Information Fusion 108 (2024) 102419

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

From the abundance perspective: Multi-modal scene fusion-based hyperspectral image synthesis

Erting Pan[1], Yang Yu[1], Xiaoguang Mei, Jun Huang *, Jiayi Ma

Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Hyperspectral image synthesis
Multi-modal fusion
Unmixing
Diffusion
Generative AI

ABSTRACT

Nowadays, data is of paramount importance for artificial intelligence. However, collecting real-world hyperspectral images (HSIs) with desired characteristics and diversity can be prohibitively expensive and time-consuming, leading to the data scarcity issue in HSI, and further limiting the potential of deep learning-based HSI applications. Existing work to tackle this issue fails to generate abundant, diverse, and reliable synthetic HSIs. This work proposes a multi-modal scene fusion-based method that diffusion from the abundance perspective for HSI synthesis, termed MSF-Diff. Concretely, highlights involve: (1) Synthesis in low-dimensional abundance space, other than original high-dimensional HSI space, greatly releases the difficulty; (2) Integration of multi-modal data greatly enriches the diversity of spatial distribution that the model can perceive; (3) Incorporation of the unmixing concept ensures that the generated synthetic HSI has reliable spectral profiles. The proposed research can generate a vast amount of HSI with a rich diversity in various categories and scenes, closely resembling realistic data. It plays a pivotal role in ensuring that the model produces reliable results and can be trusted for real-world applications. The code is publicly available at https://github.com/EtPan/MSF-Diff.

1. Introduction

Hyperspectral image (HSI), spanning the ultraviolet, visible, and infrared spectral ranges with high spectral resolution, has emerged as a powerful tool for target identification and anomalous material detection, among other applications. With the advent of artificial intelligence (AI), HSIs can bring about a technique reformation in a range of fields, including environmental monitoring, damage assessment, disease diagnosis, and agriculture [1–3]. However, the effective application of AI to HSI relies heavily on an adequate volume of high-quality data.

With the continuous advancement of sensor technology, more spaceborne and airborne hyperspectral imaging sensors have been significantly increased [4–6]. However, the following issues remain. First, as a result of the complex and sensitive imaging process along with the limitations inherent in the imaging system, real-world HSIs are largely imperfect, characterized by restricted resolution or various degradation, thereby impeding their practical value [7–9]. Second, capturing HSIs with hundreds of spectral bands demands specialized and expensive sensors beyond the scope of conventional imaging systems. Due to the huge expenses associated with such high-dimensional data collection, transmission, and storage, constructing a desirable large-scale HSI dataset still poses a considerable financial challenge [10,11]. Thirdly, current spaceborne sensors operating in routine revisit mode have the capability to collect vast amounts of HSIs. Nevertheless, when it comes to a specific task or application, most of these HSIs are irrelevant and meaningless, and sifting through all these data for useful ones can be laborious and time-consuming again [12,13]. Consequently, the availability of manually annotated HSI datasets is still limited, and they are often small, not very representative, extremely imbalanced, and potential inclusion of noisy labels [5,11,14]. The scarcity of high-quality and usable HSIs persists as a pressing concern [15,16].

To meet the ever-growing demands of large volumes and high-quality HSIs, researchers have been actively conducting investigations in various avenues [17–19]. Previous studies can be taxonomized into three categories: simulation, augmentation, and reconstruction. A succinct overview of the definition and representative works related to each category are outlined in Table 1. Simulation mainly depends on empirical or statistical models rooted in the physical imaging process to generate HSIs with radiometric and spatial accuracy [20–23]. It should be noted, however, that each simulation model is typically tailored

* Corresponding author.
E-mail addresses: panerting@whu.edu.cn (E. Pan), yuyang1995@whu.edu.cn (Y. Yu), meixiaoguang@gmail.com (X. Mei), junhwong@whu.edu.cn (J. Huang), jyma2010@gmail.com (J. Ma).
[1] Equal contribution.

https://doi.org/10.1016/j.inffus.2024.102419
Received 13 January 2024; Received in revised form 30 March 2024; Accepted 8 April 2024
Available online 10 April 2024

Table 1
Taxonomy of existing solution for data scarcity in HSIs.

Solution	Definition	Representative works	Drawbacks
Simulation	Model the physical imaging process with ray-tracing, radiative transfer, atmosphere calculation, *etc.*	DIRSIG [20], DART [21], BOA-TOA [22], NNE-S2 [23]	Designed for specific imaging system
Augmentation	Modify existing data with minor changes, *i.e.*, geometric transformations, random mask, *etc.*	Random Occlusion [24], SSDDA [25], FPGANDA [26], Flexible-Mixup [27]	Limited diversity
Reconstruction	Enhance/restore incomplete or degraded data, *i.e.*, limited spatial/spectral resolution, noisy corruption, missing information, *etc.*	MST++ [28], PoNet [29], 3DT-Net [30], STP-SOM [31], T3SC [32], SLDR [34]	Fail to generate new HSI samples

towards a specific imaging system, thereby restricting its applicability in a broader context. Augmentation techniques apply various transformations to the original data to expand the size of training HSI samples [24–27]. However, the augmented dataset is constrained by the original training data and its quality and diversity. Reconstruction methods recreate each HSI based on corresponding imperfect real-world data to improve the quality and usability of the given data [28–33]. Regrettably, reconstruction-based approaches fall short of generating entirely new HSIs. More detailed literature reviews can refer to Section 2.1.

On the other hand, with the rapid development of AI, generative AI offers a promising solution to this data scarcity issue in the realm of HSI [35]. Various techniques, such as Variational Autoencoder (VAE) [36,37], Generative Adversarial Network (GAN) [38,39], Normalizing Flow (Flow) [38,39], and the advanced Denoising Diffusion Probabilistic Model (DDPM) [40–42], have been extensively explored. These methodologies have demonstrated the capability to generate synthetic data in different modalities, including text, audio, and image. Nevertheless, the high dimensionality of spectral signatures in HSI poses a significant barrier to the effective utilization of generative AI methods for data synthesis. The complexity of models escalates proportionally with the image dimensions, making the synthesis of HSIs a challenging and computationally demanding task. Furthermore, the training of robust and dependable generative AI models mandates a substantial quantity of high-quality training data, a requirement that current HSI datasets are unable to meet. Consequently, research in this particular domain has been constrained by the complexities linked to high-dimensional data and the limited availability of HSI datasets.

To this end, this work shifts the perspective of HSI synthesis from the high-dimensional cube to the low-dimensional abundance and incorporates more easily accessible RGB images. A novel multi-modal scene fusion method that diffusion from the abundance, termed as MSF-Diff, is proposed for HSI synthesis. MSF-Diff organizes a three-phase pipeline that includes scene-based unmixing, abundance-based diffusion, and fusion-based generation to generate large volumes of high-quality HSIs that exhibit satisfactory diversity and reliability, closely resembling real-world data in terms of various categories and scenes. In concrete, MSF-Diff firstly leverages the concept of HSI unmixing, which decomposes individual HSI cubes into endmembers that hold physical significance, along with corresponding abundance maps that delineate their spatial distributions. Based on this, we introduce external RGB images that cover similar scenes and propose scene-based unmixing across multi-modal data to extract reliable shared endmembers from real-world HSIs and map HSI and RGB to the same low-dimensional abundance space. Next, with abundant and various spatial distributions from real-world RGB images, we train an abundance-based diffusion model to generate synthetic abundance maps. Finally, we fuse the shared endmembers estimated from HSI and synthetic abundance maps generated based on RGB to produce new synthetic HSI samples. The proposed MSF-Diff provides a more comprehensive and accurate method for HSI synthesis, offering a promising avenue for advancing research in the field.

The contributions can be summarized below.

- Incorporate the unmixing process and construct scene-based unmixing, empowering the perspective of HSI synthesis from the high-dimensional cube shifting to the low-dimensional abundance and deriving endmembers with reliable physical signatures.
- Integrate external RGB images and design abundance-based diffusion, providing the model with sufficient information about the distribution of real scenes in an economical and reliable manner and generating a large volume of synthetic abundance.
- Fuse the shared endmembers estimated from HSI and synthetic abundance maps generated based on RGB, producing new synthetic HSI samples with reliable spectral profiles and reasonable spatial distributions.
- To the best of our knowledge, there is no existing work for multi-modal fusion-based HSI synthesis. The underlying benefits of overcoming this matter are remarkable, as it has the potential to revolutionize a range of fields.

2. Related work

2.1. Existing hyperspectral image synthesis techniques

As listed in Table 1, existing techniques involve simulation, augmentation, and reconstruction. Each category has its unique characteristics and limitations.

Simulation-based methods mainly simulate airborne or spaceborne imaging systems using empirical or statistical models. For instance, digital imaging remote sensing image generation (DIRSIG) tool utilizes an end-to-end ray tracing physical model to calculate scattered radiance by analyzing the surface properties of objects. It can be configured to simulate the functionality of the airborne hyperspectral instrument AVIRIS [20]. Discrete anisotropic radiative transfer (DART) model accounts for BRDF effects in Earth-atmosphere radiation interaction within heterogeneous 3-D scenes, enabling forward simulations of satellite reflectance images and LiDAR [21]. The neural network emulator learns statistically the nonlinear relationships between the hyperspectral airborne sensor HyPlant and multispectral satellite Sentinel-2 (S2) to produce a realistic hyperspectral S2-like datacube [23]. Significantly, a huge amount of time and effort is required to construct the appropriate setting and computational capacity essential for the precise rendering of synthetic HSIs. Furthermore, the efficiency of these models is greatly reliant on specific essential parameters, necessitating the need to parameterize models tailored to different imaging systems.

