HSR-Diff: Hyperspectral Image Super-Resolution via Conditional Diffusion Models

Chanyue Wu[1*], Dong Wang[1*], Yunpeng Bai[2*], Hanyu Mao[1], Ying Li[1†], Qiang Shen[2]
[1]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, School of Computer Science, Northwestern Polytechnical University, China
[2]Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, U.K.
{chanyuewu, dongwang, maomhy}@mail.nwpu.edu.cn lybyp@nwpu.edu.cn {yub3, qqs}@aber.ac.uk

Abstract

Despite the proven significance of hyperspectral images (HSIs) in performing various computer vision tasks, its potential is adversely affected by the low-resolution (LR) property in the spatial domain, resulting from multiple physical factors. Inspired by recent advancements in deep generative models, we propose an HSI Super-resolution (SR) approach with Conditional Diffusion Models (HSR-Diff) that merges a high-resolution (HR) multispectral image (MSI) with the corresponding LR-HSI. HSR-Diff generates an HR-HSI via repeated refinement, in which the HR-HSI is initialized with pure Gaussian noise and iteratively refined. At each iteration, the noise is removed with a Conditional Denoising Transformer (CDFormer) that is trained on denoising at different noise levels, conditioned on the hierarchical feature maps of HR-MSI and LR-HSI. In addition, a progressive learning strategy is employed to exploit the global information of full-resolution images. Systematic experiments have been conducted on four public datasets, demonstrating that HSR-Diff outperforms state-of-the-art methods.

1. Introduction

Hyperspectral images (HSIs) contain dozens or hundreds of spectral bands, enabling them to provide more faithful knowledge of targeted scenes than conventional imaging modalities. As such, HSIs play an irreplaceable role in various computer vision tasks, including classification [35, 42], segmentation [6], and tracking [32]. Although HSIs contain rich spectral information, contemporary hyperspectral imaging sensors lack high-resolution (HR) in the spatial domain, due to the stringent constraint of typically low signal-to-noise ratios. Their widespread use is significantly hindered by this fact. Restricted by hardware limitations, a practical way to work around this problem is to fuse the low-resolution (LR) HSI with an HR multispectral image (MSI). This requires the implementation of so-called HSI super-resolution (SR)[36], as shown in Figure 1.

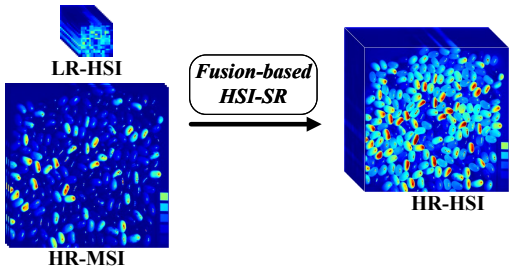


Figure 1: Illustration of fusion-based Hyperspectral Image Super-Resolution (HSI-SR).

Over the past few decades, a significant amount of research efforts have been devoted to developing HSI-SR approaches, which can be roughly classified into five categories [44]: Extensions of pansharpening [5, 25, 27], Bayesian inference-based methods [1, 29], matrix factorization-based methods [2], tensor-based methods [18], and deep learning (DL)-based methods [31]. Whilst pansharpening methods [26] have been extended to the field of HSI-SR, such approaches are prone to spectral distortion. Bayesian inference-based approaches rely on the assumption of prior knowledge, thereby having a weak flexibility in dealing with different HSI structures. Matrix factorization-based techniques reshape the 3D HSIs and MSIs into matrices, thus facing the challenge of learning the required relationship between space and spectrum. Although sev-

*Equal contribution
†Corresponding author. This work is supported by the National Natural Science Foundation of China under Grant (62271400), and in part by the Shaanxi Provincial Key R&D Program, China (2023-GHZD-02).

eral tensor-based methods have been proposed that can maintain the 3D structure of input images, they consume much more memory and computational power. Furthermore, these traditional approaches work via relying heavily on hand-crafted priors.

Recently, DL-based methods, especially convolutional neural network (CNN)-based approaches, have flooded over into the HSI-SR research community [8, 31, 10, 22, 46, 43, 19]. Rather than resorting to hand-crafted features, DL-based techniques learn prior knowledge automatically from given data. Particularly, Dong et al. proposed the first DL-based method for image SR, with the end-to-end mapping between LR images and HR images learned using a CNN [9]. Subsequently, generative adversarial networks (GANs) were introduced to the field of image SR in an effort to produce high-frequency details [17, 11]. After that, various GAN-based models have been devised, showing state-of-the-art results in the HSI-SR literature [24]. However, such work requires carefully designed regularization and optimization tricks to tame optimization instability and avoid mode collapse.

