S^2CycleDiff: Spatial-Spectral-Bilateral Cycle-Diffusion Framework for Hyperspectral Image Super-resolution

Jiahui Qu*, Jie He*, Wenqian Dong†, Jingyu Zhao

State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China
jhqu@xidian.edu.cn, jiehe@stu.xidian.edu.cn, wqdong@xidian.edu.cn, jingyuzhao@stu.xidian.edu.cn

Abstract

Hyperspectral image super-resolution (HISR) is a technique that can break through the limitation of imaging mechanism to obtain the hyperspectral image (HSI) with high spatial resolution. Although some progress has been achieved by existing methods, most of them directly learn the spatial-spectral joint mapping between the observed images and the target high-resolution HSI (HrHSI), failing to fully preserve the spectral distribution of low-resolution HSI (LrHSI) and the spatial distribution of high-resolution multispectral imagery (HrMSI). To this end, we propose a spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff) for HISR, which can step-wise generate the HrHSI with high spatial-spectral fidelity by learning the conditional distribution of spatial and spectral super-resolution processes bilaterally. Specifically, a customized conditional cycle-diffusion framework is designed as the backbone to achieve the spatial-spectral-bilateral super-resolution by repeated refinement, wherein the spatial/spectral guided pyramid denoising (SGPD) module seperately takes HrMSI and LrHSI as the guiding factors to achieve the spatial details injection and spectral correction. The outputs of the conditional cycle-diffusion framework are fed into a complementary fusion block to integrate the spatial and spectral details to generate the desired HrHSI. Experiments have been conducted on three widely used datasets to demonstrate the superiority of the proposed method over state-of-the-art HISR methods. The code is available at https://github.com/Jiahuiqu/S2CycleDiff.

Introduction

Hyperspectral image (HSI) is a data cube that contains hundreds of spectral bands. Compared with other remote sensing images, HSI can provide more detailed spectral information that reflects the properties of objects (Dong et al. 2023). This unique characteristic of HSI enables a diverse range of applications, including natural disaster monitoring (Liu et al. 2017), urban planning (Huang et al. 2017; Qu et al. 2023), and water resource management (Khan et al. 2018). However, due to the limitation of the imaging mechanism, there is a trade-off between spatial and spectral resolution in HSI. This means that a sensor capturing images with high spectral resolution at the expense of spatial resolution. The lack of spatial information limits the application of HSI in downstream tasks, making it imperative to acquire images that possess both high spectral and spatial resolution.

The fusion-based hyperspectral image super-resolution (HISR) has become a crucial approach for improving the spatial resolution of HSI, which aims to fuse HSI and multispectral (MSI) to generate the HSI with high spatial resolution. In recent years, lots of HISR methods have been introduced, which can be roughly divided into two main categories, i.e., traditional methods and deep learning-based methods.

The traditional methods for HISR mainly include component substitution (CS), multiresolution analysis (MRA), Bayesian approaches and matrix factorization (Loncan et al. 2015). These methods require some prior knowledge to design reasonable strategies to integrate spatial information from MSI into HSI. However, the traditional methods often heavily rely on manually designed feature extractor, which makes it difficult to effectively capture complex features when dealing with high-dimensional data. In recent years, deep learning (DL) has achieved significant advancements, which has been widely used in HISR (Scarpa, Vitale, and Cozzolino 2018; Xu et al. 2020; Hu et al. 2022; Yao et al. 2020). Initially, convolutional neural networks (CNNs) were employed as a power tool for HISR and were trained under the guidance of MSI to learn the nonlinear mapping from low-resolution to high-resolution HSI (HrHSI). Furthermore, deep generative models have shown remarkable success in HISR due to their capability to approximate data distribution. Variational autoencoders (VAEs) (Kingma and Welling 2022) learn latent data distribution by probability encoding and decoding. Generative adversarial networks (GANs) (Goodfellow et al. 2014) leverage adversarial training between a generator and a discriminator to generate visually realistic samples. Recently, Diffusion Probabilistic Model (DPM) (Ho, Jain, and Abbeel 2020) has shown great potential in the field of computer vision (CV). The DPM refines the process of image reconstruction into multiple steps, gradually obtaining high-quality images. In contrast to other Generative models, the training process of the DPM exhibits improved stability and efficiency, eliminating the need for intricate adversarial training like GANs.

Although existing methods have made some achieve-

*These authors contributed equally.
†The corresponding author.

ments, most of them directly learn the spatial-spectral joint mapping between the low-resolution hyperspectral image (LrHSI) and the target high-resolution hyperspectral image (HrHSI), thereby failing to fully preserve the spectral distribution of LrHSI and the spatial distribution of high-resolution multispectral imagery (HrMSI) respectively.

Inspired by recent advancements in deep generative models, in this paper, we propose a spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff), which learns the conditional distribution of spatial and spectral super-resolution processes bilaterally to generate the ideal HrHSI stably. Specifically, the degraded HrHSI can be first obtained by adding Gaussian noise to the ground truth image during the forward process. For the reverse process, the conditional cycle-diffusion framework learns the conditional distribution of the super-resolution process from the spatial and spectral dimensions through the guidance of HrMSI and LrHSI, and generates the preliminary HrHSIs from these two aspects stably. Furthermore, the complementary fusion module integrates the outputs of the conditional cycle-diffusion framework to generate the high-quality HrHSI. The contribution we have made can be summarized as follows:

- We propose a spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff) for hyperspectral super-resolution, which designs the conditional cycle-diffusion framework as the backbone to learn the conditional distribution of super-resolutions in terms of spatial and spectral aspects, so as to obtain the high-quality HrHSI by a sequence of refinements.
- The customized conditional cycle-diffusion framework is designed to encode the conditional distribution of spatial-spectral-bilateral super-resolution processes and generate the HrHSIs with complementary spatial and spectral information.
- The spatial/spectral guided pyramid denoising (SGPD) module is proposed, in which HrMSI and LrHSI are adopted as the guiding factors to guide the image denoising process, so as to minimize the spatial-spectral distortion by spatial details injection and spectral correction respectively.

