HYDRA: HYbrid knowledge Distillation and spectral Reconstruction Algorithm for high channel hyperspectral camera applications

Christopher Thirgood
University of Surrey
Surrey, UK
c.thirgood@surrey.ac.uk

Oscar Mendez
University of Surrey
Surrey, UK
o.mendez@surrey.ac.uk

Erin Chao Ling
University of Surrey
Surrey, UK
chao.ling@surrey.ac.uk

Jon Storey
i3D Robotics
Kent, UK
Jstorey@i3drobotics.com

Simon Hadfield
University of Surrey
Surrey, UK
s.hadfield@surrey.ac.uk

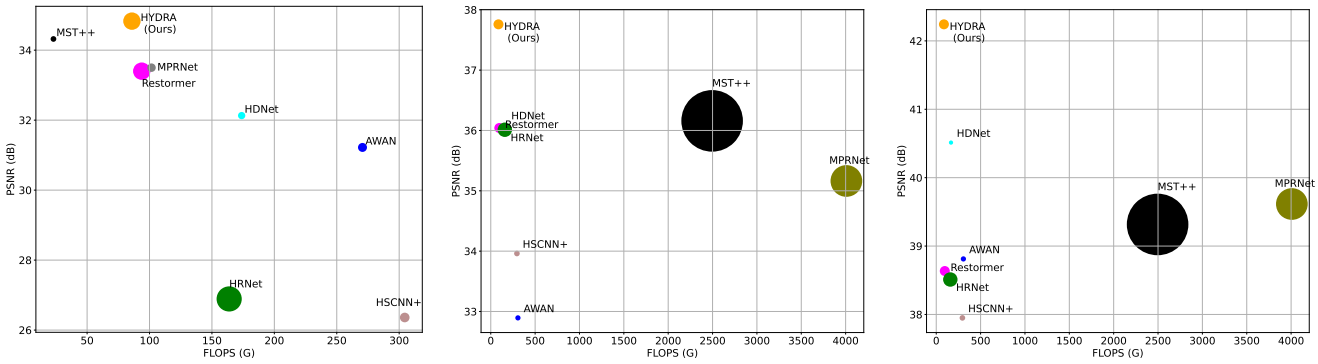


Figure 1. (a) NTIRE-2022 (b) HySpecNet-11k (c) LIB-HSI. HYDRA excels in spectral reconstruction across various channel depths, balancing computational efficiency (FLOPS) with high performance (PSNR). Circle size indicates memory cost.

Abstract

Hyperspectral images (HSI) promise to support a range of new applications in computer vision. Recent research has explored the feasibility of generalizable Spectral Reconstruction (SR), the problem of recovering a HSI from a natural three-channel color image in unseen scenarios. However, previous Multi-Scale Attention (MSA) works have only demonstrated sufficient generalizable results for very sparse spectra, while modern HSI sensors contain hundreds of channels. This paper introduces a novel approach to spectral reconstruction via our HYbrid knowledge Distillation and spectral Reconstruction Architecture (HYDRA). Using a Teacher model that encapsulates latent hyperspectral image data and a Student model that learns mappings from natural images to the Teacher's encoded domain, alongside a novel training method, we achieve high-quality spectral reconstruction. This addresses key limitations of prior SR models, providing SOTA performance across all metrics, including an 18% boost in accuracy, and faster inference times than current SOTA models at various channel depths.

1. Introduction

Hyperspectral Imaging (HSI) extends traditional Red, Green, and Blue (RGB) imaging by capturing a broader spectrum across numerous narrow bands, offering detailed wavelength information invaluable in fields like medical imaging[20] and remote sensing[28]. However, conventional HSI acquisition, reliant on spectrometers, is often impractical for dynamic or real-time applications due to its time-intensive nature and the high costs of HSI cameras. Emerging technologies in Snapshot Compressive Imaging (SCI) and computational reconstruction algorithms have

started to address these challenges [23, 35]. However, HSI still demands costly hardware and long exposures. A promising alternative is spectral reconstruction (SR) from RGB images: it leverages standard cameras to estimate hyperspectral data, slashing costs and acquisition time. SR can't replace a true HSI camera everywhere, yet it is already an industrial tool for multispectral systems that recover far fewer bands. However, SR often serves as a bridge, allowing models trained on one sensor to be deployed on another or facilitating rapid filtering during active exploration tasks. Traditional SR methods are based on sparse coding; as such, they have limited representational ability for this task and have been mostly tested on single scenes split into multiple images, such as Salinas or Indiana Pines [24]. Advances in deep learning, especially deep Convolutional Neural Networks (CNNs), have provided initial attempts at a generalizable mapping from RGB to HSI data. However, these CNN-based methods struggle with capturing long-range dependencies and inter-spectral similarities. When applying current SOTA approaches directly to SR tasks there are two primary challenges: first, an inefficiency in capturing spatial interdependencies in HSIs; second, the computational intensity of global Multi-head Self Attention (MSA), which scales quadratically, and the limitations of local window-based MSA. This is demonstrated in Fig.1 where increasing the channel depth of previous techniques leads variously to dramatic increases in model size, FLOPS and inference times or a decrease in accuracy. These remarks are echoed by the parameter count and computational complexity increasing logarithmically for larger channel datasets that SOTA models have not been tested on.

Our work introduces HYDRA, to solve these flaws while improving the overall performance. This is done by reformulating the SR problem to operate in a learned latent space with drastically reduced dimensionality. To this end, HYDRA employs a knowledge distillation framework to perform SR from RGB images via the latent HSI space of a Teacher model. The HYDRA architecture is based on a cross-modal Teacher-student architecture, which is yet to be explored in the field of SR. This approach diverges from standard knowledge distillation methods by using different input data modalities in a three-step process: First, the teacher model is trained as a compressive autoencoder. Second, a student spectral reconstruction model is trained to convert RGB images to latent codes matching those of the teacher model. Finally, a further training step in both models we call the 'refinement' step.

