A New Hyperspectral Reconstruction Method With Conditional Diffusion Model for Snapshot Spectral Compressive Imaging

Yifan Si, Zijian Lin, Xiaodong Wang, and Sailing He, *Fellow, IEEE*

***Abstract*— In the coded aperture snapshot spectral imaging (CASSI) system, the coded and compressed single-channel measurements need to be reconstructed into hyperspectral cubes. Existing discriminative models reconstruct the spectral cube by optimizing the mean squared error (MSE) between the ground truth and the predicted image, employing peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as metrics to gauge the quality of reconstruction. However, these indicators often possess significant limitations in mimicking human visual perception and in discerning the impact of image distortions on perceived visual quality. In this article, a new model named CASSIDiff is proposed to reconstruct CASSI measurements, achieving advanced results in perceptual loss-based evaluation metrics such as learned perceptual image patch similarity (LPIPS) and Fréchet inception distance (FID). The diffusion model, which enjoys high accuracy and reliability in generative tasks, is used for the first time for the hyperspectral reconstruction task. A feature fusion mechanism based on discrete wavelet transform (DWT) is used to weaken the noise interference effect in the conditional diffusion model. Considering the interspectra similarity and long-range dependencies of hyperspectral data, the spatial–spectral attention mechanism is also introduced. Experiments show that CASSIDiff not only outperforms most existing algorithms in simulation datasets but also shows robustness to real data published and collected in our home-built CASSI system. The code and models are publicly available at: https://github.com/YifanSi/CASSIDiff.**

***Index Terms*— Coded aperture snapshot spectral imaging (CASSI), conditional diffusion model, Fréchet inception distance (FID), hyperspectral reconstruction, learned perceptual image patch similarity (LPIPS).**

Received 31 December 2024; accepted 21 January 2025. Date of publication 14 March 2025; date of current version 2 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant W2412107; in part by the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang under Grant 2025C02140, Grant 2024C03045, Grant 2023C03083, and Grant 2023C03135; in part by the National Key Research and Development Program of China under Grant 2022YFB2804100, Grant 2022YFC2010003, and Grant 2022YFC3601003; in part by Ningbo Science and Technology Project under Grant 2024Z146; and in part by Ningbo Public Welfare Research Program Project under Grant 2024Z234. The Associate Editor coordinating the review process was Dr. Maryam Shamgholi. *(Corresponding author: Sailing He.)*

Yifan Si is with the Centre for Optical and Electromagnetic Research, National Engineering Research Center for Optical Instruments, Zhejiang Provincial Key Laboratory for Sensing Technologies, Zhejiang University, Hangzhou 310058, China.

Zijian Lin is with the Centre for Optical and Electromagnetic Research, National Engineering Research Center for Optical Instruments, Zhejiang Provincial Key Laboratory for Sensing Technologies, Zhejiang University, Hangzhou 310058, China, and also with Shanghai Institute for Advanced Study, Zhejiang University, Shanghai 201203, China.

Xiaodong Wang is with the School of Engineering, Westlake University, Hangzhou 310030, China.

Sailing He is with the Centre for Optical and Electromagnetic Research, National Engineering Research Center for Optical Instruments, Zhejiang Provincial Key Laboratory for Sensing Technologies, Zhejiang University, Hangzhou 310058, China, and also with the Department of Electromagnetic Engineering, School of Electrical Engineering, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: sailing@kth.se).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIM.2025.3551465, provided by the authors.

Digital Object Identifier 10.1109/TIM.2025.3551465

I. Introduction

HYPERSPECTRAL imaging is a technique with high-spectral resolution and wide wavelength coverage [1], [2], [3], [4], [5], [6], [7], [8]. Unlike single-channel grayscale images and three-channel RGB images, the number of channels of hyperspectral images is between dozens and thousands, and its spectral resolution can reach several nanometers [9], [10]. Based on the above characteristics, the hyperspectral imaging has a wide range of applications, including art identification [11], crop health [12], coastline mapping [13], forests [14], mineral exploration [15], urban and industrial infrastructure [16], production line product quality [17], environmental monitoring [18], and more.

Hyperspectral imaging principles can be roughly divided into four categories: grating spectroscopy [19], prism spectroscopy [20], tunable filter spectroscopy [21], and chip coating [22]. Most of the traditional imaging methods, such as spot scan [23], [24], [25] or line scan [26], [27], are time-consuming, which makes it difficult to acquire high-precision hyperspectral images of moving objects. The emergency of snapshot compressive imaging (SCI) system [28], [29], [30] has solved this problem. Examples include coded aperture snapshot spectral imaging (CASSI) system [31], [32], [33] and computational tomography imaging system (CTIS) [34], [35], [36]. They are spatially and spectrally modulated and compressed by different coding devices and then reconstructed to hyperspectral data by corresponding algorithms.

In the current snapshot-based compression imaging system, the CASSI technology is the most advanced and has the best imaging quality, which has become the mainstream research direction and promoted the development of related reconstruction algorithms. Some traditional model-based algorithms are based on hand-crafted regularization terms, such as total variation prior terms [37], [38], [39] and low-rank prior terms [40], [41], [42]. These algorithms suffer from poor reconstruction

quality and long optimization times. With the rapid development of deep learning, convolutional neural networks [43], [44], [45] are capable of end-to-end reconstruction and can directly restore 2-D compressed encoded data into 3-D hyperspectral data. The article [46] designed a deep convolutional network to learn Gaussian mixed scale priors and proved that it can learn more spatial–spectral correlations than hand-crafted prior. The reconstruction accuracy of CNN-based correlation network [47], [48], [49], [50] has made a qualitative leap compared with traditional algorithms, but it lacks effective solutions in the face of problems such as nonlocal correlations and long-range dependencies, and it is difficult to correlate features of different receptive fields. Relevant studies [51], [52], [53] applied the attention mechanism in natural language processing to the hyperspectral reconstruction algorithm to extract and correlate features of spatial dimension and spectral dimension, respectively. The TSA-Net [52] uses self-attention mechanisms in three directions of x, y, and z for hyperspectral feature vectors, and jointly models spatial and spectral correlations in a sequentially independent manner. MST [53] not only uses an attention mechanism for the spectral direction but also introduces real mask information to make up for the missing spatial features. However, the above algorithms still have great room for improvement in both reconstruction accuracy and quality.

All of the aforementioned algorithms aim to achieve higher peak signal-to-noise ratio (PSNR) [54] and structural similarity index (SSIM) [55] scores by minimizing the mean squared error (MSE) [56]; yet, these metrics have been recognized for their significant limitations across various scenarios [57], [58], [59], [60], [61], [62], [63]. PSNR primarily assesses pixel-level errors, quantifying image quality through the ratio of peak signal power to noise power. However, it falls short in accurately emulating the human visual system's perception of fine details and complex visual scenes, and it lacks sensitivity to subtle content changes within images [64]. SSIM improves upon PSNR by incorporating considerations of luminance, contrast, and structural information, thereby providing a more nuanced simulation of human visual perception. Nonetheless, it may still fail to fully encapsulate the spectrum of subjective human responses to image quality [65]. In addition, both indices might struggle to discern the impact of diverse image distortions—such as those caused by compression, noise, or blurring—on the perceived visual quality [57], [66]. Consequently, the quality of reconstruction is not solely contingent on pixel-level errors; the fidelity to human visual perception is equally indispensable and should not be overlooked. Nonetheless, there is a dearth of research dedicated to hyperspectral reconstruction tasks within the CASSI system.

Generative models, particularly those evaluated through perceptual loss, outperform discriminative models, yet they are underutilized in hyperspectral reconstruction tasks. Our work pioneers the application of a conditional denoising diffusion probabilistic model (DDPM) for this purpose, achieving advanced results in perceptual loss metrics (learned perceptual image patch similarity (LPIPS) [57] and Fréchet inception distance (FID) [67]). Fig. 1 illustrates our conditional DDPM, with detailed mathematical foundations provided in the Appendix. The reconstruction of hyperspectral data from CASSI system measurements is inherently challenging due to the ill-posed nature of deriving a 3-D cube from a 2-D image, compounded by optical aberrations and noise. Unlike super-resolution algorithms that focus on spatial dimension expansion, hyperspectral reconstruction must also recover spectral information. We introduce a novel approach by integrating encoded 2-D measurements into the diffusion process to guide the generation of accurate hyperspectral data.

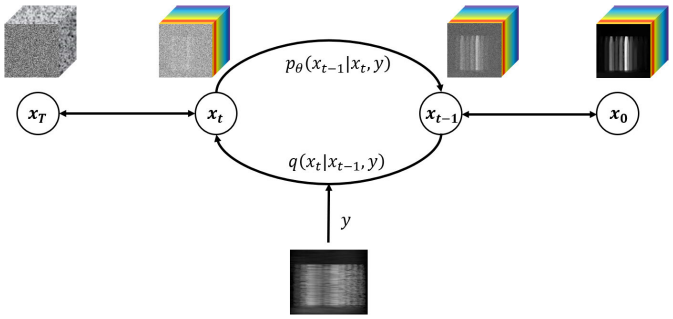


Fig. 1. Schematic of the conditional diffusion model. Also, the diffusion process is from right to left, and the reverse process is from left to right. θ in $\mathbf{p}_\theta$ is a learnable parameter, in other words, the parameter of conditional noise predictor.

Considering the spectral similarity and long-range dependencies in hyperspectral data, we have designed a spatial–spectral attention mechanism capable of capturing feature connections across different scales. However, the diffusion process's Gaussian noise often obscures spatial features with high-frequency noise, diminishing the effectiveness of direct measurement integration. To address this, we employ a wavelet transform-based structure that better fits high-frequency signals such as edges and object details, enabling the learning of both explicit and implicit feature information. Leveraging the diffusion model theory and the unique characteristics of hyperspectral data, we present CASSIDiff, a generative network that reconstructs high-fidelity hyperspectral cubes from CASSI-generated measurements. The main contributions of this article are as follows.

