DSTrans: Dual-Stream Transformer for Hyperspectral Image Restoration

Dabing Yu, Qingwu Li, Xiaolin Wang, Zhiliang Zhang, Yixi Qian and Chang Xu
Hohai university
{yudadabing, zhangzl, 211620010037, xuchang}@hhu.edu.cn, {li_qingwu, xlwang1998}@163.com

Abstract

Most CNN models exhibit two major flaws in hyperspectral image (HSI) restoration tasks. First, limited high-dimensional HSI training examples exacerbate the difficulty of deep learning methods in learning effective spatial and spectral representations. Second, the existing CNN-based methods model local relations and present limitations in capturing long-range dependencies. In this paper, we customize a novel dual-stream Transformer (DSTrans) for HSI restoration, which mainly consists of the dual-stream attention and the dual-stream feed-forward network. Specifically, we develop the dual-stream attention consisting of Multi-Dconv-head spectral attention (MDSA) and Multi-head Spatial self-attention (MSSA). MDSA and MSSA respectively calculate self-attention along the spectral and spatial dimensions in local windows to capture long-range spectrum dependencies and model global spatial interactions. Meanwhile, the dual-stream feed-forward network is developed to extract global signals and local details in parallel branches. In addition, we exploit a multi-tasking network to train the auxiliary RGB image (RGBI) task and HSI task jointly so that both numerous RGBI samples and limited HSI samples are exploited to learn parameter distribution for DSTrans. Extensive experimental results demonstrate that our method achieves state-of-the-art results on HSI restoration tasks, including HSI super-resolution and denoising. The source code can be obtained at: https://github.com/yudadabing/Dual-Stream-Transformer-for-Hyperspectral-Image-Restoration.

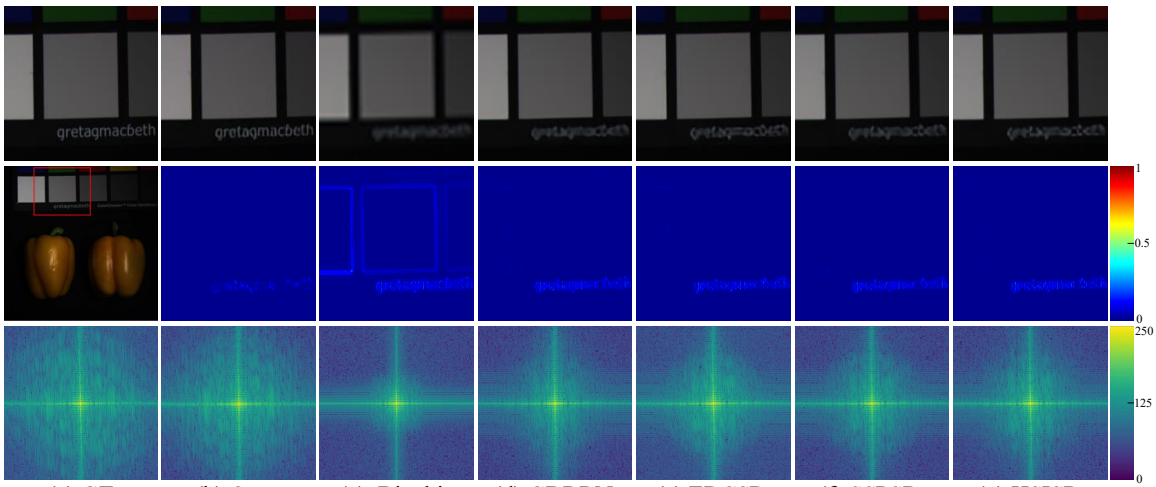
1. Introduction

Hyperspectral image (HSI) collects rich and detailed spectral information, effectively reflecting the subtle spectral difference of different objects. Relying on this contribution, the hyperspectral image has been widely promoted in a variety of tasks, *e.g.,* land-cover classification [21], target detection [70], mineral exploration [44], environmental monitoring [41] and medical diagnosis [25].

Nevertheless, mainly due to the physical limitations of spectral sensors, it is inevitable to gather degraded hyperspectral images. First, the generated hyperspectral image has a low spatial resolution, which is a trade-off result between spatial resolution and spectral resolution. The hyperspectral image sensor has to sacrifice spatial resolution to obtain a high spectral resolution with abundant spectral information [64], [14]. Second, the hyperspectral imaging systems scan the object scenes along the spatial or spectral dimension for a long time, inevitably introducing numerous noises [9]. These degradations bring the negative influence on the subsequent hyperspectral image interpretation [46]. Hyperspectral image restoration is a postprocessing technique, such as HSI super-resolution (SR) and HSI denoising, which aims to model the ill-posed problem and generate a high-quality HSI from its degraded counterpart without hardware sensors modification.

As a learning-based approach, Transformer has confirmed its superior performance on natural language tasks [12, 45] and computer vision tasks [8, 52, 15]. Transformer relies on a self-attention (SA) mechanism to model the global contextual information and has the potential to relieve the aforementioned limitations of CNN-based methods in HSI restoration. Recently, Transformer has been used for image restoration tasks [32, 68, 56]. However, these Transformers are just tailored for RGB image, while there is less attention on hyperspectral images. The main reasons are twofold. Firstly, there is a lack of large-scale HSI datasets with high-resolution (HR) and high-quality HSI samples. Generally speaking, Transformer exploits enormous amounts of training data to learn the data distribution and feature presentation. A limited amount of training examples exacerbate the undesirable behaviors, such as memorization and sensitivity to out-of-distribution samples. Secondly, traditional Transformers have an advantage in capturing the long-range dependencies in global spatial locations. In this case, directly applying the Transformer can capture spatial interactions but not model the inter-spectra similarity and correlations. However, global spectral information and global spatial information are equally important for HSI restoration.

To cope with the aforementioned challenges, we propose



(a) GT (b) Ours (c) Bicubic (d) GDRRN (e) ERCSR (f) SSPSR (g) HSISR

Figure 1. Qualitative results for $\times 4$ image SR on *real_and_fake_peppers_ms* from CAVE dataset [65]. From top to down are visual results, error maps and frequency visualization results.

dual-stream Transformer (DSTrans) for hyperspectral image restoration. Firstly, restricted to the craftsmanship of hardware, the limited HSI training samples will not be addressed in the foreseeable future. Inspired by [28], in our work, we choose a distinct route to increase training data, which selects the numerous training samples from heterogeneous datasets to auxiliarily train the Transformer model. The HSIs restoration and RGBIs restoration learning networks share the same goal of integrating information from neighboring spatial regions and spectral bands. We formulate both tasks into the same Transformer such that abundant training samples can effectively regularize parameters and achieve excellent performance. Secondly, we propose dual-stream attention that is capable of modeling global pixels connectivity and global spectra correlation. Specifically, dual-stream attention is consist of multi-Dconv-head spectral attention (MDSA) and multi-head spatial self-attention (MSSA) lying on shifted windows. The spatially global context is learned by MSSA. Importantly, MDSA ensures that the contextualized global relationships between spectra are modeled while computing covariance-based channel maps.

