Spectral-Cascaded Diffusion Model for Remote Sensing Image Spectral Super-Resolution

Bowen Chen, *Graduate Student Member, IEEE*, Liqin Liu, *Graduate Student Member, IEEE*, Chenyang Liu, *Graduate Student Member, IEEE*, Zhengxia Zou, *Member, IEEE*, and Zhenwei Shi*, *Senior Member, IEEE*

***Abstract*—Hyperspectral remote sensing images have unique advantages in urban planning, precision agriculture, and ecology monitoring since they provide rich spectral information. However, hyperspectral imaging usually suffers from low spatial resolution and high cost, which limits the wide application of hyperspectral data. Spectral super-resolution provides a promising solution to acquire hyperspectral images with high spatial resolution and low cost, taking RGB images as input. Existing spectral super-resolution methods utilize neural networks following a single-shot framework, i.e., final results are obtained by one-stage spectral super-resolution, which struggles to capture and model the complex relationships between spectral bands. In this paper, we propose Spectral-Cascaded Diffusion Model (SCDM), a coarse-to-fine spectral super-resolution method based on the diffusion model. The diffusion model fits the real data distribution through stepwise denoising, which is naturally suitable for modeling rich spectral information. We cascade the diffusion model in the spectral dimension to gradually refine the spectral trends and enrich spectral information of the pixels. The cascade solves the highly ill-posed problem of spectral super-resolution step-by-step, mitigating the inaccuracies of previous single-shot approaches. To better utilize the potential of the diffusion model for spectral super-resolution, we design Image Condition Mixture Guidance (ICMG) to enhance the guidance of image conditions and Progressive Dynamic Truncation (PDT) to limit cumulative errors in the sampling process. Experimental results demonstrate that our method achieves the state-of-the-art performance in spectral super-resolution. Codes can be found at https://github.com/Mr-Bamboo/SCDM.**

***Index Terms*—remote sensing, spectral super-resolution, diffusion model, cascade-based methods**

I. Introduction

HYPERSPECTRAL remote sensing images (HSIs) surpass RGB or multispectral images in capturing comprehensive spectral information. HSIs cover wavelength range of visible bands, short-wave infrared, mid-infrared, and even thermal infrared bands with nanometer spectral resolution. The wide wavelength range and fine spectral resolution of spectral data endow it unique advantages in various tasks, including military applications [1], mineral exploration [2], ecological conservation [3, 4], urban construction [5], and post-disaster recovery [6]. However, due to the fine spectroscopy of different imaging bands, HSIs often face rigorous challenges, such as high acquisition costs, restricted imaging quality, limited spatial resolution [7], etc. Consequently, acquiring remote sensing images that possess both high spatial and spectral resolution remains an arduous undertaking, profoundly restricting the scope of HSI applications.

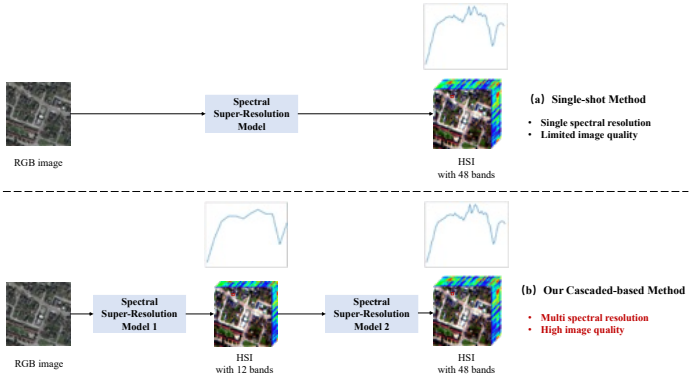


Fig. 1. The previous approach is shown in (a) and our approach is shown in (b). Our approach obtains higher-quality spectral super-resolution results and multiple spectral resolutions by cascading the models in the spectral dimension..

Spectral super-resolution methods offer a potential solution to mitigate above challenges, employing well-designed models to obtain high spatial resolution HSIs from RGB remote sensing images as input. However, transforming RGB images to HSIs is inherently challenging due to its ill-posed nature, making it a complex process to model. Some attempts aim to determine the optimal linear combination of basis functions for each hyperspectral pixel through linear model optimization [8] or reconstruct HSIs using a carefully trained sparse dictionary [9–11]. However, these methods predominantly rely on linear models to represent the mapping process, overlooking the multitude of nonlinear intricacies inherent to HSIs, making achieving satisfactory outcomes a formidable task.

With the burgeoning of deep learning techniques in the domain of natural image generation [12, 13], deep learning-based methods are progressively gaining popularity in the

The work was supported by the National Natural Science Foundation of China under Grant 62125102, the National Key Research and Development Program of China under Grant 2022ZD0160401, the Beijing Natural Science Foundation under Grant JL23005, the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20240933, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Bowen Chen, Liqin Liu, Chenyang Liu and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Liqin Liu and Chenyang Liu is also with Shen Yuan Honors College of Beihang University, Beijing 100191, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

field of spectral super-resolution. These methods extract high-level semantic features from RGB or multispectral images through neural networks and subsequently generate HSIs based on these features, effectively addressing the challenge of accommodating the non-linear characteristics inherent to HSIs. Deep learning-based approaches can be broadly classified into two architectural paradigms. The first employs well-designed convolutional neural networks (CNNs) augmented with pixel-level reconstruction loss (e.g., Mean Squared Error Loss) and incorporates advanced techniques such as residual structures [14, 15], attention mechanisms [16–19], group recovery [20, 21], structural prior [22, 23] and other innovations to improve the synthesis capacity of the network. The second category involves the development of Generative Adversarial Networks (GANs) [24], wherein discriminators guide the generators in producing more authentic outcomes [25, 26]. The former exhibits greater training stability but may encounter difficulties in generating high-quality HSIs due to the constraints imposed by the optimization objectives. In contrast, the latter demonstrates well in generating HSIs with high fidelity but may exhibit training instability and the tendency to replicate the most common situation in the training dataset, potentially failing to accurately fit the real data distribution.

In recent years, diffusion models have attracted much attention for tasks such as high-quality image generation [27] and image super-resolution [28], which match the data distribution by learning to reverse a gradual and multi-step noising process [29, 30]. Diffusion models offer desirable distribution coverage and a simple training objective, enabling them to effectively capture real data distributions while maintaining training stability [31]. These properties make diffusion models naturally suited for modeling the intricacies of spectral variability.

Although previous methods have achieved some success in spectral super-resolution tasks, they all follow the single-shot paradigm (Fig. 1 (a)), i.e., RGB remote sensing images are processed in one stage to obtain complete spectral information. However, real hyperspectral remote sensing images have complex characteristics due to the influence of the atmosphere, solar illumination, and sensor noise, which makes it difficult to model their complete spectral details at once.

To address the above issues, we propose Spectral-Cascaded Diffusion Model (SCDM), a multi-stage spectral super-resolution method based on diffusion models. First, coarse-grained spectral trends are obtained by using RGB information to determine the basic properties of ground objects. Then, the spectral trends are progressively refined by multiple diffusion models cascaded in the spectral dimension. During the refinement process, SCDM gradually fits the spectral variability caused by the real-world imaging process to improve the accuracy of spectral super-resolution. Meanwhile, our method can also output spectral super-resolution results with different spectral resolutions.

In order to fully integrate the rich information in the input image conditions, we draw inspiration from [32] and propose the Image Condition Guidance Mixture (ICMG) strategy. During the training process, the pre-trained diffusion model is fine-tuned by randomly removing the image-conditional inputs. During the sampling process, ICMG enhances the consistency between the input image conditions and the sampling results by linearly mixing the conditional and unconditional outputs.

Furthermore, in reconstructing HSIs, the sampling process in diffusion models can lead to instability, resulting in the gradual accumulation of prediction errors. Additionally, the ICMG strategy can produce high-dimensional unconditional outputs, which may amplify this instability. The existing fixed-threshold truncation methods are insufficient to handle these situations. To address these issues, we implement an adaptive truncation of the predicted noise at each step of the sampling process. The truncation threshold is related to a certain percentile of the current noise level, and this percentile increases progressively during the sampling to accommodate the difficulty of prediction at different time steps. This truncation method is termed Progressive Dynamic Truncation (PDT). Compared with the fixed-threshold truncation, PDT can effectively suppress instability and fully release the potential of the ICMG strategy.

Experiments on the IEEE *grss_dfc_2018* [33, 34] dataset show that SCDM can outperform CNN-based and GAN-based methods for high-quality spectral super-resolution. The main contributions of the paper can be summarized as follows:

1) Different from the previous single-shot paradigm, we propose a novel super-resolution approach cascading in spectral dimensions. On this basis, we propose the Spectral-Cascaded Diffusion Model (SCDM), a multi-stage method that models spectral information from coarse to fine, allowing for a better fit to the complex spectral properties of HSIs.
2) To fully utilize the input image condition, we design the Image Condition Guidance Mixture (ICMG), and to suppress instability of the sampling process, we propose Progressive Dynamic Truncation (PDT).
3) The effectiveness of the proposed method is empirically substantiated, exceeding the existing state-of-the-art spectral super-resolution methods on four fidelity metrics. Moreover, our method can output results at multiple spectral resolutions.

