the visible to the infrared bands (Xu et al. 2025a; Zhu et al. 2024), enabling accurate identification of materials by analyzing their unique reflectance properties across different wavelengths. This advanced capability has significantly transformed a wide range of image analysis applications, such as target detection (Gu et al. 2023), change detection (Jian, Ou, and Chen 2024), and scene understanding (Ren et al. 2024). However, hardware limitations in optical remote sensing systems impose an intrinsic trade-off between spatial and spectral resolution in hyperspectral cameras. Based on this premise, a cost-effective and practical approach for obtaining high-resolution hyperspectral images (HRHS) involves fusing LRHS with corresponding high-resolution multispectral images (HRMS) of the same scene. This technique, termed the HSISR task, effectively integrates the detailed spatial information of HRMS with the rich spectral content of LRHS.

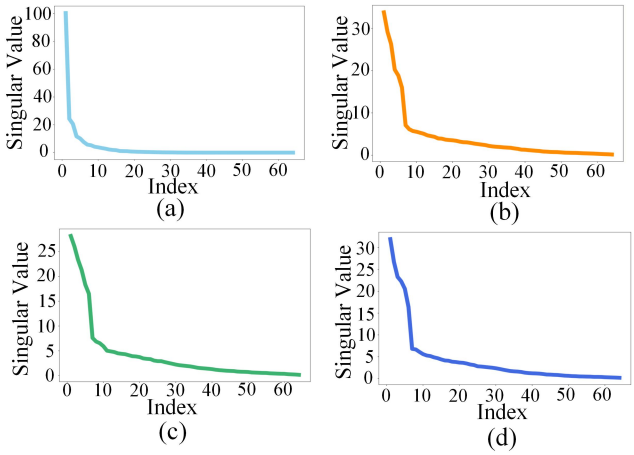



Figure 1: Distribution of Singular Values of CAVE *stuffed toys*. (a) attention map, (b-d) query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$.

Recent HSISR methods generally fall into two categories: model-based and deep learning (DL) approaches. Model-based methods (Dian and Li 2019; Xu et al. 2024a) formulate fusion as an optimization problem solved via iterative algorithms, benefiting from solid theoretical founda-

*Corresponding author.

tions. Their effectiveness mainly depends on regularization terms, which encode priors to guide the solution. Various priors have been explored, including spectral correlation (Xu et al. 2025c; Dian, Liu, and Li 2025; Xu et al. 2022), local spatial smoothness (Hou et al. 2024; Xu et al. 2021; Jiang et al. 2024), and global low-rank structures (Xu et al. 2024b). However, the complexity of HSI data suggests that relying on a single prior is often insufficient; combining multiple priors may help but makes model selection more challenging. In addition, iterative inference increases computation time, limiting the practical use of such methods.

With the advent of large-scale image datasets and the surge in computational capabilities, DL-based methods have exhibited significant advantages over traditional knowledge-driven approaches in the HSISR domain. Convolutional neural networks (CNNs), which are commonly used in DL architectures, excel at extracting local features but struggle to capture long-range dependencies due to their limited receptive fields. In response, Transformer architectures have emerged as a promising solution, offering an unparalleled capacity to model long-range relationships. Ma et al. (Ma et al. 2024) developed a dual cross Transformer model to capture spatial and spectral information. Fang et al. (Fang et al. 2024) introduced multihead feature map attention, multihead feature channel attention, and a multiscale convolutional gated feedforward network, constructing a mixture spatial-spectral Transformer. Despite these advancements, incorporating Transformer models into HSISR field still faces two substantial challenges: *1) Model Interpretability—the 'black box' dilemma. 2) The Quadratic Computational Complexity,* $O(N^2d)$. These challenges pose significant obstacles to the widespread adoption of Transformers for HSISR applications, particularly in scenarios highly demanding interpretability and computational efficiency.

To address the challenge of model interpretability, deep unfolding network (DUN) unfolds iterations of inference algorithms into an end-to-end (Xu et al. 2025a; Jiang et al. 2025), data-driven framework, decomposing the primary problem into a sequence of subproblems using deep models as priors (e.g., CNN or Transformer architectures). This approach maintains the strong theoretical foundations of optimization algorithms while achieving enhanced performance (Xu et al. 2025b). However, designing degradation matrices across spatial and spectral dimensions, along with their transposed counterparts, is labor-intensive, increasing complexity as more iterations are added. Addressing the second challenge, linear attention (Han et al. 2023) reduces computational complexity to $\mathcal{O}(Nd^2)$ by decoupling the Softmax operation, approximating it suitably, and applying a mapping function to both query $\boldsymbol{Q}$ and key $\boldsymbol{K}$. Focused linear attention (Han et al. 2023) introduces an efficient mapping function along with a rank-restoration module to enhance the expressiveness of self-attention. Other methods, such as PVT (Wang et al. 2021) and SOFT (Lu et al. 2021), use sparse attention or matrix decomposition to reduce computational costs. However, these techniques compromise the ability to model long-range dependencies, making them less effective than traditional Softmax attention, and their theoretical foundation for modeling remains less robust.

In this paper, we focus on the limitations of current DUNs and vanilla attention mechanisms, and propose a novel Tensor Ring Transformer (TRT) to unfold a newly elaborated RPCA architecture, by which both high efficiency and expressiveness can be achieved. On the one hand, through a detailed analysis of the target variable's update alongside the upsampling LRHS residual, we verified the inherent unification between residual images and noise characteristics, thus reformulating HSISR into a simplified RPCA-based denoising framework. In this context, the original approximation involving two degradation matrices can be reduced to a single matrix. More importantly, the complex degradation simulation across the spatial dimension is eliminated. On the other hand, the steep decline in the singular value curves of tokens of Transformer and the attention matrix products, as shown in Fig. 1, highlights the low-rank nature, indicating inherent informational repetition in global self-attention and redundancy in token count under the multi-head mechanism. Therefore, we introduce the concept of tensor ring multi-dimensional products into the Transformer framework. Specifically, guided by a ring rank $r$, we apply a pooling operation to map $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ into the tensor ring factor space, reducing the dimensionality from $N \times d$ to $N \times r$; moreover, the conventional dot product is replaced with a multilinear tensor-ring product. The Softmax attention is retained, operating on $\boldsymbol{Q}$ and $\boldsymbol{K}$ within the factor space, but its complexity is reduced from the original $\mathcal{O}(N^2d)$ to $\mathcal{O}(Nr^2)$. Intriguingly, the proposed Tensor Ring Transformer can be regarded as a generalization of linear attention, providing a tensor decomposition perspective on how attention balances the running efficiency and the expressive power. In other words, it seamlessly integrates the merits of both Softmax-based and compressed linear attention mechanisms.