1) The proposed CASSIDiff applies a conditional diffusion model to a CASSI optical system for hyperspectral reconstruction tasks, demonstrating advanced reconstruction performance in perceptual loss-based evaluation metrics. We analyze the limitations inherent in the current evaluation frameworks for hyperspectral reconstruction quality and introduce the assessment method that aligns more closely with human visual perception for the first time. Experimental results have also shown that our proposed CASSIDiff has significant advantages in related matrics (LPIPS and FID).
2) A feature fusion mechanism based on wavelet transform is introduced to effectively alleviate the noise interference problem in the generation model. Also, a spatial–spectral attention mechanism is designed to deal with the spectral similarity and long-range dependence of hyperspectral data.
3) Employing our home-built CASSI system, an empirical dataset has crafted and published. It comprises

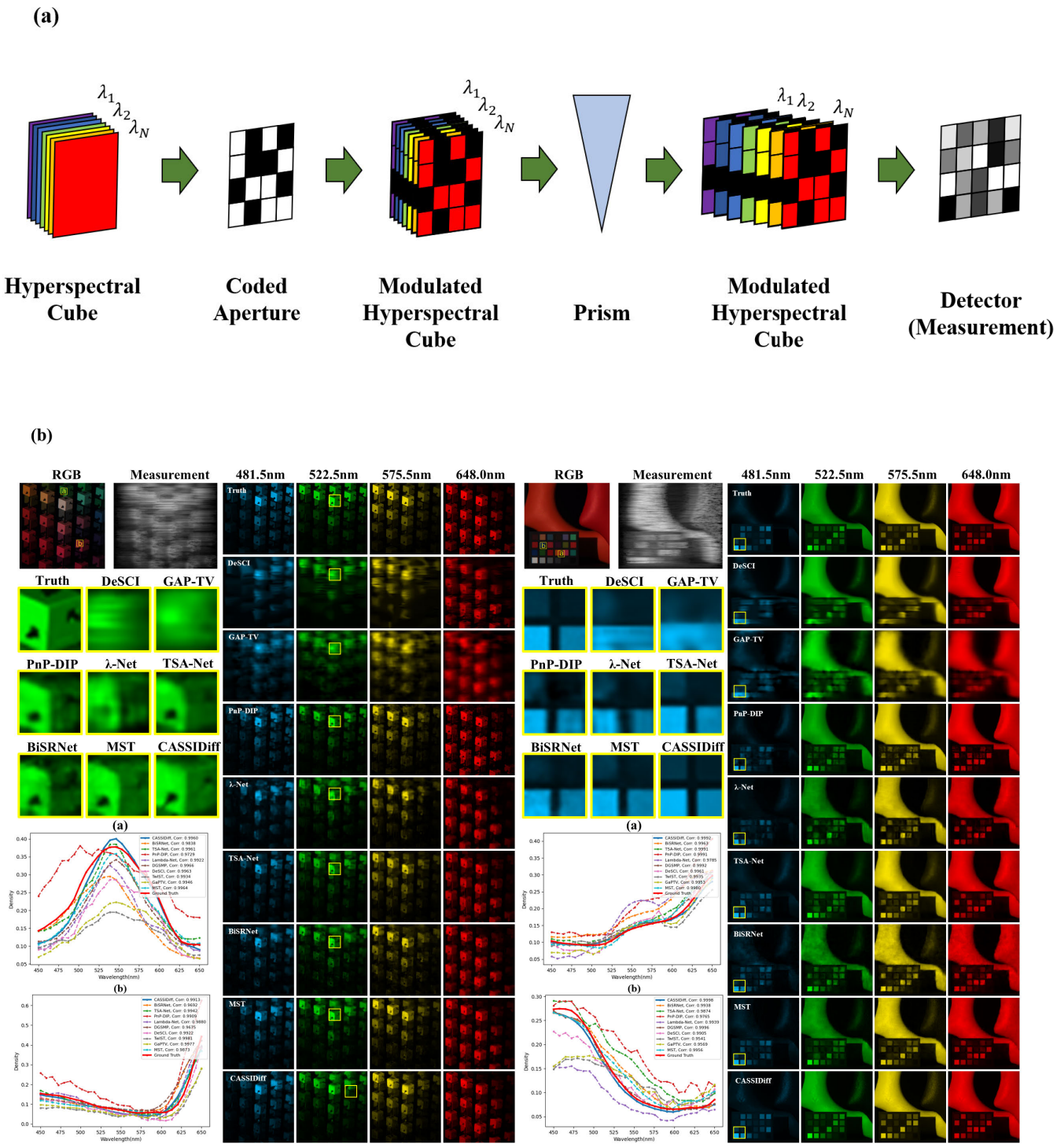


Fig. 2. (a) Means coding principle of the CASSI system. (b) Means the reconstructed pseudo-color images of the single wavelength and spectral curves in typical regions are shown. Two scenes in simulated data are included in the results. Zoomed-in-view for better view.

measurements from five different scenes, each with a resolution of 728 × 512 pixels, covering a spectral range from 450 to 650 nm. Concurrently, we have utilized a QE PRO spectrometer from Ocean Optics to acquire spectral curve data for specific regions, serving as a reference. This dataset is invaluable for validating the algorithm's robustness across various scenarios.

II. Method

A. CASSI Optical System

The principle of the CASSI system is shown in Fig. 2(a). On the left is the hyperspectral cube, denoted by $F \in \mathbb{R}^{H\times W\times N_\lambda}$, where H and W are the height and width of the hyperspectral cube, respectively, and N_λ is the number of channels. The encoding mask is recorded as $M^* \in \mathbb{R}^{H\times W}$,

where H and W are the height and width, respectively, which are equal to the spatial size. Its function is to modulate the hyperspectral cube space. The mathematical expression is as follows:

$$F'(:,:,n_\lambda) = F(:,:,n_\lambda) \bigotimes M^* \tag{1}$$

where F' represents the hyperspectral cube after spatial modulation, $n_\lambda \in [1, \ldots, N_\lambda]$ represents the spectral channel, and $\bigotimes$ represents the matrix dot product. For modulation of the spectral dimension, prisms are added in the optical path. After passing through the prism, F' is shifted along the y-axis. $F'' \in \mathbb{R}^{H\times(W+d(N_\lambda-1))\times N_\lambda}$ represents the shifted hyperspectral cube, where d represents the shifted length. Let λ_c be the reference wavelength; that is, $F''(:,:,n_{\lambda_c})$ does not shift along the y-axis after passing through the prism. Then, the spectral modulation of the prism is expressed as follows:

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

where (u, v) represents the plane coordinates of the detector, λ_n represents the wavelength of the n_λth channel of the hyperspectral cube, and $d(\lambda_n - \lambda_c)$ represents the spatial offset of the n_λth channel along the y-axis. Finally, the cube modulated by space-spectrum is received by the detector and is expressed as follows:

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

where $Y \in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ represents the compressed coded measurement received by the detector and $G \in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$ represents the noise added. Solving F–Y is a typical ill-posed problem. We use the conditional diffusion model to reconstruct F and denoising.

B. Structure of CASSIDiff

Similar to standard DDPM, we also estimate the error through a designed U-Net structure. We use the encoded measurement as a prior to guide the diffusion model to reconstruct the hyperspectral cube. The mathematical expression is as follows:

$$\epsilon_\theta(x_t, M, t) = D\left(\left(E_t^M + E_t^x, t\right), t\right) \tag{4}$$

where E_t^x is the feature map of the current timestep and E_t^M is the conditional measurement. After feature extraction and timestep embedding, the feature map is input into the decoder D for reconstruction. The overall structure of U-Net is shown in Fig. 3.

In other tasks based on conditional diffusion models such as super-resolution, the prior image is usually input as an invariant constant during feature embedding. The reason is that the low-resolution image still restores complete spatial information, which is not easily disturbed during the diffusion process. However, the measurement E_t^M encoded and compressed by the CASSI system has been destroyed and reorganized in the spatial and spectral dimensions. Therefore, directly embedding it to the feature map generated during the diffusion process will have a very poor reconstruction effect. Inspired by the work [68], we adopt a dynamic encoding method to merge hierarchically at the feature level. The effect is that different levels of coding layers can, respectively, extract low-level contour information, high-level detail feature information, and spectral information in different areas, which can enhance the features at the current step in the diffusion process and accelerate retrieve of information. The specific method is to input E_t^x and E_t^x into two encoders, respectively. One of the encoders is part of U-Net in the diffusion model, and their feature extraction layers are composed of similar blocks like ResNet. The feature maps m_k^M and m_k^x with equal size obtained after passing through the kth layer, respectively, will be fused through calculation which is similar to the attention mechanism. The mathematical expression is as follows:

$$\mathcal{A}\left(m_k^x, m_k^M\right) = \left(\mathrm{LN}\left(m_k^x\right) \bigotimes \mathrm{LN}\left(m_k^M\right)\right) \bigotimes m_k^M \tag{5}$$

where LN represents the layer normalization and $\bigotimes$ represents the matrix dot product. Two feature maps with the same size obtained by the diffusion process and prior information, respectively, are multiplied after layer normalization, and then obtain a learnable attention map, which has the ability to learn the weight distribution of features in the spatial and spectral dimensions. Finally, the attention map and m_k^M do a matrix dot product to guide the diffusion process by prior information.

C. DWT-Based Attention Mechanism

There is a phenomenon in the experiment that directly fusing the feature map in the diffusion process with the coded measurement will result in a lot of noise in the generated results, causing the object to be covered by noise. The reason is that during the fusion process, E_t^x will introduce high-frequency noise that cannot be eliminated later. In this article [68], the FF-Paser module based on the Fourier transform is used to filter high-frequency noise. The feature map can be converted from spatial domain to frequency domain through the Fourier transform, so that the model can learn the feature information covered by the noise and indirectly filter out the high-frequency noise. The wavelet transform [69], [70], [71] can express local features in the time domain (or spatial domain), and also has better expression ability in the face of mutation signals. Therefore, we designed an attention mechanism based on the discrete wavelet transform (DWT), which not only ensures high-quality reconstruction effects but also has a faster convergence speed during training. The specific structure is shown in Fig. 4.

