Progressive Self-Supervised Learning for CASSI Computational Spectral Cameras

Xiaoyin Mei, Yuqi Li, Qiang Fu, and Wolfgang Heidrich, *Fellow, IEEE*

***Abstract*—Compressive spectral imaging (CSI) is a technique used to capture high-dimensional hyperspectral images (HSIs) with a few multiplexed measurements, thereby reducing data acquisition costs and complexity. However, existing CSI methods often rely on end-to-end learning from training sets, which may struggle to generalize well to unseen scenes and phenomena. In this paper, we present a progressive self-supervised method specifically tailored for coded aperture snapshot spectral imaging (CASSI). Our proposed method enables HSI reconstruction solely from the measurements, without requiring any ground truth spectral data. To achieve this, we integrate positional encoding and spectral cluster-centroid features within a novel progressive training framework. Additionally, we employ an attention mechanism and a multi-scale architecture to enhance the robustness and accuracy of HSI reconstruction. Through extensive experiments on both synthetic and real datasets, we validate the effectiveness of our method. Our results demonstrate significantly superior performance compared to state-of-the-art self-supervised CASSI methods, while utilizing fewer parameters and consuming less memory. Furthermore, our proposed approach showcases competitive performance in terms of reconstruction quality when compared to state-of-the-art supervised methods.**

***Index Terms*—Self-supervised compressive reconstruction, positional encoding, hyperspectral Imaging.**

I. Introduction

HYPERSPECTRAL images (HSIs) are extensively employed in various fields such as remote sensing [1], tracking [2], medical image processing [3], and high color fidelity display [4]. Traditional spectral imaging systems based on Nyquist's sampling theorem demonstrate a high degree of accuracy in the reconstruction of original HSI data. However, owing to physical limitations, image sensors are restricted to acquiring two-dimensional images, leading to a need for a scanning-based strategy for HSI data acquisition [5], which can significantly impact data acquisition efficiency. Furthermore, acquiring and storing high-dimensional spectral data requires significant storage space and network transmission bandwidth. To address these challenges, some recent studies estimate high-resolution HSIs from RGB images, however these approaches are inherently limited in their ability to discriminate metameric colors i.e. different spectra that map to the same RGB triplet, which is an important consideration in many applications of HSIs [6]. RGB to HSI methods can therefore not be considered spectral measurements, but are in effect *hallucinations* of spectral information learned from a limited dataset, making them ill-suited for scientific applications. Other methods employ encoded lenses to recover HSIs but may impact the contrast of the resulting HSIs [7].

To ensure the fidelity of reconstructed HSIs on both spectral and spatial dimensions, various snapshot compressed spectral imaging (CSI) systems have been developed [8], [9], [10], [11], [12]. In these systems, upon entering the cameras, each beam is dispersed by a diffraction grating to separate its spectral components. Subsequently, the modulated spectral information is encoded by a modulation mask before reaching the sensors. The measurements gathered by the sensors are utilized to recover the HSIs using reconstruction algorithms. This technique employs the theory of compressed sensing [13], [14] that allows the reconstruction of the hyperspectral images using a smaller number of measurements than those required by traditional spectral imaging systems, thus enhancing efficiency and reducing storage and transmission costs. Unlike RGB to HSI estimation methods, compressed sensing approaches modify the capture process to encode additional spectral information, and hence can provide guaranteed signal restoration under certain sparsity assumptions. Consequently, CSI has become a viable solution for the recovery of high-quality HSIs. The underlying principle of the CSI system is to encode the high-dimensional data into a 2D measurement. Among various hyperspectral image acquisition systems, the coded aperture snapshot compressive imager (CASSI) is widely used. CASSI uses a coded aperture and a disperser to modulate the HSI cube at different wavelengths and then mixes all modulated signals to generate 2D measurements. Fig. 2 shows a schematic diagram and reconstruction process of CASSI. Given an HSI cube, denoted by $\mathbf{x} \in R^{hwn\times 1}$, where (h, w) and n represent the spatial and spectral resolutions of the HSI respectively. A sensing matrix $\Phi \in R^{hw\times hwn}$ is designed to project $\mathbf{x}$ into a measurement by

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{z}, \tag{1}$$

Received 17 March 2024; revised 12 June 2024 and 11 August 2024; accepted 2 September 2024. Date of publication 18 September 2024; date of current version 23 October 2024. This work supported in part by the Ningbo Major Research and Development Plan Project under Grant 2023Z225 and in part by the Ningbo City Public welfare Science and technology Plan Project under Grant 2024. The associate editor coordinating the review of this article and approving it for publication was Prof. Henry Arguello. *(Xiaoyin Mei and Yuqi Li contributed equally to this work.) (Corresponding author: Yuqi Li.)*

Xiaoyin Mei and Yuqi Li are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315600, China (e-mail: 2111082080@nbu.edu.cn; liyuqi1@nbu.edu.cn).

Qiang Fu and Wolfgang Heidrich are with the Visual Computing Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia (e-mail: qiang.fu@kaust.edu.sa; wolfgang.heidrich@kaust.edu.sa).

The related code and data are available at https://github.com/ccccddd1/ceinr.

Digital Object Identifier 10.1109/TCI.2024.3463478

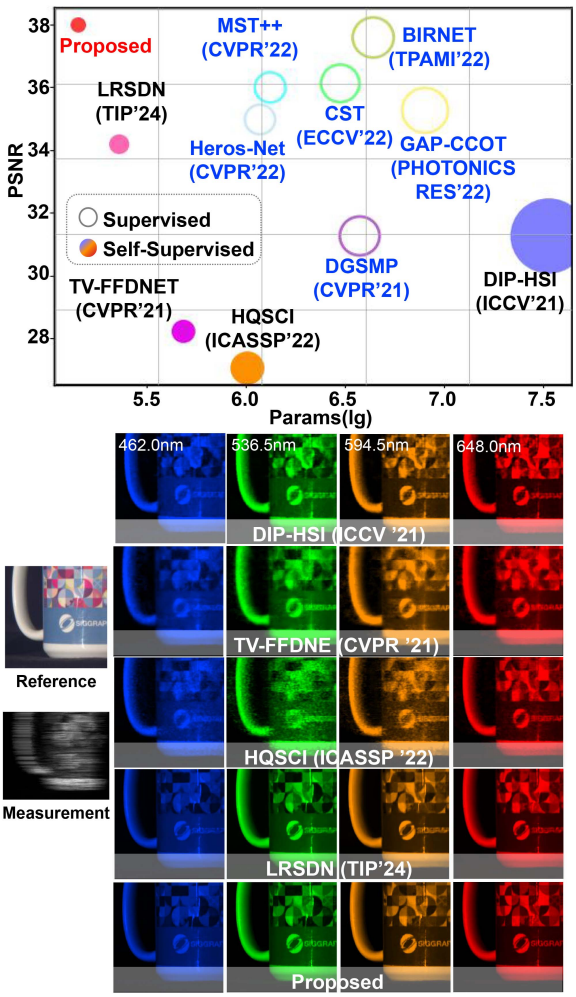

Fig. 1. The comparisons of performance and visual reconstruction quality. Upper: The reconstruction quality(PSNR), parameter amount, and the relative memory consumption(circle radius) of the recent self-supervised(solid circle) and supervised CSI methods(hollow circle). Bottom: The visual comparison of the reconstructed HSI of the five self-supervised methods from a compressed image.

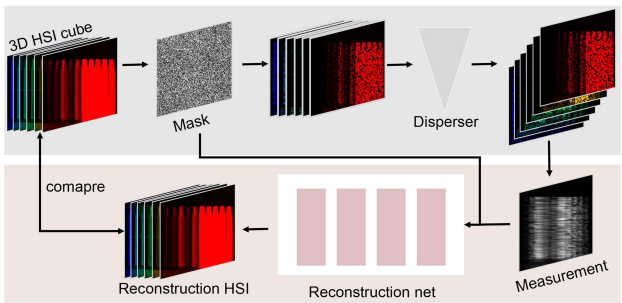

Fig. 2. The schematic diagram and reconstruction process of CASSI. It uses a disperser and a mask to modulate the HSI, and then the measurements are processed through a reconstruction network to obtain the reconstructed HSI.

where $\mathbf{z} \in R^{hw\times 1}$ denotes the noise.

The optimization of CSI is typically framed as an ill-posed inverse problem. Existing solutions either rely on hand-crafted image priors or employ deep networks to perform the HSI reconstruction in a data-driven manner. Hand-crafted image priors, such as total variation [15], [16], and sparsity [17], [18], can be utilized as regularizers to enforce constraints on the reconstructed results, thereby mitigating the risk of over-fitting and suppressing the impact of noise while enforcing a desired structure. However, such simple handcrafted priors cannot compete with recent learned representations, and often require scene-specific parameter tuning. On the other hand, deep supervised methods require a large-scale dataset to acquire implicit prior knowledge from the images [19], which is impractical in the fields of life and natural sciences, including mechanics, materials science, and biology, due to the challenge of lacking training data. Even if the data can be captured, spectral image datasets are highly sensitive to spectral bands and resolutions. Models trained on datasets ranging from 400 nm to 700 nm cannot be applied to test images ranging from 400 nm to 800 nm. It is also worth noting that recent concurrent work [20] has shown very clearly the limitations of data-driven spectral reconstruction methods, and specifically suggests that the good numerical results from current approaches are misleading and due to overfitting, resulting in very poor generalization. Additionally, due to the spatial invariance of CNNs, using CNN-based networks for spectral reconstruction can lead to the problem of spectral ambiguity. [21] addresses this issue by using positional-encoding vector as input to resolve the spatial invariance problem. However, the reconstruction process using only positional-encoding is insufficient and often generates unnecessary artifacts without ground truth. Therefore, a possible solution is to obtain features from spectral domain to assist the positional-encoding based reconstruction network. By combining spectral features and positional encoding vectors, it is possible to effectively distinguish spectra and remove artifacts.

