Journal of Computational Design and Engineering, 2026, **13**, 197–212

DOI: 10.1093/jcde/qwag008
Advance access publication date: 2 February 2026
Research Article

TDiff-HSI: Tucker-guided diffusion for high-dimensional RGB-to-HSI image generation

Jaeik Bae and Yong-Gu Lee*

Department of AI Convergence, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, 61005 Gwangju, Republic of Korea
*Correspondence: lygu@gist.ac.kr

Abstract

We introduce TDiff-HSI, a diffusion-based model that can generate hyperspectral images (HSIs) directly from RGB images and material-wise segmentation masks. HSI provides both spatial (u, v) and spectral (λ) information. The accompanying dataset that we are releasing spans wavelengths in the range from 420 to 1728 nm, digitized into 512 channels. Directly handling this immense three-dimensional dataset is computationally prohibitive and often leads to numerical errors. To address this challenge, TDiff-HSI leverages Tucker decomposition to reduce dimensionality, enabling more stable and efficient processing. Moreover, spectral precision is enhanced by combining RGB channels with a material segmentation mask.
To support this research, we constructed a new dataset using a hyperspectral camera. The dataset comprises 40 014 RGB-HSI pairs across 78 scenes, featuring 12 objects with corresponding polygonal segmentation labels. Experimental evaluation demonstrates that TDiff-HSI achieves state-of-the-art performance verified on the existing dataset. For the new dataset that we are releasing, we establish new benchmarks of MRAE 0.2169, RMSE 0.0192, PSNR 36.46 dB, SAM 0.0424, and SSIM 0.9327
Project and dataset are available at https://github.com/JaeikBae/TDiff-HSI

Keywords: diffusion model, hyperspectral image, Tucker decomposition

Nomenclature

AI:	Artificial intelligence
CNN:	Convolutional neural network
HSI:	Hyperspectral image
LiDAR:	Light detection and ranging
MRAE:	Mean relative absolute error
RMSE:	Root mean squared error
PSNR:	Peak signal-to-noise ratio
SAM:	Spectral angle mapper
SSIM:	Structural similarity index measure
u, v :	Spatial coordinates of an image pixel
λ :	Wavelength
x_0 :	Clean (ground-truth) data sample
x_t :	Noisy data sample at diffusion time step t
$\hat{x}_0$:	Predicted clean data sample
t :	Diffusion time step
T :	Total number of diffusion steps
β_t :	Noise variance at time step t
ϵ :	Gaussian noise
X :	Original hyperspectral image tensor
$\mathcal{C}$:	Core tensor in Tucker decomposition
$\mathcal{F}_n$:	Factor matrix of the nth mode in Tucker decomposition

1. Introduction

Hyperspectral images (HSIs) provide far richer spectral information per pixel than conventional RGB imaging, offering high spectral resolution across a broad wavelength range. HSIs are effective in applications such as camouflage detection and material discrimination tasks that are infeasible with RGB data alone. Consequently, HSIs have found extensive applications in agriculture, environmental monitoring, remote sensing, medical imaging, and military purposes. In this study, we focus on HSI data spanning 420–1728 nm, divided into 512 spectral channels, forming a three-dimensional structure defined by spatial coordinates (u, v) and spectrum (λ).

The widespread use of HSI remains challenging due to several reasons. Acquiring HSI data require expensive equipment. In addition, a 512-channel HSI demands nearly 170 times more storage and computation than a three-channel RGB image. Moreover, most HSI cameras employ line-scan mechanisms, restricting data collection to controlled indoor environments. Satellite-based HSI cameras offer broader coverage but suffer from limited spectral range or low spatial resolution. These challenges underscore the need for methods that enhance the accessibility and usability of high-dimensional HSI data. Recently, diffusion models have emerged as powerful generative frameworks, capable of producing high-resolution images via progressive denoising. Building on this paradigm, we propose TDiff-HSI, a model that generates HSI data directly from RGB images and material segmentation masks. Naively generating all 512 spectral channels simultaneously requires high computational cost and numerical instability. To overcome this, we employ Tucker decomposition, which factorizes the high-dimensional data into a compact representation, thereby improving stability and efficiency. By enabling the gener-

Received: October 25, 2025. **Revised:** January 26, 2026. **Accepted:** January 28, 2026

ation of HSIs in environments where direct acquisition is difficult or cost-prohibitive, the proposed method can contribute to a wide range of computational design applications. For example, it can be effectively applied to tasks such as material analysis, surface appearance modeling, and optical simulation, where spectral information is essential but direct hyperspectral data acquisition is challenging.

The main contributions of this paper are as follows:

1. To the best of our knowledge, this is the first work that directly generates complete 512-channel HSI data using diffusion models conditioned on RGB images and segmentation masks.
2. We introduce a novel application of Tucker decomposition to factorize high-dimensional HSI data into a low-dimensional core matrix and factor matrices, substantially improving both numerical stability and generation performances.
3. We construct and release a new HSI dataset comprising 78 scenarios with 12 objects across nine material categories. The dataset includes RGB images, full-channel HSI data, and material-wise segmentation in YOLO-style polygon format. This dataset uses 512 channels, which is the largest among available open datasets. We believe this dataset can be used for future RGB-to-HSI research that utilizes the largest number of channels.

2. Related Works

2.1. Diffusion models

Diffusion models based on deep learning have recently achieved remarkable progress across diverse modalities, including image, video, audio, and 3D data generation. The Denoising Diffusion Probabilistic Model (DDPM; Ho *et al.*, 2020) introduced the forward-reverse noise process, demonstrating superior generative quality compared to Generative Adversarial Networks (GANs; Goodfellow *et al.*, 2020). However, the computational time to generate images was too high. The Denoising Diffusion Implicit Model (DDIM; Song *et al.*, 2020) used deterministic, non-Markovian diffusion sampling, giving rise to implicit models that generate high-quality samples much faster. Later advancements such as Latent Diffusion Models (LDMs) (Rombach *et al.*, 2022) improved computational efficiency, while GLIDE (Nichol *et al.*, 2021) and Imagen (Saharia *et al.*, 2022) showcased strong performance in text-to-image generation. Diffusion models have also been extended to video generation (Ho *et al.*, 2022), producing high-resolution text-to-video results. Beyond media content, domain-specific applications include fire and smoke simulation in buildings (Zeng *et al.*, 2024) and AI-powered architectural design (Jo *et al.*, 2024; Shi *et al.*, 2024).

Despite these successes, most studies focused on RGB or text-conditioned images. Efforts to generate HSIs have been rare. Prior work such as R2H-CCD (Zhang *et al.*, 2023) and DDSR (Chen *et al.*, 2024) attempted HSI generation but were limited to the narrow visible spectrum (400–700 nm). Existing diffusion frameworks are primarily optimized for low-dimensional RGB data and do not scale well to high-dimensional, multi-channel spectral data, where continuous spectral consistency is critical. To address this challenge, our study applies Tucker decomposition to compress hyperspectral data, enabling accurate spectral detail preservation while improving computational stability.

2.2. Hyperspectral imaging

Hyperspectral imaging plays a crucial role in analyzing material-specific spectral properties by providing fine-grained spatial and spectral information. In geology, HSIs have been employed for soil and rock classification, chemical composition analysis, and monitoring erosion processes (Allegretta *et al.*, 2022; Son *et al.*, 2022). In medicine, it has been used for brain tumor classification (Baig *et al.*, 2021), cellular analysis using hyperspectral endoscopy (Grigoroiu *et al.*, 2020), and detection of pine wilt disease in forestry (Lee *et al.*, 2014). In the food and agriculture industries, HSIs have been applied to pungency grading of red pepper powder (Choi *et al.*, 2022) and quality assessment of beef (Park *et al.*, 2023).

In object detection tasks, HSIs have also been used for tasks that use RGB images. Representative studies include spectral-domain segmentation using CNNs (Hu *et al.*, 2015), object tracking in hyperspectral videos (Chen *et al.*, 2016; Li *et al.*, 2020; Liu *et al.*, 2021; Xiong *et al.*, 2020), and detection of camouflaged or microscopic objects (Hupel *et al.*, 2022; Ling *et al.*, 2022). To mitigate challenges of high dimensionality and sensor noise, various approaches such as denoising (Dao *et al.*, 2021), outlier detection (Kim *et al.*, 2016), wavelet-based feature extraction (Hsu *et al.*, 2007), and PCA-driven dimensionality reduction (Ren *et al.*, 2014) have been explored.

Nevertheless, most existing studies focus on analyzing or augmenting HSI data rather than directly generating it. Earlier efforts using datasets such as CAVE (Yasuma *et al.*, 2010), ARAD-1k (Arad *et al.*, 2022), Foster (Nascimento *et al.*, 2016), and KAUST-HS (Li *et al.*, 2021) targeted only 31-channel multispectral data in the 400–700 nm visible spectrum, essentially could be regarded as RGB channels augmented with interleaving finer channels. By contrast, our work directly generates 512-channel HSI spanning 420–1728 nm, addressing a substantially broader spectral range.

2.3. Sensor fusion with HSI

While spectral cues, their low spatial resolution, and structural ambiguity limit standalone utility. Consequently, researchers have explored fusing HSI with other modalities such as RGB or LiDAR. For instance, HSI-LiDAR fusion has enhanced land-cover classification (Liao *et al.*, 2014), and CNN-based hyperspectral-multispectral fusion has restored high-resolution HSI (Fu *et al.*, 2019). More recently, transformer-based cross-attention approaches have achieved efficient band selection with LiDAR guidance, balancing accuracy and computational cost (Yang *et al.*, 2024).

Despite these advances, sensor fusion research has been constrained by the scarcity of high-quality HSI datasets. In this context, our work proposes direct hyperspectral generation from RGB images and segmentation masks, eliminating dependence on costly HSI acquisition hardware and expanding the accessibility of spectral data.