We visualize the visual results, error maps, and frequency maps of reconstructed HSIs in Fig. 1. It can be seen that our DSTrans keeps the most significant visual result and error map and relieves the frequency domain discrepancy between reconstructed result and ground truth. Our contributions are summarized as follows:

1. We propose a novel DSTrans, which is a tailored Transformer for HSI restoration. To the best of our knowledge, it is the first attempt to explore the potential of Transformer in HSI restoration. Besides HSI samples, DSTrans exploits the numerous samples from heterogeneous datasets to learn the parameters distribution of DSTrans.

2. We present a novel attention mechanism, dual-stream attention, to capture global pixels and inter-spectra similarity and dependencies of HSIs in two parallel branches. Besides, we propose the dual-stream feed-forward network to extract the global signals and local details simultaneously.

3. Extensive experiments verify that our DSTrans greatly outperforms SOTA methods in terms of HSI denoise and HSI SR tasks on multiple HSI datasets.

2. Related Work

Hyperspectral Image Super-Resolution. The hyperspectral image super-resolution methods can be roughly divided into fusion-based super-resolution [20, 61, 35, 73, 13, 63, 71, 58, 11] and single super-resolution [75, 27, 24, 19, 28, 53]. Fusformer [20] is first time using the transformer to solve the hyperspectral image fusion-based super-resolution problem. The drawback of fusion-based super-resolution is the need for an well-co-registered auxiliary image with higher resolution. Therefore, single super-resolution is more popular in real scenes.

Benefit from the superior performance in many computer vision fields, deep learning method has been introduced into single HSI super-resolution task. Deep neural network learns to directly map an input low resolution HSI to a high resolution HSI, which can reduce the spectral distortion and ultimately improve the resolution performance [31, 62]. Jia *et al*. [23] proposed spectral–spatial network that joint spectral and spatial properties to effectively increase spatial res-

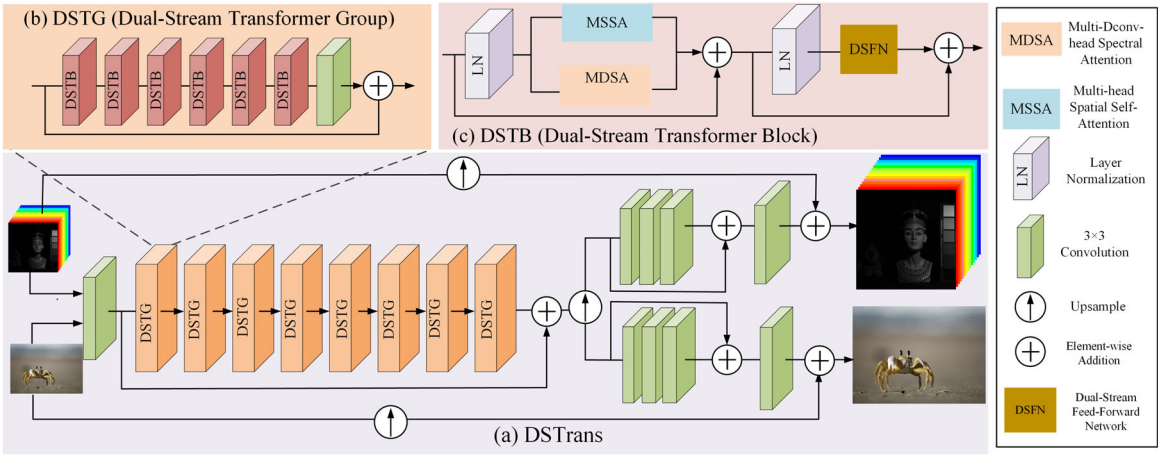


Figure 2. Network architecture of our DSTrans. DSTrans adopts the structure of the double tasks on heterogeneous datasets. Our DSTrans mainly consists of *residual in residual* design incorporating dual-stream Transformer blocks.

olution and keep spectral information. Nevertheless, 2-D convolution only focuses on the spatial information of HSIs. Recently, 3D recurrent neural network [51, 16] and mixed 2D/3D convolutional networks [29], [30] are designed to extract spatial–spectral features simultaneously. However, these methods focus on exploring local spatial–spectral information and neglect the global features of HSIs.

Hyperspectral Image Denoising. Hyperspectral image denoising task was addressed early as band-wise image denoising problem, *e.g.* BM3D [10], WNNM [17]. Due to the ignoration of the spectral continuous features, these methods generate denoised results with distortions and artifacts in the spectral domain. Recently, neural network-based methods has introduced to hyperspectral image denoising task [66, 39, 69, 42, 48]. Wei *et al.* [57] proposed an alternating directional 3D quasi-recurrent neural network to embed the structural spatial-spectral correlation and global correlation along spectrum. In the work of [47], the dual-attention denoising network is proposed to consider the global dependence between spatial and spectral information. Cao *et al.* [4] considered both the local and global information for HSI noise removal.

Vision Transformer. Transformer was first proposed by Vaswani *et al.* [49] for natural language processing (NLP). Transformer has achieved significant breakthroughs with their strong representation capacity. Recent years, Transformer has been expanded to numerous computer vision tasks and has been an effective alternative to CNN in the vision applications, *e.g.* image recognition [15], segmentation [52], object detection [5]. Transformer also has been developed to address the low-level vision tasks, such as image restoration [56, 33, 22, 8, 38]. Liang *et al.* [32] proposed the SwinIR model for image restoration based on the Swin Transformer to apply self-attention within local image regions. Restormer [68] built multi-head attention and feed-forward network to capture long-range pixel interactions and achieve excellent results.

3. Proposed Method

3.1. Overall Pipeline

Data Alignment. As shown in Fig. 2, our DSTrans learns two same restoration tasks on heterogeneous datasets together. Particularly, in this paper, we select the RGBI dataset as the auxiliary dataset. Given a HSI dataset $\Omega_{HSI} = \{x^i_{HSI}, X^i_{HSI}\}_{i=1}^{N_{HSI}}$ and RGBI dataset $\Omega_{RGBI} = \{x^i_{RGBI}, X^i_{RGBI}\}_{i=1}^{N_{RGBI}}$, where $x_{HSI} \in \mathbb{R}^{h\times w\times D}$ presents the degraded HSI, $X_{HSI} \in \mathbb{R}^{H\times W\times D}$ presents the high-quality HSI counterpart. Resembleily, $x_{RGBI} \in \mathbb{R}^{h\times w\times 3}$ is the degraded RGB image and $X_{RGBI} \in \mathbb{R}^{H\times W\times 3}$ is the high-quality counterpart. h, w, H and W stand for the width and height of the degraded image and desired image, D is the number of bands of HSI. For HSI SR, we have $H = \lambda h$, $W = \lambda w$, and λ is the scaling factor. For HSI denoise, λ is set to 1. N_{HSI} and N_{RGBI} are the number of HSI and RGBI samples. We attempt to exploit the knowledge from RGBI dataset, which means RGBI dataset provides numerous high-quality samples. Thus, we have $N_{RGBI} = vN_{HSI}$, and $v \geq 1$.