Related Work

Traditional Methods for HISR The traditional methods can be categorized into four classes, i.e., component substitution (CS), multiresolution analysis (MRA), Bayesian approaches and matrix factorization. Principal component analysis (Shettigara 1992), intensity-hue-saturation method (Choi 2006), and guided filtering (Qu, Li, and Dong 2017) belong to the CS methods, in which the spatial and spectral information of observed images is separated, and then the HSI is improved by replacing the spatial component with that from MSI. The MRA methods, such as smoothing filter-based intensity modulation (SFIM) (Liu 2000), the MTF-generalized Laplacian pyramid (MTF-GLP) (Aiazzi et al. 2006) method, and regression-based high-pass modulation (Wang et al. 2022a), extract high-frequency details of MSI and then inject them into the HSI. Despite the simplicity of CS and MRA methods, the spectral discrepancy between the MSI and HSI can lead to significant spectral distortion. The Bayesian approaches rely on Bayesian statistics to infer probabilistic parameters and design different Bayesian estimators for combining the co-registered paired MSI and HSI (Wei, Dobigeon, and Tourneret 2015). Coupled nonnegative matrix factorization (CNMF) (Yokoya, Yairi, and Iwasaki 2011) is a representative work of matrix factorization methods, which relies on the unmixing procedure. Although these two methods have made some progress, they rely heavily on artificial prior information, making it difficult to obtain the optimal solution for this ill-posed problem.

Deep Learning-based Methods for HISR The deep learning-based methods have proven to be effective in obtaining HrHSI, which employ the elaborate network architecture to efficiently extract the spatial and spectral features without manual parameter selection. MHFNet embedded the mapping relationship between the observed images and the target image into a unfolding network to gradually obtain the HrHSI (Xie et al. 2022). In addition, generative models have been proven to be an effective framework for HISR (Shi et al. 2022). FusionNet was proposed as a novel VAE framework for integrating the spatial and spectral information of LrHSI and HrMSI, combined with meta-learning to enable fast adaptation to different scenes (Wang et al. 2020). Moreover, HSSRGAN was proposed as a GAN-based method for HISR, which designed a generator to enhance the spatial feature and refine the spectral information (Wang et al. 2021). Recently, DPM emerged as a powerful approach for various image generation tasks, including text-to-image translation (Rombach et al. 2022), image restoration(Kawar et al. 2022), and other high-level image manipulation tasks (Wang et al. 2022b). The DPM stands out among other generative models due to its stable training process and ability to generate high-quality images. Inspired by the advancements, the proposed S^2CycleDiff encodes the conditional distribution of the spatial-spectral-bilateral super-resolution process into a cycle-diffusion framework to stably generate the desired HrHSI.

Proposed Method

In this section, the details of the proposed spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff) are presented, the flow chart is shown in Figure. 1, and the illustration of each time step operation is shown in Figure. 2.

Overall Framework

Given a training paired LrHSI $\mathbf{X} \in \mathbb{R}^{B\times h\times w}$, HrMSI $\mathbf{Y} \in \mathbb{R}^{b\times H\times W}$, and the corresponding ground truth HrHsI $\mathbf{Z} \in \mathbb{R}^{B\times H\times W}$, where $\{b, B\}$, $\{h, H\}$, and $\{w, W\}$ denote the number of bands, heights, and widths, respectively. Since the LrHSI contains richer spectral information and HrMSI contains more spatial details, it is generally considered that $b < B$, $h < H$, and $w < W$. We aim to train a model to aggregate the complementary information between LrHSI and HrMSI to obtain the desired HrHSI $\widetilde{\mathbf{Z}} \in \mathbb{R}^{B\times H\times W}$. In this paper, as depicted in Figure. 1, the proposed S^2CycleDiff

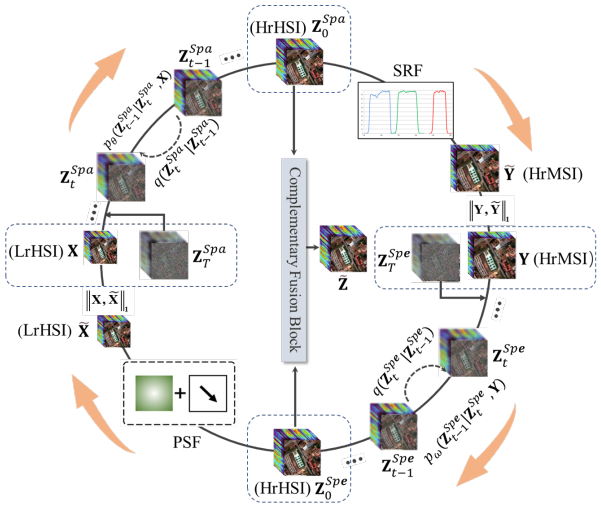

Figure 1: The flow chart of the proposed spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff) for hyperspectral image super-resolution.

designs a conditional cycle-diffusion structure as the backbone to generate the HrHSI with high spatial-spectral fidelity through a series of refinement processes. Specifically, the degraded HrHSI can be obtained by adding stochastic noise step by step during the forward process, and then the restored HrHSIs can be obtained by the spatial and spectral guided reverse diffusion process conditioned on HrMSI and LrHSI, respectively. In addition, we design consistency constraint strategies for model training to ensure the spectral distribution of the target HrHSI consistent with that of LrHSI and the spatial distribution is consistent with that of HrMSI. The ideal target HrHSI can be obtained by integrating the complementary information from the outputs of the spatial-spectral-bilateral super-resolution process. The proposed method iteratively refines the spatial details and spectral signature through the cycle-diffusion process, which not only ensures stable model training but also maximizes the spatial-spectral fidelity.

Forward Diffusion Process

Given a clean HrHSI $\mathbf{Z}_0 \sim q(\mathbf{Z}_0)$, where $\mathbf{Z}_0 = \mathbf{Z}$, the forward diffusion process gradually add the stochastic noise to $\mathbf{Z}_0$ through a Markov chain to generate noisy image $\mathbf{Z}_T$. Each time step t of the forward diffusion process can be defined as follows,

$$q(\mathbf{Z}_t|\mathbf{Z}_{t-1}) = \mathcal{N}(\mathbf{Z}_t; \sqrt{1-\beta_t}\mathbf{Z}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where $t \in [0, T]$, $\mathcal{N}(\mathbf{Z}_t; \sqrt{1-\beta_t}\mathbf{Z}_{t-1}, \beta_t\mathbf{I})$ represents that $\mathbf{Z}_t$ obeys the Gaussian distribution with mean $\sqrt{1-\beta_t}\mathbf{Z}_{t-1}$ and variance $\beta_t\mathbf{I}$, and $\beta_t \in (0, 1)$ is the variance schedule. Further, we can obtain $\mathbf{Z}_t$ from $\mathbf{Z}_0$ directly by reparametrization technique as follows,

$$q(\mathbf{Z}_t|\mathbf{Z}_0) = \mathcal{N}(\mathbf{Z}_t; \sqrt{\overline{\gamma}_t}\mathbf{Z}_0, (1-\overline{\gamma}_t)\mathbf{I}) \tag{2}$$

$$\mathbf{Z}_t = \sqrt{\overline{\gamma}_t}\mathbf{Z}_0 + \sqrt{1-\overline{\gamma}_t}\boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\gamma_t = 1 - \beta_t$, and $\overline{\gamma}_t = \prod_{i=0}^{t} \gamma_t$. The forward process transforms the data distribution into a standard Gaussian distribution viz. $\mathbf{Z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Conditional Cycle-Diffusion Process