The HSI input modality is reserved solely for the Teacher model training phase, where it is learnt channel-wise to exploit spectral information in a latent space via unsupervised training. Our Student model instead operates on RGB and uses spatial information alongside channel-wise attention mechanisms. As a result, HYDRA demonstrates a remarkable capability to efficiently handle generalisable SR of large datasets containing a significantly higher number of channels compared to current SOTA models, which have only been tested on channel depths of around 31. This novel approach outperforms existing methods at a fraction of the complexity and model sizes as seen in Fig. 1. The main contributions of our work are as follows:

1. Introduction of HYDRA, a novel approach enhancing generalizable spectral reconstruction efficiency with a unique cross-modal knowledge distillation.
2. An SR architecture combining modern attention mechanisms with squeeze-excitation blocks for improved computational efficiency at high channel depth.
3. The first SR benchmark and evaluation protocol for the spectral reconstruction of high dimensional HSI datasets, to support further developments in the field.

2. Related Work

Generalizable spectral reconstruction (SR) from RGB to hyperspectral images (HSIs) is challenging due to the wide spectral range of HSIs contrasted with the narrower RGB spectrum [2–4]. Earlier methods used specialized RGB cameras and machine learning techniques to address the complex RGB-HSI relationship [1, 2, 11].

2.1. Deep Learning for SR

Supervised SR models like those by Nguyen et al. [26] and Robles-Kelly [29] introduced white-balancing normalization and sparse coding. However, these model-based methods have limited representation capacities and poor generalization. Deep learning, particularly using convolutional neural networks (CNNs) and generative adversarial networks (GANs), has revolutionized hyperspectral reconstruction. Vision transformers have been adapted for hyperspectral SR, improving long-range dependency handling [4, 17]. Models like HSCNN+ [36] and MST++ [7] enhance accuracy and incorporate spectral attention. However, global transformers' computational demands scale quadratically with spatial and channel dimensions (see for example MST++ in Fig. 1 and Tab. 2 and Tab. 3), While local transformers have limited receptive fields, which hinders effective self-attention [4, 7]. Despite advancements, these methods still require extensive training data and accurate labelling. Previous approaches are bottlenecked in computational capacity and generalization, especially at higher channel depths (see Tables 3 and 2). Datasets like NTIRE 2022 [4] and ICVL [2] offer complex data but remain shallow in channel depth, which is impractical in modern hyperspectral applications. Many SR models are untested on deeper datasets recently released [10, 13], which offer broader spectral coverage for diverse scenes. We aim to bridge this gap by enhancing SR performance across diverse

datasets, including those with up to 204 channels like LIB-HSI [13] and HySpecNet-11k [10].

2.2. HSI Compression

Hyperspectral imagery's spatial and spectral redundancies offer ample compression opportunities. While lossless methods provide accurate reconstruction with limited size reduction [25], lossy techniques offer greater reductions with acceptable quality. Inter-band and intra-band methods reduce spectral [8] and spatial [40] redundancies, respectively, improving performance but potentially increasing inference time. JPEG2000 HSI compression standards [6, 9] exploit PCA to minimize spectral correlation. Unlike PCA-based compression, which assumes linear relationships, HYDRA's Teacher model uses a non-linear latent space to better encapsulate hyperspectral variations. Tensor decomposition reduces dimensions while preserving spatial details [39]. Techniques like DCT and 3D-DCT address spatial and spectral aspects but may introduce artefacts, mitigated by wavelet transforms [27]. These methods have improved image quality and compression ratios in applications like FuSENet [32]. Our approach combines these innovations to transform spectral signatures into a lower-dimensional space, aiming to surpass existing lossy methods in efficiency.

2.3. Teacher-Student Architectures

Knowledge Distillation (KD) [15] efficiently trains lightweight deep learning models. In KD, a complex 'teacher' model transfers knowledge to a simpler 'student' model, enhancing efficiency. Traditional KD methods often require extensive datasets and computational resources [5], leading to generalization issues. Recent advancements expand Teacher-Student architectures to knowledge expansion, adaptation, and multi-task learning [12, 34]. However, balancing teacher and student complexities remains challenging [19], prompting research into more efficient designs. In computer vision, these architectures enable compact models to replicate larger models' performance, crucial for resource-limited deployments. Adaptability issues persist in diverse or data-scarce environments [30]. Our HYDRA architecture introduces a transformative training methodology, surpassing traditional single-modality techniques. The student and teacher operate on different modalities with varying supervision and data availability. This is particularly important in precision-critical fields like hyperspectral image processing [22]. Most Teacher-Student frameworks operate within the same data modality, whereas HYDRA uniquely leverages cross-modal learning to bridge RGB and hyperspectral data.

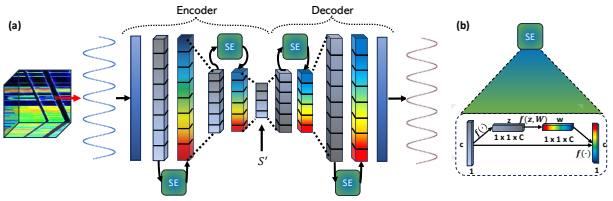


Figure 2. (a) The Teacher model, (b) the operation of an SE block.

3. Method

Fig. 4 illustrates the HYDRA model's architecture. The Teacher model compresses and reconstructs HSI spectra without compromising spectral detail, effectively encapsulating the hyperspectral camera's sensitivity function within its latent space. The Student model maps RGB inputs to this latent space, ensuring efficient optimisation over long-range channel image modalities while bounding the inference prediction space for more accurate spectral reconstruction. By optimizing the Student model within this feature-dense and regularized space, the loss function becomes more meaningful during training, facilitating the learning process and reducing the likelihood of significant errors due to predictions outside the latent space. Noisy spectral regions are unpredictable and challenging for existing models to regress from full HSI, often resulting in smoother representations in these areas (seen in Fig. 6). By leveraging the regularized latent space that smooths out these noisy bands, the Student network can learn more effectively during training. Moreover, during inference on unseen data, the embedded sensitivity function aids in predicting the overall spectral shape, even if specific band intensities are inaccurate. We detail the Teacher model in Section 3.1, the Student model in Section 3.2, and the HYDRA training procedure in Section 3.3.