The rest of the paper is organized as follows. In Section II, we introduce spectral super-resolution methods, diffusion models and the cascaded-based models. Section III details the SCDM method. Section IV provides experimental evaluations on the reconstruction quality. Section V gives the discussion about the advantages and limitations of SCDM. Finally, we draw conclusions in Section VI.

II. Related Works

In this section, we briefly review the spectral super-resolution methods, diffusion models, and cascading-based methods for image generation. Specifically, we focus on diffusion models in the field of image generation.

A. Spectral super-resolution

Techniques to obtain hyperspectral images with high spatial resolution include super-resolution [35–38], spectral super-resolution [39–41], and image fusion [42]. Spectral super-resolution complements missing spectral information in the

spectral dimension. In this paper, we leverage high-resolution RGB or multispectral images as conditional inputs to produce high-resolution HSIs.

Some traditional spectral super-resolution methods revolved around the concept of optimizing linear models, including the utilization of basis functions [8, 43] or sparse representations [9–11, 44]. However, these methods employed linear models to model HSIs, which led to suboptimal modeling of the inherent nonlinear characteristics. Furthermore, they exhibited sensitivity to the selection of basis functions or dictionaries, rendering them susceptible to potential instability.

In recent years, the rapid development of neural networks has made deep learning-based methods a hot spot in spectral super-resolution research. Such methods excel in learning the nonlinear characteristics inherent to HSIs, leading to superior synthesis outputs. In terms of model architecture, deep learning-based methods can be categorized into two primary classes. One class is grounded in CNN and pixel-level loss functions, which feature a straightforward optimization objective and robust training processes but may yield suboptimal synthesis quality. The other class relies on GAN [24], excelling in generating realistic results but often involving intricate training procedures. To alleviate the synthesis complexity, techniques like residual learning [14, 15], attention mechanisms [16–18, 45], 3D convolutions [46, 47] and signal decomposition theories [48, 49] have been introduced. Furthermore, some researchers have integrated physical priors derived from the imaging process, such as spectral response function (SRF) [50] or spectral mixture model [51], to enhance the realism and interpretability of the synthesized results.

Recently, Liu et al. also proposed a spectral super-resolution method based on diffusion modeling. They introduced HyperLDM [52], which aimed at transforming high-dimensional hyperspectral data into a lower-dimensional latent space utilizing VQGAN, subsequently diffusing it within this latent domain. Despite offering a partial solution to the challenge of challenging noise prediction, this approach still exhibits constraints in achieving high-quality spectral super-resolution.

B. Diffusion Models

Diffusion models aim to generate images from Gaussian noise via an iterative denoising process, which consists of a forward process and a reverse process. During the forward process, an image is converted to a Gaussian noise by adding random Gaussian noise with T iterations. Ho et al. first combined the diffusion model with a score-based model and proposed the denoising diffusion probabilistic model (DDPM) [29], which has achieved great success in image synthesis. In contrast to GAN, the diffusion model demonstrates more desirable distribution coverage, simpler training objectives, and enhanced scalability [27]. Consequently, an increasing number of researchers have shifted their focus towards improving DDPM, unveiling the significant potential of diffusion models within the field of image synthesis.

In recent years, a substantial body of research has been conducted on conditional diffusion models to cope with downstream tasks with different conditional inputs. For category-conditional image synthesis, Dhariwal et al. proposed the Classifier-Guided Diffusion Model [27], which utilizes pre-trained classifiers for guiding the sampling process of the diffusion model. Subsequently, Ho et al. introduced the Classifier-Free Guidance strategy [53], aimed at achieving the trade-off between fidelity and diversity, all without the dependency on a classifier. For image super-resolution, Saharia et al. proposed SR3 [28], which generates high-quality images through iterative refinement. For the text-generated image task, GLIDE [54] combines the text feature into transformer blocks in the denoising process. DALL-E 2 [55] combines a pre-trained CLIP encoder with GLIDE and introduces a cascade architecture to achieve higher-quality results. In addition, the diffusion model also achieves surprising results in tasks such as semantic image synthesis [32], inpainting [56], and colorization [57]. Diffusion model-based approaches have also achieved impressive success on remote sensing image super-resolution tasks [58–60].

However, diffusion directly for spectral super-resolution presents certain challenges. HSIs typically consist of tens or even hundreds of spectral bands compared with RGB or multispectral images. In this context, diffusion models necessitate the incorporation or removal of noise in a high-dimensional space, which makes the noise prediction problem difficult. In this paper, we investigate diffusion models for spectral super-resolution and deal with high-dimensional predictions by cascading them in spectral dimension.

C. Cascade-based Image Synthesis Methods

Cascading pipelines are frequently utilized in the generation of large-scale images. When compared to small-scale images, the synthesis of large-scale images often necessitates the utilization of complex neural networks and intricate computational techniques to effectively capture longer-range dependencies among pixels. Cascading pipelines was initially explored within the context of methods like VQ-VAE [61, 62] and autoregressive models [63], which successfully mitigated the challenges associated with large-scale image synthesis by decoupling the synthesis task along the size dimension. Subsequently, cascaded GANs have also demonstrated remarkable success in various tasks, including image dehazing [64] and sample augmentation [65].

In recent years, cascading pipelines have also been introduced to diffusion models. Ho et al. [66] used cascaded diffusion models to achieve category-conditional image generation. Saharia et al. [28] applied cascaded diffusion models for image super-resolution. In the domain of text-to-image task, DALL-E 2 [55] and Imagen [67] utilize text-conditioned diffusion models for small-scale image generation, and a series of super-resolution diffusion models to generate larger-scale images.

However, most of the existing cascading pipelines cascade models in the spatial dimension. To the best of our knowledge, there exists no image synthesis method that cascades in the spectral dimension.

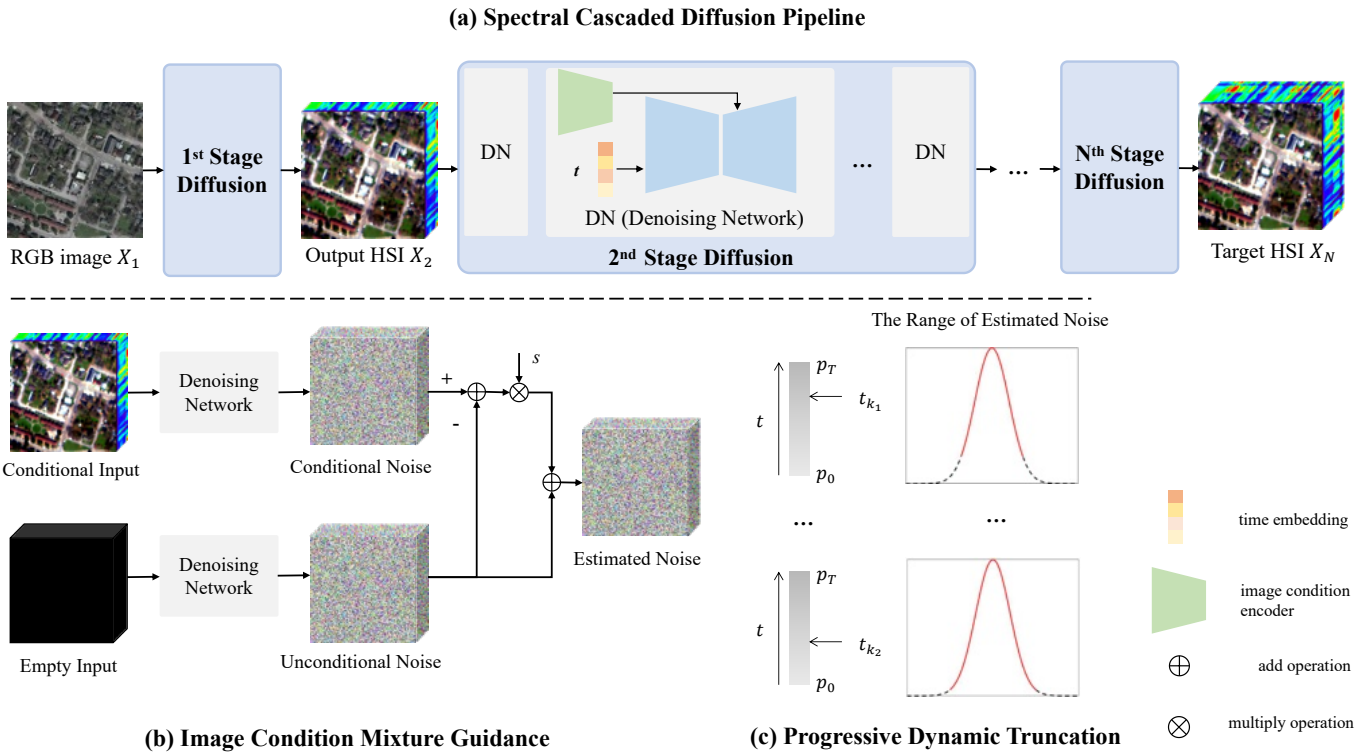


Fig. 2. An overview of the proposed method. The pipeline of SCDM is shown by (a).SCDM consists of a series of conditional diffusion models cascaded in the spectral dimension. The conditional input to the first stage of the diffusion model is the RGB remote-sensing images, and the subsequent diffusion models are conditioned on the output of the previous stage. The generation process of each stage starts with noise and iteratively denoises to yield the desired outputs for the corresponding spectral bands. (b) shows the sampling procedure with ICMG, which achieves better utilization of the input conditions by mixing conditional and unconditional outputs weighted in sampling. (c) illustrates the process of PDT, where p represents the truncation percentile. At the early stages of sampling, the p value is relatively small. As the sampling progresses (with decreasing t), p increases.