Augmentation-based techniques have been frequently utilized in numerous studies, particularly those addressing the challenge of limited labeled samples in enhancing the performance of machine learning models for HSI classification or target recognition. These techniques typically involve the modification of existing HSI samples through geometric transformations such as scaling, rotation, and shearing [25]; spatial transformations including random occlusion or noise [24]; as well as the mixup of two or more strategies [26,27]. Despite effectively increasing the data volume, augmentation often lacks in augmenting the dataset's diversity or variety. This limitation stems from the fact that the transformations implemented do not introduce novel features

or attributes but instead only modify existing ones, resulting in an over-emphasis on certain features while neglecting others. Hence, while such augmentations can offer a temporary remedy to data scarcity in quantity, their capacity to enhance the overall quality and diversity of HSIs remains constrained.

Reconstruction-based methods are often anchored in the enhancement or restoration of existing data. Largely real-world data exhibits imperfections, such as limited spatial or spectral resolution due to the imaging system, or degradation from low-light or noisy conditions. Research efforts have been made in areas such as spectral superresolution [28,29], spatial superresolution [30], image enhancement [31,33], and denoising [34]. Undoubtedly, these endeavors have led to advancements in the quality and usability of the data at hand. Nevertheless, these methods come with inherent limitations. Primarily, they may struggle to fully restore the original data's spectral signatures, particularly with materials possessing complex or unique spectral characteristics. This can result in synthetic HSIs lacking reliability. Additionally, these methods are not inherently capable of producing entirely new HSI samples, thus providing only limited assistance in diversifying hyperspectral datasets.

2.2. Generative AI-based image synthesis

As a trending area, image synthesis has gained significant attention over the past decade. The primary goal is to create new, realistic images or modify existing ones while preserving natural textures and nuances. In the last few years, it has witnessed very impressive progress thanks to the advance of generative AI, especially deep generative models [43,44].

GANs [45–48] have been at the forefront of the image synthesis revolution. GANs consist of two neural networks, a generator and a discriminator, which compete against each other in a zero-sum game scene. The generator's goal is to create images that the discriminator cannot distinguish from real images, thereby enhancing the generator's ability to produce high-quality synthetic images. The discriminator, on the other hand, attempts to distinguish the synthetic images from the real ones. While GANs have produced impressive results in terms of the perceptual quality of synthetic images, they are known for their training instability. Balancing the generator and discriminator can be a significant challenge as an overly powerful discriminator can cause the generator to fail, and vice versa. As a result, GANs often require careful hyperparameter tuning and can exhibit mode collapse where they generate a limited variety of images.

VAEs [49–51] represent another significant advancement in the field of image synthesis. VAEs belong to a class of generative models that are trained to learn the underlying probability distribution of the training data, allowing them to generate new data points with similar properties. VAEs have shown promise in terms of optimization performance and distribution estimation. They provide a lower-bound estimate of the data likelihood, which can be optimized directly. However, VAEs have been criticized for their tendency to produce blurry images, which is often attributed to the use of a pixel-wise reconstruction loss.

More recently, diffusion models [52,53] have emerged as a powerful tool for image synthesis. These models capture the data distribution by modeling it as a diffusion process, which gradually transforms a simple initial distribution into the target distribution. Diffusion models offer several desirable properties, such as a clear training objective, model stability, and easy extensibility. They have achieved state-of-the-art results in various image synthesis tasks, outperforming both GANs and VAEs. However, training diffusion models can be computationally intensive and slow due to the iterative nature of the diffusion process. To address the computational challenges associated with training diffusion models, some researchers have proposed a two-stage synthesis framework [54,55]. They combine a compression stage, which projects raw images into a lower-dimensional latent space, with a generative stage, which generates new images from the latent representations. While this approach can significantly reduce computational complexity, the quality of the synthetic images depends heavily on the performance of both the compression and generative models. Any deficiencies in either model can negatively impact the final result.

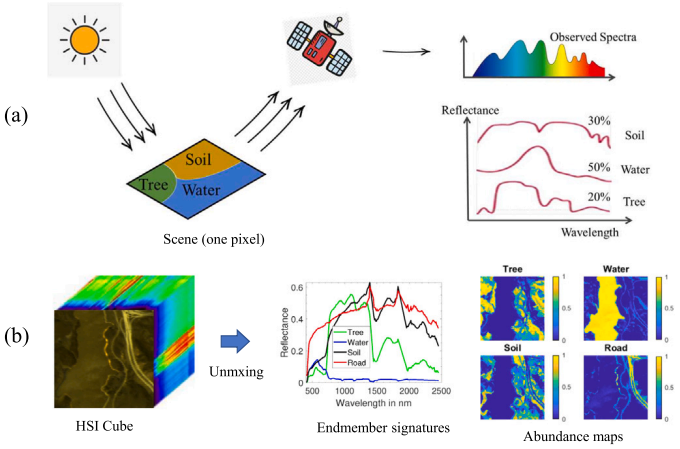


Fig. 1. (a) Illustration of the formation of a mixed pixel in HSI imaging. (b) Illustration of decomposing HSI into endmember signatures and abundance maps by unmixing.

While deep learning and generative AI have been successful in synthesizing RGB images, their application in HSI synthesis remains largely unexplored. This gap indicates a substantial potential for breakthroughs in this field. The generation of high-quality HSIs is a challenging task, necessitating innovative research efforts. Given the limited progress and inherent challenges in existing HSI synthesis techniques, our research aims to leverage the strengths of deep learning and generative AI. We strive to develop novel techniques tailored for HSI synthesis, aiming to enhance the diversity of HSIs, streamline the synthesis process, and improve the quality of synthetic images. Our research will provide a valuable contribution to the field of HSI synthesis, potentially leading to significant advancements.

3. Motivation

This work is motivated by HSI unmixing. As depicted in Fig. 1(a), scene information is collected by remote sensing HSI imaging systems. Typically, the corresponding ground sampling distance, which is the distance between the centers of two neighboring pixels measured on the ground, can easily reach tens of meters. Consequently, one pixel in the HSI may consist of several substances, forming a mixed pixel with an observed spectral curve. It has brought about the emergence of HSI unmixing [56,57], which, as depicted in Fig. 1(b), decomposes the scene into multiple endmember signatures and their corresponding abundance maps. In light of this, HSI unmixing, by quantifying the various substances (endmembers) and their corresponding spatial distribution (abundances) that constitute a given scene, stands out as an immensely valuable technique for reducing data dimensionality. Moreover, the endmembers and abundances derived from HSI unmixing possess discernible physical meanings.

In this work, we endeavor to generate new HSIs by combining endmembers estimated by unmixing and synthetic abundances generated by abundance-based diffusion. It offers two significant advantages: (1) By reducing the dimension of the generation space from hundreds of spectral channels to an infinite number of endmembers, it alleviates the high-dimensional dilemma and decreases the difficulty of HSI diffusion. (2) Using endmembers estimated from real HSIs ensures the reliability of the spectral profile of the synthetic HSIs.

Besides, as previously mentioned, off-the-shelf HSI datasets are plagued by skewed class distribution [58,59], leading to overfitting on

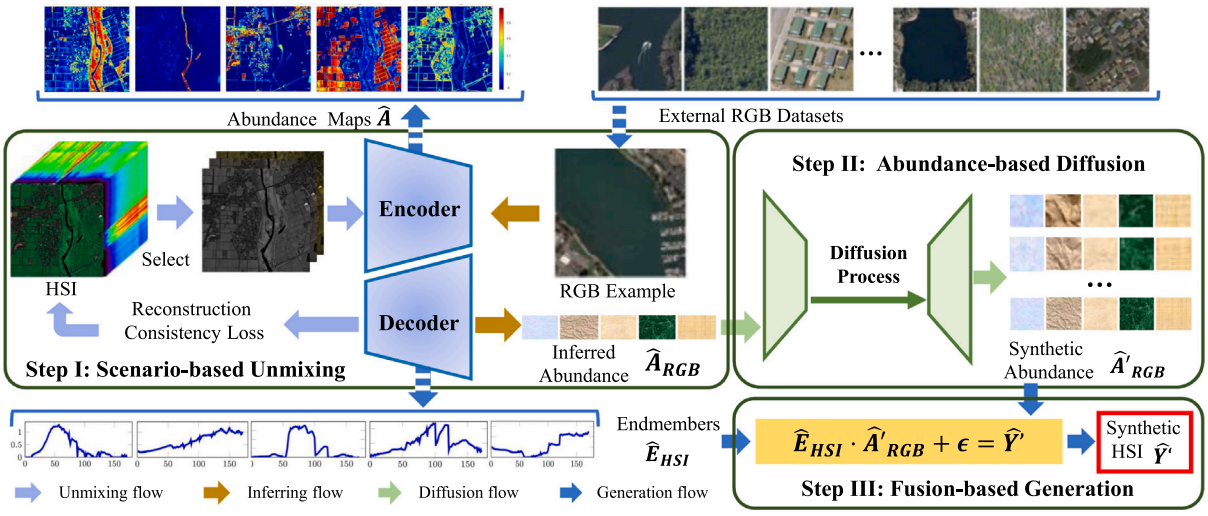


Fig. 2. The proposed pipeline for generating synthetic HSIs using a multi-modal scene fusion-based method. The pipeline involves three main steps: scene-based unmixing with training via the unmixing flow and inferring via the inferring flow, abundance-based diffusion via the diffusion flow, and fusion-based generation via the generation flow.

majority classes and inferior performance on minority classes. Moreover, if biases exist in the original datasets, any new data generated from them will inherently perpetuate these biases. Intuitively, to address this issue, this study proposes integrating external datasets that contain more diverse and accessible natural RGB images. By assuming that images of different modalities in similar scenes share representative endmembers but with characterized abundances, the study proposes a scene-based unmixing to infer the abundance maps of RGB images. With auxiliary RGB datasets, the proposed method can perceive a large and diverse object distribution. Combined with the aforementioned abundance-based diffusion model, it can generate diversified synthetic abundances, alleviating the issue of *limited diversity*.

Overall, our approach combines scene-based unmixing, abundance-based diffusion, and multi-modal fusion to generate new HSIs with improved quantity, diversity, and reliability.

4. Proposed method

In general, it is expected that the synthetic new HSI should accurately represent the spatial characteristics of objects while maintaining the physical meanings of their spectral signatures. To this end, we propose a solution outlined in Fig. 2, which involves three steps, scene-based unmixing, abundance-based diffusion, and fusion-based generation. Further details on this approach can be found below.

