

# Spatial-Spectral Transformer for Hyperspectral Image Denoising

Miaoyu Li<sup>1</sup>, Ying Fu<sup>1\*</sup>, Yulun Zhang<sup>2</sup>

<sup>1</sup>Beijing Institute of Technology

<sup>2</sup>ETH Zürich

{miaoyli, fuying}@bit.edu.cn, yulun100@gmail.com

## Abstract

Hyperspectral image (HSI) denoising is a crucial preprocessing procedure for the subsequent HSI applications. Unfortunately, though witnessing the development of deep learning in HSI denoising area, existing convolution-based methods face the trade-off between computational efficiency and capability to model non-local characteristics of HSI. In this paper, we propose a Spatial-Spectral Transformer (SST) to alleviate this problem. To fully explore intrinsic similarity characteristics in both spatial dimension and spectral dimension, we conduct non-local spatial self-attention and global spectral self-attention with Transformer architecture. The window-based spatial self-attention focuses on the spatial similarity beyond the neighboring region. While, the spectral self-attention exploits the long-range dependencies between highly correlative bands. Experimental results show that our proposed method outperforms the state-of-the-art HSI denoising methods in quantitative quality and visual results. The code is released at <https://github.com/MyuLi/SST>.

## Introduction

Hyperspectral images (HSIs) provide abundant information in spectral dimension and have been widely applied to the fields of remote sensing (Cloutis 1996), material recognition (Thai and Healey 2002), agriculture (Kersting et al. 2012), medical diagnosis (Fei 2020) and so on. However, in the sensing process, due to limited light, photon effects, and atmospheric interference, HSIs often suffer from corruption and noise, which negatively influences the subsequent HSI applications. Therefore, HSI denoising is a critical preprocessing procedure to enhance image quality for the aforementioned high-level computer vision tasks.

Compared to color images, HSIs offer pixel-level spectral features by imaging narrow spectral bands over a continuous spectral range. It means there are statistical similarities between bands. Early denoising works, such as dictionary learning method (Elad and Aharon 2006), and BM3D (Dabov et al. 2007), focused on the non-local similarity in spatial dimension but did not take the spectral features into account. Thus, more HSI denoising works exploit both the spatial similarity and spectral correlation. Multilinear tools

were used in (Renard, Bourennane, and Blanc-Talon 2008) to extract spectral components and spatial information for denoising. The parallel factor analysis model was employed in (Liu, Bourennane, and Fossati 2012) to exploit the decomposition uniqueness and single rank character of HSIs. Low-rank prior (Chen et al. 2017; Chang et al. 2020), sparse representation (Zhuang and Bioucas-Dias 2018), and total variation regularization (Du et al. 2018) have also been widely adopted to HSI denoising. Despite the various hand-crafted priors, traditional model-based HSI denoising methods are always hard to optimize and time-consuming.

With the development of deep learning, convolutional neural networks (CNNs) based HSI denoising methods (Yuan et al. 2018; Chang et al. 2018; Dong et al. 2019; Zhang et al. 2019; Shi et al. 2021) have shown distinguished advantages over traditional HSI denoising methods. CNN-based methods rely on convolution filters to model the data dependencies in spatial dimension and spectral dimension, facing the trade-off between computational efficiency and the ability to model the non-local similarity of HSIs. Besides, the learned convolution filters are with static weights, which means the filters used for feature extraction are fixed in the testing phase. The denoising process is based on the knowledge learned from training dataset, but, the inner characteristics of the target HSI are not fully exploited.

Recently, Transformer models have been applied to vision tasks (Zhang et al. 2020; Carion et al. 2020; Dai et al. 2021). Transformers apply the self-attention mechanism (Wang et al. 2018) across image regions and can well capture the internal similarity of target image. According to previous analysis (Cai et al. 2022; Zamir et al. 2022), Transformer could be a powerful alternative to CNNs in HSI denoising task. However, existing Transformer-based works mainly aim at natural images with ignorance in spectral similarity. Since HSIs have strong spectral correlations, such neglect would negatively affect denoising results. A practical way of employing Transformer to HSI denoising is applying spectral-wise attention to well utilize spectral correlation.

In this work, we propose a Spatial-Spectral Transformer to sufficiently explore the non-local spatial similarity and global spectral correlation of HSIs. First, the spatial information of HSI is restored by the shifted window-based self-attention, which conducts a non-local spatial-wise coarse denoising beyond neighboring pixels. Secondly, the high cor-

\*Corresponding author

relation between bands is exploited by spectral-wise self-attention, which conducts globally weighted denoising on each pixel with fine details. Finally, the output from self-attention module is passed through the multi-layer perceptron (MLP) and skip-connection for smooth convergence. In summary, the main contributions of our work are as follows:

- We propose a Spatial-Spectral Transformer to fully exploit both the non-local spatial similarity and global spectral correlation of target noisy HSI.
- We design an efficient denoising module by integrating window-based spatial self-attention with spectral self-attention. The coarse spatial features are finely weighted by spectral attention to obtain detailed features.
- Extensive experimental results on various noise degradations show that our proposed method outperforms state-of-the-art methods in terms of both objective metrics and subjective visual quality.

## Related Work

In this section, we briefly review two major research directions which are related to our work, including the latest progress in HSI denoising and vision Transformers.

### HSI Denoising

Existing HSI denoising methods could be roughly classified into two categories, including traditional model-based methods and deep learning-based methods.

Model-based HSI denoising methods usually utilize hand-crafted prior knowledge of HSIs. The non-local similarity (Zhang et al. 2019), total variation (He et al. 2015; Yuan, Zhang, and Shen 2012), low-rank (Zhang et al. 2013; Chang, Yan, and Zhong 2017; He et al. 2021; Chang et al. 2020), and sparse representation (Lu et al. 2015; Zhuang and Bioucas-Dias 2018) are frequently used to exploit the intrinsic spatial and spectral characteristics of HSIs. Fu *et al.* (Fu et al. 2015) presented an adaptive spatial-spectral dictionary learning method and high correlations in both domains are exploited. In (He et al. 2019), the spatial non-local similarity and global spectral low-rank property were integrated for HSI denoising. Zhang *et al.* (Zhang et al. 2021) proposed a double low-rank matrix decomposition method. These methods generally formulated HSI denoising as a complex iterative problem and required a long time to optimize.

