Real Noise Decoupling for Hyperspectral Image Denoising

Yingkai Zhang[1], **Tao Zhang**[2], **Jing Nie**[1], **Ying Fu**[1*]

[1]Beijing Institute of Technology, Beijing, China
[2]Hangzhou Dianzi University, Hangzhou, China
zhangyingkai@bit.edu.cn, tzhang@hdu.edu.cn, {3420235028, fuying}@bit.edu.cn

Abstract

Hyperspectral image (HSI) denoising is a crucial step in enhancing the quality of HSIs. Noise modeling methods can fit noise distributions to generate synthetic HSIs to train denoising networks. However, the noise in captured HSIs is usually complex and difficult to model accurately, which significantly limits the effectiveness of these approaches. In this paper, we propose a **multi-stage noise-decoupling framework** that decomposes complex noise into explicitly modeled and implicitly modeled components. This decoupling reduces the complexity of noise and enhances the learnability of HSI denoising methods when applied to real paired data. Specifically, for **explicitly modeled noise**, we utilize an existing noise model to generate paired data for pre-training a denoising network, equipping it with prior knowledge to handle the explicitly modeled noise effectively. For **implicitly modeled noise**, we introduce a high-frequency wavelet guided network. Leveraging the prior knowledge from the pre-trained module, this network adaptively extracts high-frequency features to target and remove the implicitly modeled noise from real paired HSIs. Furthermore, to effectively eliminate all noise components and mitigate error accumulation across stages, a **multi-stage learning strategy**, comprising separate pre-training and joint fine-tuning, is employed to optimize the entire framework. Extensive experiments on public and our captured datasets demonstrate that our proposed framework outperforms state-of-the-art methods, effectively handling complex real-world noise and significantly enhancing HSI quality.

Introduction

Hyperspectral image (HSI) provides a wealth of spectral information, making it indispensable in a variety of applications, *e.g.*, remote sensing (Blackburn 2007; Thenkabail and Lyon 2016; Zhang and Huang 2010), classification (Azar et al. 2020; Cao et al. 2019), and recognition (Pan et al. 2003; Kim et al. 2012). Despite the potential of HSI, its quality is often compromised by various real-world noises due to the limitations of imaging technology and the complexity of the capture environment. Therefore, a robust and effective denoising solution is essential.

Recently, deep learning has emerged as a powerful alternative for HSI denoising, enabling the direct learning of mappings from noisy to clean images (Zhang et al. 2023; Lai, Yan, and Fu 2023; Xiao et al. 2024). Despite the remarkable progress of learning-based methods, they still face learning bottlenecks. The complex mapping from real-world noisy images to their clean counterparts, driven by intricate noise distributions and the scarcity of high-quality paired training data, remains challenging to learn, thereby limiting their denoising performance on real-world HSIs.

To tackle these challenges, particularly the scarcity of data, one prominent line of research focuses on noise modeling to synthesize realistic HSIs (Zhang, Fu, and Zhang 2022). However, real-world noise is complex and difficult to model accurately, creating a domain gap between synthetic and real data that undermines the effectiveness of these methods. This difficulty persists even for well-understood types like readout and stripe noise, often due to inaccurate parameter fitting or the presence of unknown noise components. As illustrated in Figure 1, this discrepancy is significant: visually, many regions in the real noise cannot be properly fitted by the model (d-f), and quantitatively, there are notable PSNR differences between the real and synthetic HSIs. This domain gap severely degrades the quality of the synthetic data, making it challenging for models to learn mappings that generalize effectively to real-world scenarios.

To address these limitations, we propose a novel **multi-stage noise-decoupling framework**. Our key idea is to decompose complex real-world noise into two parts: an **explicitly modeled** component that can be described by physical noise models, and an **implicitly modeled** component, which includes residuals from inaccurate fitting and other unknown noise sources. By tackling these components separately, our framework simplifies the learning task at each stage and enhances the overall denoising performance. Specifically, for **explicitly modeled noise**, we pre-train a state-of-the-art (SOTA) denoising network on synthetic data generated from an SOTA noise model. We aim to equip it with prior knowledge for handling these well-defined noise patterns, which is essential for the subsequent decoupling step. Furthermore, for **implicitly modeled noise** (*i.e.*, poorly fitted or unknown components), we introduce a high-frequency wavelet guided network motivated by the high correlation in the high-frequency of residuals between real and synthetic noise (Figures 1(a-c)). Leveraging the information-preserving nature of the wavelet transform, this network adaptively extracts

*Corresponding author

high-frequency noise features from spectral-spatial information. By utilizing the prior knowledge from the pre-trained denoising network, the high-frequency wavelet guided network is capable of removing the implicitly modeled noise decoupled from the real complex noise, based on real paired data. To effectively eliminate all noise components and mitigate error accumulation across stages, we develop a **multi-stage learning strategy** to guide the network. The strategy consists of three stages: 1) pre-training for explicitly modeled noise removal, 2) learning for implicitly modeled noise removal, and 3) fine-tuning for noise accumulation error and whole noise removal. The strategy focuses on removing noise, improving spectral fidelity, and alleviating noise accumulation error. Through the synergy of our multi-stage network and learning strategy, our noise-decoupling framework robustly removes complex real-world noise and is not limited to improving upon the existing SOTA noise models and denoising networks. Extensive experiments on public and our captured datasets demonstrate the superior performance of our approach compared with other SOTA methods. In summary, the contributions of our paper are as follows:

- We propose a multi-stage noise-decoupling framework to effectively disentangle and remove complex noise components, including explicitly and implicitly modeled noise, based on the real paired data.
- We introduce a high-frequency wavelet guidance for adaptively extracting high-frequency features to suppress implicitly modeled noise.
- We develop a multi-stage learning strategy to remove the real-world noise by separating pre-training and joint fine-tuning, thereby mitigating the noise accumulation error.

Related Work

In this section, we review the two most relevant areas of work: HSI denoising methods and noise modeling methods.

