Region-Aware Sequence-to-Sequence Learning for Hyperspectral Denoising

Jiahua Xiao, Yang Liu, and Xing Wei*

School of Software Engineering, Xi'an Jiaotong University
{xjh847286495,1228754216}@stu.xjtu.edu.cn, weixing@mail.xjtu.edu.cn

Abstract. Proper spectral modeling within hyperspectral image (HSI) is critical yet highly challenging for HSI denoising. In contrast to existing methods that struggle between effectiveness and flexibility in modeling long-range spectral dependencies and directly explore spatial-spectral information without region discrimination, we introduce RAS2S—a simple yet effective sequence-to-sequence (Seq2Seq) learning framework for better HSI denoising. RAS2S treats HSI denoising as a Seq2Seq translation problem, which converts the noisy spectral sequence to its clean ones in an autoregressive fashion. In addition, spatial-spectral information exploration without region discrimination contradicts the intrinsic spatial-spectral diversity of HSIs, leading to negative interference from spatial-spectral unrelated regions. Thus we propose a novel spatial-spectral region-aware module to distinctively perceive the semantic regions with different spatial-spectral representations, maximizing the spectral modeling potential of Seq2Seq learning. With such an improved Seq2Seq learning paradigm, RAS2S not only shows huge potential in capturing long-range spectral dependencies, but also maintains the flexibility to handle diverse HSIs. Extensive experiments demonstrate that RAS2S outperforms existing state-of-the-art methods quantitatively and qualitatively with a minimal model size, merely **0.08M**. Source code is available at https://github.com/MIV-XJTU/RAS2S.

Keywords: Hyperspectral Denoising · Spectral Dependency· Sequence-to-Sequence Learning · Region Discrimination

1 Introduction

Unlike conventional images (e.g., grayscale or RGB), hyperspectral images (HSIs) capture the spectral attributes of various objects and materials within a static spatial scenario, widely applied in medical diagnosis [31], material identification [26, 48] and detection [59, 60]. However, various noise types introduced in the imaging process severely disrupt both the visual perception of HSIs and their aforementioned applications, making HSI denoising a crucial yet challenging task. Thanks to the unique spectral information inherited in HSIs [56], which reflects the inner properties of objects, combining spatial and spectral learning allows for more effective noise identification and HSI restoration.

* Corresponding author: Xing Wei.

In the deep learning era, mainstream CNN-based methods [46, 47, 51, 52] jointly extract spatial-spectral information using 2D convolutional filters over a short window of adjacent spectral bands. However, the limited spectral receptive field severely constrains the utilization of global spectral information. While 3D convolution [13, 41, 43] provides a broader spatial-spectral perspective, it introduces significantly increased complexity. More recently, transformer-based methods [5,27,28] leverage self-attention mechanisms to capture complex spatial-spectral relations and effectively remove noise within HSIs. However, they are unable to handles HSIs with arbitrary spectral numbers and need dataset-specific training. As a result, there seems to be a dilemma between the sufficiency and flexibility of exploring long-range spectral dependencies.

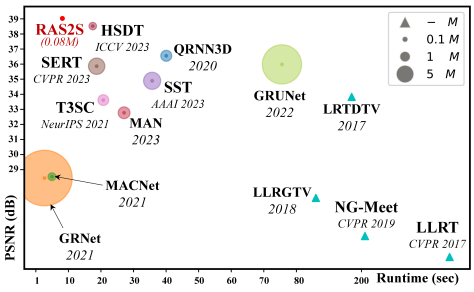


Fig. 1: Denoising performance *vs.* inference time (CPU) under CAVE [34]. The circle areas are proportional to the model size.

To solve this dilemma, we consider continuous spectral bands as a sequence and introduce a novel perspective of treating HSI denoising as a sequence translation problem. We attempt to leverage the proven capability of sequence-to-sequence (Seq2Seq) models in sequence modeling [2,21,37] to tackle this problem, which effectively capture long-range dependencies with fewer parameters (shown in Fig. 1) and spectral flexibility. However, directly applying existing Seq2Seq frameworks to spectral sequences poses challenges: (i) Existing Seq2Seq models generate sequences in a forward order (e.g., text and video sequences), hindering their application to spectral sequences, which can utilize both preceding and succeeding bands. To address this, we employ a bidirectional encoder-decoder architecture tailored for spatial-spectral modeling, where the encoder processes noisy bands to hidden states, and the decoder utilizes these states to dynamically generate denoised spectral sequences autoregressively, capturing long-range contextual dependencies bidirectionally. (ii) Existing Seq2Seq models ignore the information gap between adjacent hidden states and the current input, hindering effective hidden state aggregation. We mitigate this by introducing a hidden state compensation module that adjusts adjacent hidden states with cross-attention, alleviating the adverse effects of the information gap.

In addition, HSIs inherently contain diverse semantic variations across spatial and spectral domains. However, current mainstream convolutional methods perform information extraction using filter sharing across the spatial-spectral domain, treating the feature of each position without discrimination, which will introduces negative interference from semantically unrelated regions during parameter updates. To address this, leveraging insights from spatial non-local self-similarity [32, 58] and spectral distinctiveness (see Fig. 2), we propose discriminative semantic region learning implemented through a Spatial Spectral

Region-Aware (SSRA) module for more focused exploration of spatial-spectral characteristics. SSRA dynamically allocates learnable filters to corresponding semantic regions by utilizing encoded hidden states rich in spatial-spectral aggregation information to generate region mask predictions in both spatial and spectral branches.

The region-guided masks divide the input into distinct regions, with each region assigned a dedicated filter, eliminating irrelevant spatial-spectral interference and maximizing spectral modeling potential of Seq2Seq learning. Thus, the Region-Aware Sequence-to-Sequence (RAS2S) learning framework is proposed for HSI denoising task. Benefiting from this tailored Seq2Seq architecture, our approach effectively captures long-range spectral dependencies for arbitrary spectral numbers, and eliminates the need for dataset-specific training. Overall, this work makes the following contributions:

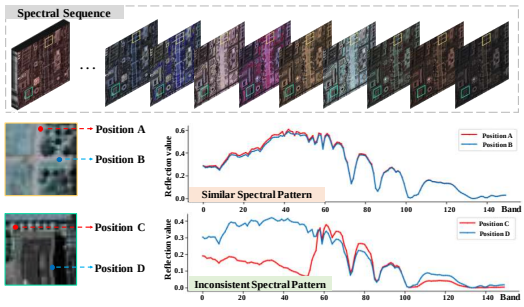


Fig. 2: Illustration of spectral distinctiveness. Spatial positions with similar materials have similar spectral reflection patterns (e.g., Position A and B), while those with different materials show divergent patterns (e.g., Position C and D). This motivates assigning learnable filters to corresponding regions with similar spatial-spectral semantic representations, enabling each region to more effectively focus on its own vital spatial-spectral characteristics. We term this discriminative semantic region learning.

