2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687944

DEEP UNFOLDING 3D NON-LOCAL TRANSFORMER NETWORK FOR HYPERSPECTRAL SNAPSHOT COMPRESSIVE IMAGING

Zheng Zhou [†,‡], *Zongxin Liu*[‡], *Yongyong Chen* [⋆], *Bingzhi Chen* [‡,∗] , *Biqing Zeng* [‡], *and Yicong Zhou* [⋆⋆]
[†]*School of Electronics and Communication Engineering, Guangzhou University,* Guangzhou, China
[‡]*School of Software, South China Normal University,* Foshan, China
[⋆]*School of Computer Science and Technology, Harbin Institute of Technology,* Shenzhen, China
[⋆⋆]*Department of Computer and Information Science, University of Macau,* Macau, China

ABSTRACT

Hyperspectral compressive imaging has shown remarkable advancements through the adoption of deep unfolding frameworks, which integrate the proximal mapping prior into the data fidelity term to formulate the reconstruction problem. However, existing technologies still face challenges in effectively capturing spatial-spectral features during the iterative deep prior learning stage, leading to unsatisfactory performance degradation. To address this issue, we propose a deep unfolding 3D non-local transformer (3DNLT) network for hyperspectral compressive imaging. A learnable half-quadratic splitting (HQS) algorithm is utilized to iteratively update the linear projection. Furthermore, a 3D non-local attention u-shaped transformer is presented as the deep proximal mapping prior module to obtain the spatial-spectral long-range dependency features, leading to enhance the network's ability to capture fine-grained hyperspectral and spatial details. Experimental results on both synthetic and real hyperspectral image reconstruction have demonstrated the superior performance of the 3DNLT network compared to state-of-the-art methods.

***Index Terms*—** Deep unfolding, non-local mechanism, transformer, hyperspectral snapshot compressive imaging.

1 Introduction

Hyperspectral Image (HSI) has gained widespread applications, *e.g.,* anomaly detection [1, 2], multimodal classification [3, 4], and image clustering [5], thanks to their unique properties of capturing detailed spectral information for each pixel in a scene [6]. However, the acquisition of 3D hyperspectral data poses significant challenges due to the limitations of traditional optical sensor imaging systems [7]. As one of the renewed imaging technologies, coded aperture snapshot spectral imaging (CASSI) utilizes coded aperture and disperser to modulate 3D HSI data, producing a compressed 2D measurement [8]. Subsequently, developing an effective reconstruction algorithm has become crucial for obtaining satisfactory HSI from a measurement.

Reconstructing HSI from compressed measurement poses a challenging ill-posed problem. To address this issue, numerous methods have been developed, which can be categorized into the following classes. ***1) Traditional hand-craft prior model*** leverages mathematical characteristics of HSI data, such as total variation [9], non-local similarity [10], low-rank [11], and sparsity [12], to incorporate prior knowledge into the reconstruction objective function. These priors rely on predefined mathematical properties and assumptions about the data, which may not always capture the complex structures and variations present in HSI data. ***2) Data-driven deep learning model*** surpasses the limitations of hand-crafted priors by learning the underlying representations and structures directly from the data [13]. For example, λ-Net [14], HD-Net [15], and TSA-Net [16] incorporate the end-to-end neural network to restore 3D data within seconds rather than hours. While deep learning models offer notable efficiency and performance advantages, challenges related to interpretability and flexibility persist [17]. ***3) Plug-and-play (PnP) model*** integrates a fixed pretrained deep prior into traditional optimization models [18] to achieve effective reconstruction. However, PnP-based methods face challenges in learning a specific denoiser tailored for reconstruction [19]. This limitation hampers their reconstruction performance, as they may not effectively adapt to the unique characteristics and complexities of each HSI dataset. ***4) Adaptive prior unfolding learning model*** iteratively learns the deep prior and updates the lin-

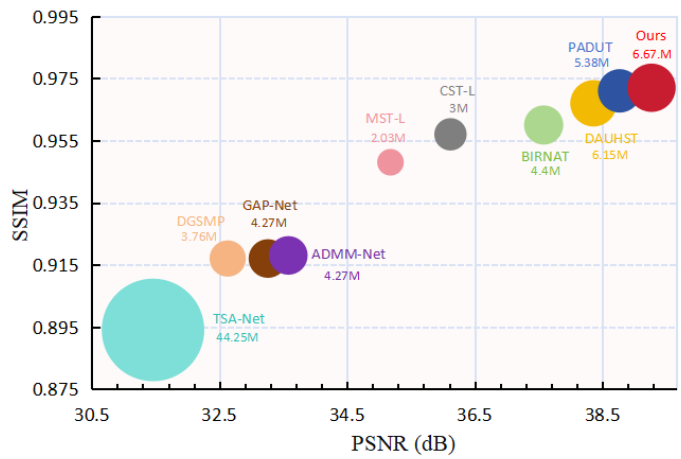

Fig. 1: Comparison of reconstruction performance (SSIM vs. PSNR) under different model training parameters. 3DNLT achieves the best performance among state-of-the-art methods.