4.1. Scene-based unmixing

Our method starts with unmixing. A common solution for HSI unmixing follows the spectral linear mixture model (LMM) theory [60], formulating as:

$$Y = E \cdot A + \epsilon, \tag{1}$$

where $Y \in \mathcal{R}^{C\times W\times H}$ indicates the observed HSI, $E \in \mathcal{R}^{C\times k}$ represents typical endmembers, k symbolizes the number of E, $A \in \mathcal{R}^{k\times W\times H}$ is their corresponding fractional abundance maps, and $\epsilon \in \mathcal{R}^{C\times W\times H}$ is the bias item.

Uniquely, this study introduces a novel idea, *i.e.,* scene-based unmixing, which assumes that different images of similar scenes can be represented by a finite number of fixed endmembers and tailored abundance maps. It should be noted that the term *endmembers* refers to typical compositional constituents in a given scene, rather than pure pixels composed of a single material in traditional unmixing.

Despite the effectiveness of traditional HSI unmixing networks, their applicability is constrained by their unsupervised training on a single HSI. It poses challenges for the scene decomposition of other modal images with varying numbers of spectral channels. Accordingly, we tailor the process for scene-based unmixing, as shown in Fig. 2. To elaborate, we first assume that the quantity k of endmembers is known and fixed. Then, we extract representative three-band data (corresponding to RGB) Y_{bs} from the HSI through band selection. Subsequently, we design an encoder $\mathcal{G}_E$ to infer the abundance and a decoder $\mathcal{G}^D$ to reconstruct the original HSI. It is imperative to note that the reconstruction target is $\hat{Y}$ and not Y_{bs}. It can be written as:

$$\hat{Y} = \mathcal{G}_D(\mathcal{G}_E(Y_{\text{bs}})), \tag{2}$$

where $\hat{Y} \in \mathcal{R}^{C\times W\times H}$ indicates the reconstructed HSI, and $Y_{\text{bs}} \in \mathcal{R}^{3\times W\times H}$ is the input of encoder. The fractional abundance maps $\hat{A}$ should be governed by the abundance non-negative constraint (ANC) and the abundance sum-to-one constraint (ASC). Hence, our $\mathcal{G}_E$ is composed of several residual spectral attention modules (RSA) [61] and ends up with a softmax layer. The decoder contains a 1×1 convolutional layer, simulating the linear mixing model (refer to Eq. (1)), of which the weights represent the extracted endmembers $\hat{E}_{\text{HSI}}$. The decoding process also can be formulated as :

$$\hat{Y} = \mathcal{G}_D(\hat{A}) = \hat{E}_{\text{HSI}} \cdot \hat{A} + b, \tag{3}$$

where b is the bias item. Such a design ensures that the output of the encoder accurately represents the spatial distribution of abundances while enabling the weights of the decoder to represent endmembers with physical significance.

The loss function of the proposed unmixing framework comprises three parts: the mean absolute error (MAE) loss $\mathcal{L}_{\text{MAE}}$ to ensure the pixel-wise reconstruction accuracy, the spectral angle distance (SAD) loss $\mathcal{L}_{\text{SAD}}$ to govern the fidelity of spectral signatures, and the endmembers total variation (ETV) loss $\mathcal{L}_{\text{ETV}}$ to preserve the spectral smoothness of the extracted endmembers. The total loss function $\mathcal{L}$ can be expressed as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{MAE}} + \alpha \cdot \mathcal{L}_{\text{SAD}} + \beta \cdot \mathcal{L}_{\text{ETV}} \\ &= \|Y, \hat{Y}\|_1 + \alpha \cdot \arccos\left(\frac{(Y, \hat{Y})}{\|Y\|_2 \|\hat{Y}\|_2}\right) + \beta \cdot \sum_{i}^{C} (e_{i+1} - e_i),\end{aligned} \tag{4}$$

where e_i represents the value of i_{th} band in each spectral vector of endmembers, and α and β are used to balance convergency for each item. α and β are setting as 0.1 and $1e{-}3$, empirically. By optimizing our network using this loss function, we can ensure stable and desirable unmixing results to a significant degree.

Our scene-based unmixing method possesses broad applicability across modalities, as similar scenes are also constructed from typical

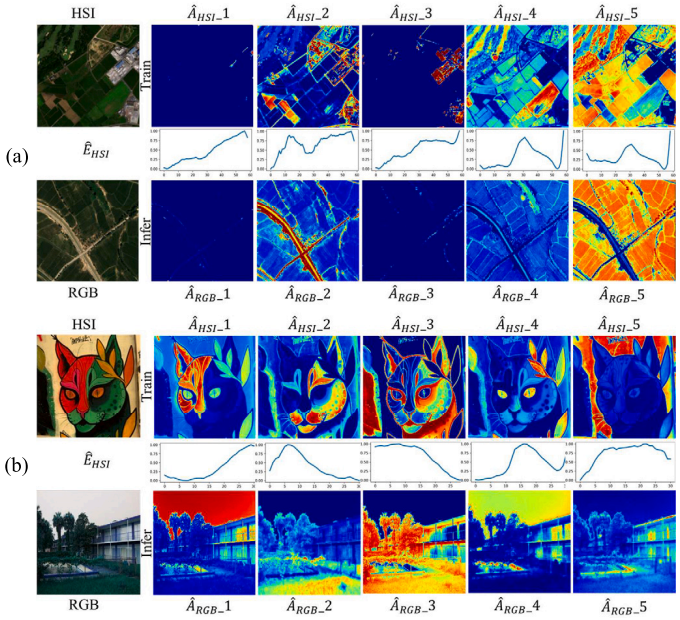


Fig. 3. Example results of scene-based unmixing from training to inference. (a) Remote sensing city scene: trained on Chikusei dataset, and inferred on AID dataset; (b) Mixed ground scene: trained on ARAD dataset, and inferred on Places 2 dataset.

Algorithm 1: Training for Abundance-based Diffusion

Input: Inferred abundances $\hat{A}_{\text{RGB}} \in \mathcal{R}^{k\times W\times H}$

// k aligns with the number of endmembers.

while *not converged* **do**

  $\hat{\mathbf{A}}^0 \sim q(\hat{\mathbf{A}}_{RGB})$;

  $t \sim \text{Uniform}(\{1,\dots,T\})$;

  $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$;

  Take gradient descent step on

  $\nabla_\theta \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\hat{\mathbf{A}}^0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \right\|^2$;

end

compositional components that exhibit different spatial patterns, represented through exclusive fractional abundance maps. Fig. 3 depicts some example results of our method, which validate our claim. Next, leveraging the well-trained unmixing network, we employ it to external RGB datasets that conform to the same scene category, to infer their abundance, denoted as $\hat{A}_{\text{RGB}}$.

4.2. Abundance-based diffusion

Our proposed abundance-based diffusion is built on the denoising diffusion probabilistic model (DDPM) [40]. As illustrated in Fig. 4, it mainly involves two processes, the forward process, and the reversal process. The forward process involves the incremental addition of Gaussian noise to the input abundance $\hat{A}_{RGB} = \{\hat{A}^0_0,\dots,\hat{A}^0_5\}$ until it resembles random noise $\{\hat{A}^T_0,\dots,\hat{A}^T_5\} \sim \mathcal{N}(0,\mathbf{I})$. In contrast, the reversal process is trained to progressively denoise the output. A well-trained denoising network is then utilized to generate synthetic abundance $\hat{A}'_{RGB}$ from a random distribution of noise.

It is imperative to recognize the distinguishing feature of our proposed abundance-based diffusion model when compared to other models, such as DDPM or LDM. The major distinction lies in the feature space used during the diffusion process. As we have emphasized earlier, the abundance space is a far superior choice when compared to diffusing in the original HSI cube space or the latent feature space. The abundance serves as an exceptionally effective low-dimensional representation, significantly reducing the computational burden of diffusing

Algorithm 2: Sampling for Abundance-based Diffusion

Output: Synthetic abundances $\hat{A}'_{\text{RGB}} \in \mathcal{R}^{k\times W\times H}$

$\hat{\mathbf{A}}^T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$;

for $t = 1$ *to* T **do**

  $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$;

  $\hat{\mathbf{A}}^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{A}}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{\mathbf{A}}^t, t)\right) + \sigma_t\mathbf{z}$;

end

return $\hat{A}'_{\text{RGB}} = \text{softmax}(\hat{A}^0)$ // softmax(·) for constraints of ANC and ASC.

in the high-dimensional cube space. Furthermore, unlike the latent feature, abundance holds a clear physical significance as it precisely describes the spatial distribution of each endmember. Noteworthy, unlike RGB images with three channels where the values within each channel are largely independent of each other, the abundance matches the number of channels with the count of endmembers and must strictly adhere to the constraints of ASC and ANC. Therefore, when we shift our focus to diffusion in the abundance space, it becomes essential to consider and adhere to these two critical constraints. As such, we introduce a softmax mapping before concluding the sampling process.

Both the training and sampling procedures of our abundance-based diffusion for synthetic abundance generation are detailed in Algorithms 1 and 2.

4.3. Fusion-based synthetic HSI generation

The generation of synthetic HSI is a fusion process of multi-modal information, *i.e.*, synthetic abundance maps from external RGB images $\hat{A}'_{\text{RGB}}$ and endmembers estimated from HSIs $\hat{E}_{\text{HSI}}$. The fusion rule still adheres to the original LMM as in Eq. (1), which can be reformulated as:

$$\hat{Y}' = \hat{E}_{\text{HSI}} \cdot \hat{A}'_{\text{RGB}} + \epsilon, \tag{5}$$

where $\hat{Y}' \in \mathcal{R}^{C\times W\times H}$ indicates the synthetic HSI, $\hat{E}_{\text{HSI}} \in \mathcal{R}^{C\times k}$ represents the estimated endmembers, and $\hat{A}_{\text{RGB}} \in \mathcal{R}^{k\times W\times H}$ symbolizes the generated abundance maps.