2.4. Tucker decomposition for high-dimensional image representation

Tucker decomposition has been widely studied as an effective technique for compactly representing high-dimensional tensor data (Kolda *et al.*, 2009). In image analysis and generation tasks, it has been employed to reduce computational complexity, improve numerical stability, and capture latent structures across multiple dimensions (Cichocki *et al.*, 2015). Prior studies have demonstrated that tensor decomposition methods, including Tucker decomposition, are particularly suitable for handling data with

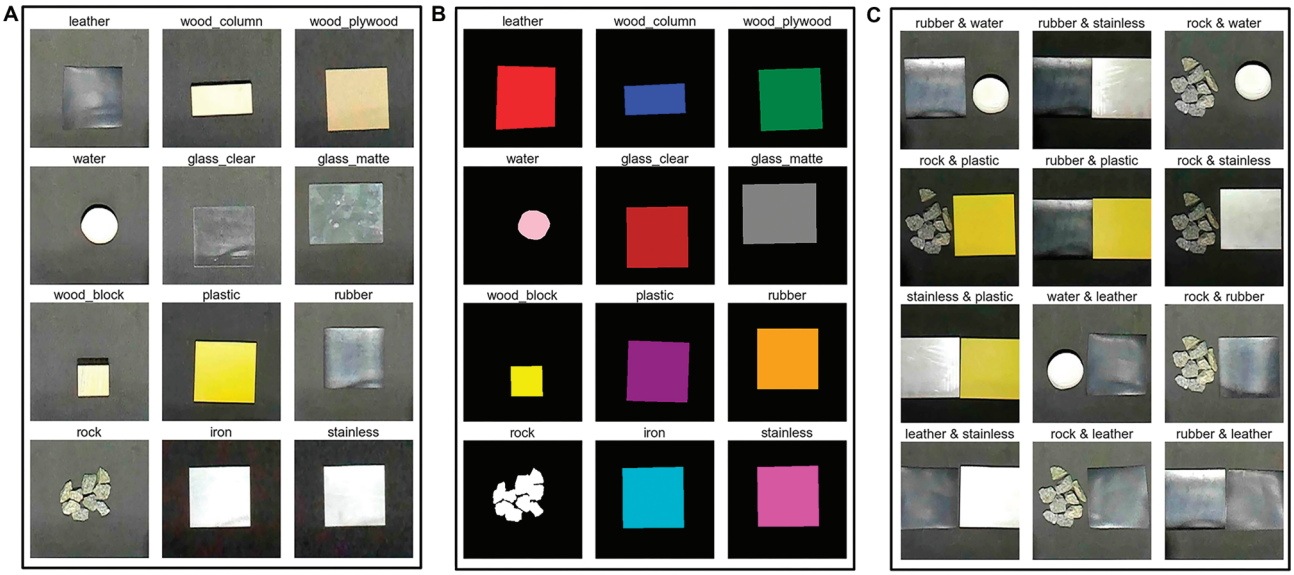

Figure 1: Examples from our dataset: (A) Single-object data, (B) Corresponding segmentation labels, and (C) Dual-object images (partially shown).

strong inter-dimensional correlations, such as spatial–spectral or spatial–temporal relationships in images (Sidiropoulos *et al.*, 2017).

Although direct applications of Tucker decomposition to HSI generation remain limited, its effectiveness in high-dimensional image representation motivates its adoption in this work. In contrast to conventional dimensionality reduction methods that flatten or collapse spectral information, Tucker decomposition preserves the multi-way structure of hyperspectral data while significantly reducing its dimensionality. This property makes it a natural choice for HSI generation, where maintaining spectral coherence across hundreds of channels is critical.

3. Dataset

To train and evaluate the proposed TDiff-HSI model, we constructed a new RGB-HSI dataset using a hyperspectral imaging system. This dataset is released publicly and is intended to serve as a benchmark for future research on RGB-to-HSI generation.

3.1. Data collection

The data used in this study were directly collected using an HSI camera, AHS-003VIR (Aval Data Corporation, Japan). The equipment divides a wavelength region of 420–1728 nm into 512 channels and stores a monochrome image corresponding to each wavelength. Therefore, one scene is represented as a $512 \times H \times W$ data cube, and the value of each pixel represents the intensity of light reflected by the material at a specific wavelength. In order to use it as an input condition for the generative model, RGB images were taken for the same scene, and material-specific segmentation was produced by hand.

The raw HSIs captured by the AHS-003VIR camera have an original spatial resolution of 640×512 ($H \times W$). To remove unnecessary background regions and to center the target object within the field of view, each hyperspectral cube was spatially cropped to 320×320. For stable training and efficient processing within the diffusion framework, the cropped data were further resized to 256×256 before being used as model inputs.

Figure 1 shows the composition of the dataset according to this design. To enable the generation of various substances, this study implemented nine types of materials. Although the current implementation only uses nine types, our method does not theoretically limit the number of types, and the user may incorporate more materials if needed. There is no restriction on the number of objects per scene. Although only two objects were used for data preparation, more can be added. Each object type refers to a distinct geometry, while material type denotes its surface composition (e.g. metal, plastic, and glass). Accordingly, the dataset includes 12 object geometries and nine material categories. Altogether, the dataset has 78 scenes. There are 12 scenes with a single object and 66 scenes with two objects. In addition, the term 'matte' is used here to denote a dull, non-reflective surface finish (e.g. matte glass), in contrast to a clear finish. While the surface appearance differs, the hyperspectral response is dominated by the spectral properties of the material.

Figure 2 illustrates why segmentation is necessary for hyperspectral reconstruction from RGB inputs. The visible spectral range (400–700 nm) can be extracted from RGB channels, but no direct information exists in the near-infrared and shortwave-infrared region (700–1728 nm). Without additional cues, these channels cannot be reconstructed. By incorporating segmentation-guided priors, the system can infer material or class-dependent spectral properties, thereby compensating for the missing information and enabling a complete hyperspectral representation.

3.2. Data features

HSI utilizes material-specific spectra. Material-specific spectral information spans the range of 512-channel images. As shown in Figure 3, when the values of a specific pixel position (u, v) in the HSI data are collected for 512 channels, it becomes spectral information having a one-dimensional arrangement of 512 elements. Using these spectral data, it is possible to distinguish which material is located at the specific pixel position. The spectrum for each material is represented as shown in Figure 4. When analyzing such a spectrum, an important observation is the abrupt decline of the intensity at a specific wavelength. This decline can be used as a signature to identify specific materials. For example, plastic exhibits spectral characteristics that absorb light with a wavelength of approximately 1200 nm, iron absorbs light from 900 to 1200 nm, and water absorbs light at 1000 nm. On the contrary,

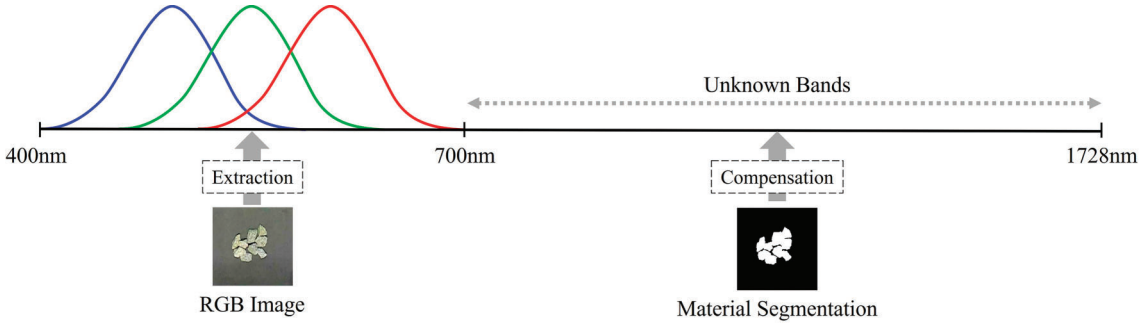

Figure 2: Conceptual illustration showing the necessity of segmentation in RGB-to-HSI conversion. While RGB images provide spectral information only in the visible range (400–700 nm), no direct information is available for the near-infrared and shortwave-infrared regions (700–1728 nm). To compensate for this missing spectral information, a material-wise segmentation map is used as an additional input. The segmentation is represented as a single-channel image, where each pixel encodes the material class at the corresponding spatial location, enabling the model to infer material-dependent spectral characteristics beyond the visible spectrum.

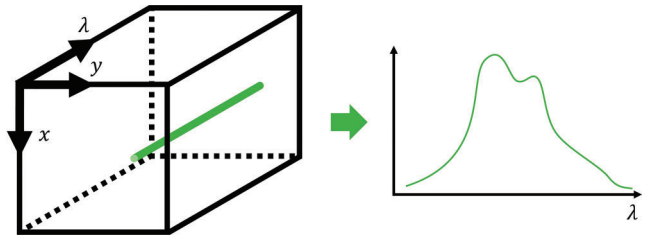

Figure 3: Illustration of how to extract two-dimensional spectral data from three-dimensional HSI data.

the bell-shaped shape that is common in each spectrum reflects the characteristics of the AHS-003VIR camera used. This is due to the fact that the sensor sensitivity of the AHS-003VIR camera is the highest at approximately 1000–1200 nm. As such, the spectral characteristics of each material appear depending on the atoms constituting the material and its molecular structure. The result is the appearance of strong signal only at a specific wavelength or over a broad wavelength. In Figure 4, leather and rubber show stronger noise than other materials because their dark color absorbs more light across the spectrum, causing the normalization process to emphasize the noise. Nevertheless, their spectral signatures remain identifiable. During acquisition, shadows of different intensity and occasional vertical stripes appeared as artifacts of illumination and line-scan sensing. However, these mainly cause intensity reduction with little effect on spectral shape. For transparent samples such as water, the spectrum is still dominated by the material itself, as Figure 4 demonstrates.

