Enhancing Hyperspectral Images via Diffusion Model and Group-Autoencoder Super-resolution Network

Zhaoyang Wang[1 2*], **Dongyang Li**[2 3], **Mingyang Zhang**[1†], **Hao Luo**[2 3], **Maoguo Gong**[1],

[1]Ministry of Education, Key Laboratory of Collaborative Intelligence Systems, Xidian University
[2]DAMO Academy, Alibaba Group, 310023, Hangzhou, China
[3]Hupan Lab, 310023, Hangzhou, China
zhaoyangwang@stu.xidian.edu.cn, yingtian.ldy@alibaba-inc.com,
myzhang@xidian.edu.cn,michuan.lh@alibaba-inc.com,gong@ieee.org

Abstract

Existing hyperspectral image (HSI) super-resolution (SR) methods struggle to effectively capture the complex spectral-spatial relationships and low-level details, while diffusion models represent a promising generative model known for their exceptional performance in modeling complex relations and learning high and low-level visual features. The direct application of diffusion models to HSI SR is hampered by challenges such as difficulties in model convergence and protracted inference time. In this work, we introduce a novel Group-Autoencoder (GAE) framework that synergistically combines with the diffusion model to construct a highly effective HSI SR model (DMGASR). Our proposed GAE framework encodes high-dimensional HSI data into low-dimensional latent space where the diffusion model works, thereby alleviating the difficulty of training the diffusion model while maintaining band correlation and considerably reducing inference time. Experimental results on both natural and remote sensing hyperspectral datasets demonstrate that the proposed method is superior to other state-of-the-art methods both visually and metrically.

Introduction

Hyperspectral images (HSIs) offer plenty of information in the spectral dimension and have been found extensive applications in various fields such as remote sensing (Cloutis 1996), material recognition (Thai and Healey 2002), agriculture (Kersting et al. 2012), medical diagnosis (Fei 2020), and many others. However, the spatial resolution of HSI images is often relatively low due to the constraints of imaging hardware, which has a negative impact on subsequent HSI applications. Therefore, the HSI super-resolution (HSI SR) task is critical and meaningful to enhance the image quality to better serve the subsequent high-level computer vision tasks.

HSI SR can be categorized into two groups depending on whether auxiliary information is required: fusion-based SR (Hu, Huang, and Deng 2021), (Liu et al. 2020) and single-image SR (SISR). The fusion-based SR methods require additional high-resolution (HR) images as an aid, making them difficult to implement in practical application scenarios and making the SISR methods the research mainstream. The current state-of-the-art methods in SISR are predominantly CNN-based (Wang et al. 2022), (Liu et al. 2022c) and the majority of existing methods still suffer from modeling complex spectral-spatial relations and characterizing local features and global features comprehensively.

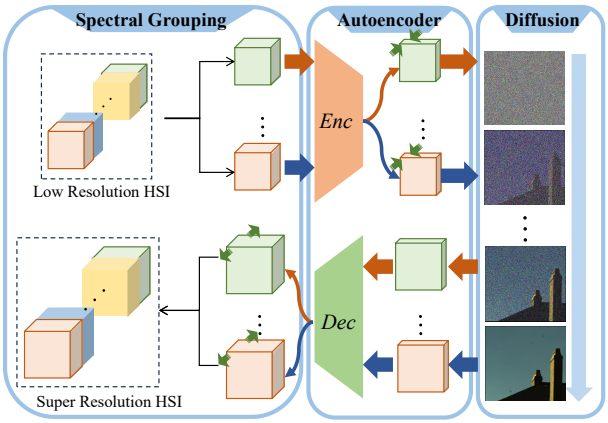

Figure 1: Our proposed framework combines three key techniques: spectral grouping and fusing techniques, autoencoder techniques and diffusion-based SR network.

Most recently, as a new emerging record-breaking model, diffusion model (Ho, Jain, and Abbeel 2020), (Nichol and Dhariwal 2021) shows superior performance in generation and reconstruction tasks and have started to be explored in the field of normal images SR (Saharia et al. 2023), with promising results being reported. In addition, the diffusion model has been applied to HSI classification (Zhou et al. 2023), demonstrating its effectiveness. The diffusion model shows great ability in acquiring and learning global information and higher-level detailed texture information (Zhou et al. 2023), which is ideally suited for handling complex spectral-spatial relationships and capturing global features and local features in dealing with HSI SR problems. The above work has inspired us to investigate the potential of the diffusion model for enhancing HSI SR.

Compared with natural images, HSI data have massive

*Work done when interning with Dongyang Li at Alibaba.
†Corresponding author.

and high dimensional characteristics, and the training samples are not as sufficient as natural images. Directly applying the diffusion model to the HSI SR task results in difficulties in model convergence, while simply performing a band-by-band SR application disrupts spectral continuity and ignores the band similarity, leading to unsatisfactory results as well (see Table 3). Additionally, the requirement for band-by-band SR necessitates multiple inferences, leading to significantly prolonged inference time (see Table 5).

To address these challenges, we propose a novel SR network for HSI data that integrates diffusion model, autoencoder techniques and spectral grouping techniques, as shown in Figure 1. We propose a novel autoencoder architecture that can encode a sheet of HSI data into several low-dimensional hidden variables for training the diffusion model. By adopting this approach, we address two crucial aspects. Firstly, it alleviates the challenges related to training and convergence in diffusion models when facing high spectral dimensionality HSI data and the "one-to-many" dimensionality reduction effectively reduces information loss during the encoding process, resulting in an enriched abundance of feature information (see Table 3). Secondly, due to the collaborative work of the autoencoders, our model efficiently narrows down the inference process to a few critical intermediate hidden variables, resulting in a substantial reduction in the inference time (see Table 5, Figure 9), making our approach more efficient and scalable for practical applications in HSI SR tasks. Our model comprises two primary training stages and following the completion of these training stages, the two modules collaborate harmoniously to execute the SR task effectively.

In summary, the main contributions of our work are as follows:

- To the best of our knowledge, our work represents the first application of the diffusion model to the field of HSI SR. We propose a novel diffusion-based SR model that facilitates the implicit capture of both high and low-level features and improves the learning ability of complex spectral-spatial relationships.
- We fuse diffusion models with autoencoder techniques to overcome the difficulty of convergence and significantly decrease the inference time in the face of high-dimensional data.
- Extensive experimental results on three publicly available HSI datasets show that our proposed method outperforms state-of-the-art methods in terms of both objective metrics and subjective visual quality.