Our major contributions are as follows.

- With the observation that residual images exhibit strong sparsity similar to noise, we reformulate HSISR as a RPCA-based denoising problem. This enables an approximation that omits the spatial degradation matrix, focusing solely on adapting to spectral responses and significantly simplifying the network design.
- Based on the low-rank properties of tokens, we demonstrate that the scaled dot-product attention in Transformers can be replaced by a multilinear tensor ring product, allowing each query, key, and value to be mapped to the ring factor domain.
- TRT establishes multilinear connections among three ring matrices, each encapsulating a set of query, key, and value factors within their respective domains. This effectively creates an elegant fusion of Softmax and linear attention, capturing sufficient attention information.

## RPCA-Based DUN Architecture

In this section, we provide a detailed exposition of the proposed RPCA-based deep unfolding framework, encompassing its evolution process, overall network design, and optimization solution.

## Problem Formulation

In this study, the main objective is to reconstruct the target HRHS $\boldsymbol{X} \in \mathbb{R}^{S\times HW}$ by combining the obtained LRHS $\boldsymbol{Y} \in \mathbb{R}^{S\times hw}$ and HRMS $\boldsymbol{Z} \in \mathbb{R}^{s\times HW}$ from the same scene. Here, $W$, $H$, $w$, and $h$ denote the spatial dimensions, where $w < W$ and $h < H$, and $S$, $s$ represent the spectral bands of HRHS and HRMS, respectively.

The specific relationship linking LRHS, HRMS, and HRHS is given by $\boldsymbol{Y} = \boldsymbol{XB}$ and $\boldsymbol{Z} = \boldsymbol{RX}$, where $\boldsymbol{B} \in \mathbb{R}^{HW\times hw}$ and $\boldsymbol{R} \in \mathbb{R}^{s\times S}$ are spatial and spectral degradation matrices, respectively. Within this framework, HSISR is reformulated as an estimation problem defined by

$$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{XB}\|_F^2 + \|\boldsymbol{Z} - \boldsymbol{RX}\|_F^2 + \lambda f(\boldsymbol{X}), \quad (1)$$

where $f(\cdot)$ represents a regularization term, and $\lambda$ serves as a balancing parameter.

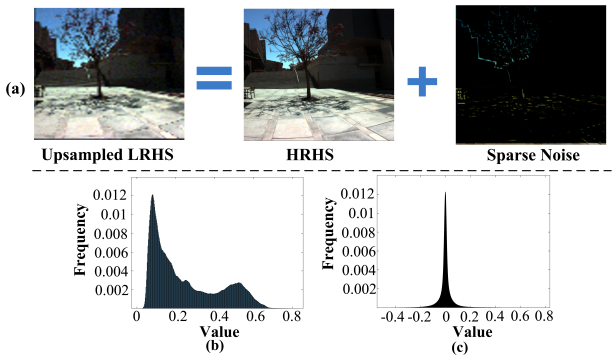



Figure 2: (a) The proposed RPCA Paradigm for HSISR. Histogram of Non-zero Element Distribution of (b) upsampled LRHS $\bar{\boldsymbol{Y}}$ and (c) the residual image $\boldsymbol{S}$.

The proximal gradient algorithm can minimize Eq. (1) by iterating the following equation until convergence:

$$\boldsymbol{X}^{(k+1)} = \arg\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{X} - (\boldsymbol{X}^{(k)} - \gamma \nabla_g(\boldsymbol{X}^{(k)}))\|_F^2 + \lambda f(\boldsymbol{X}), \quad (2)$$

where $\nabla_g(\boldsymbol{X}^{(k)}) = \left(\boldsymbol{X}^{(k)}\boldsymbol{B} - \boldsymbol{Y}\right)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{RX}^{(k)} - \boldsymbol{Z}\right)$, $\gamma$ plays the role of step size, $k$ denotes the iteration number.

Following the typical DUN framework, researchers often use CNNs to model degradation matrices $\boldsymbol{B}$, $\boldsymbol{R}$, and their transposes, i.e., $\boldsymbol{B}^{\mathrm{T}}$, $\boldsymbol{R}^{\mathrm{T}}$. Although substantial effort has been dedicated, the practical performance is barely satisfactory. Moreover, incorporating four additional CNNs for degradation modeling would drastically increase workload, which remains a challenging bottleneck.

## Inspiration from Sparsity

To comprehensively assess the fusion process, we begin by noting that the upsampled LRHS, represented as $\bar{\boldsymbol{Y}} \in \mathbb{R}^{S\times HW}$, shares consistent spectral information with the HRHS, differing solely in spatial resolution. Based on this observation, we propose the following hypothesis: $\bar{\boldsymbol{Y}}$ can be regarded as a spatially noise-contaminated version of the HRHS. To substantiate this hypothesis, it is essential to first ascertain whether the differences between the HRHS and $\bar{\boldsymbol{Y}}$ exhibit characteristics consistent with noise. As illustrated in Fig. 2(a), the residual image, defined as $\boldsymbol{S} = \bar{\boldsymbol{Y}} - \boldsymbol{X}$, displays a marked reduction in primary image features, such as color, texture, and most edges, retaining only scattered auxiliary information. This qualitative observation is further reinforced by the histogram of non-zero elements in both $\bar{\boldsymbol{Y}}$ and $\boldsymbol{S}$ in Fig. 2(b) and (c), where the non-zero elements of $\boldsymbol{S}$ are predominantly concentrated near zero, indicative of significant sparsity. Collectively, these results provide compelling evidence that $\boldsymbol{S}$ is characterized by enhanced sparsity. Thus, if the sparsity of the residual image $\boldsymbol{S}$ is adequately maintained, the reconstructed HRHS will more closely approximate the ground truth. This confirms the hypothesis that the residual image conforms to sparse noise characteristics, thereby reframing the fusion task into a RPCA problem of denoising, which can be generally expressed as