- We introduce RAS2S, the first framework to reformulate the HSI denoising problem from a sequence-to-sequence learning perspective, tailored for spectral sequence modeling.
- We first propose the discriminative semantic region learning within spatial-spectral domains and implement it through a novel SSRA module to maximize the denoising potential of the Seq2Seq framework.
- Extensive experiments on prevailing benchmarks demonstrate the superiority and generality of RAS2S, achieving state-of-the-art performance with few parameters.

2 Related Work

2.1 Hyperspectral Image Denoising

Traditional methods rely on handcrafted priors based on hyperspectral image (HSI) statistics and often struggle when faced with noise distributions inconsistent with prior assumptions. For instance, approaches like [8, 17, 40] capture global spatial-spectral correlation, effectively leveraging low-rank properties and

inherent smoothness in HSIs. Others, such as [7,35], utilize total variation priors to preserve marginal features effectively. In the era of deep learning, 2D CNN-based methods like [47,51,52] jointly extract spatial-spectral information through a short window of adjacent spectral bands, constrained by their spectral receptive field. Meanwhile, 3D CNN-based methods like [36,43,44] capture spatial-spectral dependencies but with increased computational cost. RNN-based approaches like [15] explore sequential spectrum analysis but struggle with long-range spatial-spectral dependencies. Recently, transformer-based models such as [9,27,28], leveraging self-attention mechanisms, show promising denoising performance. However, the extensive interaction among global tokens and the need for fixed-length spectral bands pose computational and flexibility challenges for modeling long-range spectral dependencies. In contrast, our model, a simple serial encoder-decoder architecture, effectively captures long-range spectral dependencies and handles HSIs with varying spectral numbers at a minimal model size.

2.2 Sequence-to-Sequence Learning

Sequence-to-Sequence (Seq2Seq) learning [37] is a pivotal framework in natural language processing (NLP) and has shown significant success in various tasks like machine translation [39] and text summarization [50]. These models use an encoder to process input sequences and a decoder to generate output sequences. They evolve with the introduction of attention mechanisms, allowing them to focus on specific parts of the input when generating output. Recently, there are a growing trend in leveraging Seq2Seq models for various computer vision tasks, such as video restoration [29] and pose estimation [54]. Unlike natural language paragraphs and video frame sequences, which follow a temporal order, hyperspectral bands can utilize both preceding and succeeding bands. Tailored modifications are necessary to achieve effective spatial-spectral modeling and maximize the potential of Seq2Seq models for HSI denoising.

3 Method

In the subsequent sections, we represent the pairs of noisy and clean HSI as $Y = \{y_i\}_1^B$ and $X = \{x_i\}_1^B$, where y_i and x_i denote individual spectral bands, and $Y, X \in \mathbb{R}^{H\times W\times B}$. Here, H and W represent the spatial height and width, respectively, while B indicates the number of spectral bands in the HSI. Therefore, the task of HSI denoising is formulated to maximize the conditional probability $P(X|Y)$.

3.1 HSI Denoising as Sequence Translation

We treat HSI denoising as a sequence translation problem and introduce RAS2S. However, existing Seq2Seq models cannot be directly transferred to HSI processing tasks. To effectively models the long-range spectral dependencies, our model

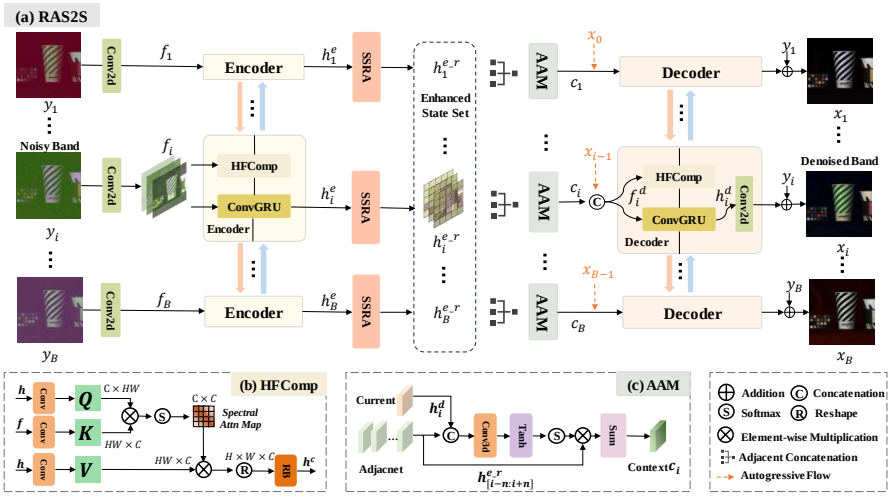


Fig. 3: RAS2S architecture. The workflow of the proposed RAS2S is depicted in (a), which models long-range spectral dependencies with a tailored bidirectional encoder-decoder architecture. The modules at the bottom display the internal details of (b) hidden state compensation (HFComp) and (c) adjacent attention mechanism (AAM).

uses the idea of bidirectional RNN to capture dependencies in both preceding bands $\{y_i\}_1^{i-1}$ and succeeding bands $\{y_i\}_{i+1}^{B}$, and then outputs the bidirectional propagated hidden state sets ($\left\{h_i^{fw}\right\}_1^B$ and $\left\{h_i^{bw}\right\}_1^B$). For brevity and clarity, we take the forward propagation process as an example to elaborate the workflow of RAS2S in the following section, which can be easily extended to a bidirectional process. As presented in Fig. 3(a), RAS2S consists of the following key components.

Encoder. In forward propagation process, the goal of encoder is to transform the shallow feature $\{f_i\}_1^B$ of each band into the hidden state set $\left\{h_i^{fw_e}\right\}_1^B \in \mathbb{R}^{H\times W\times C}$. In order to alleviate the information gap between preceding hidden features $h_{i-1}^{fw_e}$ and current visual features f_i, we propose to make the Hidden State Compensation (HFComp) for $h_{i-1}^{fw_e}$ to obtain an adjusted hidden state $h_{i-1}^{fw_ec}$. The HFComp will be introduced immediately. In implementation, we utilize the ConvGRU cell [3] to facilitate the learning of long-range spatial-spectral information, which combines f_i and $h_{i-1}^{fw_ec}$ to generates the final output $h_i^{fw_e}$. Thus, the workflow of encoder for each spectral step can be formulated as:

$$h_{i-1}^{fw_ec} = \text{HFComp}(f_i, h_{i-1}^{fw_e}), \tag{1}$$

$$h_i^{fw_e} = \text{ConvGRU}(f_i, h_{i-1}^{fw_ec}), \tag{2}$$

where the backward propagation process also produce the hidden state $h_i^{bw_e}$. The final encoded hidden state h_i^e can be generated by using a 2D convolution layer to fuse the forward and backward hidden states ($h_i^{fw_e}$ and $h_i^{bw_e}$).