Related Works

Single Hyperspectral Image Super-Resolution

In recent years, deep convolutional networks have shown impressive capabilities in recovering missing features in HSI data. For instance, in (Li et al. 2018), the authors proposed a recursive residual network (GDRRN) that utilizes all data channels as inputs and integrates Spectral Angle Mapper (SAM) into the loss function, which was a groundbreaking advancement. Other researchers, such as (Li, Wang, and Li 2020), presented the Mixed Convolutional Network (MCNet), which employs both 2-D and 3-D convolutions to reduce the computational burden of processing all bands simultaneously. Additionally, (Li, Wang, and Li 2021) explored the relationships between 2-D and 3-D convolutions. In (Jiang et al. 2020), the Spatial-Spectral Prior Network (SSPSR) was introduced, featuring a group convolution and progressive upsampling framework built upon the spectral grouping strategy. Building on this approach, (Wang et al. 2022) proposed the Group-based Embedding Learning and Integration Network (GELIN), which effectively utilizes information from neighboring spectral bands. Furthermore, many other research efforts have been based on the spectral grouping strategy, as demonstrated in (Liu, Fan, and Zhang 2022), (Liu and Dong 2022), (Liu et al. 2022a) and (Wang, Ma, and Jiang 2022). Additionally, some researchers have utilized Transformer-based architectures to learn the complex relationships between spectral and spatial information, as seen in (Gao et al. 2021), (Hu et al. 2022) and (Liu et al. 2022d). These works collectively contribute to the advancement of HSI data analysis and SR tasks.

Diffusion Based Super-Resolution Model

In recent years, the diffusion model has demonstrated its remarkable generative capabilities in various domains, including natural image SR tasks (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Li et al. 2022; Saharia et al. 2022). Notably, the SR3 model (Saharia et al. 2022), based on the diffusion model, has achieved high-performance results in super-resolving natural images. Stable Diffusion (LDM) (Rombach et al. 2022) is another top-performing diffusion method that exhibits exceptional performance in SR tasks. Moreover, diffusion models have been successfully applied to continuous SR of natural images (Gao et al. 2023). Beyond natural images, diffusion models have been found applications in various domains, such as magnetic resonance image SR (Mao et al. 2023; Chung, Lee, and Ye 2022), and have even been introduced to remote sensing imagery SR (Liu et al. 2022b). The broad applicability and impressive performance of diffusion models showcase their potential for various SR tasks across different domains.

Methodology

Our model is a two-stage training model that consists of two main parts: the autoencoder and the diffusion SR model, as shown in Figure 2. Our proposed model efficiently operates SR on the latent space. Specifically, an autoencoder consisting of an encoder $Enc(\cdot)$ and decoder $Dec(\cdot)$ is trained with a reconstruction objective. Given an HR HSI input HSI_{HR}, the encoder $Enc(\cdot)$ maps one single image to several hidden latent variables $[Z^1_{HR}, Z^2_{HR} \cdots Z^n_{HR}]$, and the decoder $Dec(\cdot)$ reconstructs the image from the hidden latent list. In this way, at each timestep t, a noisy latent list can be obtained $[Z^1_{HR,t}, Z^2_{HR,t} \cdots Z^n_{HR,t}]$. Beyond the routine training scheme, we also devise conditioning mechanisms to control the SR image by concatenating the LR hidden latent list $[Z^1_{LR}, Z^2_{LR} \cdots Z^n_{LR}]$. In inference, a list-shaped latent is sampled from standard normal distribution $\mathcal{N}(0, 1)$ and the

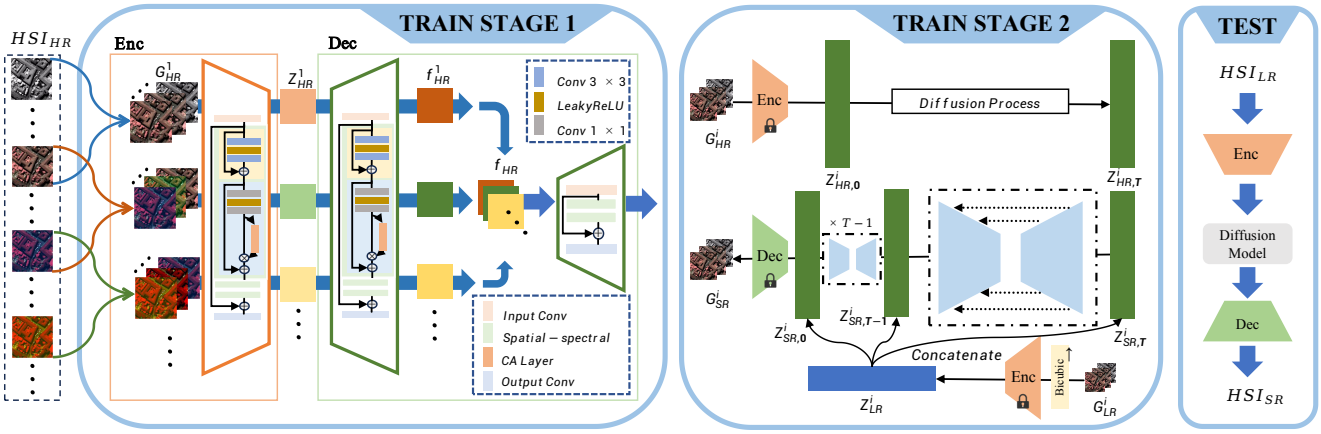

Figure 2: Overview of the proposed model, In Stage 1, the autoencoder is trained to encode the input data into a series of hidden variables ($[Z^1_{HR}, Z^2_{HR} \cdots Z^n_{HR}]$). In Stgae 2, the diffusion model is trained. The grouped data (G^i_{HR} and G^i_{LR}) are first encoded, generating hidden variables (Z^i_{HR} and Z^i_{LR}) and the z^i_{LR} is added as conditional information by directly concatenating it with the hidden variables ($Z^i_{SR,t}$) at each moment during the denoising process.

denoising model U-Net is used to iteratively remove noise to produce SR latent list. In the end, the SR latent list is passed to the decoder $Dec(\cdot)$ to generate the SR image HSI_{SR}.

In the following section, we will introduce the architectural components and principles of the two training stages as well as the detailed testing process.

Stage 1: Training the Group Autoencoder

Overall Network Architecture. Our autoencoder model, illustrated in Figure 2, incorporates a spectral grouping strategy and an asymmetric decoder design. Initially, the input data is grouped into bands, and each group is encoded to form a latent variable list. The decoding process is divided into two parts. In the first part, each sub-hidden variable is decoded separately. In the second part, the decoded sub-hidden variables are concatenated and decoded as a whole, resulting in the final decoded image. This two-stage decoding approach allows for effective feature extraction and reconstruction, enhancing the performance of our autoencoder model.