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

where $\alpha$ is the penalty parameter. The $L_1$-norm, $\|\cdot\|_1$, serves to ensure and reinforce the sparsity of the residual image $\boldsymbol{S}$. By employing task-style transfer, we effectively eliminate the cumbersome and complex spatial degradation matrix from Eq. (2), replacing it with the simpler $L_1$-norm. However, to achieve enhanced fusion performance, it remains necessary to incorporate guidance from the HRMS information. Consequently, we retain the spectral regularization term, as well as the easily simulated spectral degradation matrix, which is typically implemented through a $1 \times 1$ convolution operation. Accordingly, the entire RPCA framework for HSISR can be formulated as:

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

Following the ADMM update rules and introducing an auxiliary variable $\boldsymbol{T} = \boldsymbol{X}$, we can transform Eq. (4) into several subproblems for solution.

$$\begin{cases} \boldsymbol{S}^{k+1} = \arg\min_{\boldsymbol{S}} \alpha\|\boldsymbol{S}\|_1 + \frac{\mu_1}{2}\|\bar{\boldsymbol{Y}} - (\boldsymbol{X}^k + \boldsymbol{S}) + \frac{\boldsymbol{M}_1^k}{\mu_1}\|_F^2, \\ \boldsymbol{T}^{k+1} = \arg\min_{\boldsymbol{T}} \lambda f(\boldsymbol{T}) + \frac{\mu_2}{2}\|\boldsymbol{T} - \boldsymbol{X}^k + \frac{\boldsymbol{M}_2^k}{\mu_2}\|_F^2, \\ \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \gamma\boldsymbol{R}^T\left(\boldsymbol{RX}^k - \boldsymbol{Z}\right) - \gamma\mu_2(\boldsymbol{X}^k - \boldsymbol{T}^{k+1} \\ \quad - \boldsymbol{M}_2^k/\mu_2) - \gamma\mu_1(\boldsymbol{X}^k - \bar{\boldsymbol{Y}} + \boldsymbol{S}^{k+1} - \boldsymbol{M}_1^k/\mu_1) \end{cases}$$

Here, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ represent the Lagrange multipliers, with $\mu_1$ and $\mu_2$ denoting penalty parameters. The overall ADMM update steps are illustrated in Fig. 3, and the detailed derivation of the steps is provided in the supplementary materials. For brevity, the $\boldsymbol{S}$ subproblem can be readily solved using the $L_1$-soft thresholding, that is, $\boldsymbol{S}^{k+1} = \text{shrink}\left(\bar{\boldsymbol{Y}} - \boldsymbol{X}^k + \frac{\boldsymbol{M}_1^k}{\mu_1}, \frac{\alpha}{\mu_1}\right)$. The $\boldsymbol{T}$ subproblem is solved as a prior term, where we will use the Tensor Ring Transformer as a data-driven prior for updating, i.e.,

$$\boldsymbol{T}^{k+1} = \textbf{proxNet}_{(\lambda/\mu_2)}(\boldsymbol{X}^k - \frac{\boldsymbol{M}_2^k}{\mu_2}) \quad (5)$$

where **proxNet** is an approximate operator that would be replaced by the proposed network architecture.

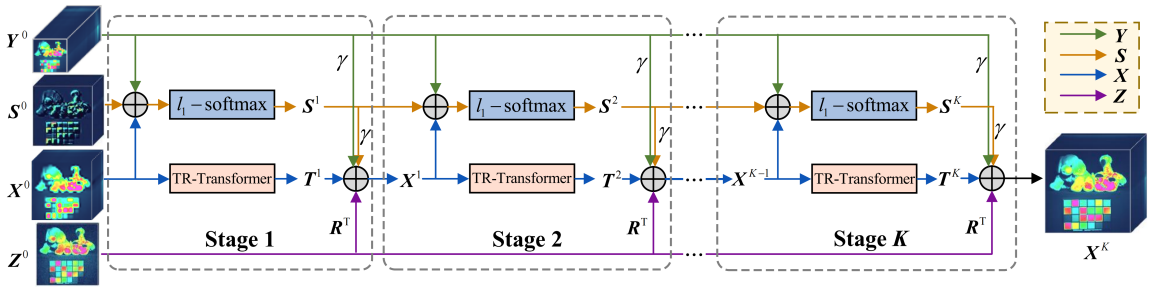



Figure 3: The illustration of RPCA unfolding framework with colors indicating variable update paths.

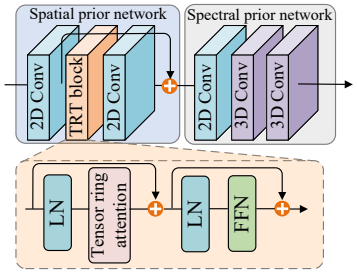



Figure 4: Illustration of the Tensor Ring Transformer.

## Spatial-Spectral Tensor Ring Transformer

The overall architecture of the Tensor Ring Transformer (TRT) is illustrated in Fig. 4. TRT adopts a spatial–spectral design with two components: a spatial prior branch and a spectral prior branch. In the spatial branch, embeddings from an initial 3×3 2D convolution pass through a TRT block and are then refined by another 2D convolution. The TRT block follows a pre-LN Transformer layout with a tensor-ring attention layer and a feed-forward network (FFN), as shown in the bottom of Fig. 4. In parallel, the spectral branch employs two 3D convolutions to capture channel-wise spectral correlations in the HSI. In what follows, we briefly review TR decomposition and then show how a standard Transformer is reformulated into the TRT block.

### Preliminaries of Tensor Ring Decomposition

Tensor ring decomposition (TRD) breaks down a tensor into a sequence of third-order latent tensors, also known as TR factors. Assume that $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is an $N$-th order tensor. Each TR factor is represented as $\mathcal{G}^{(n)} \in \mathbb{R}^{r_{n-1} \times I_n \times r_n}$ for $n = 1, \ldots, N$, with the cyclic constraint $r_0 = r_N$. The set $\boldsymbol{r} = \{r_1, r_2, \ldots, r_N\}$ denotes the TR-rank, which controls the model complexity of TRD. We refer to the first and third modes of each TR factor as the *rank modes*, while the second mode is termed the *dimension mode*. For any tensor $\mathcal{X}$, the element-wise relation of TRD is given by

$$\mathcal{X}(i_1, i_2, \ldots, i_N) = \mathrm{Trace}\left(\mathbf{G}_{i_1}^{(1)} \mathbf{G}_{i_2}^{(2)} \cdots \mathbf{G}_{i_N}^{(N)}\right), \quad (6)$$

where $\mathrm{Trace}(\cdot)$ indicates the trace operation, summing the diagonal elements of a matrix, and $\mathbf{G}_{i_n}^{(n)} \in \mathbb{R}^{r_{n-1} \times r_n}$ denotes the $i_n$-th slice of $\mathcal{G}^{(n)}$ along the dimension mode. The illustration of TRD is shown in Fig. 5(a).