HSI Denoising Methods

HSI denoising is a fundamental task in hyperspectral image processing, aiming to remove noise components while preserving the underlying clean information (He et al. 2019; Shi et al. 2021). Existing methods can be broadly categorized into two groups: traditional optimization-based methods and deep learning-based methods.

Traditional optimization-based methods typically formulate the denoising task as an optimization problem, which is addressed by imposing various handcrafted regularizations (Fu et al. 2017; Wei and Fu 2019; Peng et al. 2020). However, these optimization-based methods rely on handcrafted priors, which cannot sufficiently represent the nonlinearity and complexity of various realistic HSIs.

Recently, deep learning has developed rapidly (Zhang et al. 2025; Tian, Fu, and Zhang 2023; Jiang et al. 2024) and has been applied to learning denoising mappings in a purely data-driven manner in several works (Xiao, Liu, and Wei 2024; Zhang, Fu, and Zhang 2024; Zou et al. 2025). Despite significant advancements in network design, existing learning-based methods still face challenges in handling complex data mappings due to intricate noise distributions and the scarcity of real data, which ultimately degrade their denoising performance.

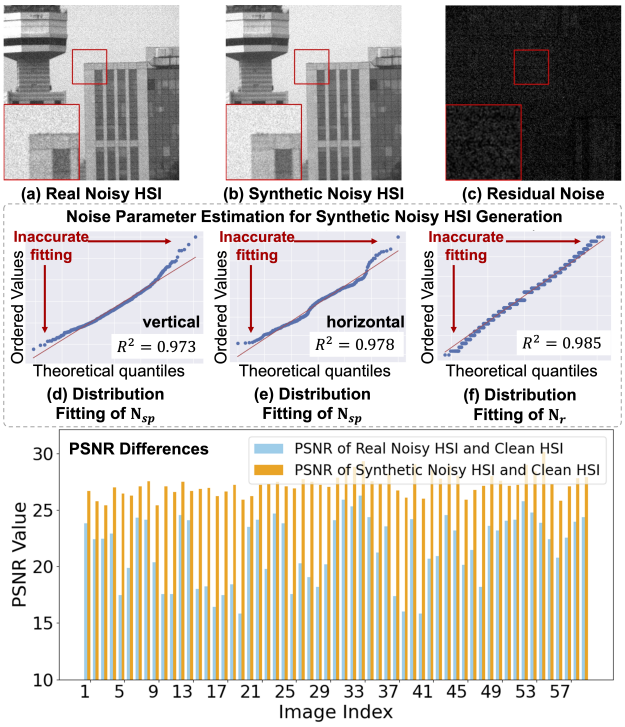

Figure 1: The significant differences between real noisy HSIs and synthetic noisy HSIs generated by the noise model (Zhang, Fu, and Zhang 2022).

Noise Modeling

Existing HSI denoising learning-based methods are typically data-driven. Thus, paired clean and noisy data are vital for training and evaluating the performance of denoising networks. There are two main strategies to address this issue. One way is to collect real paired HSI data. However, there still exists a small number of real paired HSI datasets, and even learning-based methods struggle with the complex data mapping between clean and noisy HSIs.

Another way is to generate more realistic data using noise models (Chen et al. 2017; Li et al. 2025). Recently, Zhang *et al.* (Zhang, Fu, and Zhang 2022) propose a noise model that contains more noise sources for the scanning hyperspectral camera and estimate the noise parameters for exactly simulating the original real noise distribution. However, real-world complex noise is challenging to model accurately, leading to discrepancies between synthetic and real data, which adversely affect denoising performance. Thus, we propose a noise-decoupling framework to enhance learnability for real HSI data mappings by separating the noise components into explicitly and implicitly modeled noise.

Multi-Stage Noise-Decoupling Framework

In this section, we first introduce the formulation and motivation behind our proposed framework. Then, we describe