Inspired by the recent developments in deep generative models [14, 23], in this paper, we propose an innovative approach that we refer to as HSR-Diff (HSI-SR with conditional diffusion models). It works by learning to transform the original standard normal distribution into the data distribution of HR-HSI through a sequence of refinements. In contrast to GAN-based methods which require inner-loop maximization, HSR-Diff operates by simply minimizing a well-defined loss function to learn prior knowledge. Although conditional diffusion models are straightforward to define and efficient to train, there has been no demonstration that they are capable of merging LR-HSI and HR-MSI to the best of our knowledge. We show that conditional diffusion models are capable of generating high-quality HR-HSIs, which may best the state-of-the-art results.

A key factor of HSR-Diff is its inherent denoising ability thanks to use of deep neural networks. In spite of the effectiveness of CNNs for denoising, they have shown limitations in modelling long-range dependencies. Some previous transformer-based HSI-SR works, especially Pyramid Shuffle-and-Reshuffle Transformer (PSRT) [7], Transformer-based Fusion network (Fusformer) [15], and Hyperspectral and Multispectral image Fusion Transformer (HMF-Former) [40], all of which learn long-range dependencies with attention mechanisms. These Transformers only take LR-HSI and HR-MSI as input, while the denoising Transformer is also fed with noise level γ and noisy HR-HSI. To comprehensively exploit the spatial and spectral information in the LR-HSI, HR-MSI, and noisy HR-HSI, a Conditional Denoising Transformer (CDFormer) is herein designed and trained with a denoising objective to remove various levels of noise iteratively. In addition, a progressive learning strategy is utilized to help the CDFormer learn the global statistics of full-resolution HSIs. The main contributions of this work are summarized as follows:

- We propose the novel application of conditional diffusion models in the field of HSI-SR that works by progressively destroying HR-HSI through injecting noise and subsequently learning to reverse this process, in order to perform HSI-SR. To the best of our knowledge, this is the first work to exploit diffusion models within the realm of HSI-SR.
- We introduce a CDFormer that refines a noisy HR-HSI conditioned on the deep feature maps of HR-MSI and LR-HSI, capable of modelling global connectivity with self-attention. Rather than concatenation, cross-attention is utilized as the conditioning mechanism to incorporate spatio-spectral information and noise level.
- We employ a progressive learning strategy to exploit the global information of full-resolution HSIs, with CDFormer being trained on small image patches in the early epochs with high efficiency and on the global images in the later epochs to acquire global information.
- We present experimental investigations on four public datasets, with quantitative and qualitative results illustrating the superior performance of our approach as compared with state-of-the-art methods.

2. Related Work

2.1. Diffusion models

Typical deep generative models include autoregressive models (AR), normalizing flows (NF), variational autoencoders (VAE), GANs, diffusion models, etc. Over the past decade, GANs have gained prominence as a pivotal technology in the challenging field of image synthesis. However, this landscape has seen a transformation with the emergence of diffusion models, which have rapidly become a prominent research avenue across diverse domains, such as natural language processing, image processing, and computer vision. Diffusion models have arisen as an innovative approach that operates by iteratively refining a noise vector into the desired output. This reversible process allows for the generation of high-quality samples and has proven versatile in various applications.

The development of diffusion models has seen a dramatically accelerating pace over the past three years. In the field of computer vision, diffusion models have shown promise in tasks like image segmentation [33], object detection [30], and even video processing [13]. Whilst diffusion models

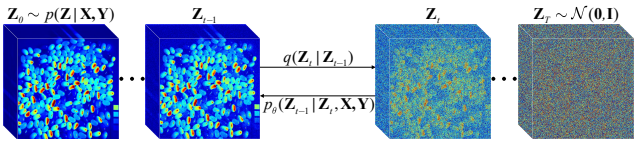


Figure 2: Forward and reverse processes of HSR-Diff, with forward process q generating an HSI sequence (left to right) by gradually adding Gaussian noise, and reverse process p iteratively refining HR-HSI (right to left).

have shown great potential for a variety of computer vision applications, none of them have yet been devoted to the problem of HSI-SR to the best of our knowledge. In this paper, we extend the utility of diffusion models to the field of HSI-SR.

2.2. Deep Learning-Based HSI-SR

In recent years, data-driven CNN architectures have been shown to outperform traditional approaches for use in the HSI-SR literature. These methods formulate the underlying fusion problem as a highly nonlinear mapping that takes HR-MSIs and LR-HSIs as input to generate an optimal HR-HSI. For example, CMHF-net [31] is an interpretable CNN, the design of which exploits the deep unfolding technique. Zhang et al. [43] proposed to reconstruct HR-HSIs with a two-stage network, while Zhang et al. [45] designed an interpretable spatial-spectral reconstruction network (SSR-NET) based on CNN. Aiming at problems of inflexible structure and information distortion, Jin et al. embedded Bilateral Activation Mechanism into ResNet, resulting in the effective model of BRResNet [16]. Thanks to the inductive bias of CNN, such as locality and weight sharing, these methods can provide good generalization performance and achieve impressive results. Nevertheless, CNNs have limitations in capturing long-range dependencies and self-similarity priors. To overcome such shortcomings, some Transformer-based HSI-SR approaches [15, 40] have been proposed.