The reverse process aims to remove the noise added during the forward process and infers $\mathbf{Z}_0$ through iterative refinements, which can be approximated by a neural network. In this paper, we design a conditional cycle-diffusion framework to generate the HrHSIs by spatial-spectral-bilateral super-resolution process. Particularly, the spatial super-resolution process is conditioned on LrHSI and guided by HrMSI, while the spectral super-resolution process is conditioned on HrMSI and guided by LrHSI. The conditional distribution of the spatial-spectral-bilateral super-resolution process, viz. the reverse process, is introduced below.

The posterior distribution $p(\mathbf{Z}_{t-1}|\mathbf{Z}_t, \mathbf{Z}_0)$ is useful to the reverse process, which can be formulated as follows,

$$p(\mathbf{Z}_{t-1}|\mathbf{Z}_t, \mathbf{Z}_0) = \mathcal{N}(\mathbf{Z}_{t-1}; \mu(\mathbf{Z}_t, \mathbf{Z}_0), \Sigma(\mathbf{Z}_t, \mathbf{Z}_0)) \tag{4}$$

where the mean $\mu(\mathbf{Z}_t, \mathbf{Z}_0)$ and variance $\Sigma(\mathbf{Z}_t, \mathbf{Z}_0)$ can be represent as,

$$\mu(\mathbf{Z}_t, \mathbf{Z}_0) = \frac{\sqrt{\gamma_t}\,(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t}\mathbf{Z}_t + \frac{\sqrt{\overline{\gamma}_{t-1}}\,(1-\gamma_t)}{1-\overline{\gamma}_t}\mathbf{Z}_0 \tag{5}$$

$$\Sigma(\mathbf{Z}_t, \mathbf{Z}_0) = \frac{(1-\overline{\gamma}_{t-1})\,(1-\gamma_t)}{1-\overline{\gamma}_t} \tag{6}$$

It is worth noting that, in this paper, the forward process generates two noisy images, namely $\mathbf{Z}_T^{Spa}$ and $\mathbf{Z}_T^{Spe}$ ($\mathbf{Z}_T^{Spa} = \mathbf{Z}_T^{Spe} = \mathbf{Z}_T$), for the spatial and the spectral super-resolution process.

For each time step, the conditional distribution of spatial-spectral-bilateral super-resolution process can be encoded into the cycle-diffusion framework as shown in Figure. 2 and produce the prediction $\mathbf{Z}_{t,0}^{Spa}$ and $\mathbf{Z}_{t,0}^{Spe}$ at each time step respectively, which can be defined as follows,

$$\mathbf{Z}_{t,0}^{Spa} = f_\theta(\mathbf{Z}_t^{Spa}, \mathbf{X}, \mathbf{Y}) \tag{7}$$

$$\mathbf{Z}_{t,0}^{Spe} = f_\omega(\mathbf{Z}_t^{Spe}, \mathbf{Y}, \mathbf{X}) \tag{8}$$

where θ and ω are the model parameters of the bilateral super-resolution processes, and $f_\theta(\mathbf{Z}_t^{Spa}, \mathbf{X}, \mathbf{Y})$ represents the spatial super-resolution process that is conditioned on LrHSI and guided by HrMSI, $f_\omega(\mathbf{Z}_t^{Spe}, \mathbf{Y}, \mathbf{X})$ represents the spectral super-resolution process that is conditioned on HrMSI and guided by LrHSI. Then, the inputs of the next time step can be obtained as follows according to Eq. (5),

$$\begin{aligned}\mathbf{Z}_{t-1}^{Spa} = &\frac{\sqrt{\overline{\gamma}_{t-1}}(1-\gamma_t)}{1-\overline{\gamma}_t} f_\theta(\mathbf{Z}_t^{Spa}, \mathbf{X}, \mathbf{Y}) \\ &+ \frac{\sqrt{\gamma_t}(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t}\mathbf{Z}_t^{Spa} + \sqrt{\frac{(1-\overline{\gamma}_{t-1})\,(1-\gamma_t)}{1-\overline{\gamma}_t}}\boldsymbol{\varepsilon}\end{aligned} \tag{9}$$

$$\begin{aligned}\mathbf{Z}_{t-1}^{Spe} = &\frac{\sqrt{\overline{\gamma}_{t-1}}(1-\gamma_t)}{1-\overline{\gamma}_t} f_\omega(\mathbf{Z}_t^{Spe}, \mathbf{Y}, \mathbf{X}) \\ &+ \frac{\sqrt{\gamma_t}(1-\overline{\gamma}_{t-1})}{1-\overline{\gamma}_t}\mathbf{Z}_t^{Spe} + \sqrt{\frac{(1-\overline{\gamma}_{t-1})(1-\gamma_t)}{1-\overline{\gamma}_t}}\boldsymbol{\varepsilon}\end{aligned} \tag{10}$$

