SAH-SCI: Self-Supervised Adapter for Efficient Hyperspectral Snapshot Compressive Imaging

Haijin Zeng[1,⋆], Yuxi Liu[2,⋆], Yongyong Chen[2(✉)], Youfa Liu[3], Chong Peng[4(✉)], and Jingyong Su[2]

[1] IMEC-Ghent University, Ghent 9000, Belgium
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
[3] School of Computer Science, Wuhan University, Wuhan, China
[4] School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
YongyongChen.cn@gmail.com; pchong1991@163.com
https://github.com/lyuxi/SAH-SCI

Abstract. Hyperspectral image (HSI) reconstruction is vital for recovering spatial-spectral information from compressed measurements in coded aperture snapshot spectral imaging (CASSI) systems. Despite the effectiveness of end-to-end and deep unfolding methods, their reliance on substantial training data poses challenges, notably the scarcity of labeled HSIs. Existing approaches often train on limited datasets, such as KAIST and CAVE, leading to biased models with poor generalization capabilities. Addressing these challenges, we propose a universal Self-Supervised Adapter for Hyperspectral Snapshot Compressive Imaging (SAH-SCI). Unlike full fine-tuning or linear probing, SAH-SCI enhances model generalization by training a lightweight adapter while preserving the original model's parameters. We propose a novel approach that combines spectral and spatial adaptation to enhance an image model's capacity for spatial-spectral reasoning. Additionally, we introduce a customized adapter self-supervised loss function that captures the consistency, group invariance and image uncertainty of CASSI imaging. This approach effectively reduces the solution space for ill-posed HSI reconstruction. Experimental results demonstrate SAH's superiority over previous methods with fewer parameters, offering simplicity and adaptability to any end-to-end or unfolding methods. Our approach paves the way for leveraging more robust image foundation models in future hyperspectral imaging tasks.

Keywords: Compressive imaging · Self-supervised learning · Adapter

1 Introduction

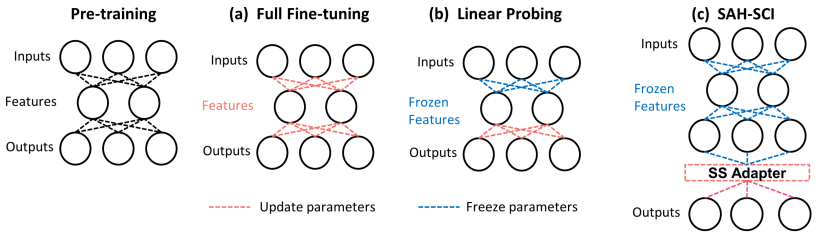
Hyperspectral images (HSIs), which acquire both spatial and spectral information, can more accurately characterize the captured scene than traditional RGB

⋆ These authors contributed equally to this work.

images. Based on spectral features with rich information, HSIs are widely used in many computer vision tasks, e.g., object detection [18, 32, 33], medical image processing [24, 30], remote sensing [4, 27, 48], etc. Inspired by conventional compressive sensing (CS) [49, 53], snapshot compressive imaging (SCI) has been proposed to capture HSIs. Compressing 3D HSI cubes into 2D measurements, SCI can acquire HSIs quickly with low cost and low bandwidth. In recent years, SCI systems have been developed by various hardware [22, 26, 45, 47], which are used for 3D spectral images [41–43]. Among many SCI systems, coded aperture snapshot spectral imaging (CASSI) [13,29,39] stands out for its outstanding performance and has become the mainstream solution for SCI. CASSI modulates spectral frames by a coded aperture, and samples them by shifting across the spectral dimension via a disperser [52].

Based on CASSI, a large number of reconstruction algorithms have been proposed to solve the inverse problem, i.e., reconstructing 3D HSIs from 2D measurements. Considering the inverse problem is ill-posed, conventional model-based approaches typically use handcrafted priors such as sparsity [19, 39], total variation [37, 40], low-rankness [23, 50], etc. These methods require manual parameter adjustments, resulting in poor quality and slow recovery rates. Thanks to the advancements in deep learning, the speed and accuracy have been significantly improved through the design of various deep network architectures [5,6,16,28,29,31]. Due to the complexity of compression in CASSI, it is very difficult to design a high-quality and fast self-supervised algorithm. Almost all existing methods in this task train a deep neural network in a supervised manner, i.e., using datasets with many paired ground truth (GT) images and their measurements. However, these deep learning methods face a challenge that needs to be addressed: the lack of available large-scale training data. Supervised networks often require large amounts of data for high-quality and stable performance. Unfortunately, in the hyperspectral domain, there is very limited image data available. Acquiring sufficient spectral data is difficult and expensive. Most reconstructed networks are trained based on the KAIST [10] and CAVE [34] datasets. Although these methods can achieve good reconstruction results on given synthetic data, the bias in the limited training data might lead to poor generalization performance of the pre-trained model, e.g., the reconstruction results of HDNet [16] on ICVL [1] dataset have large errors in spectral accuracy, which limits the application of the model to real-world datasets.

So how to improve the generalization performance of existing reconstruction models without sufficient HSI training datasets? A possible strategy is self-supervised fine-tuning. Conventional self-supervised fine-tuning methods include full fine-tuning and linear probing. However, these proposed self-supervised fine-tuning strategies fail to achieve generality and efficiency for SCI models. As shown in Fig. 1, full fine-tuning leads to model instability, resulting in worse reconstruction results with much lower PSNR and SSIM [44]. Linear probing only updates the final linear layer parameters, which fails to be applied with deep unfolding models resulting in poor generality. To address the above challenges, we firstly propose a universal self-supervised fine-tuning framework for



Method	Tunable parameters	ICVL	Harvard	NTIRE2022
Pre-training	2.37M	28.91dB/0.851	30.68dB/0.827	34.93dB/0.925
Full Fine-tuning	2.37M	25.39dB/0.734	27.16dB/0.703	27.48dB/0.736
SAH-SCI	**0.29M**	**34.78dB (↑5.87dB) /0.937 (↑0.086)**	**34.04dB (↑3.36dB) /0.882 (↑0.055)**	**37.10dB (↑2.17dB) /0.943 (↑0.018)**

Fig. 1: Comparison of self-supervised fine-tuning methodologies: In this context, the demonstrated pre-trained model is HDNet [16]. (a) Fine-tune all the model parameters. (b) Linear probing freezes the feature extractor and updates the final linear layer parameters. (c) Our SAH-SCI trains a lightweight spatial-spectral (SS) adapter to fine-tune the outputs of the pre-trained model while preserving its parameters. PSNR/SSIM and tunable parameters on three datasets demonstrate the superiority of our method.

hyperspectral SCI (SAH-SCI), which can be applied to various types of hyperspectral reconstruction models. We keep the parameters of the pre-trained model unchanged and train an adapter for the model to improve its generalization performance, which enhances its visual details of unknown data while preserving the characteristics of the original model. Since the adapter is trained in a self-supervised manner and preserves the original model's parameters, it requires cheaper memory and computational costs. Additionally, we propose a lightweight Self-Supervised Adapter (SAH) with a customized self-supervised loss function. The spatial and spectral properties of hyperspectral images are effectively exploited through SAH. Simultaneously, the self-supervised loss function mitigates the solution ambiguity and nullspace information deficiency in the inverse problem of SCI by exploring the consistency, group invariance and uncertainty in CASSI imaging.