[†]First author: alexzhouzhengz@gmail.com
[∗]Corresponding author: chenbingzhi@m.scnu.edu.cn

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

ear projection by developing an effective network within the objective function [20]. Deep unfolding methods offer powerful learnability and good interpretability, enabling effective reconstruction performance by systematically unveiling the learning stages [21].

However, existing deep unfolding-based approaches have often neglected either the spatial or spectral domain features [8]. Moreover, these approaches have treated spatial and spectral attention features as separate steps rather than considering them as a unified whole [7]. Consequently, there is a pressing need to develop methods that can effectively capture and leverage both spatial and spectral information in a cohesive manner. This holistic approach will enable more comprehensive and accurate reconstructions. In this paper, we present a deep unfolding 3D Non-Local Attention u-shaped Transformer (3DNLT) network to simultaneously consider spatial-spectral non-local attention features as a whole. To the best of our knowledge, our work is the first to investigate the 3D non-local attention mechanism in the deep unfolding methods. As demonstrated in Fig. 1, the proposed method achieves superior reconstruction performance compared to other approaches in terms of peak-signal-to-noise ratio (PSNR) [22] and structural similarity index measure (SSIM) [23], while maintaining a reasonable computation cost. Overall, our main contributions can be summarized as follows:

- We propose a deep unfolding 3D non-local transformer model including the mathematical linear projection module and deep 3D non-local attention prior module for spectral snapshot compressive imaging.
- We develop a 3D non-local mechanism to learn spatial-spectral attention features for reconstruction. The proposed 3D non-local transformer can capture intricate spatial structure and content, while also accurately modeling the correlation across the spectral bands.
- Extensive experiments of synthetic and real HSI reconstruction have validated the superiority of the proposed method over state-of-the-art approaches.

2 Methodology

2.1 Problem Formulation

In the CASSI system, the detector captures spatially modulated spectral information using an encoding aperture with a set pattern and then spectrally disperses it with a dispersion prism [24]. Considering a sequence $\{\boldsymbol{F}_b\}_{b=1}^{B} \in \mathbb{R}^{H\times W}$, where B is the number of HSI bands. These frames are modulated by a mask $\boldsymbol{M} \in \mathbb{R}^{H\times W}$:

$$\boldsymbol{F}_b^{'} = \boldsymbol{M} \odot \boldsymbol{F}_b \tag{1}$$

where $\boldsymbol{F}_b^{'}$ means the modulated HSI frames and $\odot$ denotes the element-wise multiplication. Next, the frames $\boldsymbol{F}_b^{'}$ are shifted horizontally according to the dispersion function s. Consequently, the modulated HSI frames $\boldsymbol{F}_b^{'} \in \mathbb{R}^{H\times W}$ are compressed into a form of coded measurement as follows:

$$\boldsymbol{G}(m,n) = \sum_{b=1}^{B} \boldsymbol{F}_b^{'}(m, n+s(b)) + \boldsymbol{N} \tag{2}$$

where m and n denote the spatial coordinates. Then $\boldsymbol{N} \in \mathbb{R}^{H\times(W+B-1)}$ and $\boldsymbol{G} \in \mathbb{R}^{H\times(W+B-1)}$ denote the noise and the compressed measurement. Therefore, the overall imaging model is formulated as:

$$\boldsymbol{g} = \boldsymbol{\Phi}\boldsymbol{f} + \boldsymbol{n}. \tag{3}$$

where the vectorization of a shifted version of coded aperture $\boldsymbol{M}$, $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{N}$ are denoted $\boldsymbol{\Phi} \in \mathbb{R}^{H(W+B-1)\times HWB}$, $\boldsymbol{f} \in \mathbb{R}^{HWB}$, $\boldsymbol{g} \in \mathbb{R}^{H(W+B-1)}$, and $\boldsymbol{n} \in \mathbb{R}^{H(W+B-1)}$, respectively. In the CASSI system, spatial information is partially sacrificed to capture comprehensive spectral information, resulting in a fused representation of spatial and spectral data. Consequently, it becomes crucial to carefully account for the intricate relationship between spatial-spectral information when reconstructing HSI. This forms the core of our optimization process for improving 3D reconstruction.

2.2 Unfolding Algorithm

Inspired by the half quadratic splitting (HQS) [20], the HSI reconstruction can be treated as an optimization problem:

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \frac{1}{2}\|\boldsymbol{g} - \boldsymbol{\Phi}\boldsymbol{f}\|_2^2 + \lambda T(\boldsymbol{f}) \tag{4}$$

where the first term is the data fidelity term, $T(\boldsymbol{f})$ means the image prior term, and λ denotes the trade-off regularization parameter between data and prior terms. By introducing an auxiliary variable $\boldsymbol{h}$, Eq. (4) can be reformulated as:

$$(\hat{\boldsymbol{f}}, \hat{\boldsymbol{h}}) = \arg\min_{\boldsymbol{f},\boldsymbol{h}} \frac{1}{2}\|\boldsymbol{g} - \boldsymbol{\Phi}\boldsymbol{f}\|_2^2 + \lambda T(\boldsymbol{h}) + \frac{\nu}{2}\|\boldsymbol{h} - \boldsymbol{f}\|_2^2, \tag{5}$$