HFComp. In forward Encoder, HFComp (see Fig. 3(b)) adaptively adjusts $h_{i-1}^{fw_e}$ through calculating the spectral cross-attention map [12] $\mathrm{Att}(Q,K,V) \in \mathbb{R}^{C\times C}$ with f_i . The whole procedure of HFComp is formulated as:

$$\mathrm{Att}(Q,K,V) = V \cdot Softmax(K \cdot Q), \tag{3}$$

$$h_{i-1}^{fw_ec} = \mathrm{RB}(\mathrm{Att}(Q,K,V)) + h_{i-1}^{fw_e}, \tag{4}$$

where $Q \in \mathbb{R}^{C\times HW}, V \in \mathbb{R}^{HW\times C}$ are linearly projected from $h_{i-1}^{fw_e}$ and $K \in \mathbb{R}^{HW\times C}$ is projected from f_i. to improve the feature processing ability, the adjusted hidden features $h_{i-1}^{fw_ec} \in \mathbb{R}^{H\times W\times C}$ is obtained from the concatenated residual blocks (RBs [16]).

SSRA. The SSRA aim to generate the region-aware encoded hidden state $h_i^{e_r}$ based on discriminative semantic region learning. The motivation and implementation of discriminative semantic region learning will be elaborated in subsection 3.2.

AAM. The attention mechanism is widely used in the Seq2Seq model for addressing the challenges of handling long sequences [30], facilitating the effective capture of crucial information in the input sequence. In the context of HSI denoising, we introduce the Adjacent Attention Mechanism (AAM, see Fig. 3(c)) to capture the spatial-spectral context information from the $2n+1$ adjusted adjacent encoded hidden states $\{h_j^e\}_{i-n}^{i+n}$ according to generated decoded hidden state $h_{i-1}^{fw_d}$. Thus, the context feature features c_i can be represented by the weighted sum of adjacent encoded hidden states:

$$c_i = \sum_{j=i-n}^{i+n} \lambda_{i_j} h_j^e; \quad \{\lambda_{i_j}\}_{i-n}^{i+n} = \mathrm{RS}(h_{i-1}^{fw_d}, \{h_j^e\}_{i-n}^{i+n}), \tag{5}$$

where λ is the weight of adjacent encoded hidden states. $\mathrm{RS}(\cdot,\cdot)$ means to calculate the relevant scores between two input features. In implementation, we use a 3D convolution layer for parameter learning as follows:

$$\{\lambda_{i_j}\}_{i-n}^{i+n} = Softmax(Tanh(\mathrm{Conv3D}([h_{i-1}^{fw_d}, \{h_j^e\}_{i-n}^{i+n}]))) \tag{6}$$

where $[\cdot,\cdot]$ represents concatenation along channel dimension.

Decoder. For denoised sequence generation, We employ the identical structure of encoder in the decoder. It serially generate the subsequent output x_i

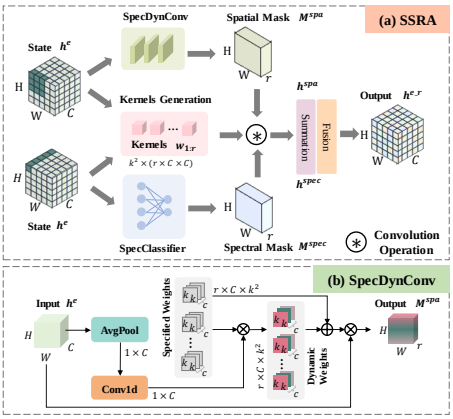


Fig. 4: Illustrations of (a) spatial-spectral region-aware module (SSRA) and (b) Spec- DynConv.

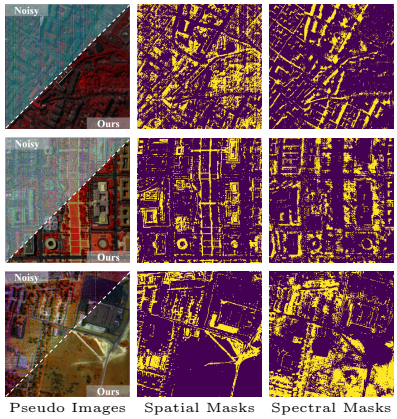


Fig. 5: Visualizations of spatial and spectral guided masks on various HSIs.

with residual learning [55] in an autoregressive manner, conditioned on context features c_i, the decoded hidden feature h_i^d and generated output x_{i-1}:

$$h_i^d = \text{Decoder}(c_i, x_{i-1}), \tag{7}$$

$$x_i = \text{Conv2D}(h_i^d) + y_i, \tag{8}$$

where c_i is generated by AAM. The h_i^d is fused from its forward and backward states ($h_{i-1}^{fw_d}$ and $h_{i-1}^{bw_d}$) and decoded feature f_i^d, where f_i^d is the concatenation feature of generated output x_{i-1} and context c_i. Thus, the inference of $\{x_i\}_1^B$ is to maximizes $P(\{x_i\}_1^B \mid \{h_i^e\}_1^B)$. Based on the chain rule, the goal of Decoder can be reformulated as:

$$\arg\max \prod_{i=1}^{B} P(x_i | x_1, \ldots, x_{i-1}; h_1^e, \ldots, h_B^e). \tag{9}$$

3.2 Discriminative Semantic Regions Learning

Taking a deep insight into the intrinsic properties of HSI (see Fig. 2), we propose the SSRA module (see Fig. 4(a)) to automatically assign learnable filters to their corresponding semantic regions for modeling diverse spatial-spectral variations. Inspired by DRConv [10] (proposed for high-level vision tasks like classification and detection) that divides shallow features into several regions based on local spatial semantics, which is not infeasible and reliable in the regions severely contaminated by noises for HSI, SSRA utilizes HSI-specific spatial-spectral properties to improve semantic region segmentation. In detail, thanks to the Seq2Seq

framework, the SSRA includes spatial and spectral mask prediction branches that operate on the hidden state list $\{h_i^e\}_1^B$ to guide the allocation of filters, which effectively aggregates spatial-spectral representation.