Existing learning-based HSI denoising methods (Chang et al. 2018; Yuan et al. 2018; Zhang et al. 2019; Sidorov and Yngve Hardeberg 2019) rely on deep CNNs to automatically learn the prior from large-scale datasets. In (Chang et al. 2018), a spatial-spectral deep residual CNN was employed for HSI restoration. To exploit the global dependency and correlation information in both dimensions, Shi *et al.* (Shi et al. 2021) proposed dual-attention denoising network that could obtain more essential feature extraction of HSI. Wei *et al.* (Wei, Fu, and Huang 2020) designed a 3D quasi-recurrent neural network (QRNN3D) to make full use of structural spatial-spectral correlation as well as global correlation along the spectrum. In (Bodrito et al. 2021), a hybrid trainable spectral-spatial sparse coding model was proposed.

Although these CNN-based HSI denoising methods have achieved excellent performance, data similarities in spectral dimension is always underestimated. Moreover, with trained parameters fixed, convolution filters follow a stereotyped paradigm to extract features and lack the adaptability to exploit the intrinsic similarity characteristic of noisy HSI.

### Vision Transformers

Transformer was first proposed in (Vaswani et al. 2017) for NLP tasks and then was successfully applied to numerous vision tasks, such as image classification (Zhang et al. 2020), image segmentation (Wang et al. 2021b), object detection (Dai et al. 2021), and face recognition (Zhong and Deng 2021). Generally, Transformer contains three essential components, including LayerNorm (LN) module, self-attention module, and multi-layer perceptron (MLP) module. When Transformer is first applied to visual tasks, the whole image was treated as a sequence of non-overlapping medium-sized image patches in ViT (Dosovitskiy et al. 2021). To make Transformer more efficient, pyramid vision Transformer was proposed in (Wang et al. 2021a). Swin Transformer (Liu et al. 2021) utilized a shift operation and had linear computational complexity to image size. In (Zamir et al. 2022), an efficient Transformer for high-resolution image restoration was proposed.

Though Transformer has achieved excellent performance, existing Transformers mainly focus on exploiting spatial similarity and lack exploration of spectral correlation. Thus, directly applying previous Transformers to the HSI task (Zhong et al. 2021; Pang, Gu, and Cao 2022) may be less effective in exploiting two-dimensional features of HSIs.

## Method

In this section, we first illustrate the problem formulation of HSI denoising and the motivation of our work. Then, we describe the proposed spatial-spectral multi-head self-attention module in detail (see Figure 1). Finally, the overall architecture of our proposed SST is provided (see Figure 2).

### Motivation and Formulation

Mathematically, the additive noise degradation model for one HSI could be formulated as follows

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \quad (1)$$

where  $\mathbf{x} \in \mathbb{R}^{H \times W \times B}$  stands for the clean HSI, and  $\mathbf{y} \in \mathbb{R}^{H \times W \times B}$  is the noisy HSI corrupted by additive noise  $\mathbf{n}$ .  $H$  and  $W$  stand for spatial height and width, respectively.  $B$  denotes the spectral resolution of HSI.

HSIs in the real-world are usually degraded by multifarious noise, including Gaussian noise, stripe noise, impulse noise, deadline noise, or a mixture of them (Zhang et al. 2013; Chang et al. 2020). The goal of HSI denoising is to restore the desired clean HSI  $\mathbf{x}$  from noisy observation  $\mathbf{y}$ .

The effectiveness of non-local spatial similarity prior has been verified by previous traditional model-based HSI denoising works (Zhang et al. 2019; Chang et al. 2020). Since these methods use hand-crafted priors, they lack the exploration of statistical information from external datasets. Fortunately, the spatial self-attention mechanism could provide

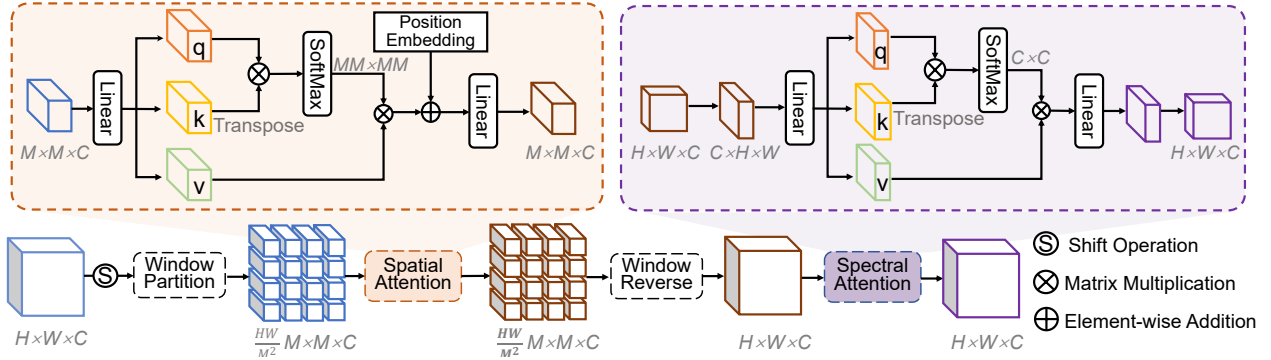


Figure 1: Illustration of Spatial-Spectral Multi-head self-Attention (SSMA). The SSMA module mainly contains non-local spatial multi-head self-attention and global spectral multi-head self-attention.

a comparative ability to obtain the non-local similarity of images. What is more, with learnable parameters to obtain diverse feature expression, a network integrated with self-attention operation can automatically learn the deep prior knowledge from large-scale datasets, and obtains better HSI denoising results. Thus, it is natural to apply the spatial self-attention operation to learning-based HSI denoising.