Overall, our contributions of this work can be summarized as follows:

- We propose a universal self-supervised fine-tuning framework for hyperspectral SCI (SAH-SCI). It is the first attempt to introduce self-supervised fine-tuning to hyperspectral SCI.
- We design a lightweight Self-Supervised Adapter SAH with a customized self-supervised loss function to capture spatial and spectral properties and effectively reduce the feasible solution space in hyperspectral SCI.
- Our SAH-SCI framework can effectively enhance the generalization of the model while requiring less time and fewer costs. Extensive results on both synthetic and real datasets verify the superiority of the proposed approach.

2 Related Work

2.1 Hyperspectral Image Reconstruction

For the HSI reconstruction, traditional methods [12,23,40,46,50] use handcrafted image priors to solve the inverse problem iteratively. However, these iterative optimization-based algorithms are limited by hand-crafted priors and the low reconstruction speed. Recently, inspired by deep learning to solve inverse problems [3], convolution neural networks (CNNs) and transformer-based networks have been used to solve the inverse problem of spectral SCI to achieve rapid reconstruction [5,6,25,28,29,31,52]. The above methods can be categorized into three main groups, i.e., end-to-end methods, deep unfolding methods and plug-and-play (PnP) methods. PnP methods plug CNN denoisers into model-based approaches to address the HSI reconstruction problem, which typically require hundreds of iterations resulting in no real-time reconstruction. Given enough training data and time, two other types of methods can achieve desirable results and output instantaneously, thus building a real-world SCI system (including sampling and reconstruction). However, their reliance on paired samples of real HSIs limits their applicability to specific systems with defined bands and characteristics, posing a challenge for generalizability across different scenarios. Consequently, there is a growing demand for efficient self-supervised algorithms capable of overcoming the limitations associated with large-scale training datasets.

Despite the challenges of implementing self-supervised algorithms for SCI systems, there have been a few studies on self-supervised deep learning for CS to remove the prerequisite on GT images. [11] trained generative adversarial networks with unpaired samples. [8, 9] proposed a self-supervised scheme based on image equivariance. [36] proposed a dual-domain self-supervised loss for handling possible ambiguity and overfitting.

2.2 Adapter Tuning

Models that are pre-trained using comprehensive general domain datasets have shown impressive capabilities in generalization, greatly enhancing a variety of applications, ranging from natural language processing (NLP) [35] tasks to multi-model tasks [21]. To tailor the model for specific downstream tasks, full fine-tuning is often applied to retrain all model parameters. However, as the size of the model and dataset increases, fine-tuning the entire model imposes unacceptable overheads, while inevitably being limited by the original model. Another method is linear probing [20], this approach only fine-tunes the last layer of the network which limits its application to deep unfolding methods.

Adapter tuning is proposed to solve the above problem by introducing additional trainable modules into the frozen backbone [14, 15]. [15] adds the linear modules in sequence to the existing layer. [14] recommends integrating these modules in parallel with the original layer to improve performance. However, due to the complexity of snapshot compressive imaging, traditional adapter tuning method faces initialization sensitivity issues that affects effectiveness.

3 Mathematical Model of CASSI

The schematic of the CASSI system is shown in Fig. 2. In the CASSI system, the 3D spectral cube is first modulated via a coded aperture and then dispersed via a disperser. Mathematically, let $\mathbf{X} \in \mathbb{R}^{H\times W\times N_\lambda}$ denote the 3D spectral cube, where H, W, and N_λ are the cube's height, width, and number of wavelengths. $\mathbf{M^0} \in \mathbb{R}^{H\times W}$ denotes the coded aperture. We use $\mathbf{X'} \in \mathbb{R}^{H\times W\times N_\lambda}$ to represent the modulated signals. Then the modulation process can be expressed as:

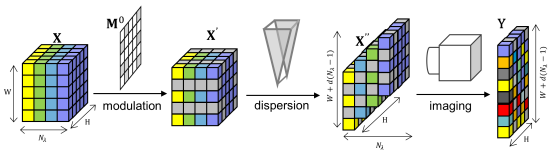


Fig. 2: The schematic of CASSI system.

$$\mathbf{X'}(:,:,\lambda) = \mathbf{X}(:,:,\lambda) \odot \mathbf{M^0}, \tag{1}$$

where $\odot$ is element-wise multiplication, $\lambda \in [1, ..., N_\lambda]$ represents the spectral wavelengths. After that, modulated HSI frames with different wavelengths pass the disperser. $\mathbf{X'}$ is tilted and is seen as sheared along the y-axis. We then use $\mathbf{X''} \in \mathbb{R}^{H\times(W+d\times(N_\lambda-1))\times N_\lambda}$ to represent the tilted data cube, where d refers to the shift step. Assuming λ_c to be the reference wavelength, image $\mathbf{X'}(:,:,\lambda_c)$ represents frame that is not sheared along the y-axis, we can formulate the dispersion as:

$$\mathbf{X''}(u, v, n_\lambda) = \mathbf{X'}(x, y + d(\lambda_n - \lambda_c), n_\lambda), \tag{2}$$

where (u, v) denotes the spatial coordinates, λ_n denotes the wavelength of the n_λ-th channel, and $d(\lambda_n - \lambda_c)$ denotes the shift distance. Based on the previous model, the measurement $\mathbf{Y} \in \mathbb{R}^{H\times(W+d\times(N_\lambda-1))}$ can be expressed as:

$$\mathbf{Y} = \sum_{n_\lambda=1}^{N_\lambda} \mathbf{X''}(:,:,n_\lambda) + \mathbf{N}, \tag{3}$$

where $\mathbf{N} \in \mathbb{R}^{H\times(W+d\times(N_\lambda-1))}$ indicates random noise generated by sampling. Then given $\mathbf{x} \in \mathbb{R}^{nN_\lambda}$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{n} \in \mathbb{R}^n$ with $n = H(W + d \times (N_\lambda - 1))$, which denote the vectorized form of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{N}$ respectively, the measurement in SCI can be modeled by:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}, \tag{4}$$

where $\mathbf{\Phi} \in \mathbb{R}^{n\times nN_\lambda}$ denotes the sampling matrix. Thus, the reconstruction task for SCI is to solve $\mathbf{x}$, given the measurement $\mathbf{y}$ (sampled by camera) and the sampling matrix $\mathbf{\Phi}$ (pre-designed masks).