### Tensor Ring Evolution of Transformer

We begin by revisiting the vanilla self-attention mechanism utilized in Vision Transformers. Given an input with $N$ tokens, represented as $x \in \mathbb{R}^{N \times d}$, the self-attention mechanism within each head can be expressed as follows:

$$\boldsymbol{Q} = xW_Q, \quad \boldsymbol{K} = xW_K, \quad \boldsymbol{V} = xW_V,$$
$$\boldsymbol{O}_i = \sum_{j=1}^{N} \frac{\mathrm{Sim}(\boldsymbol{Q}_i, \boldsymbol{K}_j)}{\sum_{j=1}^{N} \mathrm{Sim}(\boldsymbol{Q}_i, \boldsymbol{K}_j)} \boldsymbol{V}_j, \quad (7)$$

where $W_Q, W_K$, and $W_V \in \mathbb{R}^{d \times d}$ denote the projection matrices, and $\mathrm{Sim}(\cdot, \cdot)$ represents the similarity function. When using $\mathrm{Sim}(\boldsymbol{Q}, \boldsymbol{K}) = \exp(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})$ in Eq. (7), the process involves creating an attention map by calculating similarities across all query-key pairs, resulting in a substantial computational complexity of $\mathcal{O}(N^2 d)$.

As revealed in Fig. 1, both the attention map and the $\boldsymbol{Q}/\boldsymbol{K}/\boldsymbol{V}$ projections exhibit a rapidly decaying singular value spectrum, indicating a pronounced low-rank structure and substantial redundancy in token representations. This observation suggests that the essential information is concentrated in a few dominant modes, leaving ample room for compression without sacrificing expressiveness. Motivated by this property, we introduce a tensor-ring representation of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ (illustrated in Fig. 5(b)), which decomposes them into cyclically connected low-rank factors. Such a representation not only aligns with the inherent low-rank characteristics observed in Fig. 1 but also provides a natural foundation for integrating tensor decomposition into the self-attention mechanism, thereby reducing computational complexity while preserving critical structural information.

**Full Mapping Operator.** TRD represents high-dimensional embeddings via a chain of low-rank TR factors for compact yet expressive modeling. Given TR ranks $r = [r_1, r_2, r_3]$, we map the attention projections into TR cores $\mathcal{G}_Q \in \mathbb{R}^{r_1 \times I_1 \times r_2}$, $\mathcal{G}_K \in \mathbb{R}^{r_2 \times I_2 \times r_3}$, and $\mathcal{G}_V \in \mathbb{R}^{r_3 \times I_3 \times r_1}$, where the middle mode is the *dimension mode*. In our setting, $I_1 = N$ is the number of query tokens, $I_2 = M$ denotes the number of (pooled) key/value tokens, and $I_3 = d_h$ is the per-head channel dimension. As illustrated in Fig. 5(b), each core preserves one dimension mode while compressing the two rank modes to $r_n$, thereby removing redundancy.

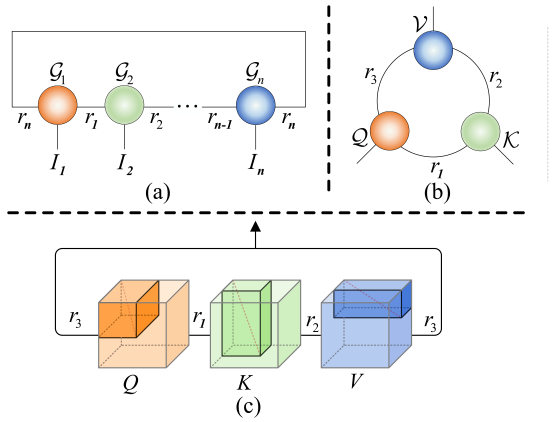



Figure 5: Diagram of TR Structure (a) Original TR, (b) TR Product of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, and (c) Token dimension compression.

**Merging of TR-Cores.** After obtaining the mapped tensor ring factors, we define the Softmax Multilinear Product, which is illustrated in Fig. 5(c), aiming to achieve the merging of TR cores while preserving the Softmax operation.

***Definition 1*** *(TR-inspired Attention).* Let the embeddings be parameterized by TR cores $\mathcal{G}_Q$, $\mathcal{G}_K$, and $\mathcal{G}_V$. The attention output is formulated as a cyclic TR contraction. Specifically, we perform a non-linear contraction of $\mathcal{G}_Q$ and $\mathcal{G}_K$ along rank $r_2$, cascaded with a contraction against $\mathcal{G}_V$ along rank $r_3$:

$$\mathcal{O} = \mathrm{Tr}_{r_1}\Big(\mathrm{Softmax}(\tilde{\mathcal{L}}) \times_{r_3} \mathcal{G}_V\Big), \text{with } \tilde{\mathcal{L}} := \mathcal{G}_Q \times_{r_2} \mathcal{G}_K, \tag{8}$$

where $\tilde{\mathcal{L}}$ denotes the interaction tensor, and $\mathrm{Tr}_{r_1}$ denotes ring-closure over the boundary rank $r_1$.

Eq. (8) outlines the computation in the tensor-ring factor domain. To theoretically characterize the structural constraint induced by the TR parameterization, we derive rank bounds for the pre-softmax score matrix (and the resulting output features), as stated below.