3.1. Teacher model

The Teacher model of HYDRA is an unsupervised autoencoder defined generally as an encoder and decoder such that $S \approx \text{Dec}(\text{Enc}(S))$. It employs squeeze-excitation (SE-Blocks) [16] for high-quality, pixel-wise compression. The choice of SEBlocks over other attention mechanisms, such as Transformers or Convolutional Block Attention Modules, is motivated by their proven effectiveness in enhancing channel interdependencies with increasing channel depth without significantly increasing computational complexity. The encoder is comprised of 1D convolution paired with 1D max-pooling layers and squeeze-excitation modules, establishing a high compression ratio (CR) for different channel depths. This ensures high-fidelity decoding by the mirrored decoder, thus maintaining the integrity of the compressed spectra. The detailed architecture of the Teacher network is shown in Fig. 2.

It is worth noting that these operations function in 1D across the spectral (i.e. channels) dimension. While

spatial information could also be captured in the Teacher model, we found this contextual information was harder to generalise given the small dataset sizes currently available compared to non-HSI task datasets. Furthermore, in development, we found it makes edges and structures of output images blurred around low-frequency details.

3.1.1. Encoder Structure

The encoder converts a 1-dimensional input signal S taken from a single pixel of an HSI containing b samples into a compressed latent representation S'. We introduce intermediate variables to simplify the equations.

Let $F^l = \text{Conv}(S^l)$ represent the output of the convolutional layer at layer l, and $A^l = E(F^l)$ denote the attention weights computed via the Squeeze-Excitation (SE) block. The encoder updates are then defined as:

$$S^{l+1} = \text{Enc}(S^l) = \text{Pool}\left(F^l \odot A^l\right) \tag{1}$$

Here, $\odot$ denotes element-wise multiplication, and S^l is the output of the l-th layer, with S^N being the final compressed form S'. In this context, Conv denotes a 1D convolution operation, and Pool is a 1D max-pooling operation that reduces the spectral dimensionality.

3.1.2. Decoder Structure

The decoder reconstructs the original signal S from its compressed form S'. Similar to the encoder, we define intermediate variables for the decoder. Let $\tilde{F}^l = \text{Conv}(\tilde{S}^l)$ be the convolutional output at layer l, and $\tilde{A}^l = E(\tilde{F}^l)$ represent the attention weights. The decoder updates are given by:

$$\tilde{S}^{l-1} = \text{Dec}(\tilde{S}^l) = \text{UpS}\left(\tilde{F}^l \odot \tilde{A}^l\right) \tag{2}$$

Since these functions operate in 1D across the spectrum, the upsampling process UpS increases the spectral dimensionality of the feature map using nearest-neighbour interpolation, facilitating reconstruction. The final output of the decoder is the reconstructed signal, matching the input dimensions of the encoder. In this modular approach, the encoder and decoder share a symmetrical structure, where $\tilde{W}$, $\tilde{S}$, and associated parameters act as the decoder's counterparts to the encoder's weights and inputs. Differently from a traditional U-Net [31], this network omits skip connections between the encoder and decoder blocks to allow the decoder to operate independently of the Student network during inference. We discovered that pixel-wise compression more effectively encapsulates hyperspectral data, subsequently refining the Student network's spectral predictions. These improvements are demonstrated in the heatmaps presented in Fig. 5b.

3.2. Student Model

Our proposed composite framework HYDRA integrates a Teacher network for efficient spectral compression and a Student network for high-fidelity spectral reconstruction. The architecture of our Student model, depicted in Fig. 3, is inspired by U-Net style attention networks [14], renowned for their channel-focused attention mechanisms, but tailored for natural image restoration tasks.

Our Student model employs the Multi-Dconv Head Transposed Attention (MDTA) [38] as its transformer backbone, integrating spatial convolution within the channel-wise attention mechanism for enhanced feature extraction and representation. Each block in the U-Net pipeline leverages these Transformer blocks to capture both spatial and spectral dependencies.

In the encoder, the feature extraction function $Y(X)$ represents the process where input feature maps X are transformed by applying a series of MDTA blocks, reducing the spatial dimensions while increasing the number of channels. In the decoder, the function $\tilde{Y}(\tilde{X})$ represents the upsampling process, where the decoder reconstructs the feature maps to their original spatial dimensions by reversing the operations of the encoder. Skip connections between corresponding encoder and decoder layers ensure that spatial details are preserved, providing rich feature representations at multiple scales. This architecture allows for efficient long-range dependency modelling, similar to the Restormer architecture. The model operation is formalised explicitly in the following subsections.

3.2.1. U-Net Pipeline

The Student model's U-Net-like encoder-decoder architecture processes an input image $I \in \mathbb{R}^{H\times W\times 3}$ through four levels. The initial convolutional layer transforms the input into feature embeddings $X^0 \in \mathbb{R}^{H\times W\times C}$, where C represents the number of channels.

In the encoder, each layer generates feature maps $Y(X^{l-1}) = X^l \in \mathbb{R}^{H/2^l\times W/2^l\times 2^l C}$ by employing an increasing number of Transformer blocks, which utilize MDTA. This process reduces the spatial dimensions (H and W) while increasing the channel count. The decoder, using upsampling operations, reconstructs the original image dimensions, with the help of skip connections between the corresponding Student encoder and decoder layers $\tilde{Y}(\tilde{X}^l) + X^l = \tilde{X}^{l-1}$. This helps retain important features from earlier stages during the reconstruction process.