Moreover, HSI is usually regarded as a data cube and is especially rich in spectral information. The similarities exist not only in the spatial dimension but also in the spectral dimension. Since most existing vision Transformers only conduct spatial self-attention to get representative spatial features, the comprehensive consideration of spatial dimension and spectral dimension is lacking. Therefore, in this work, we combine spectral self-attention with spatial self-attention to exploit both the non-local spatial similarity and spectral high correlation of HSIs. With shifted local spatial window attention, the non-local neighborhood information beyond the pixel is well exploited with computation efficiency. With global spectral attention, the spectral correlation is well modeled and utilized to benefit the denoising process.

### Spatial-Spectral Transformer

In this section, we first introduce our proposed spatial-spectral Transformer layer (SSTL) with our designed self-attention module. To effectively exploit the non-local spatial similarity and spectral correlation of target HSI, we propose a spatial-spectral multi-head self-attention (SSMA) module for HSI denoising. Moreover, according to (Dong, Cordonnier, and Loukas 2021), skip connections and MLP are beneficial to prevent the network from falling into rank collapse. Thus, we follow the structure proposed in (Dosovitskiy et al. 2021) with self-attention module followed by MLP layer, residual connections, and LayerNorm.

Concretely, let  $Z_{l-1} \in \mathbb{R}^{H \times W \times C}$  stands for the input feature embeddings of the  $l$ -th SSTL, the overall processing structure of SSTL could be expressed as:

$$\begin{aligned} Z'_l &= \text{SSMA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \\ Z_l &= \text{MLP}(\text{LN}(Z'_l)) + Z'_l, \end{aligned} \quad (2)$$

where  $Z'_l$  and  $Z_l$  denote the outputs of SSMA and SSTL.

The details of our SSMA are illustrated in Figure 1. It mainly includes a non-local spatial self-attention (NLSA) layer and a global spectral self-attention (GSA) layer.

Given normalized input features,  $Z^{in} \in \mathbb{R}^{H \times W \times C}$ , a window partition operation is first conducted in spatial dimension with a window size of  $M$ . Thus, the whole input features are divided into  $\frac{HW}{M^2}$  non-overlapping patches as  $\{Z_1^{in}, \dots, Z_i^{in}, \dots, Z_S^{in}\}$ , where  $Z_i^{in} \in \mathbb{R}^{M^2 \times C}$ . After partition, each patch  $Z_i^{in}$  individually passes through the NLSA layer to exploit the non-local similarity in the spatial dimension. This process can be expressed as follows:

$$\{Z_i^{in}\} = \text{WinPartition}(Z^{in}), i = 1, \dots, S \quad (3)$$

$$Z_i^{ns} = \text{NLSA}(Z_i^{in}), i = 1, \dots, S \quad (4)$$

$$Z^{ns} = \text{WinReverse}(\{Z_i^{ns}\}), i = 1, \dots, S. \quad (5)$$

After gathering output patches from NLSA layer together through the window reverse operation, the non-local spatial features  $Z^{ns}$  are fed directly to GSA layer. The global spectral correlation is utilized and helps the removal of noise as:

$$Z^{gs} = \text{GSA}(Z^{ns}). \quad (6)$$

**Non-Local Spatial self-Attention (NLSA).** For real-world HSI denoising applications, there is a fundamental trade-off between model capability and model flexibility. Though global spatial self-attention could bring spatial long-range dependencies in a large perspective field, it is frustratingly not suitable for the HSI denoising task since its quadratic complexity grows with spatial size. Thus, we employ self-attention with shifted window to obtain the internal non-local spatial similarity information beyond neighbors. Consequently, NLSA layer provides considerable ability to express spatial features in linear complexity.

For NLSA layer, each input patch  $Z_i^{in} \in \mathbb{R}^{M^2 \times C}$  is linearly projected into *query*  $Q_i^{ns}$ , *key*  $K_i^{ns}$ , and *value*  $V_i^{ns} \in \mathbb{R}^{M^2 \times C}$  to increase the representation ability as:

$$Q_i^{ns} = Z_i^{in} W_q^{ns}, K_i^{ns} = Z_i^{in} W_k^{ns}, V_i^{ns} = Z_i^{in} W_v^{ns}, \quad (7)$$

where  $W_q^{ns}$ ,  $W_k^{ns}$ , and  $W_v^{ns}$  are weights of size  $C \times C$ .

Subsequently, to jointly assemble information from different representation subspaces, multi-head mechanism is

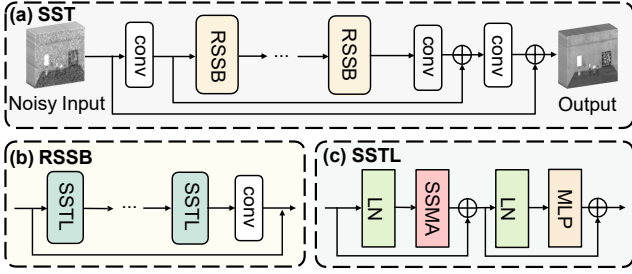


Figure 2: Overall architecture of SST. (a) The basic pipeline of SST. (b) Residual spatial-spectral block (RSSB). (c) Spatial-spectral Transformer layer (SSTL).

employed.  $\mathbf{Q}_i^{ns}$ ,  $\mathbf{K}_i^{ns}$ , and  $\mathbf{V}_i^{ns}$  are split into  $N$  heads as  $\mathbf{Q}_{ij}^{ns}$ ,  $\mathbf{K}_{ij}^{ns}$ , and  $\mathbf{V}_{ij}^{ns}$ , respectively. Thus, the non-local spatial self-attention matrix for each  $head_{ij}^{ns}$  is computed as:

$$\begin{aligned} \mathbf{A}_{ij}^{ns} &= \text{Softmax}(\mathbf{Q}_{ij}^{ns} \mathbf{K}_{ij}^{nsT} / \sqrt{d} + \mathbf{B}), j = 1, \dots, N \\ head_{ij}^{ns} &= \mathbf{A}_{ij}^{ns} \mathbf{V}_{ij}^{ns}, \quad j = 1, \dots, N, \end{aligned} \quad (8)$$

where  $d$  is the dimension of  $\mathbf{Q}_{ij}^{ns}$ , specifically as  $C/N$ . And  $\mathbf{B}$  is the relative bias defined in (Liu et al. 2021). After self-attention operation, the outputs  $head_{ij}^{ns}$  are embedded together and linearly projected to get  $\mathbf{Z}_i^{ns}$  in Eq. (3). Moreover, we also conduct a spatial shift operation between each SSTL to obtain more comprehensive spatial information. It can bring interactions between local windows. Specifically, the shift operation is conducted by shifting the input features by  $\lfloor \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor \rfloor$  pixels before partitioning.