***Theorem 1*** *(Rank bounds of scores and output).* Let $\mathcal{G}_Q \in \mathbb{R}^{r_1 \times I_1 \times r_2}$, $\mathcal{G}_K \in \mathbb{R}^{r_2 \times I_2 \times r_3}$, and $\mathcal{G}_V \in \mathbb{R}^{r_3 \times I_3 \times r_1}$. Define $\tilde{\mathcal{L}} := \mathcal{G}_Q \times_{r_2} \mathcal{G}_K$ and let $\boldsymbol{L} \in \mathbb{R}^{I_1 \times I_2}$ be the $(I_1, I_2)$-matricized pre-softmax score matrix. Then $\mathrm{rank}(\boldsymbol{L}) \le r_2$. Moreover, with $\boldsymbol{P} = \mathrm{Softmax}(\boldsymbol{L})$ (row-wise) and $\boldsymbol{O} = \boldsymbol{P}\boldsymbol{V}$ (where $\boldsymbol{V} \in \mathbb{R}^{I_2 \times I_3}$ is induced by $\mathcal{G}_V$ via linear matricization), we have $\mathrm{rank}(\boldsymbol{O}) \le \mathrm{rank}(\boldsymbol{V}) \le \mathrm{rank}(\mathcal{G}_{V(2)})$, where $\mathcal{G}_{V(2)}$ denotes the mode-2 unfolding of $\mathcal{G}_V$.

***Proposition 1*** *(Complexity of TR factor mapping).* In TR attention, $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ are represented in the tensor-ring factor domain as $\mathcal{G}_Q \in \mathbb{R}^{r_1 \times I_1 \times r_2}$, $\mathcal{G}_K \in \mathbb{R}^{r_2 \times I_2 \times r_3}$, and $\mathcal{G}_V \in \mathbb{R}^{r_3 \times I_3 \times r_1}$, with TR ranks $r = [r_1, r_2, r_3]$. Core construction/processing scales as $O(r_1 I_1 r_2)$, $O(r_2 I_2 r_3)$, and $O(r_3 I_3 r_1)$, yielding $O(r_1 r_2 I_1 + r_2 r_3 I_2 + r_3 r_1 I_3)$ overall. For fixed $r_1 \approx r_2 \approx r_3 \approx r$, this becomes $O(r^2(I_1 + I_2 + I_3))$. With $(I_1, I_2, I_3) = (N, M, d_h)$ and fixed $(M, d_h)$, the dominant cost is $O(Nr^2)$.

Practically, we set the tensor ring rank $r$ as a small hyperparameter, achieving a linear computation complexity of $O(Nr^2)$ relative to the number of input features $N$ while maintaining global context modeling capability.

In TR-based attention, cyclic contraction yields a trace-related scalar component that acts as a query-dependent offset in the pre-softmax logits. Explicit TR contraction is inefficient in the Transformer pipeline. We thus use a trace-free surrogate by subtracting the per-query mean; by softmax shift-invariance, the attention weights are preserved (Approximation 1).

***Approximation 1 (Trace-free Logit Centering).*** Explicitly computing the cyclic trace induced by the Tensor Ring interaction can be expensive. Let $\boldsymbol{L} \in \mathbb{R}^{N \times M}$ denote the pre-softmax logits, where $N$ and $M$ are the numbers of query and (pooled) key tokens, respectively. Instead of forming the cyclic contraction explicitly, we apply row-wise logit centering before softmax: $\tilde{\boldsymbol{L}}_{ij} = \boldsymbol{L}_{ij} - \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{L}_{ik}$. By softmax shift-invariance, subtracting a query-dependent constant preserves attention weights while removing row-wise bias terms (including trace-related offsets shared across keys). This yields a stable and efficient implementation without explicit cyclic contraction.

*The proof of Approximation 1 is provided in the Appendix.*

Taken together, Definition 1, Eq. (8), and Theorem 1 show that the single-head TR attention operator provides a principled low-rank formulation with linear-time complexity. TR attention thus serves as a compact yet expressive building block, where low-rank bounded outputs and row-wise logit centering (Approximation 1) jointly reduce computation without compromising representational capacity.

Building on this foundation, we extend the single-head TR formulation to the multi-head architecture used in modern Transformers. As shown in Eq. (8) and Fig. 5, compact TR-core contractions together with an approximate ring-closure (Approximation 1) yield ring-structured interactions with a low-rank bias, reducing computation and parameters. In our design, each head is equipped with its own set of tensor-ring factors, while keeping head-specific linear projections. This per-head core parameterization preserves the low-rank bias of TR attention and enables diverse interaction patterns across heads, as formalized in Proposition 2.

***Proposition 2*** *(Multi-head TR-inspired Attention).* Let $\boldsymbol{E} \in \mathbb{R}^{N \times d}$ be the input tokens and $H$ the number of heads ($d_h = d/H$). For each head $h$, we use head-specific TR cores $(\mathcal{G}_Q^{(h)}, \mathcal{G}_K^{(h)}, \mathcal{G}_V^{(h)})$ with ranks $r = [r_1, r_2, r_3]$. The head output $\mathcal{O}^{(h)}$ is obtained by applying the TR-inspired attention operator in Definition 1, i.e., $\mathcal{O}^{(h)} = \mathcal{T}(\mathcal{G}_Q^{(h)}, \mathcal{G}_K^{(h)}, \mathcal{G}_V^{(h)})$, where the ring-closure term in $\mathcal{T}(\cdot)$ is approximated by Approximation 1. Let $\boldsymbol{O}^{(h)} \in \mathbb{R}^{N \times d_h}$ denote the matricization of $\mathcal{O}^{(h)}$ along the token mode. The multi-head output is then given by

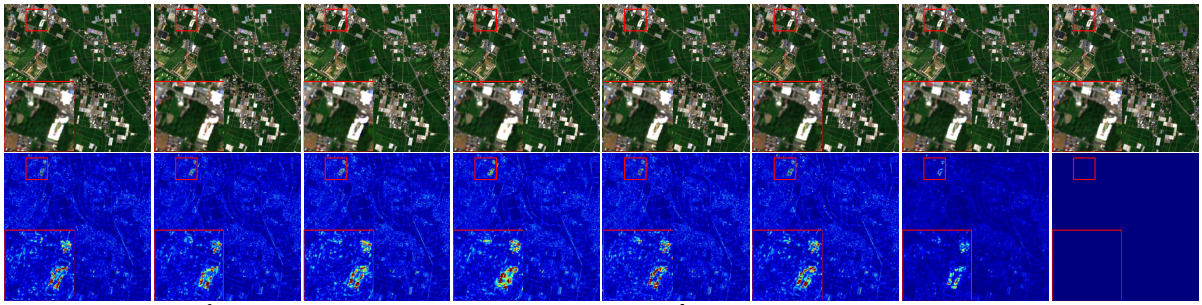
$$\boldsymbol{Y} = \mathrm{Concat}\big(\boldsymbol{O}^{(1)}, \ldots, \boldsymbol{O}^{(H)}\big)\boldsymbol{W}_O, \tag{9}$$

where $\boldsymbol{W}_O \in \mathbb{R}^{(Hd_h) \times d}$ is the output projection.