In the student decoder stage, deep features are refined through a convolutional layer, leading to the generation of the residual encoded hyperspectral image. The system outputs the final encoded image, which is a combination of the initial feature embeddings and the residual image:

$$\tilde{S} = X^0 + \text{Conv}(\tilde{X}^0) \tag{3}$$

Here, $\tilde{S}$ represents the final latent restored image, and $\tilde{X}^0$ is the output of the final decoder block The U-Net architecture [31] is particularly well-suited for tasks that require detailed spatial-channel correlation, such as super-resolution

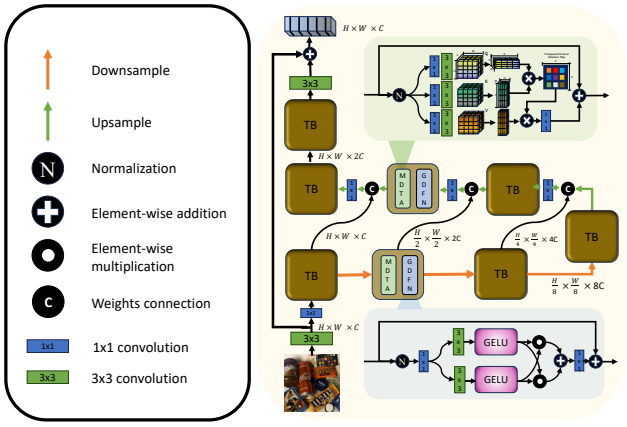


Figure 3. Our Student model: A U-Net shaped Vision Transformer network in the channel domain, utilizing Multi-Dconv Head Transposed Attention (MDTA) and Dual-Gated-Dconv Feed-Forward Network (DGDF) modules. Green 3x3 blocks are depth-wise convolutions and blue blocks 1x1 are pixel-wise convolutions to capture spatial and spectral shapes.

or spectral-reconstruction. Its encoder-decoder structure efficiently captures both high-level and low-level features.

3.2.2. MDTA blocks

The MDTA block is a key part of the $Y(X)$ and $\tilde{Y}(\tilde{X})$ functions, responsible for feature extraction within the U-Net pipeline. It operates with channel depth-wise (H_d) and pixel-wise (H_p) convolutions on layer-normalized input X:

$$H_\text{d}(X) = \mathcal{C}_D(\text{LN}(X)) \tag{4}$$
$$H_\text{p}(X) = \mathcal{C}_P(\text{LN}(X)) \tag{5}$$

where LN denotes layer normalization.

The combined operation of depth-wise and pixel-wise convolutions is then defined as:

$$\mathcal{C}_\text{dp}(X) = H_\text{d}(H_\text{p}(X)) \tag{6}$$

The Student model, unlike the Teacher model, emphasizes spatial context across RGB images. It employs MSA blocks for channel-wise attention calculation, coupled with feed-forward blocks to effectively propagate feature transformations through the network.

The attention mechanism operates by computing a dot-product interaction of the query (Q), key (K), and value (V) outputs to generate the attention map:

$$A(X) = \mathcal{C}^V_\text{dp}(X) \cdot \text{Softmax}\left(\frac{\mathcal{C}^K_\text{dp}(X) \cdot \mathcal{C}^Q_\text{dp}(X)^\top}{\alpha}\right), \tag{7}$$

where α is a learnable scaling parameter. The enhanced feature representation $MDTA$ is generated by combining the attention map A with the input feature X:

$$\text{MDTA}(X) = H_\text{p}(A(X)) + X. \tag{8}$$

The combination of depth-wise and pixel-wise convolutions in MDTA allows for a more nuanced and effective feature transformation. Depth-wise convolutions handle channel-wise interactions efficiently, while pixel-wise convolutions focus on spatial details, making the model adept at capturing both spatial and spectral features.

3.2.3. DGFN

The Dual Gated-Dconv Feed-Forward Network processes the output from preceding layers, focusing on the effective propagation of texture features. The DGFN incorporates a dual gating mechanism defined as:

$$M^g_G(X) = \phi(\mathcal{C}_\text{dp}(\text{MDTA}(X))) \tag{9}$$

where g is either the first (1) or second (2) gate, and ϕ represents an activation function.

These pathways are combined in the encoder branch as follows:

$$Y_G(X) = (M^1_G(X) \odot M^2_G(X)) + (M^1_G(X) \odot M^2_G(X)) \tag{10}$$
$$Y(X) = H_\text{p}(Y_G(X)) + \text{MDTA}(X). \tag{11}$$

This approach merges gated features with the original features, ensuring the preservation of essential information while emphasizing critical textural elements. This representation is then fed into the decoder of the Teacher model for SR. The addition of a skip connection from the encoder enhances the Student decoder's ability to recover fine details lost during compression. The Student decoder operation, along with the skip connection from the encoder, is formalized as:

$$\tilde{Y}(\tilde{X}) = H_p(\tilde{Y}_G(\tilde{X})) + \text{MDTA}(\tilde{X}) + X \tag{12}$$

where $\tilde{X}$ represents the input to the Teacher's decoder, H_p denotes pixel-wise convolutions, $\tilde{Y}_G$ is the Dual Gated-Dconv Feed-Forward Network as in eq.10, and X is the feature map from the equivalent layer of the Student's encoder provided via the skip connection. This output is then passed to the Teacher's decoder, which reconstructs the final hyperspectral image. To clarify, the Student processes RGB to latent space, and the Teacher converts this latent space into hyperspectral data.

3.3. Training Procedure

The training procedure outlined in this section is illustrated via Fig. 4. Our novel HYDRA training methodology uses a three-stage process.

Training stage one: The first stage trains the Teacher network (section 3.1 and Fig. 4.a) to compress HSIs into a compact representation, providing a feature dense latent space for the Student network. The loss function used is

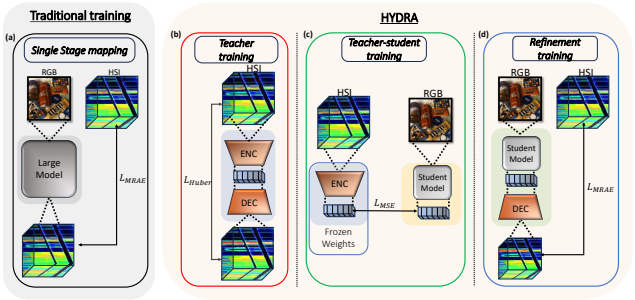


Figure 4. (a) The traditional training method for SR. (b), (c) and (d) show the stages of the HYDRA training procedure.

the Huber loss which provides a smooth loss throughout training, as it handles outliers effectively while maintaining smooth gradients.

Training stage Two: In the second stage, the Student network learns to map RGB images to the frozen Teacher's latent HSI representations (see Fig. 4(c)). The loss function, mean absolute error (MAE), is defined as:

$$L_{sr} = \frac{1}{n}\sum_{i=1}^{n}|\tilde{Y}_i - S'_i| \tag{13}$$

where $\tilde{Y}_n$ represents the Student's predictions and S' is the Teachers latent representation and n is number of data sample. We choose MAE because it penalizes all errors equally, making it robust to outliers. Unlike squared loss, which disproportionately emphasizes larger errors, MAE treats all deviations uniformly. This helps the Student network minimize discrepancies with the Teacher's outputs across latent spectral channels without being overly influenced by outliers, leading to more stable training.