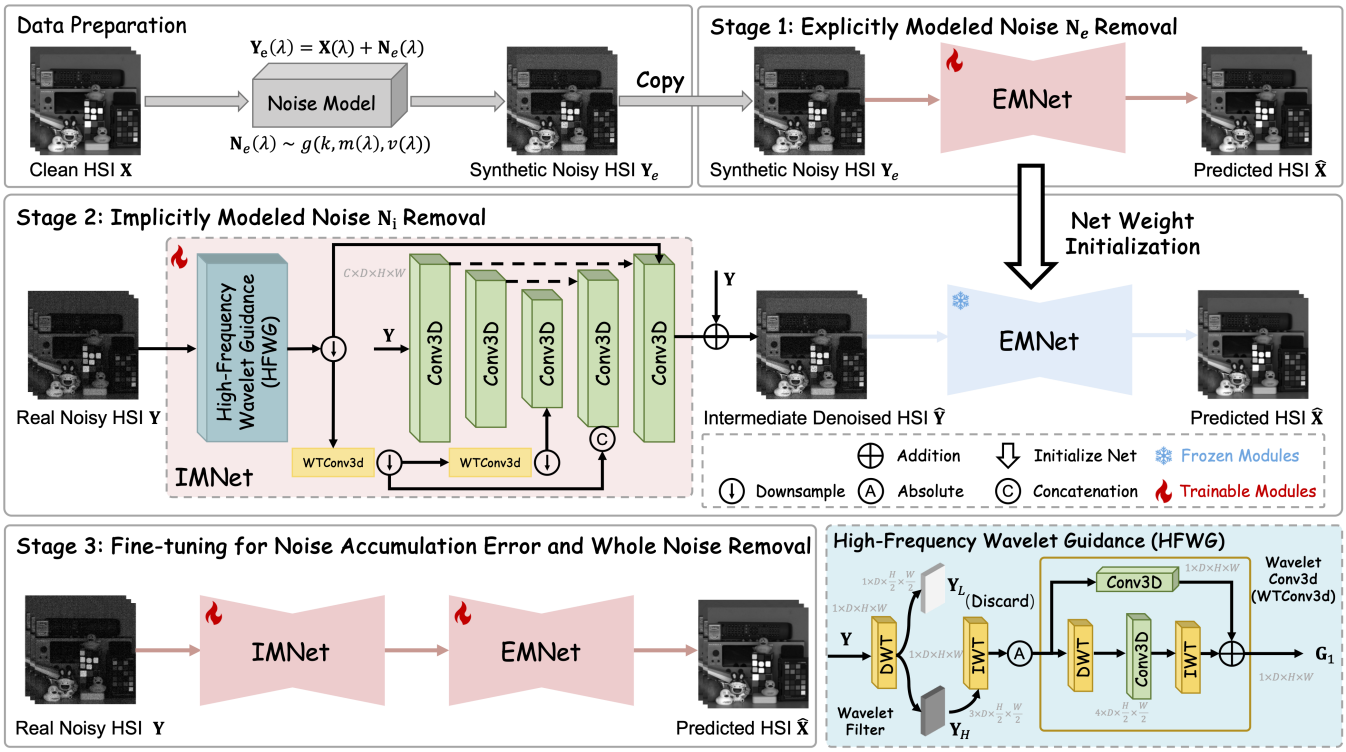

Figure 2: The overall multi-stage noise-decoupling framework. Stage 1: EMNet removes explicitly modeled noise based on the noise model, guiding the IMNet (Stage 2) to adaptively eliminate implicitly modeled noise. Stage 3: Entire fine-tuning for mitigating noise accumulation errors across stages and effectively removing real-world noise.

the framework in detail, including both explicitly and implicitly modeled noise removal, along with the multi-stage learning strategy. The framework is shown in Figure 2.

Formulation and Motivation

Suppose the clean HSI is denoted as $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$, where H, W, and D represent the height, width, and number of spectral bands, respectively. The linear model of the camera between the clean and noisy HSIs can be formulated as:

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{D\times H\times W}$ is the noisy HSI. Thus, we aim to remove the noise $\mathbf{N}$ from the noisy HSI $\mathbf{Y}$ to restore $\mathbf{X}$.

As discussed in (Healey and Kondepudy 1994; Schott 2007; Zhang, Fu, and Zhang 2022), the noise $\mathbf{N}$ is commonly categorized into types, such as photon shot noise $\mathbf{N}_s$, readout noise $\mathbf{N}_r$, and stripe noise $\mathbf{N}_{sp}$. These components are typically modeled using statistical distributions, such as the Poisson distribution for photon shot noise and the Gaussian distribution for other noise. However, these idealized models often fail to capture the full complexity of real-world noise. This discrepancy, arising from simplified assumptions or uncharacterized noise sources, is evident both visually and quantitatively. For example, as shown in Figures 1(d-f), there are significant regions where the model cannot properly fit the real noise, resulting in a notable PSNR gap between the synthetic and real noisy images. Thus, the Eq. (1) can be further formulated as:

$$\mathbf{N} = \mathbf{N}_e + \mathbf{N}_i, \tag{2}$$

where $\mathbf{N}_e$ denotes the explicitly modeled noise components, and $\mathbf{N}_i$ denotes the implicitly modeled noise components.

Therefore, we propose a multi-stage noise-decoupling framework that separates real noise $\mathbf{N}$ into explicitly modeled noise $\mathbf{N}_e$ and implicitly modeled noise $\mathbf{N}_i$, thereby reducing the difficulty of learning complex mappings. Specifically, we leverage an existing noise model to generate synthetic data for training a network, termed EMNet in our framework, to handle the removal of $\mathbf{N}_e$ in the first stage. Then, to tackle $\mathbf{N}_i$, we propose a high-frequency wavelet guided network, termed IMNet, for the second stage. The design of IMNet is motivated by our observation in Figures 1(a-c) that the residual between real and synthetic noise exhibits a high correlation in its high-frequency components. Leveraging the information-preserving property of the wavelet transform, the wavelet guidance of IMNet adaptively extracts these high-frequency features from the spatial-spectral domain of real noisy HSI. The primary role of the IMNet is to first remove the complex implicit modeled noise $\mathbf{N}_i$, preparing a cleaner image so that the EMNet can effectively remove the remaining explicit modeled noise $\mathbf{N}_e$. Finally, the entire network is jointly fine-tuned to mitigate noise accumulation errors across stages, and during the testing phase, the IMNet and EMNet are sequentially applied to effectively remove the real-world noise.

Explicitly Modeled Noise Removal

In this stage, we aim to remove noise components, $\mathbf{N}_e$ that can be explicitly modeled. Thus, we generate noisy HSIs based on noise model and pre-train the denoising network to remove noise $\mathbf{N}_e$. Besides, we can leverage this prior knowledge for further complex noise decoupling and removal.

Data Preparation. Our process begins with the utilization of the physical noise model in (Zhang, Fu, and Zhang 2022) to generate synthetic noisy HSIs. The noisy HSIs based on explicitly modeled noise can be formulated as:

$$\begin{aligned} \mathbf{Y}_e(\lambda) &= \mathbf{X}(\lambda) + \mathbf{N}_e(\lambda), \\ \mathbf{N}_e(\lambda) &\sim g(k, m(\lambda), v(\lambda)), \end{aligned} \tag{3}$$

where $\mathbf{Y}_e(\lambda)$ is the synthetic noisy HSI in λ-th band, and $\mathbf{N}_e(\lambda)$ is the explicitly modeled noise components in λ-th band. $g(k, m(\lambda), v(\lambda))$ denotes the Poisson/Gaussian distribution with parameters system gain k, mean $m(\lambda)$, and variance $v(\lambda)$ for λ-th band. More details about the noise model and calibration are shown in the supplementary material.

Noise Removal Network. Post generating synthetic noisy HSI, $\mathbf{Y}_e$, we can now pre-train the denoising network to remove explicitly modeled noise $\mathbf{N}_e$. We choose recent state-of-the-art networks, which have shown promising results in HSI denoising, as the explicitly modeled noise removal network, termed EMNet in our framework. Thus, the denoised HSI, $\hat{\mathbf{X}}$ can be formulated as:

$$\hat{\mathbf{X}} = f_{EMNet}(\mathbf{Y}_e; \theta), \tag{4}$$

where $f_{EMNet}(\circ; \theta)$ denotes the EMNet and θ is its parameters to be optimized.