where ν means the penalty parameter. Then Eq. (5) can be solved by decoupling $\boldsymbol{f}$ and $\boldsymbol{h}$ into the following two separate iterative sub-problems:

$$\boldsymbol{f}^{(k+1)} = \arg\min_{\boldsymbol{f}} \|\boldsymbol{g} - \boldsymbol{\Phi}\boldsymbol{f}^{(k)}\|_2^2 + \nu\|\boldsymbol{h}^{(k)} - \boldsymbol{f}^{(k)}\|_2^2, \tag{6}$$

$$\boldsymbol{h}^{(k+1)} = \arg\min_{\boldsymbol{h}} \frac{\nu}{2}\|\boldsymbol{h}^{(k)} - \boldsymbol{f}^{(k+1)}\|_2^2 + \lambda T(\boldsymbol{h}^{(k)}), \tag{7}$$

2.3 Deep Unfolding 3D Non-local Transformer

Network Structure. In Fig. 2, we introduce a 3D non-local transformer (3DNLT) network to investigate the non-local spatial-spectral relationship in hyperspectral imaging (HSI). The linear projection (LP) module provides an explicit sub-optimal solution, while the denoiser (DN) module utilizes a u-shaped 3D non-local transformer network to capture deep spatial-spectral correlations in the HSI data. To mitigate information loss during training, the proposed 3D non-local

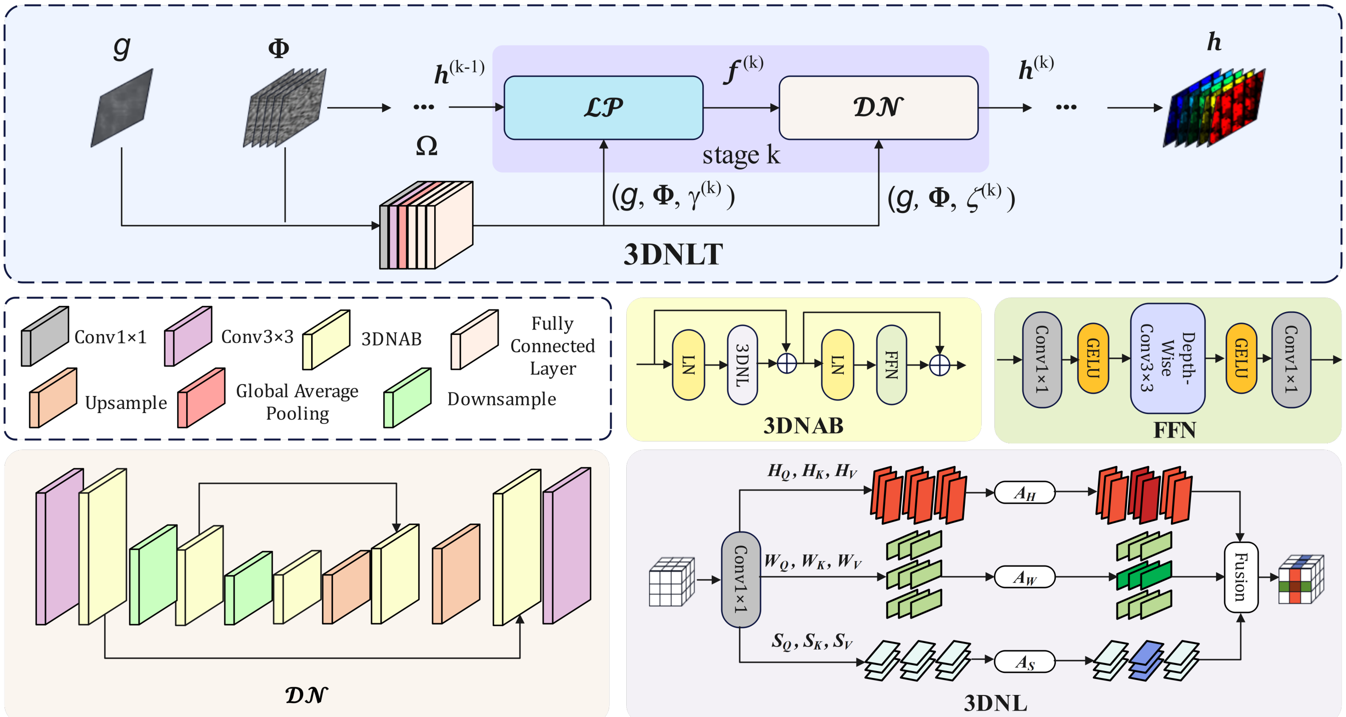

Fig. 2: The framework of deep unfolding 3D Non-Local Transformer (3DNLT) for hyperspectral compressive imaging.

attention block (3DNAB) incorporates residual connections. Within the 3DNAB, the 3D non-local (3DNL) attention module holistically combines non-local horizontal, vertical, and spectral attention mechanisms. The feed-forward network (FFN) employs various convolutional and activation layers to obtain channel-wise attention features.

