GEWDiff: Geometric Enhanced Wavelet-based Diffusion Model for Hyperspectral Image Super-resolution

Sirui Wang[1], Jiang He[1,2], Natàlia Blasco Andreo[3], Xiao Xiang Zhu [1,2*]

[1]Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany
[2]Munich Center for Machine Learning, 80333 Munich, Germany
[3]Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain
sirui.wang@tum.de, Natalia.Blasco@uab.cat, jiang.he@tum.de,xiaoxiang.zhu@tum.de

Abstract

Improving the quality of hyperspectral images (HSIs), such as through super-resolution, is a crucial research area. However, generative modeling for HSIs presents several challenges. Due to their high spectral dimensionality, HSIs are too memory-intensive for direct input into conventional diffusion models. Furthermore, general generative models lack an understanding of the topological and geometric structures of ground objects in remote sensing imagery. In addition, most diffusion models optimize loss functions at the noise level, leading to a non-intuitive convergence behavior and suboptimal generation quality for complex data. To address these challenges, we propose a Geometric Enhanced Wavelet-based Diffusion Model (GEWDiff), a novel framework for reconstructing hyperspectral images at 4-times super-resolution. A wavelet-based encoder-decoder is introduced that efficiently compresses HSIs into a latent space while preserving spectral-spatial information. To avoid distortion during generation, we incorporate a geometry-enhanced diffusion process that preserves the geometric features. Furthermore, a multi-level loss function was designed to guide the diffusion process, promoting stable convergence and improved reconstruction fidelity. Our model demonstrated state-of-the-art results across multiple dimensions, including fidelity, spectral accuracy, visual realism, and clarity.

Code — https://github.com/zhu-xlab/GEWDiff

Introduction

Hyperspectral images (HSIs) offer a unique perspective by capturing continuous spectral features of ground objects. Despite advancements in research, the high costs and low coverage of super-resolution (SR) hyperspectral data limit their applications. Currently, open-access hyperspectral airborne data are predominantly regional, typically focused on several cities, increasing their exclusivity and cost. Hyperspectral satellites have better coverage but suffer from insufficient spatial resolution. Improving the spatial resolution of hyperspectral satellite images is, therefore, crucial to allow for fully harnessing the potential of hyperspectral data in Earth observation (EO). Many fusion models that combine hyperspectral and multispectral images (MSIs) have been developed. However, the fusion model cannot generate HSIs in any region of interest without any prior knowledge offered by VHR RGB data. Most fusion models can obtain 10 m resolution hyperspectral data with MSIs, such as Sentinel 2, but to get a spatial resolution from 10 to 2.5 m, we focused on the single-image-generation methods for HSIs at 4-times super-resolution.

Traditional hyperspectral image (HSI) super-resolution approaches typically rely on interpolation techniques, such as nearest neighbor or bilinear interpolation. While straightforward and computationally efficient, these methods fail to capture the complex, nonlinear relationships present in high-dimensional spectral data. Recently, deep learning-based methods, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) (Sidorov and Yngve Hardeberg 2019; Li, Wang, and Li 2020; Hou et al. 2022; Shi et al. 2022; Li et al. 2020), have emerged as powerful alternatives. These models are capable of learning expressive spectral-spatial representations directly from data, leading to significant improvements in reconstruction accuracy over classical methods. Recent studies have shown the strong potential of transformers (Zhang et al. 2023; Chen, Zhang, and Zhang 2023; Yu et al. 2023; Su et al. 2025) in vision tasks, due to their ability to model long-range spatial and spectral dependencies. However, a common limitation shared by the aforementioned methods is their difficulty in generating rich textures and complex spatial structures, despite their strong performance in preserving spectral fidelity.

Diffusion models have recently demonstrated remarkable success in generating high-quality natural images, as seen in models such as the Stable Diffusion (Rombach et al. 2021) and DiT (Peebles and Xie 2023) frameworks. Motivated by their strong generative capabilities and robustness in modeling complex distributions, researchers have begun exploring their applicability to hyperspectral image super-resolution. For instance, SpectralDiff (Chen et al. 2023) and HSR-Diff (Wu et al. 2023) extend the diffusion paradigm directly to the hyperspectral domain, aiming to better model spectral-spatial correlations. However, despite recent progress, adapting diffusion models to hyperspectral image generation remains a significant challenge. Unlike natural or multispectral images, HSI data typically exhibit a much lower signal-to-noise ratio and higher spectral dimensionality, making it difficult to design diffusion architectures

*Corresponding author.

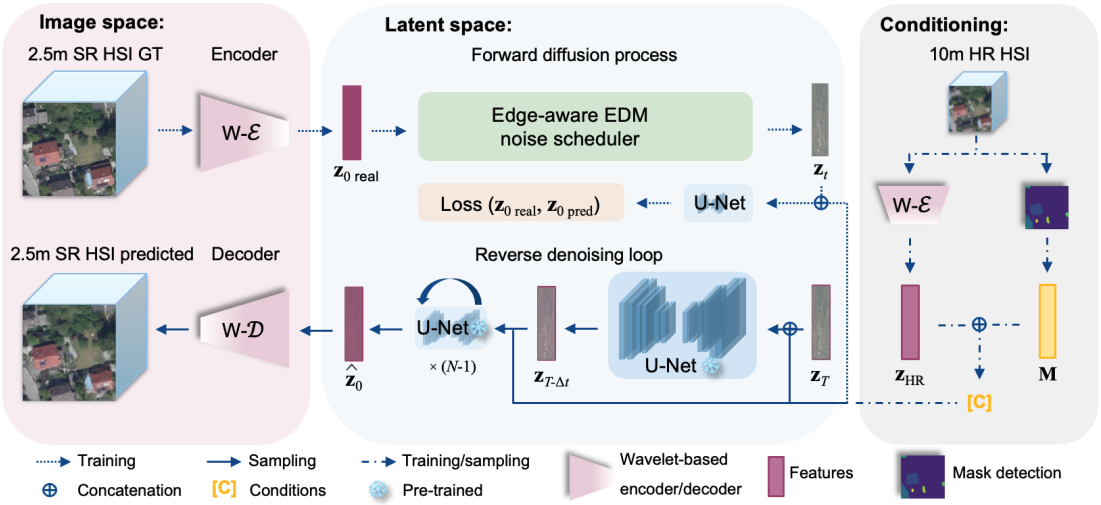

Figure 1: Illustration of the Geometric Enhanced Wavelet-based Diffusion Model pipeline.