In the spatial branch, we propose SpecDynConv (see Fig. 4(b)) that empowers the spatial convolutions with spectral modeling abilities to address the limitations of local spatial semantics and generates more precise spatial mask prediction $M_i^{spa} \in \mathbb{R}^{H\times W\times r}$ (r means output region number). Specifically, the filter of SpecDynConv $w^{dyn} \in \mathbb{R}^{k^2\times r\times C}$ for hidden state h_i^e is decomposed into the product of a base filter $w^b \in \mathbb{R}^{k^2\times r\times C}$ and a calibration weight $\alpha_i \in \mathbb{R}^{1\times 1\times C}$:

$$w^{dyn} = \alpha_i \cdot w^b + 1; \quad \alpha_i = \text{Conv1D}(\text{GAP}(h_i^e)), \tag{10}$$

where the α_i is dynamically generated from h_i^e through the global average pooling over the spatial dimension and two-layer 1D convolutions, which effectively model the spectral dynamics for all channels in w^b. Different from SENet [19], the calibration weight generated by the input feature can better adapt to itself as illustrated in [20]. Thus, the spatial mask M_i^{spa} is calculated as:

$$M_i^{spa} = \text{Scatter}(Argmax(h_i^e * w^{dyn})) \tag{11}$$

where $*$ is 2D convolution operation. $Argmax(\cdot)$ is the index of the maximum value along the channel dimension, indicating which position belongs to which region. Subsequently, the result of $Argmax(\cdot)$ is scattered to its one-hot form.

In the spectral branch, based on the nature of HSI as illustrated in Fig 2, we propose the SpecClassifier to adaptively learn the spectral inconsistency of each spatial position for semantic region segmentation. Specifically, the SpecClassifier efficiently performs spectral mask prediction using a single-layer linear network. It takes the reshaped hidden state $h_i^e \in \mathbb{R}^{HW\times C}$ as input and produces the spectral mask $M_i^{spec} \in \mathbb{R}^{HW\times r}$, which is then reshaped to match the dimensions of M_i^{spa} and executes the same index scattering operation as in the spatial branch:

$$M_i^{spec} = \text{Scatter}(Argmax(\text{Linear}(h_i^e))). \tag{12}$$

Through spatial and spectral branches, the encoded state h_i^e is divided into r distinct regions, with each region sharing only one filter [38] in $w = \{w_1, \ldots, w_r\}$. The summation is applied to obtain the enhanced features $\{h_i^{spa}, h_i^{spec}\}$ from $\{M_i^{spa}, M_i^{spec}\}$ as below:

$$\{h_i^{spa}, h_i^{spec}\} = \sum_r (h_i^e * w_i) \otimes \{M_i^{spa}, M_i^{spec}\}, \tag{13}$$

where $\otimes$ denotes the pixel-wise multiplications. Finally, $\{h_i^{spa}, h_i^{spec}\}$ are fused to generate the final result $h_i^{e-r} \in \mathbb{R}^{H\times W\times C}$ using a 2D convolution layer. To better visualize the region-aware capability of the SSRA module, we use shallow feature f_i as input instead of deep encoded hidden feature h_i^e, Fig. 5 presents the predicted spatial and spectral guided masks ($r = 2$) generated by the SSRA module, demonstrating that our method successfully assigns filters to regions

with similar spatial-spectral semantics. It needs to be noted that the spatial and spectral semantic segmentation results for the same position do not always align. The spatial branch emphasizes texture edges for partitioning, and the spectral branch focuses on target spectral characteristics. Therefore, SSRA combines both advantages and predictions for a more comprehensive feature learning.

4 Experiments

4.1 Implementation Details

Datasets. Following [28], the RAS2S is trained on ICVL [1]. To evaluate the proposed method, seven public benchmark datasets are adopted, including ICVL [1], CAVE [34], KAIST [11], RealHSI [57], and remotely sensed datasets Pavia centre [14], WDC Mall [45], and Urban [33] for testing. Evaluation metrics include PSNR, SSIM [42] and SAM [53].

Noise Settings. To simulate real-world noise interference, we introduce four types of noise (i.e., non-i.i.d Gaussian, stripe, deadline, and impulse noise) and combine them to create five different noise cases [43]. Among them, the mixture noise case is not introduced in the training process. To assess the generalization of models beyond training noise, we conduct additional validation on i.i.d Gaussian, speckle, and Poisson noise. More details and results are provided in the supplementary material.

Network Settings. We optimize the model with Adam [22]. The batch size is set to 8 with 90 epochs of training. The learning rate is initialized as 1e-3 and gradually decreased to 1e-6. We train the model on single NVIDIA GTX3090 GPU. Unless otherwise specified, all the quantitative and qualitative comparisons are conducted with 3 learnable regions within SSRA, and the sequence attention length within AAM is set to 3.

4.2 Main Results

In all experiments, the RAS2S trained on ICVL [1] is directly adopted to denoise HSI datasets with variant-numbers of bands, without necessitating dataset-specific training and fine-tuning. In contrast, methods such as GRNet [6], T3SC [4], SST [27] and SERT [28], which lack flexibility with a high number of spectral bands, we adopt a sliding window manner for denoising and calculate the average values as the ultimate results. More results are provided in the supplementary material.

Simulation Scenarios Denoising

Quantitative Comparison. (i) Noise robustness evaluations on the KAIST [11] dataset are presented in Tab. 1. RAS2S exhibits superior denoising performance across all noise cases, particularly outperforming the state-of-the-art method by

Table 1: Average quantitative results under various noise cases on KAIST [11] dataset. The **best** and second-best values are highlighted.

Methods	Non-i.i.d Gaussian			Gaussian+Stripe			Gaussian+Deadline			Gaussian+Impulse			Gaussian+Mixture		
	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Noisy	18.26	0.6145	1.0741	18.10	0.6136	1.0746	17.23	0.5770	1.0960	15.00	0.4625	1.0965	13.70	0.4216	1.1133
LLRT [8]	33.36	0.9339	0.1665	30.85	0.8725	0.6351	30.06	0.8942	0.5495	27.14	0.7126	0.7400	24.74	0.6685	0.7764
LRTDTV [40]	37.33	0.9510	0.2195	37.16	0.9498	0.2113	36.49	0.9413	0.2354	37.03	0.9505	0.2445	35.60	0.9355	0.2503
LLRGTV [18]	36.66	0.9109	0.3678	36.68	0.9152	0.3510	34.76	0.8761	0.4268	33.61	0.8502	0.5109	32.14	0.8166	0.5218
NGMeet [17]	32.90	0.9038	0.2755	32.64	0.9027	0.3022	32.06	0.9002	0.2966	25.79	0.723	0.7395	24.54	0.7042	0.7612
QRNN3D [43]	39.42	0.9874	0.2085	39.24	0.9877	0.2086	38.99	0.9868	0.2167	38.15	0.9853	0.2505	37.10	0.9831	0.2934
GRNet [6]	30.58	0.9131	0.5235	30.41	0.9093	0.5298	30.22	0.9044	0.5280	28.77	0.8870	0.5996	28.29	0.8847	0.5780
MAC-Net [49]	34.33	0.9537	0.4795	33.37	0.9520	0.4733	32.82	0.9428	0.4718	31.02	0.9205	0.5925	28.41	0.9041	0.5840
T3SC [4]	37.79	0.9915	0.1603	37.68	0.9915	0.1678	37.04	0.9903	0.1888	34.87	0.9808	0.2561	33.62	0.9750	0.2958
GRUNet [24]	39.39	0.9863	0.2232	39.21	0.9869	0.2113	38.94	0.9850	0.2294	38.11	0.9834	0.2811	36.69	0.9741	0.3525
MST [5]	32.68	0.9545	0.3942	32.53	0.9572	0.3965	32.57	0.9530	0.4050	34.85	0.9569	0.3747	34.12	0.9513	0.3884
MAN [23]	38.76	0.9887	0.2076	38.54	0.9889	0.2102	38.16	0.9881	0.2117	36.38	0.9779	0.3220	35.27	0.9704	0.3818
SST [27]	36.79	0.9774	0.2267	36.59	0.9779	0.2294	36.34	0.9762	0.2323	35.62	0.9726	0.2622	34.65	0.9638	0.3050
SERT [28]	37.77	0.9851	0.1918	37.57	0.9849	0.1956	37.35	0.9836	0.1976	36.58	0.9803	0.2262	35.71	0.9761	0.2459
HSDT-S [25]	40.18	0.9905	0.1413	40.07	0.9909	0.1406	39.76	0.9920	0.1465	39.47	0.9913	0.1537	38.31	0.9904	0.1582
RAS2S(Ours)	**40.70**	**0.9916**	**0.1373**	**40.74**	**0.9919**	**0.1364**	**40.52**	**0.9936**	**0.1293**	**40.14**	**0.9933**	**0.1362**	**39.48**	**0.9920**	**0.1532**