III. Proposed Method

We introduce the Spectral Cascade Diffusion Model (SCDM) as illustrated in Fig. 2, which consists of a series of diffusion models cascaded in the spectral dimension. The multiple diffusion models of the SCDM gradually generate images with an increasing number of bands, starting from RGB images, and subsequently up-sampling the images along the spectral dimension while incorporating spectral details. The structure of the multiple diffusion models remains consistent except for the dimensions of the input and output channels. These models can be simultaneously and independently trained, obviating the need for data generated by the previous model.

A. Priliminary of Denoising Diffusion

The diffusion model [29] consists of two processes: a forward process that perturbs data to noise and a reverse process that converts noise back to data. The forward process is typically hand-designed to transform any data distribution into a simple Gaussian distribution, while the reverse process generates data samples by inverting the forward process with deep neural networks.

Let the diffusion process consist of a total of T steps, x_0 denotes the input data sample, and x_t denotes the intermediate state of the input data in the diffusion process with a time step of $t = 1, 2, ..., T$. The forward process $q(\cdot)$ can be defined as

$$\begin{aligned} q(x_1, x_2, \ldots, x_T|x_0) &:= \prod_{t=1}^{T} q(x_t|x_{t-1}) \\ q(x_t|x_{t-1}) &:= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \end{aligned} \tag{1}$$

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes x follows a Gaussian distribution with mean μ and variance σ^2 and $\beta_t \in (0, 1)$ is a predefined sequence of variances of Gaussian noise.

As observed by Sohl-Dickstein et al. [68], we can marginalize the joint distribution in Eq. (1) to obtain the analytical form of $q(x_t|x_0)$, which is denoted as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} (1-\beta_t)$. Then given x_0, we can easily obtain a sample of x_t by sampling a Gaussian vector $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ and utilizing the reparameterization technique, without iteratively adding noise.

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{3}$$

For the reverse process, the diffusion model starts by first generating an unstructured noise from the prior distribution, then gradually removing noise therein by running a learnable Markov chain in the reverse time direction. Specifically, the

purpose of the reverse process is to model the reverse distribution $q(x_{t-1}|x_t)$ at time step t. If μ_θ and Σ_θ denote the mean and covariance of the Gaussian noise that should be removed, respectively, the reverse distribution can be approximated by a trainable neural network with parameters θ as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{4}$$

The joint distribution is

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \tag{5}$$

where $x_T \sim \mathcal{N}(0, I)$.

To optimize the neural network, we minimize the Kullback-Leibler (KL) divergence between the reverse process $p_\theta(x_{0:T})$ and the actual time reversal of the forward process $q(x_{0:T})$:

$$\begin{aligned} L &:= D_{KL}\left(q(x_{0:T}) \| p_\theta(x_{0:T})\right) \\ &= \mathbb{E}_{q(x_0)q(x_{1:T}|x_0)}\left[-log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] + \text{const}, \end{aligned} \tag{6}$$

where const represents a constant that does not depend on the model parameter θ and hence does not affect optimization. The first term of the second line in Eq. (10) is the variational lower bound (VLB) of the log-likelihood of the data x_0, that is

$$L_{vlb} = \mathbb{E}_{q(x_0)q(x_{1:T}|x_0)}\left[-log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]. \tag{7}$$

The objective of DDPM training is to minimize the negative VLB, which is particularly easy to optimize because it is a sum of independent terms, and can thus be estimated efficiently by Monte Carlo sampling and optimized effectively by stochastic optimization.

Further improvements come from variance reduction by rewriting the Eq. (7) as

$$\begin{aligned} L_{vlb} = \mathbb{E}_q\Bigg[& \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} \\ & + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \\ & - \underbrace{\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Bigg]. \end{aligned} \tag{8}$$

Noting that L_T is irrelevant to θ and L_0 is equal to L_{t-1} when $t = 1$. Therefore, the L_{vlb} is determined by the expected value of the sum of L_{t-1}. The tractable posterior distribution $q(x_{t-1}|x_t, x_0)$ conditioned on $x_0$is:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{9}$$

where $\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_t)}{1-\bar{\alpha}_t}x_t$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

Thus L_{t-1} can be written in the following form

$$\begin{aligned} L_{t-1} &= D_{KL}\left(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)\right) \\ &= \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\right] + \text{const}. \end{aligned} \tag{10}$$

Finally, utilizing the reparameterization technique to transform the mean form of L_{t-1} into the noise form yields

$$L_{simple} = \mathbb{E}_{x_0\sim q(x_0), \epsilon\sim\mathcal{N}(0,I), t\sim\mathcal{U}(1,T)}\left[\lambda(t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{11}$$

where $\lambda(t)$is a weighting function, $\mathcal{U}(1, T)$ is a uniform distribution over the set $1, 2, ..., T$, and $\epsilon_\theta(x_t, t)$ is the noise predicted by the neural network given x_t and t and parameterized by θ.

B. Spectral-Cascaded Diffusion

Hyperspectral data exhibit complex distribution and high dimensionality, making it challenging to direct synthesis using diffusion models. We propose a novel approach that decomposes the process from RGB remote sensing images to hyperspectral remote sensing images along the spectral dimension. Specifically, we employ a cascade of multiple diffusion models. The initial model takes the RGB image as a conditional input and generates an image with low spectral resolution. Subsequent models utilize the output of the preceding model as their conditional input, progressively enhancing spectral details. The final diffusion model produces the target HSI. We term this pipeline as Spectral-Cascaded Diffusion.

Based on experience, we employed a two-stage cascade model in this study, ensuring that the increasing multiplier in the number of bands remains as consistent as possible across both stages. This allocation strategy strikes a balance between the complexity of reconstruction and error accumulation. We summarize the band allocation strategy with the following formula

$$n = \lfloor 3 \times \sqrt{N/3} \rfloor, \tag{12}$$

where n is the number of HSI bands to be reconstructed in the first stage, N is the total number of bands in the complete HSI to be reconstructed, $\lfloor \cdot \rfloor$ represents rounding operation. The bands to be reconstructed in the first stage are obtained by sampling the complete HSI at equal intervals along the spectral dimension.

For each stage, we design a UNet [69] as noise prediction network. Using conditional RGB images as input, the diffusion model is iteratively denoised for T steps, gradually recovering the HSI from the noise randomly sampled from $\mathcal{N}(0, 1)$. In this section, we describe the design of the denoising network and the loss function.

1) Denoising network: We utilize the U-Net from [29] as the base noise prediction network, augmented with conditional inputs. Hence, we take the time step t and the image condition X_i required for this diffusion process as inputs to the noise prediction network. For the image conditional input, we first extract the features by an image conditional encoder. The encoded features are concatenated with the noisy image features obtained by the U-Net encoder and finally fed into the decoder of the U-Net to remove the noise. Except for the initial convolutional layer, the image conditional encoder shares weights with the other layers of the U-Net encoder.

The U-net is built mainly on Resblocks [70] and AttentionBlocks [71] with 2D convolution. For U-Net en-

coder, We perform 7 downsampling operations with multiplicity $[1, 2, 2, 2, 2, 2, 2]$, resulting in feature resolutions of $[256^2, 128^2, 64^2, 32^2, 16^2, 8^2, 4^2]$ in the specified order. Each Resblock is repeated 2 times for each resolution. The AttentionBlock is only used for the feature resolution at 16×16. The design of the U-Net decoder is almost symmetrical to the encoder, except that the number of repetitions of the Resblock has been changed to 3. U-Net uses skip connections between the features of the encoder and decoder corresponding to the resolution.

2) Loss functions: We use two objective functions to train our spectral-cascaded diffusion model. The first objective function is the DDPM denoising loss. According to Eq. (11), it can be expressed as the Mean Square Error (MSE) loss between the predicted noise and the true noise:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(1,T)} \|\epsilon - \epsilon_\theta\|_2^2. \tag{13}$$

However, using only MSE would focus more on the differences between pixel values and ignore the requirement of spectral profile accuracy, so we constrain the spectral properties of the synthesized images by introducing the spectral angle similarity loss. The spectral angle similarity loss is defined as the cosine similarity between the original and synthesized spectral vectors. The loss is written as follows:

$$\begin{aligned} \mathcal{L}_{\text{SAM}} &= 1 - \cos(x_0, x_{\theta,0}) \\ &= 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} x_{\theta,0_{ij}} \cdot x_{0_{ij}}}{\sqrt{\sum_{j=1}^{m} (x_{\theta,0_{ij}})^2} \cdot \sqrt{\sum_{j=1}^{m} (x_{0_{ij}})^2}}, \end{aligned} \tag{14}$$

where $x_{\theta,0}$ is denoted as $\frac{1}{\sqrt{\alpha_t}} \left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\right)$ and $x_{0_{ij}}$ represents the spectrum of the pixel at position (i, j) in $x_{0_{ij}}$.