A feature map in spatial domain $m \in \mathbb{R}^{H\times W\times C}$ becomes $\mathbb{C}^{(H/2)\times(W/2)\times 4C}$ after 2-D DWT, which is transformed into frequency domain with size changing simultaneously. Each channel represents the low-frequency and high-frequency feature maps in horizontal and vertical directions, respectively. Let $\mathcal{W}[\cdot]$ be the 2-D DWT, then M can be expressed as follows:

$$M = \mathcal{W}[m] \in \mathbb{C}^{\frac{H}{2}\times\frac{W}{2}\times 4C}. \tag{6}$$

For the feature map in frequency domain, we introduce a learnable attention map $A \in \mathbb{C}^{(H/2)\times(W/2)\times 4C}$, which can automatically focus on the details in the frequency domain

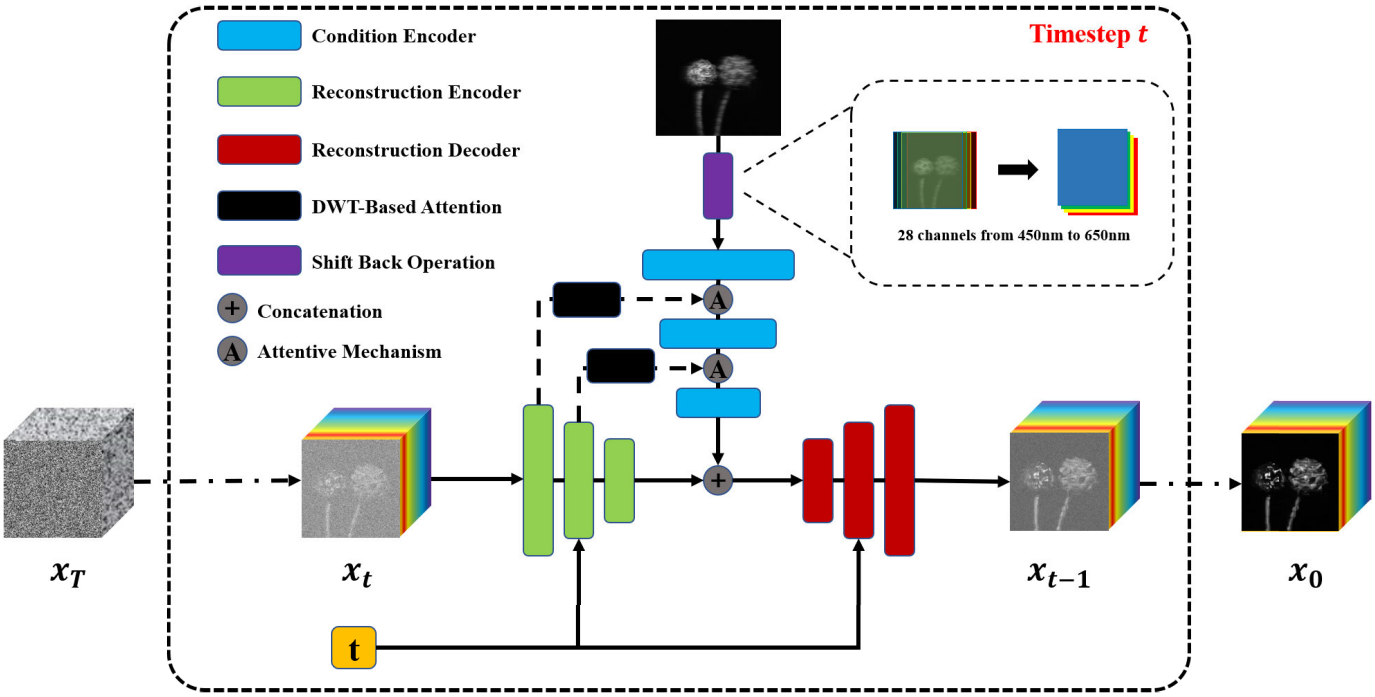


Fig. 3. Illustration of CASSIDiff. The principle of serial spatial–spectral attention mechanism and DWT-based attention mechanism in encoder and decoder are given in the following sections.

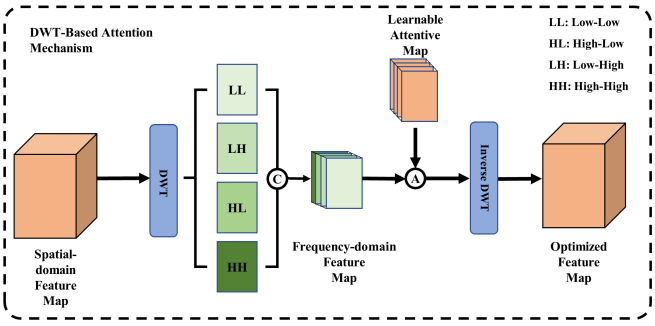


Fig. 4. Illustration of DWT-based attention mechanism, where "C" stands for concatenation operations in the channel dimension, and "A" stands for matrix dot product operation. The notations "LL," "HL," "LH," and "HH" correspond to the low–low, high–low, low–high, and high–high frequency subbands, respectively, capturing different aspects of the image's texture and details.

through training. The new feature map M' can be expressed as follows:

$$M' = A \otimes M \tag{7}$$

where $\otimes$ represents the matrix dot product. The feature map M' is transformed from frequency domain to spatial domain through 2-D discrete wavelet inverse transform. The final feature map $m' \in \mathbb{R}^{H\times W\times C}$ can be expressed as follows:

$$m' = \mathcal{W}^{-1}\left[M'\right] \tag{8}$$

where $\mathcal{W}^{-1}$ represents the 2-D discrete wavelet inverse transform. The attention mechanism based on DWT can distinguish different information in both spatial domain and frequency domain, effectively characterizes global features while capturing local details, and filters out most noise.

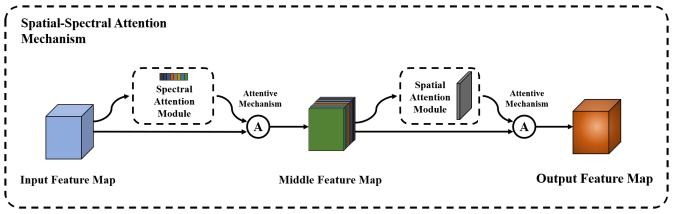


Fig. 5. Illustration of serial spatial–spectral attention mechanism. The input feature mapping is performed by attention operation in spectral and spatial dimensions, respectively, to extract the intrinsic correlation information.

D. Serial Spatial–Spectral Attention Mechanism

Different from tasks such as super-resolution and semantic segmentation based on RGB images, dual reconstruction to measurements collected by the CASSI system in spectral and spatial dimensions requires rethinking the design of the feature extraction module. The attention mechanism module in the original U-Net network cannot fit well to the spatial sparsity and spectral density of the hyperspectral cube. The channel attention mechanism proposed in SEnet [72] can not only reduce the amount of calculation and enhance the expression ability of the model but also has a strong learning ability for the weights between different wavelengths (or channels). Therefore, we designed a serial spatial–spectral attention mechanism for related characteristics. The specific structure is shown in Fig. 5.

For the feature map $m \in \mathbb{R}^{H\times W\times C}$, the channel feature $\mathcal{Z}_c \in \mathbb{R}^{1\times 1\times C}$ is first obtained after global pooling. This is a process of global spatial information compression, and its mathematical expression is as follows:

$$\mathcal{Z}_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} m(i,j). \tag{9}$$

Algorithm 1 Training

Input: Measurement and ground truth pairs $P = \left\{\left(x_M^k, x_{gt}^k\right)\right\}_{k=1}^K$, the total number of diffusion steps T

Initialize: get shifted measurement $x_s = \mathcal{S}(x_M)$ by shift back operation $\mathcal{S}$ and randomly initialize conditional noise predictor $\in_\theta$

repeat

  Sample $\left(x_s, x_{gt}\right) \sim P$

  Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $t \sim Uniform(\{1, \ldots, T\})$

  Take a gradient descent step on $\nabla_\theta \|\epsilon - \epsilon_\theta(x_t, x_s, t)\|$, where $x_t = \sqrt{\bar{\alpha}_t} x_{gt} + \sqrt{1-\bar{\alpha}_t}\epsilon$

until converged

Algorithm 2 Inference

Input: Measurement x, the total number of diffusion steps T

Initialize: get shifted measurement $x_s = \mathcal{S}(x_M)$ by shift back operation $\mathcal{S}$

Load: conditional noise predictor $\in_\theta$

Sample $x_T \sim \mathcal{N}(0, \mathbf{I})$

for $t = T, T-1, \ldots, 1$ **do**

  Sample $\mathcal{Z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathcal{Z} = 0$

  Compute x_{t-1} according to equation (12): $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{1-\bar{\alpha}_t}\epsilon_\theta(x_t, x_s, t)\right) + \widetilde{\beta}_t \ddagger = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{1-\bar{\alpha}_t}\hat{\epsilon}\right) + \widetilde{\beta}_t \ddagger$

end for

return x_0 as hyperspectral cube construction result

After further information compression and nonlinear mapping through the fully connected layer and activation function layer, the obtained weights are multiplied by the original feature map. The mathematical expression is as follows:

$$m' = \mathcal{F}[\mathcal{Z}_c] \odot m \tag{10}$$

where $\mathcal{F}[\cdot]$ represents the linear and nonlinear operations, such as full connection and activation function, and $\odot$ represents channel multiplication. The feature map $m' \in \mathbb{R}^{H\times W\times C}$ can focus on more effective information through channel weights and reconstruct the spectrum. After completing the channel (or spectral)-based attention mechanism, we implemented a spatial attention mechanism on the feature map m'. The mathematical expression is as follows:

$$m'' = \mathcal{F}'\left[\text{softmax}\left(\frac{q \cdot k^T}{\sqrt{d}}\right) \cdot v\right] \tag{11}$$

where $q, k, v \in \mathbb{R}^{H\times W\times C'}$ are three tokens. d is the spatial scale parameter, and $\mathcal{F}'[\cdot]$ is operations such as convolution and layer normalization, with the purpose of channel number matching and normalization.

The serial spatial–spectral attention mechanism not only allows the model to focus on important feature details of space and spectrum but also reduces the network parameters and computational cost. Experiments show that this mechanism can effectively handle the hyperspectral reconstruction task based on the CASSI system.

E. Algorithms

The hyperspectral reconstruction of the CASSI system based on the conditional diffusion model is divided into two stages: supervised training and inference, as shown in Algorithms 1 and 2.

In the training stage, the preprocessed ground truth with 28 channels and the corresponding simulation-generated single-channel measurements (which are subsequently restored to 28 channels through a shift-back operation) are used as the training set. Before training, the parameters of conditional noise predictor $\in_\theta$ and the sampling timestep T are initialized. In each loop during training, we randomly extract image pairs with the number of minibatch from the training set, calculate x_t and feed it to conditional noise predictor together with current t after embedding to obtain the prediction noise. We sample noise from the standard Gaussian distribution and t from the total timesteps. The noise predictor is then optimized via (19) in the Supplementary Material.