Linear Projection (LP) Module. The HQS algorithm effectively separates the data term and the regularization term, enabling the solution of these two sub-problems in an alternating manner. In essence, the f-sub-problem in Eq. (6) has a closed-form solution as:

$$\boldsymbol{f}^{(k+1)} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \nu\boldsymbol{I})^{-1}(\boldsymbol{\Phi g} + \nu\boldsymbol{h}^{(k)}) \tag{8}$$

$$= \boldsymbol{h}^{(k)} + \frac{1}{1+\nu}\boldsymbol{\Phi}^T(\boldsymbol{\Phi\Phi}^T)^{-1}(\boldsymbol{g} - \boldsymbol{\Phi h}^{(k)}) \tag{9}$$

For the linear projection and denoiser modules, we employ a simple network $\boldsymbol{\Omega}$ to fuse the compressed measurement $\boldsymbol{g}$ and the sensing matrix $\boldsymbol{\Phi}$ as input, formulated by:

$$(\boldsymbol{\gamma}, \boldsymbol{\zeta}) = \boldsymbol{\Omega}(\boldsymbol{g}, \boldsymbol{\Phi}) \tag{10}$$

where $\boldsymbol{\Omega}$ comprises a Conv1×1, a branch of Conv3×3, a global average pooling, and three fully connected layers. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$ are dynamically determined at each stage. To facilitate the learnable solution of Eq. (9), we establish a convenient correspondence between $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$ at each stage. Based on this correspondence, we generate $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$ as inputs for the linear projection ($\mathcal{LP}$) and the 3DNLT denoiser ($\mathcal{DN}$), respectively. Thus, Eq. (6) and Eq. (7) can be transformed as:

$$\boldsymbol{f}^{(k+1)} = \mathcal{LP}(\boldsymbol{g}, \boldsymbol{h}^{(k)}, \boldsymbol{\gamma}^{(k+1)}, \boldsymbol{\Phi}) \tag{11}$$

$$\boldsymbol{h}^{(k+1)} = \mathcal{DN}(\boldsymbol{f}^{(k+1)}, \boldsymbol{\zeta}^{(k+1)}) \tag{12}$$

3D Non-Local (3DNL) Attention. In the spatial-spectral domain, we firstly extract the vertical, horizontal, spectral features using the convolutional network. Then we use Conv1×1 to obtain the query, key, and value of vertical representations $\boldsymbol{H_Q}$, $\boldsymbol{H_K}$, $\boldsymbol{H_V}$, horizontal representations $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, $\boldsymbol{W_V}$, and spectral representations $\boldsymbol{S_Q}$, $\boldsymbol{S_K}$, $\boldsymbol{S_V}$. The 3D non-local attention features along the orthogonal directions are calculated as:

$$\boldsymbol{A_H}(\boldsymbol{H_Q}, \boldsymbol{H_K}, \boldsymbol{H_V}) = softmax(\boldsymbol{H_Q H_K^T})\boldsymbol{H_V} \tag{13}$$

$$\boldsymbol{A_W}(\boldsymbol{W_Q}, \boldsymbol{W_K}, \boldsymbol{W_V}) = softmax(\boldsymbol{W_Q W_K^T})\boldsymbol{W_V} \tag{14}$$

$$\boldsymbol{A_S}(\boldsymbol{S_Q}, \boldsymbol{S_K}, \boldsymbol{S_V}) = softmax(\frac{\boldsymbol{S_K^T S_Q}}{\alpha})\boldsymbol{S_V} \tag{15}$$

where $\boldsymbol{A_H}$, $\boldsymbol{A_W}$, and $\boldsymbol{A_S}$ denote the non-local self-attention for vertical, horizontal, and spectral axes, respectively. In the spectral non-local self-attention, we introduce a learnable temperature parameter α to achieve an adaptive balance in the calculation of spectral attention scores.

Finally, the computation of the 3DNL attention features is carried out in a fusion module, which is formulated as:

$$\boldsymbol{A_{3D}} = \beta(\boldsymbol{A_H} + \boldsymbol{A_W}) + \boldsymbol{A_S} \tag{16}$$

where $\boldsymbol{A_{3D}}$ denotes the 3D non-local attention features, which includes vertical $\boldsymbol{A_H}$, horizontal $\boldsymbol{A_W}$, and spectral $\boldsymbol{A_S}$ non-local attention features, β refers to the learnable trade-off weight. In the experiments, we adopt a simple shallow neural network to adaptively obtain the parameter β.

Table 1: Performance (PSNR & SSIM) on the KAIST dataset. **Boldface** and underline indicate the best and second-best.