**Global Spectral self-Attention (GSA).** After the non-local spatial self-attention operation, the features are already spatially representative for HSI. But it still lacks spectral representations. Since HSI has underlying spectral correlations, it provides sufficient similarity information for the self-attention operation to obtain long-range dependencies. Thus, we employ global spectral self-attention after non-local spatial self-attention. By doing so, the non-local spatial similarity and spectral correlation are both finely considered.

For a single GSA layer, given an input from NSLA layer,  $\mathbf{Z}^{ns} \in \mathbb{R}^{H \times W \times C}$ , the input is first transposed and reshaped into  $\mathbf{Z}^T \in \mathbb{R}^{C \times HW}$ . Then, it is also linearly projected to  $\mathbf{Q} \in \mathbb{R}^{C \times HW}$ ,  $\mathbf{K} \in \mathbb{R}^{C \times HW}$ , and  $\mathbf{V}^{gs} \in \mathbb{R}^{C \times HW}$  as:

$$\mathbf{Q}^{gs} = \mathbf{W}_q^{gs} \mathbf{Z}^T, \mathbf{K}^{gs} = \mathbf{W}_k^{gs} \mathbf{Z}^T, \mathbf{V}^{gs} = \mathbf{W}_v^{gs} \mathbf{Z}^T, \quad (9)$$

where  $\mathbf{W}_q^{gs}$ ,  $\mathbf{W}_k^{gs}$ ,  $\mathbf{W}_v^{gs}$  are weights of size  $C \times C$ .

Similar to non-local spatial attention, we split  $\mathbf{Q}^{gs}$ ,  $\mathbf{K}^{gs}$ , and  $\mathbf{V}^{gs}$  into  $N$  heads. Each  $head_j^{gs}$  is defined by:

$$\begin{aligned} \mathbf{A}_j^{gs} &= \text{Softmax}(\mathbf{K}_j^{gsT} \mathbf{Q}_j^{gs} / \sqrt{d}), j = 1, \dots, N \\ head_j^{gs} &= \mathbf{V}_j^{gs} \mathbf{A}_j^{gs}, \quad j = 1, \dots, N. \end{aligned} \quad (10)$$

The outputs  $head_j^{gs}$  are concatenated in spectral dimension and projected to  $\mathbf{Z}^{gs}$  in Eq. (6). It is worth emphasizing that  $\mathbf{Z}^{gs}$  contains more spectral details without losing critical spatial information compared to  $\mathbf{Z}^{ns}$ .

**Computational Complexity.** We analyze the computational complexity of our proposed SSMA module as:

$$\begin{aligned} \mathcal{O}(NLSA) &= (M^2 HWC), \mathcal{O}(GSA) = (C^2 HW), \\ \mathcal{O}(SSMA) &= (M^2 HWC + C^2 HW), \end{aligned} \quad (11)$$

where  $M$  and  $C$  are predefined constants. Thus, our proposed SSMA module achieves linear computation cost.

**Overall Network Architecture.** As shown in Figure 2, our whole Transformer first employs one  $3 \times 3$  convolution layer to extract low-level features embeddings  $\mathbf{F}_0 \in \mathbb{R}^{H \times W \times C}$  from noisy observation  $\mathbf{y} \in \mathbb{R}^{H \times W \times B}$ . Then, shallow features pass through  $T$  Residual Spatial-Spectral Block (RSSB) layers with a fixed feature size of  $H \times W \times C$ . The process of deep feature extraction could be denoted as:

$$\mathbf{F}_t = \mathbf{H}_t(\mathbf{F}_{t-1}), t = 1, \dots, T, \quad (12)$$

where  $\mathbf{H}_t$  is the  $t$ -th RSSB layer and  $\mathbf{F}_t$  stands for the  $t$ -th spatial-spectral feature map obtained by RSSB layer.

The designed RSSB contains 6 Spatial-Spectral Transformer layers. Each of RSSB ends up with a  $3 \times 3$  convolution layer. We have adopted a skip connection that adds residual features from the previous block.

To recover from deep feature  $\mathbf{F}_T$ , two  $3 \times 3$  convolution layers are concatenated with shallow features via skip connections. With a global skip connection that adds the noisy input to the output, the middle network actually learns a noise pattern that matches the noise distribution of input.

## Experiments

In this section, we first introduce the datasets and settings used in our experiments. The quantitative metrics and competing methods are also covered. Then, we provide denoising results quantitatively and qualitatively on simulated data and real data. Finally, ablation studies are carried out to analyze the effectiveness of the components in our model.

### Simulated Experiments

**Datasets.** We evaluate our method mainly on ICVL (Arad and Ben-Shahar 2016) dataset. It consists of 201 images with  $1392 \times 1300$  spatial resolution and 31 bands from 400 nm to 700 nm. We follow settings in (Bodrito et al. 2021), which uses 100 HSIs for training and 50 HSIs for testing to ensure pictures captured from the same scene are only used once. Specifically, we center crop training images to size  $1024 \times 1024$  and normalize them to  $[0, 1]$ . Then, we extract patches of size  $64 \times 64$  at different scales, with strides 64, 32, and 32. As for testing samples, each HSI is cropped to size  $512 \times 512 \times 31$  for better visual effects. Normalization is also conducted on testing HSIs.