In the inference stage, the measurement becomes the input x_s after shift back, and the same sampling timestep as in the training stage is determined simultaneously. Inference begins with t and sampling a latent variable x_T from the standard Gaussian distribution. In each loop of the inference, t will be reduced by 1. x_s, current x_t, and t are fed to conditional noise predictor, and x_{t-1} is obtained. When $t = 1$, the last loop is completed and x_0 is obtained, which is the final reconstruction result.

III. Results

A. Experimental Setup

Referring to the design of paper [53], we reconstructed hyperspectral cubes with a total of 28 channels in the spectral distribution from 450 to 650 nm. Two public hyperspectral datasets: CAVE and KAIST are used for our experiments. The CAVE dataset contains 32 images with a size of 512 × 512, and each image contains hyperspectral information of different objects. The KAIST dataset consists of 30 images with a size of 2704 × 3376. Each image contains hyperspectral information of different real scenes. The experiments verified

the reconstruction effect of simulated data and real data (two separate experiments).

In the stage of verifying the simulated data, we use the CAVE dataset as a training set, which is randomly cropped into the size of 256 × 256 and fed to the model after feature enhancement. The simulated data (test set) are encoded measurements obtained by selecting ten photographs from the KAIST dataset and propagating them in the CASSI system through simulation. The same operations are used in the stage of verifying the real data, and the training sets at this time are CAVE and KAIST datasets, with a crop size of 660 × 660, to match the data size of the real data. The real data (testing set) use the dataset disclosed in the literature [52], which is collected by their SD-CASSI system. During the training process, the batch size is set to 8, and a total of 500k iterations are performed. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the learning rate is set to 1×10^{-4}. For the hyperparameters of the diffusion model, the sampling timestep is set to 1000, and the beta schedule is linear. The loss function during the training process is shown in (19) in the Supplementary Material. The entire training and testing process is based on the Pytorch platform and implemented on seven Nvidia 4090Ti graphics cards with 24 GB of memory.

B. Metrics in Simulated Data and Real Data

To forge an evaluation system that closely mirrors human visual perception, we have adopted LPIPS [57] and FID [67] as the metrics for assessing hyperspectral reconstruction quality. LPIPS, a deep learning-based perceptual model, offers a more precise assessment of image visual quality by capturing the intricate features that resonate with the human visual system. Also, FID gauges the distribution variance between the synthesized and real images within the feature space, effectively detecting discrepancies in both global and local image features. The mathematical expressions of LPIPS and FID are as follows:

$$\mathrm{LPIPS}\left(I_{\mathrm{recon}}, I_{\mathrm{gt}}\right) = \sum_{l=1}^{L} w_l \cdot d\left(F_l(I_{\mathrm{recon}}), F_l\left(I_{\mathrm{gt}}\right)\right) \tag{12}$$

$$\mathrm{FID} = \left\|\mu_{\mathrm{recon}} - \mu_{\mathrm{gt}}\right\|^2 + \mathrm{Tr}\left(\Sigma_{\mathrm{recon}} + \Sigma_{\mathrm{gt}} - 2\sqrt{\Sigma_{\mathrm{recon}}\Sigma_{\mathrm{gt}}}\right). \tag{13}$$

They utilized pretrained AlexNet and InceptionV3 models to extract feature maps from the images, respectively. In the LPIPS formula, I_{recon} and I_{gt} denote the reconstructed image and the ground truth, respectively. L signifies the total number of network layers, w_l is the weight assigned to the Lth layer, F represents the AlexNet model, and d symbolizes the distance metric used to measure the divergence between feature maps. In the FID formula, μ_{recon}, μ_{gt}, Σ_{recon}, and Σ_{gt} are the mean vectors and covariance matrices of the feature distributions for the reconstructed image and the ground truth, respectively. For each hyperspectral cube, we compute the values channel by channel and then take the average to serve as the final metric.

We compared the results of GAP-TV [82], TwIST [83], DeSCI [73], Lambda-net [75], BBCU [76], HSSP [84], TSA-Net [52], PnP-DIP [74], IRNet [76], BTM [77], ReActNet [79], DNU [85] and BiSRNet [80], PnP-CASSI [86], MST [53], DAUHST [80], PADUT [81], and CASSIDiff. As the data shown in Table I, for the ten scenes in the test set, our algorithm has excellent reconstruction results, and the average LPIPS and FID in all scenes reached 0.097 and 0.292, respectively, surpassing most of the algorithms in the control group. These metrics demonstrate that CASSIDiff is capable of reconstructing images that align more closely with human visual perception, and the features of the generated hyperspectral cubes are more akin to those of the ground truths.

The provided table offers a detailed comparison of reconstruction accuracy, measured by LPIPS and FID, for various advanced algorithms, including our proposed CASSIDiff, against several state-of-the-art (SOTA) methods such as MST, DAUHST, and PADUT. The results indicate that CASSIDiff exhibits a slight gap when compared to these SOTA algorithms, particularly in terms of LPIPS and FID scores across different scenarios of simulated data.

The minor differences in performance can be attributed to the computational constraints faced during the training of the diffusion model, which forms the backbone of CASSIDiff. Despite these limitations, CASSIDiff's performance is commendably close to that of MST, DAUHST, and PADUT, suggesting that there is significant room for improvement. This gap also highlights the potential for enhancing CASSIDiff's capabilities with more robust computational resources and further model optimizations. Moreover, the comparison with the current SOTA algorithms not only validates the effectiveness of CASSIDiff but also underscores its advancement in the field of hyperspectral image reconstruction. CASSIDiff's competitive performance against established algorithms indicates that it is a viable and promising approach, particularly when considering its ability to achieve high accuracy with the given constraints.

To establish the excellence of our algorithm in pixel fidelity, we have carried out comparative experiments utilizing PSNR and SSIM metrics. The findings are presented in Table II. The presented table provides a detailed comparison of reconstruction accuracy in terms of PSNR and SSIM for various algorithms, including CASSIDiff and current SOTA methods, such as MST, DAUHST, and PADUT. CASSIDiff, while showing some differences in PSNR and SSIM scores compared with these SOTA algorithms, demonstrates that the generative model it employs prioritizes the reconstruction of the entire image canvas rather than pixel-by-pixel optimization.

The reason for the gap between CASSIDiff and SOTA algorithms in PSNR and SSIM scores is rooted in the fundamental approach of the generative model. CASSIDiff's model is designed to focus on the overall repainting of the image, which may not align perfectly with the pixel-centric evaluation metrics of PSNR and SSIM. This approach, however, does not imply inferior reconstruction quality. On the contrary, it suggests that CASSIDiff is more adept at capturing the holistic features and spectral details of the hyperspectral data, which is crucial for accurate scene representation.

It should be highlighted that in the context of intricate textures and patterns, metrics such as PSNR and SSIM might not fully account for all the visual nuances and subtle

TABLE I

RECONSTRUCTION ACCURACY (LPIPS AND FID) OF DIFFERENT ADVANCED ALGORITHMS ARE COMPARED WITH TEN SCENARIOS OF SIMULATED DATA

Algorithm \ Scene	DeSCI [73]		PnP-DIP [74]		TSA-Net [52]		Lambda-net [75]		IRNet [76]		BTM [77]		BBCU [78]		ReActNet [79]		BiSRNet [80]		MST [53]		DAUHST [80]		PADUT [81]		CASSIDiff	
	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID
1	0.340	2.865	0.148	0.278	0.126	1.137	0.227	0.972	0.339	0.268	0.283	0.335	0.361	1.052	0.311	0.373	0.164	0.166	0.055	0.307	0.056	0.251	0.049	0.447	**0.088**	**0.104**
2	0.485	2.904	0.216	1.725	0.211	0.979	0.367	0.811	0.364	0.530	0.359	0.491	0.391	1.529	0.363	0.438	0.207	0.092	0.081	0.099	0.096	0.060	0.043	0.206	**0.133**	**0.268**
3	0.237	0.600	0.106	0.908	0.106	0.219	0.155	0.197	0.395	4.085	0.346	3.334	0.403	6.273	0.376	4.033	0.170	0.965	0.048	0.036	0.055	0.071	0.041	0.179	**0.059**	**0.084**
4	0.094	0.120	0.069	0.044	0.052	0.045	0.075	0.206	0.275	4.794	0.261	3.888	0.334	7.449	0.279	4.482	0.095	0.526	0.026	0.023	0.028	0.023	0.026	0.035	**0.040**	**0.068**
5	0.350	1.304	0.147	0.992	0.106	0.150	0.242	0.205	0.378	2.295	0.323	1.956	0.383	3.532	0.352	1.848	0.153	0.411	0.054	0.019	0.082	0.045	0.036	0.137	**0.087**	**0.032**
6	0.364	1.675	0.158	0.372	0.079	0.435	0.257	0.313	0.305	1.255	0.286	1.089	0.331	2.493	0.301	1.323	0.115	0.114	0.037	0.022	0.036	0.035	0.028	0.081	**0.095**	**0.203**
7	0.389	1.803	0.136	3.568	0.103	0.480	0.275	0.396	0.387	1.728	0.318	1.364	0.416	3.132	0.356	1.162	0.174	0.579	0.046	0.097	0.061	0.152	0.033	0.151	**0.085**	**0.096**
8	0.386	2.137	0.165	0.397	0.119	0.627	0.259	0.574	0.287	0.766	0.292	0.800	0.314	1.853	0.279	0.927	0.118	0.114	0.051	0.054	0.061	0.102	0.034	0.130	**0.140**	**0.588**
9	0.312	1.555	0.085	0.649	0.146	0.504	0.246	0.476	0.357	2.795	0.350	2.091	0.397	4.500	0.360	2.510	0.207	0.501	0.068	0.043	0.073	0.061	0.039	0.127	**0.091**	**0.155**
10	0.423	5.427	0.159	0.389	0.140	1.422	0.321	2.311	0.387	0.626	0.350	0.688	0.417	1.329	0.409	0.789	0.166	0.458	0.040	0.145	0.038	0.189	0.029	0.297	**0.150**	**1.320**
Avg	0.338	2.039	0.139	0.932	0.119	0.500	0.242	0.646	0.347	1.914	0.317	1.604	0.375	3.314	0.339	1.789	0.157	0.393	0.051	0.085	0.059	0.099	0.036	0.179	**0.097**	**0.292**