Methods	Params	GFLOPs	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
TwIST [12]	-	-	25.16 0.700	23.02 0.604	21.40 0.711	30.19 0.851	21.41 0.635	20.95 0.644	22.20 0.643	21.82 0.650	22.42 0.690	22.67 0.569	23.12 0.699
GAP-TV [9]	-	-	26.82 0.754	22.89 0.610	26.31 0.802	30.65 0.852	23.64 0.703	21.85 0.663	23.76 0.688	21.98 0.655	22.63 0.682	23.10 0.584	24.36 0.669
DeSCI [11]	-	-	27.13 0.748	23.04 0.620	26.62 0.818	34.96 0.897	23.94 0.706	22.38 0.683	24.45 0.743	22.03 0.673	24.56 0.732	23.59 0.587	25.27 0.721
λ-Net [14]	64.64M	117.98	30.10 0.849	28.49 0.805	27.73 0.870	37.01 0.934	26.19 0.817	28.64 0.853	26.47 0.806	26.09 0.831	27.50 0.826	27.13 0.816	28.53 0.841
HSSP [25]	-	-	31.48 0.858	31.09 0.842	28.96 0.823	34.56 0.902	28.53 0.808	30.83 0.877	28.71 0.824	30.09 0.881	30.43 0.868	28.78 0.842	30.35 0.852
DNU [21]	1.19M	163.48	31.72 0.863	31.13 0.846	29.99 0.845	35.34 0.908	29.03 0.833	30.87 0.887	28.99 0.839	30.13 0.885	31.03 0.876	29.14 0.849	30.74 0.863
DIP-HSI [18]	33.85M	64.42	32.68 0.890	27.26 0.833	31.30 0.914	40.54 0.962	29.79 0.900	30.39 0.887	28.18 0.839	29.44 0.885	34.51 0.876	28.51 0.849	31.26 0.863
TSA-Net [16]	44.25M	110.06	32.03 0.892	31.00 0.858	32.25 0.915	39.19 0.953	29.39 0.884	31.44 0.908	30.32 0.878	29.35 0.888	30.01 0.890	29.59 0.874	31.46 0.894
DGSMP [26]	3.76M	646.65	33.26 0.915	32.09 0.898	33.06 0.925	40.54 0.964	28.86 0.882	33.08 0.937	30.74 0.886	31.55 0.923	31.66 0.911	31.44 0.925	32.63 0.917
GAP-Net [24]	4.27M	78.58	33.74 0.911	33.26 0.900	34.28 0.929	41.03 0.967	31.44 0.919	32.40 0.925	32.27 0.902	30.46 0.905	33.51 0.915	30.24 0.895	33.26 0.917
ADMM-Net [27]	4.27M	78.58	34.12 0.918	33.62 0.902	35.04 0.931	41.15 0.966	31.82 0.922	32.54 0.924	32.42 0.896	30.74 0.907	33.75 0.915	30.68 0.895	33.58 0.918
HDNet [15]	2.37M	154.76	35.14 0.935	35.67 0.940	36.03 0.943	42.30 0.969	32.69 0.946	34.46 0.952	33.67 0.926	32.48 0.941	34.89 0.942	32.38 0.937	34.97 0.943
MST-L [28]	2.03M	28.15	35.40 0.941	35.87 0.944	36.51 0.953	42.27 0.973	32.77 0.947	34.80 0.955	33.66 0.925	32.67 0.948	35.39 0.949	32.50 0.941	35.18 0.948
MST++ [29]	1.33M	19.42	35.80 0.943	36.23 0.947	37.34 0.957	42.63 0.973	33.38 0.952	35.38 0.957	34.35 0.934	33.71 0.953	36.67 0.953	33.38 0.945	35.99 0.951
CST-L [30]	3.00M	40.01	35.96 0.949	36.84 0.955	38.16 0.962	42.44 0.975	33.25 0.955	35.72 0.963	34.86 0.944	34.34 0.961	36.51 0.957	33.09 0.945	36.12 0.957
BIRNAT [31]	4.40M	2122.66	36.79 0.951	37.89 0.957	40.61 0.971	46.94 0.985	35.42 0.964	35.30 0.959	36.58 0.955	33.96 0.956	39.47 0.970	32.80 0.938	37.58 0.960
DAUHST [8]	6.15M	79.50	37.25 0.958	39.02 0.967	41.05 0.971	46.15 0.983	35.80 0.969	37.08 0.970	37.57 0.963	35.10 0.966	40.02 0.970	34.59 0.956	38.36 0.967
PADUT [7]	5.38M	90.46	37.30 0.960	**40.30** **0.975**	42.19 0.976	46.15 **0.987**	36.21 0.972	37.23 **0.972**	37.76 0.964	35.30 **0.971**	40.73 0.976	34.52 0.960	38.77 0.971
Ours (3DNLT)	6.67M	112.62	**37.85** **0.964**	40.09 0.974	**42.54** **0.977**	**47.01** 0.986	**36.66** **0.973**	**37.36** 0.971	**38.50** **0.969**	**35.95** 0.969	**41.72** **0.977**	**35.04** **0.962**	**39.27** **0.972**

3 Experiments

3.1 Experimental Settings

Following the configurations of TSA-Net [16], we utilize a set of 28 wavelengths ranging from 450 nm to 650 nm for our experiments. These wavelengths are obtained through spectral interpolation manipulation of HSI.