Table 2: Average quantitative results under mixture noise case on different datasets. The **best** and second-best values are highlighted.

Methods	ICVL [1]			CAVE [34]			Pavia Centre [14]			WDC [45]		
	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Noisy	13.97	0.3392	0.8987	14.17	0.4188	1.1371	13.89	0.3400	0.9746	13.98	0.2148	1.0120
LLRT [8]	24.79	0.6294	0.4889	24.59	0.5795	0.8171	16.63	0.4895	0.8437	22.80	0.5083	0.5718
LRTDTV [40]	34.46	0.9184	0.1127	33.82	0.9085	0.2938	29.65	0.8963	0.2445	36.33	0.8597	0.2148
LLRGTV [18]	31.39	0.8756	0.2538	27.12	0.7221	0.6567	28.41	0.8923	0.3142	34.32	0.8260	0.3150
NGMeet [17]	26.59	0.7787	0.4016	23.20	0.6229	0.7874	22.59	0.6658	0.4839	23.26	0.5031	0.4541
QRNN3D [43]	39.22	0.9904	0.0809	36.55	0.9825	0.4244	32.93	0.9698	0.1570	33.96	0.8744	0.1344
GRNet [6]	31.67	0.9557	0.1431	28.44	0.8899	0.6329	26.57	0.8536	0.2815	25.14	0.7682	0.3255
MAC-Net [49]	30.75	0.9332	0.2673	28.53	0.8920	0.6234	27.34	0.8813	0.3530	30.74	0.7740	0.5371
T3SC [4]	35.68	0.9790	0.1389	33.61	0.9728	0.4137	31.39	0.9523	0.2314	30.49	0.9064	0.1972
GRUNet [24]	38.63	0.9913	0.0651	35.98	0.9764	0.4049	31.86	0.9592	0.1598	31.37	0.8018	0.1650
MST [5]	39.12	0.9914	0.0582	33.97	0.9473	0.4657	31.21	0.9355	0.1971	33.49	0.8557	0.2289
MAN [23]	35.85	0.9691	0.1458	35.15	0.9712	0.4990	32.94	0.9711	0.1820	32.77	0.8065	0.1638
SST [27]	39.58	0.9928	0.0480	34.89	0.9616	0.4095	32.85	0.9588	0.1555	30.59	0.8924	0.1860
SERT [28]	40.44	0.9941	0.0470	35.86	0.9737	0.3403	33.28	0.9653	0.1451	31.33	0.9098	0.1609
HSDT-S [25]	40.76	0.9940	0.0505	38.52	0.9892	0.2362	34.25	0.9768	0.1295	37.79	0.9465	0.1049
RAS2S(Ours)	**41.33**	**0.9953**	**0.0467**	**39.03**	**0.9913**	**0.2221**	**36.84**	**0.9862**	**0.0958**	**40.92**	**0.9760**	**0.0735**

3.6 dB in untrained mixture noise cases. This highlights the impressive denoising robustness of RAS2S. (ii) Generalization results on multiple datasets are shown in Tab. 2. RAS2S outperforms other methods on ICVL [1] and CAVE [34] (with the same spectral number as the training set). Additionally, RAS2S exhibits stronger generalization abilities on remotely sensed datasets, Pavia Centre [14] and WDC [45], with a larger spectral number than the training set. RAS2S outperforms the state-of-the-art denoiser on all indicators by a large margin, demonstrating robust generalization and insensitivity to spectral categories.

Qualitative Comparison. Qualitative comparisons between RAS2S and other models in Fig. 6 reveal the following: (i) For ICVL [1] and CAVE [34] with the same spectral number as the training set, other methods generate inaccurate textures and exhibit inferior noise removal performance. (ii) For the KAIST [11]

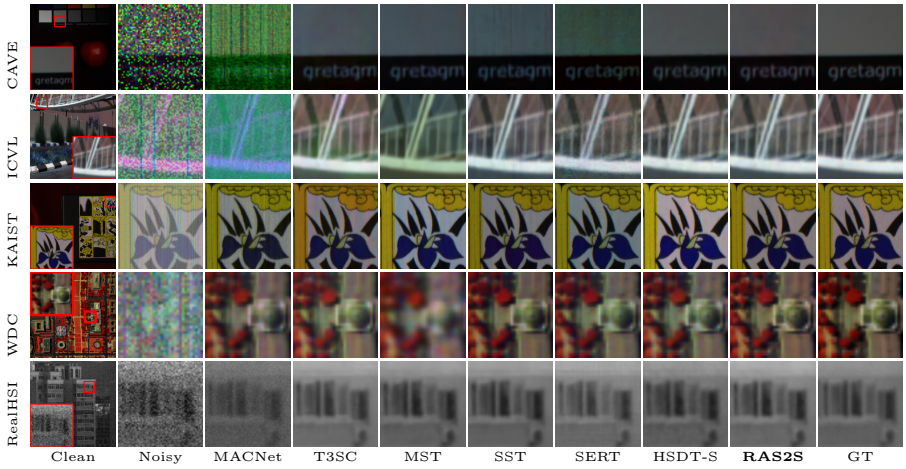


Fig. 6: Visual comparisons on prevailing paired datasets. Except for RealHSI [57], which involves real-world noise, the other datasets are contaminated with Non-i.i.d Gaussian+Mixture noise. GT denotes the reference ground-truth.