TABLE II

RECONSTRUCTION ACCURACY (PSNR AND SSIM) OF DIFFERENT ADVANCED ALGORITHMS ARE COMPARED WITH TEN SCENARIOS OF SIMULATED DATA

Algorithm \ Scene	GAP-TV [82]		TwIST [83]		PnP-DIP [74]		BiSRNet [80]		TSA-Net [52]		HSSP [84]		BTM [77]		Lambda-net [75]		DNU [85]		PnP-CASSI [86]		MST [53]		DAUHST [80]		PADUT [81]		CASSIDiff	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	26.82	0.754	25.16	0.700	31.98	0.862	30.95	0.847	31.26	0.887	31.07	0.852	28.75	0.739	31.16	0.861	31.72	0.863	29.09	0.799	34.71	0.930	33.26	0.915	36.25	0.951	**31.54**	**0.893**
2	22.89	0.610	23.02	0.604	26.57	0.767	29.21	0.791	26.88	0.855	26.30	0.798	26.91	0.674	29.13	0.795	31.13	0.846	28.05	0.708	34.45	0.925	32.09	0.898	37.92	0.963	**31.42**	**0.877**
3	26.31	0.802	21.40	0.711	30.37	0.862	29.11	0.828	30.03	0.921	29.00	0.875	26.14	0.708	31.80	0.898	29.99	0.845	30.15	0.850	35.52	0.943	33.06	0.925	39.63	0.970	**33.07**	**0.933**
4	30.65	0.852	30.19	0.851	38.71	0.930	35.91	0.903	39.90	0.964	38.24	0.926	32.74	0.794	39.08	0.944	35.34	0.908	39.17	0.939	41.50	0.967	40.54	0.964	44.55	0.985	**38.93**	**0.972**
5	23.64	0.703	21.41	0.635	29.09	0.849	28.19	0.827	28.89	0.878	27.98	0.827	25.87	0.692	28.14	0.843	29.03	0.833	27.45	0.798	31.90	0.933	28.86	0.882	34.59	0.964	**28.92**	**0.896**
6	21.85	0.663	20.95	0.644	29.85	0.848	30.22	0.863	31.30	0.895	29.16	0.823	27.37	0.739	28.48	0.846	30.87	0.887	26.16	0.752	33.85	0.943	33.08	0.937	35.58	0.965	**30.10**	**0.913**
7	23.76	0.688	22.20	0.643	27.69	0.864	27.85	0.800	25.16	0.887	24.11	0.851	26.26	0.707	28.54	0.834	28.99	0.839	26.92	0.736	32.69	0.911	30.74	0.886	35.69	0.950	**29.58**	**0.874**
8	21.98	0.655	21.82	0.650	28.96	0.843	28.82	0.843	29.69	0.8887	27.94	0.831	26.20	0.718	27.24	0.832	30.13	0.855	24.92	0.710	31.69	0.933	31.55	0.923	33.76	0.960	**27.60**	**0.888**
9	22.63	0.682	22.42	0.690	33.55	0.881	29.46	0.832	30.03	0.903	29.14	0.822	26.10	0.717	30.56	0.857	31.03	0.876	27.99	0.752	34.67	0.939	31.66	0.911	38.26	0.963	**30.88**	**0.900**
10	23.10	0.584	22.67	0.569	28.05	0.833	27.88	0.800	28.32	0.848	26.44	0.740	25.73	0.671	26.87	0.759	29.14	0.849	25.58	0.664	31.82	0.926	31.44	0.925	33.24	0.947	**27.91**	**0.831**
Avg	24.36	0.669	23.12	0.669	30.48	0.854	29.76	0.837	30.15	0.893	28.93	0.834	27.21	0.716	30.10	0.847	30.74	0.863	28.55	0.771	34.26	0935	32.63	0.917	36.95	0.962	**31.00**	**0.898**

distortions present in an image. As such, introducing human visual perception-based indicators such as LPIPS and FID provides a more holistic quality evaluation of the reconstructed hyperspectral cubes.

When the detector is used in a low-light environment, photon shot noise is the main source of noise. Meng et al. [52] proved that introducing the noise injection model [55] during training can significantly improve the reconstruction quality, so we injected 11-bit shot noise when restoring the real data. DeSCI [73], GAP-TV [82], PnP-DIP [74], Lambda-Net [75], TSA-Net [52], BiSRNet [80], MST [53], and CASSIDiff are selected to show and contrast the restoration of spatial and spectral features in detail. Reconstructed pseudo-color images and spectral curves of simulated and real data are shown in Figs. 2(b) and 6, respectively, from which we see low-frequency contours and uniformly colored areas can be reconstructed. Compared with the control group, the conditional diffusion model adopted in CASSIDiff also has significant advantages for high-frequency features with rich texture information.

C. Metrics in Our Home-Built System

The CASSI system we built is shown in Fig. 7. The high-spectral object is first relayed to the coded aperture plane by an imaging lens. The coded object is then imaged through a 4f system to the camera plane. While in-between the objective lens (Thorlabs, TL4X-SAP) and tube lens (Thorlabs, AC254-100-A), a prism (Edmund) is used to disperse multiple wavelengths of the high-spectral object in a lateral plane. Within the single exposure time of the camera (HIKROBOT, MV-CU013-A0UM), coded objects from shifted wavelengths are collapsed into one single image, creating a compressed measurement. Two bandpass filters (Thorlabs, FELH0450, and FESH0650) are mounted in front of the camera to ensure the spectral wavelength range is between 450 and 650 nm.

To verify the portability of the algorithm, we used this system to collect five samples and reconstructed their hyperspectral cubes through CASSIDiff. One of the reconstruction results are shown in Fig. 7(b)–(d). The results show that CASSIDiff can work well in different systems and has reliable portability and robustness.

D. Ablation Study of DWT-Based Module and Serial Spatial–Spectral Attention Module

To verify the effect of the DWT-based module and the serial spatial–spectral attention module in CASSIDiff, we designed an ablation experiment using simulated data to quantify their differences through the same metrics. Table III shows the experimental results when only the DWT-based module and the series attention mechanism module are retained,

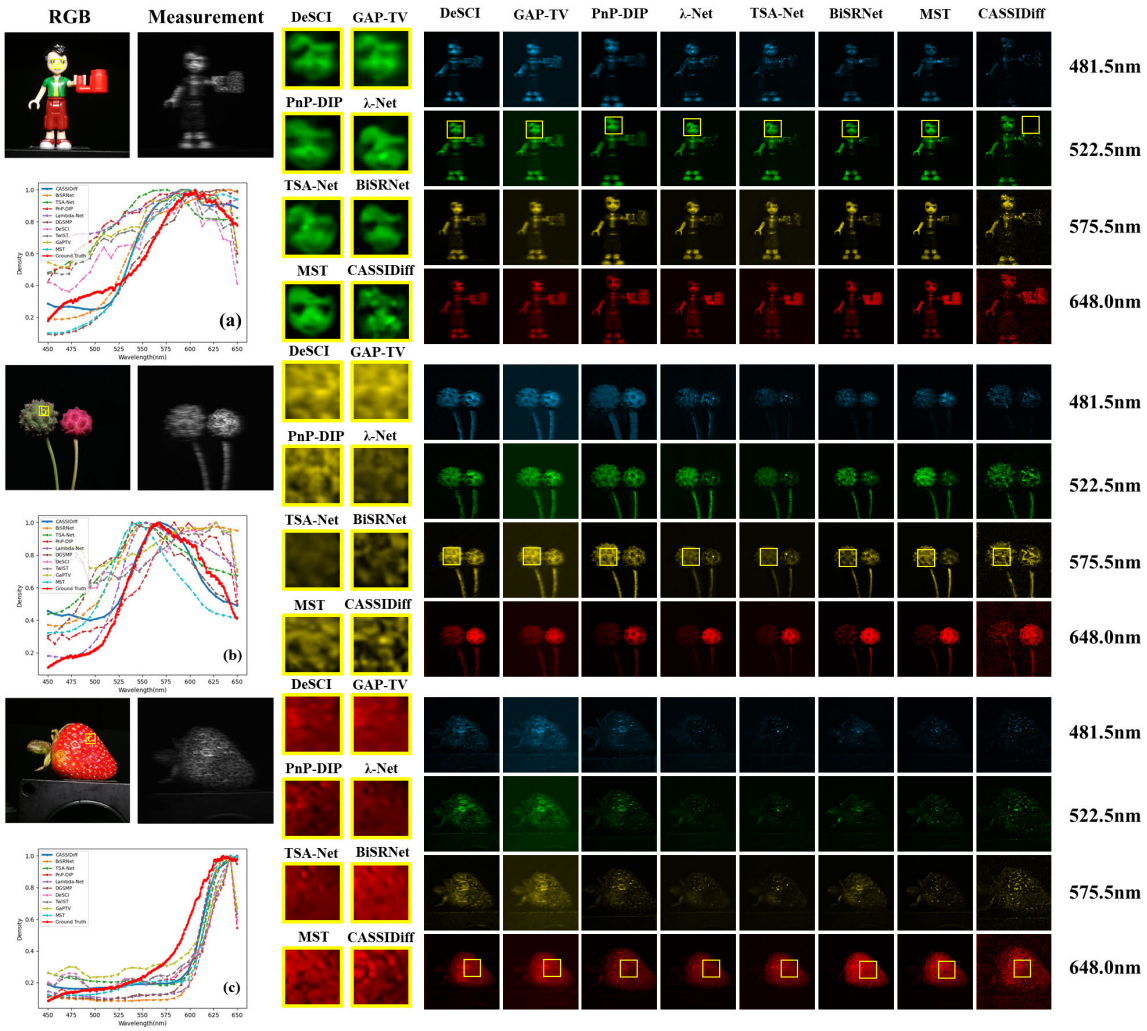


Fig. 6. Reconstructed pseudo-color images of the single wavelength and spectral curves in typical regions are shown. Three scenes in real data are included in the results. g-in-view for better view.

respectively, and other conditions are completely consistent. When the serial spatial–spectral attention module is removed, both the LPIPS and FID indicators are significantly increased, by 0.016 and 0.302, respectively. When the DWT-based module is removed, the degree of variation in LPIPS and FID is not obvious, but the decline curve of the loss function is more oscillatory during training, and it takes a longer time to converge. When faced with the hyperspectral reconstruction task based on the CASSI system, the DWT-based module and the serial spatial–spectral attention module we introduced can not only improve the reconstruction accuracy but also enhance the stability of training and accelerate the convergence speed.