In this paper, we focus on the reconstruction of compressed spectral images without labeled data for training, which holds significant relevance and application within the area where datasets are unavailable. Unlike previous works that utilize handcrafted denoising priors or explicit CNN denoising priors, we propose to reconstruct HSIs through a hybrid framework combining positional-encoding representation in the spatial domain and features of cluster-centroid in the spectral domain. The proposed method can be seen as progressive fitting processing of the target HSI within a high-dimensional space that is spanned by the pixel coordinates and the corresponding cluster centroids. The contributions of this paper can be summarized as follows:

- We design a novel deep network framework for single image CSI without any labeled data for training. To the best of our knowledge, we are the *first* to successfully combine the benefits of clustering features and positional-encoding representation in a self-supervised fashion to reconstruct HSIs.
- We put forward a novel progressive training approach that integrates clustering uncertainty to avoid over-fitting. Additionally, we leverage attention mechanisms and a multi-scale architecture to enhance the reconstruction quality of HSIs and apply noise injection to ensure the robustness of the method.
- Quantitative and qualitative experiments demonstrate that the proposed method surpasses the existing self-supervised

learning method by *6.7 dB* in PSNR and *0.06* in SSIM, the method also exhibits comparable performance to the state-of-the-art (SOTA) supervised learning methods with much fewer training parameters and lower memory consumption.

II. Related Works

There have been many recent efforts to study CSI. Most of the techniques pursued can be classified into optimization-based and feature-based.

A. Optimization-Based Methods

To address the ill-posed inverse problem in CSI, conventional approaches typically rely on hand-crafted priors, including sparsity [11], [18], total variation [15], and non-local similarity [22], to accomplish the regularization of the reconstruction process. The advent of deep learning has yielded significant advances in the application of CSI. Deep unfolding networks [23], [24], [25], [26], [27], [28], [29], [30], [31] combine the advantages of model-based optimization and deep discriminative learning, unroll the optimization algorithms utilized in compressive reconstruction, and employ diverse deep image denoising priors to substitute manually designed priors. This yields a remarkable level of flexibility in constructing compressive masks and enhances the representational capacity of a wide array of hyperspectral images (HSIs). More recently, researchers have investigated more efficient linear mapping algorithms to accelerate optimization [32], [33]. Additionally, some studies [25], [34] have utilized self-attention transformer denoising modules to enhance the reconstruction quality. Although unfolding networks provide promising reconstructed results with clear interpretability, they require a significant amount of data to be effectively trained.

B. Feature-Based Methods

Unlike optimization-based methods, feature-based methods only focus on compressive reconstruction with specific compressive masks, and reconstruct HSIs by analyzing the features extracted from the measurements. In such methods, synthetically obtained measurements serve as the input data, and the target HSIs are treated as labels for training the reconstruction models under the aforementioned masks. These network architectures include but are not limited to fully connected networks [35], convolutional neural networks [36], [37], and generative adversarial networks [38].

However, the small receptive fields of the convolutional neural networks and fully connected networks used in these methods lead to inherent limitations in capturing the non-local self-similarity and long-range dependencies that are crucial for HSI reconstruction. The emergence of Transform [39] brings new perspectives to introduce non-local attention to the deep networks. For example, MST [40] captures the long-range inter-spectra dependencies by computing the self-attention in the spectral domain. S^2-Transformer [41] exploits the spatial-spectral domain to model spatial dependencies. CST [42] attempts to embed the HSI spatial sparsity nature into a Transformer-based method. Despite yielding notable outcomes, these approaches still rely on supervised training and entail significant computational expenses due to the incorporation of the Transformer model.

C. Self-Supervised Methods

Supervised networks require sufficient training data and time and are usually application and domain-specific. Therefore, effective unsupervised algorithms are still highly desired as researchers are eager to apply CSI to unseen scenes. PnP-DIP-HSI [43] develops a self-supervised framework by integrating deep image prior (DIP) into an optimization procedure without any training set. TV-FFDNET [44] proposes a general framework that uses pretrained denoiser and total variation priors for CSI reconstruction. LR2DP [45] integrates the low-rank prior and deep image priors for SCI reconstruction, in order to exploit the strong spectural correlation and deep spatial structure of HSI. Furthermore, LRSDN [46] embeds data-driven low-rank subspace representations and data-driven attention networks within an iterative optimization framework. More recently, another self-supervised method [47] adopts an unrolling network with an ensemble process to enhance the reconstruction quality. However, the current self-supervised or unsupervised approaches lead to a large number of parameters and excessive memory consumption due to their explicit representation and may lack robustness when handling high-dimensional data such as videos [48]. This prompts us to seek other alternatives to effectively and efficiently reconstruct HSI without a training set.

D. Positional-Encoding Representation

Traditional modeling approaches, like deep image priors, often require extensive features and are susceptible to overfitting with limited supervised data [49], [50]. The emergence of coordinate-based MLPs offers a novel perspective, which takes low-dimensional coordinates as input and outputs desired representations at each position. Coordinate-based MLPs have achieved state-of-the-art results across a variety of tasks such as [51], [52], [53]. However, standard MLPs are poorly suited for these low-dimensional coordinate-based vision and graphics tasks. In particular, MLPs have difficulty learning high frequency functions, a phenomenon referred to in the literature as "spectral bias" [54], [55]. Recently, positional-encoding representation methods have gained traction, particularly in 3D vision tasks such as geometric reconstruction [56], [57] and novel-view synthesis [58]. These methods represent signals as parameterized functions of Fourier features, using a coordinate vector $p \in \mathbb{R}^d$, and reconstruct signals via a coordinate network [48]. Unlike conventional encoding, these signals are implicitly encoded within the network parameters, facilitating efficient high-dimensional data representation [21]. However, the potential of positional-encoding for spectral reconstruction remains unexplored. This paper introduces a method that combines positional-encoding and embedded clustering features for precise CSI, leveraging the capability to represent and reconstruct spectral images with fewer parameters. The approach offers robustness and adaptability for diverse data and conditions, making it valuable for real-world applications [55], [59].

III. Method

A. Motivation

To make CSI applicable across multiple application domains which may include previously unobserved phenomena and limited trained data, we focus on the reconstruction of HSI from a single compressed image measurement using a compact neural network architecture, which does not rely on any supervised training data and hand-crafted priors. Our method is a self-supervised method that uses positional-encoding representation and spectral clustering features. The combination of the clustering features and positional-encoding representation can form a complementary approach that constructs data in high dimensions. In addition, it adopts a progressive training strategy and introduces some uncertainty via clustering to avoid trapping into local minima.

The previous modeling approaches, such as deep image priors, necessitates numerous features and parameters to represent data, and are prone to overfitting when faced with inadequate supervised training data. Although deep image priors can capture the intrinsic structure of an image, it may not effectively recognize the spatial location information, particularly when dealing with images that exhibit complex spatial variations. Due to the ability to accurately represent various types of data, representation methods based on positional-encoding have become increasingly popular in recent years, especially in 3D vision tasks like geometric reconstruction [49], [56], [57], [60], [61] and novel-view synthesis [58], [62], [63]. This technique represents a specific signal as a parameterized function of Fourier features, with an input consisting of a coordinate vector $p \in \mathbb{R}^d$, and reconstructs the desired signal via a coordinate network. Unlike traditional encoding methods, the signals are implicitly encoded in the coordinate network parameters, allowing for efficient representation of high-dimensional data. The Multilayer perceptron (MLP) and 1×1 Unet are commonly utilized coordinate networks in this technique [48]. Before feeding into the coordinate network, the coordinates p need to be mapped to a higher dimensional space by positional encoding [11], [52], which allows the network to learn the high-frequency component of the desired signals, since simply taking coordinates as input can lead to learning only the low-frequency part of the input signal and exhibit global behavior [55], [59], [64], [65], [66].

To the best of our knowledge, no studies have explored the potential of positional-encoding representation to spectral reconstruction, which involves recovering spectral information from compressive images of objects or scenes. However, positional-encoding networks can be overly sensitive to high-frequency noise, which adversely affects the quality of HSI reconstruction. Therefore, we attempt to utilize the self-similarity of spectra through clustering to perform denoising. Combining spectral features with positional encoding vectors as high-dimensional inputs to the reconstruction network enables the capture of a broader range of data characteristics, enhancing the model's generalization capability. This paper aims to adopt both positional-encoding and embedded clustering features for accurate CSI, building on the ability to accurately represent and reconstruct spectral images with fewer parameters than traditional deep models. Fourier feature mapping can represent complex relationships between input features and output images, allowing for a more accurate reconstruction of spectral data by providing the image structure prior. In addition, positional-encoding in CSI can also lead to more robust and generalizable models, able to handle a wider range of data and conditions. This can be especially useful in real-world scenarios where conditions may vary or data may be noisy. Inspired by the benefits of positional-encoding, we encode a continuous function into the parameters of the neural network to facilitate the reconstruction of the target HSI.

In the following sections, we first introduce our proposed architecture and progressive training strategy. Then we present the two major modules in our method and the training details. Lastly, we evaluate the effectiveness of our method through simulations.

B. Progressive Training

As shown in Fig. 3, we propose a progressive deep neural network training strategy that involves multiple iterations. Without loss of generality, in the i-th iteration, we treat the reconstructed HSI $\hat{\mathbf{X}}^{(i-1)}$ of the last iteration and the coordinate tensor $\mathbf{\Gamma}$ as inputs and generate the HSI $\hat{\mathbf{x}}^{(i)}$ by using the proposed network. The proposed framework consists of two modules, which are as follows: (1) a spectral cluster-centroid denoising module (SCCD) $\mathcal{G}$ that can be treated as a simple but efficient non-local means denoising algorithm in spectral dimension and uses pixel-level clustering and an efficient attention block to capture long-distance relationships; (2) a compact multi-scale aggregation (MSA) module $\mathcal{F}$ with pyramid structure blocks are used to map the hybrid high-dimensional encoded tensor to an HSI.

Note that the coordinate tensor is not input directly to the MSA module, but rather mapped to a higher-dimensional space through a position encoding before being fed into the network. The network has the ability to incorporate the influence of positional information when learning in high-dimensional spaces, thereby enhancing its capability to capture intricate details and structural characteristics of the input space [67]. We utilize frequency (sinusoidal) encoding for spatial coordinates and map a coordinator p to a vector $\gamma(p) \in \mathbb{R}^{4L+2}$. The positional encoding operation of a single coordinate, denoted by $\gamma(\cdot)$, is defined as follows:

$$\begin{aligned}\gamma(p) = \big(&p, \sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right),\\ &\cdots \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\big),\end{aligned} \tag{2}$$

where $p = (u, v)$ denotes the normalized coordinate value lying in $[-1, 1]$, and L denotes the order of frequency of the vector.