that balance spatial fidelity, spectral accuracy, and visual realism. Existing architectures often suffer from a slow model convergence speed, excessive sampling steps, and a need for high computational GPU memory, which limits their practicality in real-world scenarios. We propose a Geometric Enhanced Wavelet-based Diffusion Model (GEWDiff) to address these challenges, as shown in Figure 1. The main contributions of this study are summarized as follows:

- **An efficient wavelet-based encoder-decoder** will almost losslessly transform hyperspectral data into a latent space by decomposing the input into multiple frequency levels. This method preserves spectral-spatial information while reducing channel dimensionality without long-term training.
- **Structural invariance control via the diffusion process:** An edge-aware noise scheduler was designed to improve generation efficiency and accuracy during training. Mask conditioning ensures preservation of the geometric integrity and prevents distortion.
- **A multi-level loss function comprising pixel-wise loss, perceptual loss, and gradient loss:** Each part of the loss function contributes to a balanced convergence speed. The loss function also enables the evaluation and alignment of predicted and ground-truth images on specific details and semantic understanding.

Related Work

Properties of wavelet-based diffusion models

Wavelet-based diffusion models have recently gained increasing attention due to their suitability for image generation. For example, WaveDiff (Phung, Dao, and Tran 2023) demonstrated that applying diffusion in the wavelet domain allows images to be compressed into a structured latent space, enabling almost-lossless reconstruction while significantly reducing computational overheads. This approach not only preserves high-fidelity details but also offers substantial memory savings, making it especially advantageous for large-scale diffusion models. Building on this, Zhao et al. (Zhao et al. 2024) proposed a parallel diffusion strategy that separates high-frequency and low-frequency components for underwater image restoration. Shi et al. (Shi et al. 2024) proposed WaveDiffUR, a wavelet-domain diffusion model for remote sensing ultra-resolution (UR), which involved reformulating high-magnification SR as a conditional stochastic differential equation (SDE) solved via iterative wavelet decomposition, integrating pre-trained SR modules for scalability and a cross-scale pyramid (CSP) constraint to preserve spectral-spatial fidelity. Si et al. (Si et al. 2025) proposed CASSIDiff, the first diffusion model for CASSI hyperspectral reconstruction, which integrated a DWT-based feature fusion mechanism to reduce noise and a spectral-spatial attention module to capture spectral correlations. Despite the success of various wavelet-based diffusion models in natural and remote sensing images, their application to hyperspectral image generation remains unexplored.

Diffusion model for hyperspectral image super-resolution

Recent work has proposed adapting the diffusion process to better fit the characteristics of HSI data. As such, existing approaches can be broadly categorized into three paradigms: two-stage models, grouped autoencoder models, and end-to-end frameworks (Wang et al. 2023). Two-stage models decompose the super-resolution task into two separate subtasks handled by distinct networks. For example, HSI-Gene (Pang et al. 2024b) first generates high-resolution RGB bands from the input HSI, and then fuses them with the low-resolution HSI to reconstruct the final output. This modular design helps reduce computational complexity and allows for flexible training. Grouped models partition the spectral bands into groups and process them in parallel, often using autoencoder-style architectures. DMGASR (Wang et al. 2024), for instance, employs spectral grouping and trains separate VAE-based diffusion modules for different groups, enabling scalable training across high-dimensional

spectra while preserving inter-band correlations. End-to-end frameworks perform full-spectrum reconstruction in a single model, often incorporating various strategies to manage complexity and improve the expressiveness. For example, HIR-Diff (Pang et al. 2024a) integrates a codebook with singular value decomposition to compress and guide the generation process. MTLSC-Diff (Qu et al. 2024) uses classification maps as spatial priors to improve the generation accuracy. LSDiff (Cheng et al. 2024) applies the diffusion process in a compressed latent space, which allows reducing memory usage while maintaining the generation quality. Although some methods have been explored, most rely on two-stage training or fail to simultaneously ensure spectral fidelity and visual quality, while our model addresses both challenges effectively.

Method

Wavelet-based encoder and decoder

Our training-free encoder-decoder is based on Regression wavelet analysis (RWA), first proposed for lossless hyperspectral image compression by (Amrani et al. 2016). RWA applies a predefined number of Haar wavelet (Haar 1910) decompositions intercalated with a linear regression of the spectral dimension to exploit the redundancy that still remains in the discrete wavelet transform (DWT) domain to further compress the data. RWA can compress HSIs with a more efficient, lossless, or near-lossless transform by storing the prediction error. The structure is shown in Figure 2.

Encoder. For the super-resolution task, RWA allows us to reduce the number of bands given to the diffusion model by using the Haar wavelet. Let $\mathbf{I}_{\text{LR}}$ be the input 10 m high-resolution hyperspectral image, the J-th level RWA transform can be represented as:

$$(\mathbf{V}^J_{\text{LR}}, (\mathbf{w}^j_{\text{LR}})^{1\leqslant j\leqslant J}) = \text{RWA}(\mathbf{I}_{\text{LR}}, J), \tag{1}$$

$$\hat{\mathbf{w}}^j_i = \beta^j_{i,0} + \beta^j_{i,1}\mathbf{V}^j_1 + ... + \beta^j_{i,k}\mathbf{V}^j_k, \tag{2}$$

$$\min||\mathbf{w}^j_i - \hat{\mathbf{w}}^j_i||_2, \tag{3}$$

where $\mathbf{w}^j_i$ represents the i-th details (high-coefficient) of the j-th level wavelet transform and $\hat{\mathbf{w}}^j_i$ its prediction; $\mathbf{V}^j_k$ represents the k-th low-coefficient (main-coefficient) of the j-th level wavelet transform and $\beta^j_{i,k}$ the linear regression coefficients that will be learned to adjust the linear regression. Contrary to traditional RWA, where the residuals

$$\mathbf{W}^j_{\text{LR}} = \mathbf{w}^j_{\text{LR}} - \hat{\mathbf{w}}^j_{\text{LR}}, \tag{4}$$

are computed in order to fully recover the original signal, the proposed encoder will only store $\mathbf{V}^J_{\text{LR}}$ and the weights of all the adjusted linear models $\mathbf{B}_{\text{LR}} = [\beta^{1\leqslant j\leqslant J}]$, where J is the level of wavelet transforms applied. The main coefficients $\mathbf{V}^J_{\text{LR}}$, which contain the most critical information, are used as input for the principal component analysis (PCA). The following PCA transformation enables a more efficient compression of hyperspectral imagery (HSI) by achieving a higher compaction factor while preserving more information. Furthermore, it can convert the sparse wavelet-based