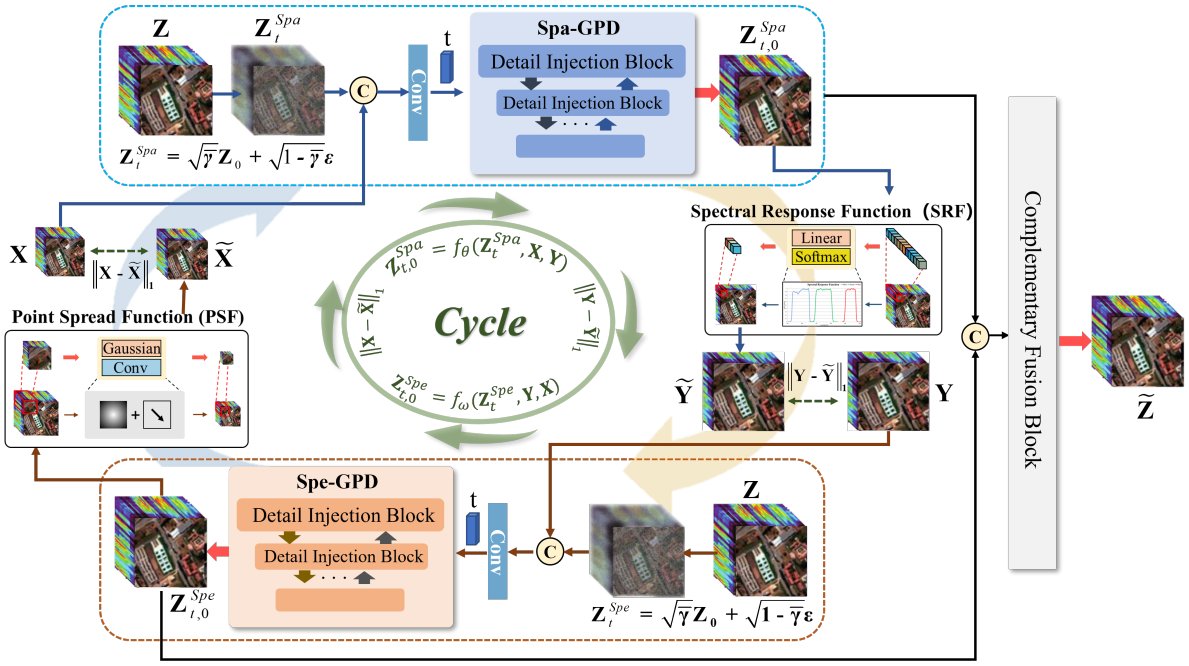

Figure 2: The illustration of each time step operation in the proposed spatial-spectral-bilateral cycle-diffusion framework (S^{2}CycleDiff) for hyperspectral image super-resolution.

Spatial/Spectral Guided Pyramid Denoising Module To prevent the spatial texture blur and spectral distortion during the super-resolution process, we design a spatial/spectral guided pyramid denoising (SGPD) module as the core component of the conditional cycle diffusion framework. The HrMSI and LrHSI are used as the guiding factors to guide the generation of high-fidelity spatial and spectral information in the bilateral super-resolution processes, respectively. The SGPDs of the bilateral super-resolution processes exhibit a symmetrical configuration, which consist of multiple layers of detail injection block (DIB). At each time step, SGPD takes the multi-scale features of the noisy image and the guiding factor as input and achieves image denoising through downsampling and upsampling operations between each layer. Therefore, the spatial-spectral-bilateral super-resolution process can be expressed as follows,

$$\begin{aligned}\mathbf{Z}_{t,0}^{Spa} &= f_\theta(\mathbf{Z}_t^{Spa}, \mathbf{X}, \mathbf{Y}) \\ &= \text{Spa-GPD}(\text{Conv}_{3\times3}(\mathbf{Z}_t^{Spa}, \mathbf{X}), \mathbf{Y}) \\ &= \text{Spa-GPD}(\widehat{\mathbf{Z}}_t^{Spa}, \mathbf{Y})\end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{Z}_{t,0}^{Spe} &= f_\omega(\mathbf{Z}_t^{Spe}, \mathbf{Y}, \mathbf{X}) \\ &= \text{Spe-GPD}(\text{Conv}_{3\times3}(\mathbf{Z}_t^{Spe}, \mathbf{Y}), \mathbf{X}) \\ &= \text{Spe-GPD}(\widehat{\mathbf{Z}}_t^{Spe}, \mathbf{X})\end{aligned} \tag{12}$$

where Spa-GPD$(\cdot)$ and Spe-GPD$(\cdot)$ represent the SGPD modules of the spatial and the spectral super-resolution process, respectively, and Conv$_{3\times3}(\cdot)$ represents the 3×3 convolutional layer used to unify dimensions before inputting the noisy image into the SGPD module.

Taking the spatial super-resolution process as an example, the illustration of the l-th DIB of SGDP is shown in Figure. 3, which leverages the cross-attention mechanism to couple the finer spatial information of HrMSI. The operation of the l-th DIB can be formulated as follows,

$$(\widehat{\mathbf{Z}}_t^{Spa,l}, \mathbf{Z}_{t,0}^{Spa,l}) = \text{DIB}_t^l(\widehat{\mathbf{Z}}_t^{Spa,l-1}, \mathbf{Z}_{t,0}^{Spa,l+1}, \mathbf{Y}) \tag{13}$$

where DIB$_t^l(\cdot)$ represents the l-th DIB of SGDP,

$$\begin{cases} \mathbf{Z}_{t,0}^{Spa,l} = R_2^l([(R_1^l(\mathbf{Y}) + \dfrac{\mathbf{Q}_\mathbf{Y}^l(\mathbf{K}_\mathbf{Y}^l)^\mathrm{T}}{\sqrt{\mathrm{B}}}\mathbf{V}_\mathbf{Y}^l); \\ \qquad\qquad (R_2^l(\widehat{\mathbf{Z}}_t^{Spa,l-1}) + \dfrac{\mathbf{Q}_\mathbf{Y}^l(\mathbf{K}_\mathbf{Z}^l)^\mathrm{T}}{\sqrt{\mathrm{B}}}\mathbf{V}_\mathbf{Z}^l); (\mathbf{Z}_{t,0}^{Spa,l+1}\uparrow)]) \\ \widehat{\mathbf{Z}}_t^{Spa,l} = R_2^l(\widehat{\mathbf{Z}}_t^{Spa,l-1})\downarrow \end{cases} \tag{14}$$

$$\begin{cases} \mathbf{Q}_\mathbf{Y}^l = \varphi_q^l(\mathbf{Y}), \mathbf{K}_\mathbf{Y}^l = \varphi_k^l(\mathbf{Y}), \mathbf{V}_\mathbf{Y}^l = \varphi_v^l(\mathbf{Y}) \\ \mathbf{K}_\mathbf{Z}^l = \rho_k^l(\widehat{\mathbf{Z}}_t^{Spa,l-1}), \mathbf{V}_\mathbf{Z}^l = \rho_v^l(\widehat{\mathbf{Z}}_t^{Spa,l-1}) \end{cases} \tag{15}$$

where $R_n^l(\cdot)$ represents a sequence of n stacked residual blocks in the l-th DIB ($R_2^l(\cdot)$ is to extract features from inputs while $R_1^l(\mathbf{Y})$ is adopted to unify the dimensions of Y and input features), B is the scale factor, $\uparrow$ represents the upsampling operation, $\downarrow$ represents the down-sampling operation, $\varphi_q^l(\cdot)$, $\varphi_k^l(\cdot)$, and $\varphi_v^l(\cdot)$ represent the specific operators that consist of a residual block and a 1×1 convolutional layer, $\rho_k^l(\cdot)$ and $\rho_v^l(\cdot)$ are the operators that contains two residual block and a 1×1 convolutional layer for dimension unification. The residual block consists of 3×3 convolutional layer, Swish activation function, and group normalization operation.