To reconstruct the desired HSI $\mathbf{x}$ from the measured image $\mathbf{y}$, we generate an initial guess $\hat{\mathbf{x}}^{(0)} = \Phi^T\mathbf{y}$, and optimize the $\hat{\mathbf{x}}$ iteratively. In order to ensure high reconstruction quality, we couple the representation of the HSI after the SCCD module $\mathcal{G}$ and the representation $\gamma(\mathbf{p})$ after positional encoding in the MSA module $\mathcal{F}$, where $\mathbf{p}$ is the coordinate image. The trainable parameters θ in the MSA module $\mathcal{F}$ and the parameters ω in the

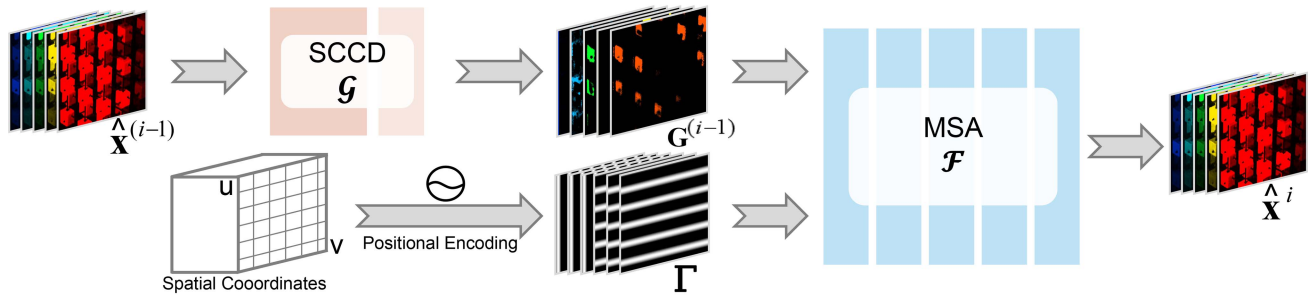

Fig. 3. Our progressive training with positional-encoding representation and embedded feature of spectral cluster-centroid for self-supervised learning. MSA is fed with the positional encoded tensor of spatial coordinates and the output tensor of the denoising module SCCD as inputs and generates the intensity of reconstructed HSI as output. The reconstruction $\hat{\mathbf{x}}^{(i)}$ serves as the input for the $x+1$ iteration of the SCCD module to obtain spectral centroid features $\mathbf{G}^{(i)}$, which are continuously optimized as training progresses, gradually increasing their contribution to the network.

SCCD module $\mathcal{G}$ are jointly optimized in the i-th iteration as:

$$\arg\min_{\theta,\omega} \left\| \mathbf{y} - \mathbf{\Phi}\mathcal{F}\left(\mathbf{G}^{(i-1)}, \mathbf{\Gamma};\ \theta\right)\right\|_1,$$
$$s.t.\quad \mathbf{\Gamma} = \gamma(\mathbf{p}),\quad \mathbf{G}^{(i-1)} = \mathcal{G}(\hat{\mathbf{x}}^{(i-1)};\omega), \tag{3}$$

where $\mathbf{\Gamma}$ denotes the positional encoding tensors of coordinates, and $\mathbf{G}^{(i-1)}$ denotes the output tensors of modules the SCCD of the i-th iteration. The goal of the objective function is to minimize the reprojection error of the reconstructed HSI $\hat{\mathbf{x}}^{(i)}$ as it can be obtained directly by:

$$\hat{\mathbf{x}}^{(i)} = \mathcal{F}\left(\mathbf{G}^{(i-1)}, \mathbf{\Gamma};\ \theta\right). \tag{4}$$

We will introduce the details of the two modules in the following sections.

C. Spectral Cluster-Centroid Denoising Module

Non-local image denoising algorithms are widely acknowledged as effective techniques for removing noise from images. These algorithms operate on the principle that patches in an image that is similar in structure also share similar noise characteristics. In this study, we exploit the observation that HSI exhibits a degree of self-similarity in their spectra and that this similarity can be exploited to estimate the true underlying structure of the image. Pixels with similar spectra can be grouped together by clustering, allowing for specialized processing of each group of similar pixels. This approach effectively enhances the denoising performance. To this end, we propose the SCCD module that first identifies similar spectra for each pixel and replaces the pixel's spectrum with the mean of these similar spectra. Following this, a convolution modulation block is employed to further reduce noise in the image and generate the denoised HSI $\mathbf{G}^{(i-1)}$ which is considered as the clustering feature representation of the image.

Rather than calculating the similarity between every pair of pixels, we utilize a clustering method directly on the HSI. Within each cluster, we consider the similarity between pixels to be 1, while for the pixel pairs in different clusters, we assign a similarity of 0. There are two key advantages to utilizing a clustering method in this context. Firstly, the computational complexity of clustering is significantly less than that of non-local means, and it can be further improved through the use of GPU computing [69]. Secondly, clustering can introduce more uncertainty than deterministic non-local means, which can help to avoid local minima and ensure robustness.

We denote the reshaped matrix of $\mathbf{x} \in \mathbb{R}^{hwn\times 1}$ as $\mathbf{X} \in \mathbb{R}^{hw\times n}$, thus the spectrum of j-th pixel is denoted as $\mathbf{X}_j \in \mathbb{R}^{1\times n}$, where $j = 1, 2, \ldots, hw$. In the i-th iteration, we apply k-means algorithm to divide the set of estimated spectrum $\{\hat{\mathbf{X}}_j^{(i-1)} | j = 1, 2, \ldots, hw\}$ into $k^{(i)}$ clusters and calculate the centroid for each cluster, where $k^{(i)}$ is the number of clusters in the i-th iteration. Suppose $\bar{\mathbf{X}}^{(i-1)} \in \mathbb{R}^{hw\times n}$ is the matrix constructed by using the spectrum of each cluster centroid, thus $\bar{\mathbf{X}}_j^{(i-1)}$ is the centroid spectrum of the cluster that $\hat{\mathbf{X}}_j^{(i-1)}$ belongs to. Here the use of the centroid matrix $\bar{\mathbf{X}}^{(i-1)}$ to represent the current reconstructed HSI matrix $\hat{\mathbf{X}}^{(i-1)}$ is based on the observation that HSI images can often be reconstructed from a few samples [70].

Note that $k^{(i)}$ is the cluster number that varies with the number of iterations. We design a simple clustering number that grows linearly with the number of iterations, as shown in:

$$k^{(i)} = \begin{cases} \left\lfloor \frac{i}{\alpha} \right\rfloor, & \mod(i,\beta) \neq 0, \\ 1, & \mod(i,\beta) = 0, \end{cases} \tag{5}$$

where α is an empirical parameter to adjust the growth rate of the number of clusters, and $\lfloor\cdot\rfloor$ is the round-down operation. To introduce more uncertainty into the reconstruction for avoiding overfitting, we set the number of clusters to 1 every β iteration to allow the network to escape from local optima. Note that at the beginning of the reconstruction iterations, the number of clusters is set to 0, and we use $\mathbf{\Phi}^T\mathbf{y}$ as the initial reconstructed HSI matrix $\hat{\mathbf{X}}^{(0)}$ in this case.

The design of this mechanism is mainly due to the fact that during the initial stage of reconstruction, the spectral information of each pixel is inaccurate and difficult to partition, whereas, with the increasing number of the iterations, the reconstructed spectrum becomes increasingly accurate and exhibits a clear distribution of clusters.

Then we treat the centroid matrix $\bar{\mathbf{X}}^{(i-1)}$ as a tensor with n channels and hw samples and feed it into a convolutional modulation (CM) block [68] to capture long-range dependencies. We denote the reshaped tensor of $\bar{\mathbf{X}}^{(i-1)} \in \mathbb{R}^{hw\times n}$ as $\bar{\mathcal{X}}^{(i-1)} \in \mathbb{R}^{h\times w\times n}$. The block simplifies the self-attention

Algorithm 1: The progressive Self-Supervised Learning for CSI.

Inputs: Mask $\mathbf{\Phi}$, measurement $\mathbf{y}$, coordinates tensor $\mathbf{\Gamma}$, maximum iteration amount max_iter

Outputs: Reconstructed HSI $\mathbf{x}$

Initialization: Random initialize θ and ω, $\hat{\mathbf{x}}^{(0)} = \Phi^T\mathbf{y}$, $\bar{\mathbf{x}}^{(0)} = \hat{\mathbf{x}}^{(0)}$, $k^{(0)} = 0$

for $i = 1, 2, ..., max\_iter$ **do**

  if $i \geq 5000$ **then**

    $k^{(i)} = \begin{cases} \left\lfloor \frac{i}{\alpha} \right\rfloor, & \mod(i, \beta) \neq 0, \\ 1, & \mod(i, \beta) = 0, \end{cases}$

    using k-means algorithm on $\hat{\mathbf{x}}^{(i-1)}$ with cluster amount $k^{(i)}$ in module $\mathcal{G}$

  else

    skip k-means and let $\bar{\mathbf{x}}^{(i)} = \bar{\mathbf{x}}^{(i-1)}$ in module $\mathcal{G}$

  end if

  $\theta, \omega = \arg\min_{\theta,\omega} \|\mathbf{y} - \mathbf{\Phi}\mathcal{F}(\mathcal{G}(\hat{\mathbf{x}}^{(i-1)}; \omega), \mathbf{\Gamma};\ \theta)\|_1$

  $\hat{\mathbf{x}}^{(i)} = \mathcal{F}(\mathcal{G}(\hat{\mathbf{x}}^{(i-1)}; \omega), \mathbf{\Gamma};\ \theta)$

end for

mechanism [40] by modulating the value $\mathbf{V} \in \mathbb{R}^{h\times w\times c}$ with the convolutional attention feature $\mathbf{A} \in \mathbb{R}^{h\times w\times c}$ obtained through group convolution:

$$\begin{aligned} \mathbf{A} &= \mathrm{DConv}_{\mathrm{k\times k}}\left(\mathbf{W}_1\bar{\mathcal{X}}^{(\mathrm{i}-1)}\right), \\ \mathbf{V} &= \mathbf{W}_1\bar{\mathcal{X}}^{(i-1)}, \\ \mathbf{G}^{(i-1)} &= \mathbf{A} \odot \mathbf{V}, \end{aligned} \tag{6}$$

where $\mathrm{DConv}_{\mathrm{k\times k}}$ denotes depthwise convolution with kernel size $k \times k$, and $\odot$ denotes the Hadamard product. We use $k = 11$ in the convolutional modulation block. The CM block employs convolution for building relationships, which is a more memory-efficient alternative to self-attention architectures, particularly when dealing with high-resolution images. Additionally, the block is capable of adapting to input content through the use of a modulation operation, in contrast to the fixed structure of classic residual blocks. The CM block with trainable parameters ω is for efficient spectral cluster-centroid denoising and maps the centroid matrix $\bar{\mathbf{X}}^{(i-1)}$ to a tensor with c channels.