Therefore, the overall objective function is:

$$\begin{aligned} \mathcal{L}_{\mathcal{D}} &= \mathcal{L}_{MSE} + \lambda \mathcal{L}_{SAM} \\ \theta^* &= \arg\min_{\theta} \mathcal{L}_{\mathcal{D}}. \end{aligned} \tag{15}$$

C. Image Condition Mixture Guidance

For the diffusion model conditioned on image c, previous studies have obtained the image at step $t-1$ by learning the conditional distribution $p_\theta(x_t|c)$ and then taking the following sampling approach similar with DDPM [29]:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t|c)) + \sigma_t \epsilon, \tag{16}$$

where $\epsilon_\theta(x_t|c)$ is the noise prediction output of the denoising network $D_i(x, t, c)$ for i th diffusion model.

However, it has been shown that conditional diffusion models can not explicitly handle conditional inputs in the sampling process of the DDPM. In addition, the sampling approach of DDPM may result in synthetic images that lack realism and exhibit reduced correlation with the conditional image, particularly when there is a significant disparity between the conditional image conditions and the final output [32]. Dhariwal et al. introduced a classifier guidance technique, which enhances the quality of generated samples by mixing the score estimates from the conditional diffusion model with the log-probability gradient of the classifier [27], which is denoted as:

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c) \cdot \nabla_{x_t} \log p(c|x_t), \tag{17}$$

where s is the guidance weight used to weigh sample quality over diversity, and $\Sigma_\theta(x_t|c)$ represents the variance. However, for image-conditioned diffusion models, relying on the classifier to provide the gradient is inappropriate.

Inspired by [53] and [32], we designed the Image Condition Mixture Guidance (ICMG) strategy for sampling image-conditioned diffusion models. It is worth noting that the dissimilarity between the estimated noise $\epsilon_\theta(x_t|c)$ in the image conditional guidance and the estimated noise $\epsilon_\theta(x_t|\emptyset)$in the unconditional case is directly related to the log-probability gradient employed for the mixing process, which is denoted as:

$$\begin{aligned} \epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset) &\propto \nabla_{x_t} \log p(x_t|c) - \nabla_{x_t} \log p(x_t) \\ &\propto \nabla_{x_t} \log p(c|x_t). \end{aligned} \tag{18}$$

Therefore, during the sampling process, we can improve the quality of the samples by mixing the output of the conditional diffusion model with the output of the jointly trained unconditional diffusion model:

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|c) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset)), \tag{19}$$

where $\emptyset$ is defined as the image with all band pixel values set to 0. The sampling process of ICMG is shown in Fig. 2 (b) and Algorithm 2.

In the training process, we jointly train unconditional diffusion models by setting the probability of randomly setting the image condition to zero. The training process for each diffusion model in Spectral-Cascaded Diffusion is shown in Algorithm 1.

Algorithm 1 Training of i th Diffusion Model in Spectral-Cascaded Diffusion

Input: $q(x_0)$, image condition c corresponding to x_0, probability of unconditional training p_{unc}, learning rate l_r.

Output: The denoising network with parameters $\theta_{\mathcal{D}_i}$ for i th diffusion model.

1: **repeat**
2: Sample $x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0, I), t \sim \mathcal{U}(\{1, \ldots, T\})$
3: $c \leftarrow \emptyset$ with probability p_{unc}
4: Predict the noise with $\epsilon_\theta(x_t|c) = \mathcal{D}_i(t, x_t, c)$, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
5: Calculate gradient descent $\nabla_{\theta_{\mathcal{D}_i}} \mathcal{L}_{\mathcal{D}}$
6: Update $\theta_{\mathcal{D}_i}$ with $\theta_{\mathcal{D}_i} \leftarrow \theta_{\mathcal{D}_i} - l_r \nabla_{\theta_{\mathcal{D}_i}} \mathcal{L}_{\mathcal{D}}$
7: **until converged**

D. Progressive Dynamic Truncation Strategy

For the distribution of bounded data (e.g., image data), truncation of the sampling process using thresholding can reduce the adverse effects of out-of-bounds errors in the sampling process, and thus guide the model to sample high-quality samples [67, 72]. Many of the previous image-conditioned

diffusion models use static thresholding to truncate the sampling results at each step. However, due to the complexity of hyperspectral images and the specificity of guided sampling, static thresholding sampling leads to insufficiently accurate spectral details in the synthesized images.

For guided sampling, Saharia et al. [67] proposed a dynamic thresholding technique that actively prevents pixels from exceeding the boundary at each step by pushing the out-of-bounds samples inward. In each sampling step of this technique, the dynamic threshold is set based on the pixel value of x_0 predicted directly. However, for hyperspectral images, it is difficult to predict x_0 directly, which leads to intolerable bias. Therefore, we modify the above truncation method by setting the truncation threshold dynamically for the noise predicted in each step.

Specifically, in each sampling step, we set τ_p to the absolute pixel value of a certain percentile p in $\epsilon_\theta(x_t|c)$. If τ_p exceeds 1, we truncate the predicted noise to the range $[-\tau_p, \tau_p]$, otherwise we truncate to the range $[-1, 1]$. In this way, we can control the truncation dynamically and adaptively at each step.

Simultaneously, the sampling process of the diffusion model introduces instability, which must be mitigated to achieve a HSI that closely resembles the ground truth. The initial steps of sampling (i.e., when t is close to T) establish the fundamental trajectory of the reconstruction and necessitate predictions at high noise levels, where instabilities are significant and may lead to catastrophic errors. To address this, we set smaller percentiles for truncation in initial sampling step to reduce instability. In the later stages of sampling, as the model predicts at lower noise levels with increased stability, the truncation percentile can be appropriately raised. Therefore, we define the percentile value p as a monotonically decreasing function of time t, denoted as $p(t)$, referred to as progressive dynamic truncation (PDT). This approach ensures that $p(t)$ is smaller at the beginning of the sampling and gradually increases as the process continues. In our method, $p(t)$ is set as a linear function of

$$p(t) = p_0 + (p_T - p_0)t. \quad p_T < p_0. \tag{20}$$

In general, we set p_0 to 1.0. Thus, the truncation threshold $\tau_p(t, \epsilon_\theta)$ is related to the time step and the noise at that time step.

The process of progressive dynamic truncation for sampling is shown in Fig. 2 (c) and Algorithm 2, where $Clamp(\hat{\epsilon}_\theta(x_t|c), \tau_p(t, \epsilon_\theta)$ represents threshold noise to the range $[-\tau_p(t, \epsilon_\theta), \tau_p(t, \epsilon_\theta)]$.

Algorithm 2 Progressive Dynamic Truncation for Sampling

Input: image condition c, guidance weight s, progressive function for threshold $\Gamma(t)$.
Output: $x_0 \sim q(x_0)$.

```
Sample x_T ~ N(0, 1);
for t = T, ..., 1 do
    Calculate conditional output ε_θ(x_t|c) = D(t, x_t, c)
    Calculate unconditional output ε_θ(x_t|∅) = D(t, x_t, ∅)
    ε̂_θ(x_t|c) = ε_θ(x_t|c) + s · (ε_θ(x_t|c) − ε_θ(x_t|∅))
    ε̃_θ(x_t|c) = Clamp(ε̂_θ(x_t|c), τ_p(t, ε_θ))
    Sample ε ~ N(0, I) if t > 1, else ε = 0
    x_{t−1} = 1/√(1−β_t) (x_t − β_t/√(1−ᾱ_t) ε̃_θ(x_t|c)) + σ_t ε
end for
```

IV. Experiment

In this section, we present the datasets and experimental settings to validate the performance of spectral super-resolution. Additionally, we showcase the synthesis results of SCDM and conduct meticulous ablation studies on it.

A. Datasets and Experimental Setup

We evaluate our method on the IEEE *grss_dfc_2018* dataset and Pavia Center dataset. IEEE *grss_dfc_2018* dataset was collected by the National Center for Airborne Laser Mapping (NCALM) from Houston University [33, 34]. The dataset contains an image scene with spatial size 4172×1202 and 48 bands, covering a wavelength range of 380-1050 nm. The bands 23,12,5 from the data are chosen as the RGB image condition input. The dataset is cropped into 27 paired 512×512 patches, where 3 patches without overlap are used for testing and others for training. All methods are trained on 256×256 patches randomly cropped from the training patches and tested on 3 patches with a size of 512×512 patches.

The Pavia Center dataset was captured by the ROSIS sensor covering a wavelength range of 430-860 nm and contains 102 bands with a total of 1096×715 effective pixels after removal of bad bands. The bands 60,30,10 from the data are chosen as the RGB image condition input. We used 1024×715 pixels in our experiments and cropped 8*6 = 48 patches with a size of 128×128. 40 of them are used for training and 8 for testing, with no overlap between the training and testing data.