The proposed approach of integrating synthetic abundance maps and estimated endmembers for generating synthetic HSI is novel and robust. It effectively leverages the strengths of both components, resulting in a more accurate and representative synthetic HSI. The use of the reformulated LMM model ensures that the fusion process remains consistent with proven theoretical models, further enhancing the reliability and validity of the generated synthetic HSIs. Examples of synthesized HSIs in different scenes are presented in the following experiments.

5. Experiments

We present a comprehensive evaluation of the proposed method through a range of experimental procedures, including ablation, comparative, and expansion experiments.

5.1. Experimental settings

5.1.1. Datasets

This study employed multiple scenes to evaluate the robustness and generalizability of the proposed method. In remote sensing scenes, we trained the unmixing model using the Chikusei HSI dataset,[2] validated the unmixing model by the HSRS-SC dataset [11], and subsequently employed the AID natural scene classification dataset,[3] which exhibits similar scenes, for abundance inference and diffusion. Similarly, in the

[2] https://naotoyokoya.com/Download.html

[3] https://captain-whu.github.io/AID/

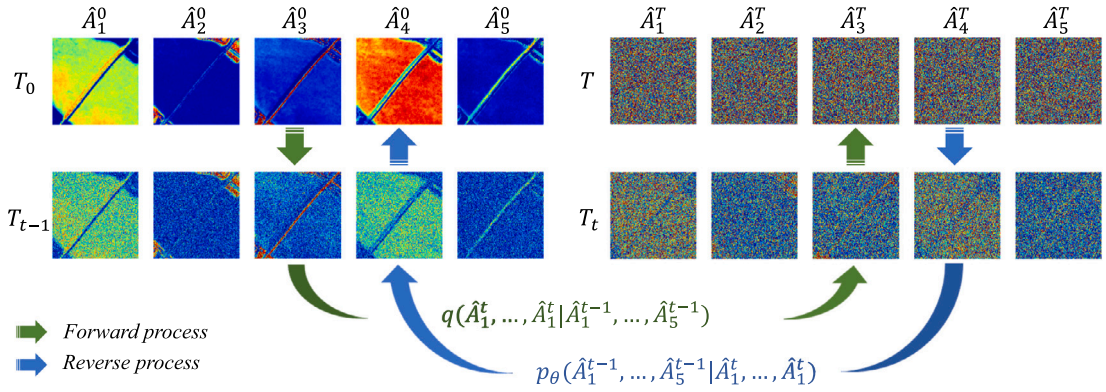
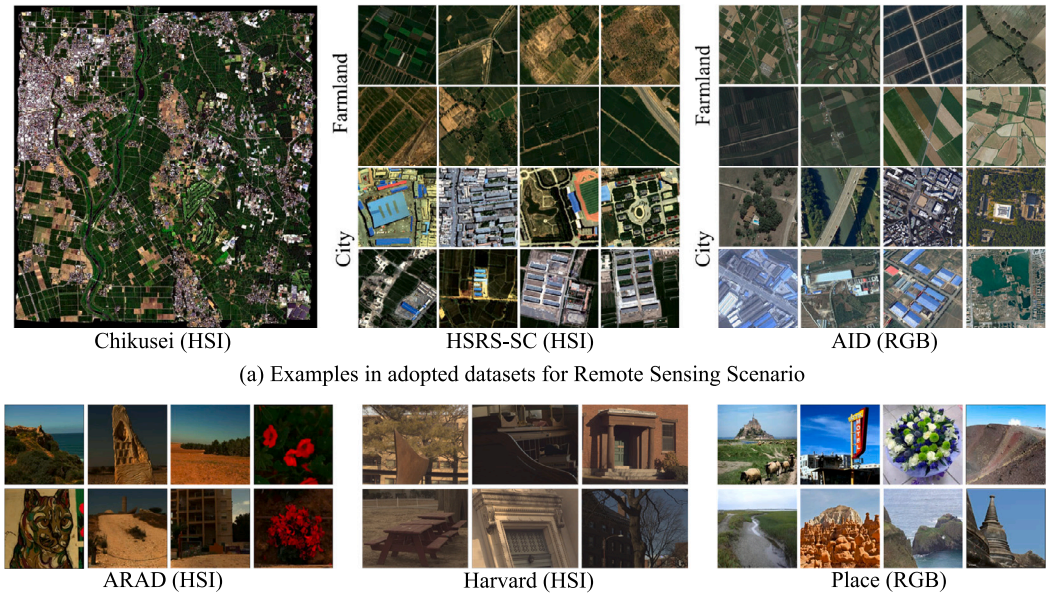


Fig. 4. The Markov chain of forward (reverse) diffusion process of generating synthetic abundance by slowly adding (removing) noise abundance-based diffusion process.



(a) Examples in adopted datasets for Remote Sensing Scenario

(b) Examples in adopted datasets for Natural Scenario

Fig. 5. RGB rendering examples of adopted datasets for different cases.

Table 2
Detailed description of datasets adopted for HSI synthesis in the remote sensing scene.

Tasks	Training the scene-based unmixing	Validation the scene-based unmixing	Training the abundance-based diffusion
Datasets	Chikusei (HSI)	HSRS-SC (HSI)	AID (RGB)
Imaging Sensor	The Headwall Hyperspec-VNIR-C imaging sensor	The Compact airborne spectro-graphic imager, CASI	The Google Earth imagery
Spectral Range	363-1018nm→398-698 nm	380-1050nm→398-698 nm	–
Spectral Resolution	128→59	48→59	3
Spatial Resolution	2.5 m	1 m	0.5-8 m
Patch Size	2517 × 2335 × 128→128 × 128 × 59	256 × 256 × 59	256 × 256 × 3
Samples	1 HSI→840 patches	700	1902

mixed ground scene, we utilized the ARAD hyperspectral dataset,[4] Harvard dataset,[5] in conjunction with the Place 2[6] natural image dataset. The RGB rendering examples of these adopted datasets for difference cases are presented in Fig. 5, and the detailed settings for the remote sensing scenes are list in Table 2.

[4] https://codalab.lisn.upsaclay.fr/competitions/721
[5] http://vision.seas.harvard.edu/hyperspec/
[6] http://places2.csail.mit.edu/

5.1.2. Metrics

The advance of synthesis HSI generation can be primarily evaluated in terms of diversity and reliability. The former is largely dependent on subjective assessments, while the latter relies on the quality of the generated HSI and the distinction of the synthesis abundance. Since the generated data lacks reference, no-reference quality evaluation metrics, such as Fréchet Inception Distance (FID), precision, and recall, are employed. Furthermore, in an ablation study to further verify the reliability of the abundance inferred from RGB, quantitative evaluation metrics, such as root mean square error (RMSE), peak signal-to-noise

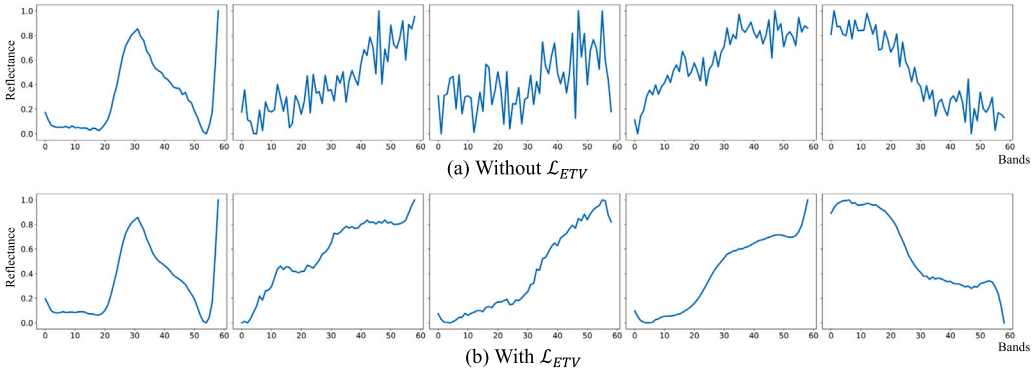


Fig. 6. Illustration of spectral profiles of the extracted endmembers under with/without $\mathcal{L}_{\text{ETV}}$.

ratio (PSNR), structural similarity index (SSIM), spectral angle distance (SAD), are also introduced.

5.1.3. Implementation details

All experiments are conducted using two NVIDIA 3090 GPUs. For scene-based unmixing, we assume a scene composed of 5 endmembers. The encoder is built using a stack of RSA modules, each containing 3×3 convolutional kernels and spectral channel attention with 1×1 convolution. The corresponding spectral channels of each layer are varied in $[3, 32, 64, 128, 96, 48, 5]$, where 3 refers to the channels of input after band selection and 5 represents the number of endmembers, corresponding to the quantities of the estimated abundance maps. Then, the decoder employs a linear layer composed of 1×1 convolutions to reconstruct HSIs. For abundance-based diffusion, we set the variance to increase linearly from $1e{-}6$ to $1e{-}2$ with a step size of 2000 during the forward process. We construct a denoising U-Net with depth multipliers of $[1, 2, 2, 4, 4]$ for the reversal process. During training, we set the initial learning rate to $1e{-}4$ and adopt the Adam optimizer. The scene-based unmixing network is trained for a total of 40 epochs, while the diffusion model is trained for 2M steps with a batch size of 8 to ensure model convergence.

5.2. Ablation study

5.2.1. Endmember TV Loss function in the proposed scene-based unmixing

Traditional unmixing networks often employ endmember extraction methods, such as VCA, to obtain the initial endmembers of the HSI, initializing the weights of the unmixing decoder. However, it requires additional processing, which not only makes network training non-end-to-end but also causes error accumulation. Hence, we do not initial the proposed scene-based unmixing in this manner. Additionally, to avoid spectral spikes due to inadequate initialization, we introduce the spectral total variation loss $\mathcal{L}_{\text{ETV}}$ to preserve the spectral smoothness of the extracted endmembers. As depicted in Fig. 6(a), without $\mathcal{L}_{\text{ETV}}$, an unstable initialization results in significant spectral spikes and distortion in the extracted endmembers. Consequently, these endmembers fail to accurately capture the fundamental spectral features of the scene. By contrast, integrating $\mathcal{L}_{\text{ETV}}$, (refer to Fig. 6(b)), effectively mitigates the spectral spikes present in the extracted endmembers. Incorporating a TV constraint enhances the smoothness of these sharing endmembers, thereby promoting a more coherent and reliable representation of the scene's spectral characteristics.