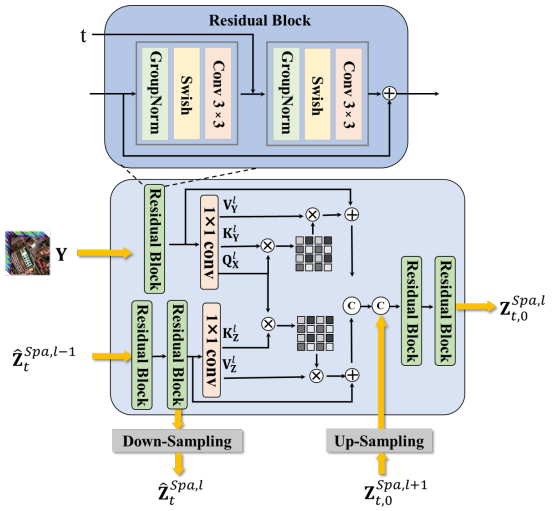

Figure 3: The illustration of the l-th detail injection block (DIB) of spatial guided pyramid denoising module (Spa-GDP).

Complementary Fusion Block To make full use of the complementary spatial and spectral information between $\mathbf{Z}_{t,0}^{Spa}$ and $\mathbf{Z}_{t,0}^{Spe}$, we use a complementary fusion block that consists of a series of convolutional layers to generate the ideal HrHSI with high spatial-spectral fidelity. The process can be represented as follows,

$$\begin{aligned}\widetilde{\mathbf{Z}} &= f_\eta(f_\theta(\mathbf{Z}_t^{Spa}, \mathbf{X}), f_\omega(\mathbf{Z}_t^{Spe}, \mathbf{Y})) \\ &= f_\eta(\mathbf{Z}_{t,0}^{Spa}, \mathbf{Z}_{t,0}^{Spe})\end{aligned} \tag{16}$$

where η is the parameter of the complementary fusion block, and $\widetilde{\mathbf{Z}}$ represents the generated ideal HrHSI.

Training Objective

Given the training sets $\{(\mathbf{X}^n, \mathbf{Y}^n, \mathbf{Z}^n)|n = 1, 2, \cdots, N\}$, where N is the number of training samples, $\mathbf{X}^n$ and $\mathbf{Y}^n$ represent the n-th training samples, and $\mathbf{Z}^n$ is the corresponding ground truth map. Furthermore, a series of consistency constraint strategies are used in the process of network training to obtain high-quality HrHSI. We define $\mathcal{L}_1$ to achieve this objective,

$$\mathcal{L}_1 = \frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{Z}^n - \mathbf{Z}_{t,0}^{Spa,n}\|_1 + \|\mathbf{Z}^n - \mathbf{Z}_{t,0}^{Spe,n}\|_1) \tag{17}$$

where $\mathbf{Z}_{t,0}^{Spa,n}$ and $\mathbf{Z}_{t,0}^{Spe,n}$ are the n-th outputs of bilateral super-resolution.

Additionally, a paired of HrMSI $\widetilde{\mathbf{Y}}^n$ and LrHSI $\widetilde{\mathbf{X}}^n$ can be obtained from $\mathbf{Z}_{t,0}^{Spa,n}$ and $\mathbf{Z}_{t,0}^{Spe,n}$ by spectral response function (SRF) and point spread function (PSF), respectively. We design $\mathcal{L}_2$ loss function to constrain the consistency between $\{\mathbf{X}^n, \mathbf{Y}^n\}$ and $\{\widetilde{\mathbf{X}}^n, \widetilde{\mathbf{Y}}^n\}$, which can be defined as follows,

$$\mathcal{L}_2 = \frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{X}^n - \widetilde{\mathbf{X}}^n\|_1 + \|\mathbf{Y}^n - \widetilde{\mathbf{Y}}^n\|_1) \tag{18}$$

$$\widetilde{\mathbf{X}}^n = \mathrm{PSF}(\mathbf{Z}_{t,0}^{Spe,n}), \widetilde{\mathbf{Y}}^n = \mathrm{SRF}(\mathbf{Z}_{t,0}^{Spa,n}) \tag{19}$$

where $\mathrm{PSF}(\cdot)$ represents the Gaussian blurring followed by a 3×3 convolutional operation, and $\mathrm{SRF}(\cdot)$ denotes the linear operation. Meanwhile, two consistent MSIs can be obtained from HrMSI and LrHSI through the above PSF and SRF, respectively, which can be constrained as follows,

$$\mathcal{L}_3 = \frac{1}{N}\sum_{n=1}^{N}\|\mathrm{PSF}(\mathbf{Y}^n) - \mathrm{SRF}(\mathbf{X}^n)\|_1 \tag{20}$$

Furthermore, it is necessary to constrain the consistency between the generated ideal HrHSI and the ground truth map, which can be formulated as follows,

$$\mathcal{L}_4 = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{Z}^n - \widetilde{\mathbf{Z}}^n\|_1 \tag{21}$$

where $\widetilde{\mathbf{Z}}^n$ represents the generated high-quality HrHSI.