***Corollary 1.*** Since each TR head follows the same low-rank formulation as in the single-head case, the multi-head TR attention inherits the low-rank structure of the TR decomposition, while its expressive capacity grows with the number of heads through the concatenation of per-head outputs.

| | Method | Chikusei | | | | Pavia | | | | Cave | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ | PSNR↑ | SSIM↑ | SAM↓ | ERGAS↓ |
| 4x | $\text{DSPNet}_{\text{TGRS23}}$ | 43.8834 | 0.9609 | 2.0223 | 3.2694 | 47.8338 | 0.9832 | 3.1502 | 1.6587 | 50.9327 | 0.9965 | 2.1665 | 0.8646 |
| | $\text{S}^2\text{CycleDiff}_{\text{AAAI24}}$ | 44.3554 | 0.9642 | 1.8954 | 3.4253 | 48.0921 | 0.9821 | 3.0784 | 1.6998 | 51.9475 | 0.9970 | 1.8963 | 0.7758 |
| | $\text{MIMO}_{\text{TGRS24}}$ | 44.2295 | 0.9604 | 2.0352 | 3.1132 | 48.2156 | 0.9837 | 3.0965 | 1.6300 | 51.3342 | 0.9965 | 2.2223 | 0.8468 |
| | $\text{DCINN}_{\text{IJCV24}}$ | 44.5477 | 0.9632 | 1.8610 | 3.0684 | 48.2246 | 0.9837 | 3.1024 | 1.6345 | 52.7989 | 0.9974 | 1.7439 | 0.6963 |
| | $\text{S}^3\text{Net}_{\text{NN25}}$ | 44.2589 | 0.9641 | 2.0483 | 3.1265 | 48.6117 | 0.9838 | 3.0678 | 1.6284 | 51.7824 | 0.9956 | 1.9391 | 0.7850 |
| | $\text{CYformer}_{\text{TCSVT25}}$ | 44.5032 | 0.9661 | 1.8841 | 3.0835 | 48.1663 | 0.9827 | 3.2200 | 1.6900 | 52.4568 | 0.9972 | 1.9380 | 0.7604 |
| | TRT | 44.7128 | 0.9683 | 1.6922 | 2.9866 | 49.4814 | 0.9851 | 2.9542 | 1.5761 | 52.9614 | 0.9976 | 1.6711 | 0.6949 |
| 8x | $\text{DSPNet}_{\text{TGRS23}}$ | 41.5310 | 0.9363 | 2.5489 | 2.0554 | 48.1835 | 0.9837 | 3.1145 | 0.8332 | 50.2080 | 0.9959 | 2.2255 | 0.4951 |
| | $\text{S}^2\text{CycleDiff}_{\text{AAAI24}}$ | 42.2425 | 0.9405 | 2.5587 | 1.9875 | 48.3521 | 0.9840 | 3.0875 | 0.8515 | 50.4458 | 0.9952 | 2.2988 | 0.4998 |
| | $\text{MIMO}_{\text{TGRS24}}$ | 41.0199 | 0.9336 | 2.6048 | 2.1838 | 46.2671 | 0.9817 | 3.3764 | 0.8879 | 48.8998 | 0.9946 | 2.6229 | 0.5666 |
| | $\text{DCINN}_{\text{IJCV24}}$ | 41.2587 | 0.9405 | 2.5424 | 2.0687 | 47.3609 | 0.9829 | 3.2287 | 0.8496 | 50.9673 | 0.9966 | 2.0112 | 0.4188 |
| | $\text{S}^3\text{Net}_{\text{NN25}}$ | 41.6357 | 0.9419 | 2.5380 | 2.1106 | 48.2836 | 0.9838 | 3.1058 | 0.8247 | 48.0214 | 0.9932 | 3.5757 | 0.6051 |
| | $\text{CYformer}_{\text{TCSVT25}}$ | 41.9340 | 0.9429 | 2.4507 | 1.9938 | 48.0945 | 0.9831 | 3.1514 | 0.8408 | 50.0274 | 0.9961 | 2.1587 | 0.5019 |
| | TRT | 42.7128 | 0.9433 | 2.3922 | 1.9374 | 48.7409 | 0.9845 | 3.0260 | 0.8139 | 51.2087 | 0.9975 | 1.9805 | 0.3987 |

Table 1: Quantitative comparison of different methods on Chikusei, Pavia, and Cave datasets. The best values and the second best values are respectively highlighted in red and blue colors.



(a) DSPNet (b) $\text{S}^2$CycleDiff (c) MIMO (d) DCINN (e) $\text{S}^3$Net (f) CYformer (g) TRT (h) Ground truth

Figure 6: Visual comparisons on typical Chikusei data.

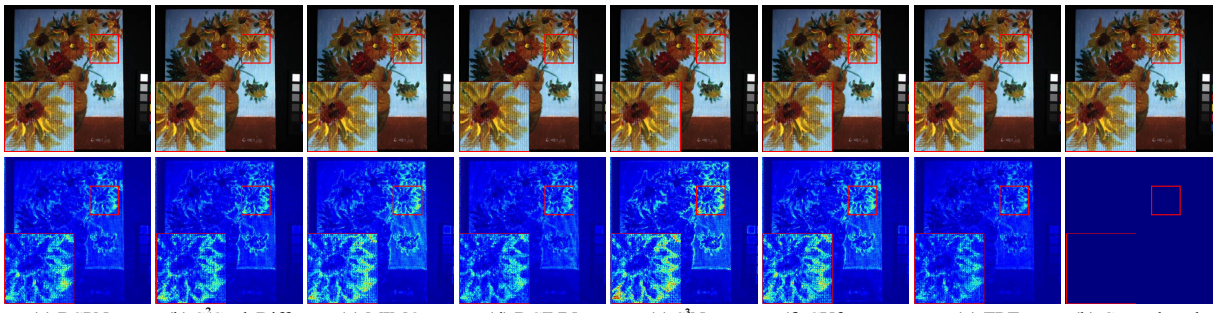
## Experiments

For TRT evaluation, several cutting-edge approaches are selected as competitors, including DSPNet (Sun et al. 2023), MIMO (Fang et al. 2024), DCINN (Wang et al. 2024), $\text{S}^2$CycleDiff (Qu et al. 2024), $\text{S}^3$Net (Wang et al. 2025), and CYformer (Chen, Zhang, and Zhang 2025). Quantitative performance is assessed using four widely adopted metrics: PSNR, SSIM, SAM, and ERGAS. Additional experimental settings are provided in the Appendix.