The proposed SCDM method is compared with six state-of-the-art methods, including the CNN-based method MSCNN [39], HSCNN+ [14], FMNet [15], HSRNet [50] and HASIC-Net [18], the GAN-based method R2HGAN [26] and the diffusion-based method HyperLDM [52]. HSCNN+ [14] uses multiple residual blocks for feature mapping. MSCNN [39] designs a multiscale deep convolution network with pixel-shuffle. FMNet [15] designs an adaptive receptive field that maintains spectral learning capability. HASIC-Net [18] utilizes the attention module to focus on structural information similarity. LTRN [23] combine low-rank tensor decomposition with CNN to adaptively reconstruction HSI. R2HGAN generates HSIs under a GAN framework with spectral and spacial discriminators [26] and HyperLDM designs a latent diffusion model for spectral super-resolution [52]. For fair competition, all these methods are optimized adequately and parameters for the best results are selected.

For a fair competition, all the experiments are conducted with the Intel (R) Xeon (R) Gold 6330 CPU @2.00GHZ and an NVIDIA GeForce RTX 4090 GPU, optimized adequately and the best parameters are selected.

We utilize four metrics to measure the quality of spectral super-resolution, including root-mean-square error (RMSE), mean peak signal-to-noise ratio (MPSNR), mean structural

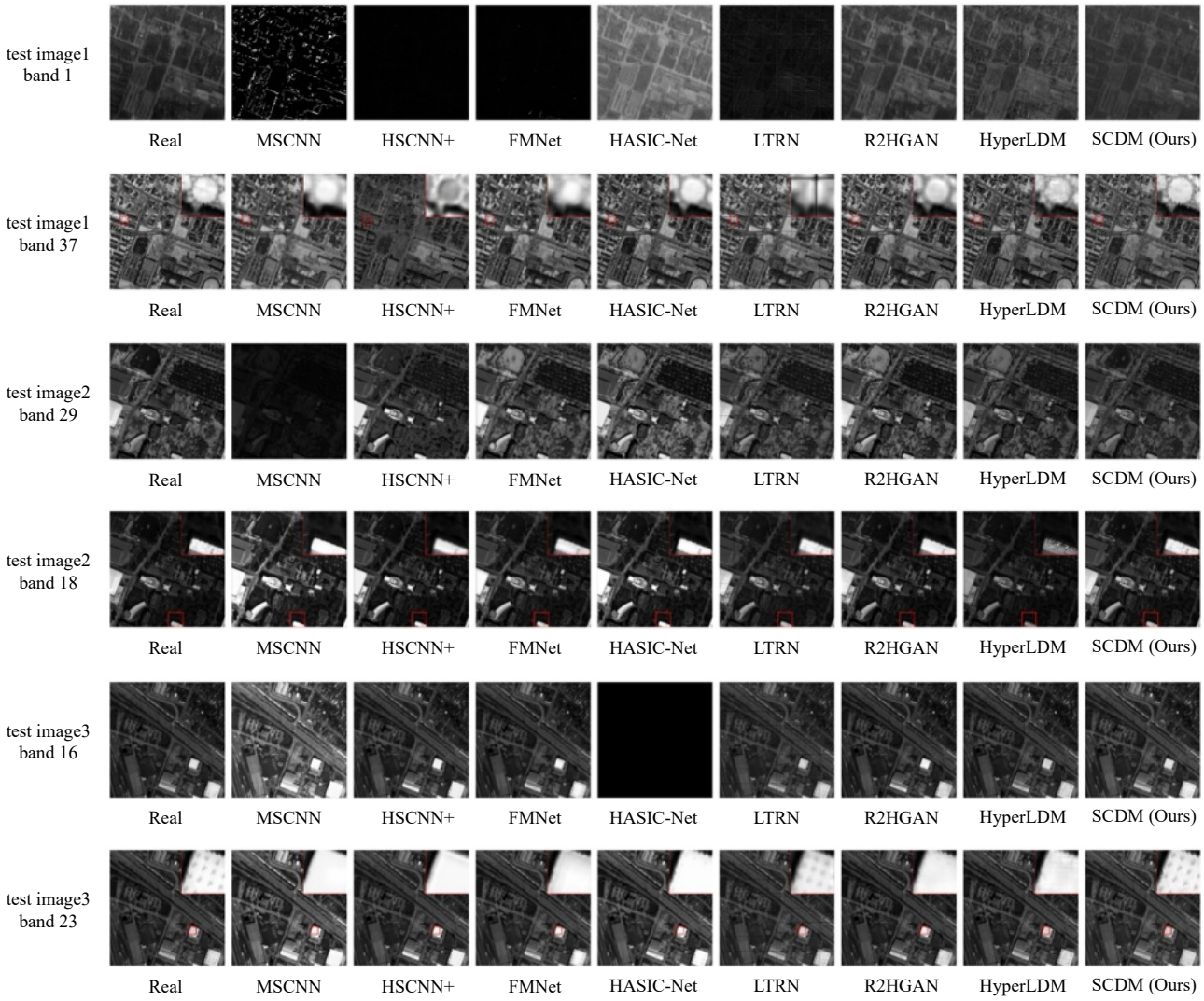


Fig. 3. Results of each spectral super-resolution method on different bands. Rows 1, 3, and 5 show the global results, and rows 2, 4, and 6 show the local details.

similarity (MSSIM) [73], and SAM, to quantitatively evaluate the performance of all the compared methods. The RMSE measures the difference between the reconstructed image and the true value. The MPSNR and MSSIM are metrics that show the spatial fidelity of the reconstructed HSI which are computed on each band and averaged over all spectral bands, with larger values of the results indicating that the method is more effective in preserving spatial detail. Meanwhile, SAM evaluates the spectral retention of all compared algorithms, with improved spectral fidelity when SAM is small.

B. Implementation Details

1) Parameter Settings: In the spectral-cascaded diffusion model, the time-embedding $e_t \in \mathbb{R}^{1\times 512}$ and the probability of unconditional training $p_{unc} = 0.1$. The loss weight of the spectral angle similarity loss $\lambda = 1$.

For IEEE *grss_dfc_2018* dataset, we employ a two-stage cascade process. In the first stage, we reconstruct a 12-band image with a spectral resolution of 60nm, equivalent to a 4x downsampling of the full HSI in the spectral dimension. In the second stage, we reconstruct the complete 48-band image. The PDT percentile threshold p_0 is set to 0.96 in the first stage and 0.88 in the second stage.

For Pavia Center dataset, we employ a two-stage cascade process. In the first stage, we reconstruct a 17-band image with a spectral resolution of 24nm, equivalent to a 6x downsampling of the full HSI in the spectral dimension. In the second stage, we reconstruct the complete 102-band image. The PDT percentile threshold p_0 is set to 0.97 in the first stage and 0.90 in the second stage.

2) Training details: For both datasets, during the training, we optimize each diffusion model with the Adam optimizer [74], and the learning rate is initially set to 10^{-4}. The first and second stage are both trained 100000 iterations.

C. Comparison with Other Methods

Fig. 3 shows the output results of each spectral super-resolution method. It can be seen that our SCDM still performs

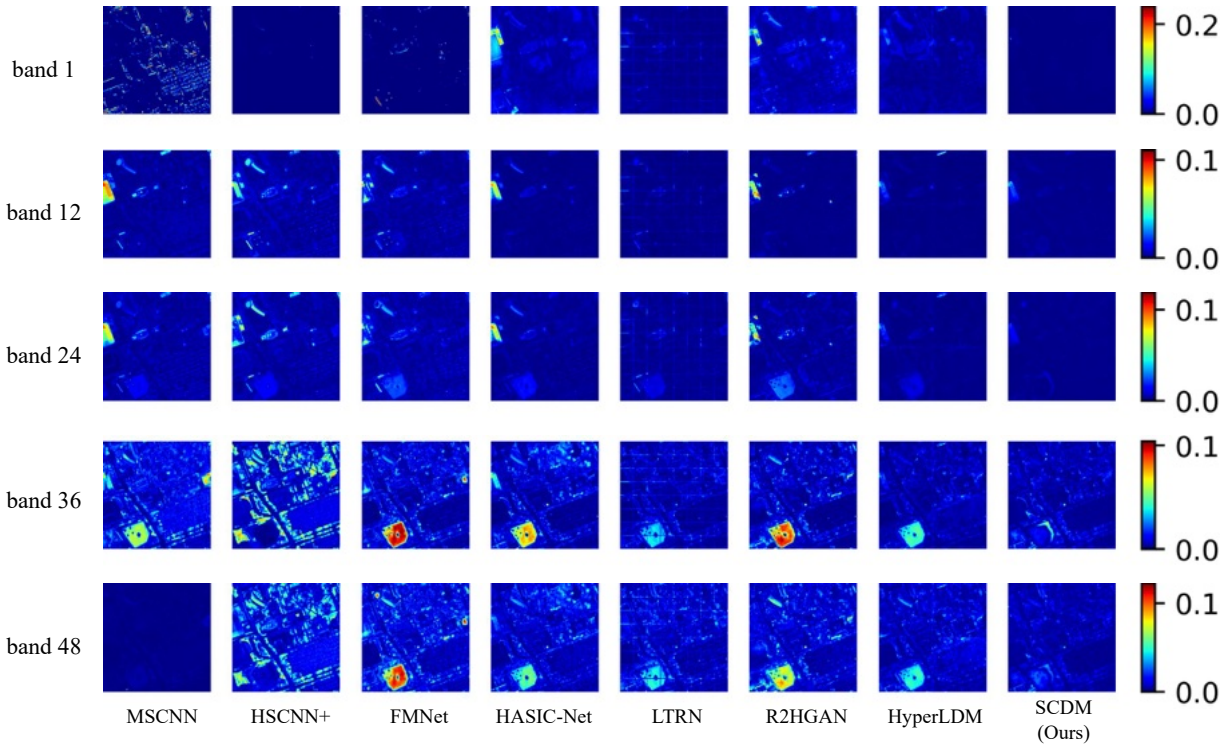


Fig. 4. The absolute differences between the reconstructed images and the ground truth at bands 1, 12, 24, 36, and 48 on IEEE *grss_dfc_2018* dataset. The scale values on the color bar on the right side represent the absolute difference divided by the maximum possible value in the reconstructed HSI.

better when other methods fail to reconstruct certain bands correctly. In addition, in terms of the spatial details of the reconstruction, our SCDM method also obtains optimal results. For example, in the results shown in row 6 of Fig. 3, all other spectral super-resolution methods are hard to obtain the point-like textures on the building, while SCDM can super-resolve these texture details clearly as well.