5.2.2. HSI reconstruction based on traditional AE or the proposed scene-based unmixing

The success of the proposed HSI synthesis solution relies on the performance of the reconstructed HSI method. To assess the efficacy of our approach, we conduct an ablation study comparing HSI reconstruction based on traditional Autoencoder (AE) and scene-based unmixing, as depicted in Fig. 7. The traditional AE here employs a symmetric U-net structure, while the unmixing network, as previously mentioned, has an asymmetric structure due to the introduction of the LMM, and its decoder only included a linear layer. To ensure a fair comparison, the encoders in both models are consistent, and the latent features in AE and the abundance in the unmixing model have the same number of channels. The comparative results in Fig. 7 demonstrate the significant advantage of the unmixing model. From a feature perspective, it is evident that abundance can better characterize the spatial distribution of specific materials with more explicit physical meaning. Additionally, based on the reconstruction performance, the HSI reconstructed by our scene-based unmixing has superior fidelity in both the spatial and spectral dimensions.

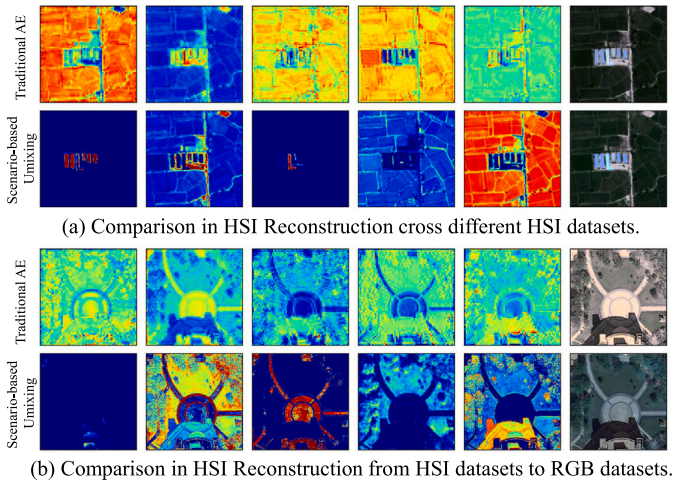


Fig. 7. Illustration of HSI reconstruction comparison based on traditional AE or our scene-based unmixing. The validation data comes from (a) the HSRS dataset (HSI) and (b) the AID dataset (RGB), respectively.

5.2.3. Reliability analysis for inferred abundances of RGB datasets

To demonstrate the reliability of inferred abundances of external datasets by the scene-based unmixing network, we introduce two additional hyperspectral imaging (HSI) datasets: the Harvard dataset[7] with the size of $1300 \times 1300 \times 31$ and the HSRS-SC hyperspectral scene classification dataset [11] with the size of $256 \times 256 \times 48$. Subsequently, we selectively extract their RGB bands, infer their abundances, and present a qualitative and quantitative assessment of the

[7] http://vision.seas.harvard.edu/hyperspec/explore.html

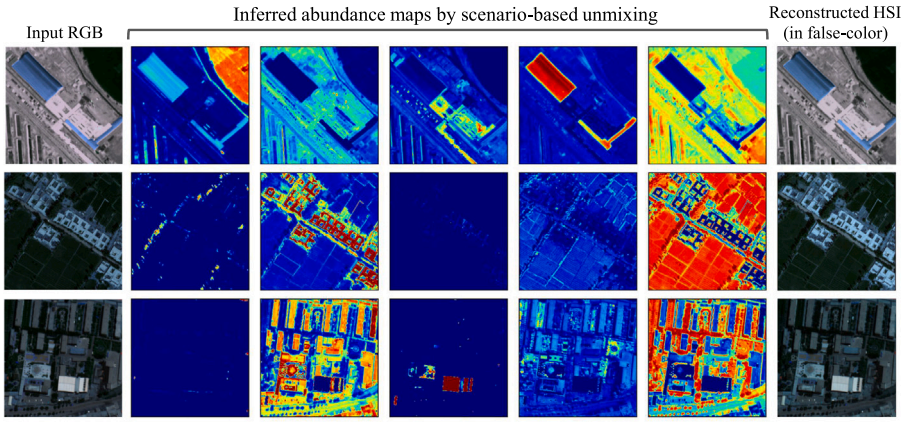


Fig. 8. Illustration of inferred abundance maps and reconstructed HSIs on HSRS-SC dataset.

Table 3
Quantitative evaluation for scene-based unmixing on cross different HSI datasets in typical scenes.

Senarios	RMSE	PSNR	SSIM	SAD
Remote Sensing City[a]	0.034	34.26	0.93	5.47
Remote Sensing Framland[b]	0.035	32.01	0.93	6.59
Natural scene[c]	0.031	33.61	0.94	7.75

[a] Chikusei (Training) → HSRS-SC city (Validation).
[b] Chikusei (Training) → HSRS-SC agriculture (Validation).
[c] ARAD(Training) → Harvard (Validation).

reconstructed HSIs in Fig. 8 and Table 3. The quantitative results are evaluated on average of 350 HSIs on the HSRS-SC dataset for each remote sensing scene and 45 HSIs on the Harvard dataset for the natural scene. Accordingly, although the training and validation are conducted cross-datasets, the vision quality of the inferred abundances and the reconstructed HSIs by our proposed method reveals robust and reliable performance in different scenes. It empowers an objective evaluation of the generalization performance of the proposed unmixing paradigm across multi-modal images.

5.2.4. HSI generation by diffusion in different feature spaces

We undertook a comparative analysis of the performance of the abundance-based diffusion model against the original methods that were based on the original HSI and latent features, as shown in Fig. 9. The original HSI-based diffusion model was trained using the Chikusei dataset, which contained 128 bands. However, due to the high dimensionality of this dataset, we faced significant challenges in generating meaningful images. Indeed, even after making 5 million generator iterations, the resulting image was still nothing more than meaningless random noise, as shown in Fig. 9(a). This, unfortunately, made it impossible for us to effectively evaluate the quality of the generated images in Table 4. Upon completing the training, the generation time of the abundance map is virtually unaffected by the number of endmembers. The generation time for a 256 × 256 size image requires approximately 90 s.

In an attempt to alleviate the so-called curse of dimensionality, we utilize the RGB dataset AID and a traditional U-net-like AE to derive latent features. This latent feature-based diffusion does show some improvement over the original method. However, according to results shown in Table 4, despite taking about 2.8 million diffusion steps, the quality of the synthesized latent features and the reconstructed HSIs was still far from satisfactory. Furthermore, the latent features lack physical meaning, and their instability poses a significant hindrance to the effective reconstruction and generation of high-quality HSIs. As shown in Fig. 9(b), the spectral curves of the generated HSIs exhibit significant distortion.

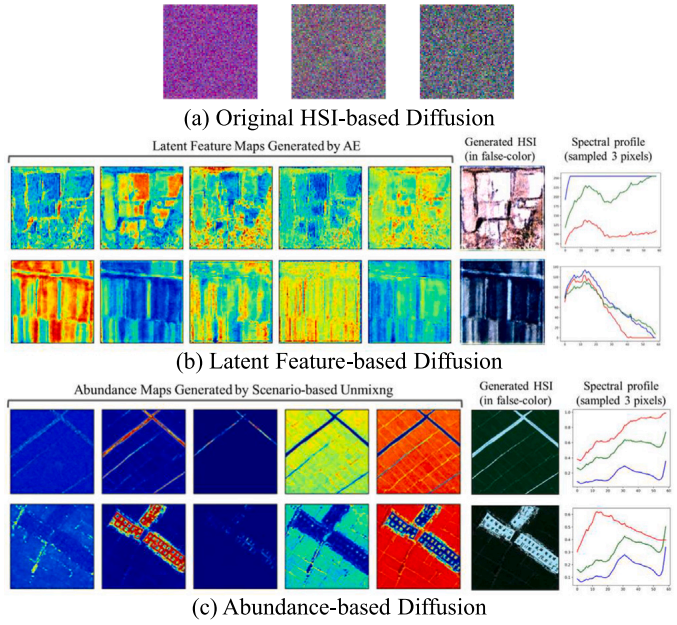


Fig. 9. Illustration of HSI generation comparison on diffusion in different feature spaces.

On a more encouraging note, the diffusion in the abundance space significantly mitigated these problems. Our newly proposed abundance-based diffusion model succeeded in generating high-quality HSIs that not only had physical significance but also adopted a more realistic style, as shown in Fig. 9(c). By exploiting the abundance, we are able to navigate the high-dimensional space and generate high-quality HSIs, overcoming the challenges that had previously hindered our progress. It is important to note that the dimension of the abundance space is equal to the number of endmembers. For this reason, we also include a comparison of the diffusion generation of abundance under the assumption of varying numbers of endmembers, as shown in Table 4. A close examination of the results in this table reveals that setting the number of endmembers to 5 provides the most effective balance. This setting offers an optimal compromise between the quality of generation and the consumption of computational resources.

5.3. Comparative experiments

5.3.1. HSI synthesis in remote sensing scene

Fig. 10 illustrates the false color images of HSIs synthesized by various algorithms. The results of the Gaussian mixture model (GMM) [62]

Table 4
Quantitative evaluation of ablation studies on the dimension of feature space used for diffusion generation.