Inspired by [24], we divide each HSI input into samples with overlapping groups of bands. More specifically, we divide the D bands of HSI into groups of S bands. For RGBI samples, we increase the channels to S via the spectral band interpolation strategy [28]. So the generated RGBI dataset $\bar{\Omega}_{RGBI} = \{\bar{x}^i_{RGBI}, \bar{X}^i_{RGBI}\}_{i=1}^{N_{RGBI}}$ and HSI dataset $\bar{\Omega}_{HSI} = \{\bar{x}^i_{HSI}, \bar{X}^i_{HSI}\}_{i=1}^{N_{HSI}}$ have similar format, where $\bar{x}_{RGBI} \in \mathbb{R}^{h\times w\times S}$, $\bar{X}_{RGBI} \in \mathbb{R}^{H\times W\times S}$ and $\bar{x}_{HSI} \in \mathbb{R}^{h\times w\times S}$, $\bar{X}_{HSI} \in \mathbb{R}^{H\times W\times S}$.

Feature Extraction. Give the degraded HSI input $\bar{x}_{HSI}$

and RGBI input $\bar{x}_{RGBI}$, our DSTrans first applied a convolutional layer to extract shallow feature maps $F_{SF}^{HSI} \in \mathbb{R}^{h\times w\times L}$ and $F_{SF}^{RGBI} \in \mathbb{R}^{h\times w\times L}$,

$$\left(F_{SF}^{HSI}, F_{SF}^{RGBI}\right) = H_{SF}\left(\bar{x}_{HSI}, \bar{x}_{RGBI}\right) \tag{1}$$

where L is the number of channels of the shallow feature. $H_{SF}(\cdot)$ is the 3×3 convolutional layer that maps the input image to a high-dimensional feature space. Then the shallow features are transport to the shared encoder Φ_{EN} to extract deep features $F_{DF}^{HSI} \in \mathbb{R}^{h\times w\times L}$ and $F_{DF}^{RGBI} \in \mathbb{R}^{h\times w\times L}$

$$\left(F_{DF}^{HSI}, F_{DF}^{RGBI}\right) = \Phi_{EN}\left(F_{SF}^{HSI}, F_{SF}^{RGBI}\right) \tag{2}$$

where $\Phi_{EN}(\cdot)$ is consists of *residual in residual* design incorporating dual-stream Transformer blocks.

Image Reconstruction. Then the aggregating deep feature, shallow feature and degraded image are mapping to desired high-quality output. Naturally, there are two branches matching the HSI restoration task and the RGBI restoration task.

We take the HSI SR task as an example to describe the process. Before passing to the residual enhancing module, we exploit the concatenation operation $Cat(\cdot)$ to concatenate the extracted features of all the groups of $\bar{x}_{HSI}$ based on their original spectral band position,

$$I_{ER}^{HSI} = \Phi_{REM}^{HSI}\left(Cat\left(H_{\uparrow}\left(F_{DF}^{HSI} + F_{SF}^{HSI}\right)\right)\right), \tag{3}$$

$$I_{ER}^{RGBI} = \Phi_{REM}^{RGBI}\left(H_{\uparrow}\left(F_{DF}^{RGBI} + F_{SF}^{RGBI}\right)\right), \tag{4}$$

where $\Phi_{REM}^{HSI}(\cdot)$ and $\Phi_{REM}^{RGBI}(\cdot)$ are the residual enhancing modules for HSI and RGBI tasks and keep the consistent structure. $H_{\uparrow}(\cdot)$ denotes an upscale module, in this paper, we upsample the aggregated features exploiting the operation of sub-pixel convolution [24]. The residual enhancing module contains three 3×3 convolutional layers and the residual connection.

Loss Function. We combine the L_1 loss and the spatial-spectral total variation (SSTV) loss [24] to optimize the parameters of DSTrans. **More details are presented in the supplementary material**.

3.2. Dual-Stream Attention

Locally finding similar external patches is exploited in HSI restoration by CNN-based methods, but they have ignored the long-range feature-wise similarities in HSIs. Recently, Transformer has achieved impressive performance benefiting from the capability to capture long-range dependencies. Unlike natural images, HSIs have numerous narrow bands. Capturing long-range spectrum dependencies and modeling global spatial interactions are equally essential. Hence, we propose dual-stream attention, which consists of Multi-head Spatial Self-Attention (MSSA) and Multi-Dconv-head Spectral Attention (MDSA), to model long-range dependencies in spatial and spectral dimensions, respectively.

As shown in Fig. 3, we follow [36] and apply dual-stream attention to the shifted window to reduce the computing burden. Given an input $X_{in} \in \mathbb{R}^{h\times w\times C}$. Dual-Stream Transformer partitions the input into non-overlapping local windows $X_t \in \mathbb{R}^{M\times M\times C}$, $t \in \left[1, \frac{hw}{M^2}\right]$. Then, it computes the dual-stream attention separately for each window.

Multi-Dconv-head Spectral Attention. Multi-Dconv-head spectral attention is intent on applying self-attention across spectral channels. As shown in Fig. 3(b) , MDSA computes cross-covariance across channels to generate an attention map encoding the global spectral signal. X_t is first projected and reshaped into $query\ Q_{spe} \in \mathbb{R}^{C\times M^2}$, $key\ K_{spe} \in \mathbb{R}^{C\times M^2}$ and $value\ V_{spe} \in \mathbb{R}^{C\times M^2}$ by applying 1×1 point-wise convolutions W_P followed by 3×3 depth-wise convolutions W_D to encode spectral-wise spatial context,

$$\begin{aligned} &Q_{spe} = W_P^Q W_D^Q X_t, K_{spe} = W_P^K W_D^K X_t, \\ &V_{spe} = W_P^V W_D^V X_t. \end{aligned} \tag{5}$$

Next, the spectral attention map is computed by the self-attention mechanism in a local window. We apply dot-product interaction on Q_{spe} and K_{spe} to generate the spectral attention map $A_{spe} \in \mathbb{R}^{C\times C}$,

$$A_{spe} = Softmax\left(\frac{Q_{spe}\cdot K_{spe}}{\varepsilon} + B\right), \tag{6}$$

$$Attention\left(Q_{spe}, K_{spe}, V_{spe}\right) = W_P \cdot V_{spe} \cdot A_{spe}, \tag{7}$$

where ε is a learnable parameter to reweight the dot product of Q_{spe} and K_{spe} before applying the softmax function. B is the learnable relative positional encoding.