4 Method

4.1 Overall Architecture

The overall architecture of SAH-SCI is shown in Fig. 3. Firstly, Mask $\mathbf{\Phi}$ and measurements $\mathbf{y} \in \mathbf{R}^{H\times(W+d(N_\lambda-1))}$ are input into the pre-trained model to

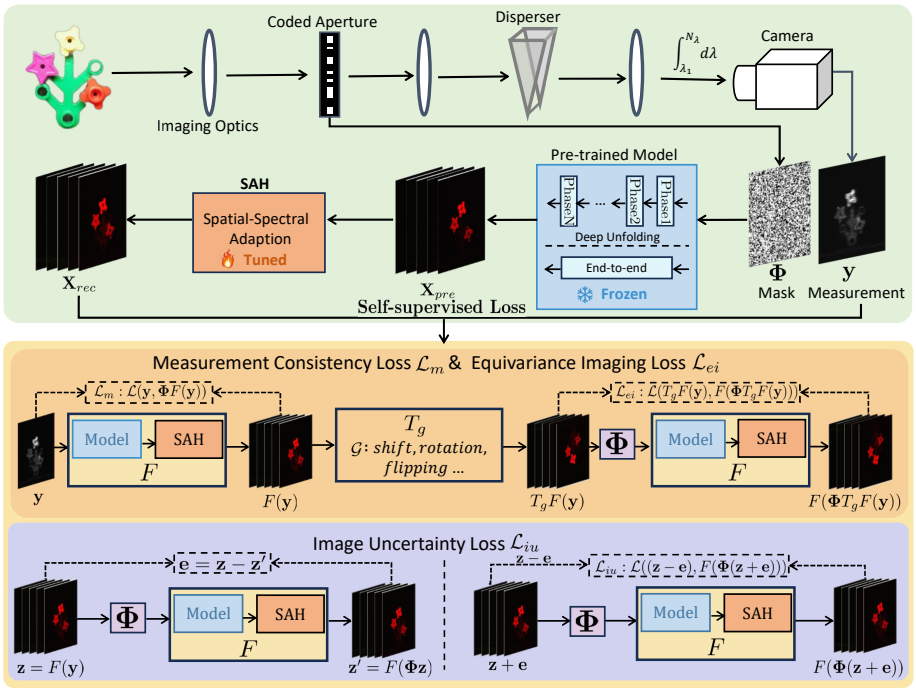


Fig. 3: Top: Framework of SAH-SCI. Mask $\mathbf{\Phi}$ and measurement $\mathbf{y}$ are input into the pre-trained model to obtain the initial results $\mathbf{X}_{pre}$, where the pre-trained model parameters are frozen. Then $\mathbf{X}_{pre}$ is fed into the self-supervised adapter SAH to obtain the HSI reconstruction result $\mathbf{X}_{rec}$. **Bottom:** Self-supervised loss component details.

obtain the initial results $\mathbf{X}_{pre} \in \mathbf{R}^{H \times W \times N_\lambda}$. It should be noted that pre-trained model is frozen. Keeping the parameters of the pre-trained model unchanged saves the overhead of updating model, which is faster to train and occupies less memory. It also ensures the spectral knowledge and features learned by the pre-trained model are not forgotten. Secondly, $\mathbf{X}_{pre}$ is fed into the self-supervised adapter SAH to obtain the fine-tuned HSI reconstruction results $\mathbf{X}_{rec}$. The SAH-SCI process can be written as:

$$\begin{aligned} \mathbf{X}_{pre} &= \text{Pretrained}(\mathbf{\Phi}, \mathbf{y}), \\ \mathbf{X}_{rec} &= \text{Adapter}(\mathbf{X}_{pre}). \end{aligned} \tag{5}$$

The SAH adopts a U-shaped structure where a lightweight spatial-spectral convolution, and a self-supervised loss are customized. More details about SAH implementation and self-supervised loss are shown in Sec. 4.2 and Sec. 4.3.

4.2 Architecture of Proposed SAH

As depicted in Fig. 4 (a), SAH adopts the spatial-spectral convolution as the basic module where different scales of modules are stacked in a U-shaped struc-

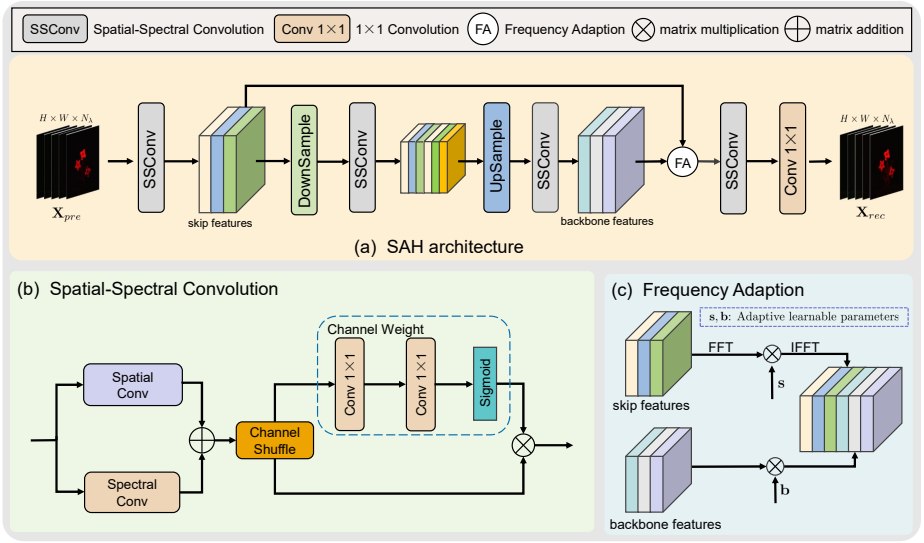


Fig. 4: The architecture of SAH. (a) SAH stacks spatial-spectral convolutions with U-shaped structure. (b) The structure of the spatial-spectral convolution. (c) Frequency Adaption. Skip features and backbone features are concatenated by adaptive learnable weights based on frequency domain information.

ture. By fusing the extracted spatial and spectral information, spatial-spectral convolution is able to effectively exploit hyperspectral features while requiring less parameters. Besides, we design a frequency-adaptive feature fusion module, frequency adaption, to balance the denoising ability of the backbone network and the ability of skip connections to introduce high-frequency details.