Spectral Grouping. Compressing all the spectral information directly into low-dimensional space can lead to significant information loss due to the high dimensionality of HSI data (Jiang et al. 2020). Finding the right compression scale is crucial, as a scale too large may result in excessive information loss, while a scale too small may cause convergence issues during training of the diffusion model. To strike a balance between information loss and effective coordination with the diffusion model, we introduced the spectral grouping strategy. This strategy involves dividing adjacent bands into groups with certain overlaps between them, allowing the encoder to learn the correlation between bands. In this way, we achieve efficient dimensionality reduction of the spectral dimension, reduces information loss, and enhances the richness of the hidden variables' features. Additionally, it considers the similarity between bands, which facilitates a more effective collaboration with the diffusion model. This approach enables our model to effectively handle the challenges of HSI SR tasks.

Asymmetric Architecture. Due to the list-shaped latent variable, we designed an asymmetric decoder with a larger model compared to the encoder to better capture the features. The decoder comprises two primary components: the local decoding part and the global decoding part. In the local decoding part, the variables in each list are initially decoded to decipher the local information features. These decoded variables are then concatenated to match the size of the actual HSI data. Subsequently, the concatenated data is passed through the global decoding part, which decodes the smooth details for the connected parts and enhances the overall effect. This two-stage decoding process allows our model to effectively reconstruct the HSI data and improve feature representation.

Loss Function. Our loss function consists of four main components: the L1 loss, the spectral angle mapper (SAM) loss, the gradient loss and the perceptual loss (Johnson, Alahi, and Fei-Fei 2016). The L_1 loss can be formulated as

$$L_1 = \frac{1}{N}\sum_{n=1}^{N} ||I^n_{Re} - I^n_{HR}|| \tag{1}$$

where I^n_{Re} and I^n_{HR} are the nth reconstructed HSI and original HSI_{HR} and N is the number of images in one training batch. In addition, to maintain the consistency of the spectral information while performing spatial SR, we introduce SAM as part of the loss function, which can be formulated as

$$L_{SAM} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{N_p}\sum_{i=1}^{N_p}\frac{1}{\pi} arccos\left(\frac{I^{n,i}_{Re}\cdot I^{n,i}_{HR}}{||I^{n,i}_{Re}||_2 \cdot ||I^{n,i}_{HR}||_2}\right) \tag{2}$$

where N_p is equal to $H \times W$ and $I^{n,i}_{HR}$ refers to the ith spectral vector of the nth image. Furthermore, inspired by (Wang et al. 2022), we add a gradient loss to preserve the sharpness

of the reconstructed images in both spatial and spectral domains, which is shown below:

$$L_g = \frac{1}{N}\sum_{n=1}^{N} ||M(I_{HR}^n) - M(I_{Re}^n)|| \tag{3}$$

where M computes the gradient value along the horizontal, vertical, and spectral dimensions of the image. Finally, we add perceptual loss to better take into account the perception of the human visual system, to better preserve the high-level features of the images and to better avoid over-smoothing. The perceptual loss can be formulated as

$$L_p = \frac{1}{N}\sum_{n=1}^{N} ||VGG(I_{HR}^n) - VGG(I_{Re}^n)|| \tag{4}$$

where VGG represents the pre-trained VGG19 model (Simonyan and Zisserman 2014).

In summary, the total loss for our proposed GAE model can be formulated as follows:

$$\text{Loss} = L_1 + \lambda_1 L_{\text{SAM}} + \lambda_2 L_g + \lambda_3 L_p \tag{5}$$

In our experiments, we set the weights as $\lambda_1 = 0.3$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.001$ to balance the contributions of the different loss components. This formulation allows our model to effectively optimize and achieve superior results in HSI SR tasks.

Stage 2: Training the Diffusion Model

In our approach, the complete training process of stage 2 is visually represented in Figure 2. The GAE plays a key role by simultaneously encoding both HSI_{HR} and HSI_{LR} images, generating separate latent variable lists for each type ($[z_{LR}^1, z_{LR}^2 \cdots z_{LR}^n]$, $[z_{HR}^1, z_{HR}^2 \cdots z_{HR}^n]$). This unique capability allows us to create novel training data pairs by combining corresponding data points from each variable list (z_{LR}^i, z_{HR}^i). Subsequently, these paired data points are used to train the diffusion SR model, enabling it to learn and generate high-quality SR results.

For our diffusion SR model, we adopt the architecture from the SR3 framework (Saharia et al. 2022), which has demonstrated excellent performance in natural image SR. By combining the strengths of both the autoencoder and the diffusion model, our proposed approach achieves significant improvements in HSI SR tasks, both qualitatively and quantitatively. To address the SR task, we incorporate the LR image as conditional information by directly concatenating it with the HR latent image. We choose this straightforward approach over more complex methods like cross-attention mechanisms (Rombach et al. 2022) because our conditional information has already been encoded and compressed once. Directly concatenating all the information allows the network to learn more effectively, building upon the demonstrated superiority in the SR3 framework. This effective utilization of conditional information enhances the performance of our model in handling HSI SR tasks.

Algorithm 1: Testing process

Input: encoder: $Enc(\cdot)$, decoder: $Dec(\cdot)$, noise prediction model: $\epsilon_\theta(\cdot)$, LR images: HSI_{LR}, time step: t
Output: SR images: HSI_{SR}

1: Encode LR images:

$$[z_{LR}^1, z_{LR}^2 \cdots z_{LR}^n] = Enc(HSI_{LR})$$

2: Perform SR on each element in the $[z_{LR}^1, z_{LR}^2 ... z_{LR}^n]$:
for $i = 0, 1 \cdots n$ **do**
$z_T^i \sim \mathcal{N}(0, I)$
for $t = T \cdots 1$ **do**
$\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$
$z_{t-1}^i = \frac{1}{\sqrt{\alpha_t}}(z_t^i - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t^i, t, z_{LR}^i)) + \sqrt{1-\alpha_t}\epsilon$
$z_{SR}^i = z_0^i$

3: Decode the obtained list $List_{z_{SR}}$:

$$HSI_{SR} = Dec([z_{SR}^1, z_{SR}^2 \cdots z_{SR}^n])$$

4: **Return** HSI_{SR}

Testing Process

The complete process of using the trained model for SR tasks is illustrated in Algorithm 1. First, we encode HSI_{LR} to obtain a list of latent variables ($[z_{LR}^1, z_{LR}^2 ... z_{LR}^n]$). Next, the diffusion model performs SR on each element inside the list, resulting in a new list of SR latent variables ($[z_{SR}^1, z_{SR}^2 ... z_{SR}^n]$). Finally, we decode the SR latent variables to obtain the final HSI_{SR} image. This sequential process enables our model to effectively enhance the spatial resolution of HSI data and achieve high-quality SR results.