4. TDiff-HSI

4.1. Preliminaries

4.1.1. *Diffusion process*

Compared to existing generative AI algorithms such as GANs, the diffusion algorithm demonstrates more efficient and stable learning performance. Diffusion basically operates by predicting the noise present in the current image. Specifically, the diffusion model consists of a forward process and a backward process. In the forward process, noise is gradually added to a clean data Z_0, and finally, complete noise data Z_T that follows a nearly normal distribution is generated. This process is illustrated in Figure 5, and it is mathematically defined in Equation 1:

$$q\left(Z_t|Z_{t-1}\right) := \mathcal{N}\left(Z_t; \sqrt{1-\beta_t}Z_{t-1}, \beta_t I\right). \tag{1}$$

Here, β_t represents a small noise dispersion (intensity) according to the time step t which is not a learnable parameter, and the model uses a general scheduling method like linear or cosine to vary this through each step. When the final step T is reached, the data becomes completely noisy. Conversely, the reverse process is a process of restoring a sample Z_0 close to the original data distribution while gradually removing noise, starting with this noisy data Z_T. The reverse process learns the conditional distribution through the neural network $\mu_\theta(Z_t, t)$, and it is expressed as Equation 2:

$$p_\theta\left(Z_{t-1}|Z_t\right) := \mathcal{N}\left(Z_{t-1}; \mu_\theta\left(Z_t, t\right), \sigma_t^2 I\right). \tag{2}$$

The noise variance σ_t^2 can also be interpreted as the small noise dispersion at the time step t. The noise variance can either be learned or predefined. In this study, σ_t^2 is fixed as $\sigma_t^2 = \beta_t$, which still ensures a valid diffusion process. Through the use of this equation, the diffusion algorithm has previously shown excellent performance in image generation applications. Based on this success, the diffusion algorithm was used as the base generative algorithm. In addition, Equation 3, which is the method of DDIM (Song *et al.*, 2020), a follow-up study of DDPM, was used to quickly generate images and achieve high consistent generation quality. Using this method, it is possible to generate images of comparable quality even when the number of sampling iterations T is reduced by 10–100 times. Furthermore, in long sampling iterations, Z_0 was used as a conditional dictionary probability for the same initial noise state, which enabled consistent generation:

$$q_\sigma\left(Z_{1:T}|Z_0\right) := q_\sigma\left(Z_t|Z_0\right)\prod_{t=2}^{T} q_\sigma\left(Z_{t-1}|Z_t, Z_0\right). \tag{3}$$

4.1.2. *Tucker decomposition*

As shown in Figure 6, Tucker decomposition decomposes a high-dimensional tensor into a product of a core tensor and several factor matrices. The HSI covered in this paper may be expressed in the form of a three-dimensional tensor, and it can be expressed as shown in Equation 4:

$$\mathcal{X} \in \mathbb{R}^{\lambda \times H \times W}. \tag{4}$$

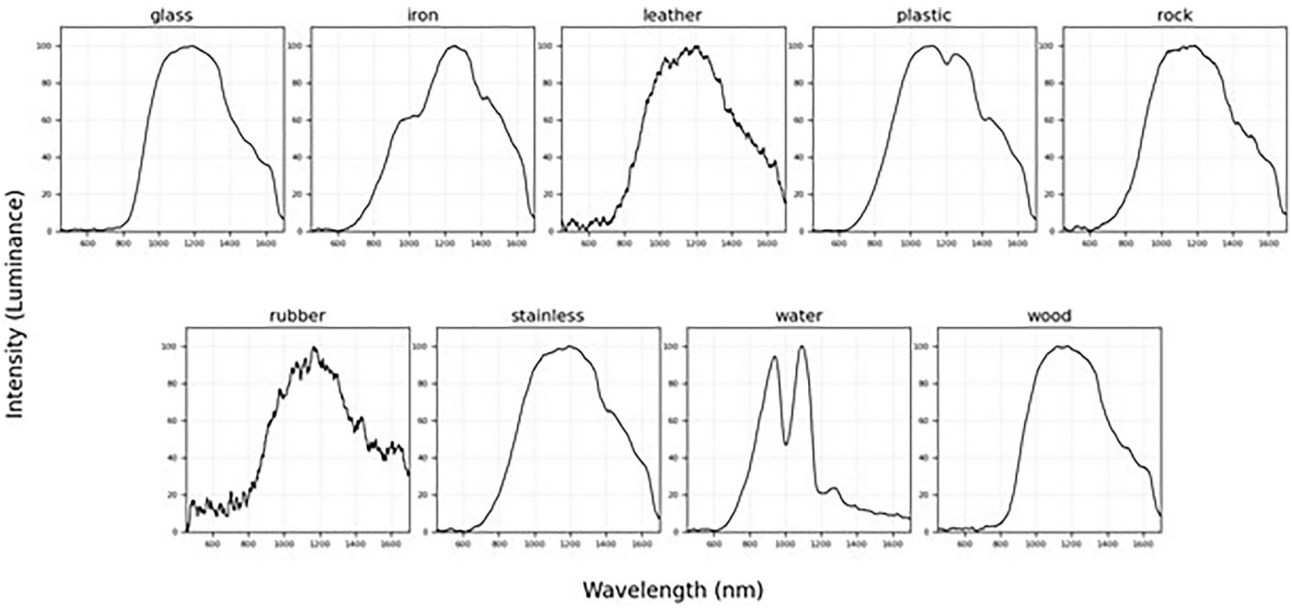

Figure 4: Average spectral responses of each material, normalized to a 0–100 scale. Distinct reflectance patterns appear across wavelengths.

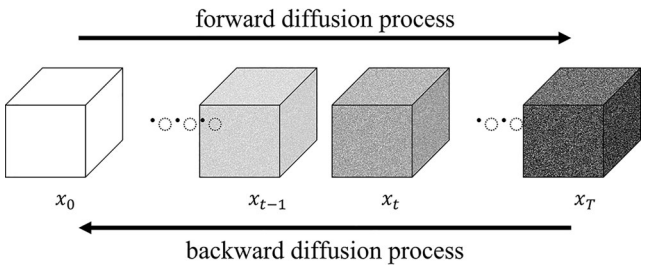

Figure 5: Forward and backward diffusion processes.

Here, λ, H, and W represent the number of spectral bands, image height, and image width of the HSI, respectively. After Tucker decomposition, the original data X is approximated through Singular Value Decomposition (SVD) as follows:

$$\mathcal{X} \approx \mathcal{C} \times \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3. \tag{5}$$

Here, $\mathcal{C} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is a low-dimensional core tensor, which implicitly contains the core structure and important information of the original data. $\mathcal{F}_n \in \mathbb{R}^{I_n \times R_n}$ $(n = 1, 2, 3)$ is a factor matrix corresponding to each dimension and mainly contains the common characteristics of the original dataset. R_n $(n = 1, 2, 3)$ represents each dimension, and generally serves to reduce the dimension of the data by choosing a $R_n \ll I_n$ where I_n denotes the original size of the nth dimension in the data tensor.

Through the above process, the size of the original data can be effectively reduced, and data similar to the original can be restored by matrix-multiplying the core matrix with the factor matrix. In this paper, data were generated efficiently by using this Tucker decomposition which significantly reduced the size of the required data. For example, if trying to generate a tensor of $512 \times 256 \times 256$, this amounts to 33, 554, 432 floats. However, if the size of the core R_n is set to $64 \times 128 \times 128$, which is the value used in this study, only $(512 \times 64) + (256 \times 128) + (256 \times 128) = 98,304$ real values are required when applying Tucker decomposition. The reduction in the size of floats amounts to approximately 341 × . This has the effect of significantly reducing the overall computational and memory burden.

The motivation for adopting Tucker decomposition lies in its ability to compactly represent high-dimensional hyperspectral data while preserving its intrinsic multidimensional structure. HSIs consist of strong correlations across spectral, spatial height, and spatial width dimensions, and directly modeling all channels simultaneously often leads to high computational cost and numerical instability.

By decomposing the hyperspectral tensor into a low-dimensional core and factor matrices, Tucker decomposition effectively reduces the dimensionality of the data while retaining its essential spectral-spatial characteristics. This not only stabilizes the diffusion-based generation process but also enables efficient training and inference under limited computational resources, making it particularly well suited for high-channel HSI generation.

4.2. Architecture

The architecture of the TDiff-HSI model proposed in this study is shown in Figure 7. In particular, it is designed to generate HSIs efficiently and accurately using RGB images and segmentation. The model can divided mainly into four components: input, context embedding, diffusion model (U-Net), and output.

First, RGB images, segmentation maps, and random noise are given as inputs. RGB images contain three-channel information. Segmentation consists of a single channel representing the spatial distribution of the material. These two inputs are used to identify the spatial distribution of each object and material. The random noise serves as the initialization for the diffusion process.

Second, in the context embedding process, the RGB image and segmentation information are combined to form contextual information. This contextual information helps the model to accurately understand each object's shape, location, and material properties. At the same time, segmentation passes through a lightweight residual and attention-based encoder to create multiple factors designed to capture the intrinsic spectral and structural properties of the material.

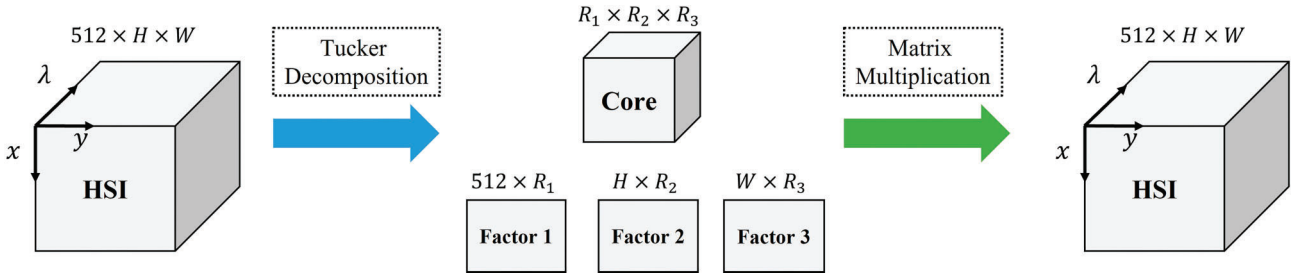

Figure 6: Illustration of Tucker decomposition and reconstruction.

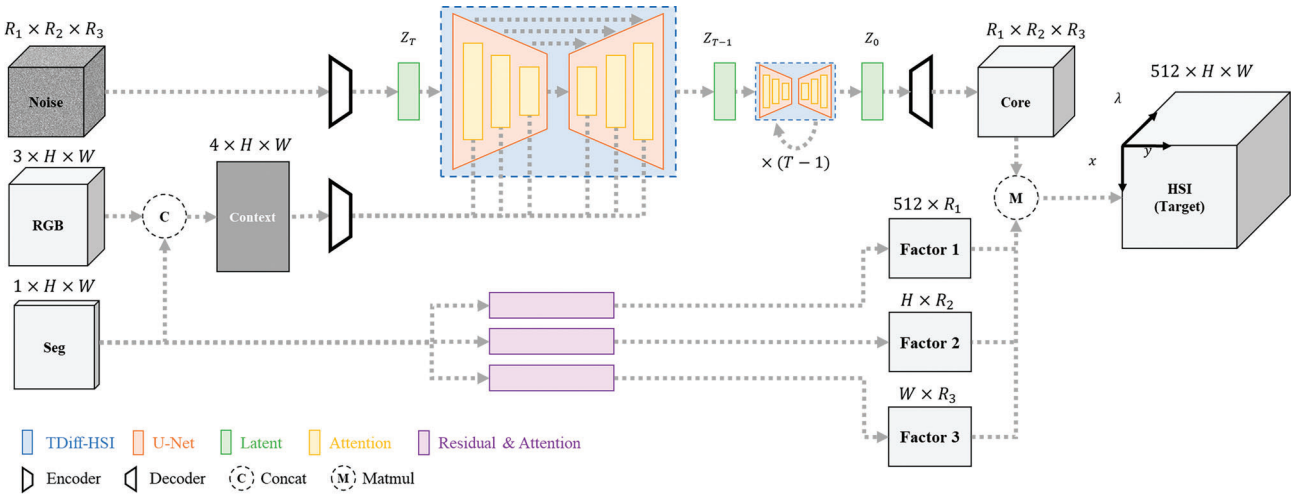

Figure 7: Architecture of proposed model (TDiff-HSI).