Spatial-Spectral Convolution. Conventional 2D convolution fails to effectively extract the spectral features of HSIs, while 3D convolution results in unacceptably large parameter counts. To address above limitations, we propose a novel spatial-spectral convolution module to exploit the unique spatial and spectral properties of HSIs while ensuring the lightweight structure of the adapter, as shown in Fig. 4 (b). For the given input features, spatial features and spectral features are extracted by a spatial convolution and a spectral convolution, respectively. We choose 3×3 group convolution as spatial convolution and 1×1 convolution as spectral convolution. To minimize the number of parameters, we set the number of groups to be the greatest common divisor of the input and output channels in group convolution. Then the extracted features are summed and fused by channel shuffle module [51]. The process is formulated as follows:

$$\begin{aligned}
\mathbf{F}_{spatial} &= \text{GroupConv}_{3\times3}(\mathbf{F}_{input}), \\
\mathbf{F}_{spectral} &= \text{Conv}_{1\times1}(\mathbf{F}_{input}), \\
\mathbf{F}_{fusion} &= \text{ChannelShuffle}(\mathbf{F}_{spatial} + \mathbf{F}_{spectral}).
\end{aligned} \tag{6}$$

Finally, the fusion features are channel weighted to get the output, where the channel weights are computed by a sequence of 1×1 convolution, 1×1 convolution and sigmoid, which is expressed as:

$$\mathbf{F}_{out} = \mathbf{F}_{fusion} \otimes \text{Sigmoid}(\text{Conv}_{1\times1}(\text{Conv}_{1\times1}(\mathbf{F}_{fusion}))), \tag{7}$$

where $\otimes$ means matrix multiplication operation.

Frequency Adaption. In Sec. 5, it is observed that pre-trained models generally succeed in reconstructing the main structural information of HSI, but they often fall short in preserving finer details and may introduce artifacts. Consequently, adapters must prioritize the recovery of details and the reduction of artifacts, both of which are typically associated with high-frequency components. However, the incorporation of skip connections could potentially compromise the inherent reconstructing capability of the backbone network, resulting in the generation of aberrant image details [38]. To address this concern, we propose a frequency adaptation module, depicted in Figure 4 (c). This module introduces adaptive learnable weights, denoted as b and s, where b is utilized to augment the backbone features while s is employed to diminish the contribution of skip features. Furthermore, to counteract the potential oversmoothing of textures resulting from enhanced denoising, skip features undergo spectral modulation in the Fourier domain to suppress low-frequency information. With adaptive weighting, skip features and backbone features are then concatenated to get the integration features, which is formulated as:

$$\begin{aligned}
&\mathbf{F}'_{backbone} = \mathbf{F}_{backbone} \otimes b, \\
&\mathbf{F}'_{skip} = \text{IFFT}(\text{FFT}(\mathbf{F}_{skip}) \otimes s), \\
&\mathbf{F}_{integration} = \text{Concatenate}(\mathbf{F}'_{backbone}, \mathbf{F}'_{skip}),
\end{aligned} \tag{8}$$

where FFT$(\cdot)$ and IFFT$(\cdot)$ are Fourier transform and inverse Fourier transform.

4.3 Self-Supervised Loss Function for Adapter

Solving the severe linear inverse problem of SCI requires an understanding of the underlying hyperspectral signal model, which has to be learned from data. However, learning the model from observations obtained by a single incomplete measurement mask is impossible, as the nullspace of the operator contains no information about the reconstructed signal model. Thus, methods based on the incomplete mask are limited by observational data and fail to provide a complete portrayal of the signal characteristics of HSIs. Besides, as the inverse problem of SCI is ill-posed, the infinite solutions caused by the nullspace of operator lead to ambiguity in the solution space. To overcome the above limitations, we propose a self-supervised loss function that captures the consistency, group invariance and image uncertainty of CASSI imaging, as shown in Fig. 3. By assuming that the hyperspectral signal is invariant to certain group action, our self-supervised loss greatly enriches the completeness of the measurement mask, mitigating the deficiency of information in the reconstructed signal model resulting from the

nullspace of incomplete mask. Furthermore, we effectively reduce the solution ambiguity by modeling the uncertainty in the linear inverse problem of SCI.

Measurement Consistency Loss. Due to the fact that the measurement $\mathbf{y}$ is encoded from the hyperspectral image $\mathbf{x}$ by $\mathbf{\Phi x}$, we use $\mathbf{y}$ as a label for self-supervised training. We first ensure that the inverse mapping f is consistent in the measurement domain:

$$\mathbf{\Phi} f(\mathbf{y}) = \mathbf{y}. \tag{9}$$

Thus the measurement consistency loss is defined as follows:

$$\mathcal{L}_m(\theta) = \frac{1}{N_d} \sum_{i=1}^{N_d} \|\mathbf{y}_i - \mathbf{\Phi} F_\theta(\mathbf{y}_i)\|_2^2, \tag{10}$$

where $\{\mathbf{y}_i\}_{i=1}^{N_d}$ indicates the input measurement, θ indicates the parameters to be updated by adapter. $F_\theta : \mathbb{R}^{H\times(W+d(N_\lambda-1))} \rightarrow \mathbb{R}^{H\times W\times N_\lambda}$ denotes the mapping from measurement $\mathbf{y}$ to reconstructed signal $\mathbf{x}$, which can be viewed as a composite of pre-trained model mapping and adapter mapping.