We also designed related experiments to conduct preliminary exploration on how the sampling timestep in the diffusion model affects the reconstruction accuracy in the CASSI system. With other conditions unchanged, we set the sampling timestep to 500, 1000, 1500, and 2000, respectively, and reconstructed the simulation data after training. Table IV shows the reconstruction accuracy in four cases, also using LPIPS-FID and PSNR-SSIM as metrics. There is no positive correlation between the sampling timestep and reconstruction accuracy, which is consistent with the conclusions in other literature. Also, we also randomly selected areas with rich feature information, averaged them in the spatial dimension, and used cosine similarity to compare the differences with the ground truth, as shown in Fig. 8. In a comprehensive comparison, the reconstruction effect is better when the sampling timestep is set to 1000, so we will maintain this hyperparameter setting in all experiments.

E. Computational Efficiency Analysis

In our analysis of computational efficiency, we have evaluated the floating point operations (FLOPs) of our proposed method, CASSIDiff, against several SOTA reconstruction algorithms, including DAUHST [80], MST [53], DGSMP [46], BIRNAT [88], HDNet [89], and our own CASSIDiff. The comparison includes a measure of computational complexity, FLOPs, which is crucial for understanding the practicality and efficiency of each method, especially in resource-constrained environments.

Our findings, as presented in Table V, reveal that CASSIDiff exhibits a competitive FLOPs count of 153.76 giga-FLOPs

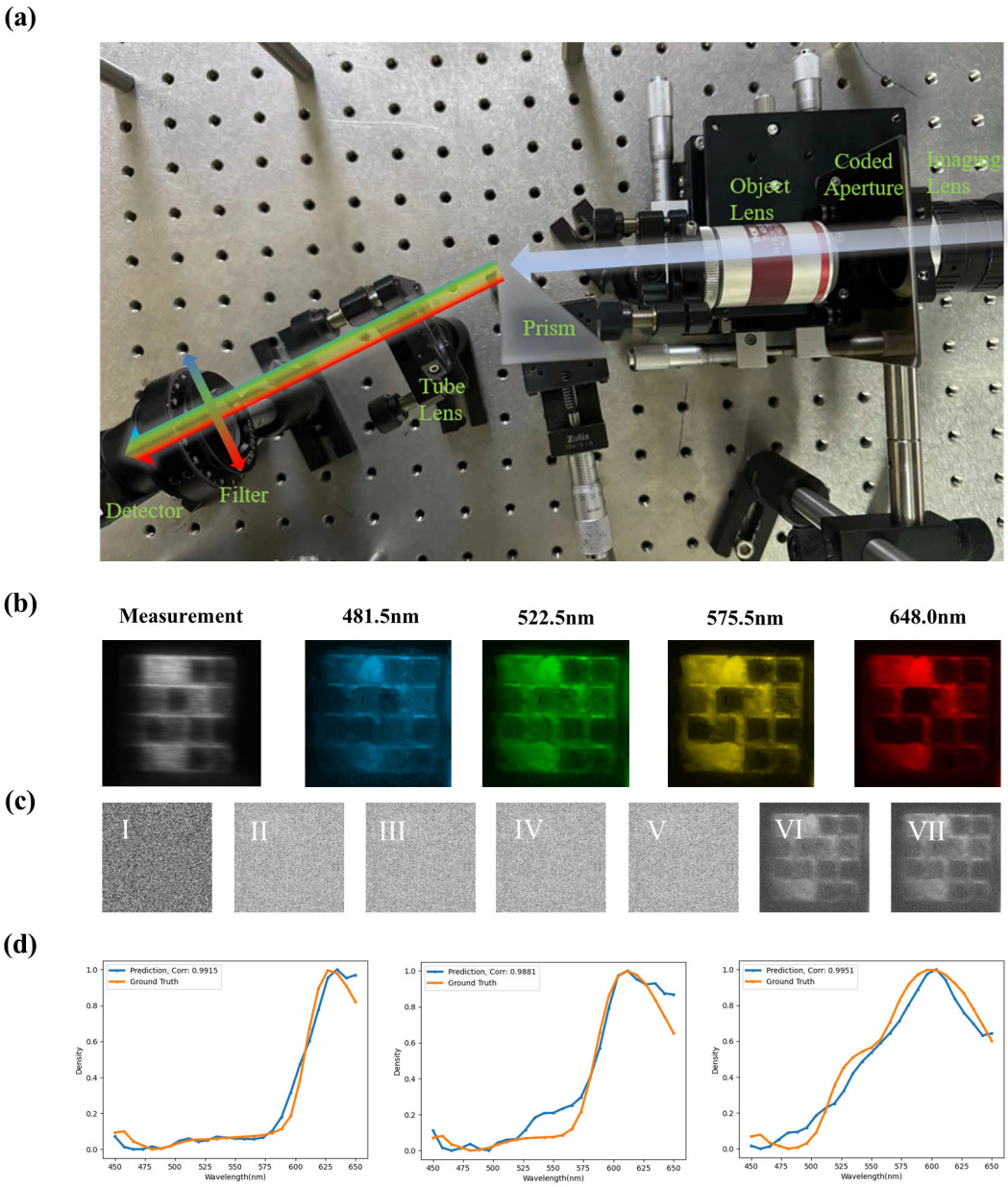


Fig. 7. (a) Schematic of the CASSI system we built. When polychromatic light is incident, the prism disperses the light. The double arrow represents the direction of dispersion and the large arrow represents the direction of light propagation. (b)–(d) Reconstruction results of CASSIDiff in the CASSI system we built ourselves. They are consistent with the meanings represented in Figs. 6, 9, and 8, respectively, and the relevant settings remain unchanged. The ground truth in (d) is measured with a spectrometer (QE PRO, Ocean Optics).

TABLE III

ABLATION STUDY ON DWT-BASED TRANSFORM MODULE AND SPATIAL–SPECTRAL ATTENTION MECHANISM. LPIPS, FID, PSNR, AND SSIM ARE USED AS THE METRICS

DWT-based Transform	Serial Spatial-Spectral Attention	LPIPS	FID	PSNR	SSIM
	✓	0.105	**0.181**	30.56	0.890
✓		0.116	0.302	30.13	0.881
✓	✓	**0.097**	0.292	**31.00**	**0.898**

(G), which is significantly lower than that of BIRNAT (2122.66 G) and DGSMP (646.65 G), and comparable with HDNet (154.76 G) and MST (28.15 G). This indicates that CASSIDiff is more computationally efficient than BIRNAT and DGSMP while maintaining a similar operational intensity to HDNet and MST. The relatively lower FLOPs count of CASSIDiff can be attributed to the strategic integration of the DWT-based attention mechanism and serial spatial–spectral attention mechanism. These mechanisms not only enhance the model's ability to capture features from the data but also improve computational efficiency by reducing unnecessary computations.

TABLE IV
ABLATION STUDY ON SAMPLING TIMESTEP. LPIPS, FID, PSNR, AND SSIM ARE USED AS THE METRICS

Scene \ Timestep	500				1000				1500				2000			
	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM
1	0.106	0.104	30.72	0.873	**0.088**	0.104	**31.54**	**0.893**	0.137	**0.709**	31.36	0.884	0.107	0.081	30.52	0.863
2	0.145	0.163	30.65	0.860	**0.133**	0.268	**31.42**	**0.877**	0.275	0.903	30.24	0.856	0.151	**0.101**	31.00	0.863
3	0.082	0.084	33.22	0.926	**0.059**	0.084	33.07	**0.933**	0.084	0.163	32.07	0.914	0.062	**0.043**	**33.27**	0.932
4	0.047	**0.034**	36.42	0.966	**0.040**	0.068	**38.93**	**0.972**	0.067	0.061	36.34	0.961	0.046	0.053	38.02	0.970
5	0.144	0.335	28.83	0.893	**0.087**	**0.032**	28.92	0.896	0.136	0.155	28.70	0.886	0.133	0.245	**29.11**	**0.900**
6	0.108	0.254	29.30	0.899	**0.095**	0.203	**30.10**	**0.913**	0.176	0.693	29.28	0.897	0.101	**0.090**	29.61	0.905
7	0.134	0.359	28.50	0.841	**0.085**	**0.096**	**29.58**	**0.874**	0.132	0.430	28.36	0.831	0.121	0.196	28.57	0.853
8	**0.131**	0.550	**27.62**	0.883	0.140	0.588	27.60	**0.888**	0.164	0.836	27.55	0.879	0.136	**0.234**	27.61	0.882
9	0.106	**0.062**	30.13	0.877	**0.091**	0.155	**30.88**	**0.900**	0.204	0.572	29.39	0.868	0.125	**0.062**	29.39	0.870
10	0.163	1.182	27.22	0.812	0.150	1.320	**27.91**	**0.831**	0.200	2.284	27.55	0.823	**0.149**	**0.555**	27.52	0.825
Avg	0.117	0.313	30.26	0.883	**0.097**	0.292	**31.00**	**0.898**	0.158	0.681	30.09	0.880	0.113	**0.166**	30.46	0.886

TABLE V
COMPARATIVE ANALYSIS OF FLOPS (G) AND PARAMS (M) FOR CASSIDIFF AND OTHER SOTA RECONSTRUCTION ALGORITHMS

Algorithms	DAUHST [80]	MST [53]	DGSMP [46]	BIRNAT [88]	HDNet [89]	CASSIDiff
FLOPs(G)	79.50	28.15	646.65	2122.66	154.76	153.76
Params(M)	6.15	2.03	3.76	4.40	2.37	103.91

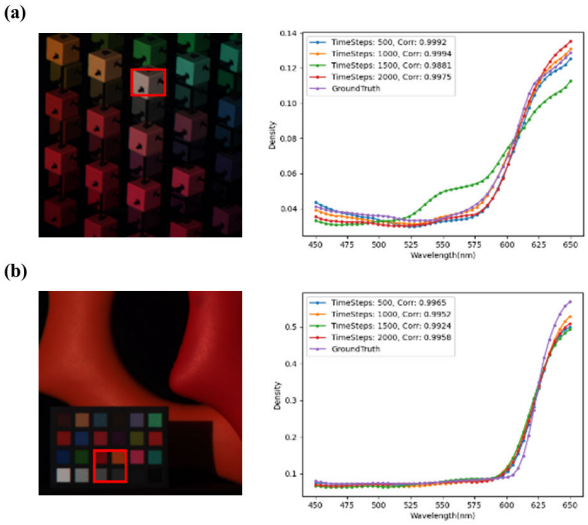


Fig. 8. Ablation study on sampling timestep. (a) and (b) Two scenes in the simulated data, and the red boxes represent randomly selected regions with rich spectral features, respectively. In the graph, the predicted spectral curve and the ground truth of different sampling timesteps are provided, and the correlation coefficient is also given. The cosine similarity [87] is taken as an example.