Implicitly Modeled Noise Removal

In this stage, we aim to remove noise components that cannot be explicitly modeled, *i.e.*, $\mathbf{N}_i$, with the assistance of the prior knowledge from the pre-trained EMNet. Specifically, we first apply a wavelet filter to extract high-frequency information from the real noisy HSI. Next, we perform wavelet 3D convolution in the wavelet domain, integrating spatial and spectral information to generate multi-scale high-frequency guidance. This multi-scale guidance is then incorporated into the decoder of a 3D U-Net (Ronneberger, Fischer, and Brox 2015) as IMNet to guide the noise removal.

Inspired by the high correlation in high-frequency components, we introduce the high-frequency wavelet guidance (HFWG) to adaptively extract high-frequency components to guide the suppression of implicitly modeled noise. As shown in Figure 2, we first apply a wavelet filter to extract the high-frequency features:

$$\begin{aligned} \{\mathbf{Y}_L, \mathbf{Y}_H\} &= \text{DWT}(\mathbf{Y}), \\ \hat{\mathbf{G}} &= \text{abs}(\text{IWT}(\mathbf{Y}_H)), \end{aligned} \tag{5}$$

where DWT (Discrete Wavelet Transformation) is used to filter the low- and high-frequency components of the noisy image $\mathbf{Y}$ and IWT denotes the Inverse Wavelet Transform. $\mathbf{Y}_L, \mathbf{Y}_H$ denotes the low- and high-frequency components. Inspired by the design of wavelet 2D convolution in (Finder et al. 2024), we adopt this approach to further adjust high-frequency guidance in the wavelet domain, incorporating 3D spectral-spatial noise information. The process of wavelet 3D convolution (WTConv3d) can be formulated as:

$$\begin{aligned} \{\hat{\mathbf{G}}_L, \hat{\mathbf{G}}_H\} &= \text{DWT}(\hat{\mathbf{G}}), \\ \mathbf{G}_1 &= \text{IWT}(Conv3d(\hat{\mathbf{G}}_L, \hat{\mathbf{G}}_H) \cdot \theta_w), \end{aligned} \tag{6}$$

where θ_w is the learnable parameters to be optimized. We then utilize the WTConv3d to generate the multi-scale noise feature guidance $\mathbf{G}_i, i \in \{1, 2, 3\}$ incorporated into the decoder of the backbone, 3D Unet, to suppress the implicitly modeled noise. The whole process can be formulated as:

$$\hat{\mathbf{Y}} = f_{IMNet}(\mathbf{Y}, \mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3; \theta), \tag{7}$$

where $f_{IMNet}(\circ; \theta)$ denotes the IMNet to learn a residual for noise removal with the guidance of HFWG.

Dataset	Paired	Sensor	Bands	Size	Ratios
Urban	✗	Hydice	210	1	*N/A*
LHSI	✓	SPECIM FX10/IQ	64	14	15
RealHSI	✓	SOC710-VP	34	59	50
MEHSI	✓	SOC710-VP	34	303	20,50,100

Table 1: Summary of existing datasets and our multi-exposure dataset.

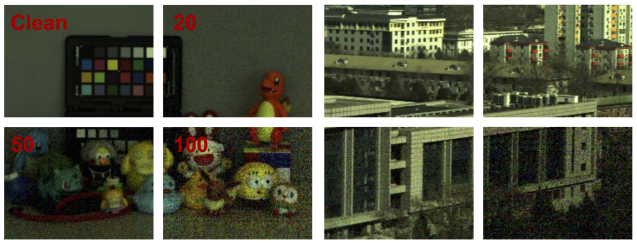

Figure 3: Example scenes in MEHSI dataset.

Multi-Stage Learning Strategy

Pre-training for Explicitly Modeled Noise Removal. We apply the Charbonnier loss (Charbonnier et al. 1994) to constrain the EMNet to focus on explicitly modeled noise removal:

$$\mathcal{L}_c = \sqrt{||\mathbf{X} - \hat{\mathbf{X}}||^2 + \epsilon^2}, \tag{8}$$

where $\mathbf{X}$ is the ground truth clean image, $\hat{\mathbf{X}}$ is the predicted clean image, and $\epsilon = 10^{-3}$ is a constant.

Learning for Implicitly Modeled Noise Removal. Due to the unknown noise distribution, we cannot acquire the ground truth only with explicitly modeled noise, but we can utilize the pre-trained EMNet as a discriminator with prior knowledge to distinguish the distribution of explicitly modeled noise based on real paired data. Thus, we freeze the EMNet during this training stage and also apply $\mathcal{L}_c$ loss function to constrain the learning process to update the parameters in IMNet. Besides, to further constrain the output of IMNet, we adopt the Kullback-Leibler divergence loss function to enforce distribution consistency between its output and the corresponding synthetic data:

$$\mathcal{L}_k = \sum_i p(\mathbf{Y}_i) \log \frac{p(\mathbf{Y}_i)}{q(\hat{\mathbf{Y}}_i)}, \tag{9}$$