Equivariance Imaging Loss. To address the limitations imposed by a single incomplete measurement mask in the SCI system, we enrich the completeness of the mask by introducing group transformations that are invariant to the hyperspectral signal model applied to the operator $\mathbf{\Phi}$:

$$T_g F_\theta(\mathbf{\Phi x}) = F_\theta(\mathbf{\Phi} T_g \mathbf{x}), \tag{11}$$

where $\{T_g\}_{g=1,\ldots,|\mathcal{G}|}$ denotes a group of transformations (e.g., shifts, rotations, etc.) which satisfies the invariance assumption. By enforcing the equivariance in Eq. (11), learning hyperspectral model from a single mask becomes possible as the transformations $T_1, ..., T_{|\mathcal{G}|}$ allow access to more virtual masks $\mathbf{\Phi}T_1, ..., \mathbf{\Phi}T_{|\mathcal{G}|}$ with different nullspaces. The equivariance imaging loss is defined as follows:

$$\mathcal{L}_{ei}(\theta) = \frac{1}{|\mathcal{G}|N_d} \sum_{i=1}^{N_d} \sum_{\mathrm{g}=1}^{|\mathcal{G}|} \|T_g F_\theta(\mathbf{y}_i) - F_\theta(\mathbf{\Phi} T_g F_\theta(\mathbf{y}_i))\|_2^2 . \tag{12}$$

Image Uncertainty Loss. The nullspace of the operator leads to infinite solutions thus making the reconstructed hyperspectral images ambiguous. Image uncertainty loss solves the solution ambiguity by modeling uncertainty [36]. Since $\mathbf{y}$ contains no information about $\mathbf{x}$ in the nullspace of $\mathbf{\Phi}$, we use the output $F_\theta(\mathbf{y})$ as an uncertain version of GT for training. Denote $F_\theta(\mathbf{y})$ as $\mathbf{z}$, consider $\mathbf{z}' = \mathbf{z}+\mathbf{n}$ with random noise $\mathbf{n}$, we have:

$$F_\theta(\mathbf{\Phi z}') \rightarrow \mathbf{z} = \mathbf{x} + \mathbf{e}. \tag{13}$$

Assume that $\mathbf{z}'$ is a estimate of $\mathbf{x}$, the residual $\mathbf{e} = \mathbf{z} - \mathbf{z}'$ can be considered as an approximate calculation of the real residual, which reveals the ambiguity in nullspace. By minimizing the residual $\mathbf{e}$, the ambiguity in nullspace can be effectively mitigated. Noting that there is a correlation between $\mathbf{e}$ and $\mathbf{z}$, which

Table 1: PSNR/SSIM performance comparisons with/without SAH on 5 scenes of ICVL dataset (S1-S5).

Models	Category	S1	S2	S3	S4	S5	Avg
λ-Net [31]	CNN	31.31/0.898	30.40/0.816	26.45/0.774	22.70/0.720	28.28/0.839	27.83/0.810
λ-Net-SAH		33.11/0.935	32.92/0.865	28.71/0.814	24.09/0.761	31.06/0.876	29.98/0.850
TSA-Net [29]	CNN	27.97/0.733	28.52/0.717	25.70/0.748	23.59/0.753	28.47/0.843	26.85/0.759
TSA-Net-SAH		33.39/0.925	30.52/0.827	30.60/0.853	26.11/0.826	31.44/0.902	30.41/0.867
DGSMP [17]	Deep Unfolding	29.24/0.862	31.35/0.854	26.28/0.786	23.87/0.738	25.17/0.765	27.18/0.801
DGSMP-SAH		34.97/0.957	33.83/0.888	31.65/0.897	27.75/0.867	30.92/0.920	31.82/0.906
GAP-Net [28]	Deep Unfolding	30.22/0.911	32.73/0.882	25.63/0.824	23.80/0.769	25.40/0.813	27.55/0.840
GAP-Net-SAH		34.01/0.952	35.04/0.899	30.32/0.871	26.28/0.837	28.90/0.893	31.11/0.890
DAUHST [6]	Deep Unfolding	33.80/0.958	38.72/0.946	30.84/0.905	26.77/0.860	30.95/0.916	32.22/0.917
DAUHST-SAH		34.65/0.963	39.00/0.948	31.77/0.913	27.48/0.871	31.51/0.921	32.88/0.923
HDNet [16]	Transformer	31.26/0.912	33.69/0.876	27.32/0.837	25.02/0.797	27.28/0.832	28.91/0.851
HDNet-SAH		37.99/0.971	37.31/0.930	34.46/0.928	30.51/0.913	33.64/0.941	34.78/0.937
MST++ [5]	Transformer	32.13/0.937	35.29/0.901	28.67/0.880	25.60/0.826	28.42/0.881	30.02/0.885
MST++-SAH		36.68/0.966	38.40/0.941	33.70/0.927	29.34/0.898	34.04/0.940	34.43/0.935

is not consistent with the law of real residual, $\mathbf{e}$ is generated by random sign-flipping the calculation. The image uncertainty loss is defined as follows:

$$\mathcal{L}_{iu}(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d}\left\|F_\theta\left(\mathbf{\Phi}(\mathbf{z}_i+\mathbf{e})\right)-\mathbf{z}_i+\mathbf{e}\right\|_2^2, \tag{14}$$

where $\mathbf{z}_i = F_\theta(\mathbf{y}_i)$. Finally, we add total variation loss [37] to remove noise and eliminate artifacts that might caused during HSI reconstruction:

$$\mathcal{L}_{tv}(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d} TV(F_\theta(\mathbf{y}_i)). \tag{15}$$

In summary, the overall self-supervised loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_m(\theta) + \alpha\mathcal{L}_{ei}(\theta) + \beta\mathcal{L}_{iu}(\theta) + \gamma\mathcal{L}_{tv}(\theta), \tag{16}$$

where α, β and γ denote the equilibrium constants.

5 Experiments

5.1 Experiment Settings

Datasets. Different from the training datasets KAIST [10] of most pre-trained models, the simulation experiments are conducted on three publicly available HSI datasets ICVL [1], Harvard [7], and NTIRE2022 [2] with randomly clipped patches of size 256×256×28, i.e., 28 spaces of size 256×256. For real datasets, we use the 660 × 660 × 28 data captured in TSA-Net [29].