In the experiments, we use the CAVE and KAIST datasets. The CAVE dataset consists of 32 HSIs with a spatial size of 512×512, while the KAIST dataset includes 30 HSIs with a spatial size of 2704×3376. The CAVE dataset is used for training, and 10 scenes from the KAIST dataset are used for testing. For real HSI reconstruction, we trained a separate model from scratch using the combined CAVE and KAIST datasets. To simulate real-world conditions, we introduce 11-bit shot noise to the simulated measurements during training. For evaluation, we use 5 authentic HSIs acquired with the CASSI system. We implemented the 3DNLT models using PyTorch with the Adam optimizer on RTX 3090 GPUs and trained the model 300 epochs with a learning rate of 4×10^{-4}.

For the comparison methods, we select several model-based methods including TwIST [12], GAP-TV [9], DeSCI [11], and HSSP [25]; deep learning-based methods such as λ-Net [14], TSA-Net [16], HDNet [15], MST-L [28], MST++ [29], CST-L [30], and BIRNAT [31]; PnP-based methods including DIP-HSI [18]; and deep unfolding-based methods such as HSSP [25], DNU [21], DGSMP [26], GAP-Net [24], ADMM-Net [27], DAUHST [8], and PADUT [7].

3.2 Simulation Reconstruction Results

The quantitative results of synthetic HSI reconstruction are presented in Table 1. We use PSNR and SSIM for reconstruction evaluation, and training parameters and GFLOPs for model complexity. For a fair comparison, deep unfolding methods undergo 12 stages of iteration under identical conditions. 3DNLT outperforms state-of-the-art methods in HSI reconstruction, achieving improvements of 0.5 dB and 0.91 dB compared to PADUT [7] and DAUHST [8] in PSNR, with

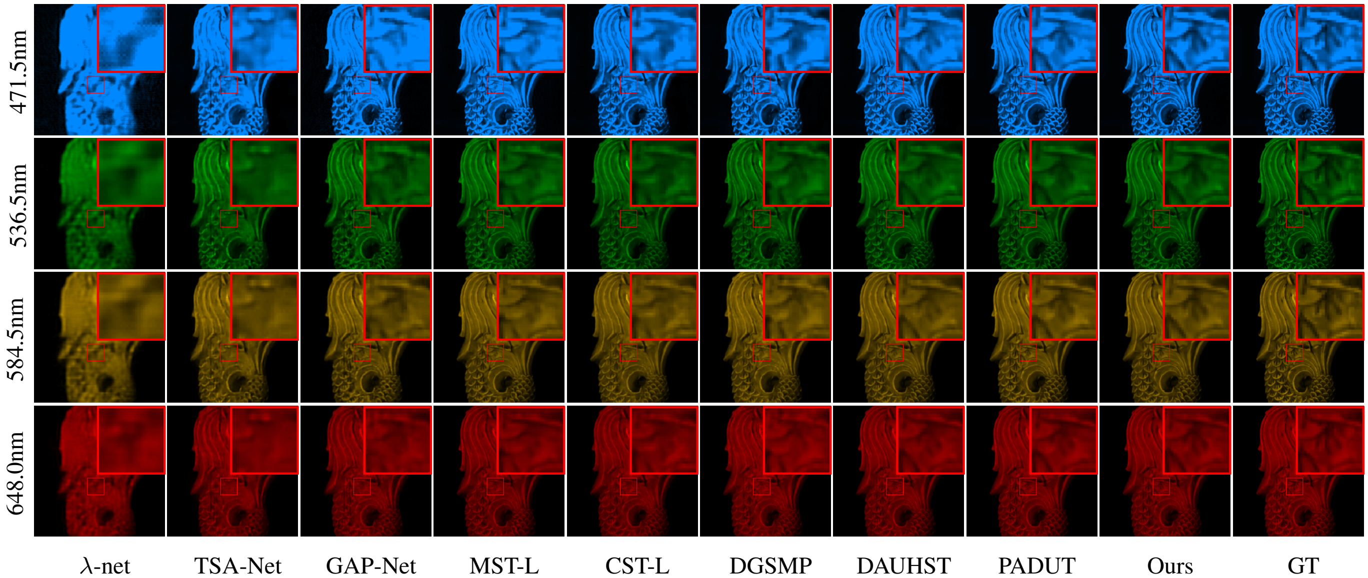

Fig. 3: Visual results of hyperspectral image reconstruction for Scene 10 on the KAIST dataset. Zoom in for better viewing.

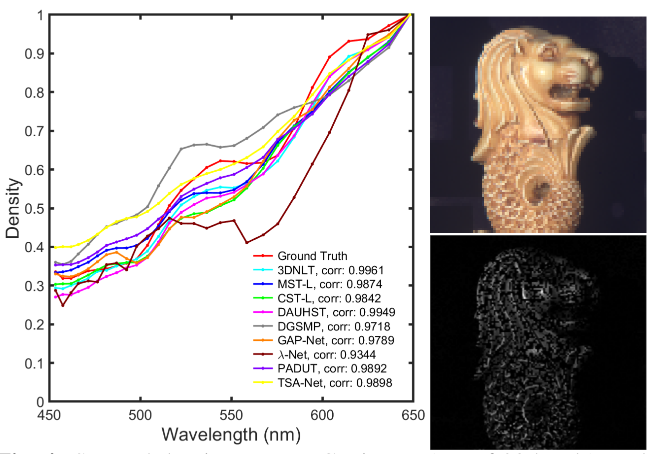

Fig. 4: Spectral density curve, RGB image (out of 28 bands), and compressed measurement of Scene 10 on the KAIST dataset, arranged from left to right and top to bottom.