Diffusion space (dimension)	Generation quality			Cost		
	FID↓	Recall ↑	Precision ↑	Params↓	Training (steps)	Sampling
HSI Cube (128)	–	–	–	99.88M	5 million	91s
Latent Feature (5)	49.47	0.165	0.181	103.57M	2.8 million	90s
Abundance (8)	15.19	0.361	0.310	99.78M	4.2 million	90s
Abundance (5)	**8.23**	**0.509**	**0.484**	99.76M	2.8 million	90s
Abundance (3)	8.77	0.503	0.415	**99.75M**	**2 million**	89s

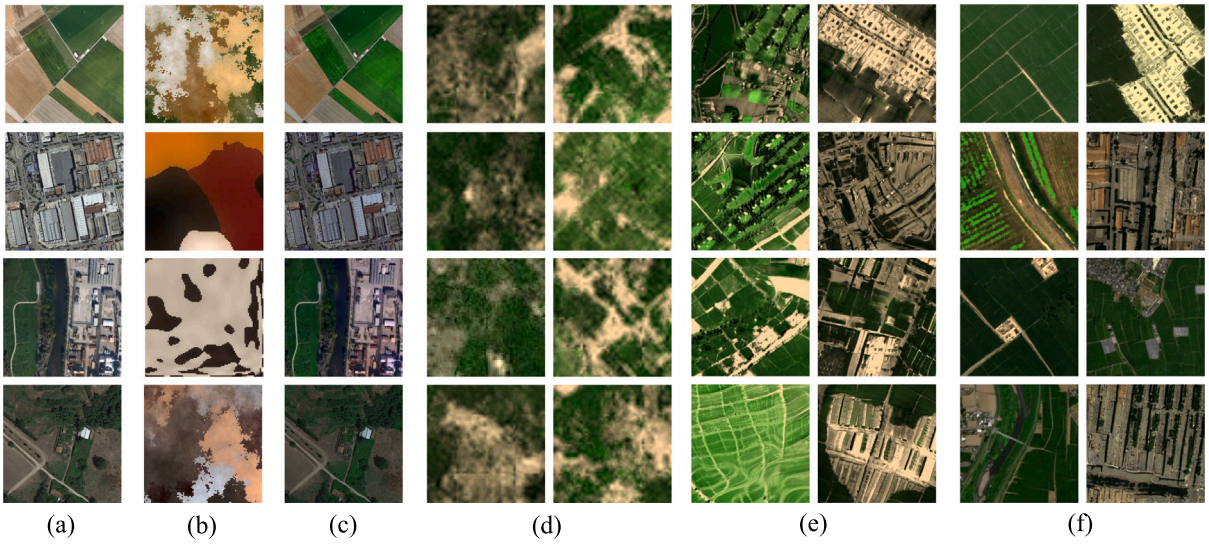


Fig. 10. Comparative experimental results. (a) Input RGB; (b) GMM; (c) MST++; (d) Abundance-based E-VDVAE; (e) Abundance-based BigGAN; (f) The proposed MSF-Diff.

Table 5
Quantitative evaluation for synthetic HSIs generated by comparative generative AI-based methods for different scenes (evaluated on 5000 synthetic HSIs for each scene).

Scenes	Remote sensing					Natural				
Metrics	FID↓	Recall ↑	Precision ↑	Params	Sampling	FID↓	Recall ↑	Precision ↑	Params	Sampling
Abundance-based E-VDVAE	39.13	0.209	0.184	178.78M	0.15s	51.43	0.146	0.133	178.77M	0.15s
Abundance-based BigGAN	11.35	0.456	0.392	**58.41M**	**0.03s**	10.04	0.275	0.647	**58.40M**	**0.03s**
The proposed MSF-Diff	**9.33**	**0.517**	**0.479**	99.76M	90s	**13.56**	**0.346**	**0.755**	99.75M	90s

are clearly distant from the real scene as it relies on ideal mathematical distributions to synthesize endmembers, rather than accurately simulating the true richness of the scene. On the other hand, the output of the Multi-stage Spectral Transformer (MST++) closely aligns with the input, although the spatial quality of the reconstructed data is slightly lacking. Worse still, spectral super-resolution methods like MST++ have a notable limitation, requiring a large number of paired RGB-HSI images for training.

In addition to traditional algorithms, we also conduct comparisons with advanced deep generative models, involving Efficient VDVAE (E-VDVAE) [51] and BigGAN [47]. Notably, to guarantee a fair comparison, we tailor them to the abundance domain in our experiments. The outputs produced by the Abundance-based E-VDVAE exhibit blurred textures, limited information content, and an overall subpar quality. The Abundance-based BigGAN, while showing slight improvements over E-VDVAE, is hampered by the protracted training time due to the unstable nature of the generative adversarial process. Apparently, texture distortions, unclear land cover boundaries resulting from localized blurring and deformation, as well as unreasonable spatial distributions exist in the related synthetic HSIs. These deficiencies can largely be attributed to the inherent instability associated with adversarial training. In contrast, the HSI synthesized by the proposed MSF-Diff effectively captures the spatial distribution characteristics of authentic remote sensing scenes, yielding a diverse range of rich HSI samples. The quantitative assessments presented in Table 5 further underscore the superiority of our proposed method in terms of quality and diversity in HSI generation, notwithstanding a minor trade-off in efficiency.

Furthermore, Fig. 11 shows the abundance generated by the proposed abundance-based diffusion, the synthesized HSI, and the spectral curves of selected sampling points, providing comprehensive evidence of the reliability in generating HSIs. This indicates that the proposed method holds significant advantages in terms of both quantity and quality.

We also attempt to evaluate some no-reference metrics on synthetic HSIs across different scenes. Typically, NIQE, PIQE, and BRISQUE metrics are employed to assess spatial quality in natural images, and in this study, we apply them to evaluate the spatial quality of each band image in the synthetic HSI. The outcomes are visualized in Fig. 12, depicting metric curves plotted along the spectral dimension. The findings indicate that each band in the synthetic HSIs exhibits relatively consistent and robust spatial quality.

5.3.2. HSI synthesis in natural scene

We also conducted comparison experiments in a natural mixed ground scene, from which we obtained promising results that further validate the effectiveness and generalization ability of our proposed method. Notably, natural scenes have a very complex composition with more unstable factors, making it impossible to list all the substances they contain. Based on the synthetic examples in Fig. 13 and the corresponding evaluations of the no-reference quality metrics in Table 5,

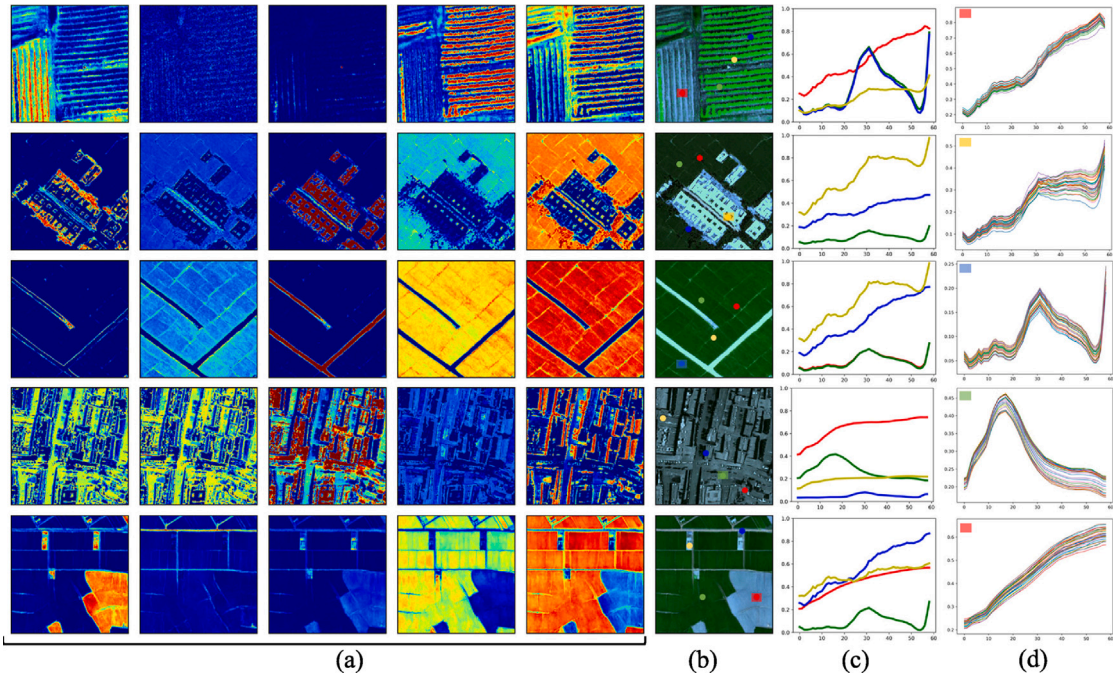


Fig. 11. Generation results in the remote sensing scene. (a) Synthetic abundance maps; (b) Generated HSIs in false color; (3) Spectral profile of several sampled pixels; (d) Spectral profile of sampled regions in the size of 6 × 6.

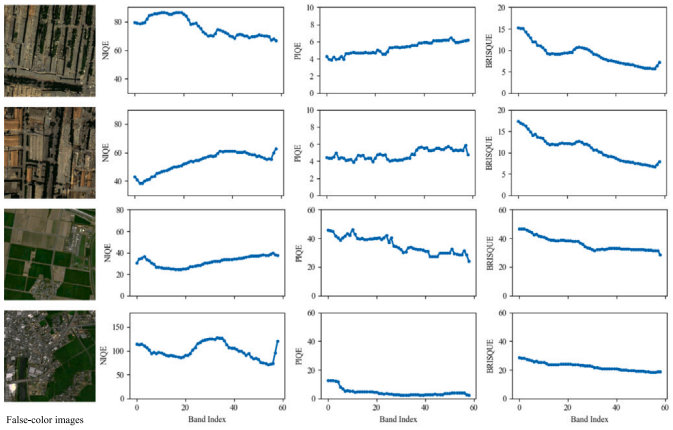


Fig. 12. Illustration of typical non-reference quality evaluation on synthetic HSIs along the spectral dimension.

traditional HSI synthesis methods continue to show poor performance. Similar to the findings observed in remote sensing scenarios, compared to the proposed MSF-Diff, VAE- and GAN-based methods exhibit shortcomings such as inadequate and indistinct textures, severe warping and distortions, and unreliable spatial relationships.