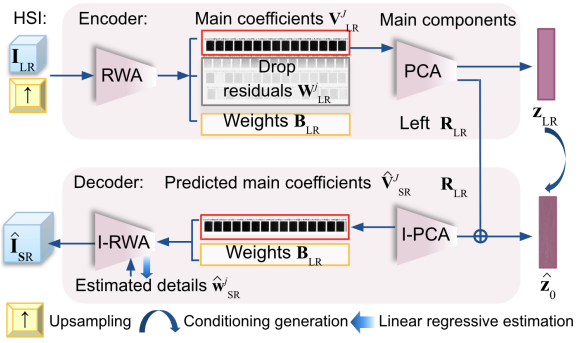

Figure 2: Illustration of the wavelet-based encoder-decoder.

coefficients into a dense and orthogonal matrix, facilitating more coherent spectral analysis:

$$(\mathbf{z}_{\text{LR}}, \mathbf{R}_{\text{LR}}) = \text{PCA}(\mathbf{V}^J_{\text{LR}}), \tag{5}$$

where $\mathbf{z}_{\text{LR}}$ is the feature that will be input to the diffusion latent space, and $\mathbf{R}_{\text{LR}}$ represents the remaining components that will be kept and reused in the decoder equation (6).

Decoder. For the reconstruction, the inverse PCA recovers predicted features $\hat{\mathbf{z}}_0$ from the diffusion process to the super-resolution HSI main coefficients $\hat{\mathbf{V}}^J_{\text{SR}}$.

$$(\hat{\mathbf{V}}^J_{\text{SR}}) = \text{I-PCA}(\hat{\mathbf{z}}_0, \mathbf{R}_{\text{LR}}). \tag{6}$$

The final super-resolved image is obtained by an inverse RWA, having set the residuals $\mathbf{W}^j_{\text{LR}}$ to zero, since these are not available for the super-resolution HSI image:

$$(\hat{\mathbf{I}}_{\text{SR}}) = \text{I-RWA}(\hat{\mathbf{V}}^J_{\text{SR}}, \mathbf{B}_{\text{LR}}, \mathbf{W}^j_{\text{LR}}, J). \tag{7}$$

The details $\hat{\mathbf{w}}^j_{\text{SR}}$ will be predicted by the adjusted linear regression model $\mathbf{B}_{\text{LR}}$ to recover the information lost in the wavelet transform. Once the diffusion outputs the super-resolution components $\hat{\mathbf{V}}^J_{\text{SR}}$, inverse-RWA reconstructs the predicted main coefficients to the spectral dimension.

Geometric enhanced diffusion process

Hyperspectral image generation usually requires larger reverse sampling time steps. To solve this problem, we used EDM (Elucidating Diffusion Models) (Karras et al. 2022) as our baseline model. EDM adds noise in one step in the training process. A probability flow ordinary differential equation (ODE) then continuously increases the noise level of the image when moving forward in time (Karras et al. 2022). Instead of using a discrete time step to add noise, we used a concrete number σ to represent the strength of the noise:

$$\sigma \sim \exp(\text{N}(P_{\text{mean}} = -1.2, P_{\text{std}} = 1.2)), \tag{8}$$

where P_{mean} represents the mean value, P_{std} is the standard deviation, and N is a Gaussian distribution.

The "time variable" t used in our model is a continuous variable, which has the advantage of mapping the noise scale to an approximately linear interval. In this way, the noise strength that will be added at the t moment can correspond to the size of t, and the relationship can be represented as:

$$t = -\log(\sigma_t). \tag{9}$$

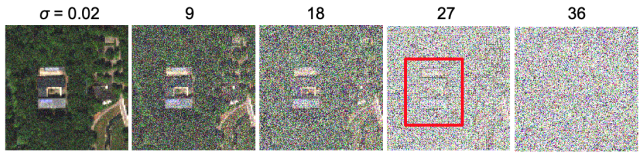

Figure 3: Edge perturbed noisy image over time.

Edge-aware noise scheduler. General diffusion models have an equal generation ability for each pixel. In the remote sensing scenario, we wanted to clarify the contour of buildings and other ground objects. Inspired by (Vandersanden et al. 2024), we designed an edge-aware noise scheduler in our training stage to increase the generation ability of the diffusion model for the pixels around the edges. The edge is preserved during the forward diffusion process. The noise around the edge is smaller than the general noise:

$$\mathbf{z}_{\mathrm{t}} = \mathbf{z}_0 + \sigma_t \epsilon \odot (1 - \mathbf{E}(1 - \sigma^2_{\mathrm{norm}})\eta), \tag{10}$$

where $\mathbf{z}_{\mathrm{t}}$ represents the noisy features at moment t, $t \in (0, T)$, $\mathbf{z}_0$ is the features from the ground truth, σ_t represents the noise strength at moment t, $\epsilon \sim \mathrm{N}(0, 1)$ is random noise, $\mathbf{E}$ is the binary edge map obtained from the input image, σ_{norm} is the normalized sigma at moment t, $\odot$ highlights that the operation is a matrix multiplication, and $\eta = 0.5$ adjusts the perturbation strength influenced by the edge.

Mask controllable training and sampling. In our preliminary research, generating geometric objects without distortion was a big challenge. To address this, a mask is introduced as a condition that improves the ability to generate buildings. Segmentation is calculated from low-resolution RGB channels from hyperspectral images with a segment-anything model (Kirillov et al. 2023). We used one minus the average of $\mathrm{NDVI}_{\mathrm{norm}}$,$\mathrm{NDVI}_{\mathrm{norm}} \in [0, 1]$ (Kriegler et al. 1969) as the value of the mask to highlight the attention of buildings. Let S_s be the pixel number of the s th segmentation region, then the value M_s of the mask is defined as:

$$M_s = 1 - \frac{1}{|S_s|} \sum_{(x,y) \in S_s} \mathrm{NDVI}_{\mathrm{norm}}(x, y). \tag{11}$$

In the training stage, the predicted result $\hat{\mathbf{z}}_0$ is calculated in one step:

$$\hat{\mathbf{z}}_0 = f_\theta \left(\mathbf{z}_t, \mathbf{C}, \sigma_t \right), \mathbf{C} = [\mathbf{z}_{\mathrm{LR}},\ \mathbf{M}], \tag{12}$$

where $\hat{\mathbf{z}}_0$ represents the predicted features when t = 0; f_θ represents the objective function 3D U-Net with spectral fidelity enhancer (SFE) (Dong et al. 2021), as shown in figure 1; $\mathbf{z}_{\mathrm{LR}}$ represents the low-resolution condition; $\mathbf{M} \in (0, 1)^{H \times W}$ is the mask condition; and $\mathbf{C}$ is the concatenation of all conditions.