Multi-head Spatial Self-Attention. MSSA aims to apply self-attention across global spatial location and generates an attention map modeling the long-range dependencies and spatial interactions. As illustrated in Fig. 3(c), in MSSA branch, X_t is first linearly projected into $query\ Q_{spa} \in \mathbb{R}^{M^2\times C}$, $key\ K_{spa} \in \mathbb{R}^{M^2\times C}$ and $value\ V_{spa} \in \mathbb{R}^{M^2\times C}$,

$$Q_{spa} = W^Q X_t, K_{spa} = W^K X_t, V_{spa} = W^V X_t, \tag{8}$$

where W^Q, W^K and $W^V \in \mathbb{R}^{C\times C}$ are learnable projection matrices that shared across local windows. The atten-

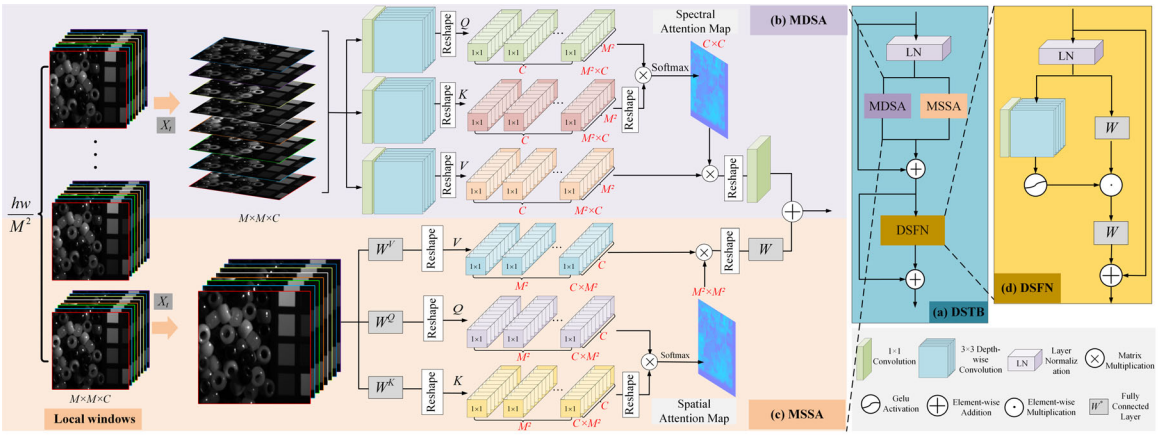


Figure 3. Illustration of the dual-stream Transformer blocks (DSTB). The core modules of **(a)** DSTB are Dual-Stream attention (DSA) and **(d)** Stream Feed-Forward Network (DSFN). DSA consisting of **(b)** Multi-Dconv-head Spectral Attention (MDSA) and **(c)** Multi-head Spatial Self-Attention (MSSA) tends to capture long-range spectrum dependencies and spatial interactions in parallel branches. DSFN performs controlled feature flow, *i.e.,* the detailed signal is activated and flows in the desired direction.

tion matrix is thus computed by the self-attention mechanism in a local window. We apply dot-product interaction on Q_{spa} and K_{spa} to generate the spatial attention map $A_{spa} \in \mathbb{R}^{M^2 \times M^2}$,

$$A_{spa} = Softmax\left(\frac{Q_{spa} \cdot K_{spa}}{\sqrt{C}} + B\right), \tag{9}$$

$$Attention\left(Q_{spa}, K_{spa}, V_{spa}\right) = W_{out} \cdot V_{spa} \cdot A_{spa}, \tag{10}$$

where $W_{out} \in \mathbb{R}^{C \times C}$ is also learnable projection matrices. Following multi-head SA [49], MDSA and MSSA divide the number of channels into 'heads', then perform the attention function for 'heads' times in parallel and concatenate the results for Multi-head results.

3.3. Dual-Stream Feed-Forward Network

In the traditional feed-forward network, the two fully connected layers are applied to expand the input feature channels and map the output channels back to the original input dimension. The fully connected layer operates token information identically point-wise; thus, it neglects the local information. In our work, we propose the dual-stream feed-forward network, which aims at complementing local information by encoding information from spatially neighboring pixel positions. As shown in Fig. 3(d), we extract the global signals and local details in two parallel paths. We exploit the fully connected layer to model the global feature information in the regular branch. The depth-wise convolution is added to complement the local details in the additional branch, followed by the GELU non-linearity to activate the local signal. Then, the gating mechanism is formulated as the element-wise product of outputs in two parallel paths. Given an input feature $\hat{x} \in \mathbb{R}^{h \times w \times O}$, DSFN is formulated as:

$$\begin{aligned} x^{'} &= W^{1}\left(LN\left(\hat{x}\right)\right) \odot H_{Gelu}\left(W_{P} W_{D} LN\left(\hat{x}\right)\right), \\ x^{''} &= \hat{x} + W^{2} x^{'}, \end{aligned} \tag{11}$$

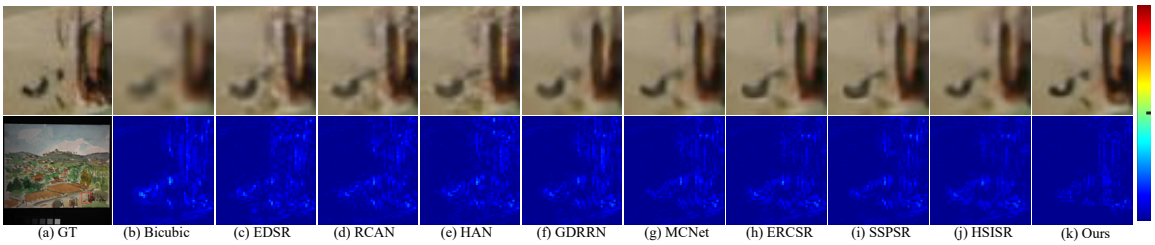
where $\odot$ denotes element-wise multiplication, H_{Gelu} represents the Gelu non-linearity, W^1 and W^2 denotes the fully connected layers. Overall, the DSFN controls the information flow through the activated local signal in our pipeline, thereby allowing each level to focus on the fine details.

4. Experiments and Analysis

4.1. Experimental Settings

Datasets. We evaluate our DSTrans on benchmark datasets and experimental settings for two HSI restoration tasks: HSI super-resolution and HSI denoising. The datasets considered are four nature HSI datasets: CAVE dataset [65], Harvard dataset [6], ICVL dataset [3] and HSIDwRD dataset [72]. HSI super-resolution experiments are conduct on CAVE and Harvard datasets and HSI denoising experiments are perform on ICVL (Gaussian denoising) and HSIDwRD datasets (Real-world denoising).