Thus, the total loss function $\mathcal{L}$ can be represented as follows,

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 \tag{22}$$

Experiment

Datasets and Implementation Details

To illustrate the effectiveness of the proposed method, we conduct the comparative experiments with several competing methods on three public datasets, namely CAVE, Pavia Center, and Chikusei. The CAVE dataset consists of 32 images with a size of 512 × 512 × 31, where 22 images are selected as the training set, while the remaining 10 images are allocated to the test set. The size of the Chikusei dataset is 2304 × 2048 × 110. The top area with 1792 × 2048 pixels is selected as the training data, while the remaining area is used as the test data that is split into four patches with the size of 512 × 512. The Pavia Center dataset consists of 102 bands with the size of 960 × 640, which is cropped into patches with 160 × 160 pixels for training and testing. We use the Wald's protocol (Wald, Ranchin, and Mangolini 1997) to generate pairs of LrHSI and HrMSI for training. The spectral bands of HrMSI in CAVE, Pavia Center, and Chikusei are 3, 4, and 3, respectively. These three datasets are degraded spatially with a factor of 4. In order to match the spatial dimension of LrHSI with that of HrMSI and HrHSI, we adopt bicubic interpolation so that the concatenation of noisy HrHSI and LrHSI can be regarded as inputs for spatial super-resolution in S^{2}CycleDiff. We conducted the experiments with the PyTorch framework and trained on two NVIDIA GeForce RTX 3090 GPU. The experiments were conducted with a batch size of 8 and 100k iterations on all datasets. The Adam optimizer is employed for the optimization process, with a maximum learning rate set at 0.0001.The time step T is set to 2000, and the hyperparameter sequence $\{\beta_1, \beta_2, ..., \beta_n\}$ was defined with uniform growth ranging from 0 to 0.02.

Methods	CAVE				Chikusei				Pavia Center			
	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM
GSA	39.5631	5.3966	2.6803	0.9782	33.2088	2.1924	2.3728	0.9052	33.4475	5.8496	3.2015	0.9538
FUSE	33.9303	4.4017	4.7760	0.9600	35.5124	1.9726	2.3549	0.9478	28.8145	7.9142	5.5524	0.8820
CNMF	38.8704	4.9593	3.0723	0.9727	33.3137	2.2277	2.5688	0.9355	31.6690	5.0485	4.2487	0.9168
SSR-NET	42.3322	4.1381	1.8452	0.9830	39.5345	1.4703	1.5435	0.9715	42.9138	2.9478	1.6410	0.9837
MoG-DCN	40.9808	3.2583	2.1891	0.9803	39.1090	1.4171	1.4380	0.9696	42.9275	2.8845	1.6354	0.9843
LAGC-NET	42.9198	7.9494	1.7216	0.9542	40.5422	1.2895	1.7007	0.9772	40.0506	3.3257	1.7127	0.9818
LightNet	41.5682	3.5185	2.0093	0.9856	40.1045	1.3091	1.5081	0.9747	42.1629	3.0056	1.8158	0.9802
PSRT	43.4795	2.8763	1.6422	0.9896	39.7903	1.4464	1.7904	0.9729	39.7629	3.0586	2.4049	0.9827
S^2CycleDiff	**43.9264**	**2.7907**	**1.5834**	**0.9898**	**43.8817**	**1.2756**	**1.3402**	**0.9909**	**43.3698**	**2.7814**	**1.5903**	**0.9847**

Table 1: Quantitative results obtained by different methods on CAVE, Chikusei, and Pavia Center datasets.

Methods	CAVE $\times 16$			
	PSNR	SAM	ERGAS	SSIM
SSR-NET	37.8502	5.6890	3.0349	0.9694
MoG-DCN	35.2812	7.6792	3.9560	0.9455
LAGC-NET	39.6383	5.9763	2.4719	0.9753
LightNet	39.0308	5.3803	2.7409	0.9781
PSRT	39.3620	4.5799	2.9440	0.9716
S^2CycleDiff	**39.9191**	**4.5523**	**2.4022**	**0.9840**

Table 2: Quantitative results of competing methods on CAVE dataset with a upsampling factor of 16.

Dataset	Model	PSNR	SAM	ERGAS	SSIM
CAVE	Variant-Spe	41.0376	3.9461	2.2326	0.9832
	Variant-Spa	41.3527	3.7541	2.0873	0.9860
	S^2CycleDiff	**43.9264**	**2.7907**	**1.5834**	**0.9898**
Chikusei	Variant-Spe	41.6815	1.9208	2.2181	0.9794
	Variant-Spa	41.4683	1.8818	1.7288	0.9854
	S^2CycleDiff	**43.8817**	**1.2756**	**1.3402**	**0.9909**
Pavia Center	Variant-Spe	40.5205	3.3896	2.1224	0.9775
	Variant-Spa	41.9572	3.1807	1.8424	0.9809
	S^2CycleDiff	**43.3698**	**2.7814**	**1.5903**	**0.9847**

Table 3: Effectiveness of conditional cycle-diffusion framework in the proposed method.

Competing Methods and Evaluation Metrics

We selected three traditional methods and five deep learning-based methods for comparison. The traditional methods included GSA (Aiazzi, Baronti, and Selva 2007), FUSE (Wei, Dobigeon, and Tourneret 2015) and CNMF (Yokoya, Yairi, and Iwasaki 2011). The Deep Learning-based methods include PSRT (Deng et al. 2023), LightNet (Chen et al. 2022), LAGC-NET (Jin et al. 2022), MoG-DCN (Dong et al. 2021) and SSR-NET (Zhang et al. 2021). Four widely used indexes are used for quantitative evaluation, including peak signal-to noise ratio (PSNR), spectral angle mapper (SAM), root mean squared error (RMSE), and erreur relative global adimensionnelle de synthese (ERGAS).

Quantitative and Qualitative Evaluation Table. 1 shows the quantitative assessment of different methods on the three datasets. Values in the table are averaged metrics across all test data. Compared with other competing methods, the proposed method can obtain superior performance on three datasets, achieving the best values across all quantitative evaluation metrics. Some representative visual results of different approaches are depicted in Figure. 4. To provide a clear demonstration of the superiority of the proposed method, we have included the residual result beneath each experimental result. In contrast, the proposed method can generate the HrHSI with exceptional detail fidelity. The residual images obtained by the proposed method exhibit the smallest deviation from the reference images. As a result, S^2CycleDiff can achieve outstanding results in both quantitative and qualitative aspects. To further validate the performance of the proposed method at higher upsampling rate, we conduct experiments on CAVE dataset with a upsampling factor of 16. Table. 2 shows that the proposed method still maintains the optimal performance.

Ablation Study

Effectiveness of Conditional Cycle-diffusion Framework To verify the effectiveness of the proposed conditional cycle-diffusion framework, we set up two variants, i.e., Variant-Spa that with the single spatial super-resolution branch and Variant-Spe that with the single spectral super-resolution branch, to achieve HISR, and conduct comparative experiments with the proposed method on three datasets. The experimental results are presented in Table. 3. In comparison to single branch, the proposed method demonstrates superior performance in generating HrHSI. This substantiates that the conditional cycle-diffusion framework can improve the spatial-spectral fidelity of the generated HrHSI.