Method	Venue	Ratio 20			Ratio 50			Ratio 100			Average		
		PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Noisy	-	28.18	0.919	13.451	23.59	0.777	24.677	19.51	0.543	38.028	23.76	0.746	25.385
QRNN3D	TNNLS'20	36.33	0.984	2.563	32.71	0.973	3.534	30.66	0.936	5.529	33.23	0.965	3.875
MAC-Net	TGRS'21	35.54	0.981	2.639	32.54	0.967	3.576	29.78	0.926	5.147	32.62	0.958	3.787
T3SC	NeurIPS'21	32.11	0.969	3.055	30.84	0.960	3.934	29.07	0.928	4.595	30.67	0.953	3.861
GRNet	TGRS'21	35.08	0.970	2.953	32.55	0.944	3.821	29.02	0.878	4.821	32.22	0.931	3.865
SST	AAAI'23	37.23	0.986	2.430	34.98	0.979	3.367	31.68	0.952	5.147	34.63	0.973	3.648
SERT	CVPR'23	37.59	0.986	2.361	35.27	0.980	3.351	31.68	0.955	4.989	34.85	0.974	3.567
HSDT	ICCV'23	35.99	0.984	2.654	32.78	0.973	3.672	31.37	0.946	5.613	33.38	0.967	3.980
TDSAT	TGRS'24	37.45	0.986	2.636	34.72	0.978	3.229	32.56	0.959	4.643	34.91	0.975	3.503
HIRDiff	CVPR'24	31.80	0.959	7.412	29.59	0.928	8.922	25.90	0.805	13.506	29.09	0.897	9.946
VolFormer	CVPR'25	37.37	0.986	2.198	34.90	0.980	2.867	32.40	0.955	4.095	34.89	0.974	3.053
Ours (TDSAT)	-	**38.44**	**0.988**	**1.822**	**36.64**	**0.983**	**2.149**	**34.01**	**0.972**	**2.714**	**36.36**	**0.981**	**2.228**

Table 2: Averaged results of different methods under different noise levels on the MEHSI dataset. The best and second-best results are highlighted in bold and underlined, respectively.

where $p(\mathbf{Y}_i)$ is the actual noise distribution and $q(\hat{\mathbf{Y}}_i)$ is the predicted noise distribution. The $p(\mathbf{Y}_i)$ can be generated by the explicitly modeled noise during the training process, *i.e.*, online, or before the training process, *i.e.*, offline.

Fine-tuning for Noise Accumulation Error and Whole Noise Removal. To further remove whole noise and eliminate the accumulated noise errors from each previous stage, we perform joint fine-tuning on the pre-trained IMNet and EMNet. Except for the above loss functions, we also apply the spectral consistency loss to guide the network to focus on spectral fidelity:

$$\mathcal{L}_s = 1 - \frac{1}{N}\sum_i \frac{\mathbf{X}_i \cdot \hat{\mathbf{X}}_i}{\|\mathbf{X}_i\|^2 \|\hat{\mathbf{X}}_i\|^2}, \tag{10}$$

where N is the number of pixels in the image. The overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_c + \lambda_k \mathcal{L}_k + \lambda_s \mathcal{L}_s, \tag{11}$$

where λ_k and λ_s are hyperparameters.

Experiments

In this section, we first introduce the datasets, implementation details, metrics for quantitative evaluation, and compared methods. Then, we provide quantitative and qualitative results. We further conduct ablation studies and discussions to evaluate the effectiveness of the proposed approach.

Datasets and Experimental Settings

Datasets. Real experiments are conducted on two datasets: the RealHSI dataset (Zhang, Fu, and Zhang 2022) and our collected Multi-Exposure real HSI denoising (MEHSI) dataset. The **RealHSI** dataset consists of 59 paired clean and noisy images, each captured under a single exposure ratio. Each HSI contains 34 bands from 400 nm to 700 nm with a size of 696×520 pixels. Due to the limited data volumes and limited noise level in the RealHSI dataset, we additionally capture a larger real dataset, **MEHSI**, with various noise levels. We capture 101 indoor and outdoor scenes with 3 different noise levels, totaling 303 pairs by the SOC710-VP hyperspectral camera. The image pre-processing is the same as that of the RealHSI dataset to obtain paired HSIs with 34 bands. The comparisons of datasets are shown in Table 1 and example scenes of MEHSI dataset are shown in Figure 3. More details can be found in the supplementary material.

Method	Venue	PSNR↑	SSIM↑	SAM↓
Noisy	-	23.26	0.760	17.329
QRNN3D	TNNLS'20	30.42	0.953	3.939
MAC-Net	TGRS'21	28.82	0.936	5.227
T3SC	NeurIPS'21	27.79	0.901	4.220
GRNet	TGRS'21	28.44	0.909	3.956
SST	AAAI'23	28.50	0.948	3.885
SERT	CVPR'23	29.01	0.939	3.202
HSDT	ICCV'23	31.24	0.958	3.751
TDSAT	TGRS'24	30.70	0.958	3.241
HIRDiff	CVPR'24	30.34	0.943	4.923
VolFormer	CVPR'25	29.33	0.930	3.231
Ours (HSDT)	-	**32.31**	**0.967**	**2.742**

Table 3: Averaged results of different methods on the RealHSI dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Implementation Details. We implement the proposed framework with Pytorch (Paszke et al. 2019). In detail, we crop overlapped 128×128 spatial regions from the paired data and augment them by random flipping and/or rotation. Following the settings in SERT (Li et al. 2023), we randomly choose 44 HSIs for training and the remaining 15 for testing in the RealHSI dataset. For MEHSI dataset, we randomly select 273 pairs for training and the remaining 30 for testing. The models are trained with Adam (Kingma and Ba 2014) (β_1 = 0.9 and β_2 = 0.999) for 200 epochs in RealHSI and 400 epochs in our MEHSI dataset. The initial learning rate and batch size are set to 1×10^{-4} and 1, respectively. Experiments are conducted on a single NVIDIA RTX 4090 GPU.