Experiments

Dataset

In our experiments, we used three publicly available datasets to validate the performance of our model. These datasets include two remote-sensing HSI datasets: Pavia Center (PaviaC) dataset and Chikusei dataset(Yokoya and Iwasaki 2016), and one natural image HSI dataset: Harvard dataset (Chakrabarti and Zickler 2011). These datasets were selected to cover different scenarios and challenges for HSI SR tasks.

Evaluation Measures

In our comprehensive experiments, we employed six widely-used evaluation indices: Peak Signal-to-Noise Ratio (PSNR), Spectral Angle Mapper (SAM), Structural Similarity (SSIM), Cross Correlation (CC), Relative Dimensionless Global Error in Synthesis (ERGAS), and Root-Mean-Squared Error (RMSE). For PSNR and SSIM, we calculated their mean values over all spectral bands. The best values for these indices are $+\infty$, 0, 1, 1, 0, and 0, respectively.

Implementation Details

We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for training, with a batch size of 8 for the Harvard dataset and 4 for the PaviaC and Chikusei datasets. The learning

PaviaC Dataset & Scale = 2							PaviaC Dataset & Scale = 3					
Models	MPSNR ↑	MSSIM ↑	CC ↑	RMSE ↓	SAM ↓	ERGAS ↓	MPSNR ↑	MSSIM ↑	CC ↑	RMSE ↓	SAM ↓	ERGAS ↓
Bicubic	30.998	0.899	0.940	0.0292	4.675	4.567	27.865	0.785	0.877	0.0424	5.855	6.472
EDSR	31.342	0.904	0.940	0.0276	6.746	4.548	28.748	0.826	0.898	0.0376	7.714	5.945
GDRRN	31.559	0.905	0.947	0.0271	5.186	4.324	29.172	0.835	0.908	0.0360	6.531	5.615
SSPSR	32.335	0.922	0.953	0.0247	5.346	4.015	29.507	0.850	0.916	0.0346	5.987	5.416
MCNet	32.068	0.921	0.953	0.0262	5.219	4.174	28.193	0.802	0.889	0.0415	7.171	6.408
CEGATSR	31.746	0.909	0.939	0.0269	5.741	4.337	28.489	0.807	0.894	0.0392	6.816	6.051
GELIN	33.326	0.937	0.963	0.0222	4.099	3.553	29.611	0.850	0.918	0.0347	**5.294**	5.334
Ours	**34.491**	**0.950**	**0.971**	**0.0195**	**4.080**	**3.140**	**30.035**	**0.867**	**0.925**	**0.0328**	5.715	**5.091**
Chikusei Dataset & Scale = 2							Chikusei Dataset & Scale = 3					
Bicubic	35.008	0.932	0.965	0.0229	1.718	3.995	31.460	0.847	0.921	0.0345	2.547	5.935
EDSR	35.489	0.941	0.961	0.0198	2.444	4.525	31.962	0.868	0.925	0.0305	3.356	6.244
GDRRN	35.958	0.939	0.971	0.0206	1.561	3.606	32.383	0.866	0.935	0.0305	2.398	5.402
SSPSR	35.723	0.944	0.965	0.0197	2.275	4.187	33.015	0.890	0.942	0.0280	2.558	5.175
MCNet	36.371	0.948	0.971	0.0198	1.784	3.650	32.380	0.872	0.934	0.0309	2.496	5.581
CEGATSR	35.866	0.938	0.957	0.0204	2.212	3.994	31.685	0.856	0.923	0.0325	3.010	5.981
GELIN	37.747	0.959	0.979	0.0170	**1.384**	3.011	33.796	0.900	0.952	0.0267	**2.022**	4.539
Ours	**38.748**	**0.966**	**0.982**	**0.0161**	1.638	**2.738**	**34.192**	**0.909**	**0.954**	**0.0264**	2.637	**4.364**
Harvard Dataset & Scale = 2							Harvard Dataset & Scale = 3					
Bicubic	44.813	0.972	0.974	0.00840	2.623	3.183	41.717	0.946	0.951	0.01230	3.081	4.446
EDSR	44.945	0.976	0.972	0.00714	3.588	3.376	43.102	0.957	0.957	0.00994	3.687	4.024
GDRRN	45.868	0.975	0.976	0.00748	2.556	2.927	43.129	0.952	0.959	0.01060	**2.970**	3.949
SSPSR	44.939	0.977	0.975	0.00737	3.489	3.433	43.409	0.959	0.960	0.00974	3.478	3.868
MCNet	46.367	0.973	0.939	0.00713	3.300	3.738	42.745	0.952	0.953	0.01060	3.475	4.463
GELIN	**47.024**	**0.981**	0.981	**0.00623**	**2.530**	2.576	43.653	0.957	0.961	0.00992	3.000	3.765
Ours	46.953	0.979	**0.982**	0.00678	3.079	**2.527**	**44.028**	**0.959**	**0.966**	**0.00965**	3.596	**3.525**

Table 1: Quantitative results on the PaviaC dataset, Chikusei dataset and Harvard dataset at different scales. Bold represents the best result and underline represents the second best.

rate was set to $1e^{-4}$ during GAE training and reduced to $1e^{-5}$ for the diffusion model. During the training process, we utilized a pre-trained SR3 diffusion model. In the GAE module, bands were divided into subgroups of size 16 for PaviaC and Chikusei datasets, and 8 for the Harvard dataset, with one-quarter overlap between subgroups.

Results and Comparison with SOTA

To validate the effectiveness of our model, we conducted a comprehensive comparison with various state-of-the-art SISR methods. We used standard bicubic interpolation as a baseline and evaluated our results against six SOTA SISR methods: EDSR (Lim et al. 2017), GDRRN (Li et al. 2018), SSPSR (Jiang et al. 2020), MCNet (Li, Wang, and Li 2020), CEGATSR (Liu and Dong 2022), and GELIN (Wang et al. 2022). This comparison was conducted on three different datasets at three different scales. However, due to network design and CUDA memory limitations, the CEGATSR network could not be compared on the Harvard dataset. Nonetheless, we made every effort to reproduce the performance of each compared network to ensure a fair evaluation.