Effectiveness of Detial Injection Block Detial Injection Block (DIB) aims to incorporate the spectral or spatial information of the guiding factors into the denoising process. We devise two variants, i.e., Variant1 (the module without guiding factors) and Variant2 (the module concatenates guiding factors directly with intermediate features along the channel dimension), for experiments on three datasets to verify the effectiveness of DIB. The experimental results are presented in Table. 4. It is evident that our approach, which leverages the cross-attention mechanism to integrate finer spectral or spatial information from guiding factors, can generate superior experimental outcomes.

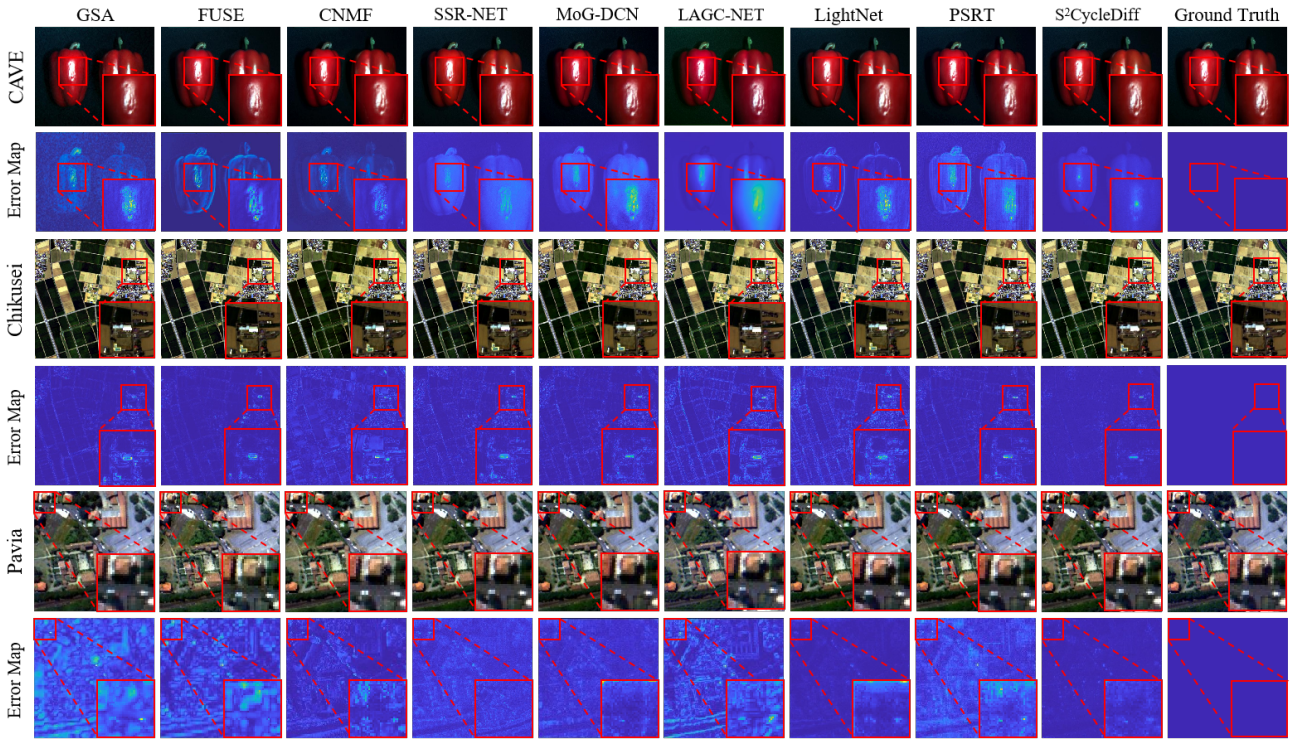

Figure 4: Visual results and reconstruction error maps obtained by different methods on CAVE, Chikusei and Pavia datasets. The darker error maps indicate better performance.

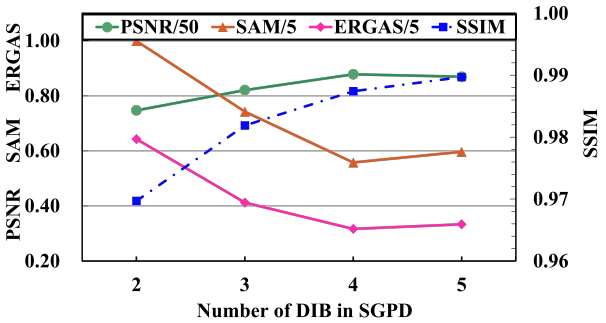

Figure 5: Experimental results of the proposed method with different number of DIBs.

Dataset	Model	PSNR	SAM	ERGAS	SSIM
CAVE	Variant1	36.2690	5.8561	3.4709	0.9652
	Variant2	43.4294	2.7107	1.6931	0.9897
	S^2CycleDiff	**43.9264**	**2.7907**	**1.5834**	**0.9898**
Chikusei	Variant1	29.1468	3.5522	5.4893	0.9351
	Variant2	42.9069	1.3852	1.3401	0.9904
	S^2CycleDiff	**43.8817**	**1.2756**	**1.3402**	**0.9909**
Pavia Center	Variant1	39.3987	3.7453	2.3496	0.9720
	Variant2	42.4301	2.9805	1.7412	0.9833
	S^2CycleDiff	**43.3698**	**2.7814**	**1.5903**	**0.9847**

Table 4: Effectiveness of DIB in the proposed method.

The Number of DIBs in SGPD The number of DIBs in SGPD directly affects the results of HISR. Therefore, we conducted a series of experiments on CAVE dataset to select the number of DIBs that can make the model perform optimally. The experimental results are normalized according to their corresponding manners, as shown in Figure. 5. It is demonstrate that the performance of the proposed method is optimal when the number of DIBs in SGPD is set to 4.

Conclusion

In this paper, we propose a spatial-spectral-bilateral cycle-diffusion framework (S^2CycleDiff) for hyperspectral super-resolution, which encodes the conditional distribution of spatial and spectral super-resolution processes into a conditional cycle-diffusion framework, so as to obtain the high-quality HrHSI by a sequence of refinements. The customized conditional cycle-diffusion framework can produce the HrHSIs with complementary spatial and spectral information through the bilateral super-resolution in spatial and spectral domains, in which the spatial/spectral guided pyramid denoising (SGPD) module separately adopts HrMSI and LrHSI as the guiding factors to guide the image restoration, effectively minimizing the spatial-spectral distortion by spatial details injection and spectral correction respectively. Extensive experiments are conducted to demonstrate the superior performance of the proposed framework compared to state-of-the-art methods.