During the sampling stage, DPM-Solver++ (Lu et al. 2022) accelerates the generation by employing a second-order approximation to solve the underlying ODE, while utilizing adaptive time stepping to significantly reduce the number of function evaluations. In our sampler, $t \in [0, T]$ will be separated into N steps. The step size is $\Delta t = t_{n+1} - t_n, n = 0, 1, ..., N - 1$. The initial noisy image and noise strength at step n can then be calculated with:

$$\mathbf{z}_T = \sigma_T \cdot \epsilon, \tag{13}$$

$$\sigma_n = \left(\sigma^{1/\rho}_{\mathrm{max}} + \frac{n}{N-1} \left(\sigma^{1/\rho}_{\mathrm{min}} - \sigma^{1/\rho}_{\mathrm{max}} \right) \right)^\rho, \tag{14}$$

where ρ is the scheduling curvature parameter, and $\sigma_{\mathrm{max/min}}$is the maximum/minimum noise strength. The n+1 step denoised features $\mathbf{z}_{n+1}$ can be calculated with:

$$\gamma = -\frac{1}{2} \cdot \frac{t_{n+1} - t_n}{t_n - t_{n-1}}, \tag{15}$$

$$\tilde{f}_\theta = (1 - \gamma) f_\theta(\hat{\mathbf{z}}_n, \mathbf{C}, \sigma_n) + \gamma f_\theta(\hat{\mathbf{z}}_{n-1}, \mathbf{C}, \sigma_{n-1}), \tag{16}$$

$$\mathbf{z}_{n+1} = \frac{\sigma_{n+1}}{\sigma_n} \hat{\mathbf{z}}_n - \sigma_{n+1}(e^{-\Delta t} - 1) \cdot \tilde{f}_\theta. \tag{17}$$

Multi-level loss function

Multi-level loss equations, such as equation 18, can ensure that the generated image is accurate in all aspects.

$$\mathcal{L} = \lambda(t) \cdot (\lambda_1 \mathcal{L}_{\mathrm{pixel}} + \lambda_2 \mathcal{L}_{\mathrm{perc}} + \lambda_3 \mathcal{L}_{\mathrm{grad}}), \tag{18}$$

To balance the convergence speed, we set $\lambda_1 = 0.8, \lambda_2 = 0.1, \lambda_3 = 0.1$. $\lambda(t)$ indicates the loss weighting based on t. Considering pixel loss can ensure that the absolute value of the spectral information of each pixel is accurate. Here we use the combination of L2 norm loss and Spectral Angle Mapper (SAM) (Yuhas, Goetz, and Boardman 1992) loss:

$$\mathcal{L}_{\mathrm{pixel}} = (\|\mathbf{z}_0 - \hat{\mathbf{z}}_0\|^2 + \mathrm{SAM}(\mathbf{z}_0, \hat{\mathbf{z}}_0))/2. \tag{19}$$

Perceptron loss (Johnson, Alahi, and Fei-Fei 2016) can ensure that the generated image is similar in high-level feature space:

$$\mathcal{L}_{\mathrm{perc}} = \|\phi \mathrm{VGG}(\hat{\mathbf{z}}_0) - \phi \mathrm{VGG}(\mathbf{z}_0)\|^2_2 . \tag{20}$$

Gradient loss (Lu and Chen 2022) can ensure that the image gradient information is consistent with the real image. The images generated by DPM Solver++ have high-contrast characteristics. Gradient loss is defined as follows:

$$\mathcal{L}\mathrm{grad} = \frac{1}{2}(\|\bigtriangledown_x \hat{\mathbf{z}}_0 - \bigtriangledown_x \mathbf{z}_0\|^1 + \|\bigtriangledown_y \hat{\mathbf{z}}_0 - \bigtriangledown_y \mathbf{z}_0\|^1), \tag{21}$$

where $\bigtriangledown_{x/y}$ represents the gradient of the image in the x/y direction.

Experiments

Datasets and implementation details

A group of EeteS simulated EnMap hyperspectral data, called the EnMap Campaign (realistic EnMAP-like high-resolution data), was used for training. EnMap is a hyperspectral satellite managed by the DLR Earth Observation Center, offering 30 m spatial resolution since June 2022. Additionally, the EnMAP Campaign Portal (access via supplementary material) captures aerial hyperspectral imagery and simulates EnMap-like data, boasting a spatial resolution of 2.5-4 m and containing data from 2009 to 2016. We gathered