For SR task, we crop images to patches as 64×64 pixels with 32 pixels overlapping and patches as 128×128 pixels with 64 pixels overlapping for upsampling factors $\times 4$ and $\times 8$. The corresponding LR images are generated by Bicubic downsampling. For the auxiliary RGBI SR task, we adopt the DIV2K Dataset [1]. The training samples of DIV2K are about 30, and 12 times larger than CAVE [65] and Harvard [6], respectively. Especially, we extract image patches as 64×64 pixels with 32 pixels overlapping for denoise task. For Gaussian denoising, the RGBI denoise task



(a) GT (b) Bicubic (c) EDSR (d) RCAN (e) HAN (f) GDRRN (g) MCNet (h) ERCSR (i) SSPSR (j) HSISR (k) Ours

Figure 4. Visual comparison for HSI SR on the representative test image *watercolors_ms* from CAVE dataset with spectral bands 23-15-7 as R-G-B with the scale factor 4.

Scale	Method	CAVE [65] SAM ↓	CC ↑	ERGAS ↓	RMSE ↓	MPSNR ↑	MSSIM ↑	Harvard [6] SAM ↓	CC ↑	ERGAS↓	RMSE ↓	MPSNR ↑	MSSIM ↑
×4	Bicubic	4.176	0.9868	5.272	0.0212	34.721	0.9303	2.588	0.9758	3.871	0.0177	37.505	0.9122
	EDSR [34]	3.965	0.9926	3.738	0.0155	37.738	0.9519	2.527	0.9825	3.201	0.0146	39.183	0.9306
	RCAN [74]	4.010	0.9928	3.666	0.0289	37.952	0.9515	2.810	0.9803	3.467	0.0156	38.560	0.9240
	HAN [43]	3.726	0.9761	6.859	0.0147	38.457	0.9541	2.891	0.9790	3.611	0.0161	38.246	0.9215
	GDRRN [31]	3.726	0.9927	3.735	0.0155	37.687	0.9525	2.581	0.9807	3.369	0.0152	38.750	0.9267
	MCNet [29]	3.412	0.9843	4.222	0.0146	37.870	0.9540	2.558	0.9811	3.356	0.0147	38.924	0.9289
	ERCSR [30]	3.273	0.9847	4.153	0.0144	38.009	0.9553	2.530	0.9820	3.304	0.0147	38.992	0.9295
	SSPSR [24]	3.360	0.9930	3.543	0.0146	38.302	0.9566	2.474	0.9834	3.063	0.0142	39.484	0.9326
	HSISR [28]	3.319	0.9945	3.204	0.0131	39.060	0.9618	2.471	0.9837	3.056	0.0141	39.572	0.9340
	Ours	**3.169**	**0.9953**	**2.861**	**0.0118**	**40.073**	**0.9659**	**2.459**	**0.9846**	**3.007**	**0.0129**	**40.096**	**0.9359**
×8	Bicubic	5.896	0.9666	8.435	0.0346	30.206	0.8494	2.981	0.9533	5.606	0.0261	34.357	0.8534
	EDSR [34]	7.036	0.9764	6.887	0.0289	31.956	0.8746	3.425	0.9588	5.278	0.0242	34.847	0.8626
	RCAN [74]	7.288	0.9761	6.857	0.0289	32.015	0.8711	3.579	0.9585	5.298	0.0240	34.833	0.8620
	HAN [43]	6.429	0.9783	6.465	0.0275	32.635	0.8817	3.795	0.9567	5.422	0.0243	34.687	0.8593
	GDRRN [31]	5.858	0.9731	7.346	0.0307	31.430	0.8709	3.047	0.9608	5.080	0.0235	35.147	0.8666
	MCNet [29]	5.407	0.9695	3.573	0.0278	32.417	0.8873	2.892	0.9640	4.963	0.0232	35.309	0.8766
	ERCSR [30]	5.210	0.9630	3.440	0.0267	32.602	0.8901	2.884	0.9683	4.957	0.0232	35.391	0.8801
	SSPSR [24]	4.722	0.9800	6.050	0.0257	33.217	0.8936	2.853	0.9753	4.760	0.0233	35.613	0.8867
	HSISR [28]	5.108	0.9821	5.971	0.0246	34.096	0.9101	2.829	0.9667	4.625	0.0219	35.856	0.8901
	Ours	**4.623**	**0.9831**	**5.723**	**0.0239**	**34.797**	**0.9167**	**2.746**	**0.9787**	**4.430**	**0.0202**	**36.537**	**0.9003**

Table 1. Quantitative evaluation on CAVE and Harvard datasets of state-of-the-art SR methods by SAM, CC, ERGAS, RMSE, MPSNR, and MSSIM for scaling factors 4 and 8. Best results are **highlighted**.

	GDRRN [31]	MCNet [29]	ERCSR [30]	SSPSR [24]	HSISR [28]	Ours
×4	0.4M	17M	12.5M	6M	8.7M	12.2M
×8	0.8M	23.5M	16.5M	7.6M	9.9M	14.4M

Table 2. Comparison of the number of parameters of state-of-the-art SR methods.

is performed on DIV2K by adding Gaussian noise, the training samples of DIV2K are about 10 times larger than ICVL [3]. For real HSI denoising, the auxiliary RGBI denoise task is performed on RENOIR [2], the training samples of RENOIR are about 20 times larger than HSIDwRD dataset.

Experimental Parameters. For SR task, the RSTG number, DSTB number, window size, attention head number are generally set to 8, 6, 6, and 6, respectively. The RSTG number is set to 6 for HSI denoising. We use ADAM optimizer and the initial learning rate is set to 10^{-4}. The batch size is set to 12 and the epoch is set to 20.

Metrics. We evaluate the performance of all methods qualitatively by six standard metrics: spectral angle mapper (SAM)[67], cross correlation (CC)[37], erreur relative globale adimensionnelle de synthese (ERGAS)[50], root mean squared error (RMSE), mean peak signal-to-noise ratio (MPSNR), and mean structure similarity (MSSIM)[55].

4.2. HSI Super-Resolution Results

We compare our DSTrans with state-of-the-art methods: EDSR[34], RCAN[74], HAN[43], GDRNN[31], MCNet[29], ERCSR[30], SSPSR[24], HSISR[28]. The quantitative results, including SAM, CC, ERGAS, RMSE, MPSNR, and MSSIM, are respectively listed in Table 1, in which the best results are highlighted. Table 1 indicates that our DSTrans achieves significant performance gains over existing approaches on CAVE and Harvard datasets in terms of all evaluation metrics. Compared to the recent best method HSISR, DSTrans achieves 1.013dB and 0.701dB improvement on CAVE dataset. In order to intuitively show the performance of our method, we further exhibit the qualitative results of different SR methods. The visual results and error maps with upsampling factor $\times 4$ are shown in Fig. 4. Our method has a great performance in constructing edges and structures than those of the other algorithms.