Quantitative Experimental Results. It can be noticed that our model achieves the best performance on the PaviaC dataset, outperforming other methods in various evaluation metrics. We further validated its effectiveness on the Chikusei and Harvard datasets, where it consistently demonstrated excellent results in most metrics, as shown in Table 1 and Table 2. Although the SAM index improvement was less significant, which may be attributed to the segmented data processing caused by the two-stage training approach, our model still showed superiority in other evaluation indicators. Overall, it achieved an approximate improvement of 0.5-0.8 dB in PSNR compared to other methods, confirming its effectiveness in HSI SR tasks.

Qualitative Experimental Results. Our model exhibits remarkable superiority in SR results, as demonstrated in Figure 3 for the PaviaC dataset, Figure 4 for the Chikusei dataset, and Figure 5 for the Harvard dataset. The visual comparisons and error maps vividly showcase our model's ability to preserve and enhance texture details, resulting in clearer and more accurate reconstructions compared to other methods. The outstanding visual performance aligns with the superior metric results, solidifying the effectiveness of our approach for HSI SR tasks.

Spectral Distortion Comparison. Finally, we also compared the spectral distortion among the different methods. As shown in Figure 6, Figure 7 and Figure 8, the HSIs reconstructed by our proposed method exhibit the highest spectral fidelity, with minimal spectral distortion. The ability to maintain high spectral fidelity is crucial for HSI tasks, and our model excels in this aspect, making it a robust and reliable solution for HSI SR problems.

Models	MPSNR ↑	MSSIM ↑	CC ↑	RMSE ↓	SAM ↓	ERGAS ↓
PaviaC Dataset & Scale = 4						
Bicubic	26.270	0.678	0.820	0.0514	6.623	7.762
EDSR	27.098	0.741	0.851	0.0458	8.236	7.130
GDRRN	27.445	0.749	0.864	0.0443	6.557	6.822
SSPSR	27.768	0.771	0.876	0.0428	6.320	6.562
MCNet	27.854	0.770	0.877	0.0424	6.302	6.504
CEGATSR	27.278	0.730	0.860	0.0454	6.425	6.934
GELIN	27.592	0.770	0.870	0.0439	**6.265**	6.679
Ours	**27.928**	**0.785**	**0.880**	**0.0423**	7.406	**6.428**
Chikusei Dataset & Scale = 4						
Bicubic	29.676	0.770	0.882	0.0425	3.161	7.275
EDSR	29.976	0.799	0.893	0.0386	4.127	7.547
GDRRN	30.658	0.801	0.905	0.0374	2.913	6.551
SSPSR	30.858	0.823	0.914	0.0355	3.196	6.651
MCNet	31.189	0.821	0.916	0.0354	2.955	8.284
CEGATSR	30.569	0.806	0.908	0.0374	3.082	6.757
GELIN	31.095	0.838	0.914	0.0366	**2.834**	6.102
Ours	**32.248**	**0.860**	**0.929**	**0.0332**	3.507	**5.378**
Harvard Dataset & Scale = 4						
Bicubic	39.940	0.926	0.934	0.0152	3.345	5.354
EDSR	41.330	0.940	0.945	0.0119	4.039	4.748
GDRRN	40.369	0.933	0.922	0.0127	4.120	5.866
SSPSR	41.929	0.941	0.952	0.0115	3.513	4.410
MCNet	41.986	0.939	0.947	0.0120	3.333	4.482
GELIN	42.673	0.945	0.959	0.0110	**3.156**	4.032
Ours	**43.132**	**0.948**	**0.961**	**0.0109**	3.534	**3.883**

Table 2: Quantitative results on the PaviaC dataset, Chikusei dataset and Harvard dataset at different scales. Bold represents the best result and underline represents the second best.

Models	MPSNR ↑	MSSIM ↑	CC ↑	RMSE ↓	SAM ↓	ERGAS ↓
$Scale = 2$						
Ours	34.491	0.950	0.971	0.0195	4.080	3.140
Ours - w/o GD	33.761	0.946	0.966	0.0214	4.714	3.410
Ours - w/o GS	32.442	0.929	0.954	0.0246	5.626	4.021
Diff -w PB	31.319	0.827	0.878	0.0277	6.715	5.122
Diff -w FB	10.948	0.043	0.084	0.3501	51.461	56.450
$Scale = 3$						
Ours	30.035	0.867	0.925	0.0328	5.715	5.091
Ours - w/o GD	29.685	0.862	0.919	0.0341	6.442	5.296
Ours - w/o GS	29.469	0.854	0.915	0.0347	6.073	5.467
Diff -w PB	29.172	0.756	0.833	0.0353	7.383	6.202
Diff -w FB	10.920	0.038	0.065	0.3483	53.411	54.513
$Scale = 4$						
Ours	27.928	0.785	0.880	0.0423	7.406	6.428
Ours - w/o GD	27.624	0.778	0.874	0.0436	8.227	6.678
Ours - w/o GS	27.432	0.759	0.867	0.0433	7.193	6.822
Diff -w PB	27.262	0.757	0.867	0.0451	10.492	6.969
Diff -w FB	10.906	0.043	0.088	0.3521	53.061	57.968

Table 3: Ablation results of quantitative performance on the PaviaC dataset at scale 2,3,4. GD: global decoding, GS: spectral grouping strategy, PB: without GAE and trained with partial bands step by step, FB: without GAE and trained with full-bands.

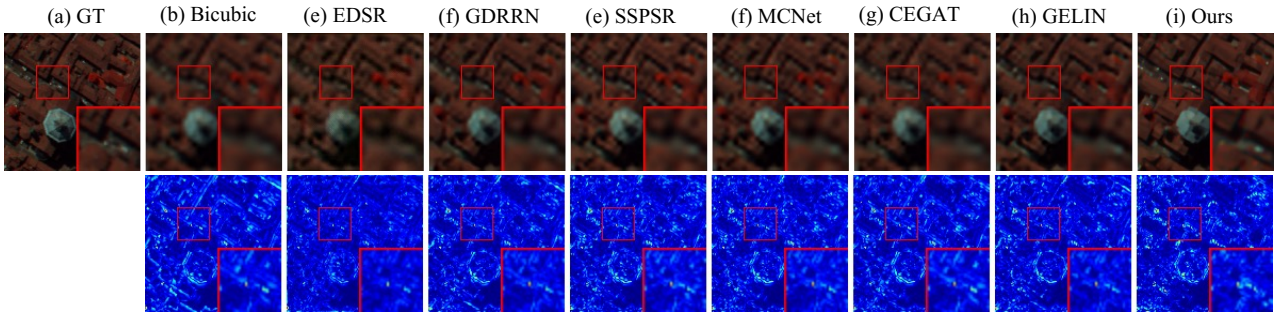

Figure 3: Qualitative results of different models at scale 4 with the corresponding error maps of the PaviaC dataset. The false-color image is used for clear visualization (red: 100, green: 30, and blue: 10).