The synthetic abundance and reconstructed spectral profile shown in Fig. 14 provide evidence that our method can generate abundance maps that match the actual spatial distribution and synthesize HSIs with good and discriminative spectral signatures. This also indicates that even for natural scenes with numerous uncertain factors, our proposed method can effectively perceive the spatial distribution of different substances. In particular, for the images in the third row of Fig. 14, which are similar to backlit scenes, we sampled some backlit areas and plotted spectral curves. The characteristics of these curves are consistent with our target cognition, indicating that the spectral reflectance of real backlit areas usually contains less information. The spectral characteristics of the sampling area in Fig. 14(d) indicate that the same substance shows good spectral consistency. This suggests that the HSI generated by our proposed method has reliable spectral characteristics and satisfactory signature attributes.

Table 6
Parameters utilized for fine-tuning CNN-based networks.

Batch size	Iterations	Learning rate		Weight decay	Momentum
		Former layers	The last layer		
32	1000	1e−4	1e−2	5e−4	0.9

Table 7
The overall scene classification accuracy (%) of different models with/without synthetic HSIs.

Method	Training set	
	40% HSRS-SC	40% HSRS-SC + Synthetic HSIs
AlexNet [63]	89.51 ± 0.21	93.65 ± 0.19
VGGNet-16 [64]	63.81 ± 0.45	65.69 ± 0.38
ResNet-18 [65]	39.05 ± 0.17	42.22 ± 0.26

5.4. Extension experiments to downstream task

We conduct extension experiments on the scene classification task. It is noted that existing scene classification datasets often have limited scale and suffer from imbalances in class distribution. This trend also can be observed even in the recently released HSI dataset, HSRS-SC [11], which includes five categories: farmland, city building, building, idle region, and water. The number of samples for each category in this dataset ranges from 154 to 485. When selecting 40% of the dataset for training purposes, the largest class contains less than 200 samples. To address this issue, the dataset was augmented using the proposed MSF-Diff technique, ensuring that each category contained 1000 samples. To demonstrate the impact on scene classification, we employ three typical CNN-based networks and fine-tune them in our experiments with or without our synthetic HSIs. The specific pa-

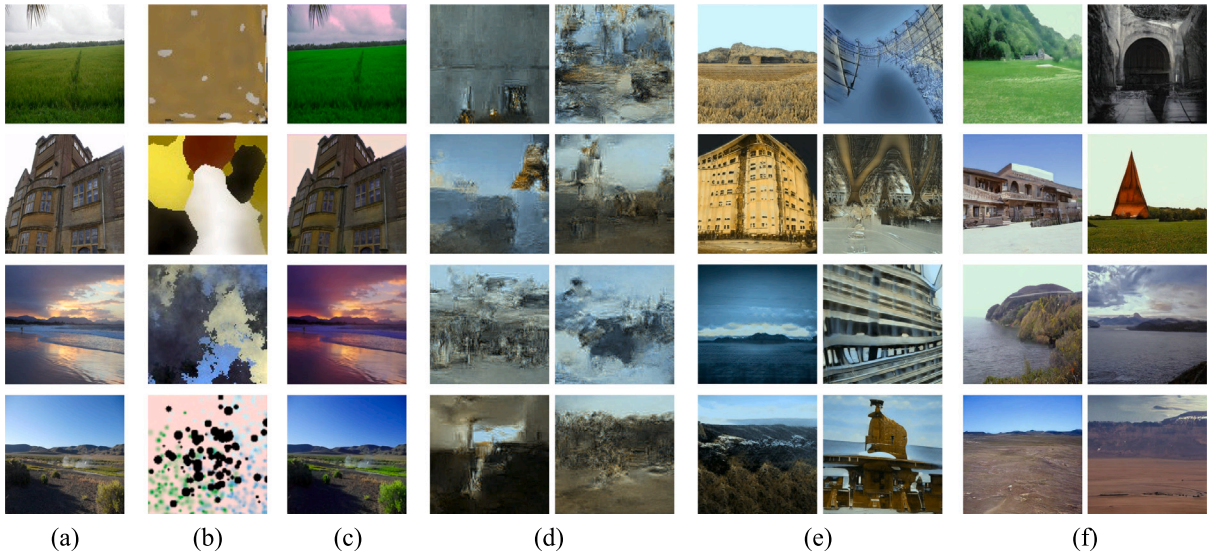


Fig. 13. Comparative experimental results. (a) Input RGB; (b) GMM; (c) MST++; (d) Abundance-based E-VDVAE; (e) Abundance-based BigGAN; (f) The proposed MSF-Diff.

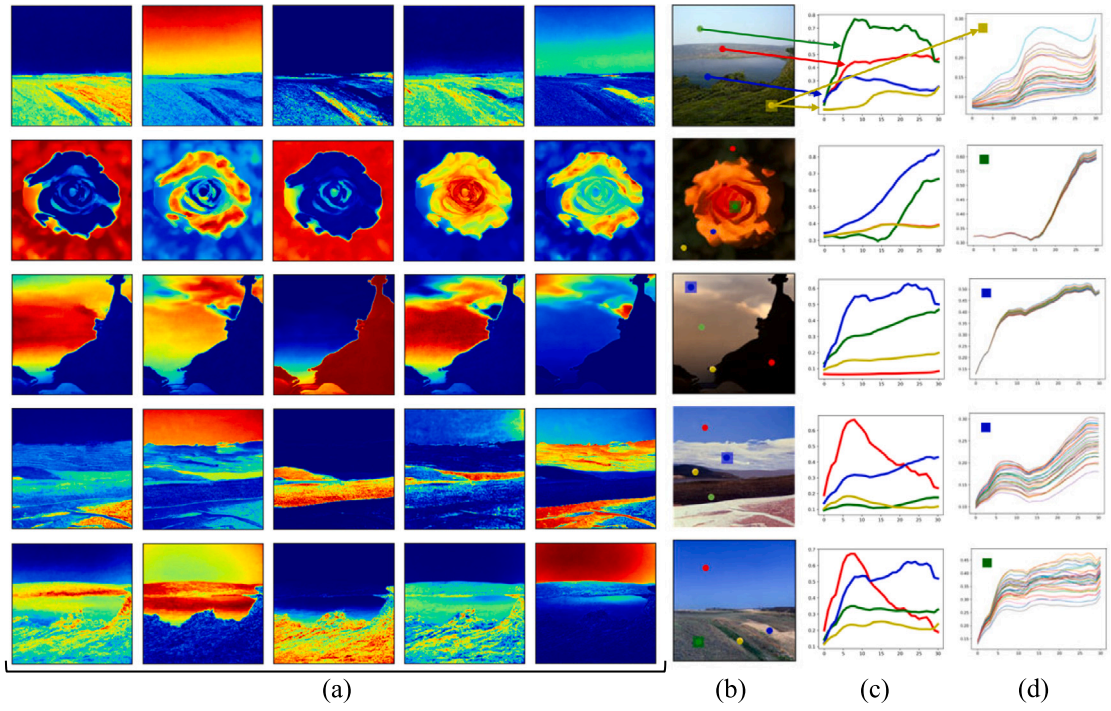


Fig. 14. Generation results in the natural mixed-ground scene. (a) Synthetic abundance maps;(b) Generated HSIs in false color; (3) Spectral profile of several sampled pixels; (d) Spectral profile of sampled regions in the size of 6 × 6.

rameters utilized for this fine-tuning process are outlined in Table 6. Notably, a higher learning rate of 0.01 was assigned to the final layer to facilitate efficient convergence and prevent entrapment in local optima, while a lower learning rate of 0.001 was assigned to the remaining layers to ensure steady progress during fine-tuning without disrupting the initializations. The results, presented in Table 7, show an improvement in the classification model performance due to synthetic data supplementation. This suggests that synthetic data can enhance the diversity and scale of existing datasets, mitigate issues like sample scarcity and class imbalance, and potentially benefit other downstream tasks.

6. Conclusion

This work proposes MSF-Diff, which shifts the focus from high-dimensional cubes to low-dimensional abundances and incorporates easily accessible RGB images for HSI synthesis. The underlying assumption is that images captured of a specific scene share similar compositions that can be effectively represented by a small set of endmembers and corresponding abundance maps. Building upon this premise, we have developed a method comprising scene-based unmixing, abundance-based diffusion, and fusion-based generation, collectively referred to as MSF-Diff. Through this methodology, the generation of large volumes of diverse and high-quality synthetic HSIs that

closely mirror real-world data becomes possible. Extensive experimentation showcases the efficacy and superiority of the proposed approach. Moreover, the validation on the downstream scene classification task demonstrates its ability to enrich the existing HSI dataset by producing a wide range of high-quality and diverse synthetic HSIs. This research constitutes a notable contribution to the HSI synthesis domain, offering a promising solution to tackling data scarcity challenges and driving progress in artificial intelligence applications across various sectors.

CRediT authorship contribution statement

Erting Pan: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Yang Yu:** Methodology, Validation, Visualization. **Xiaoguang Mei:** Methodology, Supervision, Writing – review & editing. **Jun Huang:** Supervision, Writing – review & editing, Methodology. **Jiayi Ma:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported the National Natural Science Foundation of China under Grant Nos. U23B2050, 62075169 and 62276192, the Industry-University-Research Cooperation Program of Zhuhai under Grant No. 2220004002828, and the Natural Science Foundation of Guangdong Province, China under Grant No. 2023A1515012834.

References

[1] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, X. Zhu, Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing, IEEE Geosci. Remote Sens. Mag. 9 (2) (2021) 52–87.

[2] M. Imani, H. Ghassemian, An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges, Inf. Fusion 59 (2020) 59–83.

[3] E. Pan, X. Mei, Q. Wang, Y. Ma, J. Ma, Spectral-spatial classification for hyperspectral image based on a single GRU, Neurocomputing 387 (2020) 150–160.