The high parameter count of CASSIDiff reflects its potential to capture fine-grained details and complexities in data, which is particularly valuable in hyperspectral image reconstruction where the ability to discern subtle spectral differences is crucial. However, this also implies a tradeoff between model performance and computational efficiency. CASSIDiff's higher parameter count and FLOPs (153.76 G) compared to other algorithms like MST and HDNet, which have lower FLOPs and parameters, indicates a higher computational cost and potentially increased memory usage during training and inference.

In summary, CASSIDiff's design strikes a balance between the need for high model capacity, as indicated by its parameter count, and the practical requirement for computational efficiency, as demonstrated by its FLOPs performance. This balance positions CASSIDiff as a strong contender among SOTA reconstruction algorithms, offering a competitive edge in both accuracy and efficiency. The choice of CASSIDiff as a reconstruction algorithm should consider not only its potential to deliver high-quality results but also its practical implications in terms of computational resource utilization, making it a viable option for applications, where both performance and efficiency are of paramount importance.

F. Analysis of Diffusion Processes

To intuitively reflect the inference of the conditional diffusion model, we show a sampling diagram of the 529.5-nm wavelength in hyperspectral images. As shown in Fig. 9. The total diffusion timestep is 1000, and we give visualization diagrams when the number of the diffusion timestep is 1000, 868, 682, 496, 248, 62, and 0. The reconstruction of the conditional diffusion model is a process of removing noise while generating a probability distribution that meets the expectations, guided by measurements. Therefore, for the hyperspectral reconstruction task in the CASSI system, the diffusion model has an additional denoising function. Moreover, facing high-frequency areas with rich details that are difficult to recover; in most cases, CASSIDiff can give a reasonable result after learning the correlation between global and local features through training.

Fig. 10 illustrates the convergence curves of our proposed CASSIDiff model during the validation process on both simulated and real data. It is observable that in the simulated data, the model approaches convergence after approximately

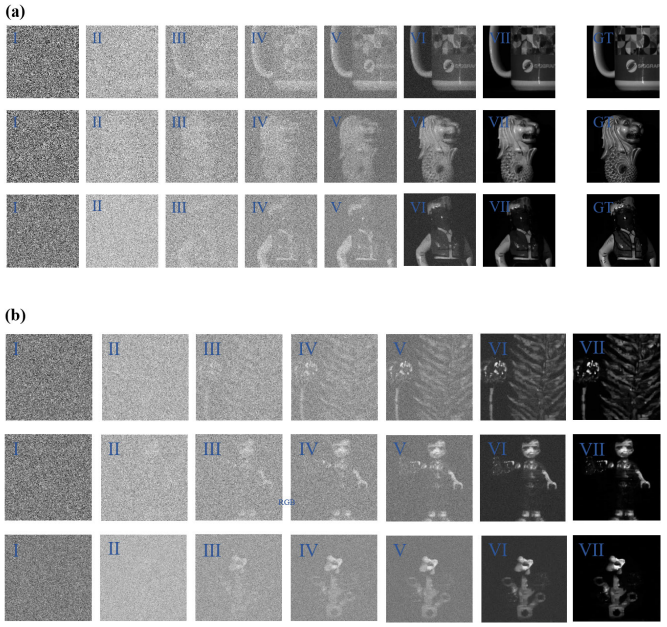


Fig. 9. Sampling process of CASSIDiff for 529.5-nm wavelength channel in hyperspectral data is shown. (a) and (b) Two scenarios of simulated and real data, respectively, and each scenario provides the results for sampling timestep of 1000, 868, 682, 496, 248, 62, and 0. In addition, for the convenience of comparison, the ground truth of simulated data is also given.

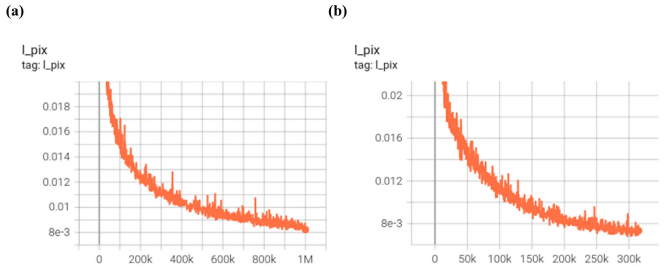


Fig. 10. Convergence curves on simulated data and real data. (a) Simulated data. (b) Real data.

1 million iterations; whereas with real data, convergence reaches around 300 000 iterations. The figures demonstrate a significant reduction in loss as the number of iterations increases, indicating that our method is effectively learning and enhancing its performance. However, it is also evident that a relatively large number of iterations are necessary for the model to fully converge. This insight should provide a clearer understanding of the training dynamics of our approach.

IV. Conclusion

In this article, we have highlighted the limitations inherent in the SSIM and PSNR metrics, commonly employed for hyperspectral reconstruction via the CASSI optical setup. And we introduce an evaluation framework that is more closely aligned with human visual perception, integrating LPIPS and FID as our metrics of choice. This approach effectively remedies the inadequacies of conventional assessment tools in capturing visual details and distortions, thus facilitating a more holistic and insightful evaluation.

Also, we have proposed a hyperspectral reconstruction algorithm based on a conditional diffusion model for the CASSI system: CASSIDiff. This approach has shown remarkable superiority when assessed through perceptual loss-based metrics. Many experimental results in simulated data and real data have shown that CASSIDiff not only reaches the first-class level for LPIPS-FID and PSNR-SSIM metrics but also has better reconstruction results when facing detailed areas that are difficult to recover. We have introduced the DWT-based module and the serial spatial–spectral attention module to solve the problem of multilevel feature fusion and accelerate convergence. As an innovative hyperspectral reconstruction algorithm, CASSIDiff holds significant potential for generating spectral cubes that align with human visual perception, outperforming other similar algorithms in this regard. However, it currently faces the drawback of lengthy training and sampling times, particularly when dealing with substantial datasets. This issue is well-documented in related literature [90], which offers theoretical insights into reducing the sampling timestep. Also, there is ample room for improvement in optimizing the model and refining the training approaches. In addition, applying CASSIDiff in other applications is also worthy of further exploration.

Acknowledgment

The authors gratefully thank Dr. Julian Evans of Zhejiang University, Hangzhou, China, for the helpful discussion.

References

[1] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.

[2] Z. Xu, Y. Jiang, J. Ji, E. Forsberg, Y. Li, and S. He, "Classification, identification, and growth stage estimation of microalgae based on transmission hyperspectral microscopic imaging and machine learning," *Opt. Exp.*, vol. 28, no. 21, p. 30686, 2020.

[3] H. Erives and N. B. Targhetta, "Implementation of a 3-D hyperspectral instrument for skin imaging applications," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 3, pp. 631–638, Mar. 2009.

[4] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1646–1657, Jul. 2017.

[5] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 11, pp. 4481–4492, Nov. 2019.

[6] M. Zucco, V. Caricato, A. Egidi, and M. Pisani, "A hyperspectral camera in the UVA band," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 6, pp. 1425–1430, Jun. 2015.

[7] C. Harrison Brodie, J. Devasagayam, and C. M. Collier, "A hyperspectral imaging instrumentation architecture based on accessible optical disc technology and frequency-domain analyses," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 7, pp. 2531–2538, Jul. 2019.

[8] B. Tu, X. Liao, C. Zhou, S. Chen, and W. He, "Feature extraction using multitask superpixel auxiliary learning for hyperspectral classification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–16, 2021.

[9] H. Zhu, J. Luo, J. Liao, and S. He, "High-accuracy rapid identification and classification of mixed bacteria using hyperspectral transmission microscopic imaging and machine learning," *Prog. Electromagn. Res.*, vol. 178, pp. 49–62, 2023.

[10] F. Succetti, A. Rosato, F. Di Luzio, A. Ceschini, and M. Panella, "A fast deep learning technique for Wi-Fi-based human activity recognition," *Prog. Electromagn. Res.*, vol. 174, pp. 127–141, 2022.

[11] A. Polak et al., "Hyperspectral imaging combined with data classification techniques as an aid for artwork authentication," *J. Cultural Heritage*, vol. 26, pp. 1–11, Jul. 2017.

[12] J. Qin et al., "A hyperspectral plant health monitoring system for space crop production," *Frontiers Plant Sci.*, vol. 14, Jul. 2023, Art. no. 1133505.

[13] J. Marcello, F. Eugenio, J. Martín, and F. Marqués, "Seabed mapping in coastal shallow waters using high resolution multispectral and hyperspectral imagery," *Remote Sens.*, vol. 10, no. 8, p. 1208, Aug. 2018.

[14] M. Dalponte, H. O. Ørka, T. Gobakken, D. Gianelle, and E. Næsset, "Tree species classification in boreal forests with hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2632–2645, May 2013.

[15] S. Peyghambari and Y. Zhang, "Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: An updated review," *J. Appl. Remote Sens.*, vol. 15, no. 3, Jul. 2021, Art. no. 031501.

[16] U. Heiden, W. Heldens, S. Roessner, K. Segl, T. Esch, and A. Mueller, "Urban structure type characterization using hyperspectral remote sensing and height information," *Landscape Urban Planning*, vol. 105, no. 4, pp. 361–375, Apr. 2012.

[17] A. Khan, M. T. Munir, W. Yu, and B. R. Young, "A review towards hyperspectral imaging for real-time quality control of food products with an illustrative case study of milk powder production," *Food Bioprocess Technol.*, vol. 13, no. 5, pp. 739–752, May 2020.

[18] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, "Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems," *Sensors*, vol. 19, no. 14, p. 3071, Jul. 2019.