Finally, we combine $\mathbf{G}^{(i-1)}$ and $\mathbf{\Gamma}$ to form the input of the MSA as shown in (4). With the two branches of data flow, our network ensures that only two pixels with similar spectra and close spatial locations can have a close distance in the high-dimensional space of the result.

D. Multi-Scale Aggregation Module for Reconstruction

The aggregation module aims to reconstruct the target HSI by fusing the extracted features of the clustering centroids $\bar{\mathbf{x}}^{(i-1)}$ and the positional encoding tensor $\mathbf{\Gamma}$. Note that the purpose of the module is equivalent to that of the MLPs in traditional implicit neural networks [71].

The proposed method not only uses the positional encoding vector of spatial coordinates (u, v) as input but also includes the denoised result of the cluster-centroid for the current reconstructed spectrum. This is for an adaptive fitting to train the network $\mathcal{F}(\cdot)$ in the high-dimensional space of "spatial-centroid" with the image structural prior. Additionally, the uncertainty of the clustering in each iteration can add a certain degree of randomness to prevent the network from falling into a local optimum.

The detailed architecture of the aggregation module is shown in Fig. 4. Due to the fact that in traditional implicit neural networks, the positional-encoded tensor $\mathbf{\Gamma}$ is not processed through convolution operations, but rather the encoding vector for each coordinate is fed into an MLP to generate image intensities, in the main structure of our aggregation module, we employ a convolutional block consisting of 1×1 convolutional layer with instance normalization and GELU activation function to replace the linear layers in the MLP equivalently. This is reasonable since previous work has shown that 1×1 U-Net architecture with positional-encoding inputs is more appropriate to deal with specific vision tasks such as image superresolution, than the widely used MLP [48]. Another benefit of this approach is the significant reduction in network parameters, enabling the construction of a compact and efficient network.

Multi-scale Centroid Feature: Clustering errors may introduce noise into the reconstruction, particularly during the early iterations. To address this issue, we propose using multiscale spectral cluster-centroid features. This is because noise levels tend to be smoothed and reduced at smaller scales in the image. Additionally, coupling features across multiple scales helps to achieve higher noise impedance.

As shown in Fig. 4, we apply the features at four scales, where each downsampled scale includes a pair of downsampling blocks and upsampling blocks. Note that all the output feature maps of the four scales share the same resolution. The first input channel amount of the first block is 28, and it is subsequently mapped to $L \times s$ after s-th block. The multiscale features are concatenated with the feature maps produced by each 1×1 convolutional layer and then fed to the next 1×1 convolutional layer. The downsampling block at the s-th scale($s = 1, 2, 3$) consists of a 3×3 convolutional layer and a $2^s \times 2^s$ max-pooling layer with stride 2^s, and the upsampling block at the s-th scale magnifies the feature maps by 2^s times using bilinear interpolation. In addition, we add a CM block after the last convolutional block for further improvement with long-range dependencies.

Noise Injection: The measurements of CSI are typically subject to imaging noise, which can significantly compromise the quality of image reconstruction in the presence of overfitting. To mitigate overfitting and improve the generalization capability of our network, we adopt a noise injection method that introduces noise to the feature maps before each 1×1 convolution block as a regularizer. Specifically, Gaussian noise is used for injection, which is updated independently to produce slightly different feature maps at each iteration. By causing the feature maps to "jitter" in the feature space, the neural network is made less likely to fit the measurements too closely. As a result, the performance of the network during training exhibits a noisy trajectory, thereby ensuring generalization capability to scenes with unknown noise priors.

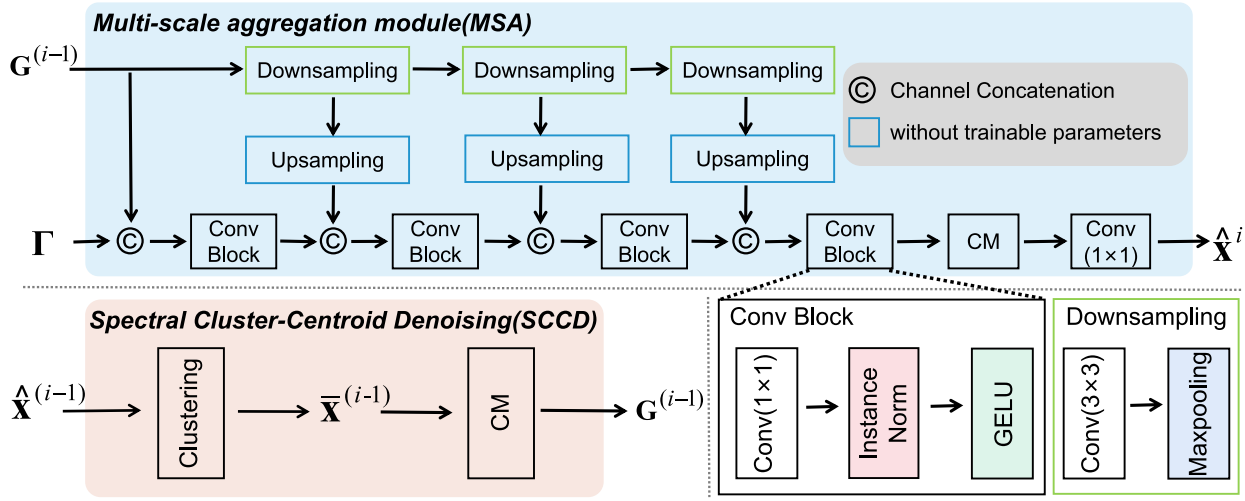

Fig. 4. The architecture of Multi-scale Aggregation(MSA) module and spectral cluster-centroid denoising(SCCD) module. Note that we use a convolutional modulation(CM) [68] block in both modules.

E. Training Details

We implement the proposed method by Pytorch and the models are trained on one GTX 1060 GPU. In the i-th iteration, the training objective is to minimize the Smooth Mean Absolute Error Loss between the reconstructed measurement $\Phi\hat{\mathbf{x}}^{(i)}$ and measurement $\mathbf{y}$ as shown in (3). Note that differing from the deep image prior methods [43] where parameters are randomly reset in each iteration, our network parameters are updated based on the previous iteration, resulting in faster computational efficiency. The models are trained with Adam optimizer, and the learning rate is set to 5×10^{-5}. We observe that $L = 64$ or larger frequencies would increase the model parameter numbers without significantly improving the reconstruction results, therefore the frequency of positional encoding embedding L is set to 32 in the whole training stage. In the SCCD, β is set to 4000 and α is set to 3000. To ensure the accuracy of clustering, we use the spectral centroid matrix obtained through the clustering algorithm until 5000 iterations. In the MSA module, the feature maps are added with Gaussian noise ($\mu = 0$ and $\sigma^2 = 0.1$) before feeding to each convolutional block. The detailed algorithm of the experiment is presented as Algorithm 1.

IV. Experiments

In this section, we compare our method with the SOTA unsupervised and supervised CSI techniques and evaluate the effectiveness of each proposed module via the ablation study.

A. Simulation HSI Reconstruction

We conduct simulations on KAIST [76] and CAVE [77]. The CAVE dataset comprises 32 hyperspectral images with a spatial resolution of 512×512, while the KAIST dataset includes 30 hyperspectral images with a spatial resolution of 2704×3376 . To evaluate the proposed method, we select ten scenes from the KAIST HSI dataset as the test set. As in previous works [27], [30], [33], [35], [37], [40], we obtain the HSIs by interpolating at the 28 wavelengths ranging from 450 nm to 650 nm with 10 nm intervals. In simulations, the 3D HSIs are cropped into patches with a spatial resolution of 256×256. The movement step size d in dispersion is set to 2. Therefore, the size of the measurements in the test set is 256×310. Table I shows the comparison of our proposed methods with four SOTA unsupervised HSI reconstruction algorithms (DIP-HSI [43],[1] HQSCI [47],[2] TV-FFDNET [44])[3] and LRSDN [46] [4] on the ten simulation scenes. In the LRSDN experiments, to ensure fairness, we conducted two experiments: one strictly follows the paper using GAP-TV [15] for initialization, and the other used $\hat{\mathbf{x}}^{(0)} = \Phi^T\mathbf{y}$ for initialization. As the source code for another self-supervised method LR2DP [45] is unavailable, here we do not include it in our comparisons. However, in Table II, we present the reconstruction errors of our method on the five images sourced from the CAVE dataset, for the purpose of comparison with the reported results of LR2DP.

In this paper, all algorithms are tested with the same settings as presented in [35], [43]. The PSNR and SSIM calculations are consistent with [33], [40]. We also present the results of spectral angle mapping (SAM) to evaluate the reconstructed spectra of each method. Our method has clearly achieved the optimal performance among various self-supervised methods in terms of evaluation metrics.

The results of our method are also competitive with those of supervised methods, and its image quality on PSNR and SSIM is very close to that of the SOTA supervised methods MST++ and BIRNET (see Fig. 1), even with much fewer trainable parameters and lighter network structure. The comparative results with supervised methods are shown in the Table III. Note that the code and the SAM of DGSM-Swin [30] are not provided.