Training stage Three: The final stage involves fine-tuning both the Teacher's decoder and the Student model together (Fig. 4.c), To this end, we unfreeze the Teachers decoder, and train the system end to end using an MSE loss to minimize the difference between the predicted and ground truth HSI. The refinement step aligns the latent space with the end goal of spectral reconstruction, addressing minor discrepancies that may arise due to independent training in earlier stages.

4. Experiments

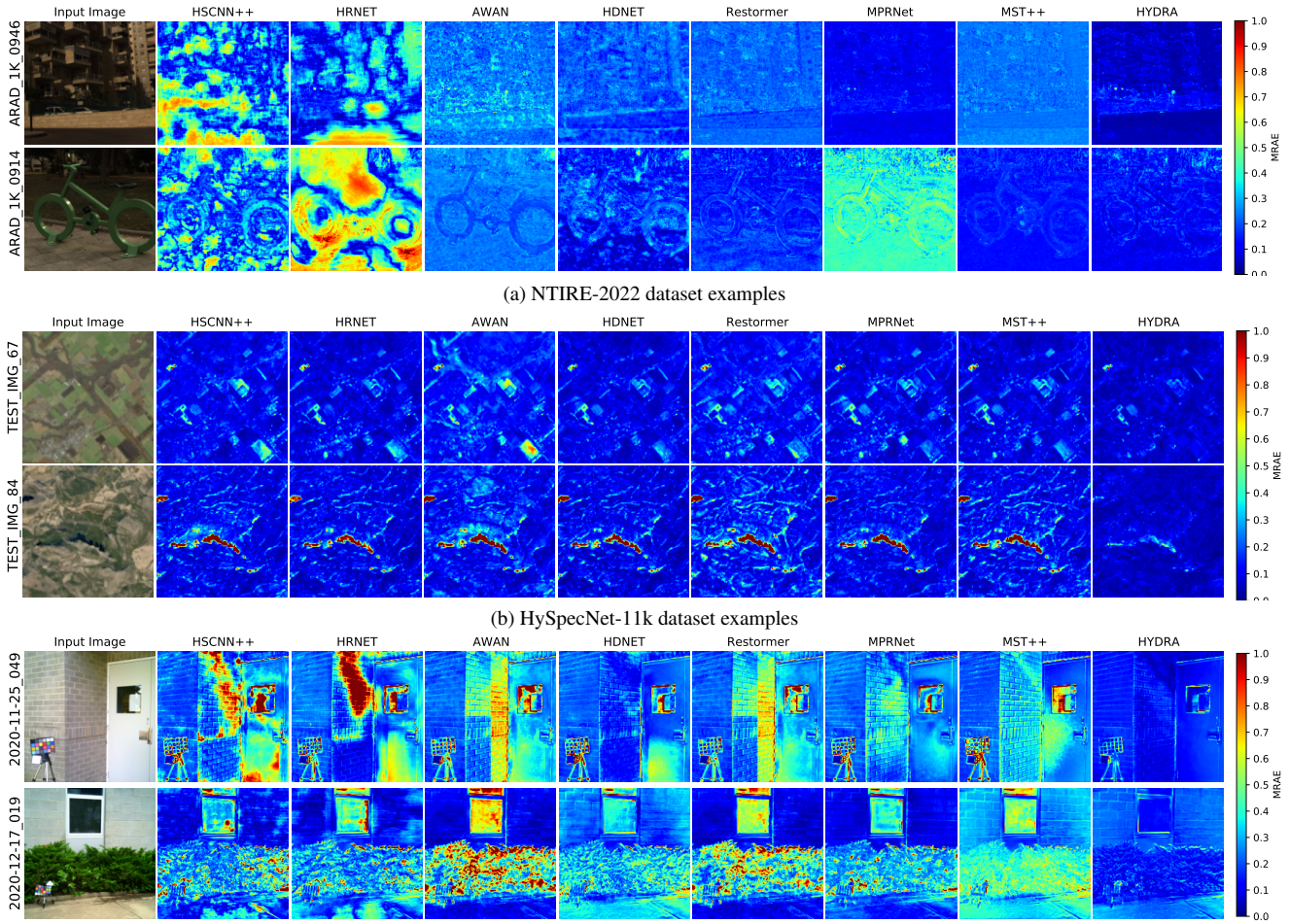
We trained our models on a single NVIDIA A100 80GB GPU, though HYDRA remains compatible with smaller GPUs due to its low memory usage. We used the ARAD1K/NTIRE2022 [4], LIB-HSI [13], and HySpecNet11K [10] datasets with their specified training splits. All methods were implemented in PyTorch, utilising the recommended optimisation and learning rates from their original papers, with minor noise removal adjustments made for the larger LIB-HSI and HySpecNet11K datasets. Each model was trained for 300 epochs with a batch size of 20, and we report the average accuracy per dataset to provide consistency.

As seen from Tables 1, 2, and 3, HYDRA outperforms existing current SOTA SR models in the SR task. This improvement is attributed to our innovative Teacher-Student framework, which efficiently leverages a compact latent space for the Student model to explore. This effect provides a defined space that improves optimisation via a bounded space via a more descriptive latent space modality. Additionally, this modality is inherently smoother and avoids regression over noisy areas of spectra, which can be handled more elegantly via the Teacher network's latent space. In our quantitative analysis (refer to Tables 1, 2, 3 and Fig. 1), our approach exhibits significant improvements across all three accuracy evaluation metrics and inference times against other transformer models (Table 4, particularly in datasets with far deeper channel depths like HyspecNet-11k and LIB-HSI being over six times that of NTIRE-22. Notably, we maintain a stable parameter count and FLOPS compared to other methods. Methods with smaller parameters or FLOPS tend to perform poorly across all metrics. FLOPS are computed using the fvcore library for consistency. In HYDRA, the Teacher's decoder is included with the Student during inference to calculate total FLOPS.