	Metric	MCNet	MSDFormer	ESSAFormer	DMGASR	HIR Diff	SNLSR	Ours
(a)	PSNR↑	28.300±0.0480	28.284±0.0000	27.483±0.1392	26.986±0.2052	24.833±0.1079	28.531±0.0001	**28.863±0.2940**
	SSIM↑	0.6658±0.0025	0.6592±0.0000	0.5915±0.0191	0.5831±0.0118	0.6401±0.0024	0.6718±0.0000	**0.7104±0.0212**
	SAM↓	8.3332±0.1243	8.7442±0.0000	9.2114±0.2482	11.340±0.0743	8.9538±0.0053	**7.8911±0.0000**	8.4283±0.3073
	CC↑	0.7440±0.0000	0.7645±0.0000	0.7374±0.0004	0.6767±0.0128	0.7543±0.0011	0.7527±0.0000	**0.7945±0.0165**
	RMSE↓	0.0557±0.0002	0.0544±0.0000	0.0560±0.0002	0.0627±0.0006	0.0810±0.0015	0.0552±0.0000	**0.0548±0.0030**
	FID↓	116.14±0.0856	103.74±0.0000	97.438±14.779	49.026±2.3984	50.596±1.0436	125.75±0.0691	**44.464±17.627**
	LV↑	0.0004±0.0000	0.0004±0.0000	0.0004±0.0000	0.0037±0.0004	0.0021±0.0002	0.0003±0.0000	**0.0041±0.0022**
(b)	PSNR↑	24.216±0.0157	24.359±0.0000	24.103±0.0132	23.021±0.0146	21.567±0.5351	24.305±0.0000	**24.933±0.0079**
	SSIM↑	0.5355±0.0008	0.5536±0.0000	0.5210±0.0010	0.4925±0.0025	0.4987±0.0133	0.5404±0.0000	**0.6337±0.0106**
	SAM↓	11.663±0.0482	11.912±0.0000	12.166±0.0156	16.158±0.0117	12.348±0.1154	11.418±0.0001	**11.323±0.0456**
	CC↑	0.7050±0.0004	0.7238±0.0000	0.7077±0.0004	0.6575±0.0007	0.7347±0.0003	0.7102±0.0000	**0.7771±0.0003**
	RMSE↓	0.0685±0.0001	0.0669±0.0000	0.0682±0.0002	0.0779±0.0011	0.0940±0.0061	0.0680±0.0000	**0.0668±0.0020**
	FID↓	257.45±0.1949	272.78±0.0000	288.85±7.1990	120.06±11.769	375.96±23.877	267.88±0.0032	**64.333±4.2810**
	LV↑	0.0007±0.0000	0.0006±0.0000	0.0007±0.0000	0.0034±0.0005	0.0005±0.0000	0.0005±0.0000	**0.0087±0.0003**
(c)	PSNR↑	33.389±0.2641	28.709±0.0000	25.504±0.1340	32.864±0.2049	34.473±0.0069	35.734±0.0000	**35.837±0.1176**
	SSIM↑	0.7441±0.0029	0.4766±0.0000	0.4120±0.0312	0.6802±0.0159	0.7362±0.0017	0.7525±0.0000	**0.7747±0.0045**
	SAM↓	8.5500±0.0896	12.213±0.0000	18.724±0.5899	11.476±0.3843	8.3601±0.0446	7.6613±0.0000	**7.4735±0.0532**
	CC↑	0.6495±0.0145	0.6300±0.0000	0.6326±0.0090	0.5001±0.0161	0.7102±0.0001	0.7733±0.0000	**0.7906±0.0055**
	RMSE↓	0.0476±0.0006	0.0525±0.0000	0.0690±0.0006	0.0542±0.0013	**0.0420±0.0001**	0.0471±0.0000	0.0468±0.0006
	FID↓	464.13±5.9021	738.62±0.0000	701.35±16.290	245.38±63.176	363.23±6.9705	470.34±0.0000	**238.12±16.970**
	LV↑	0.0003±0.0000	0.0003±0.0000	0.0010±0.0000	0.0031±0.0015	0.0002±0.0000	0.0002±0.0000	**0.0011±0.0000**
(d)	Tr time (s)	1.33×10^4	3.99×10^4	7.65×10^4	3.16×10^5	–	2.80×10^4	3.10×10^5
	Te time (s)	18.13	10.40	7.98	334.00	212.90	4.10	28.70
	NFE	256^2	256^2	256^2	20×8	20	256^2	50
	Model size	6.50 MB	57.7 MB	3.70 MB	1.18 GB	1.56 GB	7.70 MB	4.55 GB

Table 1: Quantitative comparison with SOTA SR models of PSNR, SSIM, SAM, CC, RMSE, FID, and LV on (a) MDAS sample 1, (b) MDAS sample 2, and (c) WDC dataset. (d) Model efficiency was evaluated with the training/testing time, number of function evaluations (NFE), and model size. (Best performance value is highlighted in bold. Noise-affected values are underlined.)

Figure 4: 4-times visual comparisons with SOTA SR models on (a) MDAS sample 1, (b) MDAS sample 2, and (c) WDC dataset.

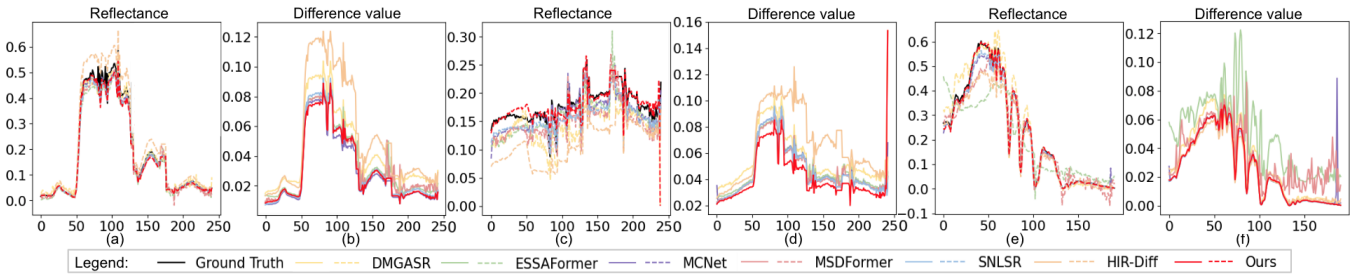

Figure 5: Reflectance of a random pixel and mean band-wise differences for (a–b) MDAS 1, (c–d) MDAS 2, and (e–f) WDC.

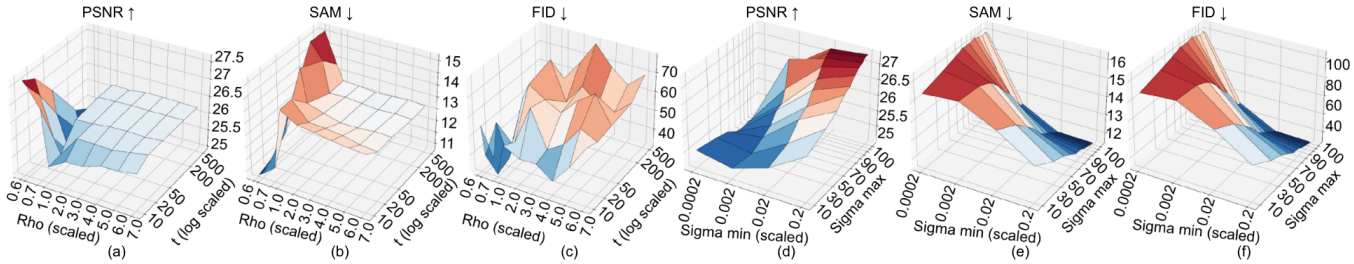

Figure 6: Qualitative comparison with different numbers of ρ and time steps for (a) PSNR, (b) SAM, and (c) FID. Qualitative comparison with different numbers of $\sigma_{\max}$ and $\sigma_{\min}$ for (d) PSNR, (e) SAM, and (f) FID on validation dataset.