### Comparisons with State-of-the-art Methods

Table 1 lists the numerical results of all competing methods under downsampling factors of 4 and 8. A detailed analysis is conducted for the case when $sf = 8$. For the Chikusei dataset, TRT generates the best results against other approaches in terms of all quantitative metrics. Specifically, the PSNR achieved by TRT is 42.7128 dB, marking an increase of approximately 1.86% over the CYformer method and 4.13% over MIMO, underscoring its superior capability in preserving high image quality. For the Pavia dataset, the proposed TRT exhibits notable strengths across various metrics, further establishing robustness and effectiveness. The improvement is particularly evident in PSNR, where TRT surpasses DSPNet, $\text{S}^2$CycleDiff, MIMO, DCINN, $\text{S}^3$-Net, and CYformer by 0.56 dB, 0.39 dB, 2.47 dB, 1.38 dB, 0.46 dB, and 0.42 dB, respectively. Reviewing the CAVE dataset results shows the proposed method performs exceptionally well on most metrics. Overall, the integration of the tensor ring Transformer and RPCA ensures stable convergence and strong reconstruction performance.

The visual comparison on the Chikusei dataset (Fig. 6) shows that all algorithms deliver strong fusion performance. Yet subtle differences remain: DSPNet, $\text{S}^2$CycleDiff, and MIMO produce slightly blurred contours and noticeable smoothing in vegetation regions. Their error-map zoom-ins also contain evident red artifacts. In contrast, our method presents the most extensive blue areas with vegetation contours closely aligned with the GT, reflecting superior restoration fidelity. For the CAVE *oil painting* scene (Fig. 7), the pseudo-color images suggest comparable overall performance. However, the error maps expose pronounced sunflower-petal contours in DSPNet and MIMO, indicating limited capability in recovering fine edges. Similar deficiencies appear in other baselines when reconstructing texture-rich petals, as seen in the magnified regions. By comparison, our approach yields minimal contour errors and better-preserved textures, demonstrating the enhanced detail-recovery ability of the tensor-ring–based Transformer.

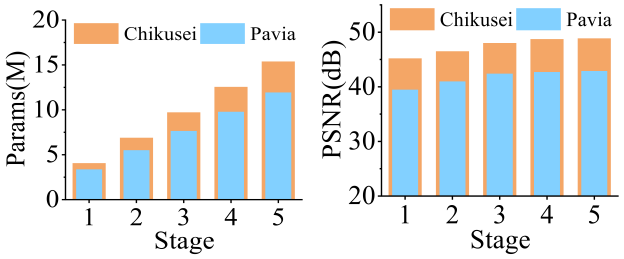


(a) DSPNet (b) S$^2$CycleDiff (c) MIMO (d) DCINN (e) S$^3$Net (f) CYformer (g) TRT (h) Ground truth

Figure 7: Visual comparisons on typical Cave *oil painting* data.

## Analysis and Discussion

**Effect of unfolding stages.** To examine performance variability as iteration count increases, Fig. 8(a) presents TRT outcomes across stages from 1 to 5. Thanks to the learning capabilities of the implicit prior extractor and the low-rank tensor ring Transformer, only a small number of iterations—typically 3 or 4—are needed to reach and stabilize performance. As depicted in Fig. 8(b), a higher iteration count naturally incurs increased costs. Consequently, setting $K = 4$ is advised as a default, with additional iterations optionally explored to further enhance results.



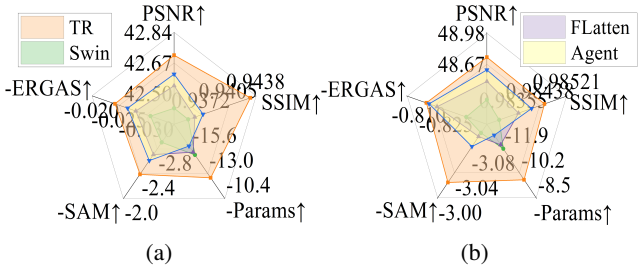

(a) (b)

Figure 8: Efficacy and efficiency of TRT with different numbers of stages on (a) Chikusei and (b) Pavia data sets.

| | PSNR↑ | SAM↓ | ERGAS↓ |
|---|---|---|---|
| 3×3 | 48.6395$_{\downarrow 0.21\%}$ | 3.1698$_{\uparrow 4.75\%}$ | 0.8280$_{\uparrow 1.73\%}$ |
| 5×5 | 48.6401$_{\downarrow 0.21\%}$ | 3.1621$_{\uparrow 4.50\%}$ | 0.8298$_{\uparrow 1.95\%}$ |
| 3×&5× | 48.6980$_{\downarrow 0.09\%}$ | 3.1241$_{\uparrow 3.24\%}$ | 0.8207$_{\uparrow 0.84\%}$ |
| **Ours** | **48.7409** | **3.0260** | **0.8139** |

Table 2: Ablation results on the RPCA paradigm.

**Effect of Tensor Ring Transformer Module.** To assess the effectiveness of the proposed TR-Transformer, the ablation results in Fig. 9 for the Chikusei and Pavia datasets with $sf = 8$ reveal two key observations. First, compared to the conventional Swin-Transformer, TR-Transformer demonstrates superior capability in capturing features, leading to improvements across all metrics. Second, when compared to models utilizing linear attention mechanisms, such as FLatten (Han et al. 2023) and Agent (Han et al. 2024) Transformer, TR-Transformer achieves a reduction in computational complexity without compromising—and indeed enhancing—performance.




(a) (b)

Figure 9: Efficacy and efficiency of different Transformer variants on (a) Chikusei and (b) Pavia data sets.

**Effectiveness of RPCA style transfer.** This ablation compares our RPCA framework with a deep unfolding variant that preserves the degraded spatial matrix. We use different convolution schemes (3×3, 5×5, and 3×&5× multi-scale) to simulate matrices $\boldsymbol{B}$ and $\boldsymbol{B}^{\mathrm{T}}$. As reported in Table 2, our sparse noise–reduction model achieves the best performance across all metrics. This gain mainly stems from its ability to remove complex spatial degradation while keeping a simpler, lower-dimensional spectral matrix, thereby reducing simulation loss. Additionally, the strong sparsification effect of $l_1$-norm further improves robustness and stability.