Third, U-Net, the core of the diffusion model, performs a denoising process by receiving noisy HSI input and contextual information, simultaneously. This study is based on the DDPM and further incorporates the DDIM for stable and efficient training and sampling. The U-Net structure shows excellent performance in gradually removing noise while preserving the satisfactory spatial resolution of the input data.

Finally, in the output step, the intermediate hyperspectral representation generated by the U-Net is combined with segmentation-derived factors through a matrix multiplication operation (denoted as *M*). This fusion process ensures that the reconstructed hyperspectral cube is not only spectrally accurate but also spatially consistent with the semantic information. The final output is a clean HSI, which closely approximates the ground-truth HSI.

By using the method represented through the illustrated architecture, this study overcomes the limitations of existing RGB-based methods. It demonstrates that high-resolution, precise HSIs can be generated efficiently by leveraging multimodal inputs and matrix-based fusion.

4.3. Training

4.3.1. *Train setup*

The training of the diffusion model proposed in this study was performed using eight NVIDIA A6000 (64GB) GPUs. For efficient multi-GPU learning, distributed data parallelism (DDPs) of CUDA 12.8 and PyTorch 2.3.1 were constructed. In addition, to reduce GPU memory usage and improve learning speed, the training was performed with the FP16 (semi-precision floating point) operation by applying the Automatic Mixed Precision (AMP). Through this setting, high-speed and stable model learning was achieved while efficiently utilizing GPU resources, and the proposed configuration ensured scalability for handling large-scale hyperspectral datasets.

4.3.2. *Data preprocessing and training*

This model aims to generate an HSI by inputting an RGB image and segmentation. The data used in this study has a total of 512 channels, and each channel represents information in a specific wavelength range of 420–1728 nm.

The core process of data preprocessing is to reduce the dimension of HSI data and preserve key information by using Tucker decomposition. To this end, the HSI data in each folder is configured as a stack of 512-channel data, normalized, and low-dimensional data (core and factors) are generated through Tucker decomposition and later used for learning.

RGB image and segmentation are also finally used as inputs to the learning model through a normalization process. For inference, only the RGB image and segmentation data are used without Tucker decomposition to increase efficiency.

4.4. Inference

4.4.1. *Denoising steps*

The diffusion model proposed in this study uses a U-Net-based structure that is directly implemented to generate a HSI by using RGB image and segmentation as the context information. In the diffusion process, the time step *T* was set to 500, but for a balance between computational efficiency and performance during actual

inference, the step interval Δ was set to 5, and only a total of 100 steps were used.

In this study, the DDIM structure was applied for efficient inference of the diffusion model. DDIM is a method to solve the constraints that the existing DDPM must go through T steps, by adding random noise at each step. DDIM has the advantage of generating excellent quality images while reducing inference time through random noise addition and deterministic generation.

The core of DDIM is that at a given time step t, the model directly predicts a clean image Z_0^t from which noise has been removed. This enables inference beyond an arbitrary intermediate step, effectively reducing the total number of steps. In our experiments, noise intensity was balanced using cosine scheduling, as it was similarly done in the learning stage.

5. Results

5.1. Quantitative results

In order to quantitatively evaluate the performance of the proposed model, five metrics that are most commonly used in the field of image generation and HSI evaluation were used. The first metric, mean relative absolute error (MRAE), is an index obtained by normalizing the absolute error $|x_i - \hat{x}_i|$ of the actual value x_i and the predicted value $\hat{x}_i$ to the actual value x_i of the corresponding pixel. It can be calculated using Equation 7:

$$MRAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{x_i}. \tag{7}$$

Because a ratio is used, it is not affected by the unit or scale of the original data. It can be seen that the performance becomes better when the values are closer to zero.

The second metric, root mean squared error (RMSE), is the square root of the average value of the squared error between the predicted value $\hat{x}_i$ and the actual value x_i. It is defined in Equation 8:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}. \tag{8}$$

The overall stability of the prediction is evaluated using the squared error. This metric captures both small errors and infrequent large errors. Lower values indicate better prediction performance.

The third metric, the peak signal-to-noise ratio (PSNR), is a log scale representation of the ratio of noise (error) intensity to maximum signal intensity based on RMSE. It can be calculated using Equation 9:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right). \tag{9}$$

Here, MAX is the maximum value that a pixel can have, and $\mathrm{MSE} = \mathrm{RMSE}^2$. A larger value indicates less noise and better image quality.

The fourth metric, structural similarity index measure (SSIM), is a statistical measure that compares the structural similarity between two images by dividing them as shown in Equation 10:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)}. \tag{10}$$

Here, μ and σ^2 are the mean and variance of the actual value x and the predicted value $\hat{x}$, respectively. C_1 and C_2 are small constants of about 0.01 and the reason for them are for enforcing stability in the numerical calculations. SSIM denotes higher structural similarity when the value is closer to unity.

The fifth metric, the spectral angle mapper (SAM), considers the spectrum of each pixel as a vector, measures the spectral angle between the actual vector x_i and the prediction vector $\hat{x}_i$, and is calculated as shown in Equation 11 as an average value for all pixels:

$$\mathrm{SAM} = \frac{1}{N}\sum_{i=1}^{N}\arccos\left(\frac{\langle x_i, \hat{x}_i\rangle}{\|x_i\|\,\|\hat{x}_i\|}\right). \tag{11}$$

This metric uses the angles between vectors. As a result, it emphasizes the similarity of spectral shape rather than absolute brightness. A smaller SAM value is desirable for the matching spectral shape. In this study, the angle is measured in radians.

As shown in Table 1, TDiff-HSI achieved consistent improvements on the KAUST and CAVE datasets, and outperformed prior methods on ARAD-1k in terms of RMSE (0.0159), PSNR (36.12 dB), and SAM (0.0439). In contrast, its MRAE (0.1720) and SSIM (0.9212) were slightly lower than MST++. On MRAE, it was second best slightly (only 3.1%) lower than the best result given by MST++. On SSIM, it was fifth among six cases. Although our method is ranked second from the bottom in terms of SSIM (0.9212), it is important to emphasize that the differences among the top five methods are relatively small. Specifically, the SSIM values range from 0.9212 to 0.9588, meaning that the absolute gap between our approach and the highest-performing method is only 0.0376. Although our method is numerically ranked second from the bottom, the relative difference from the best-performing method is less than 4%. This suggests that, in practical terms, all the top methods-including ours-achieve highly comparable levels of structural similarity. Importantly, given the closeness of these results, we argue that other metrics should be more emphasized. From this broader perspective, our method's balanced performance better in terms of RMSE, PSNR, and SAM, makes it a reasonable and promising choice despite its marginally lower SSIM.

The relatively lower performance observed on the ARAD-1k dataset can be attributed to its higher level of complexity. Unlike datasets such as CAVE, which are captured under controlled illumination conditions, ARAD-1k consists of large-scale real-world measurements with diverse materials, complex textures, and non-uniform lighting. As a result, the performance degradation on ARAD-1k reflects the increased difficulty of the dataset rather than an inherent limitation of the proposed model.

On our 512-channel dataset, TDiff-HSI obtained MRAE 0.2169, RMSE 0.0192, PSNR 36.46 dB, SAM 0.0424, and SSIM 0.9327. Despite the number of channels being more than 16 times larger than that of the 31-channel benchmarks, the performance degradation was minor. For example, when we compare with the 31-channel ARAD-1k, the percentile difference was as follows: MRAE (0.1720→0.2169, 25.58% degraded), RMSE (0.0159→0.0192, 20.75% degraded), PSNR (36.12→36.46, 0.94% improved), SAM (0.0439→0.0424, 3.42% improved), SSIM (0.9212→0.9327, 1.25% improved). Here, (%1→%2, %3) represents the change from the baseline value (%1) to the new value (%2), and %3 is the absolute percentile difference between them. This demonstrates that Tucker decomposition remains effective even under large-channel conditions and provides stable reconstruction capability. Since the exact implementations of the other comparison methods were not available, and since they did not incorporate the material mask layer, we were unable to perform a benchmark test on our new dataset.

Table 1: The average quantitative results on each dataset. The best values are in bold and underlined, and the second-best values are underlined. For each dataset, the number in parentheses beneath its name denotes the number of image channels.

Method	Datasets	MRAE↓	RMSE↓	PSNR↑	SAM↓	SSIM↑
EDSR (Lim *et al.*, 2017)	KAUST (31ch)	0.5315	**0.0939**	**21.22**	0.2298	0.5366
MPRNet (Mehri *et al.*, 2021)		0.5352	0.1035	19.87	0.1998	0.5101
HINet (Chen *et al.*, 2021)		0.6946	0.2336	12.84	**0.1100**	0.2093
Restormer (Zamir *et al.*, 2022)		0.5605	0.1209	18.59	0.2400	0.5080
MST++ (Cai *et al.*, 2022)		**0.5018**	0.0954	20.89	0.2698	**0.5888**
TDiff-HSI (Ours)		**0.3944**	**0.0377**	**28.56**	**0.0857**	**0.7048**
EDSR (Lim *et al.*, 2017)	CAVE (31ch)	**0.4271**	**0.0419**	**27.56**	0.0794	**0.7800**
MPRNet (Mehri *et al.*, 2021)		0.4364	0.0433	27.27	0.0792	0.7580
HINet (Chen *et al.*, 2021)		0.6724	0.1169	18.64	**0.0773**	0.4176
Restormer (Zamir *et al.*, 2022)		0.4814	0.0424	27.44	0.0789	0.7077
MST++ (Cai *et al.*, 2022)		0.4782	0.0463	26.69	0.0789	0.7159
TDiff-HSI (Ours)		**0.4264**	**0.0306**	**30.42**	**0.0648**	**0.8468**
EDSR (Lim *et al.*, 2017)	ARAD-1k (31ch)	0.3335	0.0440	28.23	0.0988	0.8744
MPRNet (Mehri *et al.*, 2021)		0.1881	0.0274	33.36	**0.0902**	**0.9515**
HINet (Chen *et al.*, 2021)		0.2049	0.0305	32.48	0.0941	0.9381
Restormer (Zamir *et al.*, 2022)		0.1867	0.0277	33.34	0.0984	0.9474
MST++ (Cai *et al.*, 2022)		**0.1665**	**0.0249**	**34.37**	0.0941	**0.9588**
TDiff-HSI (Ours)		**0.1720**	**0.0159**	**36.12**	**0.0439**	0.9212
TDiff-HSI (Ours)	Ours (512ch)	0.2169	0.0192	36.46	0.0424	0.9327

In summary, Tucker-based compression resulted in only limited degradations on a few sensitive metrics, while TDiff-HSI surpassed state-of-the-art methods in overall reconstruction accuracy (RMSE, PSNR) and spectral consistency (SAM). It also exhibited robustness to channel scalability.