HYDRA's impressive computational efficiency is due to the offloading of initial computation to a compact Teacher model, which shrinks the channel space for the MSA Student model. In most baseline models, larger channel depth datasets cause models to expand significantly, whereas HYDRA remains much smaller due to the reduced channel size provided by the Teacher model. By offering a compact, lower-dimensional target, the latent space reduces the computational load during stage 2 of training, simplifying the process and facilitating more efficient learning of spectral distributions from the full architecture. Our architecture's ability to capture complex spectral relationships through the latent space enables accurate reconstruction even in high-dimensional settings. Importantly, HYDRA maintains a stable parameter count and low FLOPS due to the Teacher's efficient spectral compression and the Student's focused reconstruction. In contrast, methods with fewer parameters or lower FLOPS cannot often model intricate spectral details, resulting in poorer performance.

HYDRA surpasses MSA-based models on complex datasets, as MSA models often struggle with high-dimensional spectral data. By effectively modelling inter-channel dependencies in the Teacher's latent space, HYDRA guides the Student in a smoother latent space. This approach handles high-dimensional data more efficiently and robustly than other models.

These results highlight the efficacy of our Teacher-Student cross-modality approach. By mapping RGB in-



(a) NTIRE-2022 dataset examples

(b) HySpecNet-11k dataset examples

(c) LIB-HSI dataset examples

Figure 5. MRAE heatmap errors for different datasets.

Table 1. Results on NTIRE-22 [4] (31 channels). Best-in-column in **bold**, second best underlined. The * indicates current SOTA transformer backbone methods.

Model	FLOPS (G)	MRAE (↓)	RMSE (↓)	PSNR (↑)
*MST++[7]	**23.05**	0.1645	0.0248	34.32
*MPRNet [37]	101.59	0.1817	0.0270	33.50
*Restormer [38]	93.77	0.1833	0.0274	33.40
HDNet [18]	173.81	0.2048	0.0317	32.13
AWAN [21]	270.61	0.2500	0.0367	31.22
HRNet [41]	163.81	0.3476	0.0550	26.89
HSCNN++ [33]	304.45	0.3814	0.0588	26.36
HYDRA (ours)	85.90	**0.1556**	**0.0221**	**34.83**

puts to the Teacher's latent space, the Student is guided by rich spectral information, improving generalisation and enabling more accurate HSI reconstruction than methods without cross-modality learning.

In table 4, we record the time taken to do inference on a single image from each dataset for the top-performing methods in previous tests. Here, this highlights how HYDRA's lower FLOP counts translate to high-accuracy applications.

Table 2. Results on HySpecNet-11k [10] (202 channel) dataset.

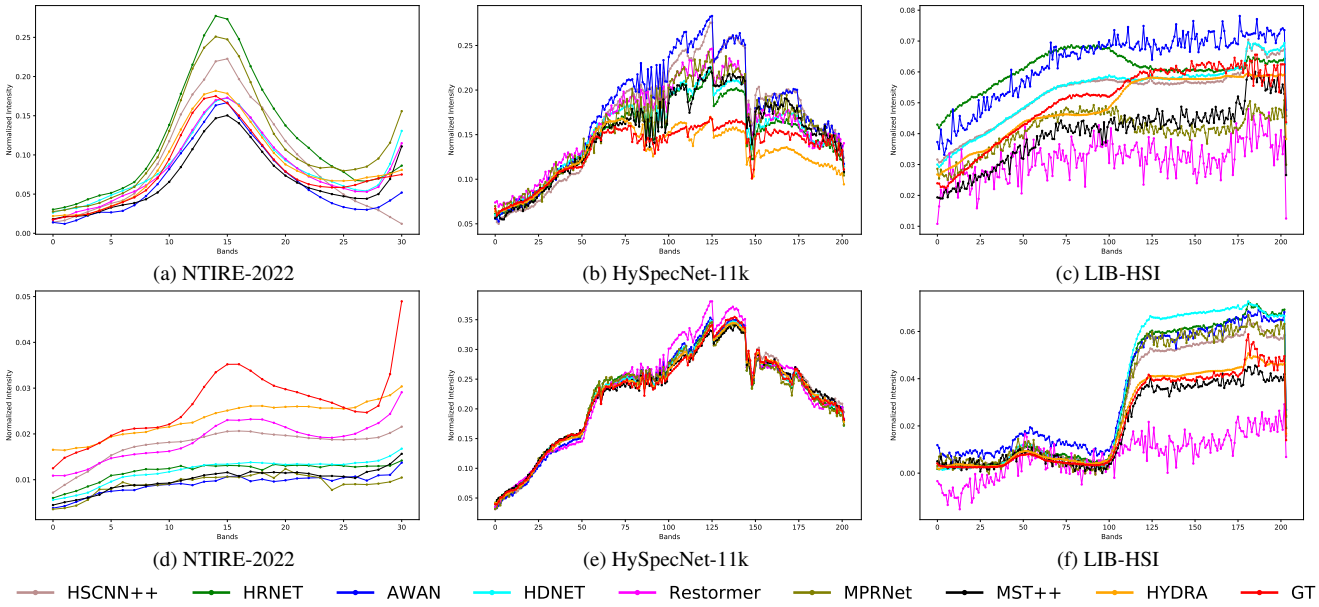
Model	FLOPS (G)	MRAE (↓)	RMSE (↓)	PSNR (↑)
*MST++ [7]	2595.2	0.3133	0.0196	36.159
*MPRNet [37]	4007.32	0.3283	0.0210	35.163
*Restormer [38]	96.34	0.2960	0.0187	36.041
HDNet [18]	166.37	0.2856	0.0186	36.09
AWAN [21]	306.07	0.2218	0.0268	32.894
HRNet [41]	158.57	0.2627	0.0192	36.012
HSCNN++ [33]	294.89	0.2953	0.0242	33.961
HYDRA (ours)	**86.59**	**0.1563**	**0.0160**	**37.759**

Table 3. Results on LIB-HSI [13] (204 channel) dataset.