**Benchmark Models.** We compare our proposed method with four traditional methods, including BM4D (Maggioni et al. 2012), LLRT (Chang, Yan, and Zhong 2017), TSLRLN (He et al. 2021), and NG-Meet (He et al. 2019). Three deep learning methods are also used for comparison, including QRNN3D (Wei, Fu, and Huang 2020), HSID-CNN (Yuan et al. 2018), and T3SC (Bodrito et al. 2021). It is worth mentioning that HSID-CNN method traversed the noisy HSI

Method	10			30			50			70			10-70		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	28.13	0.879	18.72	18.59	0.552	37.9	14.15	0.348	49.01	11.23	0.230	56.45	17.24	0.478	41.94
BM4D (Maggioni et al. 2012)	40.78	0.993	2.99	37.69	0.987	5.02	34.96	0.985	6.81	33.15	0.955	8.40	36.62	0.977	5.51
LLRT (Chang, Yan, and Zhong 2017)	46.72	0.998	1.60	41.052	0.992	2.52	38.24	0.983	3.47	36.23	0.973	4.46	40.06	0.986	3.24
TSLRLN(He et al. 2021)	46.07	0.998	1.82	41.26	0.994	3.03	38.37	0.989	4.36	36.44	0.983	5.69	40.22	0.991	3.58
NGMeet (He et al. 2019)	47.90	0.999	1.39	42.44	0.982	2.06	39.69	0.966	2.49	38.05	0.953	2.83	41.67	0.994	2.19
HSID-CNN (Yuan et al. 2018)	43.14	0.992	2.12	40.30	0.985	3.14	37.72	0.975	4.27	34.95	0.952	5.84	39.04	0.978	3.71
QRNN3D (Wei, Fu, and Huang 2020)	45.61	0.998	1.80	42.18	0.996	2.21	39.70	0.9924	3.00	38.09	0.988	3.42	41.34	0.994	2.42
T3SC (Bodrito et al. 2021)	45.81	0.998	2.02	42.44	0.996	2.44	40.39	0.993	2.85	38.80	0.990	3.26	41.64	0.994	2.61
SST (Ours)	<b>48.28</b>	<b>0.999</b>	<b>1.30</b>	<b>43.32</b>	<b>0.997</b>	<b>1.87</b>	<b>41.09</b>	<b>0.995</b>	<b>2.19</b>	<b>39.55</b>	<b>0.992</b>	<b>2.46</b>	<b>42.57</b>	<b>0.996</b>	<b>1.99</b>

Table 1: Denoising comparisons under Gaussian noise with known variance on ICVL dataset. The best results are in bold.

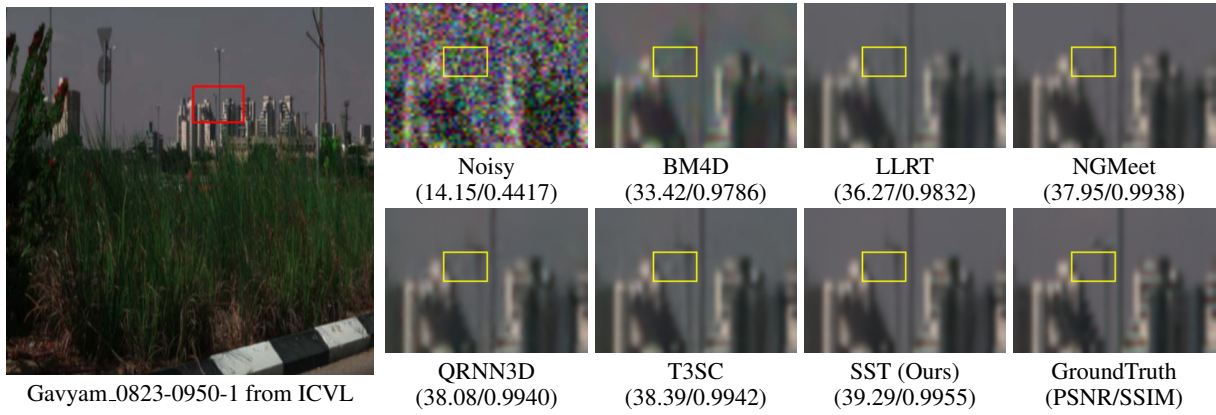


Figure 3: Visual quality comparison under Gaussian noise level  $\sigma=50$  on ICVL dataset using pseudo color image.

through a one-by-one manner, specifically, inputs of the network are the current noisy band and its adjacent bands.

**Metrics.** To quantitatively evaluate our method, we employ three performance metrics, including peak signal-to-noise ratio (PSNR), structure similarity (SSIM), and spectral angle mapper (SAM). Larger values of PSNR and SSIM imply better performance, while smaller values of SAM indicate the high fidelity of denoising results.

**Noise Patterns.** Following same settings in previous HSI denoising works (Wei, Fu, and Huang 2020; Bodrito et al. 2021), we evaluate our method on different noise patterns:

- i.i.d Gaussian noise with known variance  $\sigma$  from 10 to 70. The noise level is the same on all bands.
- Unknown non-i.i.d Gaussian noise with variance  $\sigma$  from 10 to 70. Different bands contain different level of noise.
- Non-i.i.d Gaussian noise with deadline noise, impulse noise, stripe noise, or mixture of them. The detailed settings could be found in (Wei, Fu, and Huang 2020).

**Implementation Details.** We use Adam (Kingma and Ba 2014) to optimize the network with parameters initialized by Xavier initialization (Glorot and Bengio 2010). The batch size is set to 8 with 100 epochs of training. The learning rate is set to  $1 \times 10^{-4}$  and is divided by 10 after 60 epoch.

Competing deep learning methods (HSID-CNN, QRNN3D, and T3SC) and our proposed Transformer are implemented with PyTorch and run with a GeForce RTX 3090. Traditional methods, including BM4D, LLRT, TSLRLN, and NG-Meet, are implemented with Matlab and run with an Intel Core i9-10850K CPU. All parameters involved in these competing algorithms were optimally assigned or automatically chosen as described in the reference papers.

**Gaussian Noise with Known Variance.** In this case, zero mean additive white Gaussian noises with different variance  $\sigma$  are added to the HSI to generate the noisy observations.

The quantitative results of our method and compared methods on ICVL dataset are shown in Table 1. Our method significantly outperforms all compared methods. With a noise level of  $\sigma=70$ , our method increases the PSNR by more than 0.7 dB. Furthermore, it can be seen that compared deep learning methods can achieve comparable results to traditional methods under high noise levels, but they are less effective to handle HSIs with low noise. Our proposed Transformer still achieve the best results under low level noise, showing its robustness and generalization ability.