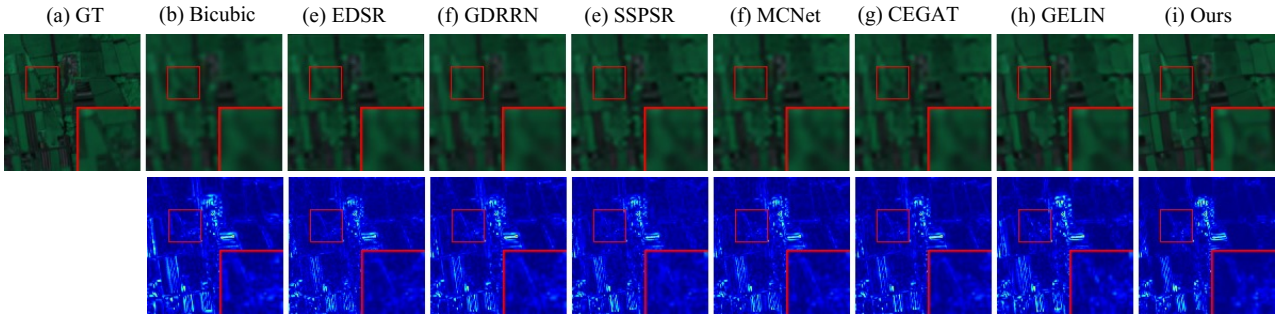

Figure 4: Qualitative results of different models at scale 4 with the corresponding error maps of the Chikusei dataset. The false-color image is used for clear visualization (red: 70, green: 100, and blue: 36).

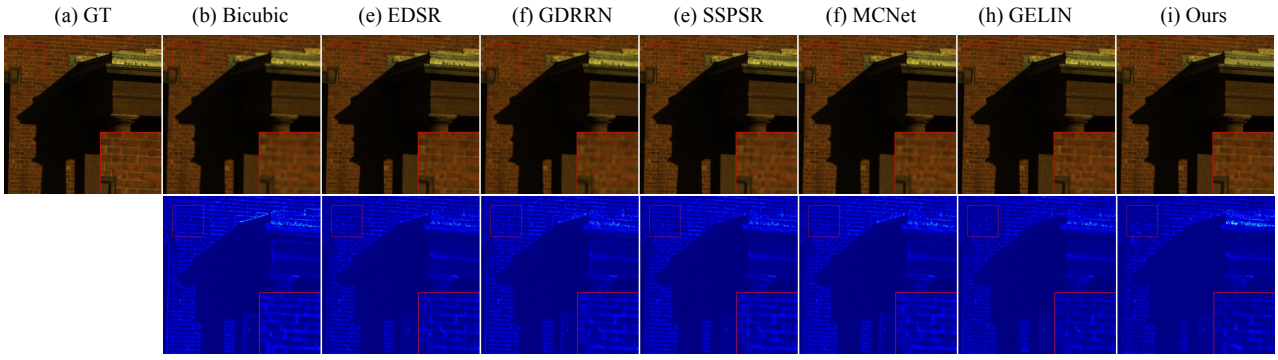

Figure 5: Qualitative results of different models at scale 4 with the corresponding error maps of the Harvard dataset. The false-color image is used for clear visualization (red: 25, green: 15, and blue: 2).

Ablation Study

We conducted ablation experiments to assess the effectiveness of each component in our model. Table 3 presents the results, where we evaluated the impact of the asymmetric autoencoder, the spectral grouping strategy, and the overall performance of our autoencoder. These experiments allowed us to verify the significance of each component in contributing to the overall performance of our model.

Asymmetric Structure. The design of the asymmetric autoencoder was essential to accommodate the list-shaped latent variable in our model. As shown in Table 3, removing the global decoding part led to a significant decrease in results, highlighting the crucial role of this structure in our model's effectiveness.

Spectral Grouping. The spectral dimension grouping strategy effectively reduces unnecessary computational costs by grouping similar spectral bands together, allowing the model to more efficiently utilize the spectral information of the image leading to effective coordination between the autoencoder and the diffusion model. As evident in Table 3, the results without the spectral grouping strategy were even worse, further underscoring its significance in achieving successful and improved results in our model.

Autoencoder. Removing the autoencoder module and directly using the diffusion SR model resulted in poor perfor-

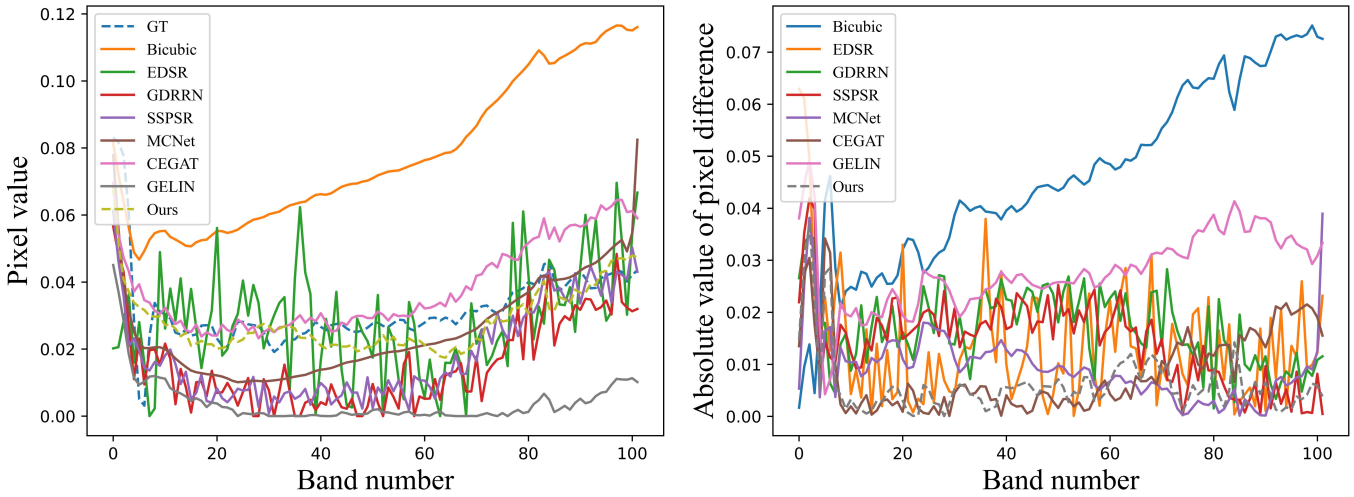

Figure 6: Example spectral curves and difference curves on a selected pixel value of PaviaC datasets with a scale factor of 4.