A comparison of the absolute differences in each band of the HSIs synthesized by our proposed SCDM and other methods with respect to the ground truth is shown in Fig. 4. The visualized results show that the SCDM method has smaller differences in each of the displayed bands compared to the other methods. It is worth noting that nearly all comparative methods struggle to achieve satisfactory reconstruction in the later bands, particularly in the near-infrared range beyond 750nm. However, the SCDM method demonstrates a significant advantage in synthesizing these bands. To illustrate this, we compared the PSNR values for each band, as shown in Fig. 5. The results indicate that our method exhibits a notable advantage in reconstruction quality, especially in the near-infrared bands beyond 750nm. This is because the input RGB image covers only the visible range and does not include information from the near-infrared range. In such cases, images reconstructed using single-stage methods often show lower quality in the near-infrared range. In contrast, the SCDM method, which progressively increases in the spectral dimension, better captures information from the near-infrared range.

We also compared the spectral curves of some key objects, as shown in Fig. 6. It can be seen that SCDM has more

TABLE I
ACCURACY OF DIFFERENT METHODS ON IEEE *grss_dfc_2018* DATASET. FOR RMSE AND SAM, A LOWER SCORE INDICATES BETTER, WHILE FOR MSSIM AND MPSNR, A HIGHER SCORE IS BETTER.

	Method	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
CNN-based	MSCNN	2117.14	0.1696	0.9555	37.945
	HSCNN+	1015.51	0.1559	0.9439	39.089
	FMNet	697.94	0.0875	0.9729	42.829
	HASIC-Net	887.05	0.1811	0.9834	44.312
	LTRN	644.22	0.1066	0.9585	41.297
GAN-based	R2HGAN	467.06	**0.0596**	0.9861	46.840
Diffusion-based	HyperLDM	515.99	0.0724	0.9797	44.736
	SCDM(Ours)	**359.48**	0.0672	**0.9887**	**47.207**

TABLE II
ACCURACY OF DIFFERENT METHODS ON PAVIA CENTER DATASET. FOR RMSE AND SAM, A LOWER SCORE INDICATES BETTER, WHILE FOR MSSIM AND MPSNR, A HIGHER SCORE IS BETTER.

	Method	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
CNN-based	MSCNN	1971.30	0.1686	0.8899	22.860
	HSCNN+	315.66	0.1435	0.8923	32.142
	FMNet	1080.80	**0.1011**	0.9344	25.779
	HASIC-Net	233.07	0.1080	0.9284	34.118
	LTRN	239.22	0.1268	0.9060	32.795
GAN-based	R2HGAN	852.83	0.1327	0.8926	28.3438
Diffusion-based	HyperLDM	/	/	/	/
	SCDM(Ours)	**172.42**	0.1042	**0.9509**	**36.496**

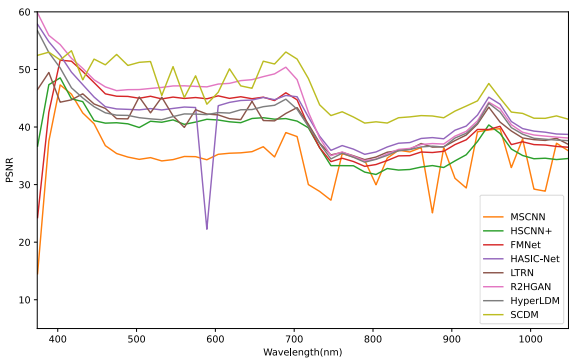


Fig. 5. PSNR values of each band achieved by different methods on IEEE *grss_dfc_2018* dataset.

accurate reconstruction results for these key features. For example, for the plant shown in Fig. 6(a) and the white car shown in Fig. 6(d), SCDM can more accurately fit the spectral values between the near-infrared bands (700 nm to 900 nm). And Fig. 7 shows that our SCDM is a process of continuous refinement of spectral details from spectral trends, gradually fitting the real spectral curve through a coarse-to-fine paradigm.

TABLE III
ABLATION STUDIES OF THE NUMBER OF CASCADE STAGES ON IEEE *grss_dfc_2018* DATASET. † DENOTES THE FINAL SETTING.

	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
1 stage	626.48	0.0764	0.9857	42.922
2 stage†	**359.48**	**0.0672**	**0.9887**	**47.207**
3 stage	572.75	0.0826	0.9845	42.844

TABLE IV
ABLATION STUDIES OF THE NUMBER OF CASCADE STAGES ON PAVIA CENTER DATASET. † DENOTES THE FINAL SETTING.

	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
1 stage	200.12	0.1118	0.9500	35.588
2 stage†	**172.42**	**0.1042**	**0.9509**	**36.496**
3 stage	191.92	0.1184	0.9440	35.277

TABLE V
ABLATION STUDIES OF THE INCREASING MULTIPLIER OF EACH STAGE ON IEEE *grss_dfc_2018* DATASET. † DENOTES THE FINAL SETTING.

increasing multiplier	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
2×,8×	595.14	0.0668	0.9815	42.467
4×,4×†	**359.48**	**0.0672**	**0.9887**	**47.207**
8×,2×	560.44	0.0792	0.9839	42.822

Finally, we report a comparison of similarity metrics on the two datasets and the results are presented in Tables I and II. It can be found that our proposed SCDM method achieves the optimum in three metrics of RMSE, MSSIM, and MPSNR compared to other methods. In terms of SAM metric, our method consistently achieved the second-best results. This demonstrates that the SCDM method can reconstruct HSIs of higher quality and fidelity compared to other methods.

Additionally, we do not conduct experiments with the HyperLDM method on the Pavia Center dataset. This decision is made because HyperLDM relies on a spectral library for reconstruction, which requires precise wavelength information for each band. However, the Pavia Center dataset has removed certain bad bands and does not provide the specific wavelengths corresponding to these bands. As a result, we are unable to match band numbers with wavelengths. To ensure fairness, we chose not to include a comparison of HyperLDM.

D. Ablation Studies

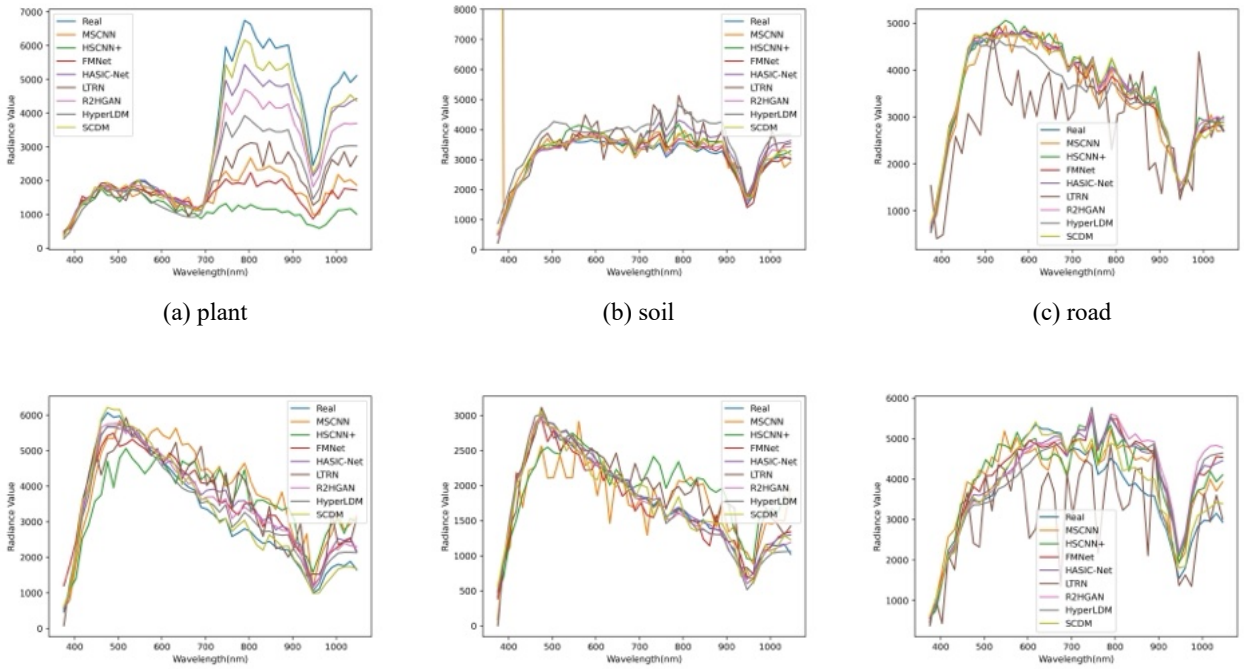
In this section, we further investigate some of the key designs in the method. First, we explore the effect of the number of diffusion models in the cascade on the super-resolution performance to verify the effectiveness of the spectral cascade and to determine the optimal cascade strategy. Afterward, we performe experimental validation of the effectiveness of PDT and ICMG at each stage in the cascade. Finally, we analyze the parameter settings of the PDT.