Acknowledgments

This work was supported in part by the the National Natural Science Foundation of China under Grant 62101414 and Grant 62201423, Young Talent Fund of Xi'an Association for Science and Technology under Grant 095920221320 and Grant 959202313052, the China Postdoctoral Science Special Foundation under Grant 2022T150508 and 2023T160502, the Youth Innovation Team of Shaanxi Universities, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117, and the China Postdoctoral Science Foundation under Grant 2021M702546 and 2021M702548.

References

Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2006. MTF-tailored Multiscale Fusion of High-resolution MS and Pan Imagery. *Photogrammetric Engineering and Remote Sensing*, 72: 591–596.

Aiazzi, B.; Baronti, S.; and Selva, M. 2007. Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3230–3239.

Chen, Z.-X.; Jin, C.; Zhang, T.-J.; Wu, X.; and Deng, L.-J. 2022. SpanConv: A New Convolution via Spanning Kernel Space for Lightweight Pansharpening. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 841–847.

Choi, M. 2006. A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6): 1672–1682.

Deng, S.-Q.; Deng, L.-J.; Wu, X.; Ran, R.; Hong, D.; and Vivone, G. 2023. PSRT: Pyramid Shuffle-and-Reshuffle Transformer for Multispectral and Hyperspectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Dong, W.; Yang, Y.; Qu, J.; Xiao, S.; and Li, Y. 2023. Local Information-Enhanced Graph-Transformer for Hyperspectral Image Change Detection With Limited Training Samples. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Dong, W.; Zhou, C.; Wu, F.; Wu, J.; Shi, G.; and Li, X. 2021. Model-Guided Deep Hyperspectral Image Super-Resolution. *IEEE Transactions on Image Processing*, 30: 5754–5768.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hu, J.-F.; Huang, T.-Z.; Deng, L.-J.; Jiang, T.-X.; Vivone, G.; and Chanussot, J. 2022. Hyperspectral Image Super-Resolution via Deep Spatiospectral Attention Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7251–7265.

Huang, X.; Wen, D.; Li, J.; and Qin, R. 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote Sensing of Environment*, 196: 56–75.

Jin, Z.-R.; Zhang, T.-J.; Jiang, T.-X.; Vivone, G.; and Deng, L.-J. 2022. LAGConv: Local-context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening. *AAAI Conference on Artificial Intelligence (AAAI)*, 36(1): 1113–1121.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. arXiv:2201.11793.

Khan, M. J.; Khan, H. S.; Yousaf, A.; Khurshid, K.; and Abbas, A. 2018. Modern Trends in Hyperspectral Image Analysis: A Review. *IEEE Access*, 6: 14118–14129.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Liu, J. G. 2000. Smoothing Filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21: 3461 – 3472.

Liu, S.; Chi, M.; Zou, Y.; Samat, A.; Benediktsson, J. A.; and Plaza, A. 2017. Oil Spill Detection via Multitemporal Optical Remote Sensing Images: A Change Detection Perspective. *IEEE Geoscience and Remote Sensing Letters*, 14(3): 324–328.

Loncan, L.; de Almeida, L. B.; Bioucas-Dias, J. M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G. A.; Simões, M.; Tourneret, J.-Y.; Veganzones, M. A.; Vivone, G.; Wei, Q.; and Yokoya, N. 2015. Hyperspectral Pansharpening: A Review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3): 27–46.

Qu, J.; Li, Y.; and Dong, W. 2017. Hyperspectral Pansharpening With Guided Filter. *IEEE Geoscience and Remote Sensing Letters*, 14(11): 2152–2156.

Qu, J.; Zhao, J.; Dong, W.; Xiao, S.; Li, Y.; and Du, Q. 2023. Feature Mutual Representation Based Graph Domain Adaptive Network for Unsupervised Hyperspectral Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Scarpa, G.; Vitale, S.; and Cozzolino, D. 2018. Target-Adaptive CNN-Based Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9): 5443–5457.

Shettigara, V. K. 1992. A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set. *Photogrammetric Engineering and remote sensing*, 58(5): 561–567.

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.

Wald, L.; Ranchin, T.; and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *PE&RS*, 63(6): 691–699.

Wang, B.; Zhang, S.; Feng, Y.; Mei, S.; Jia, S.; and Du, Q. 2021. Hyperspectral Imagery Spatial Super-Resolution Using Generative Adversarial Network. *IEEE Transactions on Computational Imaging*, 7: 948–960.

Wang, P.; Yao, H.; Li, C.; Zhang, G.; and Leung, H. 2022a. Multiresolution Analysis Based on Dual-Scale Regression for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Wang, W.; Bao, J.; gang Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; and Li, H. 2022b. SinDiffusion: Learning a Diffusion Model from a Single Natural Image. *ArXiv*, abs/2211.12445.

Wang, Z.; Chen, B.; Lu, R.; Zhang, H.; Liu, H.; and Varshney, P. K. 2020. FusionNet: An Unsupervised Convolutional Variational Network for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Image Processing*, 29: 7565–7577.

Wei, Q.; Dobigeon, N.; and Tourneret, J.-Y. 2015. Fast Fusion of Multi-Band Images Based on Solving a Sylvester Equation. *IEEE Transactions on Image Processing*, 24(11): 4109–4121.

Xie, Q.; Zhou, M.; Zhao, Q.; Xu, Z.; and Meng, D. 2022. MHF-Net: An Interpretable Deep Network for Multispectral and Hyperspectral Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1457–1473.

Xu, S.; Amira, O.; Liu, J.; Zhang, C.-X.; Zhang, J.; and Li, G. 2020. HAM-MFN: Hyperspectral and Multispectral Image Multiscale Fusion Network With RAP Loss. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7): 4618–4628.

Yao, J.; Hong, D.; Chanussot, J.; Meng, D.; Zhu, X.; and Xu, Z. 2020. Cross-Attention in Coupled Unmixing Nets for Unsupervised Hyperspectral Super-Resolution. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 208–224. Cham: Springer International Publishing.

Yokoya, N.; Yairi, T.; and Iwasaki, A. 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528–537.

Zhang, X.; Huang, W.; Wang, Q.; and Li, X. 2021. SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 5953–5965.