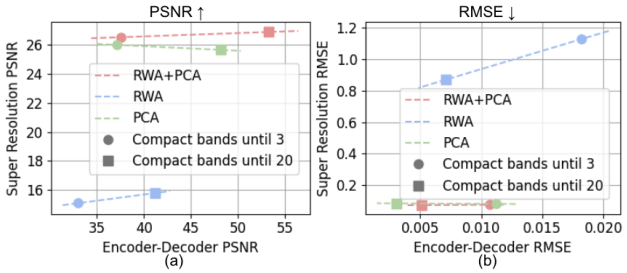

Figure 7: Encoder–decoder reconstruction quality compared with the performance of the whole SR process for the validation dataset.

8000 pairs of 256 × 256 × 242 patches of super-resolution HSI images and aligned them with 4-times downsampled images from EnMap Campaign and MDAS (Hu et al. 2023). This dataset covers 15 cities in Europe and the Americas, representing diverse ground objects. Among 8,000 pairs, one pair was split into the validation dataset and used for parameter selection, and a group of ablation study experiments. Two pairs are selected for EnMap simulation testing. The WDC (Biehl et al. 2015) dataset is used as one of the test datasets to see the transmission on other datasets. We deployed the model on four NVIDIA A100 GPUs with a learning rate of 1×10^{-4} and trained it for 200 epochs.

Quantitative metrics

Fidelity was evaluated using two standard metrics: the Peak Signal-to-Noise Ratio (PSNR) (Huynh-Thu and Ghanbari 2012), which measures the pixel-wise quality, and the Structural Similarity Index (SSIM) (Wang et al. 2004), which captures structural consistency. Both were averaged over spectral bands. **Realism and clarity** were addressed using the Fréchet Inception Distance (FID) (Heusel et al. 2017), noting that it is computed in an RGB-trained feature space and thus was only used for relative comparisons. We also included Local Variation (LV) (Pertuz, Puig, and Garcia 2013) to evaluate the local texture sharpness. **Spectral accuracy** was measured via Spectral Angle Mapper (SAM) (Yuhas, Goetz, and Boardman 1992), Cross-Correlation (CC), and Root Mean Square Error (RMSE). These metrics assess the angular, correlational, and pixel-wise spectral alignment.

State-of-the-art image generation

We evaluated the performance and efficiency of our model by comparing it with the six SOTA models: MCNet (Li, Wang, and Li 2020), MSDFormer (Chen, Zhang, and Zhang 2023), ESSAFormer (Zhang et al. 2023), DMGASR (Wang et al. 2024), HIR Diff (Pang et al. 2024a), and SNLSR (Hu et al. 2024). We trained these models on our dataset with the implementation details provided in each paper, except for the unsupervised method HIR Diff. The HIR Diff model provides a pre-trained checkpoint based on their fully prepared HSI dataset. Our model demonstrated strong performance in generating medium to large-scale ground objects. In Figure 4 and Table 1, we can see that both the visualization result and quantitative result outperformed the other models. The WDC dataset result shows our model can adapt to another dataset with a different spectral profile. However, our model may struggle with accuracy when the input conditional image lacks sufficient semantic information; for example, the rooftop of the MDAS sample 1 reconstructed image.

Effect of the encoder-decoder alone vs. with super-resolution. We compared our encoder-decoder (RWA +

Method	Baseline	A	B	C	D	E	F	G	H	I	J	Ours
w/RWA		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
w/PCA			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
w/Mask		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
w/Edge		✓	✓		✓	✓	✓	✓	✓	✓	Inverse	✓
w/L pix	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
w/L perc		✓	✓	✓	✓		✓		✓	✓	✓	✓
w/L geo		✓	✓	✓	✓			✓	✓	✓	✓	✓
w/Unet3D		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
w/SFE		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
PSNR↑	2.0476	15.788	25.640	26.579	26.681	26.101	21.664	26.467	26.388	26.181	26.8713	**27.013**
SSIM↑	-0.0150	-0.0173	0.5267	0.6503	0.6530	0.6140	0.1580	0.6443	0.6240	0.6131	0.6667	**0.6573**
SAM ↓	124.15	85.239	15.125	11.766	12.160	12.804	25.528	12.445	12.788	14.085	11.7688	**11.501**
CC↑	0.2643	0.4763	0.5530	0.6803	0.6882	0.6441	0.0145	0.6741	0.6681	0.6500	0.6999	**0.7008**
RMSE ↓	1.1879	0.8687	0.0850	0.0758	0.0749	0.0804	0.1342	0.0769	0.0783	0.0803	0.0739	**0.0726**
FID ↓	5019.1	484.16	83.627	43.445	36.269	40.303	701.94	40.195	47.475	65.985	34.9402	**30.110**
LV↑	7.5344	0.6231	0.0160	0.0079	0.0077	0.0096	0.1684	0.0093	0.0074	0.0089	0.0080	**0.0083**

Table 2: Ablation study. Quantitative comparison for the effect of each module on the validation dataset. Results are averaged over 4 runs. The baseline used an EDM backbone and DPM-Solver++ sampler on 242 bands. A: no PCA in the encoder; B: no RWA in the encoder; C: no edge perturbation; D: no mask conditioning; E/F/G: retained pixel/perceptual/geometric loss; H: no spectral fidelity module in the encoder; I: used 2D U-Net instead of 3D; J: more noise on edge (Zhang et al. 2025).

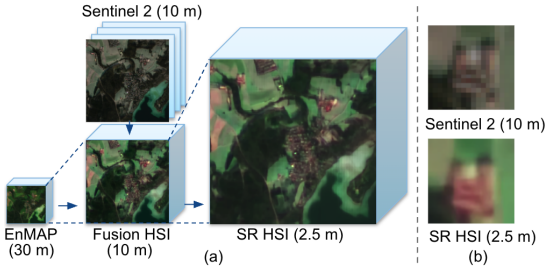

Figure 8: (a)Visualization of EnMap, Sentinel 2, Fusion image, Super resolution image. (b) Effect of super-resolution.