Model	FLOPS (G)	MRAE (↓)	RMSE (↓)	PSNR (↑)
*MST++ [7]	2595.2	0.4041	0.01254	39.316
*MPRNet [37]	4008.53	0.3922	0.01181	39.133
*Restormer [38]	96.34	0.3905	0.01309	38.668
HDNet [18]	166.37	0.3102	0.01067	40.51
AWAN [21]	306.07	0.4230	0.01248	38.812
HRNet [41]	158.57	0.3881	0.01280	38.777
HSCNN++ [33]	294.89	0.4334	0.01350	38.069
HYDRA (ours)	**86.59**	**0.3004**	**0.00905**	**42.242**

4.1. Qualitative

We visualised error rates with MRAE heat-maps and input test RGB images (Fig.5); blue marks lower errors, red higher. HYDRA's superiority is evident in NTIRE-2022,



(a) NTIRE-2022 (b) HySpecNet-11k (c) LIB-HSI

(d) NTIRE-2022 (e) HySpecNet-11k (f) LIB-HSI

Figure 6. Visualisation of 6 random pixel reconstructions at inference from selected test datasets and approaches

Table 4. Inference time (ms) for a single image for the three best SOTA methods and HYDRA (ours).

Model	NTIRE-22	HySpecNet-11k	LIB-HSI
MST++ [7]	102.52	233.53	3119.04
MPRNet [37]	74.97	212.25	2861.39
Restormer [38]	147.30	32.42	219.98
HYDRA (ours)	154.11	16.09	172.70

with darker blue hues outshining other models. This stems from our Teacher-Student approach, which constrains the Student's predictions within the latent space, reducing outliers compared to reconstructing the full high-dimensional spectra directly. Models like Restormer, lacking such guidance, are more prone to errors. In the larger HyspecNet-11k dataset (Fig. 5b), HYDRA excels in land shade reconstruction, benefiting from the increased dataset size. The LIB-HSI dataset exhibits artifacts and noise, analyzed further in our pixel reconstruction study (Fig. 6). We compare random pixel spectra from reconstructed HSIs to ground truth (GT in red) across all datasets. Unlike a single global metric, these plots provide a visual analysis of spectral shapes across the visible-to-infrared range. This approach lets us assess how well each model captures the structure of the spectrum without being biased by aggregate accuracy values—i.e., a model might score better on a global error metric yet fail to replicate the correct overall shape. By visually inspecting these spectra, we gain a more nuanced understanding of each model's strengths and weaknesses in reconstructing hyperspectral data. As seen in Fig. 6, the test NTIRE-2022 images show the best reconstruction quality, due to their shallower channel depth. Although the general spectral shapes are well-predicted, some discrepancies in absolute intensity levels occur. Figures 6a, 6b, and 6c illustrate that models with attention more closely match the ground truth. Further analysis in figures 6d, 6e, and 6f highlights HYDRA's superior accuracy over other models, with more consistent reconstructions across all datasets.

Table 5. Ablation study for HYDRA with different latent-channel sizes and training stages. The Stage 1 column shows the Teacher model's HSI auto-encoding performance (upper bound for Stages 2–3).

Dataset	Latent Size	Stage 1 (Teacher)			Stage 2 (Teach.–Stud.)			Stage 3 (Refinement)		
		MRAE	RMSE	PSNR	MRAE	RMSE	PSNR	MRAE	RMSE	PSNR
NTIRE-2022	4	0.0419	0.0057	46.51	0.2412	0.0328	31.89	0.2207	0.0386	32.53
	6	0.0198	0.0027	53.17	0.1772	0.0250	33.97	**0.1556**	**0.0221**	**34.83**
	8	0.0125	0.0017	56.70	0.2164	0.0252	33.77	0.1723	0.0245	34.02
HySpecNet-11k	13	0.1011	0.0047	47.35	0.1701	0.0178	37.11	0.1658	0.0168	37.25
	17	0.0756	0.0032	50.90	0.1621	0.0166	37.38	**0.1563**	**0.0160**	**37.76**
	34	0.0683	0.0025	52.95	0.1685	0.0171	37.32	0.1597	0.0161	37.64
LIB-HSI	13	0.0228	0.0012	59.29	0.4373	0.0139	38.55	0.4211	0.0120	39.64
	17	0.0187	0.0010	60.49	0.3171	0.0110	41.91	**0.3004**	**0.0091**	**42.24**
	34	0.0143	0.0008	62.52	0.3278	0.0141	40.18	0.3201	0.0125	41.84

4.2. Ablation Study

Below we analyse how different latent sizes and training stages affect HYDRA's performance (see table 5). The three training stages are: Teacher-only, Teacher-Student, and Refinement. The Teacher-only column encodes and decodes test HSIs, illustrating the model's upper bound in stages 2 and 3, but does not address the SR task. HYDRA is sensitive to latent size changes: size 6 is optimal for NTIRE-2022, while 17 is best for HySpecNet-11k and LIB-HSI. This indicates the need to balance representation capacity with minimal latent size for accurate RGB-to-HSI mapping. Further comparison of three training variations—Stage 1+3 only, Student-only, and the full Three-Stage training is detailed in Table 6. The comprehensive Three-Stage training approach surpasses the alternatives, underscoring its critical role in enhancing the HYDRA model's performance across all tested datasets.

Table 6. 17-channel latent-space comparison of refinement only, student only, and three-stage training.

Dataset	Stage 1 + 3			Student Only			Three-Stage		
	MRAE	RMSE	PSNR	MRAE	RMSE	PSNR	MRAE	RMSE	PSNR
NTIRE-2022	0.1689	0.0273	32.92	0.1833	0.0274	33.40	**0.1556**	**0.0221**	**34.83**
HySpecNet-11k	0.2689	0.0215	35.11	0.2960	0.0187	36.04	**0.1563**	**0.0160**	**37.76**
LIB-HSI	0.8820	0.0229	33.74	0.3905	0.0131	38.67	**0.3004**	**0.0091**	**42.24**

5. Conclusion

We introduce HYDRA, a novel spectral reconstruction approach that leverages knowledge distillation, Teacher-Student architectures, and cross-modal learning to outperform existing SOTA models. HYDRA not only achieves superior accuracy on primary benchmarks but also demonstrates exceptional computational efficiency across varying hyperspectral channel sizes. It excels in diverse channel-depth scenarios, positioning it as a novel technique for future HSI applications.