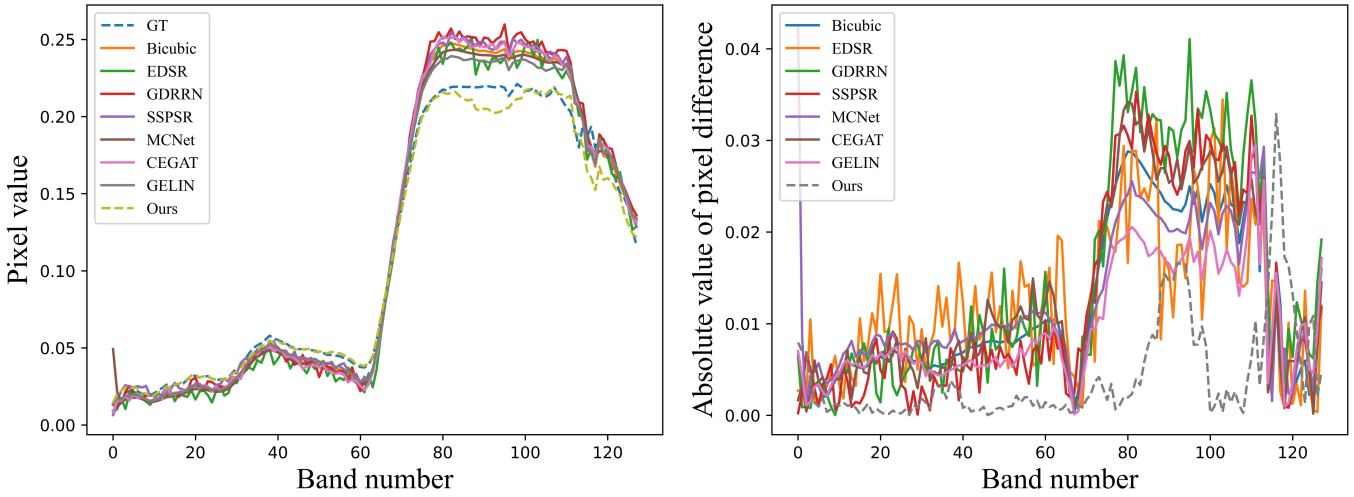

Figure 7: Example spectral curves and difference curves on a selected pixel value of Chikusei datasets with a scale factor of 4.

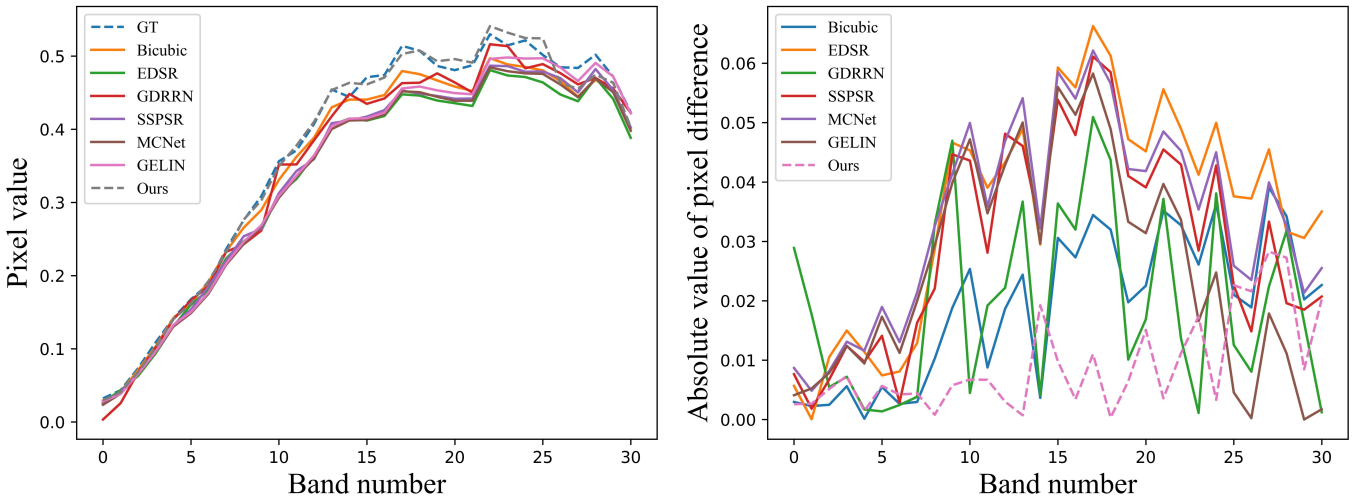

Figure 8: Example spectral curves and difference curves on a selected pixel value of Harvard datasets with a scale factor of 4.

Settings	MPSNR ↑	MSSIM ↑	ERGAS ↓
$n_{subs} = 12, n_{ovls} = 4$	27.784	0.784	6.537
$n_{subs} = 16, n_{ovls} = 4$	**27.928**	**0.785**	**6.428**
$n_{subs} = 24, n_{ovls} = 6$	27.759	0.781	6.563
$n_{subs} = 32, n_{ovls} = 8$	27.736	0.778	6.596

Table 4: Ablation results for different settings of the number of subgroups bands (n_{subs}) and the number of overlaps bands (n_{ovls}) on the PaviaC dataset at scale 4.

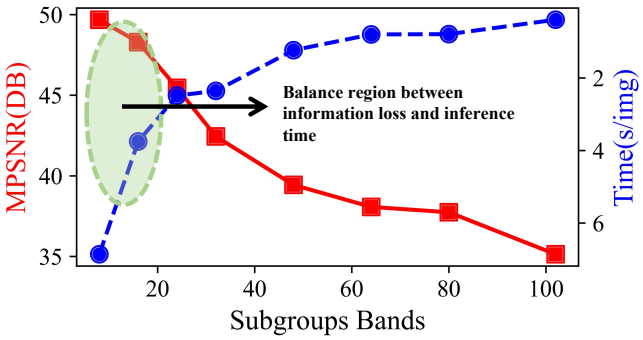

Figure 9: Comparison results of information loss and inference time on different subgroups numbers.

mance due to the large amount of redundant data in HSI and limited training samples. Training with partial bands step by step led to fragmented learning, resulting in inferior performance as well. The autoencoder plays a crucial role in handling the rich HSI information and facilitating effective encoding for superior results with the diffusion SR model.