PCA) against RWA-only and PCA-only baselines (Figure 7), under 3- and 20-band compression. Without super-resolution, our encoder-decoder achieved almost lossless reconstruction with a PSNR up to 56. With super-resolution, using RWA + PCA compression to 20 bands outperformed all the baselines, demonstrating strong spectral compaction and generation. As shown, retaining more bands improves the quality but increases the cost. We chose 20 bands to ensure a balance between performance and efficiency.

Effect of ρ, number of timesteps N, $\sigma_{\max}$ and $\sigma_{\min}$ on sampling. The EDM noise schedule is shaped by ρ, which controls how noise levels decay from $\sigma_{\max}$ to $\sigma_{\min}$ via N steps. Higher ρ sharpens early denoising but increases the randomness in later steps, potentially degrading spectral consistency. We found that lower ρ values [0.6–0.7] yielded smoother noise schedules via 50 steps, which also better preserve spectral fidelity (Figure 6). We also studied the effects of $\sigma_{\max}$ and $\sigma_{\min}$. A larger $\sigma_{\max}$ allows more diversity but risks over-noising; a smaller one limits the detail. We set $\sigma_{\max} = 80$ for a balance. For $\sigma_{\min}$, smaller values extend the denoising phase, improving the detail but increasing artifacts. We found the best results when $\sigma_{\min} \in [0.02, 0.2]$.

Real-world application. We tried to combine EnMAP and Sentinel-2 imagery to 10 m resolution hyperspectral data via the unsupervised method HySure (Simões et al. 2015). Leveraging our 4-times super-resolution generation model, GEWDiff, we could finally produce EnMAP hyperspectral images with a 2.5-meter resolution. The no-reference image quality assessment MetaIQA (Zhu et al. 2020) was improved from 0.1997 to 0.2029 (Figure 8 (b)).

Ablation study

The results of our ablation studies, presented in Table 2, offer significant insights into the contributions of various components of the GEWDiff model. The use of a suitable encoder plays a crucial role in our model design. The multi-level loss function achieves better performance than an L2 loss. The design of a 3D objective function, U-Net, and its spectral fidelity enhancer makes progress toward stabilizing the results. The geometric enhancement strategies, such as edge perturbation and mask conditioning, did not show a significant improvement in the global metrics. However, some effects could be observed from Figure 4, whereby the edges are clear, and there was no obvious building distortion.

Conclusion

We proposed GEWDiff, which improves the spatial resolution of hyperspectral images by a factor of 4. Our method integrates wavelet-domain transforms and geometric priors to effectively preserve both spectral fidelity and spatial textures while accelerating convergence. The experimental results showed that GEWDiff outperformed the current SOTA baselines. One limitation of GEWDiff is that the result relies too much on the input conditions. This limitation could be addressed in future work through the integration of classifier-free guidance, which would enable the model to better generalize under weak or ambiguous conditioning. Moreover, we would also contribute to model distillation for further lightweight alternatives.

Acknowledgments

Please refer to the extended version via https://arxiv.org/html/2511.07103v1.

This work was supported in part by Munich Center for Machine Learning and in part by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C). The authors sincerely thank GFZ Helmholtz-Zentrum for providing the EnMAP Campaign datasets (Buddenbaum and Hill 2020; Beamish et al. 2020; Brell et al. 2020; Milewski et al. 2020; Cooper et al. 2020; Hank et al. 2015; Jarmer and Siegmann 2017; Neumann, Weiss, and Itzerott 2015; Okujeni, van der Linden, and Hostert 2016; Foerster et al. 2015; Boesche et al. 2016) used in this paper, and Dr. Jingliang Hu for providing the MDAS dataset.

References

Amrani, N.; Serra-Sagristà, J.; Laparra, V.; Marcellin, M. W.; and Malo, J. 2016. Regression Wavelet Analysis for Lossless Coding of Remote-Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9): 5616–5627.

Beamish, A.; Chabrillat, S.; Brell, M.; Heim, B.; and Sachs, T. 2020. Toolik Lake Research Natural Area AISA-Eagle hyperspectral Mosaic.

Biehl, L.; Maud, A. R.; Hsu, W.-K.; and Yeh, T. T. 2015. MultiSpec.

Boesche, N. K.; Mielke, C.; Segl, K.; Chabrillat, S.; Rogass, C.; Thomson, D.; Lundeen, S.; Brell, M.; and Guanter, L. 2016. EnGeoMAP Test Data: Simulated EnMAP Satellite Data for Mountain Pass, USA and Rodalquilar, Spain.

Brell, M.; Spengler, D.; Ruhtz, T.; Ward, K.; Chabrillat, S.; Segl, K.; Foerster, S.; and Itzerott, S. 2020. Demmin, Germany (October 2015) - an EnMAP Preparatory Flight Campaign.

Buddenbaum, H.; and Hill, J. 2020. Gerolstein, 2016-09-08 - An EnMAP Preparatory Flight Campaign.

Chen, N.; Yue, J.; Fang, L.; and Xia, S. 2023. SpectralDiff: A generative framework for hyperspectral image classification with diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Chen, S.; Zhang, L.; and Zhang, L. 2023. MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Cheng, Y.; Ma, Y.; Fan, F.; Ma, J.; Yao, Y.; and Mei, X. 2024. Latent spectral-spatial diffusion model for single hyperspectral super-resolution. *Geo-spatial Information Science*, 1–16.

Cooper, S.; Okujeni, A.; Jänicke, C.; Segl, K.; van der Linden, S.; and Hostert, P. 2020. 2013 Simulated EnMAP Mosaics for the San Francisco Bay Area, USA.

Dong, W.; Hou, S.; Xiao, S.; Qu, J.; Du, Q.; and Li, Y. 2021. Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7303–7317.

Foerster, S.; Brosinsky, A.; Wilczok, C.; and Bauer, M. 2015. Isábena 2011 - An EnMAP Preparatory Flight Campaign (Datasets).

Haar, A. 1910. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69: 331–371.

Hank, T. B.; Locherer, M.; Richter, K.; and Mauser, W. 2015. Neusling (Landau a.d. Isar) 2012 - A Multitemporal and Multisensoral Agricultural EnMAP Preparatory Flight Campaign (Datasets).