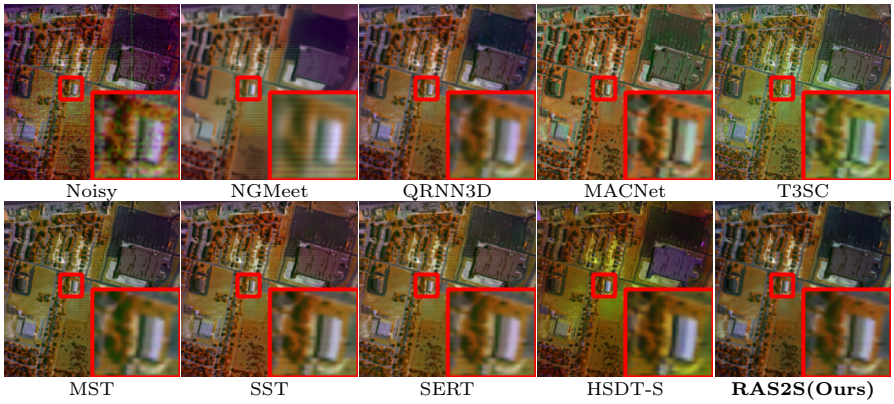


Fig. 7: Visual comparison on the real-world dataset Urban.

dataset with high spatial resolution, RAS2S effectively removes complex noises while retaining detailed textures and edges. In contrast, other models introduce color discrepancies and detail loss (T3SC [4] and SERT [28]) or leave behind residual noise (MACNet [49] and SST [27]). (iii) For the WDC [45] and Pavia Centre datasets with a large spectral number, RAS2S exhibits outstanding noise removal while preserving main structures with fewer artifacts.

Real-World Scenarios Since the ultimate goal is to deal with complicated real-world noises and produce visually pleasing HSIs, we conduct experiments on the real-world datasets, i.e., RealHSI [57] and Urban [33].

Table 3: Efficiency results on CAVE [34] and denoising results on RealHSI [57].

Methods	Efficiency			RealHSI [57]	
	Params(M)	CPU(s)↓	GPU(s)↓	PSNR(dB)↑	SAM↓
LLRT [8]	-	807.21	-	26.60	0.1004
LRTDTV [40]	-	97.65	-	26.58	0.0620
LLRGTV [18]	-	171.72	-	26.14	0.0626
NGMeet [17]	-	86.03	-	27.00	0.0592
QRNN3D [43]	0.86	40.06	0.32	26.60	0.0629
GRNet [6]	41.44	**2.66**	**0.12**	26.80	0.1168
MACNet [49]	0.43	4.95	2.53	24.25	0.1480
T3SC [4]	0.83	20.73	0.51	26.72	0.0576
GRUNet [24]	14.28	75.58	0.50	26.60	0.0587
MAN [23]	0.87	27.11	0.45	26.93	0.0703
SST [27]	4.10	35.72	2.06	26.95	0.0711
SERT [28]	1.91	18.69	0.53	26.51	0.0850
HSDT-S [25]	0.13	17.49	0.31	27.25	0.0544
RAS2S(Ours)	**0.08**	8.16	0.58	**27.39**	**0.0535**

Table 4: Ablation studies to verify module effectiveness.

AAM	HFComp	SSRA	Params(KB)↓	Time(s)↓	PSNR(dB)↑
-	-	-	79.11	7.13	37.12
✓	-	-	79.43	7.74	37.56
✓	✓	-	99.62	8.32	38.45
✓	✓	✓	84.11	8.16	39.03

Table 5: Evaluation of component effectiveness within SSRA.

Methods	Params(KB)↓	Time(s)↓	PSNR(dB)↑
w/o SSRA	76.46	7.62	37.51
w/ Conv2D	82.87	7.92	37.67
w/ SpecDynConv	84.06	8.04	37.92
w/ SpecClassifier	82.53	7.96	38.23
w/ SSRA (Ours)	84.11	8.16	39.03

Quantitative Comparison. The quantitative comparison of different methods on the RealHSI [57] is presented in Tab. 3. Model parameters and inference time also can be observed. The CPU and GPU inference time are measured on Intel(R) Xeon(R) Gold 6226R and NVIDIA GTX3090, respectively. As can be seen, our RAS2S outperforms SERT [28] by a significant margin, achieving up to 0.88 dB PSNR scores, while employing even fewer parameters (only 4.18% of SERT [28] parameters) and runtime. The results indicate that despite the simplicity, our RAS2S present effectiveness in dealing with real-world noise.

Qualitative Comparison. For visual comparison of denoised Urban [33] in Fig. 7, the competing methods have obvious residual stripe noise and blurring. In contrast, RAS2S still generates more realistic and visually pleasing textures without residual noises and artifacts, demonstrating the robustness of our method in real-world applications.

4.3 Ablation Studies

Unless otherwise specified, we apply the CAVE dataset under the mixture noise for ablation studies. We modify different baselines to a similar model size by adjusting the feature channels for fair comparison.

Component Analysis. Tab. 4 summarizes the impact of different modules. The base Seq2Seq model achieves competitive performance with minimal model size, confirming the effectiveness of Seq2Seq learning. AAM improves denoising by 0.44 dB, HFComp enhances results by addressing information gaps between hidden states and current spectral features, while SSRA significantly boosts performance by 0.58 dB in PSNR, demonstrating the effectiveness of semantic region learning for spatial-spectral information exploration. Tab. 5 compares SSRA designs, with SpecDynConv enhancing denoising by 0.41 dB and SpecClassifier achieving a 0.72 dB PSNR gain by leveraging spectral properties. SSRA combines these advantages for comprehensive feature learning, resulting in substantial improvement.

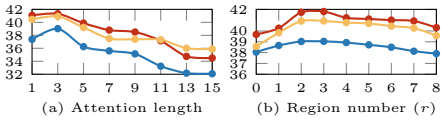


Fig. 8: Performance *vs.* the hyperparamters with different number. The —●—, —●—, and —●— curves denote the denoising performance on CAVE, ICVL and WDC datasets, respectively.

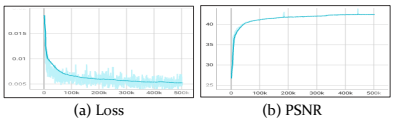


Fig. 9: Convergence curve. We show the average training loss (a) and training batch PSNR (b) of the final model.

Hyperparamter Analysis. We assess our proposed model under various settings for the sequence attention length in AAM and the learnable-region number in SSRA across multiple datasets. Fig. 8(a) shows that the model achieves optimal performance across all datasets when the sequence attention length in AAM is set to 3, suggesting that the nearest hidden state provides maximum assistance in capturing spatial-spectral relations for the current state. Longer attention lengths introduce excessive redundant spatial-spectral information, adversely affecting denoising performance. Fig. 8(b) indicates optimal performance when the learnable-region number in SSRA is 2 or 3, as the simplified SSRA design limits capacity for more regions, and increasing regions may lead to imprecise segmentation, degrading model performance.