*Due to page constraints, the parameter count and tensor ring parameter analysis have been moved to the appendix.*

## CONCLUSION

In this work, we propose a novel approach to HSISR by recasting it as an RPCA-based denoising problem, thereby circumventing the challenges associated with spatial degradation matrix approximation. Besides, the elaborated Tensor Ring Transformer prior leverages multilinear tensor products to significantly reduce the computational burden of traditional Transformer, while preserving the expressiveness. By integrating both Softmax and linear attention mechanisms, we achieve an efficient balance between interpretability and deep learning efficacy. The experimental results validate the final effectiveness, achieving state-of-the-art performance in both image quality and computational efficiency across diverse datasets.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62276232, 62406285, and 62576319; in part by the Key Program of the Natural Science Foundation of Zhejiang Province under Grant LZ24F030012; in part by the Food Science and Engineering, the Most Important Discipline of Zhejiang Province under Grant ZCLY24F0301; and in part by the Natural Science Foundation of Ningbo Municipality under Grants 2023J403.

## References

Chen, S.; Zhang, L.; and Zhang, L. 2025. Cyclic Cross-Modality Interaction for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1): 741–753.

Dian, R.; and Li, S. 2019. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. *IEEE Transactions on Image Processing*, 28(10): 5135–5146.

Dian, R.; Liu, Y.; and Li, S. 2025. Hyperspectral Image Fusion via a Novel Generalized Tensor Nuclear Norm Regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 7437–7448.

Fang, J.; Yang, J.; Khader, A.; and Xiao, L. 2024. MIMO-SST: Multi-Input Multi-Output Spatial-Spectral Transformer for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–20.

Gu, Y.; Xu, H.; Quan, Y.; Chen, W.; and Zheng, J. 2023. Orsi salient object detection via bidimensional attention and full-stage semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Han, D.; Pan, X.; Han, Y.; Song, S.; and Huang, G. 2023. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5961–5971.

Han, D.; Ye, T.; Han, Y.; Xia, Z.; Pan, S.; Wan, P.; Song, S.; and Huang, G. 2024. Agent attention: On the integration of softmax and linear attention. In *European Conference on Computer Vision*, 124–140. Springer.

Hou, J.; Liu, X.; Wang, H.; and Guo, K. 2024. Tensor recovery from binary measurements fused low-rankness and smoothness. *Signal Processing*, 221: 109480.

Jian, P.; Ou, Y.; and Chen, K. 2024. Uncertainty-Aware Graph Self-Supervised Learning for Hyperspectral Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–19.

Jiang, J.; Chen, J.; Xu, H.; He, Y.; and Zheng, J. 2025. Information-coupled MRI acceleration via multi-modal mapping and progressive masking. *Pattern Recognition*, 112061.

Jiang, J.; Xu, Y.; Xu, H.; Shen, G.; and Zheng, J. 2024. Multi-dimensional visual data completion via weighted hybrid graph-Laplacian. *Signal Processing*, 216: 109305.

Lu, J.; Yao, J.; Zhang, J.; Zhu, X.; Xu, H.; Gao, W.; Xu, C.; Xiang, T.; and Zhang, L. 2021. Soft: Softmax-free transformer with linear complexity. *Advances in Neural Information Processing Systems*, 34: 21297–21309.

Ma, Q.; Jiang, J.; Liu, X.; and Ma, J. 2024. Reciprocal transformer for hyperspectral and multispectral image fusion. *Information Fusion*, 104: 102148.

Qu, J.; He, J.; Dong, W.; and Zhao, J. 2024. S2cyclediff: Spatial-spectral-bilateral cycle-diffusion framework for hyperspectral image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4623–4631.

Ren, T.; Shen, Q.; Fu, Y.; and You, S. 2024. Point-Supervised Semantic Segmentation of Natural Scenes via Hyperspectral Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1357–1367.

Sun, Y.; Xu, H.; Ma, Y.; Wu, M.; Mei, X.; Huang, J.; and Ma, J. 2023. Dual Spatial–Spectral Pyramid Network With Transformer for Hyperspectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Wang, W.; Deng, L.-J.; Ran, R.; and Vivone, G. 2024. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4): 1029–1054.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, X.; Huang, Z.; Zhu, J.; Wang, X.; and Feng, L. 2025. S3-Net: Learning spectral-spatio self-similarity for hyperspectral image super-resolution. *Neural Networks*, 107490.

Xu, H.; Fang, C.; Ge, Y.; Gu, Y.; and Zheng, J. 2024a. Cascade-Transform-Based Tensor Nuclear Norm for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Xu, H.; Fang, C.; Wang, R.; Chen, S.; and Zheng, J. 2024b. Dual-Enhanced High-Order Self-Learning Tensor Singular Value Decomposition for Robust Principal Component Analysis. *IEEE Transactions on Artificial Intelligence*, 5(7): 3564–3578.

Xu, H.; Fang, C.; Wang, Y.; Wu, J.; and Zheng, J. 2025a. Laboring on less labors: RPCA Paradigm for Pan-sharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11393–11402.

Xu, H.; Li, Y.; Jia, Y.; Fang, C.; Chen, W.; and Zheng, J. 2025b. Collaborative Cross-Complementary Unfolding Network for Pan-sharpening Remote Sensing Image. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 1607–1616.

Xu, H.; Qin, M.; Chen, S.; Zheng, Y.; and Zheng, J. 2021. 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.

Xu, H.; Quan, Y.; Qin, M.; Wang, Y.; Fang, C.; Li, Y.; and Zheng, J. 2025c. Nonlinear Learnable Triple-Domain Transform Tensor Nuclear Norm for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–17.

Xu, H.; Zheng, J.; Yao, X.; Feng, Y.; and Chen, S. 2022. Fast tensor nuclear norm for structured low-rank visual inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2): 538–552.

Zhu, C.; Zhang, T.; Wu, Q.; Li, Y.; and Zhong, Q. 2024. An Implicit Transformer-based Fusion Method for Hyperspectral and Multispectral Remote Sensing Image. *International Journal of Applied Earth Observation and Geoinformation*, 131: 103955.