We show the visual results of one denoised HSI from different methods under  $\sigma=50$  in Figure 3. To further illustrate the result of the spectral fidelity, we use pseudo color im-

Method	Non-i.i.d			Deadline			Impulse			Stripe			Mixture		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
Noisy	18.29	0.512	46.20	17.50	0.477	47.55	14.93	0.376	46.98	17.51	0.487	46.98	13.91	0.340	51.53
BM4D (Maggioni et al. 2012)	36.18	0.977	5.78	33.77	0.962	6.85	29.79	0.861	21.59	35.63	0.973	6.26	28.01	0.842	23.59
LLRT (Chang, Yan, and Zhong 2017)	34.18	0.962	4.88	32.98	0.956	5.29	28.85	0.882	18.17	34.27	0.963	4.93	28.06	0.870	19.37
TSLRLN (He et al. 2021)	41.95	0.994	2.89	36.56	0.977	5.52	33.72	0.893	22.89	38.41	0.985	4.99	27.25	0.852	25.23
NGMeet (He et al. 2019)	34.90	0.975	5.37	33.41	0.967	6.55	27.02	0.788	31.20	34.88	0.967	5.42	26.13	0.778	31.89
HSID-CNN (Yuan et al. 2018)	39.28	0.982	3.80	38.33	0.978	3.99	36.21	0.966	5.48	38.09	0.977	4.59	35.30	0.959	6.29
QRNN3D (Wei, Fu, and Huang 2020)	42.18	0.995	2.84	41.69	0.994	2.61	40.32	0.991	4.31	41.68	0.994	2.97	<b>39.08</b>	0.989	4.80
T3SC (Bodrito et al. 2021)	41.95	0.992	4.18	39.59	0.992	4.86	37.85	0.984	6.53	41.32	0.994	3.27	35.53	0.977	8.12
SST (Ours)	<b>43.57</b>	<b>0.997</b>	<b>2.05</b>	<b>42.74</b>	<b>0.996</b>	<b>2.21</b>	<b>41.66</b>	<b>0.995</b>	<b>2.60</b>	<b>42.97</b>	<b>0.996</b>	<b>2.15</b>	38.78	<b>0.991</b>	<b>2.99</b>

Table 2: Denoising comparisons under five complex noise cases on ICVL dataset. The best results are in bold.

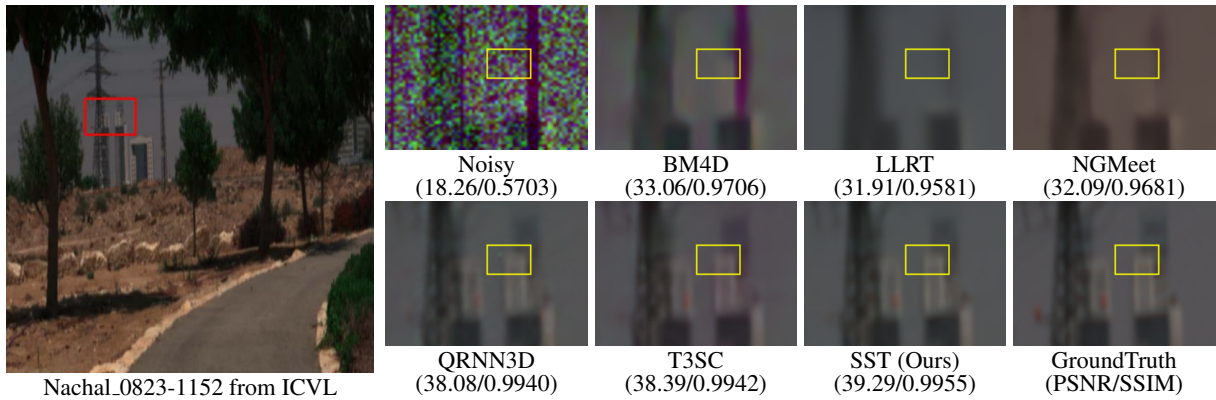


Figure 4: Visual quality comparison under deadline noise on ICVL dataset using pseudo color image.

ages that is composed of bands 9, 15 and 28 for the red, green, and blue channels. BM4D results in spatial discontinuity and loses many high frequency patterns. NG-Meet and LLRT obtain relatively good results, but they still lose fine textures as shown in the detailed figures. The results of QRNN3D and T3SC have obvious artifacts near the edge of the building. Our method obtains the most satisfying image along the spatial dimension and spectral dimension.

**Complex Noise with Unknown Non-i.i.d Gaussian Noise.** Average results of all the competing methods on ICVL with complex noise are shown in Table 2. The visual comparison under non-i.i.d Gaussian noise + deadline noise case are shown in Figure 4. Our method outperforms other methods under various types of noise degradation. An interesting observation is that traditional methods almost all fail to restore clean HSIs under impulse noise and mixed noise, while deep learning methods achieve considerable results. We speculate that the introduction of complex noise to HSI makes it lose certain informative characteristics across the spatial and spectral dimensions. Without the guidance of clean HSI, hand-crafted priors are not strong enough to work well on noisy HSI. Our method not only focuses on the similarity information of HSI, but also has better adaptability to extract features through training, thus obtaining better results.

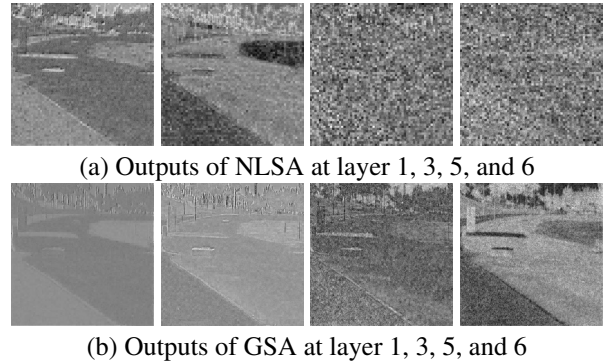


Figure 5: Feature maps before and after GSA module.