[1]https://github.com/mengziyi64/CASSI-Self-Supervised
[2]https://github.com/XinranQin/HQSCI
[3]https://github.com/ucker/SCI-TV-FFDNet
[4]https://github.com/ChenYong1993/LRSDN

TABLE I
COMPARISONS OF RECONSTRUCTION QUALITY AND PARAMETER AMOUNT BETWEEN THE PROPOSED AND SOTA SELF-SUPERVISED ON TEN SIMULATION SCENES(S1-S10)

Methods	Venue	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg	Params
HQSCI [47]	ICASSP'22	28.54	24.27	29.50	39.44	26.53	27.78	25.62	22.46	27.62	23.97	27.57	1.02M
		0.730	0.581	0.830	0.939	0.673	0.680	0.691	0.623	0.776	0.551	0.707	
		13.24	19.57	9.65	11.99	11.61	23.54	13.36	27.19	13.98	21.67	16.58	
DIP-HSI [43]	ICCV'21	32.68	27.76	31.30	40.54	29.79	30.39	28.18	29.44	34.51	28.51	31.26	33.85M
		0.890	0.833	0.914	0.962	0.900	0.877	0.913	0.874	0.927	0.851	0.894	
		10.44	15.57	10.47	16.26	13.53	21.09	6.32	23.82	11.90	18.45	14.79	
TV-FFDNET [44]	CVPR'21	30.12	28.53	27.70	34.76	28.18	26.30	27.15	25.72	27.27	26.47	28.22	0.48M
		0.842	0.776	0.824	0.917	0.833	0.799	0.785	0.788	0.808	0.724	0.810	
		11.31	15.89	13.39	13.65	11.40	18.34	12.55	20.86	13.17	17.70	14.83	
LRSDN [46] with $\Phi^T\mathbf{y}$	TIP'24	33.39	34.22	36.63	41.74	30.29	33.30	35.52	29.39	38.70	28.59	34.18	0.23M
		0.945	0.866	0.945	0.965	0.951	0.946	0.787	0.895	0.950	0.923	0.917	
		10.42	11.29	5.18	10.45	8.00	15.24	4.24	22.70	7.61	21.04	11.60	
Proposed	-	**37.19**	**38.85**	**41.52**	**47.13**	**36.57**	34.09	**38.58**	**33.46**	**41.18**	**31.67**	**38.03**	**0.143M**
		0.941	**0.963**	**0.969**	**0.988**	**0.968**	**0.947**	**0.967**	**0.945**	**0.968**	**0.916**	**0.957**	
		5.33	**6.99**	**4.12**	**6.89**	**4.64**	**8.85**	**4.02**	**9.25**	**5.97**	**8.14**	**6.42**	

The PSNR(upper cell), SSIM(middle cell), SAM(bottom cell), and parameter amount are reported. The best results are highlighted in boldface. Note that we show two cases of LRSDN with different initialization.

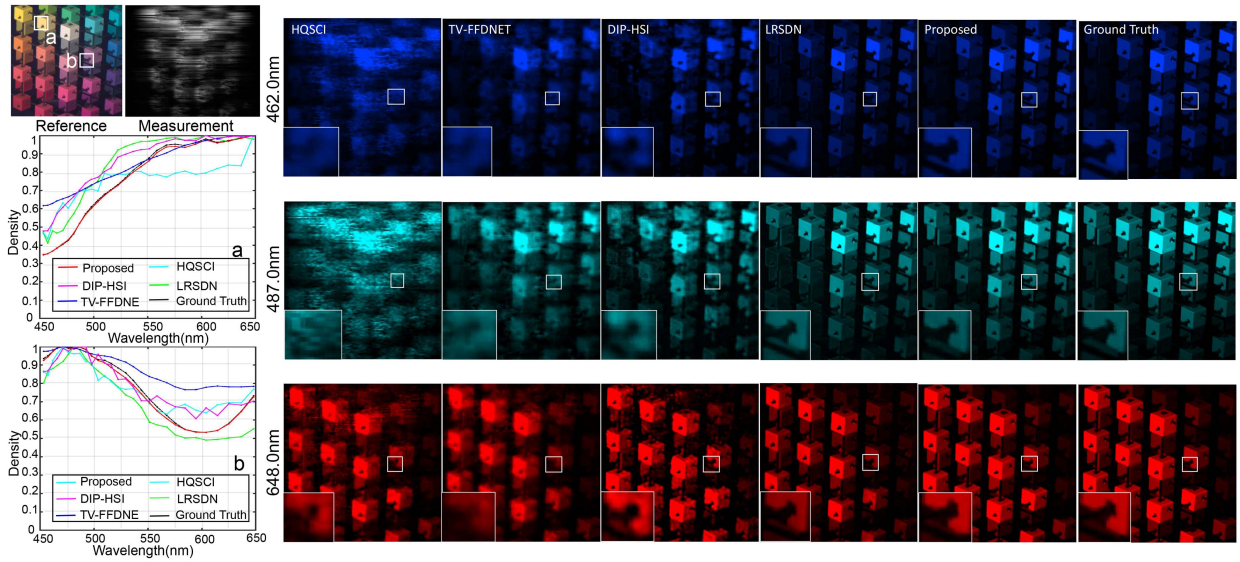

Fig. 5. Reconstructed simulation HSI comparisons of Scene 2 with 4 out of 28 spectral channels. Three SOTA self-supervised methods and our method are included. The spectral density plots (left-bottom) are corresponding to the selected white box of the reference image.

TABLE II
COMPARISONS OF RECONSTRUCTION QUALITY BETWEEN THE PROPOSED AND LR2DP ON CAVE

	Face	Toy	Clay	Egyptian	Cloth
LR2DP	39.24	33.25	42.53	40.40	29.92
	0.967	0.933	0.968	0.969	**0.848**
	9.98	10.98	**12.05**	22.06	**6.09**
Proposed	**41.53**	**33.47**	**46.87**	**41.33**	**31.32**
	0.975	**0.955**	**0.985**	**0.978**	0.766
	7.35	**10.77**	13.76	**15.06**	10.41

The PSNR(upper cell), SSIM(middle cell), SAM(bottom cell) are reported. The best results are highlighted in boldface.

We also give visual comparisons of the self-supervised methods. The simulated HSI reconstruction comparisons of Scene 2 with 3 (out of 28) spectral channels are shown in Fig. 5. As observed in the reconstruction, our method demonstrates the best visual performance on HSIs reconstruction, with fewer reconstruction artifacts, clearer boundaries, and textures compared to previous methods, particularly in reconstructing high-frequency structural content and maintaining spectral consistency. Our approach, which utilizes clustering features and positional-encoding representation, exhibits significant advantages over three methods that rely on the structure priors of CNN and handcrafted priors.

B. Real HSI Reconstruction

We further evaluate the effectiveness of our method in real HSI reconstruction. The dataset is collected by the real HSI system designed in TSA-Net [35]. Each HSI has 28 spectral channels with wavelengths ranging from 450 nm to 650 nm and has 54-pixel dispersion in the column dimension. The measurement used as input is at a spatial size of 660×714. Similar to [33], [35], we use the same real mask as [35] for HSIs reconstruction. Fig. 6 shows the visual comparisons between our method and three unsupervised SOTA methods. The reconstructed results of HQSCI exhibit noise due to the presence of measurement noise,

TABLE III
COMPARISONS OF RECONSTRUCTION QUALITY AND PARAMETER AMOUNT BETWEEN THE SOTA **SUPERVISED** METHODS AND THE PROPOSED SELF-SUPERVISED METHOD ON TEN SIMULATION SCENES(S1-S10)

Methods	Venue	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg	Params
BiSRNET [72]	NeurIPS'23	30.95 0.847 11.68	29.21 0.791 15.80	29.11 0.828 11.86	35.91 0.903 18.19	28.19 0.827 11.55	30.22 0.863 18.11	27.85 0.800 12.28	28.82 0.843 17.99	29.46 0.832 12.85	27.88 0.800 17.09	29.76 0.837 14.74	**0.06M**
HerosNet [24]	CVPR'22	35.95 0.944 8.03	35.17 0.939 11.94	35.45 0.932 8.02	40.41 0.965 12.81	33.23 0.942 8.37	34.99 0.953 12.62	34.12 0.925 7.42	33.40 0.940 13.87	33.77 0.931 9.32	33.24 0.937 12.58	34.97 0.941 10.50	1.18M
HDNet [37]	CVPR'22	35.14 0.935 7.91	35.67 0.940 9.53	36.03 0.943 6.56	42.30 0.969 10.95	32.69 0.946 6.51	34.46 0.952 9.96	33.67 0.926 6.77	32.48 0.941 11.21	34.89 0.942 8.00	32.38 0.937 10.46	34.97 0.943 8.79	2.37M
MST-L [40]	CVPR'22	35.47 0.942 7.63	36.13 0.946 9.40	36.39 0.951 6.86	41.87 0.971 12.08	32.95 0.947 6.94	34.70 0.953 10.65	34.12 0.928 6.50	32.92 0.946 12.89	35.08 0.941 8.77	32.77 0.939 11.36	35.24 0.946 9.31	2.03M
GAP-CCoT [25]	PHOTONICS RES'22	35.17 0.938 7.80	35.90 0.948 8.44	36.91 0.958 5.76	42.25 0.977 11.88	32.61 0.948 5.70	34.95 0.957 7.79	33.46 0.923 9.13	33.13 0.952 9.66	33.75 0.954 7.21	32.43 0.941 8.46	35.26 0.950 8.18	8.04M
MST++ [73]	CVPRW'22	35.80 0.943 7.34	36.23 0.947 8.93	37.34 0.957 5.61	43.63 0.973 12.25	33.38 0.952 6.19	35.38 0.957 9.81	34.35 0.934 6.17	33.71 0.953 11.22	36.67 0.953 7.38	33.38 0.945 10.43	35.99 0.951 8.53	1.33M
CST [42]	ECCV'22	35.96 0.949 7.15	36.84 0.955 8.15	38.16 0.962 5.28	42.24 0.975 11.50	33.25 0.955 6.22	35.72 0.963 9.24	34.86 0.944 5.72	34.34 0.961 9.76	36.51 0.957 7.12	33.09 0.945 9.56	36.12 0.957 7.97	3.00M
DGSM-Swin [30]	TPAMI'23	36.40 0.952 -	37.14 0.958 -	38.16 0.963 -	44.41 0.982 -	33.89 0.958 -	35.79 0.964 -	35.16 0.945 -	34.20 0.962 -	37.21 0.959 -	33.17 0.947 -	36.55 0.959 -	9.38M
BIRNET [74]	TPAMI'22	36.79 0.951 5.64	37.89 0.957 8.28	40.61 0.971 4.38	46.94 0.985 8.44	35.42 0.964 5.28	35.30 0.959 9.23	36.58 0.955 5.18	33.96 0.956 9.92	39.47 0.970 5.58	32.80 0.938 9.25	37.58 0.960 7.12	4.40M
DAUHST [33]	NeurIPS'22	37.25 0.958 5.95	39.02 0.967 7.62	41.05 0.971 4.46	46.15 0.983 10.84	35.80 0.969 5.15	37.08 0.970 9.60	37.57 0.963 4.47	35.10 0.966 10.35	40.02 0.970 6.18	34.59 0.956 9.81	38.36 0.967 7.44	6.15M
RDLUF-MixS^2 [34]	CVPR'23	37.94 0.966 4.98	40.95 0.977 6.09	43.25 0.979 3.40	47.83 0.990 6.36	37.11 0.976 4.03	37.47 0.975 6.98	38.58 0.969 4.07	35.50 0.970 7.79	41.83 0.978 4.46	35.23 0.962 6.65	39.57 0.974 5.48	1.89M
DERNN-LNLT 9STG [75]	arXiv'23	**38.49** **0.968** **4.57**	**41.27** **0.980** **5.49**	**43.97** **0.980** **3.23**	**48.61** **0.992** **5.88**	**38.29** **0.981** **3.62**	**37.81** **0.977** **6.60**	**39.40** **0.973** **3.85**	**36.51** **0.974** **7.30**	**43.38** **0.983** **4.21**	**35.61** **0.966** **6.01**	**40.33** **0.977** **5.08**	1.04M
Proposed	-	37.19 0.941 5.33	38.85 0.963 6.99	41.52 0.969 4.12	47.13 0.988 6.89	36.57 0.968 4.64	34.09 0.947 8.85	38.58 0.967 4.02	33.46 0.945 9.25	41.18 0.968 5.97	31.67 0.916 8.14	38.03 0.957 6.42	0.143M

The PSNR(upper cell), SSIM(middle cell), SAM(bottom cell)), and parameter amount are reported. The best results are highlighted in boldface.

whereas the results obtained from DIP-HSI contain water-wave-like artifacts. This is because DIP-HSI employs not only a deep image prior but also a handcrafted total-variation regularization. As shown in Figs. 6 and 7, our reconstructed results exhibit smooth and natural appearance, while maintaining significantly higher color fidelity compared to the three methods we compared against.