1) Design of cascade strategy: We first performed an ablation study on the number of cascade stages using the IEEE *grss_dfc_2018* dataset dataset and the Pavia Center dataset. For IEEE *grss_dfc_2018* dataset dataset, the two-stage setup involved band number increase multipliers of 4× in each stage, while the three-stage setup used multipliers of 4×, 2×, and 2×, respectively. For Pavia Center dataset, the two-stage configuration involved reconstructing 17 bands (5.7× increase) first, followed by 102 bands (6× increase). The three-stage configuration involved 12 bands (4× increase), 34 bands (2.8× increase), and 102 bands (3× increase).

The results, shown in Tables III and IV, reveal that the two-stage setup achieved the best results across four metrics on both datasets. In contrast, the three-stage method even showed a decline in three metrics compared to the single-stage approach. This decline is due to gaps in the training and testing processes. In addition to the first stage, each stage's input consists of true bands during training, while during testing, the input is the output from the previous stage. With an excessive number of stages, errors from earlier stages accumulate, leading to a decline in the quality of the final output.

Additionally, we conducted experiments on the IEEE *grss_dfc_2018* dataset, examining the effect of the band number increase multiplier in each stage. The results, presented in Table V, show that maintaining a consistent multiplier for the number of bands across both stages generally yields higher quality results.

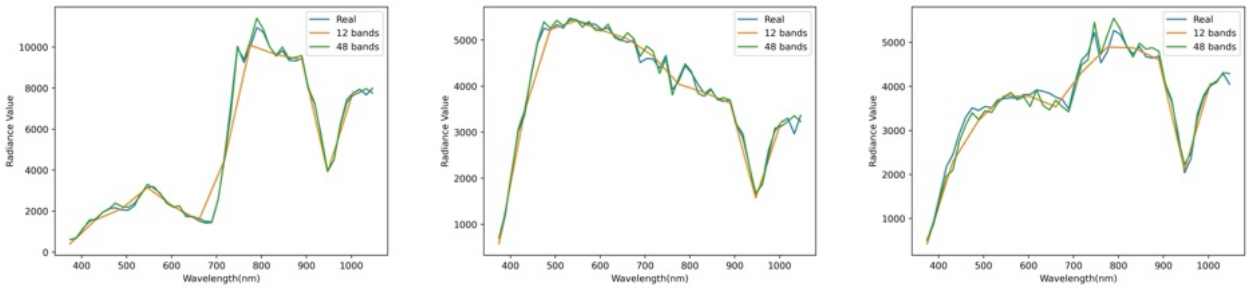
In conclusion, we recommend using a two-stage approach, ensuring that the multiplier for the increase in the number of bands is as consistent as possible across both stages. This strategy effectively breaks down the complexity of reconstructing HSIs, while also avoiding the error accumulation associated with having too many stages.



(a) plant (b) soil (c) road

(d) white car (e) parking lot (f) building

Fig. 6. Spectral curves on six objects generated by different methods.



(a) grass (b) road (c) soil

Fig. 7. Comparison between the output and the true value of SCDM for different spectral resolutions. It can be seen that SCDM is a process that goes from coarse to fine in the spectral dimension, and the prediction results are obtained from the gradual refinement of the spectral trend.

TABLE VI
ABLATION STUDIES OF THE SCDM ON IEEE *grss_dfc_2018* DATASET, † DENOTES THE FINAL SETTING.

	Stage 1		Stage 2		Metrics			
Name	PDT	ICMG	PDT	ICMG	RMSE	SAM	MSSIM	MPSNR
1					696.73	0.0729	0.9847	42.600
2	✓		✓		538.49	**0.0631**	0.9874	43.682
3 †	✓		✓	✓	**359.48**	0.0672	**0.9887**	**47.207**
4		✓		✓	647.09	0.0916	0.9818	42.732
5		✓	✓	✓	597.93	0.0769	0.9850	43.351
6	✓	✓	✓		565.89	0.0711	0.9855	43.471
7	✓	✓		✓	574.91	0.0833	0.9825	42.853
8	✓	✓	✓	✓	402.55	0.0728	0.9868	45.951

TABLE VII
ANALYSIS OF THE PDT PERCENTILE THRESHOLD p_T ON IEEE *grss_dfc_2018* DATASET. † DENOTES THE FINAL SETTING.

stage	p_T	RMSE ↓	SAM ↓	MSSIM ↑	MPSNR ↑
1	0.99	489.74	0.0946	0.9733	48.796
	0.98	376.64	0.0611	0.9882	49.656
	0.97	323.36	0.0479	0.9916	**49.911**
	0.96 †	**317.06**	**0.0473**	**0.9919**	49.684
	0.95	319.82	0.0484	0.9918	49.420
2	0.94	376.91	0.0791	0.9872	46.712
	0.92	361.23	0.0693	0.9885	47.179
	0.90	**359.07**	0.0674	0.9886	**47.237**
	0.88 †	359.48	**0.0672**	**0.9887**	47.207
	0.86	360.09	0.0673	0.9887	47.179

2) Effectiveness of ICMG and PDT: We explored the impact of the proposed ICMG strategy and PDT method on the IEEE *grss_dfc_2018* dataset, as shown in Table VI. It can be observed that, compared to fixed-threshold truncation, PDT effectively mitigates the instability caused by the diffusion sampling process and the unconditional outputs from ICMG, resulting in higher fidelity reconstructions (according to 2,6). The combined effect of PDT and ICMG allows for a more thorough extraction of information from the input image conditions, thereby improving the quality of the reconstructed results (according to 3,5,8).

Additionally, we found that employing PDT in the first stage and PDT combined with ICMG in the second stage yields the best overall performance. When both stages are set to use PDT and ICMG, although the results are not optimal, there are also improvements across four metrics compared to the baseline.

3) Parameter analysis: We conducted experiments on the IEEE *grss_dfc_2018* dataset to investigate the impact of different PDT percentile thresholds p_T at each stage. Table VII shows how varying p_T values in the first and second stages affect the reconstruction quality. The results indicate that the optimal p_T for the first stage is 0.96, while for the second stage, it is 0.88. This suggests that when fewer bands need to be reconstructed, a higher p_T should be used, and as the number of bands to be reconstructed increases, p_T should be lowered.

The rationale behind this observation is that in the early stages, our method reconstructs images with lower spectral resolution, which primarily reflect the basic properties of ground objects. At this stage, the instability is minimal, and extensive truncation during early sampling is unnecessary. However, in the later stages, the reconstructed images need to include a substantial amount of additional spectral detail, which also encompasses complex environmental effects and sensor noise, leading to higher instability. Therefore, stronger truncation during early sampling is required.

V. DISCUSSION

Compared to previous diffusion-based spectral super-resolution methods, SCDM shows significant improvements in reconstruction quality and outperforms existing CNN-based and GAN-based methods in similarity metrics. Cascading in the spectral dimension is a promising approach as it decomposes the complexity of spectral reconstruction. Moreover, as observed in Figure 5, this strategy may enhance the reconstruction quality in the infrared wavelength range, which is not covered by RGB images. This design may also have a positive effect on models with other architectures, and targeted research can be conducted in the future to explore this potential.

However, our model incurs significant time overhead. Firstly, the diffusion model requires hundreds or even thousands of sampling steps during inference. Secondly, the multi-stage cascading approach further multiplies the inference time. Thirdly, the ICMG strategy, which calculates both conditional and unconditional outputs, doubles the inference time. Therefore, accelerating the inference process of SCDM is a key area for future research.

VI. CONCLUSION

In this paper, we propose a novel spectral super-resolution method by stepwise refinement for reconstructing hyperspectral images from RGB remote sensing images, called SCDM. The pipeline of SCDM consists of multiple diffusion models, which progressively synthesize images with an increasing number of bands, starting from the RGB image, sequentially spectrally upsampling and gradually refining the spectral details. This design enhances the ability of diffusion models to predict high-dimensional noise. In order to make the conditional diffusion model better utilize the input conditions, we introduced ICMG, which reconstructs the hyperspectral image with a stronger correlation to the input conditions by interpolating the fraction of the results with and without the image condition inputs in total during the sampling process. To avoid the error caused by the instability of the sampling process, which achieved better results than employing fixed threshold truncation. Experiments show that our proposed SCDM method can reach the advanced level in the spectral super-resolution task.

VII. ACKNOWLEDGEMENT

The authors would like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory, University of Houston, for acquiring and providing the data used in this study and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

REFERENCES

[1] S.-E. Qian, "Hyperspectral satellites, evolution, and development history," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7032–7056, 2021.