We also compare the number of parameters between our DSTrans and state-of-the-art super-resolution algorithms. Table 2 shows the number of parameters, where the results

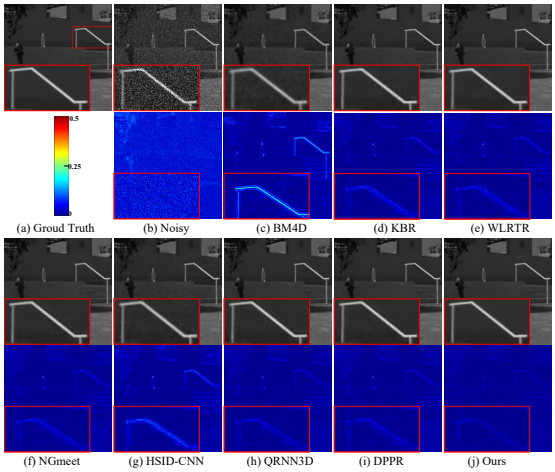


Figure 5. Denoising results and error maps at the *20th* band of image *eve_0331-1549* under Gaussian noise.

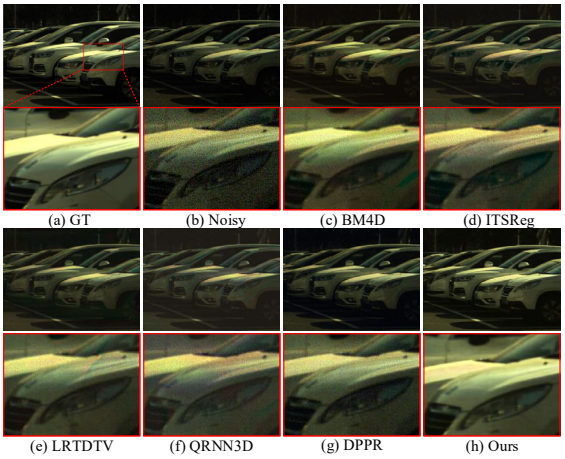


Figure 6. Denoising results of *image 46* under real-world noise with spectral bands 23-15-7 as R-G-B.

are evaluated for $4\times$ and $8\times$ upscaling factors. The results demonstrate that our DSTrans has a better tradeoff between model size and performance.

4.3. HSI Denoising Results

Gaussian denoising. To demonstrate the equally superior performance on HSI denoise, we compare our DSTrans with six state-of-the-art HSI denoising algorithms, including three traditional methods, BM4D [40], KBR [59], WLRTR [7], and NGmeet [18], and three recently developed deep learning methods, including HSID-CNN [66], QRNN3D [57] and DPPR [26].

The HSI denoising results under different noise levels for ICVL dataset are presented in Table 3. We evaluate the performance of our DSTrans on simulated Gaussian noise. Following [26], additive Gaussian white noise is added to each input HSI with different strengths, including 30, 50, 70, and random strengths ranging from 30 to 70. As one can see, the proposed DSTrans outperforms most of the competing methods in terms of MPSNR, SAM, and MSSIM at all noise levels. The visual results and error maps are presented in Fig. 5, "Noisy" is obtained by adding the additive Gaussian white noise with noise levels 50. It is evident that our method is superior to the other methods, which restores more details and achieves pleasing results.

Real-world denoising. Table 4 shows the quantitative comparisons between DSTrans and state-of-the-art HSI denoising algorithms: BM4D [40], ITSReg [60], LRTDTV [54], QRNN3D [57] and DPPR [26]. As we can see, the proposed DSTrans has MPSNR gains of 1.7dB compared to the recent best method DPPR. This is benefited from the proposed Transformer structure and modeling long-range spectrum and spatial dependencies and the learning knowledge from heterogeneous datasets. It can be seen in Fig. 6 that our DSTrans removes heavy noise corruption and generates clean HSI without compromising fine texture.

4.4. Ablation Studies

To validate the effectiveness of proposed components, we perform ablation experiments to analyze the contribution of all components and the parameter choices. For ablation experiments, we train the HSI SR model with scaling factor $\times 4$ on CAVE dataset.

Discussion on Auxiliary Task. In our DSTrans, we train the HSI restoration task and RGBI restoration task together. Both tasks share the same dual-stream Transformer to encode features such that there are an enormous amount of training samples to learn the parameter distribution. To verify the effectiveness of this strategy, we first remove the RGBI restoration task. Meanwhile, we further perform the ablation study on the number of RGBI samples. The quantity of HSI samples is limited, and we gradually increase the number of RGBI samples. As shown in Table 5, where v represents the ratio of the number of RGBI samples to HSI samples. "w/o RGBI" denotes that the RGBI samples are discarded, thus the quantitative results get worse. With the introduction of RGBI samples, the auxiliary RGBI SR strategy leads to a significant performance improvement. As expected, the quantitative results gradually improved as the sample number increased, and the performance gain becomes saturated gradually.

Discussion on the DSA and DSFN. We conduct an ablation study of different self-attention mechanisms and feed-forward networks. For self-attention mechanisms, we compare our DSA with vanilla MSA [15]. We further analyze the performance of two key components, MSSA and MDSA. Table 6(d) demonstrates that our DSA provides a favorable gain of 0.62 dB over the baseline. Furthermore, a single MSSA or MDSA brings expected improvement (see

Sigma	Metric	Method								
		Noise	BM4D [40]	KBR [59]	WLRTR [7]	NGmeet [18]	HSID-CNN [66]	QRNN3D [57]	DPPR [26]	Ours
30	MPSNR ↑	18.589	38.451	41.478	42.622	42.988	38.704	42.217	43.056	**43.534**
	MSSIM ↑	0.1100	0.9341	0.9840	0.9878	0.9889	0.9493	0.9883	0.9900	**0.9934**
	SAM ↓	0.807	0.126	0.088	0.056	0.050	0.103	0.062	0.052	**0.047**
50	MPSNR ↑	14.154	35.641	39.156	39.722	40.260	36.167	40.151	40.911	**41.360**
	MSSIM ↑	0.0462	0.8890	0.9743	0.9781	0.9784	0.9189	0.9820	0.9843	**0.9889**
	SAM ↓	0.991	0.169	0.101	0.073	0.059	0.134	0.074	0.059	**0.056**
70	MPSNR ↑	11.231	33.677	36.714	37.520	38.656	34.312	38.303	38.817	**39.667**
	MSSIM ↑	0.0254	0.8450	0.9605	0.9667	0.9743	0.8856	0.9742	0.9763	**0.9789**
	SAM ↓	1.105	0.207	0.113	0.095	0.067	0.161	0.093	0.087	**0.081**
[30, 70]	MPSNR ↑	17.338	37.662	40.681	41.664	42.230	37.811	41.369	42.231	**42.589**
	MSSIM ↑	0.1144	0.9141	0.9790	0.9825	0.9852	0.9350	0.9847	0.9873	**0.9914**
	SAM ↓	0.859	0.143	0.087	0.064	0.053	0.116	0.068	0.056	**0.049**

Table 3. Quantitative evaluation results of state-of-the-art denoise methods on ICVL dataset. Best results are **highlighted**.