C. Running Time

For real HSI reconstruction, it takes approximately one to two hours, which is close to other self-supervised methods [43]. We employ a compact network for the reconstruction task, with the majority of running time in the clustering phase. Using deep efficient clustering technique might accelerate the reconstruction [78], [79].

D. Ablation

To investigate the effectiveness of each module in the proposed method, we first conduct simulations to compare our method with the naive implicit neural representation (INR) methods solely using positional encoding, and then give the performance of the proposed method with and without the attention block, SCCD module, multi-scale architecture, and noise injection.

Comparison with naive INRs: A naive 2D approach for sorely utilizing INR to reconstruct the desired HSI involves incorporating positional encoding into the 2D spatial coordinates (u, v) and feeding the resulting encoded vector $\gamma(u, v)$ into an MLP to generate the spectrum for each spatial position within the HSI. An alternative 3D approach is treating the spectrum as a separate independent dimension and employing positional encoded vectors of 3D spatial-spectral coordinates (u, v, λ) as inputs to generate intensity values for each 3D position within the HSI.

To quantify the improvements brought about by the coupled features and training strategy, we implement the INR networks that directly adopt 2D and 3D coordinate MLPs to represent HSI and train with our loss function. In the 2D INR network, the frequency of their positional encoding L is set to 64; while in the 3D INR network, the frequency of their positional encoding L is also set to 64 for each dimension. Consequently, the input dimensions for the 2D and 3D INR networks are $4 \times L + 2 = 258$ and $6 \times L + 3 = 387$ respectively.

As shown in Table IV, the proposed method outperforms the 2D and 3D INR by 4.63 dB and 5.53 dB in PSNR, even

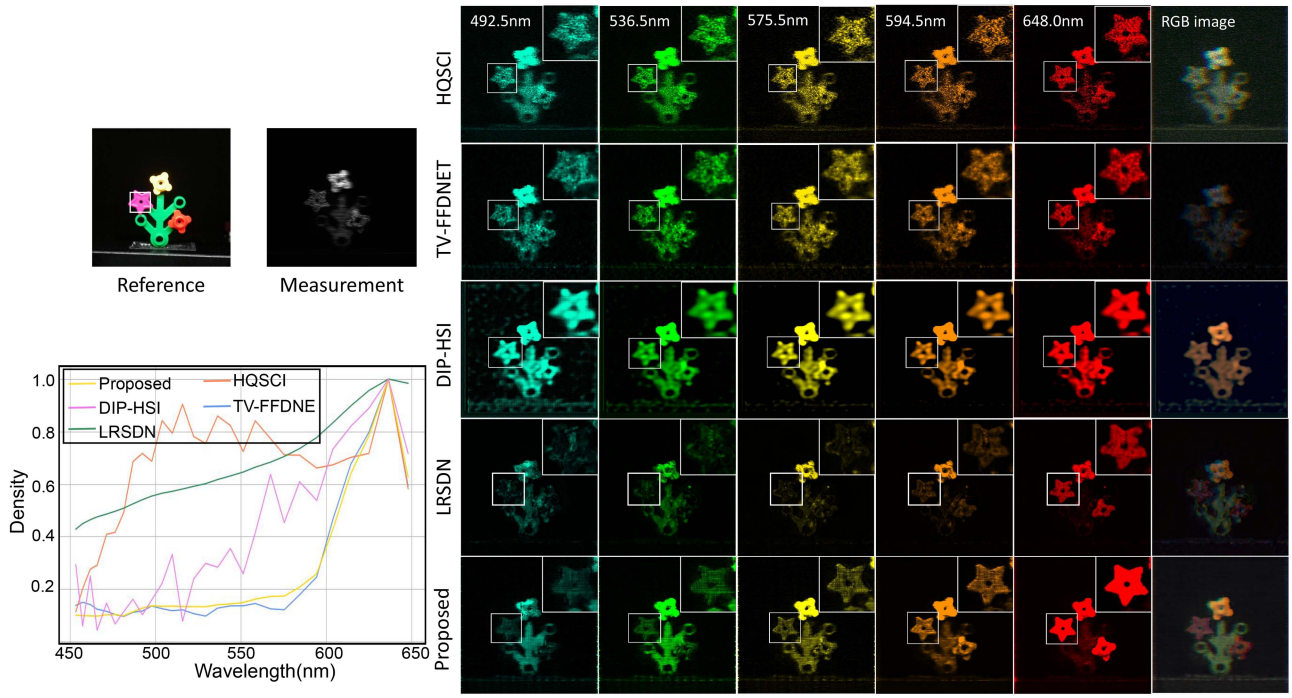

Fig. 6. The comparisons of the reconstructed HSI of the three SOTA self-supervised methods. The HSIs are reconstructed from the real measured images of Scene 1. Here 5 out of 28 spectral channels are shown. The comparison of rendered RGB images of the three self-supervised methods is also given in the last column. The spectral density plots (left-bottom) are corresponding to the selected white box of the reference image. The RGB images are rendered using CIE 1931 RGB color matching functions.

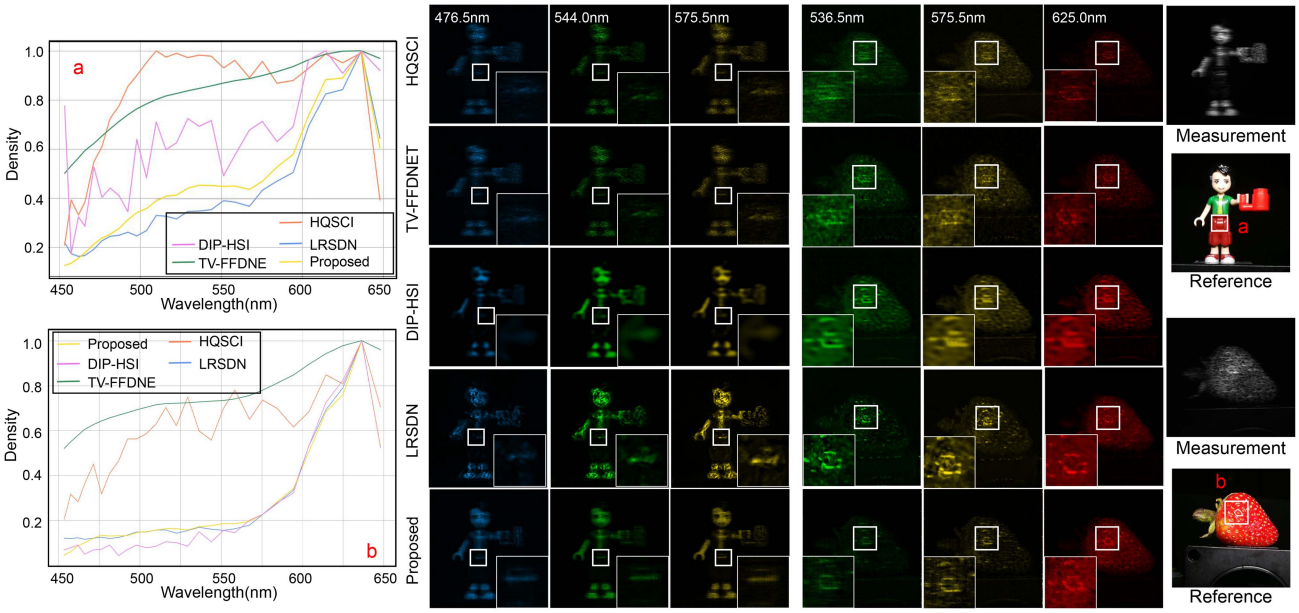

Fig. 7. The additional visual comparisons of the reconstructed HSI of the three SOTA self-supervised methods. Here 3 out of 28 spectral channels are shown.

TABLE IV
COMPARISON WITH NAIVE INRS

	PSNR	SSIM	SAM	Params
Proposed	**38.03**	**0.957**	**6.42**	**0.143M**
2D INR	33.35	0.871	11.54	0.181M
3D INR	32.45	0.848	13.12	0.222M

with fewer training parameters. Fig. 8 shows the visualized comparisons of reconstruction results adopting 2D and 3D INR on simulation scenes, respectively. While 2D and 3D imaging techniques employ differing numbers of dimensions to represent data, their reconstructed images often display significant artifacts and exhibit similarities. In contrast, our proposed method

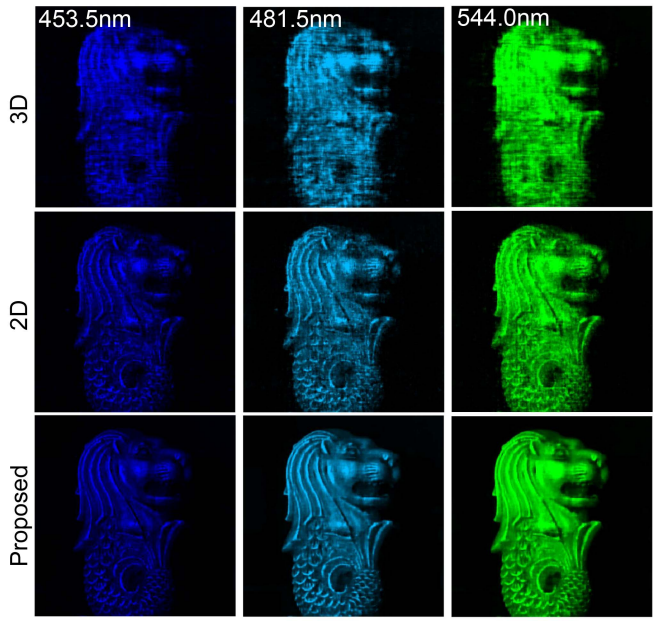

Fig. 8. Comparisons of reconstructed images of 3D and 2D naive INR methods and ours on Scene 7 and 8 with 3 (out of 28) spectra. The spectral density plots (left) are corresponding to the selected white box of the reference image.