From a technical perspective, the proposed method benefits from the generative nature of diffusion models, which can effectively capture complex nonlinear and high-dimensional correlations between RGB and hyperspectral data through a progressive denoising process. Compared to single-step regression-based approaches, this iterative generation mechanism allows the model to internalize the underlying spectral structure better, resulting in improved spectral consistency and reconstruction accuracy.

In addition, the use of Tucker decomposition significantly reduces the dimensionality of the hyperspectral representation, which alleviates memory and computational constraints during training. This dimensionality reduction makes it feasible to employ a larger diffusion model under limited GPU resources, indirectly contributing to the overall performance improvements observed in Table 1.

In addition to the quantitative evaluation, qualitative comparisons between the generated HSIs and real HSI data are provided in Appendix A (Figures 11 and 12), where both single-object and dual-object scenarios are visually analyzed.

6. Ablation Study

6.1. Tucker decomposition shapes

Figure 8 shows how the reconstruction error (RMSE) and size reduction change according to different compression configurations determined by the selected core size in the Tucker decomposition. The compression complexity is calculated on a logarithmic scale, where R_1, R_2, and R_3 denote the core dimensions of the Tucker decomposition as defined in Figure 6. For example, if the core size is (64 128 128), the compression complexity is $\log_2(64 \times 128 \times 128) = 20$.

As the compression complexity increases (shown in the logarithmic scale), a monotonic decrease in RMSE and a corresponding reduction in compression efficiency are observed. Each point in the plot represents the mean RMSE and size-reduction ratio averaged across all rank combinations at the same compression level. The configuration marked with a red star, corresponding to (64, 128, 128), achieves RMSE $\approx$ 0.9777 and 96.32% size reduction at the log two-base, $\log_2 = 20$. The improvement in accuracy per unit loss of compression efficiency drops below 0.01 RMSE per 1% size-reduction loss. This point therefore represents a quantitative threshold at which further accuracy gains are no longer efficient, serving as a practical trade-off between reconstruction fidelity and computational efficiency.

6.2. Effect of tucker decomposition

Table 2 summarizes the impact of Tucker decomposition on tensor size and inference efficiency, and accuracy. Without Tucker decomposition, the model contained 33.6 million tensor elements, leading to an average inference time of 51.75 seconds and an accuracy of 0.2589. By contrast, Tucker decomposition reduced the effective tensor size to just 1.05 million elements (≈3.1% of the original), which enabled inference within only 2.22 seconds with an accuracy of 0.2169. Under limited GPU resources (A6000, 48 GB), the smaller input tensor made it possible to use a larger model, which increased the number of model parameters (1.98 B → 3.01 B). The gain in generative ability from the larger model was greater than the slight error increase caused by the Tucker approximation, resulting in overall better MRAE performance. These results clearly demonstrate that Tucker decomposition yields substantial improvements in both memory utilization and computational efficiency, achieving about 23 × faster while also improving reconstruction accuracy (MRAE ↓).

7. Conclusions

This study proposes TDiff-HSI, which directly generates HSIs only using the RGB and material-specific segmentation masks. Instead

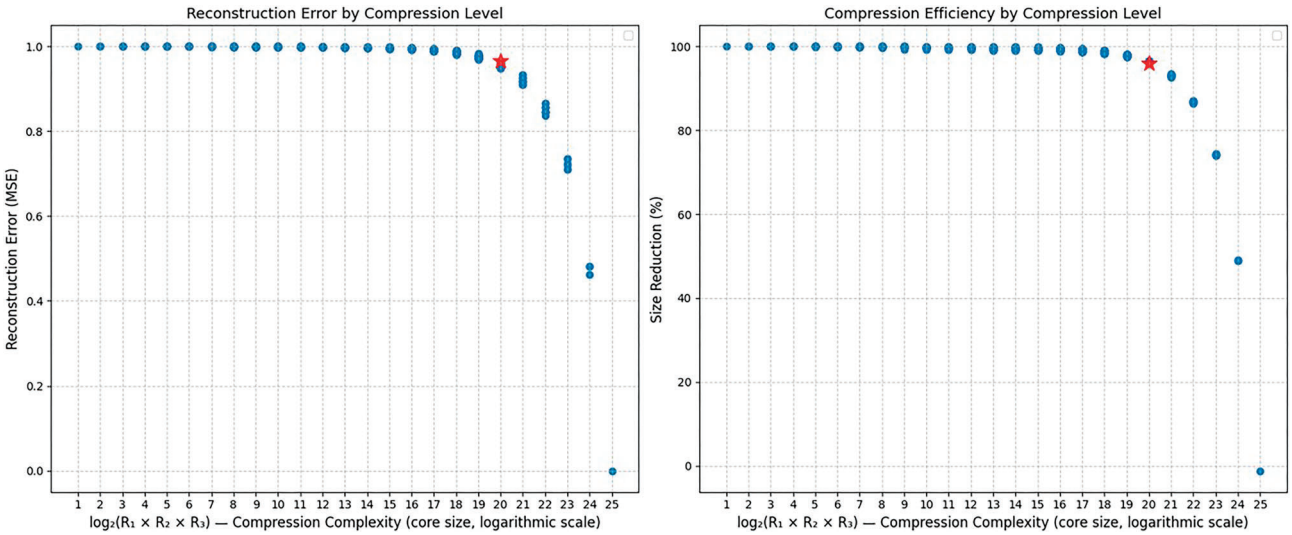

Figure 8: Mean squared error (RMSE) and size reduction vs parameter count for Tucker decomposition. Blue dots represent unique core size configurations and red star indicates the selected configuration (64, 128, 128) that demonstrates an optimal balance between model complexity and reconstruction accuracy.

Table 2: Effect of Tucker decomposition on efficiency. Tucker decomposition compresses the tensor representation to 3.1% of its original size and achieves over 20 × faster inference compared to the baseline, while also improving accuracy.

	Without Tucker	With Tucker	Improvement
Model parameters	1.98B	3.01B	–
Tensor elements	33 554 432	1 048 576	↑96.9%
Inference time	51.7504	2.2020	~23x faster
Accuracy (MRAE)	0.2589	0.2169	↑0.0420

of generating high-dimensional HSIs in naïve procedures, it utilizes low-dimensional expressions in the form of cores and factors by Tucker decomposition to secure learning stability and efficiency. It presents a practical solution to the computational and stability problems when applying a diffusion-based generator to multi-channel data such as HSI.

Through experiments, the overall stable quantitative performance was shown in single and dual object scenes. The model maintained stable reconstruction performance across various materials, with only slight local differences observed in a few cases (see Appendix A). Error differences appeared only in some local areas, and there was no consistent trend across wavelengths. Overall, the reconstructed spectra were consistent with the ground truth for each material.

In terms of reasoning, DDIM was applied, and a practical reasoning path was presented by balancing quality and speed within 100 steps. In addition, an RGB-HSI-polygon segmentation dataset consisting of a total of 78 scenarios consisting of nine materials and 12 objects is constructed and disclosed to support reproducibility and scalability of subsequent studies.

Taken together, TDiff-HSI suggests a new pathway for lightweight inputs and information-rich outputs. This direct HSI generation with only RGB with segmentation information demonstrated its practicality through comprehensive data, model, and inference design.

8. Limitation and Discussion

This study is limited to data collected in a laboratory environment with close proximity. In other words, it does not consider the wavelength-dependent light attenuation due to scattering and absorption in the air in the far-field. As the experimental scene remains in a single or dual object composition, the generalization verification for complex multi-object and composite material scenes is additionally required. Furthermore, the evaluation was conducted for thin objects and with perpendicular orthogonal projection. Performance degradation may occur under conditions where geometric optical effects such as thick translucent objects or self-shaded and strong reflections are dominant. In addition, this study aims to generate a HSI of 420–1728 nm as an RGB image in the visible light region. Precise material-wise segmentation is essential to generate information on wavelengths outside the visible light region, and currently, only manual annotations are available.

Conflicts of 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.

Author Contributions

Jaeik Bae: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing-original draft. **Yong-Gu Lee:** Conceptualization, Methodology, Supervision, Writing-review and editing.

Funding

This work was partially supported by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD) (UD230017TD); the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01842, Arti-

ficial Intelligence Graduate School Program (GIST)); Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) [P0020535, The Competency Development Program for Industry Specialist]; the InnoCORE program of the Ministry of Science and ICT (25-InnoCORE-01); and by the Regional Innovation System & Education (RISE) program through the Gwangju RISE Center, funded by the Ministry of Education (MOE) and the Gwangju Metropolitan City, Republic of Korea (2025-RISE-05-001).

Data Availability

The data underlying this article are available in the TDiff-HSI project repository at https://github.com/JaeikBae/TDiff-HSI

References

Allegretta, I., Legrand, S., Alfeld, M., Gattullo, C. E., Porfido, C., Spagnuolo, M., Janssens, K., & Terzano, R. (2022). SEM-EDX hyperspectral data analysis for the study of soil aggregates. *Geoderma*, **406**, 115540. https://doi.org/10.1016/j.geoderma.2021.115540.

Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Van Gool, L., Liu, S., Li, Y., Feng, C., Lei, L., Li, J., Du, S., Wu, C., & Mansoor Roomi, S. M. (2022). NTIRE 2022 spectral recovery challenge and data set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*(pp. 863–881). https://doi.org/10.1109/CVPRW56347.2022.00102.

Baig, N., Fabelo, H., Ortega, S., Callico, G. M., Alirezaie, J., & Umapathy, K. (2021). Empirical mode decomposition based hyperspectral data analysis for brain tumor classification. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*(pp. 2274–2277). IEEE. https://doi.org/10.1109/EMBC46164.2021.9629676.

Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., & Van Gool, L. (2022). MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*(pp. 745–755). https://doi.org/10.48550/arXiv.2204.07908.

Chen, L., Lu, X., Zhang, J., Chu, X., & Chen, C. (2021). Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*(pp. 182–192). https://doi.org/10.1109/CVPRW53098.2021.00027.

Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE transactions on geoscience and remote sensing*, **54**, 6232–6251. https://doi.org/10.1109/TGRS.2016.2584107.