Pre-trained Models. We follow the same experiment setting and pre-train 7 base models including two CNN-based methods (λ-Net [31], TSA-Net [29]), two

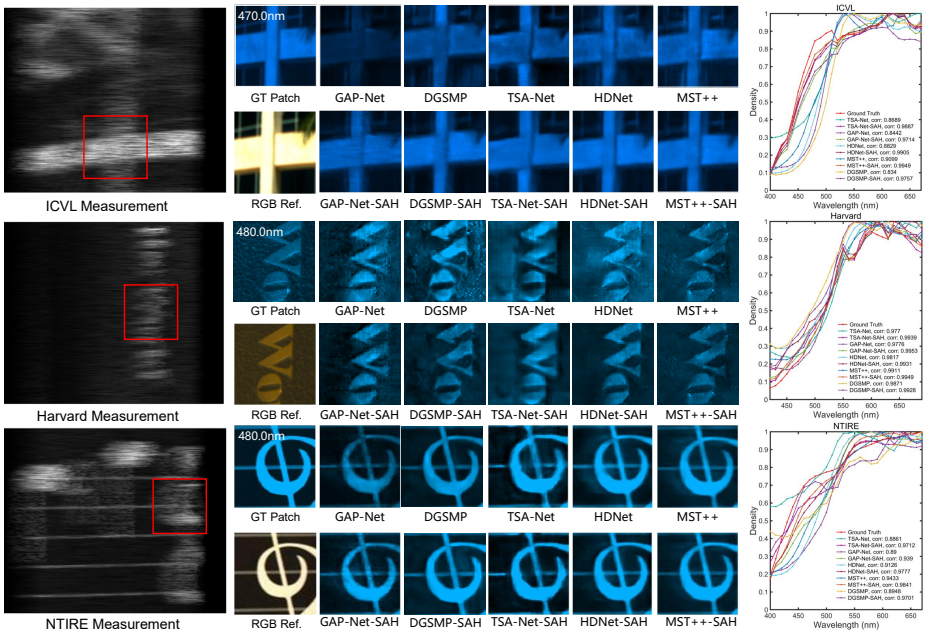


Fig. 5: Simulated HSI reconstruction comparisons on ICVL, Harvard and NTIRE2022 datasets. The right shows the spectral curves corresponding to the selected region. Choose patch for better visualization. Please zoom in for better view.

transformer-based methods (HDNet [16], MST++ [5]) and three deep unfolding methods (DGSMP [17], GAP-Net [28], DAUHST [6]).

Implementation Details. The implementation details of SAH are given in the supplementary materials. The PSNR and SSIM [44] are employed to assess HSI reconstruction quality.

5.2 Quantitative Results

We apply our adapter SAH to 7 pre-trained models described above. The PSNR and SSIM performance results on 5 scenes of ICVL dataset are listed in Tab. 1. It can be seen that our approach is universal and works for both end-to-end and deep unfolding methods. The HSI reconstruction results of models can be improved by over 2dB in average PSNR and over 0.4 in average SSIM with adaption. For HDNet, using SAH, even the best reconstruction quality can be obtained with 5.87dB/0.086 improvement where the pre-trained model does not perform well. To verify the generality of adapter on different datasets, we conduct experiments on other two datasets and the results are shown in Tab. 2. We can see that fine-tuning with our SAH brings significant improvements in the HSI reconstruction of all datasets. Besides, additional parameters and flops are just 0.29M and 5.18G, which proves the lightness of the proposed adapter.

Table 2: The average PSNR/SSIM performance, Params and GFLops with/without SAH on two datasets. Each dataset chooses 5 scenes.

Models	Category	Params	GFLops	Harvard [7]	NTIRE2022 [2]
λ-Net [31]	CNN	62.64M	117.98	28.85/0.778	30.08/0.836
λ-Net-SAH		62.93M	123.16	31.89/0.839	30.54/0.841
TSA-Net [29]	CNN	44.25M	110.06	27.68/0.726	28.18/0.772
TSA-Net-SAH		44.54M	115.24	31.73/0.835	30.49/0.833
DGSMP [17]	Deep Unfolding	3.76M	646.65	28.85/0.770	31.97/0.884
DGSMP-SAH		4.05M	651.83	32.87/0.867	33.22/0.904
GAP-Net [28]	Deep Unfolding	4.27M	78.58	31.59/0.851	33.40/0.890
GAP-Net-SAH		4.56M	83.76	34.19/0.882	34.31/0.903
DAUHST [6]	Deep Unfolding	3.44M	44.61	33.76/0.885	38.44/0.957
DAUHST-SAH		3.73M	49.79	34.77/0.901	38.71/0.959
HDNet [16]	Transformer	2.37M	154.76	30.68/0.827	34.93/0.925
HDNet-SAH		2.66M	159.94	34.04/0.882	37.10/0.943
MST++ [5]	Transformer	1.33M	19.42	32.91/0.868	35.86/0.938
MST++-SAH		1.62M	24.60	34.77/0.891	37.45/0.947

5.3 Qualitative Results

Results on Simulation Dataset. Fig. 5 compares HSI reconstruction visualization with and without adaption on ICVL, Harvard, NTIRE2022 datasets. We choose patch for better view. It can be seen that the pre-trained model produces bad reconstruction results with a lot of noise and artifacts as it has not been trained on similar datasets. In contrast, using our adapter SAH significantly reduces the noise and has fewer artifacts while recovering sharper details. The results of GAP-Net in the lower band show that our method can correct the problem of HSI reconstruction band shift caused by the limitation of training data. This is because our adapter enhances the generalization of the model to extract more spectral information through designed self-supervised loss. Besides, we calculate the spectral density curves. As shown in Fig. 5, with SAH, the HSI reconstruction results become more similar and correlated with GT, which proves the ability to improve the spectral-dimension consistency of models.

Results on Real Dataset. We further validate our method on real datasets. Following the same setting in [16, 29], 11-bit shot noise is injected into the measurements to retrain the pre-trained model. Since our method does not require GT, the real measurements are input as labels for self-supervised training of the adapter. We apply fine-tuning with SAH on λ-net [31] and HDNet [16]. The specific visualization results are shown in Fig. 6. Obviously, our method can further suppress reconstruction noise and restore clearer details based on the pre-trained models. These results demonstrate the robustness and universality of our method, which also has good prospects for real world.

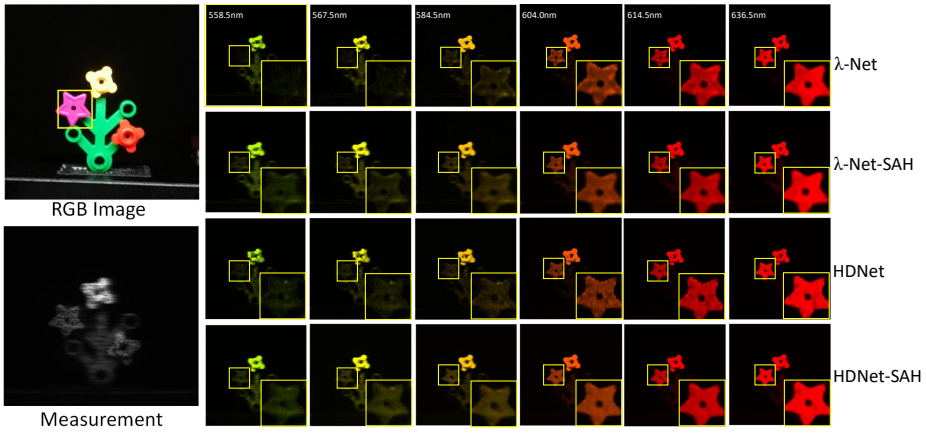


Fig. 6: Real HSI reconstruction results comparisons on Scene 1 with 6 (out of 28) spectral channels. Our SAH-SCI can effectively suppress more noise and recover clearer content. Please zoom in for better visualization performance.