TABLE V
COMPARISON OF VARIOUS CLUSTERING STRATEGIES

	PSNR	SSIM	SAM
w/o 1	36.25	0.932	7.88
fixing 20	36.06	0.930	7.87
ours	**38.03**	**0.957**	**6.42**

TABLE VI
ABLATION STUDY OF CM BLOCKS

	PSNR	SSIM	SAM	Params
w/t	**38.03**	**0.957**	**6.42**	0.143M
w/o	35.81	0.926	7.85	**0.132M**

is capable of producing resultant images with less noise and greater fidelity.

Comparison of various clustering strategies: The proposed method applies the clustering strategy as shown in Algorithm 1 to adjust the cluster amount in each iteration. To investigate the influence of the used cluster amount in our model, we apply two additional clustering strategies, including continuously increasing the number of clusters without jitter to 1, as well as fixing it to 20. Both two methods decrease the randomness of the model, which may lead to getting trapped in local minima. As shown in Table V, our method achieved optimal results, outperforming others by 1.73 and 1.92 in PSNR, respectively.

With and without CM blocks: We compare the performance of the network with and without the CM blocks. The utilization of CM blocks, which enlarges the receptive field of the network and takes into account spatial long-distance relationships, results in significantly improved performance compared to methods that do not incorporate them, with improvements of 2.17 dB and 0.031 in PSNR and SSIM (see Table VI), respectively. Fig. 9 shows the visualization of the feature maps with and without

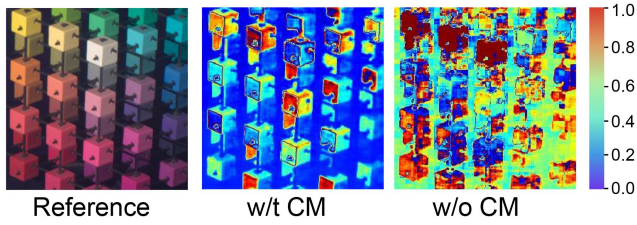

Fig. 9. The feature visualization of the last CM block in MSA. The left column shows the original reference image. The middle and right columns respectively exhibit the feature maps with and without CM. Note that the feature maps are obtained after 50,000 iterations.

TABLE VII
BREAK-DOWN ABLATION

Baseline	SCCD	MSA	PSNR	SSIM	SAM	Params
✓			32.52	0.858	12.08	0.042M
✓	✓		35.03	0.921	8.71	0.048M
✓		✓	35.09	0.917	8.37	0.136M
✓	✓	✓	**38.03**	**0.957**	**6.42**	**0.143M**

TABLE VIII
THE ABLATION OF NOISE INJECTION

	PSNR	SSIM	SAM
w/o noise injection	36.96	0.949	7.03
w/t noise injection	**38.03**	**0.957**	**6.42**

the CM blocks. In the case without CM block, we employ a single convolutional layer with instance normalization and GELU activation instead. Models that lack CM blocks generate feature maps that are noisy, whereas models that incorporate CM blocks produce feature maps with sharp edges and demonstrate resistance to noise. Therefore, the effectiveness of CM blocks in providing long-distance relationships within modules can be verified.

Break-down Ablation: We conducted a breakdown ablation analysis to investigate the impact of each component on performance. The results of the reconstruction image quality are presented in Table VII. The baseline model excludes the SCCD module and only employs Γ as input. It can be considered as an INR network followed by a CM block, but it does not encompass multi-scale features. We also evaluated the method without either the SCCD module or multi-scale architecture in the aggregation module to assess the enhancements provided by these two components. Our results indicate that both of these components contribute to an improvement of nearly 2.5 dB in PSNR, and the combination of the two components yields more than a 5.4 dB improvement. To further illustrate the effectiveness of the proposed approach, we provide a comparison of the ground truth, the reconstructed HSIs of the proposed method, and the method without the SCCD module, multi-scale features, and positional encoding tensor $\mathbf{\Gamma}$ in Fig. 10. The complete network shows the best image reconstruction quality among the compared methods, which effectively verifies the advance of the architecture in the CSI task.

Ablation of Noise Injection: To illustrate the effectiveness of our noise injection, we present the quantitative results of the reconstruction image quality in Table VIII and visualize

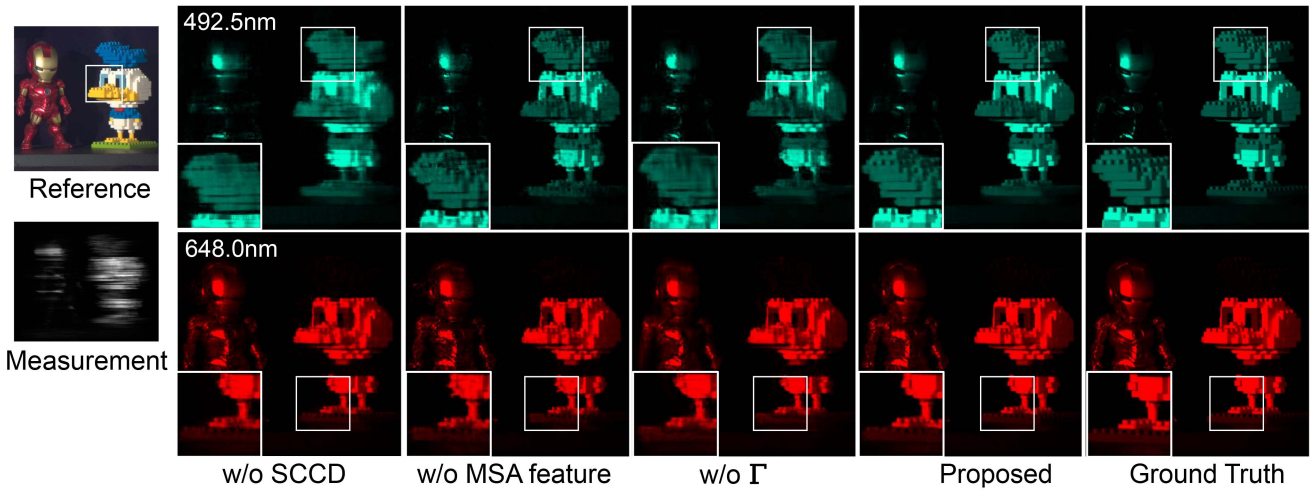

Fig. 10. Simulation HSI reconstruction results of without SCCD, without MSA feature, without positional encoding tensor $\mathbf{\Gamma}$ and the complete network on Scene 8 with 2 (out of 28) spectra.

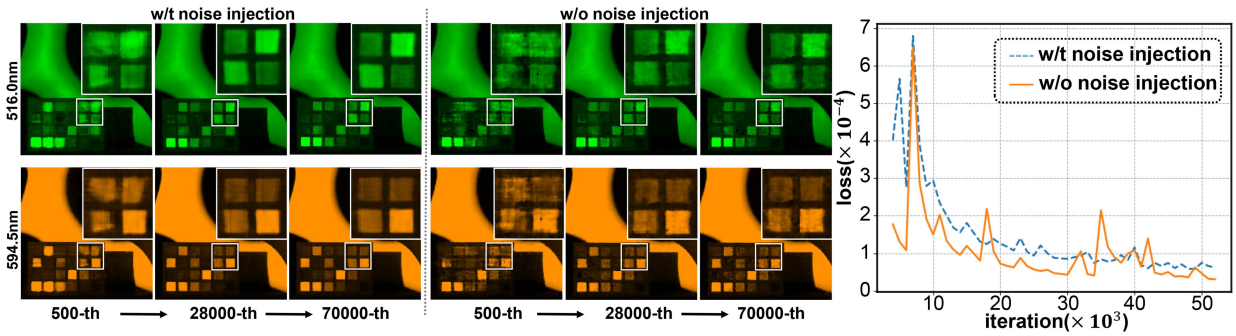

Fig. 11. The comparisons of using noise injection and no noise injection. Left: the simulated reconstructed HSIs of Scene 3 with 2 (out of 28) spectral channels. Right: Comparison of loss function curves with the varying number of iterations.

the intermediate results during training with and without noise injection. Noise injection can effectively alleviate overfitting and bring about an improvement of nearly 1 dB in PSNR and 0.01 in SSIM since it introduces some uncertainty into the training. Although the loss curve with noise injection appears higher than the one without noise injection, its loss function value remains consistently smoother. Fig. 11 displays the intermediate results obtained at iterations 5,000, 28,000, and 70,000. Notably, the implementation of noise injection techniques facilitates the prompt acquisition of the correct image structure in comparison to results obtained without noise injection.

Ablation of initial guess: To investigate how different initializations of $\hat{\mathbf{x}}^{(0)}$ influence its results, we examine the impact of various initialization techniques on the performance through a comparative analysis of three distinct initialization settings: zero-filled, random, and $\mathbf{\Phi}^T\mathbf{y}$. Each of the strategies is reconstructed for 100,000 iterations. The reconstruction using zero initialization is mainly dominated by the positional-encoding branch during the initial training phase, while multiscale centroid features contribute less to the reconstruction. Random initialization introduces some randomness into the initial guess and increases the likelihood of reconstruction. As shown in Table IX, the proposed initialization achieves the most remarkable results, surpassing the other two strategies by more than 2 dB in PSNR.