Chen, Y., & Zhang, X. (2024). DDSR: Degradation-aware diffusion model for spectral reconstruction from RGB images. *Remote Sensing*, **16**, 2692. https://doi.org/10.3390/rs16152692.

Choi, J. Y., Cho, J. S., Park, K. J., Kim, S. S., & Lim, J. H. (2022). Grading the pungency of red pepper powder using hyperspectral imaging coupled with multivariate analysis. *Food Science and Preservation*, **29**, 918–931. https://doi.org/10.11002/kjfp.2022.29.6.918.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., & Phan, H. A. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, **32**, 145–163. https://doi.org/10.1109/MSP.2013.2297439.

Dao, P. D., Mantripragada, K., He, Y., & Qureshi, F. Z. (2021). Improving hyperspectral image segmentation by applying inverse noise weighting and outlier removal for optimal scale selection. *ISPRS Journal of Photogrammetry and Remote Sensing*, **171**, 348–366. https://doi.org/10.1016/j.isprsjprs.2020.11.013.

Fu, Y., Zhang, T., Zheng, Y., Zhang, D., & Huang, H. (2019). Hyperspectral image super-resolution with optimized RGB guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*(pp. 11661–11670). https://doi.org/10.1109/CVPR.2019.01193.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, **63**, 139–144. https://doi.org/10.1145/3422622.

Grigoroiu, A., Yoon, J., & Bohndiek, S. E. (2020). Deep learning applied to hyperspectral endoscopy for online spectral classification. *Scientific Reports*, **10**, 3947. https://doi.org/10.1038/s41598-020-60574-6.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, **33**, 6840–6851. https://doi.org/10.48550/arXiv.2006.11239.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. *Advances in Neural Information Processing Systems*, **35**, 8633–8646. https://doi.org/10.48550/arXiv.2204.03458.

Hsu, P. H. (2007). Feature extraction of hyperspectral images using wavelet and matching pursuit. *ISPRS Journal of Photogrammetry and Remote Sensing*, **62**, 78–92. https://doi.org/10.1016/j.isprsjprs.2006.12.004.

Hu, W., Huang, Y., Wei, L., Zhang, F., & Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, **2015**, 258619. https://doi.org/10.1155/2015/258619.

Hupel, T., & Stütz, P. (2022). Adopting hyperspectral anomaly detection for near real-time camouflage detection in multispectral imagery. *Remote Sensing*, **14**, 3755. https://doi.org/10.3390/rs14153755.

Jo, H., Lee, J. K., Lee, Y. C., & Choo, S. (2024). Generative artificial intelligence and building design: Early photorealistic render visualization of façades using local identity-trained models. *Journal of Computational Design and Engineering*, **11**, 85–105. https://doi.org/10.1093/jcde/qwae017.

Kim, H., & Kim, S. (2016). Band selection for plastic classification using NIR hyperspectral image. In *2016 16th International Conference on Control, Automation and Systems (ICCAS)*(pp. 302–304). IEEE. https://doi.org/10.1109/ICCAS.2016.7832335.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51**, 455–500. https://doi.org/10.1137/07070111X.

Lee, J. B., Kim, E. S., & Lee, S. H. (2014). An analysis of spectral pattern for detecting pine wilt disease using ground-based hyperspectral camera. *Korean Journal of Remote Sensing*, **30**, 665–675. https://doi.org/10.7780/kjrs.2014.30.5.11.

Li, Y., Fu, Q., & Heidrich, W. (2021). Multispectral illumination estimation using deep unrolling network. In *Proceedings of the IEEE/CVF international conference on computer vision*(pp. 2672–2681). https://doi.org/10.1109/ICCV48922.2021.00267.

Li, Z., Xiong, F., Zhou, J., Wang, J., Lu, J., & Qian, Y. (2020). BAE-Net: A band attention aware ensemble network for hyperspectral object tracking. In *2020 IEEE international Conference on image processing (ICIP)*(pp. 2106–2110). IEEE. https://doi.org/10.1109/ICIP40778.2020.9191105.

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern*

recognition workshops(pp. 136–144). https://doi.org/10.1109/CVPRW.2017.151.

Ling, Q., Li, K., Li, Z., Lin, Z., & Wang, J. (2022). Hyperspectral detection and unmixing of subpixel target using iterative constrained sparse representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **15**, 1049–1063. https://doi.org/10.1109/JSTARS.2022.3140389.

Liu, Z., Wang, X., Shu, M., Li, G., Sun, C., Liu, Z., & Zhong, Y. (2021). An anchor-free Siamese target tracking network for hyperspectral video. In *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*(pp. 1–5). IEEE. https://doi.org/10.1109/WHISPERS52202.2021.9483958.

Mehri, A., Ardakani, P. B., & Sappa, A. D. (2021). MPRNet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*(pp. 2704–2713). https://doi.org/10.1109/WACV48630.2021.00275.

Nascimento, S. M., Amano, K., & Foster, D. H. (2016). Spatial distributions of local illumination color in natural scenes. *Vision Research*, **120**, 39–44. https://doi.org/10.1016/j.visres.2015.07.005.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *Proceedings of the 39th International Conference on Machine Learning*, **162**, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html.

Park, S., Yang, M., Yim, D. G., Jo, C., & Kim, G. (2023). VIS/NIR hyperspectral imaging with artificial neural networks to evaluate the content of thiobarbituric acid reactive substances in beef muscle. *Journal of Food Engineering*, **350**, 111500. https://doi.org/10.1016/j.jfoodeng.2023.111500.

Pizurica, A., Bellens, R., Gautama, S., & Philips, W. (2014). Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. *IEEE Geoscience and Remote Sensing Letters*, **12**, 552–556. https://doi.org/10.1109/LGRS.2014.2350263.

Ren, J., Zabalza, J., Marshall, S., & Zheng, J. (2014). Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. *IEEE Signal Processing Magazine*, **31**, 149–154. https://doi.org/10.1109/MSP.2014.2312071.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*(pp. 10684–10695). https://doi.org/10.1109/cvpr52688.2022.01042.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Seyed Ghasemipour, S. K., Karagol Ayan, B., Mahdavi, S. S., Gontijo Lopes, R., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, **35**, 36479–36494. https://doi.org/10.48550/arXiv.2205.11487.

Shi, M., Seo, J., Cha, S. H., Xiao, B., & Chi, H. L. (2024). Generative AI-powered architectural exterior conceptual design based on the design intent. *Journal of Computational Design and Engineering*, **11**, 125–142. https://doi.org/10.1093/jcde/qwae077.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., & Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, **65**, 3551–3582. https://doi.org/10.1109/TSP.2017.2690524.

Son, Y. S., Noh, S. G., Bang, E. S., Kim, K. E., Cho, S. J., & Baik, H. (2022). Ground-based visible-near infrared hyperspectral imaging for monitoring cliff weathering of a volcanic island in Dokdo, South Korea. *Engineering Geology*, **309**, 106854. https://doi.org/10.1016/j.enggeo.2022.106854.

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *In International Conference on Learning Representations*. https://openreview.net/forum?id=St1giarCHLP.

Tamilarasan, K., Anbazhagan, S., Maheswaran, S. U., Ranjithkumar, S., Kusuma, K., & Rajesh, V. (2022). Reflectance spectra and AVIRIS-NG airborne hyperspectral data analysis for mapping ultramafic rocks in igneous terrain. *Journal of Spectral Imaging*, **11**, a9. https://doi.org/10.1255/jsi.2022.a9.

Van Nguyen, H., Banerjee, A., & Chellappa, R. (2010). Tracking via object reflectance using a hyperspectral video camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*(pp. 44–51). IEEE. https://doi.org/10.1109/CVPRW.2010.5543780.

Xiong, F., Zhou, J., & Qian, Y. (2020). Material based object tracking in hyperspectral videos. *IEEE Transactions on Image Processing*, **29**, 3719–3733. https://doi.org/10.1109/TIP.2020.2965302.

Yang, J. X., Zhou, J., Wang, J., Tian, H., & Liew, A. W. C. (2024). LiDAR-guided cross-attention fusion for hyperspectral band selection and image classification. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1–15. https://doi.org/10.1109/TGRS.2024.3389651.

Yasuma, F., Mitsunaga, T., Iso, D., & Nayar, S. K. (2010). Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, **19**, 2241–2253. https://doi.org/10.1109/TIP.2010.2046811.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*(pp. 5728–5739). https://doi.org/10.1109/CVPR52688.2022.00564.

Zeng, Y., Zheng, Z., Zhang, T., Huang, X., & Lu, X. (2024). AI-powered fire engineering design and smoke flow analysis for complex-shaped buildings. *Journal of Computational Design and Engineering*, **11**, 359–373. https://doi.org/10.1093/jcde/qwae053.

Zhang, L., Luo, X., Li, S., & Shi, X. (2023). R2H-CCD: Hyperspectral imagery generation from RGB images based on conditional cascade diffusion probabilistic models. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* (pp. 7392–7395). IEEE. https://doi.org/10.1109/IGARSS52108.2023.10281589.

Appendix A. Qualitative Results

The qualitative performance of the proposed TDiff-HSI was examined through spectral and spatial analyses. In Figure 9, the model reliably reproduced characteristic absorption bands. For instance, the absorption of plastic near 1200 nm, iron between 900 and 1200 nm, and water around 1000 nm was consistently recovered, demonstrating close agreement with the ground truth, which is denoted by a red circle. Although RGB images contain information within the visible spectrum (approximately 400–700 nm), this information is inherently compressed into three channels and does not fully preserve the fine-grained spectral variations required for hyperspectral reconstruction at 2–3 nm resolution. In addition, the mapping between RGB and hyperspectral data is not one-to-one, which can lead to residual discrepancies from the ground truth even within the visible wavelength range.