Evaluation Metrics. We utilize three quantitative image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Spectral Angle Mapping (SAM). Higher PSNR and SSIM values, combined with

Strategy	Loss	Pre-trained	Online	Offline	EMNet	IMNet	HFWG	PSNR↑	SSIM↑	SAM↓
End-to-End	$\mathcal{L}_c$	✗	✗	✗	✓	✗	✗	34.91	0.975	3.503
	$\mathcal{L}_c + \mathcal{L}_s$	✗	✗	✗	✓	✗	✗	35.33	0.975	2.471
	$\mathcal{L}_c + \mathcal{L}_s$	✗	✗	✗	✓	✓	✗	35.56	0.976	2.459
Multi-Stage	$\mathcal{L}_c$	✓	✗	✗	✓	✓	✗	35.66	<u>0.978</u>	3.240
	$\mathcal{L}_c + \mathcal{L}_s$	✓	✗	✗	✓	✓	✗	35.83	0.977	**2.228**
	$\mathcal{L}_c + \mathcal{L}_k + \mathcal{L}_s$	✓	✓	✗	✓	✓	✗	<u>36.16</u>	0.977	<u>2.253</u>
	$\mathcal{L}_c + \mathcal{L}_k + \mathcal{L}_s$	✓	✗	✓	✓	✓	✓ (3K Params)	36.04	0.976	2.364
(Ours)	$\mathcal{L}_c + \mathcal{L}_k + \mathcal{L}_s$	✓	✓	✗	✓	✓	✓ (3K Params)	**36.36**	**0.981**	**2.228**

Table 4: Ablation studies on the learning strategy and network designs towards higher performance.

lower SAM values, indicate better performance.

Compared Methods. We compare our proposed framework against SOTA deep learning-based HSI denoising methods, including QRNN3D (Wei, Fu, and Huang 2020), MAC-Net (Xiong et al. 2021), T3SC (Bodrito et al. 2021), GR-Net (Cao et al. 2021), SST (Li, Fu, and Zhang 2023), SERT (Li et al. 2023), HSDT (Lai, Yan, and Fu 2023), TD-SAT (Zhang et al. 2024), HIRDiff (Pang et al. 2024), and VolFormer (Yu and Gao 2025). All compared methods are trained with the same settings as ours for fair comparison.

Results on MEHSI Dataset

Quantitative Comparison. We show the quantitative results of various methods on the MEHSI dataset in Table 2. Among all methods, HIRDiff is an unsupervised method that can partially denoise compared to the original noisy image. However, its performance is suboptimal when denoising real noise, especially under varying noise levels. In contrast, our proposed method, based on TDSAT, achieves an average improvement of at least 1.45 dB in PSNR and 0.825 in SAM across multiple noise levels, compared to other learning-based approaches. Notably, our method effectively removes noise across different noise levels, demonstrating stability in real-world complex noise scenarios. This highlights the effectiveness of our framework.

Qualitative Comparison. We visualize the results in the upper part of Figure 4, where the exposure ratio of the input noisy image is 50. Most methods still leave significant residual noise and blur, particularly in the results of T3SC and GRNet. HIRDiff exhibits over-smoothing, which damages texture details. In comparison, our method effectively removes noise, resulting in a smaller error map. The superior performance is primarily attributed to our multi-stage noise-decoupling approach, which effectively reduces noise complexity, with the multi-stage learning strategy to enhance the learning ability of networks.

Results on RealHSI Dataset

Quantitative Comparison. We conduct experiments on the RealHSI dataset to evaluate the performance of our proposed method, as shown in Table 3. It can be observed that, compared to the second-best method, HSDT, our method achieves higher PSNR and SSIM and lower SAM values, demonstrating superior denoising performance.

Qualitative Comparison. A visual comparison of results for a scene is presented in the lower part of Figure 4. As shown, the comparison methods either retain more residual noise or overly smooth the image, leading to a larger error map. In contrast, our method achieves more effective noise removal while preserving texture details, making the result closer to the reference image.

Model	P(M)	F(G)	PSNR↑	Δ↑	SAM↓	Δ↓
HSDT	0.52	207	33.38	-	3.980	-
HSDT(*Pt*)	0.52	207	33.46	+0.08	3.989	+0.009
HSDT-L	0.98	389	33.60	+0.22	3.875	-0.105
Ours(HSDT)	0.62	260	**34.18**	**+0.80**	**2.443**	**-1.537**
VolFormer	2.41	109	34.89	-	3.053	-
VolFormer(*Pt*)	2.41	109	35.33	+0.44	2.820	-0.090
VolFormer-L	3.24	263	35.00	+0.11	3.089	+0.036
Ours(VolFormer)	2.80	320	**36.06**	**+1.17**	**2.164**	**-0.889**
TDSAT	1.09	501	34.91	-	3.503	-
TDSAT(*Pt*)	1.09	501	35.26	+0.35	3.251	-0.252
TDSAT-L	1.69	774	35.16	+0.25	3.399	-0.104
TDSAT(HSDT)	1.62	708	35.60	+0.69	2.343	-1.160
Ours(TDSAT)	1.31	621	**36.36**	**+1.45**	**2.228**	**-1.275**

Table 5: Comparison of the model complexity. 'P(M)' means Parameters(M) and 'F(G)' means Flops(G). '*Pt*' denotes that the model is pre-trained on synthetic data, and '$*$-L' means a larger model with more parameters.

Ablation Study

Learning Strategy. As shown in Table 4, we conduct ablation studies on the multi-stage learning strategy based on the network, TDSAT. First, we adopt an end-to-end learning strategy, applying $\mathcal{L}_c$ and $\mathcal{L}_s$ loss functions as constraints. We observe that $\mathcal{L}_s$ can effectively improve the SAM metric, enhancing spectral fidelity with a slight increase in PSNR performance. Next, we pre-train TDSAT and incorporate it into our framework as EMNet, resulting in at least a 0.3 dB improvement in PSNR. Additionally, as shown in Table 5, directly using the '*Pt*', *i.e.*, pre-training strategy, leads to performance gains, and incorporating it into our framework can make consistent improvement. Subsequently, we employ $\mathcal{L}_k$ to constrain the training of IMNet. By comparing noise synthetic strategies, we use the online strategy to enhance the learning performance of the framework. More details are shown in the supplementary material.