Method	MPSNR ↑	MSSIM ↑	SAM ↓
Noise	20.907	0.3186	25.299
BM4D [40]	25.318	0.8156	6.302
ITSReg [60]	25.460	0.8400	5.143
LRTDTV [54]	25.564	0.7859	6.488
QRNN3D [57]	23.832	0.791	10.019
DPPR [26]	25.879	0.5244	16.398
Ours	**27.642**	**0.8406**	**4.629**

Table 4. Quantitative evaluation results of state-of-the-art denoise methods on HSIDwRD. Best results are **highlighted**.

	v	SAM ↓	MSSIM ↑	MPSNR ↑
w/o RGBI	$v = 0$	3.471	0.9587	38.921
	$v = 10$	3.332	0.9612	39.587
with RGBI	$v = 20$	3.181	0.9649	39.921
	$v = 30$	3.169	0.9659	40.073

Table 5. Ablation study of the auxiliary task. We add or remove the RGBI samples to modify the ratio v of the number of RGBI samples to HSI samples.

Network	Component	SAM ↓	MSSIM ↑	MPSNR ↑
Multi-head attention	(a) MSA+FN	3.327	0.9483	39.273
	(b) MSSA+FN	3.232	0.9620	39.683
	(c) MDSA+FN	3.235	0.9611	39.632
	(d) DSA+FN	3.186	0.9639	39.893
Feed-forward network	(e) MSA+DSFN	3.252	0.9561	39.497
	(f) DSA+DSFN	3.169	0.9659	40.073

Table 6. Ablation study of different self-attention mechanisms and feed-forward networks.

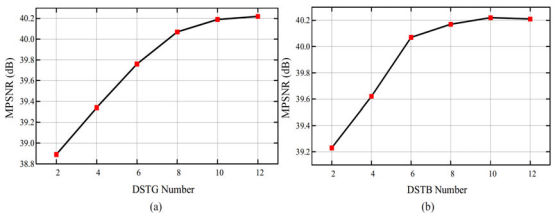


Figure 7. Ablation study on different settings of DSTB number and DSTG number.

Table 6(b) and Table 6(c), and the combination of both is the optimal choice. Experimental results confirm that the proposed DSA capture long-range dependencies in spatial and spectral dimensions. For feed-forward networks, we compare our proposed DSFN with the standard FN [49]. Table 6(e)shows that introducing local mechanisms to FN also brings performance advantages. Our DSFN also brings a MPSNR gain of 0.19 dB over the standard FN (see Table 6(f) for DSA. Overall, our Dual-stream Transformer block contributions lead to a significant gain of 0.80 dB over the standard Transformer block.

Discussion on DSTB number and DSTG number. We show the effects of RSTG number and RSTB number on model performance in Fig. 7(a) and Fig. 7(b). It can be observed that the MPSNR is positively correlated with RSTB number and RSTG number. As for RSTB number and RSTG number, the performance gain becomes saturated gradually. To balance the performance and model size, the RSTG number and RSTB number are set to 8 and 6 to obtain a relatively effective and small model.

5. Conclusion

In this work, we have customized a hyperspectral image restoration Transformer model DSTrans. Motivated by the HSI characteristics, we introduce key designs to the core components of the Transformer block for capturing inter-spectrum and inter-pixel similarity and long-range dependencies. Specifically, our Multi-Dconv-head spectral attention (MDSA) and Multi-head Spatial self-attention (MSSA) model local and global context by applying self-attention across spectral and the spatial dimension on local windows. The proposed Dual-stream feed-forward network (DSFN) introduces a gating mechanism to activate the detailed feature. Moreover, we train our DSTrans with an auxiliary RGBI restoration task. This strategy exploits the enormous high-quality RGBI samples and sparse HSI samples to optimize our DSTrans. We establish a series of experiments for HSI SR and denoise. Quantitative and qualitative comparisons demonstrate that our DSTrans surpasses state-of-the-art methods and obtains more pleasant visual results.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017.

[2] Josue Anaya and Adrian Barbu. Renoir–a dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, 51:144–154, 2018.

[3] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016.

[4] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng. Deep spatial-spectral global reasoning network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[6] Ayan Chakrabarti and Todd Zickler. Statistics of real-world hyperspectral images. In *CVPR 2011*, pages 193–200, 2011.

[7] Yi Chang, Luxin Yan, Xi-Le Zhao, Houzhang Fang, Zhijun Zhang, and Sheng Zhong. Weighted low-rank tensor recovery for hyperspectral image restoration. *IEEE transactions on cybernetics*, 50(11):4558–4572, 2020.

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[9] Yong Chen, Ting-Zhu Huang, Wei He, Xi-Le Zhao, Hongyan Zhang, and Jinshan Zeng. Hyperspectral image denoising using factor group sparsity-regularized nonconvex low-rank approximation. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[10] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[11] Xin Deng and Pier Luigi Dragotti. Deep coupled ista network for multi-modal image super-resolution. *IEEE Transactions on Image Processing*, 29:1683–1698, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Weisheng Dong, Fazuo Fu, Guangming Shi, Xun Cao, Jinjian Wu, Guangyu Li, and Xin Li. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, 25(5):2337–2352, 2016.

[14] Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 30:5754–5768, 2021.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[16] Ying Fu, Zhiyuan Liang, and Shaodi You. Bidirectional 3d quasi-recurrent neural network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2674–2688, 2021.

[17] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[18] Wei He, Quanming Yao, Chao Li, Naoto Yokoya, Qibin Zhao, Hongyan Zhang, and Liangpei Zhang. Non-local meets global: An integrated paradigm for hyperspectral image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[19] Jing Hu, Xiuping Jia, Yunsong Li, Gang He, and Minghua Zhao. Hyperspectral image super-resolution via intrafusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7459–7471, 2020.

[20] Jin-Fan Hu, Ting-Zhu Huang, and Liang-Jian Deng. Fusformer: A transformer-based fusion approach for hyperspectral image super-resolution. *arXiv preprint arXiv:2109.02079*, 2021.

[21] Maryam Imani and Hassan Ghassemian. An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. *Information fusion*, 59:59–83, 2020.

[22] Haobo Ji, Xin Feng, Wenjie Pei, Jinxing Li, and Guangming Lu. U2-former: A nested u-shaped transformer for image restoration. *arXiv preprint arXiv:2112.02279*, 2021.