**Feature Representation.** In Figure 5, we provide grayscale features maps after NLSA layer and GSA layer at different stages of our proposed Transformer, respectively. The output of the GSA layer obtains more detailed information and structural texture than the outputs of NLSA layer. Since there is a close relationship between bands, different bands could complement each other by global spectral self-attention, resulting in a better-refined feature expression.

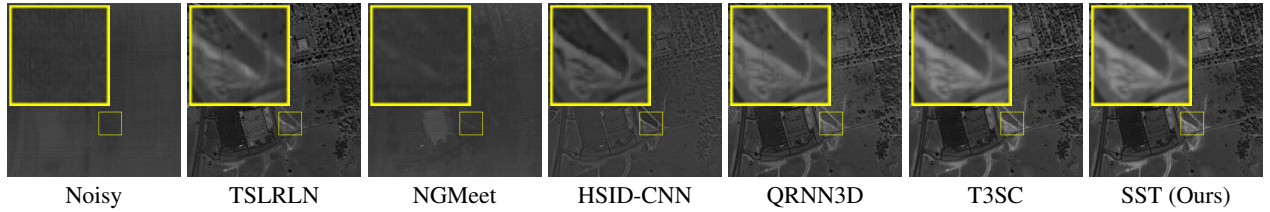


Figure 6: Visual comparison on real dataset Urban of band 105.

Method	Param (M)	GFLOPs	PSNR (dB)
QRNN3D	0.86	19.6	39.70
QRNN3D-L	1.34	30.6	39.82
HSID-CNN	0.40	50.8	37.72
T3SC	0.83	N/A	40.39
Restormer	26.15	9.5	40.53
Restormer-L	45.14	21.9	40.68
SST (Ours)	4.14	20.7	41.09

Table 3: Complexity comparisons with size  $64 \times 64 \times 31$ .

**Time Complexity.** Parameters and GFLOPs are provided in Table 3 under  $\sigma=50$ . QRNN3D-L stands for QRNN3D with deeper layers and more channels compared to the one in (Wei, Fu, and Huang 2020). Moreover, we also include Restormer (Zamir et al. 2022) for comparison. Similarly, Restormer-L stands for a larger one. Besides, since HSID-CNN conducts denoising process in band-by-band manner, the GFLOPs is calculated by multiplying the band number and time that is required for one band. Our method achieves better result under comparable computation cost.

## Real Data Experiments

**Setup.** We evaluate our method on a real-world noisy HSI dataset named Urban. Urban dataset consists of  $307 \times 307$  pixels with 210 bands. Following (Bodrito et al. 2021), our Transformer and compared deep models are pre-trained on the APEX (Itten et al. 2008) dataset, which has similar spectral coverage and band number to the Urban dataset.

**Visual Comparison.** As there is no clean image for real data, we only present grayscale images before and after denoising to visually evaluate competing methods. The results of one noisy band are shown in Figure 6. We can observe that the original HSI suffers from complex noise, which seriously affects the image quality. NGMeet fails to preserve structural content and detail texture due to over-smoothing. Though QRNN3D and T3SC can recover a relatively complete image from the noisy band, vertical stripes still exist in the restored images. Our method achieves satisfactory denoising results and restores the image texture well.

## Ablation Study

To verify the effectiveness of our method, we perform ablation studies with noise level  $\sigma=50$  on ICVL dataset.

**Component Analysis.** In Table 4, we investigate the effect of subcomponents in the SSMA module, which includes a

Method	Params (M)	GFLOPs	PSNR (dB)
w/o NLSA	3.00	14.3	34.67
w/o GSA	2.98	13.1	40.44
NLSA-NLSA	4.23	20.1	40.56
GSA-GSA	4.08	21.4	39.82
GSA-NLSA	4.14	20.7	40.69
NLSA-GSA (Ours)	4.14	20.7	41.09

Table 4: Ablation to study the effect of SSMA module.

Window Size	2	4	8	16
GFLOPs	16.87	17.50	20.70	30.24
PSNR (dB)	42.13	42.38	42.57	42.59
SSIM	0.9951	0.9953	0.9955	0.9955

Table 5: Analysis of the effect of window size.

NLSA layer and a GSA layer. Without NLSA or GSA, the PSNR is 0.5 dB lower. To mitigate the impact of computational complexity, we replace GSA layer with NLSA layer, resulting in a NLSA-NLSA module. Similarly, we perform the experiment on the GSA-GSA module. The results reveal that our attention module outperforms other methods in terms of efficiency. GSA-GSA obtains the worst result among the last four methods. With NLSA layer before GSA layer, the feature extraction of GSA is more reliable.

**Hyperparameter Analysis.** To investigate the influence of window size  $M$  in the NLSA layer, we conducted experiments with different sizes of  $M$  in Table 5. As  $M$  increases, the network gets better performance with higher computation cost. To make a better trade-off between performance and computation cost, we choose  $M=8$  in our experiments.

## Conclusion

In this paper, we propose a Spatial-Spectral Transformer for hyperspectral image denoising. The proposed Transformer considers both spatial similarity and spectral correlation in HSIs through the spatial non-local self-attention and spectral global self-attention. The spatial non-local self-attention exploits the coarse features beyond neighboring pixels. Then, the spectral self-attention enriches the representation with more details. Extensive experiments verify the superiority of our method over state-of-the-art methods under various noise degradations quantitatively and visually. In the future, it is worth investigating how to use noise estimation as guidance to boost the denoising performance of Transformer.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants No. 62171038, No. 61827901, and No. 62088101.