[2] J. M. Meyer, R. F. Kokaly, and E. Holley, "Hyperspectral remote sensing of white mica: A review of imaging and point-based spectrometer studies for mineral resources, with spectrometer design considerations," *Remote Sensing of Environment*, vol. 275, p. 113000, 2022.

[3] J. Cai, J. Chen, X. Dou, and Q. Xing, "Using machine learning algorithms with in situ hyperspectral reflectance data to assess comprehensive water quality of urban rivers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[4] H. Yao, R. Chen, W. Chen, H. Sun, W. Xie, and X. Lu, "Pseudo-label-based unreliable sample learning for semi-supervised

hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
[5] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[6] S. Tilon, F. Nex, N. Kerle, and G. Vosselman, "Post-disaster building damage detection from earth observation imagery using unsupervised and transferable anomaly detecting generative adversarial networks," *Remote sensing*, vol. 12, no. 24, p. 4193, 2020.
[7] J. He, Q. Yuan, J. Li, Y. Xiao, D. Liu, H. Shen, and L. Zhang, "Spectral super-resolution meets deep learning: Achievements and challenges," *Information Fusion*, p. 101812, 2023.
[8] F. Agahian, S. A. Amirshahi, and S. H. Amirshahi, "Reconstruction of reflectance spectra using weighted principal component analysis," *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 33, no. 5, pp. 360–371, 2008.
[9] A. Robles-Kelly, "Single image spectral reconstruction for multimedia applications," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 251–260.
[10] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–34.
[11] X. Han, W. Leng, H. Zhang, W. Wang, Q. Xu, and W. Sun, "Spectral library based spectral super-resolution under incomplete spectral coverage conditions," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
[14] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 939–947.
[15] L. Zhang, Z. Lang, P. Wang, W. Wei, S. Liao, L. Shao, and Y. Zhang, "Pixel-aware deep function-mixture network for spectral super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 821–12 828.
[16] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7251–7265, 2021.
[17] S. A. Magid, Y. Zhang, D. Wei, W.-D. Jang, Z. Lin, Y. Fu, and H. Pfister, "Dynamic high-pass filtering and multi-spectral attention for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4288–4297.
[18] J. Li, S. Du, R. Song, C. Wu, Y. Li, and Q. Du, "Hasic-net: Hybrid attentional convolutional neural network with structure information consistency for spectral super-resolution of rgb images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[19] T. Li and Y. Gu, "Progressive spatial–spectral joint network for hyperspectral image reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
[20] R. Hang, Q. Liu, and Z. Li, "Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery," *IEEE Transactions on Image Processing*, vol. 30, pp. 7256–7265, 2021.
[21] S. Mei, Y. Geng, J. Hou, and Q. Du, "Learning hyperspectral images from rgb images via a coarse-to-fine cnn," *Science China Information Sciences*, vol. 65, pp. 1–14, 2022.
[22] X. Han, H. Zhang, J.-H. Xue, and W. Sun, "A spectral–spatial jointed spectral super-resolution and its application to hj-1a satellite images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
[23] R. Dian, Y. Liu, and S. Li, "Spectral super-resolution via deep low-rank tensor representation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[25] K. G. Lore, K. K. Reddy, M. Giering, and E. A. Bernal, "Generative adversarial networks for spectral super-resolution and bidirectional rgb-to-multispectral mapping," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019, pp. 926–933.
[26] L. Liu, S. Lei, Z. Shi, N. Zhang, and X. Zhu, "Hyperspectral remote sensing imagery generation from RGB images based on joint discrimination," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7624–7636, 2021.
[27] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
[28] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[30] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.
[31] S. Chai, L. Zhuang, and F. Yan, "Layoutdm: Transformer-based diffusion model for layout generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 349–18 358.
[32] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," *arXiv preprint arXiv:2207.00050*, 2022.
[33] 2018 IEEE GRSS Data Fusion Contest. Online: http://www.grss-ieee.org/community/technical-committees/data-fusion.
[34] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 6, pp. 1709–1724, 2019.
[35] J. Wang, Y. Lu, S. Wang, B. Wang, X. Wang, and T. Long, "Two-stage spatial-frequency joint learning for large-factor remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[36] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "Ttst: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Transactions on Image Processing*, 2024.
[37] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.
[38] K. Chen, W. Li, S. Lei, J. Chen, X. Jiang, Z. Zou, and Z. Shi, "Continuous remote sensing image super-resolution based on context interaction in implicit function space," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–16, 2023.

[39] Y. Yan, L. Zhang, J. Li, W. Wei, and Y. Zhang, "Accurate spectral super-resolution from single rgb image using multi-scale CNN," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 206–217.

[40] R. Wu, W.-K. Ma, X. Fu, and Q. Li, "Hyperspectral super-resolution via global-local low-rank matrix estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7125–7140, 2020.

[41] W. Chen, X. Zheng, and X. Lu, "Semisupervised spectral degradation constrained network for spectral super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[42] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2021.

[43] T. Okamoto and I. Yamaguchi, "Simultaneous acquisition of spectral image information," *Optics letters*, vol. 16, no. 16, pp. 1277–1279, 1991.

[44] J. Aeschbacher, J. Wu, and R. Timofte, "In defense of shallow learned spectral reconstruction from rgb images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 471–479.

[45] X. Zheng, W. Chen, and X. Lu, "Spectral super-resolution of multispectral images using spatial–spectral residual attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[46] J. Li, C. Wu, R. Song, W. Xie, C. Ge, B. Li, and Y. Li, "Hybrid 2-D-3-D deep residual attentional network with structure tensor constraints for spectral super-resolution of RGB images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2321–2335, 2021.

[47] J. Li, C. Wu, R. Song, Y. Li, W. Xie, L. He, and X. Gao, "Deep hybrid 2-d-3-d cnn based on dual second-order attention with camera spectral sensitivity prior for spectral super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[48] R. Dian, T. Shan, W. He, and H. Liu, "Spectral super-resolution via model-guided cross-fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[49] L. Tan, R. Dian, S. Li, and J. Liu, "Frequency-spatial domain feature fusion for spectral super-resolution," *IEEE Transactions on Computational Imaging*, 2024.

[50] J. He, J. Li, Q. Yuan, H. Shen, and L. Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[51] L. Liu, W. Li, Z. Shi, and Z. Zou, "Physics-informed hyperspectral remote sensing image synthesis with deep conditional generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[52] L. Liu, B. Chen, H. Chen, Z. Zou, and Z. Shi, "Diverse hyperspectral remote sensing image synthesis with diffusion models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[53] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[54] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photo-realistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[55] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[56] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[57] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[58] Y. Zhang, H. Liu, Z. Li, X. Gao, G. Shi, and J. Jiang, "Tcdm: effective large-factor image super-resolution via texture consistency diffusion," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[59] K. He, Y. Cai, S. Peng, and M. Tan, "A diffusion model-assisted multi-scale spectral attention network for hyperspectral image super-resolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[60] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[61] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[62] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[63] J. Menick and N. Kalchbrenner, "Generating high fidelity images with subscale pixel networks and multidimensional upscaling," *arXiv preprint arXiv:1812.01608*, 2018.

[64] X. Sun and J. Xu, "Remote sensing images dehazing algorithm based on cascade generative adversarial networks," in *2020 13th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*. IEEE, 2020, pp. 316–321.

[65] X. Liu, Y. Qiao, Y. Xiong, Z. Cai, and P. Liu, "Cascade conditional generative adversarial nets for spatial-spectral hyperspectral sample generation," *Science China Information Sciences*, vol. 63, pp. 1–16, 2020.

[66] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2249–2281, 2022.

[67] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[68] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[69] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[72] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

Bowen Chen received his B.S. degree from China University of Petroleum East China, Qingdao, Shandong, China, in 2022. He is currently working toward his doctor's degree in the School of Astronautics, Beihang University.

His research interests include remote sensing image processing and computer vision.

Liqin Liu received her B.S. degree and her PhD degree from the Image Processing Center, School of Astronautics, Beihang University in 2018 and 2024, respectively. She is currently a Postdoctoral Research Fellow at the Image Processing Center, School of Astronautics, Beihang University. Her research interests include hyperspectral image processing, remote sensing and deep learning.

Chenyang Liu received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently working towards the Ph.D. degree in the Image Processing Center, School of Astronautics, Beihang University.

His research interests include machine learning, computer vision and multimodal learning. His personal website is https://chen-yang-liu.github.io/.

Zhengxia Zou (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2013 and 2018, respectively. He is currently a Professor with the School of Astronautics, Beihang University. From 2018 to 2021, he was a Postdoctoral Research Fellow at the University of Michigan, Ann Arbor, MI, USA. His research interests include remote sensing image processing and computer vision. He has published more than 60 peer-reviewed papers in top-tier journals and conferences, including Nature Communications, Proceedings of the IEEE, IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, and IEEE/CVF Computer Vision and Pattern Recognition. Dr. Zou serves as an Associate Editor for IEEE Transactions on Image Processing (TIP). His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or coauthored over 200 scientific articles in refereed journals and proceedings. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Prof. Shi serves as an Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Pattern Recognition, ISPRS Journal of Photogrammetry and Remote Sensing, Infrared Physics and Technology, and etc.