[23] Jinrang Jia, Luyan Ji, Yongchao Zhao, and Xiurui Geng. Hyperspectral image super-resolution with spectral–spatial network. *International journal of remote sensing*, 39(22):7806–7829, 2018.

[24] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6:1082–1096, 2020.

[25] Uzair Khan, Sidike Paheding, Colin P Elkin, and Vijaya Kumar Devabhaktuni. Trends in deep learning for medical hyperspectral image analysis. *IEEE Access*, 9:79534–79548, 2021.

[26] Zeqiang Lai, Kaixuan Wei, and Ying Fu. Deep plug-and-play prior for hyperspectral image restoration. *Neurocomputing*, 2022.

[27] Jiaojiao Li, Ruxing Cui, Bo Li, Rui Song, Yunsong Li, Yuchao Dai, and Qian Du. Hyperspectral image super-resolution by band attention through adversarial learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(6):4304–4318, 2020.

[28] Ke Li, Dengxin Dai, and Luc Van Gool. Hyperspectral image super-resolution with rgb image super-resolution as an auxiliary task. In *Proceedings of the IEEE/CVF Winter Confer-*

ence on Applications of Computer Vision, pages 3193–3202, 2022.

[29] Qiang Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, 12(10):1660, 2020.

[30] Qiang Li, Qi Wang, and Xuelong Li. Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(10):8693–8703, 2021.

[31] Yong Li, Lei Zhang, Chen Dingl, Wei Wei, and Yanning Zhang. Single hyperspectral image super-resolution with grouped deep recursive residual network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–4. IEEE, 2018.

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[33] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 29:563–567, 2022.

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[35] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. Remote sensing image fusion based on two-stream fusion network. *Information Fusion*, 55:1–15, 2020.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[37] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015.

[38] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution. *arXiv preprint arXiv:2108.11084*, 2021.

[39] Alessandro Maffei, Juan M Haut, Mercedes Eugenia Paoletti, Javier Plaza, Lorenzo Bruzzone, and Antonio Plaza. A single model cnn for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2516–2529, 2019.

[40] Matteo Maggioni, Vladimir Katkovnik, Karen Egiazarian, and Alessandro Foi. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE transactions on image processing*, 22(1):119–133, 2012.

[41] Gabriela Takahashi Miyoshi, Nilton Nobuhiro Imai, Antonio Maria Garcia Tommaselli, Eija Honkavaara, Roope Näsi, and Érika Akemi Saito Moriya. Radiometric block adjustment of hyperspectral image blocks in the brazilian environment. *International journal of remote sensing*, 39(15-16):4910–4930, 2018.

[42] Han V Nguyen, Magnus O Ulfarsson, and Johannes R Sveinsson. Hyperspectral image denoising using sure-based unsupervised convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3369–3382, 2020.

[43] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020.

[44] Sima Peyghambari and Yun Zhang. Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: an updated review. *Journal of Applied Remote Sensing*, 15(3):031501, 2021.

[45] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[46] Akrem Sellami and Salvatore Tabbone. Deep neural networks-based relevant latent representation learning for hyperspectral image classification. *Pattern Recognition*, 121:108224, 2022.

[47] Qian Shi, Xiaopei Tang, Taoru Yang, Rong Liu, and Liangpei Zhang. Hyperspectral image denoising using a 3-d attention denoising network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10348–10363, 2021.

[48] Oleksii Sidorov and Jon Yngve Hardeberg. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[50] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.

[51] Qi Wang, Qiang Li, and Xuelong Li. Spatial-spectral residual network for hyperspectral image super-resolution. *arXiv preprint arXiv:2001.04609*, 2020.

[52] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[53] Xinya Wang, Jiayi Ma, Junjun Jiang, and Xiao-Ping Zhang. Dilated projection correction network based on autoencoder for hyperspectral image super-resolution. *Neural Networks*, 146:107–119, 2022.

[54] Yao Wang, Jiangjun Peng, Qian Zhao, Yee Leung, Xi-Le Zhao, and Deyu Meng. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4):1227–1243, 2017.

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[56] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[57] Kaixuan Wei, Ying Fu, and Hua Huang. 3-d quasi-recurrent neural network for hyperspectral image denoising. *IEEE transactions on neural networks and learning systems*, 32(1):363–375, 2020.

[58] Bihan Wen, Ulugbek S Kamilov, Dehong Liu, Hassan Mansour, and Petros T Boufounos. Deepcasd: An end-to-end approach for multi-spectral image super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6503–6507. IEEE, 2018.

[59] Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1888–1902, 2018.

[60] Qi Xie, Qian Zhao, Deyu Meng, Zongben Xu, Shuhang Gu, Wangmeng Zuo, and Lei Zhang. Multispectral images denoising by intrinsic tensor sparsity regularization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2016.

[61] Qi Xie, Minghao Zhou, Qian Zhao, Zongben Xu, and Deyu Meng. Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[62] Weiying Xie, Xiuping Jia, Yunsong Li, and Jie Lei. Hyperspectral image super-resolution using deep feature matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):6055–6067, 2019.

[63] Yang Xu, Zebin Wu, Jocelyn Chanussot, and Zhihui Wei. Nonlocal patch tensor sparse representation for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 28(6):3034–3047, 2019.

[64] Jize Xue, Yong-Qiang Zhao, Yuanyang Bu, Wenzhi Liao, Jonathan Cheung-Wai Chan, and Wilfried Philips. Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 30:3084–3097, 2021.

[65] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[66] Qiangqiang Yuan, Qiang Zhang, Jie Li, Huanfeng Shen, and Liangpei Zhang. Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1205–1218, 2018.

[67] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992.

[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

[69] Haijin Zeng, Xiaozhen Xie, Haojie Cui, Yuan Zhao, and Jifeng Ning. Hyperspectral image restoration via cnn denoiser prior regularized low-rank tensor recovery. *Computer Vision and Image Understanding*, 197:103004, 2020.

[70] Gaigai Zhang, Shizhi Zhao, Wei Li, Qian Du, Qiong Ran, and Ran Tao. Htd-net: A deep convolutional neural network for target detection in hyperspectral imagery. *Remote Sensing*, 12(9):1489, 2020.

[71] Lei Zhang, Wei Wei, Chengcheng Bai, Yifan Gao, and Yanning Zhang. Exploiting clustering manifold structure for hyperspectral imagery super-resolution. *IEEE Transactions on Image Processing*, 27(12):5969–5982, 2018.

[72] Tao Zhang, Ying Fu, and Cheng Li. Hyperspectral image denoising with realistic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2248–2257, 2021.

[73] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2020.

[74] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[75] Ke Zheng, Lianru Gao, Qiong Ran, Ximin Cui, Bing Zhang, Wenzhi Liao, and Sen Jia. Separable-spectral convolution and inception network for hyperspectral image super-resolution. *International Journal of Machine Learning and Cybernetics*, 10(10):2593–2607, 2019.