6 Ablation Study

Loss Weights. The loss weights α and β in Eq. (16) are introduced to adjust the importance of $\mathcal{L}_{ei}$ and $\mathcal{L}_{iu}$ in the loss function. We compare the performance of different loss weights on HDNet and the results are shown in Tab. 3. Considering the primacy of the consistency loss $\mathcal{L}_m$, α and β are maximised to 1.0. It can be seen that better performance is usually achieved when $\alpha \geq \beta$, which indicates that $\mathcal{L}_{ei}$ plays a more important role in the loss function. When $\alpha = 0.7$ and $\beta = 0.4$, the model achieves the highest PSNR and SSIM performance. Since our loss function is primarily designed to reduce the feasible solution space, the weight γ is set to 0.001 to embellish the HSI reconstruction results.

Table 3: Comparisons of different loss weights on HDNet.

α \ β	0.1	0.4	0.7	1.0
0.1	33.86 0.928	34.02 0.934	34.38 0.936	34.17 0.935
0.4	34.17 0.931	34.57 0.937	34.35 0.937	33.64 0.932
0.7	34.48 0.936	**34.78** **0.937**	34.41 0.936	34.30 0.935
1.0	34.29 0.934	34.46 0.937	34.52 0.936	34.24 0.937

Model analysis of SAH. We analyze the performance of different components of SAH in Tab. 4, where FA refers to frequency adaption, SSconv refers to Spatial-Spectral convolution and CW refers to channel weight. Case (a) uses the original UNet structure. Tab. 4 (a,b) shows that our frequency adaption can bring an improvement of 0.74dB. Compared to a 2-layer structure, the result is only improved by 0.2dB when using a 3-layer structure, while requires more than four times parameters (case (b,c)). Thus

Table 4: Components analysis in SAH on λ-Net.

Case	Layers	FA	SSconv	CW	PSNR	SSIM	Params	GFLops
(a)	2	×	×	×	29.21	0.836	2.09M	35.14
(b)	2	✓	×	×	29.95	0.849	2.09M	35.14
(c)	3	✓	×	×	30.15	0.854	8.58M	50.86
(d)	2	✓	✓	×	29.45	0.838	0.26M	4.78
(e)	2	✓	✓	✓	29.98	0.850	0.29M	5.18

the 2-layer structure is chosen as the infrastructure of SAH. We test the use of the preliminary SSconv without channel weighting (case (d)), the results indicate that using SSconv instead of normal convolution can reduce parameters and GFLops to about one-eighth, while only dropping the performance by 0.5dB, which demonstrates the lightweight nature of the SSConv. Finally, we add the channel weighting module to further enhance the reconstruction quality by utilizing inter-spectral information from the shuffled results (case e), SSconv with channel weights can achieve better performance while remaining lightweight. The above results demonstrate the effectiveness of our SAH structure.

Loss ablations. To evaluate the contribution of different loss components, we conduct the ablation study on ICVL by turning them on and off. Tab. 5 shows the effects of each loss composition on the HSI reconstruction performance, where ✓ indicates that the method applies this loss component. We can observe that $\mathcal{L}_m$ significantly acts as the main loss component, improving the pre-trained model by 1-4dB. Compared to $\mathcal{L}_{iu}$, $\mathcal{L}_{ei}$ can improve the reconstruction quality more effectively by about 1dB, which is in line with what we mentioned in the precious ablation study. The results of case (e) show that $\mathcal{L}_{iu}$ and $\mathcal{L}_{ei}$ jointly contribute to reducing the feasible solution space, which is consistent with the starting point of the loss function design. Our method achieves the best performance when including all the loss components, as shown in Tab. 5 (f).

Table 5: Ablation study of different loss components on ICVL.

Case	Adaption	L_m	L_{iu}	L_{ei}	L_{tv}	λ-Net [31]	GAP-Net [46]	HDNet [16]	MST++ [5]
(a)	×	×	×	×	×	27.83/0.810	27.55/0.840	28.91/0.851	30.02/0.885
(b)	✓	✓	×	×	×	28.89/0.823	29.86/0.864	33.35/0.926	33.05/0.922
(c)	✓	✓	✓	×	×	28.92/0.823	29.92/0.867	33.79/0.930	33.58/0.926
(d)	✓	✓	×	✓	×	29.82/0.849	30.78/0.888	34.31/0.934	34.28/0.932
(e)	✓	✓	✓	✓	×	29.95/0.850	31.03/0.889	34.50/0.936	34.40/0.933
(f)	✓	✓	✓	✓	✓	**29.98/0.850**	**31.11/0.890**	**34.78/0.937**	**34.43/0.935**

7 Conclusion

In this paper, we consider the common problem with existing SCI reconstruction algorithms, i.e., poor generalization ability due to reliance of training data. To address this problem, a universal self-supervised fine-tuning framework SAH-SCI is proposed for HSI restoration. SAH-SCI preserves the parameters of the pre-trained model and enhances model generalization via training a lightweight self-supervised adapter. Furthermore, we develop a lightweight adapter, SAH, incorporating spatial-spectral convolution to simultaneously capture spatial and spectral characteristics and reduce the adapter's complexity. Additionally, we customize a self-supervised loss function for the adapter to address the issues of solution ambiguity and nullspace information deficiency. With these novel techniques, the generalization of pre-trained models can be effectively enhanced by adaption. Extensive experiments demonstrate SAH-SCI superiority over previous methods with fewer parameters, offering simplicity and adaptability to diverse image pre-trained models, which paves the way for leveraging more robust image foundation models in future hyperspectral imaging tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62106063 and 62106081, by the Guangdong Natural Science Foundation under Grants 2022A1515010819 and 2022A1515010 800, and Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000010.