[4] Y. Zhong, X. Wang, S. Wang, L. Zhang, Advances in spaceborne hyperspectral remote sensing in China, Geo-spatial Inf. Sci. 24 (1) (2021) 95–120.

[5] N. Yokoya, A. Iwasaki, Airborne Hyperspectral Data Over Chikusei, Tech. Rep. SAL-2016-05-27, Space Application Laboratory, University of Tokyo, Japan, 2016.

[6] J. Jia, Y. Wang, J. Chen, R. Guo, R. Shu, J. Wang, Status and application of advanced airborne hyperspectral imaging technology: A review, Infrared Phys. Technol. 104 (2020) 103115.

[7] Y. Xu, J. Gong, X. Huang, X. Hu, J. Li, Q. Li, M. Peng, Luojia-HSSR: A high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet, Geo-Spatial Inf. Sci. 26 (3) (2023) 289–301.

[8] R. Dian, S. Li, B. Sun, A. Guo, Recent advances and new guidelines on hyperspectral and multispectral image fusion, Inf. Fusion 69 (2021) 40–51.

[9] B. Fan, Y. Yang, W. Feng, F. Wu, J. Lu, H. Liu, Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features, IEEE Trans. Multimed. (2022).

[10] Z. Shao, W. Wu, D. Li, Spatio-temporal-spectral observation model for urban remote sensing, Geo-Spatial Inf. Sci. 24 (3) (2021) 372–386.

[11] K. Xu, P. Deng, H. Huang, HSRS-SC: a hyperspectral image dataset for remote sensing scene classification. Journal of image and graphics, J. Image Graph. 26 (8) (2021) 1809–1822.

[12] J. Amieva, A. Austoni, M. Brovelli, L. Ansalone, P. Naylor, F. Serva, B.L. Saux, Deep-learning-based change detection with spaceborne hyperspectral PRISMA data, 2023, arXiv preprint arXiv:2310.13627.

[13] D. Li, M. Wang, J. Jiang, China's high-resolution optical remote sensing satellites and their mapping applications, Geo-Spatial Inf. Sci. 24 (1) (2021) 85–94.

[14] J. Jiang, J. Ma, X. Liu, Multilayer spectral–spatial graphs for label noisy robust hyperspectral image classification, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2020) 839–852.

[15] Y. Xu, T. Bai, W. Yu, S. Chang, P.M. Atkinson, P. Ghamisi, AI security for geoscience and remote sensing: Challenges and future trends, IEEE Geosci. Remote Sens. Mag. 11 (2) (2023) 60–85.

[16] W. Han, X. Zhang, Y. Wang, L. Wang, X. Huang, J. Li, S. Wang, W. Chen, X. Li, R. Feng, et al., A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities, ISPRS J. Photogramm. Remote Sens. 202 (2023) 87–113.

[17] E.J. Ientilucci, S.D. Brown, Advances in wide-area hyperspectral image simulation, in: Targets and Backgrounds IX: Characterization and Representation, vol. 5075, SPIE, 2003, pp. 110–121.

[18] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.

[19] S. Han, J.P. Kerekes, Overview of passive optical multispectral and hyperspectral image simulation techniques, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (11) (2017) 4794–4804.

[20] M.K. Jakubowski, D. Pogorzala, T.J. Hattenberger, S.D. Brown, J.R. Schott, Synthetic data generation of high-resolution hyperspectral data using DIRSIG, in: Imaging Spectrometry XII, vol. 6661, SPIE, 2007, pp. 153–163.

[21] E. Grau, J.-P. Gastellu-Etchegorry, Radiative transfer modeling in the earth–Atmosphere system with DART model, Remote Sens. Environ. 139 (2013) 149–170.

[22] X. He, X. Xu, Physically based model for multispectral image simulation of earth observation sensors, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (5) (2017) 1897–1908.

[23] M. Morata, B. Siegmann, A. Pérez-Suay, J.L. García-Soria, J.P. Rivera-Caicedo, J. Verrelst, Neural network emulation of synthetic hyperspectral sentinel-2-like imagery with uncertainty, IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 16 (2022) 762–772.

[24] J.M. Haut, M.E. Paoletti, J. Plaza, A. Plaza, J. Li, Hyperspectral image classification using random occlusion data augmentation, IEEE Geosci. Remote Sens. Lett. 16 (11) (2019) 1751–1755.

[25] H. Gao, J. Zhang, X. Cao, Z. Chen, Y. Zhang, C. Li, Dynamic data augmentation method for hyperspectral image classification based on siamese structure, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 8063–8076.

[26] M. Zhang, Z. Wang, X. Wang, M. Gong, Y. Wu, H. Li, Features kept generative adversarial network data augmentation strategy for hyperspectral image classification, Pattern Recognit. 142 (2023) 109701.

[27] J. Wang, M. Zhang, W. Li, R. Tao, A multistage information complementary fusion network based on flexible-mixup for HSI-X image classification, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[28] Y. Cai, J. Lin, Z. Lin, H. Wang, Y. Zhang, H. Pfister, R. Timofte, L. Van Gool, MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 745–755.

[29] J. He, Q. Yuan, J. Li, L. Zhang, PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images, Inf. Fusion 80 (2022) 205–225.

[30] Q. Ma, J. Jiang, X. Liu, J. Ma, Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution, Inf. Fusion 100 (2023) 101907.

[31] H. Zhang, J. Ma, STP-SOM: Scale-transfer learning for pansharpening via estimating spectral observation model, Int. J. Comput. Vis. 131 (12) (2023) 3226–3251.

[32] T. Bodrito, A. Zouaoui, J. Chanussot, J. Mairal, A trainable spectral-spatial sparse coding model for hyperspectral image restoration, Adv. Neural Inf. Process. Syst. 34 (2021) 5430–5442.

[33] H. Liu, F. Jin, H. Zeng, H. Pu, B. Fan, Image enhancement guided object detection in visually degraded scenes, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[34] E. Pan, Y. Ma, X. Mei, J. Huang, Q. Chen, J. Ma, Hyperspectral image destriping and denoising from a task decomposition view, Pattern Recognit. 144 (2023) 109832.

[35] S. Bond-Taylor, A. Leach, Y. Long, C.G. Willcocks, Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[36] A. Razavi, A. Van den Oord, O. Vinyals, Generating diverse high-fidelity images with vq-vae-2, Adv. Neural Inf. Process. Syst. 32 (2019).

[37] J. Peng, D. Liu, S. Xu, H. Li, Generating diverse structure for image inpainting with hierarchical VQ-VAE, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10775–10784.

[38] A. Aggarwal, M. Mittal, G. Battineni, Generative adversarial network: An overview of theory and applications, Int. J. Inf. Manage. Data Insights 1 (1) (2021) 100004.

[39] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: Algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3313–3332.

[40] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[41] D. Kingma, T. Salimans, B. Poole, J. Ho, Variational diffusion models, Adv. Neural Inf. Process. Syst. 34 (2021) 21696–21707.
[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, 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. 10684–10695.
[43] G. Harshvardhan, M.K. Gourisaria, M. Pandey, S.S. Rautaray, A comprehensive survey and analysis of generative models in machine learning, Comp. Sci. Rev. 38 (2020) 100285.
[44] A. Oussidi, A. Elhassouny, Deep generative models: Survey, in: 2018 International Conference on Intelligent Systems and Computer Vision, ISCV, IEEE, 2018, pp. 1–8.
[45] A.N. Wu, F. Biljecki, InstantCITY: Synthesising morphologically accurate geospatial data for urban form analysis, transfer, and quality control, ISPRS J. Photogramm. Remote Sens. 195 (2023) 90–104.
[46] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, Z. Lian, Controllable person image synthesis with attribute-decomposed gan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5084–5093.
[47] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2018, arXiv preprint arXiv:1809.11096.
[48] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Adv. Neural Inf. Process. Syst. 34 (2021) 852–863.
[49] A. Razavi, A. Van den Oord, O. Vinyals, Generating diverse high-fidelity images with vq-vae-2, Adv. Neural Inf. Process. Syst. 32 (2019).
[50] R. Child, Very deep {vae}s generalize autoregressive models and can outperform them on images, in: International Conference on Learning Representations, 2021, URL https://openreview.net/forum?id=RLRXCV6DbEJ.
[51] L. Hazami, R. Mama, R. Thurairatnam, Efficient-VDVAE: Less is more, 2022, arXiv preprint arXiv:2203.13751.
[52] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[53] A.Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.
[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, 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. 10684–10695.
[55] R. Rombach, P. Esser, B. Ommer, Network-to-network translation with conditional invertible neural networks, Adv. Neural Inf. Process. Syst. 33 (2020) 2784–2797.
[56] J.M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, J. Chanussot, Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (2) (2012) 354–379.
[57] Y. Yu, Y. Ma, X. Mei, F. Fan, J. Huang, H. Li, Multi-stage convolutional autoencoder network for hyperspectral unmixing, Int. J. Appl. Earth Obs. Geoinf. 113 (2022) 102981.
[58] Q. Zhu, W. Deng, Z. Zheng, Y. Zhong, Q. Guan, W. Lin, L. Zhang, D. Li, A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification, IEEE Trans. Cybern. 52 (11) (2021) 11709–11723.
[59] B. Xi, J. Li, Y. Diao, Y. Li, Z. Li, Y. Huang, J. Chanussot, DGSSC: A deep generative spectral-spatial classifier for imbalanced hyperspectral imagery, IEEE Trans. Circuits Syst. Video Technol. 33 (4) (2022) 1535–1548.
[60] C.-I. Chang, S.-S. Chiang, J.A. Smith, I.W. Ginsberg, Linear spectral random mixture analysis for hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 40 (2) (2002) 375–392.
[61] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.
[62] Y. Zhou, A. Rangarajan, P.D. Gader, A Gaussian mixture model representation of endmember variability in hyperspectral unmixing, IEEE Trans. Image Process. 27 (5) (2018) 2242–2256.
[63] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[64] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.