Network Design. As shown in Table 4, we first conduct an ablation study on the modules of our framework. The results demonstrate that incorporating the IMNet in a multi-

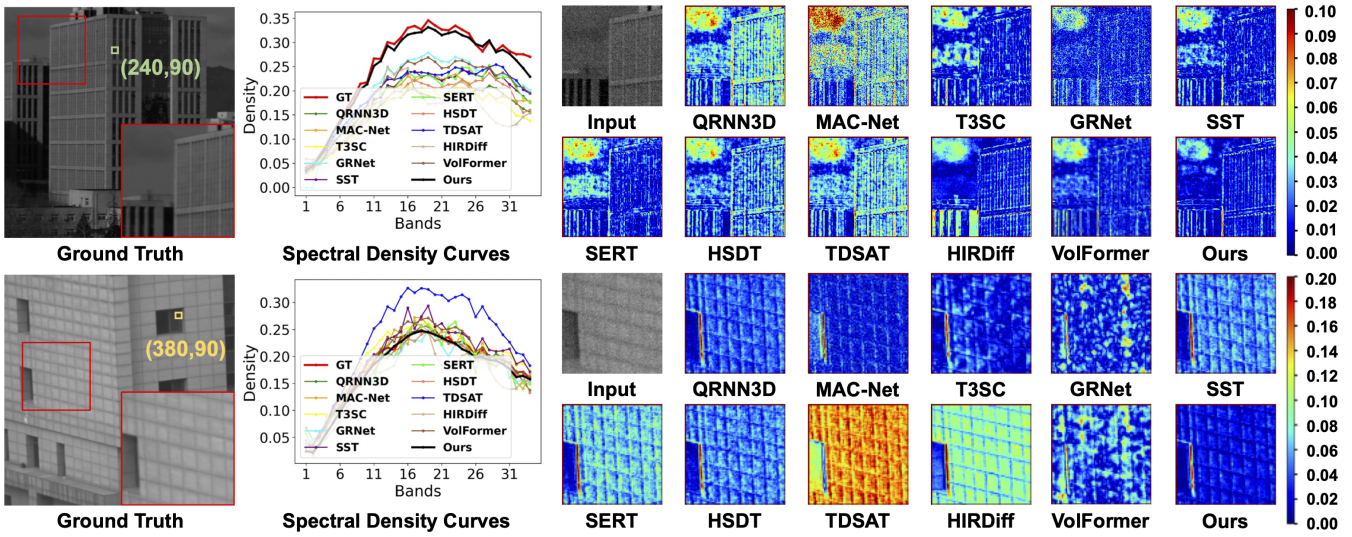

Figure 4: Visual quality comparisons of sample scene on MEHSI (up) and RealHSI (bottom) datasets with the spectral band in 510 nm, and 550 nm, respectively. The 'corr' of spectral density curves in the color box represents the correlation coefficient.

Methods	PSNR↑	SSIM↑	(λ_k, λ_s)	PSNR↑	SSIM↑
HSDT	34.62	0.940	(0.01,100)	31.41	0.965
TDSAT	36.38	0.942	(0.01,1)	31.92	0.967
VolFormer	34.42	0.925	(0.1,10)	31.06	0.965
Ours	**37.09**	**0.946**	(0.01,10)	**32.31**	**0.967**

Table 6: Comparison results on the LHSI dataset and analysis on the effect of hyperparameters.

stage configuration enhances denoising performance, even within an end-to-end training. Subsequently, we perform a further ablation on the High-Frequency Wavelet Guidance (HFWG). The results show that with a slight increase in parameters, PSNR improves by 0.2 dB. Additionally, in Table 5, using HSDT as the IMNet outperforms the single TDSAT, which validates our multi-stage noise-decoupling and learning strategy. Furthermore, our proposed IMNet with HFWG surpasses the larger HSDT model, demonstrating the effectiveness of our network design for noise removal.

Discussion

Spectral Consistency. In Figure 4, we plot the spectral density curves corresponding to the small green-boxed and yellow-boxed region on MEHSI and RealHSI datasets. The high correlation and significant overlap between our curve and the ground truth demonstrate the effectiveness of our method in preserving spectral consistency.

Model Complexity. As shown in Table 5, the 'model-L' refers to a variant with a larger parameter count. The results show that while a larger model achieves better performance compared to its smaller counterpart, our proposed framework based on these smaller models, with fewer parameters and computation, still outperforms the larger model, highlighting the effectiveness of our framework.

Generalization. Beyond demonstrating the generalization of the framework with various backbones on the MEHSI dataset (Table 5), we conduct further experiments on the LHSI dataset with other sensors, spectral resolution, and acquisition conditions (Table 6). The results also outperform competing methods, which underscores the generalization and generality of our model. It demonstrates that our framework is generic, exhibiting adaptability with diverse HSI hardware setups, and is not restricted to a specific sensor. More details about the generalization discussion on noise models are shown in the supplementary material.

Hyperparameters. As indicated in Table 6, performance exhibits sensitivity to the λ_k and λ_s, leading to variations. Based on the experimental results, we choose $\lambda_k = 0.01$ and $\lambda_s = 10$ in our experiments.

Conclusion

In this paper, we propose a multi-stage noise-decoupling framework for real hyperspectral image denoising. The framework integrates prior knowledge of a physical noise model to effectively decouple complex noise into explicitly and implicitly modeled components, thereby reducing the learning difficulty at each stage. We introduce high-frequency wavelet guidance for adaptive high-frequency feature extraction, enabling our network to focus on eliminating implicitly modeled noise by leveraging the explicit noise patterns learned in the previous stage. We develop a multi-stage learning strategy that comprises separate pre-training and joint fine-tuning of the networks to ensure effective training and mitigate error accumulation The quantitative and qualitative results on real datasets demonstrate that our method achieves superior denoising performance compared with state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62331006, 62171038 and 62506108), and the Fundamental Research Funds for the Central Universities.

References

Azar, S. G.; Meshgini, S.; Rezaii, T. Y.; and Beheshti, S. 2020. Hyperspectral image classification based on sparse modeling of spectral blocks. *Neurocomputing*, 407: 12–23.

Blackburn, G. A. 2007. Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany*, 58(4): 855–867.

Bodrito, T.; Zouaoui, A.; Chanussot, J.; and Mairal, J. 2021. A trainable spectral-spatial sparse coding model for hyperspectral image restoration. *Advances in Neural Information Processing Systems*, 34: 5430–5442.