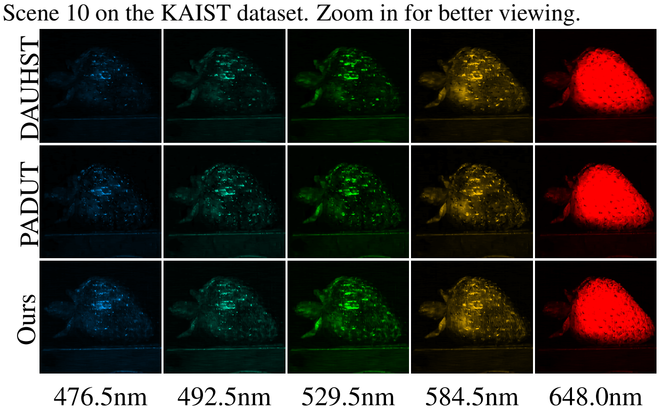

Fig. 5: Visual results of Real HSI reconstruction.

Table 2: Ablation studies of attention in the proposed method.

Baseline	HNA	VNA	SNA	Params (M)	GFLOPs	PSNR	SSIM
✓	✗	✗	✗	0.87	10.74	34.42	0.939
✓	✓	✗	✗	1.11	14.07	35.11	0.947
✓	✗	✓	✗	1.11	14.07	35.30	0.949
✓	✗	✗	✓	1.11	13.20	35.90	0.952
✓	✓	✓	✗	1.11	16.31	35.45	0.951
✓	✓	✗	✓	1.11	16.52	36.00	0.954
✓	✗	✓	✓	1.11	16.52	36.25	0.956
✓	✓	✓	✓	1.11	18.77	36.70	0.960

only a slight increase in computation cost.

As depicted in Fig. 3, compared to the blur results of DAUST and over-smoothness on PADUT, our 3DNLT effectively captures fine-grained texture details and accurately reconstructs the contents. The visualization results of the spectral density curve, as shown in Fig. 4, highlight the superior performance of our proposed method in recovering the spectral bands. Overall, our 3DNLT approach successfully restores spatial details and accurately reconstructs spectral bands, achieving satisfying results.

3.3 Real Reconstruction Results

The visualization results in Fig. 5 demonstrate the exceptional ability of our 3DNLT model to restore intricate structural details in real HSI reconstruction. In comparison to deep unfolding-based methods such as DAUHST and PADUT, our approach leverages the power of 3D non-local attention transformer, resulting in superior reconstruction performance. This represents a significant advancement in HSI reconstruction, highlighting the effectiveness of our proposed method.

3.4 Ablation Studies

Table 2 demonstrates the effectiveness of our proposed 3DNLT network. Incorporating the spectral non-local attention (SNA) module improves PSNR by 1.48 dB and SSIM by 0.013 compared to the baseline. SNA outperforms the horizontal non-local attention (HNA) and vertical non-local attention (VNA), highlighting the importance of spectral band features. The proposed 3D non-local attention mechanism achieves a gain of 2.28 dB in PSNR and 0.021 in SSIM, surpassing the individual non-local attention mechanisms (HNA, VNA, and SNA). These results confirm the superiority of our 3D non-local attention mechanism in restoring spectral and spatial details for HSI reconstruction.

4 Conclusion

In this paper, we proposed a deep unfolding 3D non-local transformer (3DNLT) network for hyperspectral compressive imaging. By incorporating a learnable half-quadratic splitting (HQS) algorithm and a 3D non-local attention u-shaped transformer, the network effectively captures spatial-spectral features and enhances reconstruction performance. Experimental results on synthetic and real hyperspectral images demonstrate the superior performance of the 3DNLT network compared to state-of-the-art methods. In future work, we will develop more deep priors or constraints into the network to enhance the reconstruction quality and robustness.

5 Acknowledgments

This work was supported in part by the NSFC fund (Grant No. 62302172 and 62106063), in part by the Guangdong University Young Innovative Talents Program Project (Grant No. 2023KQNCX020).

6 References

[1] Fengyi Zhang, Lin Zhang, Tianjun Zhang, and Dongqing Wang, "Adaptively hashing 3dluts for lightweight real-time image enhancement," in *ICME*, 2023, pp. 2771–2776.

[2] Yuyun Lian, Yongshan Zhang, Xuxiang Feng, Xinwei Jiang, and Zhihua Cai, "Low-rank constrained memory autoencoder for hyperspectral anomaly detection," in *ICASSP*, 2023, pp. 1–5.

[3] Xiaohui Huang, Yunfei Zhou, Xiaofei Yang, Xianhong Zhu, and Ke Wang, "Ss-tmnet: Spatial–spectral transformer network with multi-scale convolution for hyperspectral image classification," *RS*, vol. 15, no. 5, pp. 1206, 2023.