Comparison of Subgroups Numbers

We provide two types of evidence to support the rationality of our subgroups selection. Firstly, we present the overall SR performance of the model under different subgroups bands settings, where we compare the results in terms of several commonly used evaluation metrics. The results are shown in Table 4. Secondly, we evaluate the information reconstruction loss (measured by PSNR) and the inference time under different subgroups bands settings, as shown in Figure 9. It can be observed that our selection not only achieves high SR performance, but also strikes the right balance between information loss and inference time.

Comparison of Inference Time

We also conducted experiments to compare the inference time with the pure diffusion model (without our GAE), as presented in Table 5. Our method demonstrates a remarkable reduction in the inference time while achieving superior performance. It can be observed that directly applying the diffusion model for full-band SR results in poor performance, despite the advantage of shorter inference time. On the other hand, performing SR on each band separately using the diffusion model requires traversing the entire band spectrum, resulting in lengthy inference steps and ignoring the spectral correlations in HSI data, which in turn leads to suboptimal

Models	MPSNR	T_{64} (s/img)	T_{128} (s/img)	T_{256}(s/img)
Ours	27.928	3.752	3.876	6.174
Diff-PB	27.262	13.629	14.124	23.018
Diff-FB	10.906	0.396	0.435	0.914

Table 5: Comparison results of inference times on the PaviaC dataset. T_{64} represents the average inference time of 64×64 images.

results. In contrast, our approach significantly reduces the number of inference steps by performing SR on only a few intermediate hidden variables, resulting in much faster inference. Furthermore, the integration of the GAE takes into account the spectral similarity, which further enhances the SR results.

Overall, our method not only achieves better performance in terms of SR quality but also dramatically reduces the inference time, making it a highly efficient and effective solution for HSI SR tasks.

Conclusion

In this paper, we proposed a novel two-stage diffusion-based framework for HSI SR tasks. Our approach effectively addresses the challenge of the diffusion model's convergence with high-dimensional data and significantly reduces the inference time by integrating a autoencoder with the diffusion model. This combination enables efficient computation and facilitates superior SR results. and achieves significant improvements both visually and metrically. Going forward, we plan to explore more applications of diffusion models in HSI tasks, aiming to further enhance the field of HSI research.

Acknowledgments

The authors gratefully acknowledge the support from the National Natural Science Foundation of China under Grant No. 62376205 and Grant No. 62036006, as well as the support from Alibaba Group through Alibaba Innovative Research Program. Their contributions have been instrumental to the success of this research; the views and findings herein are those of the authors and do not necessarily reflect the views of the funders.

References

Chakrabarti, A.; and Zickler, T. 2011. Statistics of real-world hyperspectral images. In *CVPR 2011*, 193–200.

Chung, H.; Lee, E. S.; and Ye, J. C. 2022. MR image denoising and super-resolution using regularized reverse diffusion. *IEEE Transactions on Medical Imaging*, 42(4): 922–934.

Cloutis, E. A. 1996. Review Article Hyperspectral geological remote sensing: evaluation of analytical techniques. *International Journal of Remote Sensing*, 17(12): 2215–2242.

Fei, B. 2020. *Hyperspectral imaging in medical applications*, 523–565. Data Handling in Science and Technology. Elsevier Ltd.

Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; and Qian, Y. 2021. STransFuse: Fusing swin transformer

and convolutional neural network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 10990–11003.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit Diffusion Models for Continuous Super-Resolution.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hu, J. F.; Huang, T. Z.; and Deng, L. J. 2021. Fusformer: A Transformer-based Fusion Approach for Hyperspectral Image Super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*.

Hu, J.-F.; Huang, T.-Z.; Deng, L.-J.; Dou, H.-X.; Hong, D.; and Vivone, G. 2022. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Jiang, J.; Sun, H.; Liu, X.; and Ma, J. 2020. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6: 1082–1096.

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

Kersting, K.; Xu, Z.; Wahabzada, M.; Bauckhage, C.; Thurau, C.; Roemer, C.; Ballvora, A.; Rascher, U.; Léon, J.; and Pluemer, L. 2012. Pre-symptomatic prediction of plant drought stress using dirichlet-aggregation regression on hyperspectral images. In *National Conference on Artificial Intelligence*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

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

Li, Q.; Wang, Q.; and Li, X. 2021. Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(10): 8693–8703.

Li, Y.; Zhang, L.; Dingl, C.; Wei, W.; and Zhang, Y. 2018. Single Hyperspectral Image Super-Resolution with Grouped Deep Recursive Residual Network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 1–4.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Liu, C.; and Dong, Y. 2022. CNN-Enhanced graph attention network for hyperspectral image super-resolution using non-local self-similarity. *International Journal of Remote Sensing*, 43(13): 4810–4835.

Liu, C.; Fan, Z.; and Zhang, G. 2022. Gjtd-lr: A trainable grouped joint tensor dictionary with low-rank prior for single hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

Liu, J.; Wu, Z.; Xiao, L.; Sun, J.; and Yan, H. 2020. A Truncated Matrix Decomposition for Hyperspectral Image Super-Resolution. *IEEE Transactions on Image Processing*, PP(99): 1–1.

Liu, J.; Wu, Z.; Xiao, L.; and Wu, X.-J. 2022a. Model inspired autoencoder for unsupervised hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Liu, J.; Yuan, Z.; Pan, Z.; Fu, Y.; Liu, L.; and Lu, B. 2022b. Diffusion model with detail complement for super-resolution of remote sensing. *Remote Sensing*, 14(19): 4834.

Liu, Y.; Hu, J.; Kang, X.; Luo, J.; and Fan, S. 2022c. Interactformer: Interactive Transformer and CNN for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Liu, Y.; Hu, J.; Kang, X.; Luo, J.; and Fan, S. 2022d. Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Mao, Y.; Jiang, L.; Chen, X.; and Li, C. 2023. DisC-Diff: Disentangled Conditional Diffusion Model for Multi-Contrast MRI Super-Resolution. *arXiv preprint arXiv:2303.13933*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8162–8171. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2023. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Thai, B.; and Healey, G. 2002. Invariant subpixel material detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(3): 599–608.

Wang, X.; Hu, Q.; Jiang, J.; and Ma, J. 2022. A Group-Based Embedding Learning and Integration Network for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Wang, X.; Ma, J.; and Jiang, J. 2022. Hyperspectral Image Super-Resolution via Recurrent Feedback Embedding and Spatial–Spectral Consistency Regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Yokoya, N.; and Iwasaki, A. 2016. Airborne hyperspectral data over Chikusei.

Zhou, J.; Sheng, J.; Fan, J.; Ye, P.; He, T.; Wang, B.; and Chen, T. 2023. When Hyperspectral Image Classification Meets Diffusion Models: An Unsupervised Feature Learning Framework. *arXiv preprint arXiv:2306.08964*.