Figure 10 shows the reconstruction error maps of various material samples. Most samples exhibit low reconstruction errors, indicating that the Tucker-based approximation accurately captures the overall diffusion structure. In Figure 10 (A), the iron sample shows slightly higher errors along its edge (around RMSE 0.3162). This is not a problem with the model itself, but rather

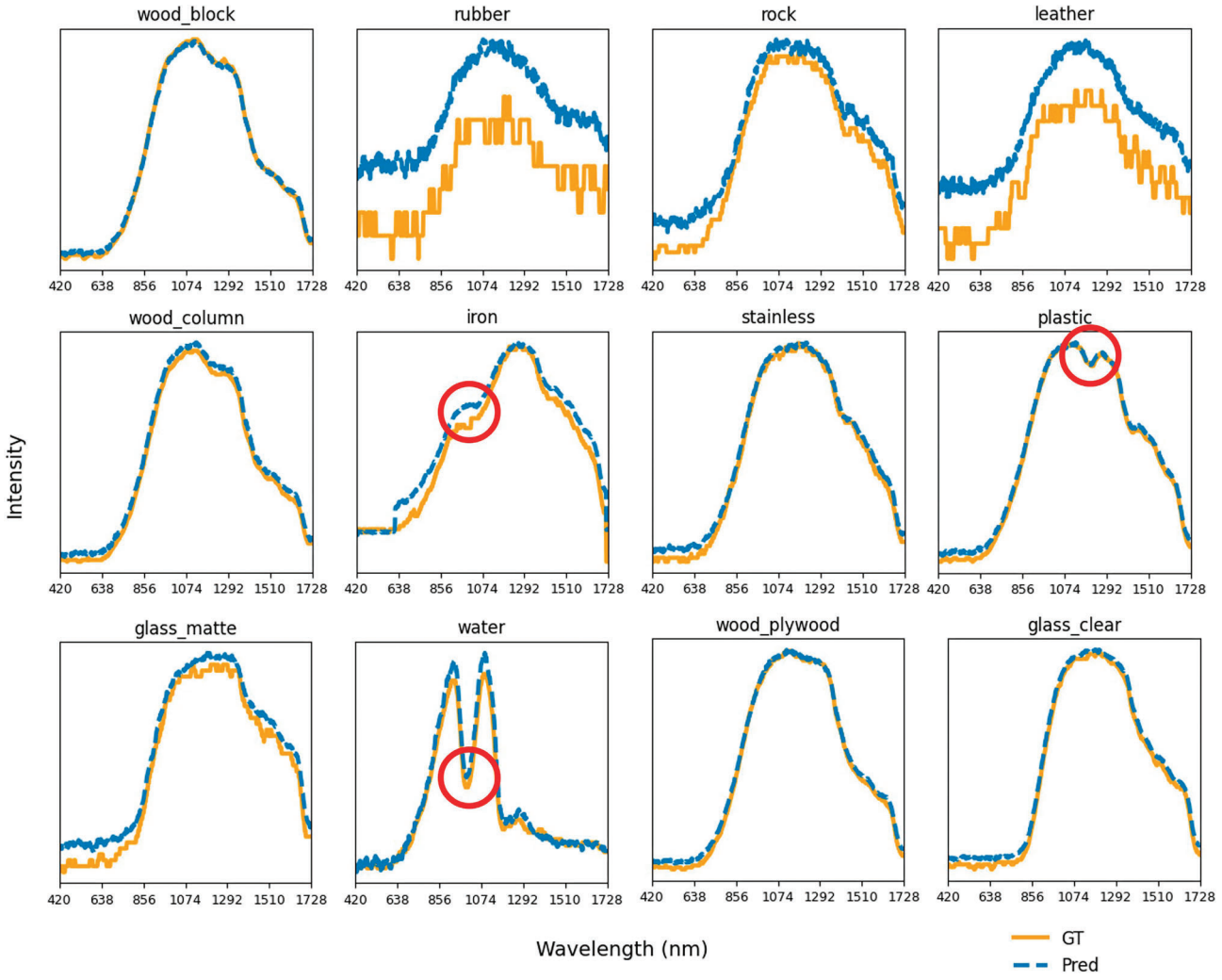

Figure 9: Comparison of generated spectra for each material. The orange curve denotes the ground truth, and the blue curve denotes the generated spectrum. Red circles indicate three representative regions where the characteristic spectral features of each material are most distinct.

a minor local error that appeared during reconstruction. In Figure 10 (B), horizontal and vertical line patterns appear across several samples. These lines formed as small errors become amplified during the multiplication of the decomposed core tensor and factor matrices, which corresponds to the 'Matrix Multiplication' process illustrated in Figure 6. Overall, the results show that while the model reconstructs the global structure well, minor direction-aligned artifacts can appear as a side effect of the Tucker decomposition.

Figure 11 shows channel-wise spatial visualization of all single object scenes. In each row, upper side shows ground truth. And lower side shows generated image. Here, the ground truth (GT) images correspond to HSIs directly captured by the real HSI camera. It demonstrates the model's ability to generate high-quality images even at object edges and in shadow regions.

Figure 12 shows channel-wise spatial visualization of all dual object scenes. It also demonstrates the overall model's ability to generate high-quality images.

In summary, three main conclusions emerge. First, the core spectral signatures of each material exhibited strong fidelity to the ground truth. Second, spatial quality remained stable overall, with only localized degradation under challenging illumination or boundary conditions. These findings are consistent with the quantitative evaluation and confirm that the proposed model effectively restores meaningful spectral cues from RGB and segmentation inputs.

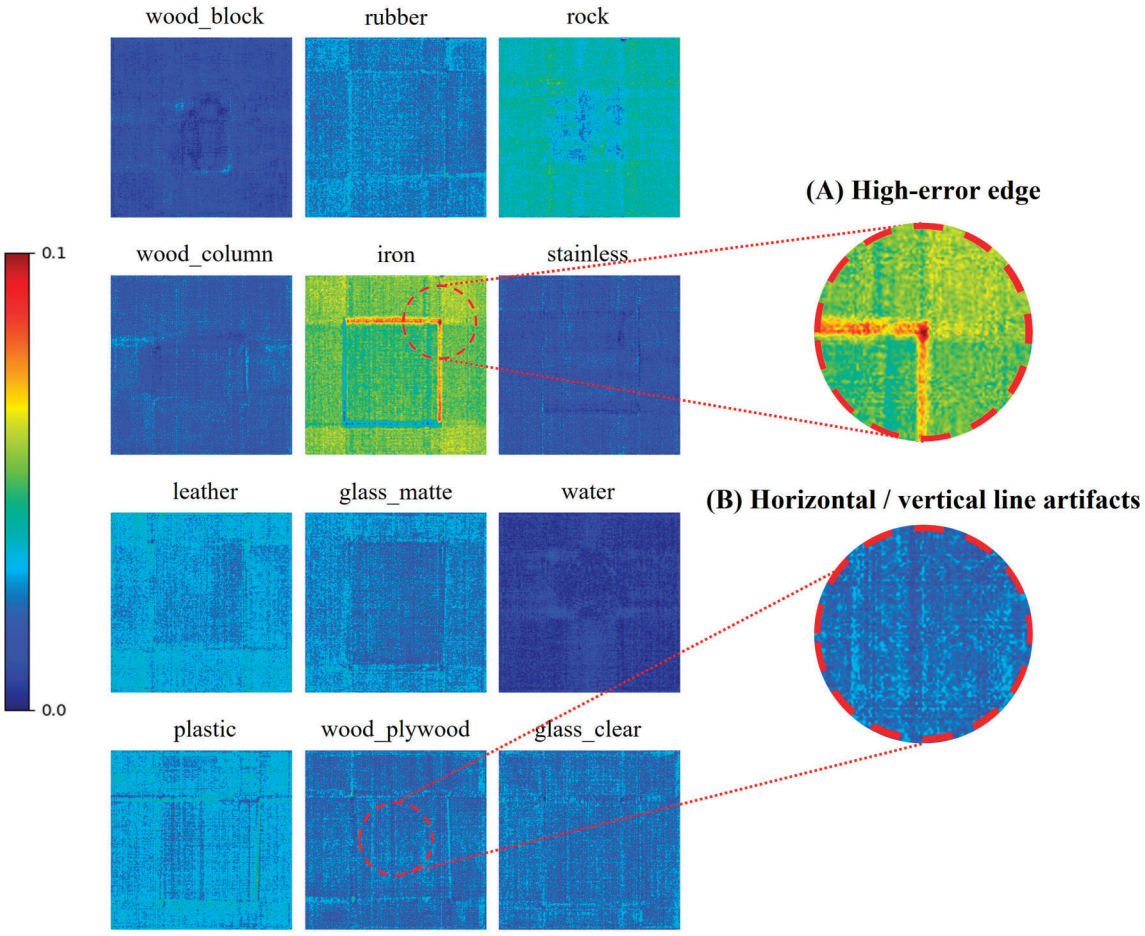

Figure 10: Reconstruction error maps for different materials in a single object. (A) High-error edge in the iron sample (RMSE ≈ 0.3162). (B) Horizontal and vertical line artifacts are commonly observed across samples.