References

[1] F Agahian, SA Arnirshahi, and SH Amirshahi. Reconstruction of reflectance spectra using weighted principal component analysis. Color Research & Application, 33(5):360–371, 2008. 2

[2] B Arad and O Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In European Conference on Computer Vision, pages 19–34. Springer, 2016. 2

[3] B Arad, R Timofte, O Ben-Shahar, Y-T Lin, and G D Finlayson. Ntire 2020 challenge on spectral reconstruction from an rgb image. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 446–447, 2020.

[4] B Arad et al. Ntire 2022 spectral demosaicing challenge and data set. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 881–895, 2022. 2, 6, 7

[5] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? Advances in neural information processing systems, 27, 2014. 3

[6] Ian Blanes and Joan Serra-Sagristà. Cost and scalability improvements to the karhunen–loève transform for remote-sensing image coding. IEEE Transactions on Geoscience and Remote Sensing, 48(7):2854–2863, 2010. 3

[7] Y Cai et al. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 745–755, 2022. 2, 7, 8

[8] Chein-I Chang. Hyperspectral Data Processing: Algorithm Design and Analysis. Wiley, 2013. 3

[9] Qian Du and James E Fowler. Hyperspectral image compression using jpeg2000 and principal component analysis. IEEE Geoscience and Remote Sensing Letters, 4(2):201–205, 2007. 3

[10] M H P Fuchs and B Demir. Hyspecnet-11k: a large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods. In IEEE International Geoscience and Remote Sensing Symposium, pages 1779–1782, 2023. 2, 3, 6, 7

[11] M Goel, E Whitmire, A Mariakakis, T S Saponas, N Joshi, D Morris, et al. Hypercam: hyperspectral imaging for ubiquitous computing applications. In ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 145–156, 2015. 2

[12] J. Gou, B. Yu, S. Maybank, and D. Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819, 2021. 3

[13] Nariman Habili, Ali Armin, Jeremy Oorloff, Weihao Li, Christfried Webers, Ernest Kwan, and Lars Petersson. Lib-hsi: Rgb and hyperspectral images of building facades. CSIRO Data Collection, 2022. 2, 3, 6, 7

[14] Sheng He, Rina Bao, Patricia Ellen Grant, and Yangming Ou. U-netmer: U-net meets transformer for medical image segmentation. ArXiv, abs/2304.01401, 2023. 4

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. 3

[16] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. ArXiv, 2017. 3

[17] J-F Hu, T-Z Huang, L-J Deng, H-X Dou, D Hong, and G Vivone. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2022. 2

[18] 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. 7

[19] B. Kang, L. Xie, X. Yang, and C. Zhang. Exploring the impact of neural architecture search on knowledge distillation. arXiv preprint arXiv:2104.01777, 2021. 3

[20] Shahid Karim, Akeel Qadir, Umar Farooq, Muhammad Shakir, and Asif A Laghari. Hyperspectral imaging: a review and trends towards medical imaging. Current Medical Imaging, 19(5):417–427, 2023. 1

[21] Jiaojiao Li, Chaoxiong Wu, Rui Song, Yunsong Li, and Fei Liu. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1894–1903, 2020. 7

[22] Zhizhong Li and Derek Hoiem. Learning without forgetting. European Conference on Computer Vision, pages 614–629, 2016. 3

[23] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. ACM Trans. Graph., 33(6), 2014. 2

[24] Larry L. Biehl Marion F. Baumgardner and David A. Landgrebe. 220 band aviris hyperspectral image data set: Indian pine test site 3. 2015. 2

[25] J Mielikainen and B Huang. Lossless compression of hyperspectral images using clustered linear prediction with adaptive prediction length. IEEE Geoscience and Remote Sensing Letters, 9:1118–1121, 2012. 3

[26] R M Nguyen, D K Prasad, and M S Brown. Training-based spectral reconstruction from a single rgb image. In European Conference on Computer Vision, pages 186–201. Springer, 2014. 2

[27] Tong Qiao, Jinchang Ren, Meijun Sun, Jiangbin Zheng, and Stephen Marshall. Effective compression of hyperspectral imagery using an improved 3d dct approach for land-cover analysis in remote-sensing applications. International Journal of Remote Sensing, 35(20):7316–7337, 2014. 3

[28] Lankapalli Ravikanth, Digvir S Jayas, Noel DG White, Paul G Fields, and Da-Wen Sun. Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products. Food and bioprocess technology, 10:1–33, 2017. 1

[29] A Robles-Kelly. Single image spectral reconstruction for multimedia applications. In Proceedings of the 23rd ACM International Conference on Multimedia, pages 251–260, 2015. 2

[30] Adrian Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2015. 3

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. ArXiv, 2015. 4

[32] SK Roy, SR Dubey, S Chatterjee, and B Baran Chaudhuri. Fusenet: fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification. IET Image Processing, 14:1653–1661, 2020. 3

[33] Zhan Shi, Chang Chen, Zhiwei Xiong, Dong Liu, and Feng Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In CVPRW, 2018. 7

[34] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. 3

[35] Yuehao Wu, Iftekhar O. Mirza, Gonzalo R. Arce, and Dennis W. Prather. Development of a digital-micromirror-device-based multishot snapshot spectral imaging system. Opt. Lett., 36(14):2692–2694, 2011. 2

[36] Z Xiong, Z Shi, H Li, L Wang, D Liu, and F Wu. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In IEEE International Conference on Computer Vision Workshops, pages 518–525, 2017. 2

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14821–14831, 2021. 7, 8

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In CVPR, 2022. 4, 7, 8

[39] L Zhang, L Zhang, D Tao, X Huang, and B Du. Compression of hyperspectral remote sensing images by tensor approach. Neurocomputing, 147:358–363, 2015. 3

[40] Dongyu Zhao, Shiping Zhu, and Fengchao Wang. Lossy hyperspectral image compression based on intra-band prediction and inter-band fractal encoding. Computers & Electrical Engineering, 54:494–505, 2016. 3

[41] Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin. Hierarchical regression network for spectral reconstruction from rgb images. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1695–1704, 2020. 7