TABLE IX
THE ABLATION OF INITIAL

	PSNR	SSIM	SAM
zero-filled matrix	37.50	0.935	8.64
randomly matrix	37.46	0.935	8.25
$\mathbf{\Phi}^T\mathbf{y}$	**39.63**	**0.961**	**6.52**

V. CONCLUSION

In this paper, we introduce an effective self-supervised method for accurate single HSI reconstruction without training data. Our approach uses a progressive learning framework to aggregate hybrid features from implicit neural representation and clustering-centroid, which can capture both spatial nonlocal relationships underlying structure of HSI. Additionally, the backbone network is further we augmented with noise injection to enhance the robustness of the reconstruction. Through empirical experiments, we have demonstrated the efficacy of our proposed method by comparing it with SOTA self-supervised

and supervised methods. The proposed method is a generic framework with the potential to be applied in various domains, such as medical image reconstruction and video compression reconstruction.

REFERENCES

[1] M. Borengasser, W. S. Hungate, and R. Watkins, *Hyperspectral Remote Sensing: Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2007.

[2] Y. Fu, Y. Zheng, I. Sato, and Y. Sato, "Exploiting spectral-spatial correlation for coded hyperspectral image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3727–3736.

[3] V. Backman et al., "Detection of preinvasive cancer cells," *Nature*, vol. 406, no. 6791, pp. 35–36, 2000.

[4] Y. Li, A. Majumder, D. Lu, and M. Gopi, "Content-Independent multi-spectral display using superimposed projections," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 337–348, 2015.

[5] S. Leavesley, Y. Jiang, V. Patsekin, B. Rajwa, and J. P. Robinson, "An excitation wavelength–scanning spectral imaging system for preclinical imaging," *Rev. Sci. Instrum.*, vol. 79, no. 2, 2008, Art. no. 023707.

[6] Y. Li, C. Wang, and J. Zhao, "Locally linear embedded sparse coding for spectral reconstruction from RGB images," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 363–367, Mar. 2018.

[7] S.-H. Baek et al., "Single-shot hyperspectral-depth imaging with learned diffractive optics," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2651–2660.

[8] X. Cao et al., "Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 95–108, Sep. 2016.

[9] H. Du, X. Tong, X. Cao, and S. Lin, "A prism-based system for multispectral video acquisition," in *Proc. 2009 IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 175–182.

[10] P. Llull et al., "Coded aperture compressive temporal imaging," *Opt. Exp.*, vol. 21, no. 9, pp. 10526–10545, 2013.

[11] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, no. 10, pp. B44–B51, 2008.

[12] Z. Yu et al., "Deep learning enabled reflective coded aperture snapshot spectral imaging," *Opt. Exp.*, vol. 30, no. 26, pp. 46822–46837, 2022.

[13] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[14] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[15] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *Proc. 2016 IEEE Int. Conf. image Process.*, 2016, pp. 2539–2543.

[16] L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu, "Dual-camera design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 54, no. 4, pp. 848–858, 2015.

[17] D. Kittle, K. Choi, A. Wagadarikar, and D. J. Brady, "Multiframe image estimation for coded aperture snapshot spectral imagers," *Appl. Opt.*, vol. 49, no. 36, pp. 6824–6833, 2010.

[18] X. Lin, Y. Liu, J. Wu, and Q. Dai, "Spatial-spectral encoded compressive hyperspectral imaging," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–11, 2014.

[19] Y. Li, Q. Fu, and W. Heidrich, "Multispectral illumination estimation using deep unrolling network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2672–2681.

[20] Q. Fu, M. Souza, E. Choi, S. Shin, S.-H. Baek, and W. Heidrich, "Limitations of data-driven spectral reconstruction–an optics-aware analysis," 2024, *arXiv:2401.03835*.

[21] J. Li, Y. Li, C. Wang, X. Ye, and W. Heidrich, "BUSIfusion: Blind unsupervised single image fusion of hyperspectral and RGB images," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 94–105, 2023.

[22] Y. Liu, X. Yuan, J. Suo, D. J. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2990–3006, Dec. 2019.

[23] L. Wang, C. Sun, M. Zhang, Y. Fu, and H. Huang, "DNU: Deep non-local unrolling for computational spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1661–1671.

[24] X. Zhang, Y. Zhang, R. Xiong, Q. Sun, and J. Zhang, "Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17532–17541.

[25] L. Wang, Z. Wu, Y. Zhong, and X. Yuan, "Snapshot spectral compressive imaging reconstruction using convolution and contextual transformer," *Photon. Res.*, vol. 10, no. 8, pp. 1848–1858, 2022.

[26] L. Wang, C. Sun, Y. Fu, M. H. Kim, and H. Huang, "Hyperspectral image reconstruction using a deep spatial-spectral prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8032–8041.

[27] Z. Meng, S. Jalali, and X. Yuan, "Gap-net for snapshot compressive imaging," 2020, *arXiv:2012.08364*.

[28] J. Yang, T. Lin, F. Liu, and L. Xiao, "Learning degradation-aware deep prior for hyperspectral image reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531515.

[29] J. Ma, X.-Y. Liu, Z. Shou, and X. Yuan, "Deep tensor ADMM-net for snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10223–10232.

[30] T. Huang, X. Yuan, W. Dong, J. Wu, and G. Shi, "Deep Gaussian scale mixture prior for image reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10778–10794, Sep. 2023.

[31] J. Dong et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 37749–37761, 2022.

[32] Y. Li et al., "End-to-end video compressive sensing using anderson-accelerated unrolled networks," in *Proc. 2020 IEEE Int. Conf. Comput. Photogr.*, 2020, pp. 1–12.

[33] Y. Cai et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," 2022, *arXiv:2205.10102*.

[34] Y. Dong, D. Gao, T. Qiu, Y. Li, M. Yang, and G. Shi, "Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22262–22271.

[35] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial-spectral self-attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 187–204.

[36] Z. Meng, M. Qiao, J. Ma, Z. Yu, K. Xu, and X. Yuan, "Snapshot multispectral endomicroscopy," *Opt. Lett.*, vol. 45, no. 14, pp. 3897–3900, 2020.

[37] X. Hu et al., "Hdnet: High-resolution dual-domain learning for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17542–17551.

[38] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "L-net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4059–4069.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[40] Y. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17502–17511.

[41] J. Wang, K. Li, Y. Zhang, X. Yuan, and Z. Tao, "S˙2-transformer for mask-aware hyperspectral image reconstruction," 2022, *arXiv:2209.12075*.

[42] Y. Cai et al., "Coarse-to-fine sparse transformer for hyperspectral image reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 686–704.

[43] Z. Meng, Z. Yu, K. Xu, and X. Yuan, "Self-supervised neural networks for spectral snapshot compressive imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2622–2631.

[44] H. Qiu, Y. Wang, and D. Meng, "Effective snapshot compressive-spectral imaging via deep denoising and total variation priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9127–9136.

[45] Y. Chen, X. Gui, J. Zeng, X.-L. Zhao, and W. He, "Combining low-rank and deep plug-and-play priors for snapshot compressive imaging," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 21, 2023, doi: 10.1109/TNNLS.2023.3294262.

[46] Y. Chen, W. Lai, W. He, X.-L. Zhao, and J. Zeng, "Hyperspectral compressive snapshot reconstruction via coupled low-rank subspace representation and self-supervised deep network," *IEEE Trans. Image Process.*, vol. 33, pp. 926–941, 2024.

[47] Y. Quan, X. Qin, M. Chen, and Y. Huang, "High-quality self-supervised snapshot hyperspectral imaging," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1526–1530.

[48] N. Shabtay, E. Schwartz, and R. Giryes, "Pip: Positional-encoding image prior," 2022, *arXiv:2211.14298*.

[49] G. Littwin and L. Wolf, "Deep meta functionals for shape representation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1824–1833.

[50] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3 d representations without 3 d supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3504–3515.

[51] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5939–5948.
[52] B. Mildenhall, P.P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NERF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.
[53] S. Liu, S. Saito, W. Chen, and H. Li, "Learning to infer implicit surfaces without 3 D supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
[54] R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman, "Frequency bias in neural networks for input of non-uniform density," in *Proc. Int. Conf. Mach. Learning. PMLR*, 2020, pp. 685–694.
[55] N. Rahaman et al., "On the spectral bias of neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5301–5310.
[56] P.-S. Wang, Y. Liu, Y.-Q. Yang, and X. Tong, "Spline positional encoding for learning 3D implicit signed distance fields," 2021, *arXiv:2106.01553*.
[57] K. Park et al., "Nerfies: Deformable neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5865–5874.
[58] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNERF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4578–4587.
[59] B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman, "The convergence rate of neural networks for learned functions of different frequencies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
[60] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10318–10327.
[61] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9421–9431.
[62] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P.P. Srinivasan, "Mip-NERF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5855–5864.
[63] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5752–5761.
[64] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 264–274.
[65] A. Bietti and J. Mairal, "On the inductive bias of neural tangent kernels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
[66] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, "Towards understanding the spectral bias of deep learning," 2019, *arXiv:1912.01198*.
[67] M. Tancik et al., "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 7537–7547.
[68] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style convnet for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.
[69] M. Zechner and M. Granitzer, "Accelerating K-means on the graphics processor via CUDA," in *Proc. 2009 IEEE 1st Int. Conf. Intensive Appl. Serv.*, 2009, pp. 7–15.
[70] Y. Li, C. Wang, J. Zhao, and Q. Yuan, "Efficient spectral reconstruction using a trichromatic camera via sample optimization," *Vis. Comput.*, vol. 34, no. 12, pp. 1773–1783, 2018.
[71] L. Shen, J. Pauly, and L. Xing, "NeRP: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 770–782, Jan. 2024.
[72] Y. Cai, Y. Zheng, J. Lin, X. Yuan, Y. Zhang, and H. Wang, "Binarized spectral compressive imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024.
[73] Y. Cai et al., "Mst : Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 745–755.
[74] Z. Cheng et al., "Recurrent neural networks for snapshot compressive imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2264–2281, Feb. 2023.
[75] Y. Dong, D. Gao, Y. Li, G. Shi, and D. Liu, "Degradation estimation recurrent neural network with local and non-local priors for compressive spectral imaging," *IEEE Trans. Geosci. Remote Sens.*, 2024.
[76] I. Choi, M. Kim, D. Gutierrez, D. Jeon, and G. Nam, "High-quality hyperspectral reconstruction using a spectral prior," 2017.
[77] J.-I. Park, M.-H. Lee, M. D. Grossberg, and S. K. Nayar, "Multispectral imaging using multiplexed illumination," in *Proc. 2007 IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[78] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu, "Deep spectral clustering using dual autoencoder network," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 4066–4075.
[79] U. Shaham et al., "Spectralnet: Spectral clustering using deep neural networks," 2018, *arXiv:1801.01587*.