Cao, X.; Fu, X.; Xu, C.; and Meng, D. 2021. Deep spatial-spectral global reasoning network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Cao, X.; Ge, Y.; Li, R.; Zhao, J.; and Jiao, L. 2019. Hyperspectral imagery classification with deep metric learning. *Neurocomputing*, 356: 217–227.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, 168–172. IEEE.

Chen, Y.; Cao, X.; Zhao, Q.; Meng, D.; and Xu, Z. 2017. Denoising hyperspectral image with non-iid noise structure. *IEEE Transactions on Cybernetics*, 48(3): 1054–1066.

Finder, S. E.; Amoyal, R.; Treister, E.; and Freifeld, O. 2024. Wavelet convolutions for large receptive fields. In *Proceedings of the European Conference on Computer Vision*, 363–380. Springer.

Fu, Y.; Lam, A.; Sato, I.; and Sato, Y. 2017. Adaptive spatial-spectral dictionary learning for hyperspectral image restoration. *International Journal of Computer Vision*, 122: 228–245.

He, W.; Yao, Q.; Li, C.; Yokoya, N.; and Zhao, Q. 2019. Non-local meets global: An integrated paradigm for hyperspectral denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6868–6877.

Healey, G. E.; and Kondepudy, R. 1994. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3): 267–276.

Jiang, T.; Chen, W.; Zhou, H.; He, J.; and Qi, P. 2024. Towards semi-supervised classification of abnormal spectrum signals based on deep learning. *Chinese Journal of Electronics*, 33(3): 721–731.

Kim, M. H.; Harvey, T. A.; Kittle, D. S.; Rushmeier, H.; Dorsey, J.; Prum, R. O.; and Brady, D. J. 2012. 3D imaging spectroscopy for measuring hyperspectral patterns on solid objects. *ACM Transactions on Graphics*, 31(4): 1–11.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, Z.; Yan, C.; and Fu, Y. 2023. Hybrid spectral denoising transformer with guided attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13065–13075.

Li, H.; Wu, Z.; Shao, R.; Zhang, T.; and Fu, Y. 2025. Noise calibration and spatial-frequency interactive network for stem image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21287–21296.

Li, M.; Fu, Y.; and Zhang, Y. 2023. Spatial-spectral transformer for hyperspectral image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1368–1376.

Li, M.; Liu, J.; Fu, Y.; Zhang, Y.; and Dou, D. 2023. Spectral enhanced rectangle transformer for hyperspectral image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5805–5814.

Pan, Z.; Healey, G.; Prasad, M.; and Tromberg, B. 2003. Face recognition in hyperspectral images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12): 1552–1560.

Pang, L.; Rui, X.; Cui, L.; Wang, H.; Meng, D.; and Cao, X. 2024. HIR-Diff: Unsupervised Hyperspectral Image Restoration Via Improved Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3005–3014.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Peng, J.; Xie, Q.; Zhao, Q.; Wang, Y.; Yee, L.; and Meng, D. 2020. Enhanced 3DTV regularization and its applications on HSI denoising and compressed sensing. *IEEE Transactions on Image Processing*, 29: 7889–7903.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.

Schott, J. R. 2007. *Remote sensing: the image chain approach*. Oxford University Press.

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.

Thenkabail, P. S.; and Lyon, J. G. 2016. *Hyperspectral remote sensing of vegetation*. CRC press.

Tian, Y.; Fu, Y.; and Zhang, J. 2023. Transformer-based under-sampled single-pixel imaging. *Chinese Journal of Electronics*, 32(5): 1151–1159.

Wei, K.; and Fu, Y. 2019. Low-rank Bayesian tensor factorization for hyperspectral image denoising. *Neurocomputing*, 331: 412–423.

Wei, K.; Fu, Y.; and Huang, H. 2020. 3-D quasi-recurrent neural network for hyperspectral image denoising. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1): 363–375.

Xiao, J.; Liu, Y.; and Wei, X. 2024. Region-Aware Sequence-to-Sequence Learning for Hyperspectral Denoising. In *Proceedings of the European Conference on Computer Vision*, 218–235. Springer.

Xiao, J.; Liu, Y.; Zhang, S.; and Wei, X. 2024. Bridging fourier and spatial-spectral domains for hyperspectral image denoising. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8489–8497.

Xiong, F.; Zhou, J.; Zhao, Q.; Lu, J.; and Qian, Y. 2021. MAC-Net: Model-aided nonlocal neural network for hyperspectral image denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Yu, D.; and Gao, Z. 2025. VolFormer: Explore More Comprehensive Cube Interaction for Hyperspectral Image Restoration and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28091–28101.

Zhang, L.; and Huang, X. 2010. Object-oriented subspace analysis for airborne hyperspectral remote sensing imagery. *Neurocomputing*, 73(4-6): 927–936.

Zhang, Q.; Dong, Y.; Zheng, Y.; Yu, H.; Song, M.; Zhang, L.; and Yuan, Q. 2024. Three-Dimension Spatial-Spectral Attention Transformer for Hyperspectral Image Denoising. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, Q.; Zheng, Y.; Yuan, Q.; Song, M.; Yu, H.; and Xiao, Y. 2023. Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, T.; Fu, Y.; and Zhang, J. 2022. Guided hyperspectral image denoising with realistic data. *International Journal of Computer Vision*, 130(11): 2885–2901.

Zhang, T.; Fu, Y.; and Zhang, J. 2024. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1): 303–312.

Zhang, Y.; Lai, Z.; Zhang, T.; Fu, Y.; and Zhou, C. 2025. Unaligned RGB Guided Hyperspectral Image Super-Resolution with Spatial-Spectral Concordance. *International Journal of Computer Vision*, 133(9): 6590–6610.

Zou, Y.; Fu, Y.; Zhang, Y.; Zhang, T.; Yan, C.; and Timofte, R. 2025. Calibration-Free Raw Image Denoising via Fine-Grained Noise Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.