[19] T. Erdogan, "Fiber grating spectra," *J. Lightw. Technol.*, vol. 15, no. 8, pp. 1277–1294, Aug. 1997.

[20] W. Wu, P. Han, M. Shi, F. Su, and F. Wu, "Design and performance analysis of a single-unit polarizing beam-splitting prism based on negative refraction in a uniaxial crystal," *Appl. Opt.*, vol. 58, no. 26, p. 7063, 2019.

[21] F. Schmitt, "Multispectral color image capture using a liquid crystal tunable filter," *Opt. Eng.*, vol. 41, no. 10, pp. 2532–2548, Oct. 2002.

[22] A. Chitnis, J. Ibbetson, and B. Keller, "Flip-chip phosphor coating method and devices fabricated utilizing method," U.S. Patent 8 878 219, Nov. 4, 2014.

[23] W. L. Chan, K. Charan, D. Takhar, K. F. Kelly, R. G. Baraniuk, and D. M. Mittleman, "A single-pixel terahertz imaging system based on compressed sensing," *Appl. Phys. Lett.*, vol. 93, no. 12, Sep. 2008, Art. no. 121105, doi: 10.1063/1.2989126.

[24] D. Shrekenhamer, C. M. Watts, and W. J. Padilla, "Terahertz single pixel imaging with an optically controlled dynamic spatial light modulator," *Opt. Exp.*, vol. 21, no. 10, pp. 12507–12518, 2013.

[25] M.-J. Sun et al., "Single-pixel three-dimensional imaging with time-based depth resolution," *Nature Commun.*, vol. 7, no. 1, p. 12010, Jul. 2016.

[26] C. Jiao, Z. Lin, Y. Xu, and S. He, "Noninvasive Raman imaging for monitoring mitochondrial redox state in septic rats," *Prog. Electromagn. Res.*, vol. 175, pp. 149–157, 2022.

[27] J. G. Dwight et al., "Compact snapshot image mapping spectrometer for unmanned aerial vehicle hyperspectral imaging," *J. Appl. Remote Sens.*, vol. 12, no. 4, Dec. 2018, Art. no. 044004.

[28] D. Sabatke, "Snapshot imaging spectropolarimeter," *Opt. Eng.*, vol. 41, no. 5, pp. 1048–1054, May 2002.

[29] 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.

[30] J. T. Mayhan, M. L. Burrows, K. M. Cuomo, and J. E. Piou, "High resolution 3D 'snapshot' ISAR imaging and feature extraction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 630–642, Apr. 2001.

[31] H. Li, X.-L. Zhao, J. Lin, and Y. Chen, "Low-rank tensor optimization with nonlocal plug-and-play regularizers for snapshot compressive imaging," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 581–593, 2022.

[32] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Opt. Exp.*, vol. 15, no. 21, pp. 14013–14027, Oct. 2007.

[33] Y. Zhao, S. Zheng, and X. Yuan, "Deep equilibrium models for video snapshot compressive imaging," 2022, *arXiv:2201.06931*.

[34] M. Descour and E. Dereniak, "Computed-tomography imaging spectrometer: Experimental calibration and reconstruction results," *Appl. Opt.*, vol. 34, no. 22, pp. 4817–4826, 1995.

[35] C. E. Volin, J. P. Garcia, E. L. Dereniak, M. R. Descour, D. T. Sass, and C. G. Simi, "MWIR computed-tomography imaging spectrometer: Calibration and imaging experiments," *Proc. SPIE*, vol. 3753, pp. 192–202, Oct. 1999.

[36] W. R. Johnson, D. W. Wilson, G. H. Bearman, and J. Backlund, "An all-reflective computed tomography imaging spectrometer," *Proc. SPIE*, vol. 5660, pp. 88–97, Dec. 2004.

[37] S. Derin Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian blind deconvolution using a total variation prior," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 12–26, Jan. 2009.

[38] A. A. Wagadarikar, N. P. Pitsianis, X. Sun, and D. J. Brady, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Opt. Exp.*, vol. 17, no. 8, pp. 6368–6388, 2009.

[39] J. Liu, Y. Sun, X. Xu, and U. S. Kamilov, "Image restoration using total variation regularized deep image prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7715–7719.

[40] W. He, H. Zhang, L. Zhang, and H. Shen, "Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 178–188, Jan. 2015.

[41] X. Lu, Y. Wang, and Y. Yuan, "Graph-regularized low-rank representation for destriping of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 4009–4018, Jul. 2013.

[42] C. Li, Y. Ma, J. Huang, X. Mei, and J. Ma, "Hyperspectral image denoising using the robust low-rank tensor recovery," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 32, no. 9, pp. 1604–1612, 2015.

[43] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*.

[44] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[45] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.

[46] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep Gaussian scale mixture prior for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16216–16225.

[47] R. Timofte et al., "NTIRE 2018 challenge on single image super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 852–863.

[48] B. Arad, R. Timofte, O. Ben-Shahar, Y.-T. Lin, and G. D. Finlayson, "NTIRE 2020 challenge on spectral reconstruction from an RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 446–447.

[49] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "HSCNN+: Advanced CNN-based hyperspectral recovery from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 939–947.

[50] Y. Zhao, L.-M. Po, Q. Yan, W. Liu, and T. Lin, "Hierarchical regression network for spectral reconstruction from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 422–423.

[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[52] 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.* Cham, Switzerland: Springer, Jan. 2020, pp. 187–204.

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

[54] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *Proc. 4th Int. Workshop Quality Multimedia Exper.*, Jul. 2012, pp. 37–38.

[55] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[56] M. S. Error, *Mean Squared Error*. Boston, MA, USA: Springer, 2010, p. 653.

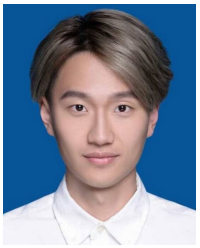
[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[58] J.-F. Pambrun and R. Noumeir, "Perceptual quantitative quality assessment of JPEG2000 compressed ct images with various slice thicknesses," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–6.

[59] W. Hou and X. Gao, "Be natural: A saliency-guided deep framework for image quality," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[60] C. M. Ward, J. Harguess, S. Parameswaran, and B. Crabb, "Image quality assessment for determining efficacy and limitations of super-resolution convolutional neural network (SRCNN)," *Proc. SPIE*, vol. 10396, pp. 19–30, Sep. 2017.

[61] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[62] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[63] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Jan. 2019, pp. 63–79.
[64] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Nov. 2020.
[65] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[66] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[67] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of ImageNet classes in Fréchet inception distance," 2022, *arXiv:2203.06026*.
[68] J. Wu, H. Fang, Y. Zhang, Y. Yang, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *Proc. Med. Imag. with Deep Learn.*, Jan. 2022, pp. 1623–1639.
[69] M. Farge, "Wavelet transforms and their applications to turbulence," *Annu. Rev. Fluid Mech.*, vol. 24, no. 1, pp. 395–458, Jan. 1992.
[70] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, 1995.
[71] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes Geophys.*, vol. 11, nos. 5–6, pp. 561–566, Nov. 2004.
[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[73] 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.
[74] 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. (ICCV)*, Oct. 2021, pp. 2622–2631.
[75] 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.*, Oct. 2019, pp. 4059–4069.
[76] H. Qin et al., "Forward and backward information retention for accurate binary neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2250–2259.
[77] X. Jiang, N. Wang, J. Xin, K. Li, X. Yang, and X. Gao, "Training binary neural network without batch normalization for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1700–1707.
[78] B. Xia et al., "Basic binary convolution unit for binarized image restoration network," 2022, *arXiv:2210.00405*.
[79] Z. Liu, Z. Shen, M. Savvides, and K.-T. Cheng, "ReActNet: Towards precise binary neural network with generalized activation functions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 143–159.
[80] Y. Cai et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 37749–37761.
[81] M. Li, Y. Fu, J. Liu, and Y. Zhang, "Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12959–12968.
[82] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2539–2543.
[83] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.
[84] 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. (CVPR)*, Jun. 2019, pp. 8024–8033.
[85] 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. (CVPR)*, Jun. 2020, pp. 1658–1668.
[86] S. Zheng et al., "Deep plug-and-play priors for spectral snapshot compressive imaging," *Photon. Res.*, vol. 9, no. 2, p. B18, 2021.
[87] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Model.*, vol. 53, nos. 1–2, pp. 91–97, Jan. 2011.
[88] 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.
[89] X. Hu et al., "HDNet: High-resolution dual-domain learning for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Oct. 2022, pp. 17542–17551.
[90] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.



Yifan Si was born in Dongying, Shandong, China, in 1997. He received the B.S. degree from Changchun University of Science and Technology, Changchun, China, in 2019. He is currently pursuing the Ph.D. degree in optical engineering with Zhejiang University, Hangzhou, China.

His current research interests include computational imaging, deep learning, and hyperspectral imaging.

Zijian Lin was born in Yingkou, Liaoning, China, in 1998. He received the bachelor's degree in optoelectronic information science and engineering from Nankai University, Tianjin, China, in 2020. He is currently pursuing the Ph.D. degree in optical engineering with Zhejiang University, Hangzhou, China.

His research interests include microphotonics/nanophotonics, liquid crystal tunable filters, and spectral imaging.

Xiaodong Wang was born in Ganzhou, Jiangxi, China, in 1997. He received the B.S. degree from Changchun University of Science and Technology, Changchun, China, in 2019, and the M.S. degree from the Southern University of Science and Technology, Shenzhen, China, in 2022. He is currently pursuing the Ph.D. degree in computer science with Westlake University, Hangzhou, China.

His current research interests include computational imaging, low-level tasks, and 3-D reconstruction.

Sailing He (Fellow, IEEE) received the Licentiate of Technology and Ph.D. degrees in electromagnetic theory from the KTH Royal Institute of Technology, Stockholm, Sweden, in 1991 and 1992, respectively.

Since then, he has been at the KTH Royal Institute of Technology, as an Assistant Professor, an Associate Professor, and a Full Professor. He also works as the Director for the Joint Research Center of Photonics (JORCEP) between KTH and Zhejiang University, Hangzhou, China. He has first-authored one monograph (Oxford University Press) and authored or co-authored about 800 articles in refereed international journals. He has given many invited/plenary talks at international conferences and has served in the leadership for many international conferences. His current research interests include subwavelength photonics, and smart sensing and imaging.