4.4 Discussion

Model Complexity. Although RAS2S includes the quadratic attention module (HFComp), it is simpler than Transformer-based methods as shown in Tab. 6.

(i) Fewer parameters: Transformer methods often stack numerous attention layers, while RAS2S utilizes only a single layer of cross-attention. (ii)Lower computational complexity: In most denoising transformers (e.g., SST, HSDT), the spectral feature channels $\boldsymbol{C}$ with quadratic computational overhead are typically more than the spectral bands $\boldsymbol{B}$. In contrast, the channels $\boldsymbol{C}$ of RAS2S is more less than bands $\boldsymbol{B}$, resulting in lower model size and computational complexity. Moreover, the smaller parameter space increases the likelihood of finding the global or local optimum during optimization, making it easier for the final model to converge. The convergence curve is shown in the Fig. 9.

Table 6: Complexity analysis.

Model	PSNR(dB)	Params(M)	Complexity
SST	34.89	4.10	$HW(M^2C + C^2)$
HSDT-S	38.52	0.13	HWC^2B
Ours	**39.03**	**0.08**	HWC^2B

Unidirection *vs.* Bidirection. Existing Seq2Seq models follow unidirectional forward temporal modeling, which is unsuitable for hyperspectral spectral sequences. We validate forward, backward, and bidirectional spectral modeling approaches. Tab. 7 shows that forward and backward spectral modeling yield similar denoising performance, as spectral sequence are not sensitive to the order of spectral bands. In contrast, the bidirectional modeling framework enhances

Table 7: Ablation studies to verify bidirection framework.

Model	Params(KB)↓	Times(s)↓	PSNR(dB)↑
Forward	91.48	7.67	38.42
Backward	91.48	7.67	38.29
Bidirection	84.11	8.16	39.03

Table 8: Ablation studies to verify autoregressive paradigm.

Model	Params(KB)↓	Times(s)↓	PSNR(dB)↑
w/o AR	80.56	7.83	38.62
w/ AR	84.11	8.16	39.03

denoising performance by integrating preceding and succeeding spectral information, effectively capturing long-range spectral dependencies.

Autoregressive Paradigm. Current per-band denoising methods restore each band independently using spectral information from adjacent bands, but overlook leveraging spectral similarity constraints. Our autoregressive framework addresses this by incorporating information from previously generated bands to enhance the reconstruction of the current band, enabling more effective learning, maintaining consistency across bands, and reducing noise impact. Without the autoregressive paradigm, Equation 9 simplifies to:

$$\arg\max \prod_{i=1}^{B} P(x_i|h_1^e, \ldots, h_B^e). \tag{14}$$

We make an ablation study in Tab. 8. It is evident that adopting the autoregressive paradigm lead to overall better denoising performance for our model.

5 Conclusion

We propose RAS2S, a simple yet effective framework for HSI denoising, leveraging the power of Seq2Seq learning. We regard HSI denoising as a noisy-to-clean sequence translation problem, and maximize the denoising potential of the Seq2Seq model through discriminative semantic region learning, utilizing insights from the intrinsic spatial-spectral properties of HSI. RAS2S uses a simple bidirectional encoder-decoder architecture and serially models spatial-spectral correlations without the need for dataset-specific training, presenting the huge potential for capturing long-range dependencies. More about, RAS2S effectively handles the diverse spatial-spectral semantic variations, delivering more precise denoising results. Extensive experiments prove RAS2S outperforms other mainstream methods and achieves state-of-the-art on prevailing benchmark datasets with few parameters.

Acknowledgements

This work was supported by National Key R&D Program of China under Grant No. 2021ZD0110400, the Fundamental Research Funds for the Central Universities No. xxj032023020, and sponsored by the CAAI-MindSpore Open Fund, developed on OpenI Community.

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: ECCV (2016) 9, 10
2. Bai, Y., Zhao, Z., Gong, Y., Wei, X.: Artrackv2: Prompting autoregressive tracker where to look and how to describe. In: CVPR (2024) 2
3. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving deeper into convolutional networks for learning video representations. In: ICLR (2016) 5
4. Bodrito, T., Zouaoui, A., Chanussot, J., Mairal, J.: A trainable spectral-spatial sparse coding model for hyperspectral image restoration. In: NeurIPS (2021) 9, 10, 11, 12
5. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: CVPR (2022) 2, 10
6. Cao, X., Fu, X., Xu, C., Meng, D.: Deep spatial-spectral global reasoning network for hyperspectral image denoising. IEEE TGRS **60**, 1–14 (2021) 9, 10, 12
7. Chang, Y., Yan, L., Fang, H., Luo, C.: Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping. IEEE TIP **24**(6), 1852–1866 (2015) 4
8. Chang, Y., Yan, L., Zhong, S.: Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising. In: CVPR (2017) 3, 10, 12
9. Chen, H., Zhao, W., Xu, T., Shi, G., Zhou, S., Liu, P., Li, J.: Spectral-wise implicit neural representation for hyperspectral image reconstruction. IEEE TCSVT (2023) 4
10. Chen, J., Wang, X., Guo, Z., Zhang, X., Sun, J.: Dynamic region-aware convolution. In: CVPR (2021) 7
11. Choi, I., Kim, M., Gutierrez, D., Jeon, D., Nam, G.: High-quality hyperspectral reconstruction using a spectral prior. ACM TOG **36**, 13 (2017) 9, 10
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020) 6
13. Fu, G., Xiong, F., Lu, J., Zhou, J., Qian, Y.: Nonlocal spatial–spectral neural network for hyperspectral image denoising. IEEE TGRS **60**, 1–16 (2022) 2
14. Gamba, P.: A collection of data for urban area characterization. In: IGARSS (2004) 9, 10
15. Guan, J., Lai, R., Li, H., Yang, Y., Gu, L.: Dnrcnn: Deep recurrent convolutional neural network for hsi destriping. IEEE TNNLS (2022) 4
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 6
17. He, W., Yao, Q., Li, C., Yokoya, N., Zhao, Q.: Non-local meets global: An integrated paradigm for hyperspectral denoising. In: CVPR (2019) 3, 10, 12
18. He, W., Zhang, H., Shen, H., Zhang, L.: Hyperspectral image denoising using local low-rank matrix recovery and global spatial–spectral total variation. IEEE J-STARS **11**(3), 713–729 (2018) 10, 12
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018) 8
20. Huang, Z., Zhang, S., Pan, L., Qing, Z., Tang, M., Liu, Z., Ang Jr, M.H.: Tada! temporally-adaptive convolutions for video understanding. In: ICLR (2022) 8
21. Keneshloo, Y., Shi, T., Ramakrishnan, N., Reddy, C.K.: Deep reinforcement learning for sequence-to-sequence models. IEEE TNNLS **31**(7), 2469–2489 (2019) 2