## References

- Arad, B.; and Ben-Shahar, O. 2016. Sparse recovery of hyperspectral signal from natural RGB images. In *ECCV*, 19–34.
- Bodrito, T.; Zouaoui, A.; Chanussot, J.; and Mairal, J. 2021. A Trainable Spectral-Spatial Sparse Coding Model for Hyperspectral Image Restoration. *NeurIPS*.
- Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *CVPR*, 17502–17511.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.
- Chang, Y.; Yan, L.; Fang, H.; Zhong, S.; and Liao, W. 2018. HSI-DeNet: Hyperspectral image restoration via convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2): 667–682.
- Chang, Y.; Yan, L.; Zhao, X.-L.; Fang, H.; Zhang, Z.; and Zhong, S. 2020. Weighted low-rank tensor recovery for hyperspectral image restoration. *IEEE Transactions on Cybernetics*, 50(11): 4558–4572.
- Chang, Y.; Yan, L.; and Zhong, S. 2017. Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising. In *CVPR*, 4260–4268.
- Chen, Y.; Guo, Y.; Wang, Y.; Wang, D.; Peng, C.; and He, G. 2017. Denoising of hyperspectral images using nonconvex low rank matrix approximation. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9): 5366–5380.
- Cloutis, E. A. 1996. Review article hyperspectral geological remote sensing: evaluation of analytical techniques. *International Journal of Remote Sensing*, 17(12): 2215–2242.
- Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE TIP*, 16(8): 2080–2095.
- Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, 1601–1610.
- Dong, W.; Wang, H.; Wu, F.; Shi, G.; and Li, X. 2019. Deep spatial-spectral representation learning for hyperspectral image denoising. *IEEE Transactions on Computational Imaging*, 5(4): 635–648.
- Dong, Y.; Cordonnier, J.-B.; and Loukas, A. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2793–2803.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
- Du, B.; Huang, Z.; Wang, N.; Zhang, Y.; and Jia, X. 2018. Joint weighted nuclear norm and total variation regularization for hyperspectral image denoising. *International Journal of Remote Sensing*, 39(2): 334–355.
- Elad, M.; and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE TIP*, 15(12): 3736–3745.
- Fei, B. 2020. Hyperspectral imaging in medical applications. In *Data Handling in Science and Technology*, volume 32, 523–565.
- Fu, Y.; Lam, A.; Sato, I.; and Sato, Y. 2015. Adaptive spatial-spectral dictionary learning for hyperspectral image denoising. In *ICCV*, 343–351.
- Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.
- He, C.; Sun, L.; Huang, W.; Zhang, J.; Zheng, Y.; and Jeon, B. 2021. TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising. *Signal Processing*, 184: 108060.
- He, W.; Yao, Q.; Li, C.; Yokoya, N.; and Zhao, Q. 2019. Non-local meets global: An integrated paradigm for hyperspectral denoising. In *CVPR*, 6868–6877.
- He, W.; Zhang, H.; Zhang, L.; and Shen, H. 2015. Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1): 178–188.
- Itten, K. I.; Dell’Endice, F.; Hueni, A.; Kneubühler, M.; Schläpfer, D.; Odermatt, D.; Seidel, F.; Huber, S.; Schopfer, J.; Kellenberger, T.; et al. 2008. APEX-the hyperspectral ESA airborne prism experiment. *Sensors*, 8(10): 6235–6259.
- Kersting, K.; Xu, Z.; Wahabzada, M.; Bauckhage, C.; Thureau, C.; Roemer, C.; Ballvora, A.; Rascher, U.; Leon, J.; and Pluemer, L. 2012. Pre-symptomatic prediction of plant drought stress using dirichlet-aggregation regression on hyperspectral images. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Liu, X.; Bourennane, S.; and Fossati, C. 2012. Denoising of hyperspectral images using the PARAFAC model and statistical performance analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10): 3717–3724.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.
- Lu, T.; Li, S.; Fang, L.; Ma, Y.; and Benediktsson, J. A. 2015. Spectral-spatial adaptive sparse representation for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*.
- Maggioni, M.; Katkovnik, V.; Egiazarian, K.; and Foi, A. 2012. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE TIP*, 22(1): 119–133.



- Pang, L.; Gu, W.; and Cao, X. 2022. TRQ3DNet: A 3D Quasi-Recurrent and Transformer Based Network for Hyperspectral Image Denoising. *Remote Sensing*, 14(18): 4598.
- Renard, N.; Bourennane, S.; and Blanc-Talon, J. 2008. Denoising and dimensionality reduction using multilinear tools for hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 5(2): 138–142.
- Shi, Q.; Tang, X.; Yang, T.; Liu, R.; and Zhang, L. 2021. Hyperspectral image denoising using a 3-D attention denoising network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12): 10348–10363.
- Sidorov, O.; and Yngve Hardeberg, J. 2019. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *ICCV Workshops*.
- Thai, B.; and Healey, G. 2002. Invariant subpixel material detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(3): 599–608.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, volume 30.
- Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021a. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 568–578.
- Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*, 7794–7803.
- Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021b. End-to-end video instance segmentation with transformers. In *CVPR*, 8741–8750.
- Wei, K.; Fu, Y.; and Huang, H. 2020. 3-D quasi-recurrent neural network for hyperspectral image denoising. *TNNLS*, 32(1): 363–375.
- Yuan, Q.; Zhang, L.; and Shen, H. 2012. Hyperspectral image denoising employing a spectral-spatial adaptive total variation model. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10): 3660–3677.
- Yuan, Q.; Zhang, Q.; Li, J.; Shen, H.; and Zhang, L. 2018. Hyperspectral image denoising employing a spatial-spectral deep residual convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2): 1205–1218.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 5728–5739.
- Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020. Feature pyramid transformer. In *ECCV*, 323–339.
- Zhang, H.; Cai, J.; He, W.; Shen, H.; and Zhang, L. 2021. Double low-rank matrix decomposition for hyperspectral image denoising and destriping. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.
- Zhang, H.; He, W.; Zhang, L.; Shen, H.; and Yuan, Q. 2013. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8): 4729–4743.
- Zhang, Q.; Yuan, Q.; Li, J.; Liu, X.; Shen, H.; and Zhang, L. 2019. Hybrid noise removal in hyperspectral imagery with a spatial-spectral gradient network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10): 7317–7329.
- Zhong, Y.; and Deng, W. 2021. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*.
- Zhong, Z.; Li, Y.; Ma, L.; Li, J.; and Zheng, W.-S. 2021. Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.
- Zhuang, L.; and Bioucas-Dias, J. M. 2018. Fast hyperspectral image denoising and inpainting based on low-rank and sparse representations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 730–742.