[4] Chen Luo, Shanshan Feng, Xiaofei Yang, Yunming Ye, Xutao Li, Baoquan Zhang, Zhihao Chen, and Yingling Quan, "Lwcdnet: A lightweight network for efficient cloud detection in remote sensing images," *IEEE TGRS*, vol. 60, pp. 1–16, 2022.

[5] Xiaohong Chen, Yongshan Zhang, Xuxiang Feng, Xinwei Jiang, and Zhihua Cai, "Spectral-spatial superpixel anchor graph-based clustering for hyperspectral imagery," *IEEE GRSL*, 2023.

[6] Yi Huang, Jiangtao Peng, Weiwei Sun, Na Chen, Qian Du, Yujie Ning, and Han Su, "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE TGRS*, vol. 60, pp. 1–13, 2022.

[7] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang, "Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction," in *ICCV*, 2023, pp. 12959–12968.

[8] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc V Gool, "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," *NeurIPS*, vol. 35, pp. 37749–37761, 2022.

[9] Xin Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *ICIP*, 2016, pp. 2539–2543.

[10] Yang Chen, Wenfei Cao, Li Pang, and Xiangyong Cao, "Hyperspectral image denoising with weighted nonlocal low-rank model and adaptive total variation regularization," *IEEE TGRS*, vol. 60, pp. 1–15, 2022.

[11] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai, "Rank minimization for snapshot compressive imaging," *IEEE TPAMI*, vol. 41, no. 12, pp. 2990–3006, 2018.

[12] José M Bioucas-Dias and Mário AT Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE TIP*, vol. 16, no. 12, pp. 2992–3004, 2007.

[13] Zheng Zhou, Yue Wu, and Yicong Zhou, "Consistent arbitrary style transfer using consistency training and self-attention module," *IEEE TNNLS*, pp. 1–12, 2023.

[14] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos, "l-net: Reconstruct hyperspectral images from a snapshot measurement," in *ICCV*, 2019, pp. 4059–4069.

[15] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool, "Hdnet: High-resolution dual-domain learning for spectral compressive imaging," in *CVPR*, 2022, pp. 17542–17551.

[16] Ziyi Meng, Jiawei Ma, and Xin Yuan, "End-to-end low cost compressive spectral imaging with spatial-spectral self-attention," in *ECCV*. Springer, 2020, pp. 187–204.

[17] Zheng Zhou, Yongyong Chen, and Yicong Zhou, "Deep dynamic memory augmented attentional dictionary learning for image denoising," *IEEE TCSVT*, vol. 33, no. 9, pp. 4784–4797, 2023.

[18] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan, "Self-supervised neural networks for spectral snapshot compressive imaging," in *ICCV*, 2021, pp. 2622–2631.

[19] Xiangyu Rui, Xiangyong Cao, Qi Xie, Zongsheng Yue, Qian Zhao, and Deyu Meng, "Learning an explicit weighting scheme for adapting complex hsi noise," in *CVPR*, 2021, pp. 6739–6748.

[20] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao, "Memory-augmented deep unfolding network for guided image super-resolution," *IJCV*, vol. 131, no. 1, pp. 215–242, 2023.

[21] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang, "Dnu: Deep non-local unrolling for computational spectral imaging," in *CVPR*, 2020, pp. 1661–1671.

[22] Zheng Zhou, Yue Wu, Xiaofei Yang, and Yicong Zhou, "Neural style transfer with adaptive auto-correlation alignment loss," *SPL*, vol. 29, pp. 1027–1031, 2022.

[23] Bo Jiang, Yao Lu, Xiaosheng Chen, Xinhai Lu, and Guangming Lu, "Graph attention in attention network for image denoising," *IEEE TSMC*, vol. 53, no. 11, pp. 7077–7088, 2023.

[24] Ziyi Meng, Xin Yuan, and Shirin Jalali, "Deep unfolding for snapshot compressive imaging," *IJCV*, vol. 131, no. 11, pp. 2933–2958, 2023.

[25] Lizhi Wang, Chen Sun, Ying Fu, Min H Kim, and Hua Huang, "Hyperspectral image reconstruction using a deep spatial-spectral prior," in *CVPR*, 2019, pp. 8032–8041.

[26] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi, "Deep gaussian scale mixture prior for spectral compressive imaging," in *CVPR*, 2021, pp. 16216–16225.

[27] Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan, "Deep tensor admm-net for snapshot compressive imaging," in *ICCV*, 2019, pp. 10223–10232.

[28] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool, "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *CVPR*, 2022, pp. 17502–17511.

[29] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool, "Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *CVPR*, 2022, pp. 745–755.

[30] Wulian Yun, Mengshi Qi, Chuanming Wang, Huiyuan Fu, and Huadong Ma, "Coarse-to-fine video denoising with dual-stage spatial-channel transformer," *arXiv preprint arXiv:2205.00214*, 2022.

[31] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan, "Birnat: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging," in *ECCV*. Springer, 2020, pp. 258–275.