22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
23. Lai, Z., Fu, Y.: Mixed attention network for hyperspectral image denoising. arXiv preprint arXiv:2301.11525 (2023) 10, 12
24. Lai, Z., Wei, K., Fu, Y.: Deep plug-and-play prior for hyperspectral image restoration. Neurocomputing **481**, 281–293 (2022) 10, 12
25. Lai, Z., Yan, C., Fu, Y.: Hybrid spectral denoising transformer with guided attention. In: ICCV (2023) 10, 12
26. Li, J., Zhang, J., Chen, F., Zhao, K., Zeng, D.: Adaptive material matching for hyperspectral imagery destriping. IEEE TGRS **60**, 1–20 (2022) 1
27. Li, M., Fu, Y., Zhang, Y.: Spatial-spectral transformer for hyperspectral image denoising. In: AAAI (2023) 2, 4, 9, 10, 11, 12
28. Li, M., Liu, J., Fu, Y., Zhang, Y., Dou, D.: Spectral enhanced rectangle transformer for hyperspectral image denoising. In: CVPR (2023) 2, 4, 9, 10, 11, 12
29. Lin, J., Hu, X., Cai, Y., Wang, H., Yan, Y., Zou, X., Zhang, Y., Van Gool, L.: Unsupervised flow-aligned sequence-to-sequence learning for video restoration. In: ICML (2022) 4
30. Liu, T., Wang, K., Sha, L., Chang, B., Sui, Z.: Table-to-text generation by structure-aware seq2seq learning. In: AAAI (2018) 6
31. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. Journal of biomedical optics **19**(1), 010901–010901 (2014) 1
32. Maggioni, M., Katkovnik, V., Egiazarian, K., Foi, A.: Nonlocal transform-domain filter for volumetric data denoising and reconstruction. IEEE TIP **22**(1), 119–133 (2012) 2
33. Mnih, V., Hinton, G.E.: Learning to detect roads in high-resolution aerial images. In: ECCV (2010) 9, 11, 12
34. Park, J.I., Lee, M.H., Grossberg, M.D., Nayar, S.K.: Multispectral imaging using multiplexed illumination. In: ICCV (2007) 2, 9, 10, 12
35. Peng, J., Xie, Q., Zhao, Q., Wang, Y., Yee, L., Meng, D.: Enhanced 3dtv regularization and its applications on hsi denoising and compressed sensing. IEEE TIP **29**, 7889–7903 (2020) 4
36. Shi, Q., Tang, X., Yang, T., Liu, R., Zhang, L.: Hyperspectral image denoising using a 3-d attention denoising network. IEEE TGRS **59**(12), 10348–10363 (2021) 4
37. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NeurIPS (2014) 2, 4
38. Wang, H., Xu, K., Lau, R.W.: Local color distributions prior for image enhancement. In: ECCV (2022) 8
39. Wang, W., Jiao, W., Hao, Y., Wang, X., Shi, S., Tu, Z., Lyu, M.: Understanding and improving sequence-to-sequence pretraining for neural machine translation. ACL (2022) 4
40. Wang, Y., Peng, J., Zhao, Q., Leung, Y., Zhao, X.L., Meng, D.: Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. IEEE J-STARS **11**(4), 1227–1243 (2017) 3, 10, 12
41. Wang, Z., Ng, M.K., Zhuang, L., Gao, L., Zhang, B.: Nonlocal self-similarity-based hyperspectral remote sensing image denoising with 3-d convolutional neural network. IEEE TGRS **60**, 1–17 (2022) 2
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004) 9
43. Wei, K., Fu, Y., Huang, H.: 3-d quasi-recurrent neural network for hyperspectral image denoising. IEEE TNNLS **32**(1), 363–375 (2020) 2, 4, 9, 10, 12

44. Wei, X., Xiao, J., Gong, Y.: Blind hyperspectral image denoising with degradation information learning. Remote Sensing **15**(2), 490 (2023) 4
45. Wu, Z., Wang, Q., Wu, Z., Shen, Y.: Total variation-regularized weighted nuclear norm minimization for hyperspectral image mixed denoising. JEI **25**(1), 013037–013037 (2016) 9, 10, 11
46. Xiao, J., Ji, Y., Wei, X.: Hyperspectral image denoising with spectrum alignment. In: ACMMM. pp. 5495–5503 (2023) 2
47. Xiao, J., Wei, X.: Hyperspectral image denoising using uncertainty-aware adjustor. In: IJCAI (2023) 2, 4
48. Xiong, F., Zhou, J., Qian, Y.: Material based object tracking in hyperspectral videos. IEEE TGRS **29**, 3719–3733 (2020) 1
49. Xiong, F., Zhou, J., Zhao, Q., Lu, J., Qian, Y.: Mac-net: Model-aided nonlocal neural network for hyperspectral image denoising. IEEE TGRS **60**, 1–14 (2021) 10, 11, 12
50. Xu, S., Zhang, X., Wu, Y., Wei, F.: Sequence level contrastive learning for text summarization. In: AAAI (2022) 4
51. Yuan, Q., Zhang, Q., Li, J., Shen, H., Zhang, L.: Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network. IEEE TGRS **57**(2), 1205–1218 (2018) 2, 4
52. Yuan, Y., Ma, H., Liu, G.: Partial-dnet: A novel blind denoising model with noise intensity estimation for hsi. IEEE TGRS **60**, 1–13 (2021) 2, 4
53. Yuhas, R.H., Boardman, J.W., Goetz, A.F.: Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques. In: JPL (1993) 9
54. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: CVPR (2022) 4
55. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE TIP **26**(7), 3142–3155 (2017) 7
56. Zhang, Q., Zheng, Y., Yuan, Q., Song, M., Yu, H., Xiao, Y.: Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven. IEEE TNNLS (2023) 1
57. Zhang, T., Fu, Y., Li, C.: Hyperspectral image denoising with realistic data. In: ICCV (2021) 9, 11, 12
58. Zhou, S., Zhang, J., Zuo, W., Loy, C.C.: Cross-scale internal graph neural network for image super-resolution. In: NeurIPS. vol. 33, pp. 3499–3509 (2020) 2
59. Zhu, D., Du, B., Zhang, L.: Two-stream convolutional networks for hyperspectral target detection. IEEE TGRS **59**(8), 6907–6921 (2020) 1
60. Zhuang, L., Gao, L., Zhang, B., Fu, X., Bioucas-Dias, J.M.: Hyperspectral image denoising and anomaly detection based on low-rank and sparse representations. IEEE TGRS **60**, 1–17 (2020) 1