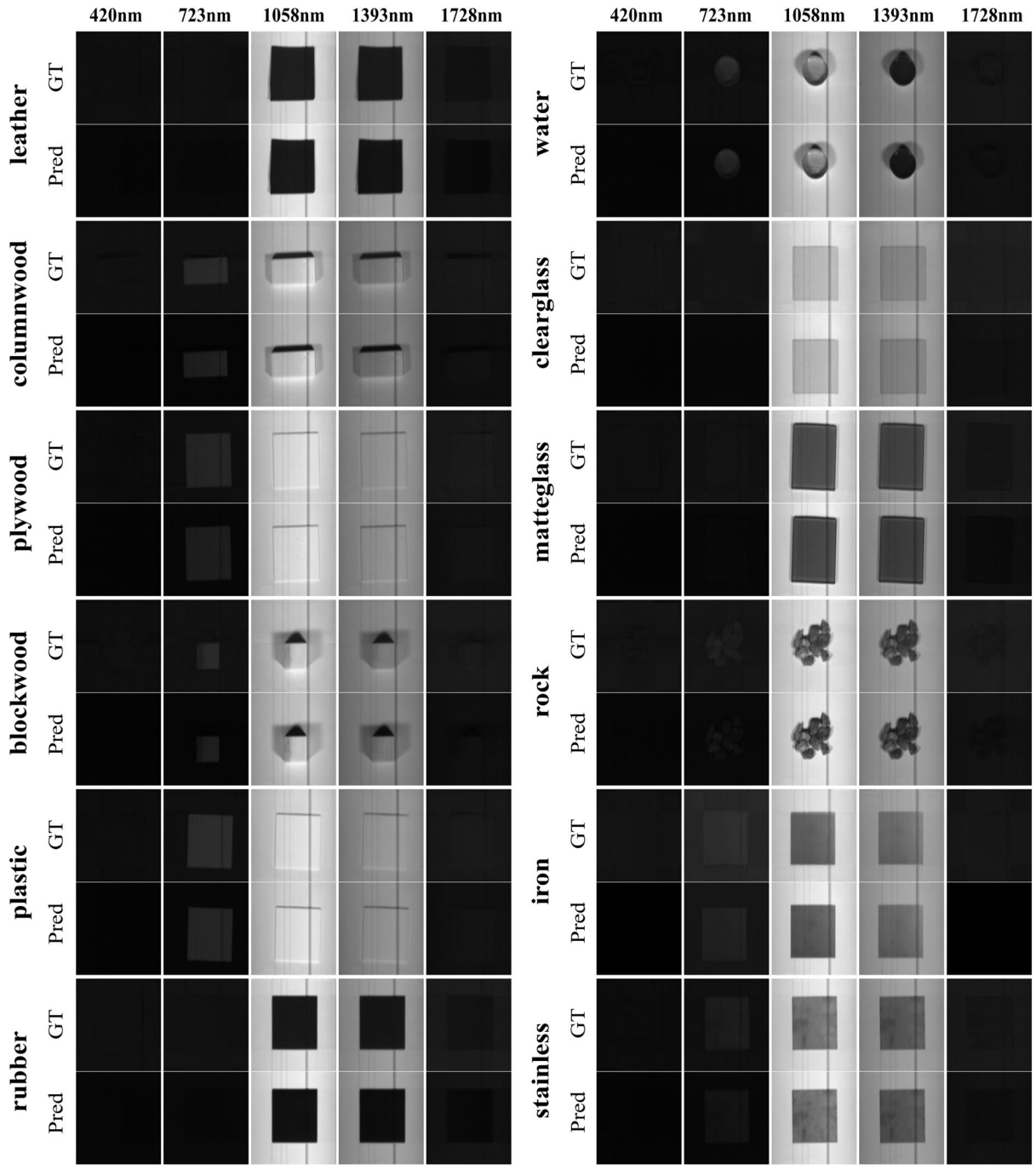

Figure 11: Channel-wise visualization of HSI generation. Each row corresponds to different materials, and each column shows representative spectral channels. Ground truth and predicted images are compared to illustrate the restoration performance of TDiff-HSI.

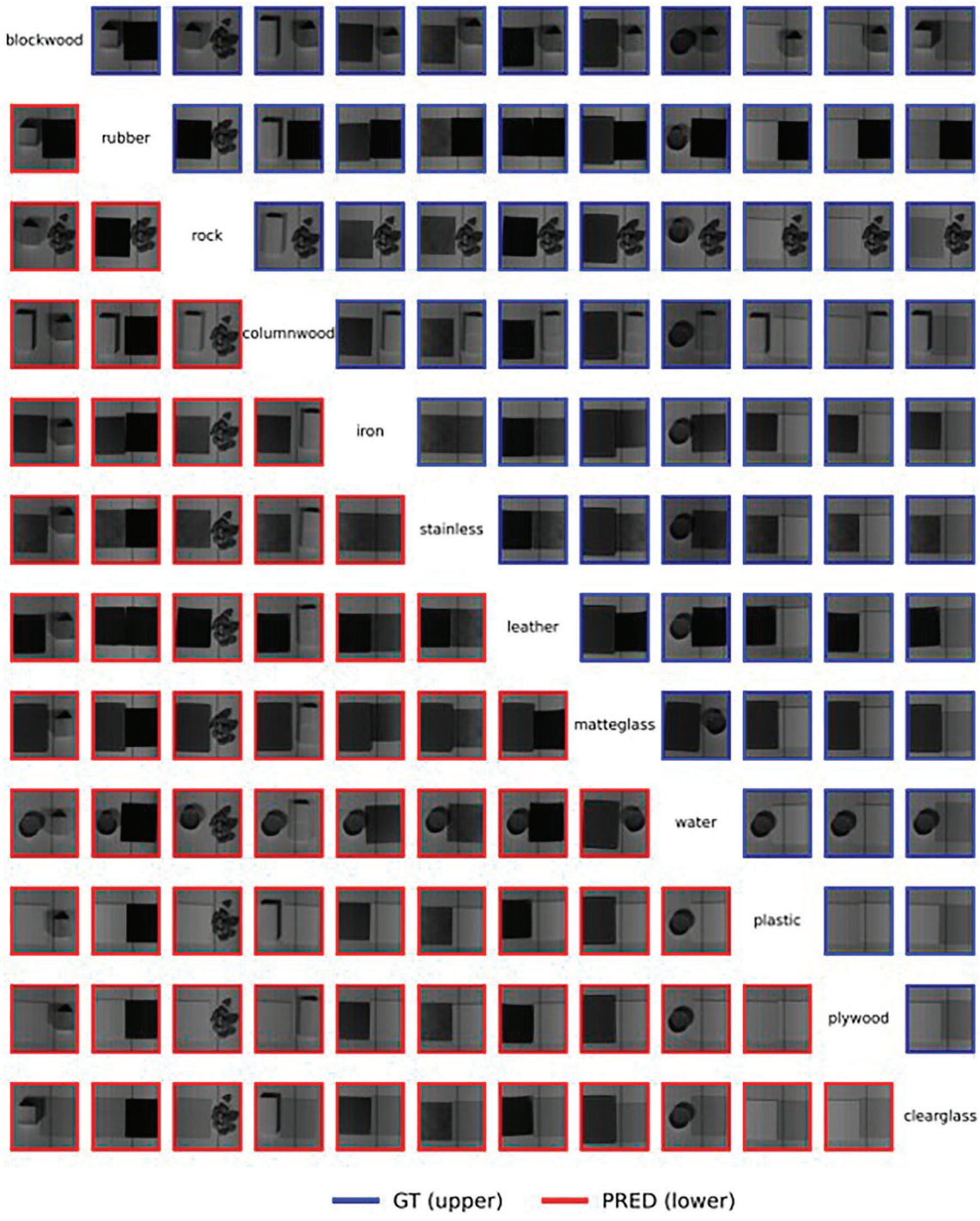

Figure 12: Material-specific spatial comparisons between ground truth (GT, upper) and predicted spectra (PRED, lower). Blue and red bounding boxes denote GT and prediction, respectively. The results demonstrate that the proposed model stably recovers material-dependent absorption while preserving object boundaries.

Appendix B. Analysis of Quantitative Results by Material

Figure 13 presents the quantitative reconstruction performance of the proposed method across different materials. Unlike Table 1, which summarizes results at the sample level, this figure provides material-level statistics to analyze how reconstruction quality varies by material type.

In this analysis, dual-object samples are included in the results of both constituent materials to ensure fair comparison. For example, a sample labeled plastic_rubber contributes to both the plastic and rubber categories when computing the material-wise averages. As a result, some numerical differences compared to Table 1 are expected, since overlapping samples are intentionally counted more than once for material-level evaluation.

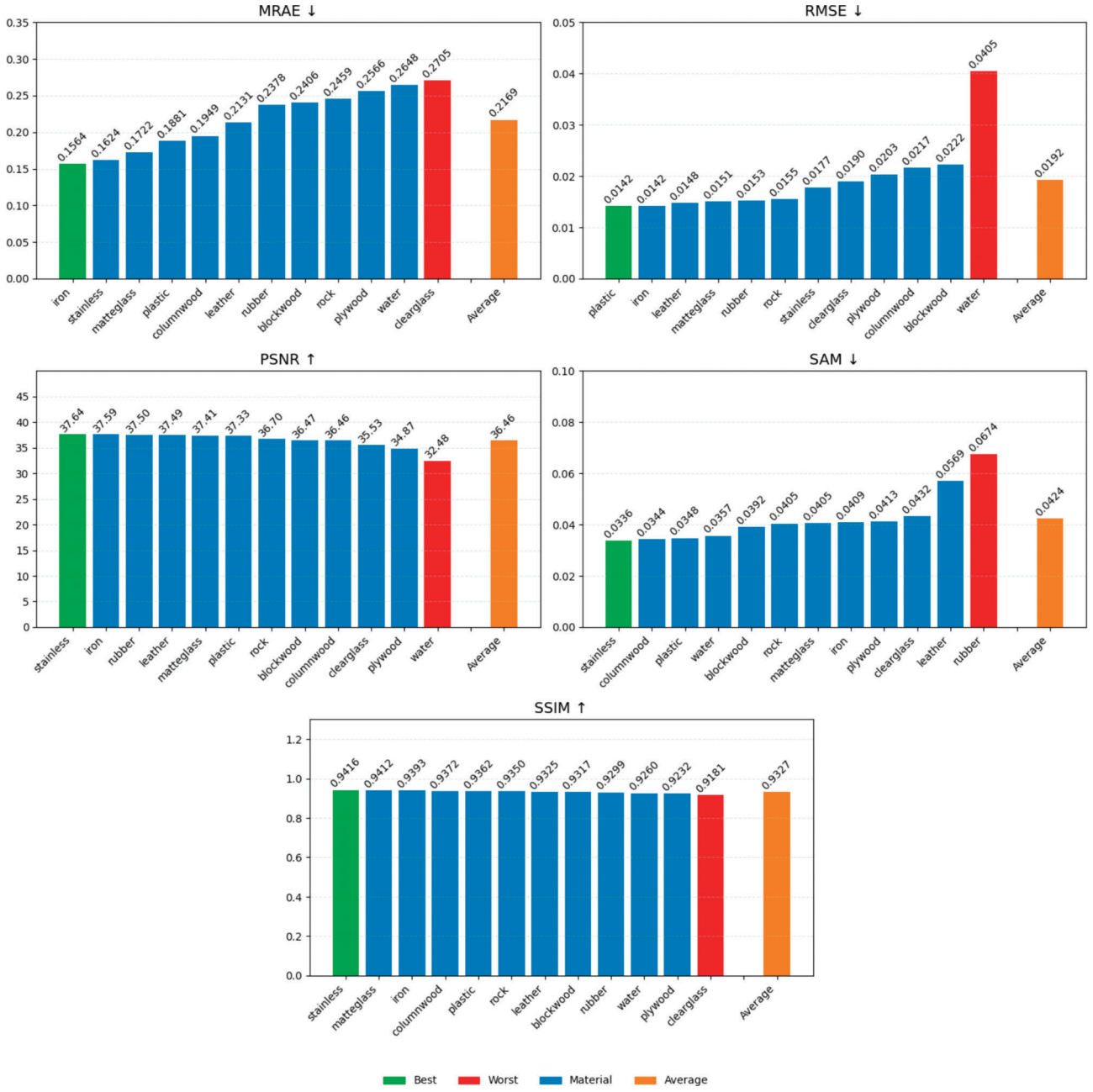

Figure 13: Quantitative evaluation results for dual-object samples. Each bar represents the reconstruction quality for a specific material. Dual-object samples (e.g. plastic_rubber) are counted once for each constituent material to evaluate per-material performance trends. Metrics include MRAE ↓, RMSE ↓, PSNR ↑, SAM ↓, and SSIM ↑, where arrows indicate the desirable direction of improvement.

Received: October 25, 2025. **Revised:** January 26, 2026. **Accepted:** January 28, 2026