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: European Conference on Computer Vision. pp. 19–34. Springer (2016)
2. Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y.: Ntire 2022 spectral recovery challenge and data set. In: CVPRW. pp. 862–880 (2022)
3. Barbastathis, G., Ozcan, A., Situ, G.: On the use of deep learning for computational imaging. Optica **6**(8), 921–943
4. Borengasser, M., Hungate, W.S., Watkins, R.: Hyperspectral remote sensing: Principles and applications. CRC press (2007)
5. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Gool, L.V.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: CVPR (2022)
6. Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., Gool, L.V.: Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging (2022)
7. Chakrabarti, A., Zickler, T.: Statistics of real-world hyperspectral images. In: CVPR (2011)
8. Chen, D., Tachella, J., Davies, M.E.: Equivariant imaging: Learning beyond the range space. In: ICCV (2021)
9. Chen, D., Tachella, J., Davies, M.E.: Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements. In: CVPR (2022)
10. Choi, I., Jeon, D.S., Nam, G., Gutierrez, D., Kim., M.H.: High-quality hyperspectral reconstruction using a spectral prior. ACM TOG (2017)
11. Cole, E.K., Pauly, J.M., Vasanawala, S.S., Ong, F.: Unsupervised mri reconstruction with generative adversarial networks. arXiv preprint arXiv:2008.13065 (2020)
12. Fu, Y., Zheng, Y., Sato, I., Sato, Y.: Exploiting spectral-spatial correlation for coded hyperspectral image restoration. In: CVPR. pp. 3727–3736 (2016)
13. Gehm, M.E., John, R., Brady, D.J., Willett, R.M., Schulz, T.J.: Single-shot compressive spectral imaging with a dual-disperser architecture. Optics Express **15**(21), 14013–14027 (2007)
14. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021)
15. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (2019)
16. Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Gool, L.V.: Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In: CVPR (2022)
17. Huang, T., Dong, W., Yuan, X., Wu, J., Shi, G.: Deep gaussian scale mixture prior for spectral compressive imaging. In: CVPR (2021)

18. Kim, M.H., Harvey, T.A., Kittle, D.S., Rushmeier, H., Dorsey, J., Prum, R.O., Brady, D.J.: 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. ACM Transactions on on Graphics **31**(38), 1–11 (2012)
19. Kittle, D., Choi, K., Wagadarikar, A., Brady, D.J.: Multiframe image estimation for coded aperture snapshot spectral imagers. Applied optics **49**(36), 6824–6833 (2010)
20. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: CVPR (June 2019)
21. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. pp. 12888–12900 (2022)
22. Lin, X., Liu, Y., Wu, J., Dai, Q.: Spatial-spectral encoded compressive hyperspectral imaging. ACM TOG (2014)
23. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank minimization for snapshot compressive imaging. IEEE TPAMI **41**(12), 2990–3006 (2019)
24. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. Journal of Biomedical Optics (2014)
25. Ma, J., Liu, X.Y., Shou, Z., Yuan, X.: Deep tensor admm-net for snapshot compressive imaging. In: ICCV. pp. 10222–10231 (2019)
26. Ma, X., Yuan, X., Fu, C., Arce, G.R.: Led-based compressive spectral-temporal imaging. Optics Express (2021)
27. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing (2004)
28. Meng, Z., Jalali, S., Yuan, X.: Gap-net for snapshot compressive imaging (2020)
29. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: ECCV (2020)
30. Meng, Z., Qiao, M., Ma, J., Yu, Z., Xu, K., Yuan, X.: Snapshot multispectral endomicroscopy. Optics Letters (2020)
31. Miao, X., Yuan, X., Pu, Y., Athitsos, V.: lambda-net: Reconstruct hyperspectral images from a snapshot measurement. In: ICCV (2019)
32. Nguyen, H.V., Banerjee, A., Chellappa, R.: Tracking via object reflectance using a hyperspectral video camera. In: CVPR. pp. 44–51 (2010)
33. Pan, Z., Healey, G., Prasad, M., Tromberg, B.: Face recognition in hyperspectral images. IEEE TPAMI **25**(12), 1552–1560 (2003)
34. Park, J.I., Lee, M.H., Grossberg, M.D., Nayar, S.K.: Multispectral imaging using multiplexed illumination. In: ICCV (2007)
35. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is chatgpt a general-purpose natural language processing task solver? arXiv preprint arXiv:2302.06476 (2023)
36. Quan, Y., Qin, X., Pang, T., Ji, H.: Dual-domain self-supervised learning and model adaption for deep compressive imaging. In: ECCV (2022)
37. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena **60**(1), 259–268 (1992)
38. Si, C., ang Yuming Jiang, Z.H., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: CVPR (2023)
39. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. Applied Optics (2008)
40. Wang, L., Xiong, Z., Gao, D., Shi, G., Wu, F.: Dual-camera design for coded aperture snapshot spectral imaging. Applied optics **54**(4), 848–858 (2015)

41. Wang, L., Xiong, Z., Gao, D., Shi, G., Zeng, W., Wu, F.: High-speed hyperspectral video acquisition with a dual-camera architecture. In: CVPR (2015)
42. Wang, L., Xiong, Z., Huang, H., Shi, G., Wu, F., Zeng, W.: High-speed hyperspectral video acquisition by combining nyquist and compressive sampling. IEEE TPAMI (2018)
43. Wang, L., Xiong, Z., Shi, G., Wu, F., Zeng, W.: Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. IEEE TPAMI **39**(10), 2104–2111 (2017)
44. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
45. Wu, Y., Mirza, I.O., Arce, G.R., Prather, D.W.: Development of a digital-micromirror-device-based multishot snapshot spectral imaging system. Optics letters (2011)
46. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. In: ICIP. pp. 2539–2543 (2016)
47. Yuan, X., Tsai, T.H., Zhu, R., Llull, P., Brady, D., Carin, L.: Compressive hyperspectral imaging with side information. IEEE Journal of Selected Topics in Signal Processing (2015)
48. Yuan, Y., Zheng, X., Lu, X.: Hyperspectral image superresolution by transfer learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2017)
49. Zhang, J., Zhao, C., Zhao, D., Gao, W.: Image compressive sensing recovery using adaptively learned sparsifying basis via l0 minimization. Signal Processing (2014)
50. Zhang, S., Wang, L., Fu, Y., Zhong, X., Huang, H.: Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In: ICCV. pp. 10182–10191 (2019)
51. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. CVPR pp. 6848–6856 (2018)
52. Zhang, X., Zhang, Y., Xiong, R., Sun, Q., Zhang, J.: Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In: CVPR (2022)
53. Zhao, C., Ma, S., Zhang, J., Xiong, R., Gao, W.: Video compressive sensing reconstruction via reweighted residual sparsity. IEEE Transactions on Circuits and Systems for Video Technology **27**(6), 1182–1195 (2017)