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hou, J.; Zhu, Z.; Hou, J.; Zeng, H.; Wu, J.; and Zhou, J. 2022. Deep posterior distribution-based embedding for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 31: 5720–5732.

Hu, J.; Liu, R.; Hong, D.; Camero, A.; Yao, J.; Schneider, M.; Kurz, F.; Segl, K.; and Zhu, X. X. 2023. MDAS: a new multimodal benchmark dataset for remote sensing. *Earth System Science Data*, 15(1): 113–131.

Hu, Q.; Wang, X.; Jiang, J.; Zhang, X.-P.; and Ma, J. 2024. Exploring the Spectral Prior for Hyperspectral Image Super-Resolution. *IEEE Transactions on Image Processing*, 33: 5260–5272.

Huynh-Thu, Q.; and Ghanbari, M. 2012. The accuracy of PSNR in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49(1): 13–27.

Jarmer, T.; and Siegmann, B. 2017. Köthen 2011/2012 - An EnMAP Preparatory Flight Campaign.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv:1603.08155.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arXiv:2304.02643.

Kriegler, F. J.; Malila, W. A.; Nalepka, R. F.; and Richardson, W. 1969. Preprocessing transformations and their effects on multispectral recognition. In *Proceedings of the Sixth International Symposium on Remote Sensing of Environment*, 97–131.

Li, J.; Cui, R.; Li, B.; Song, R.; Li, Y.; Dai, Y.; and Du, Q. 2020. Hyperspectral Image Super-Resolution by Band Attention Through Adversarial Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(6): 4304–4318.

Li, Q.; Wang, Q.; and Li, X. 2020. Mixed 2D/3D convolutional network for hyperspectral image super-resolution. *Remote sensing*, 12(10): 1660.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. ArXiv preprint, arXiv:2211.01095.

Lu, Z.; and Chen, Y. 2022. Single image super-resolution based on a modified U-net with mixed gradient loss. *signal, image and video processing*, 16(5): 1143–1151.

Milewski, R.; Chabrillat, S.; Brell, M.; Behling, R.; and Eichstaedt, H. 2020. Omongwa Pan, Namibia (June 2015) - an EnMAP Preparatory Flight Campaign.

Neumann, C.; Weiss, G.; and Itzerott, S. 2015. Döberitzer Heide 2008/2009 - An EnMAP Preparatory Flight Campaign (Datasets).

Okujeni, A.; van der Linden, S.; and Hostert, P. 2016. Berlin-Urban-Gradient dataset 2009 - An EnMAP Preparatory Flight Campaign (Datasets).

Pang, L.; Rui, X.; Cui, L.; Wang, H.; Meng, D.; and Cao, X. 2024a. HIR-Diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3005–3014.

Pang, L.; Tang, D.; Xu, S.; Meng, D.; and Cao, X. 2024b. HSIGene: A Foundation Model For Hyperspectral Image Generation. arXiv:2409.12470.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. arXiv:2212.09748.

Pertuz, S.; Puig, D.; and Garcia, M. A. 2013. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5): 1415–1432.

Phung, H.; Dao, Q.; and Tran, A. 2023. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10199–10208.

Qu, J.; Xiao, L.; Dong, W.; and Li, Y. 2024. MTLSC-Diff: Multitask learning with diffusion models for hyperspectral image super-resolution and classification. *Knowledge-Based Systems*, 303: 112415.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shi, Y.; Han, L.; Dancy, D.; and Han, L. 2024. WaveDiffUR: A diffusion SDE-based solver for ultra magnification super-resolution in remote sensing images. arXiv:2412.18996.

Shi, Y.; Han, L.; Han, L.; Chang, S.; Hu, T.; and Dancey, D. 2022. A Latent Encoder Coupled Generative Adversarial Network (LE-GAN) for Efficient Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Si, Y.; Lin, Z.; Wang, X.; and He, S. 2025. A New Hyperspectral Reconstruction Method With Conditional Diffusion Model for Snapshot Spectral Compressive Imaging. *IEEE Transactions on Instrumentation and Measurement*, 74: 1–14.

Sidorov, O.; and Yngve Hardeberg, J. 2019. Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Simões, M.; Bioucas-Dias, J.; Almeida, L. B.; and Chanussot, J. 2015. A Convex Formulation for Hyperspectral Image Superresolution via Subspace-Based Regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6): 3373–3388.

Su, X.; Shen, X.; Wan, M.; Nie, J.; Chen, L.; Liu, H.; and Zhou, X. 2025. EigenSR: Eigenimage-Bridged Pre-Trained RGB Learners for Single Hyperspectral Image Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7033–7041.

Vandersanden, J.; Holl, S.; Huang, X.; and Singh, G. 2024. Edge-preserving noise for diffusion models. ArXiv preprint, arXiv:2410.01540.

Wang, X.; Hu, Q.; Cheng, Y.; and Ma, J. 2023. Hyperspectral image super-resolution meets deep learning: A survey and perspective. *IEEE/CAA Journal of Automatica Sinica*, 10(8): 1668–1691.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wang, Z.; Li, D.; Zhang, M.; Luo, H.; and Gong, M. 2024. Enhancing hyperspectral images via diffusion model and group-autoencoder super-resolution network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5794–5804.

Wu, C.; Wang, D.; Bai, Y.; Mao, H.; Li, Y.; and Shen, Q. 2023. HSR-Diff: Hyperspectral image super-resolution via conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7083–7093.

Yu, D.; Li, Q.; Wang, X.; Zhang, Z.; Qian, Y.; and Xu, C. 2023. Dstrans: Dual-stream transformer for hyperspectral image restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3739–3749.

Yuhas, R. H.; Goetz, A. F. H.; and Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm. In *Summaries of the Third Annual JPL Airborne Geoscience Workshop*, volume 1, 147–149. Pasadena, CA: JPL.

Zhang, L.; You, W.; Shi, K.; and Gu, S. 2025. Uncertainty-guided Perturbation for Image Super-Resolution Diffusion Model. arXiv:2503.18512.

Zhang, M.; Zhang, C.; Zhang, Q.; Guo, J.; Gao, X.; and Zhang, J. 2023. Essaformer: Efficient transformer for hyperspectral image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23073–23084.

Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2024. Wavelet-based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8281–8291.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment. arXiv:2004.05508.