Within this study, the CDFormer serves as the denoising network in conditional diffusion models. It harnesses self-attention to learn global statistics from full-resolution HSIs and MSIs and employs cross-attention as the conditioning mechanism.

3. Proposed Methodology

3.1. Problem Formulation

Without losing generality, the observation models for the HR-MSI and LR-HSI of interest can be mathematically formulated as

$$\begin{aligned} \mathbf{X} &= \mathbf{RZ} \\ \mathbf{Y} &= \mathbf{ZD}, \end{aligned} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{b\times HW}$ denotes the HR-MSI which consists of b spectral bands with a resolution of HW in the spatial domain; $\mathbf{R} \in \mathbb{R}^{b\times B}$ represents the spectral response function of HR-MSI; $\mathbf{Y} \in \mathbb{R}^{B\times hw}$ denotes the LR-HSI; and $\mathbf{Z} \in \mathbb{R}^{B\times HW}$ is the HR-HSI. In the above, b and B are the numbers of bands, with h and H being the band height, and w and W the width, where $b \ll B$, $h \ll H$, and $w \ll W$. $\mathbf{D} \in \mathbb{R}^{HW\times hw}$ is the spatial response of the LR-HSI, which can be modelled with blurring and down-sampling operations. The HSI-SR can be interpreted as an inverse problem for merging a practically collected $\mathbf{X}$ and an observed $\mathbf{Y}$ to produce a latent $\mathbf{Z}$. In this paper, the ideal $\mathbf{Z}$ is restored with HSR-Diff conditioned on spatio-spectral information of $\mathbf{X}$ and $\mathbf{Y}$, the details of which are described below.

3.2. HSI-SR with Conditional Diffusion Models

Given a dataset $\mathcal{D}_{train} = \{\mathbf{X}^i, \mathbf{Y}^i, \mathbf{Z}^i\}_{i=1}^N$ satisfying a certain joint probability distribution $p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, many pairs of $(\mathbf{X}, \mathbf{Y})$ may be consistent with the same $\mathbf{Z}$. Thus, the HR-HSI $\mathbf{Z}$ can be obtained with iterative refinement that provide an approximate to $p(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$. In this work, we implement the process of iterative refinement with HSR-Diff, where the optimized HR-HSI is presumed to be produced in T refinement steps. In HSR-Diff, the target HR-HSI is initialized with a pure noise $\mathbf{Z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as shown in Figure 2. The HSI is then refined iteratively according to learned conditional distributions $p_\theta(\mathbf{Z}_{t-1}|\mathbf{Z}_t, \mathbf{X}, \mathbf{Y})$. In so doing, the image sequence $(\mathbf{Z}_{T-1}, \mathbf{Z}_{T-2}, \ldots, \mathbf{Z}_0)$ can be attained and ultimately $\mathbf{Z}_0 \sim p(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$.

The HSR-Diff makes use of two processes: a forward process that perturbs HR-HSI $\mathbf{Z}_0$ to noise, and a reverse process converting noise back to HR-HSI $\mathbf{Z}_0$. In the forward process, the intermediate images, i.e., $\mathbf{Z}_{T-1}$, $\mathbf{Z}_{T-2}$, $\cdots$, and $\mathbf{Z}_1$, are generated according to a Markov chain with fixed transition probability $q(\mathbf{Z}_t|\mathbf{Z}_{t-1})$. We are interested in reversing the process via iterative refinement, in which the noise is reduced iteratively with a reverse Markov chain conditioned on $\mathbf{X}$ and $\mathbf{Y}$. The reverse chain is learned with the CDFormer f_θ. Further details of HSR-Diff's working are given below.

3.2.1 Forward Process

Inspired by [14], forward process q iteratively adds Gaussian noise to $\mathbf{Z}_0$ over T iterations:

$$\begin{aligned} q(\mathbf{Z}_{1:T} \mid \mathbf{Z}_0) &= \prod_{t=1}^{T} q(\mathbf{Z}_t \mid \mathbf{Z}_{t-1}) \\ q(\mathbf{Z}_t \mid \mathbf{Z}_{t-1}) &= \mathcal{N}(\mathbf{Z}_t; \sqrt{\alpha_t}\mathbf{Z}_{t-1}, (1-\alpha_t)\mathbf{I}), \end{aligned} \tag{2}$$

where $\alpha_{1:T} \in (0, 1)$ are scalar hyper-parameters. Note that in the forward process, the distribution of $\mathbf{Z}_t$ given $\mathbf{Z}_0$ can

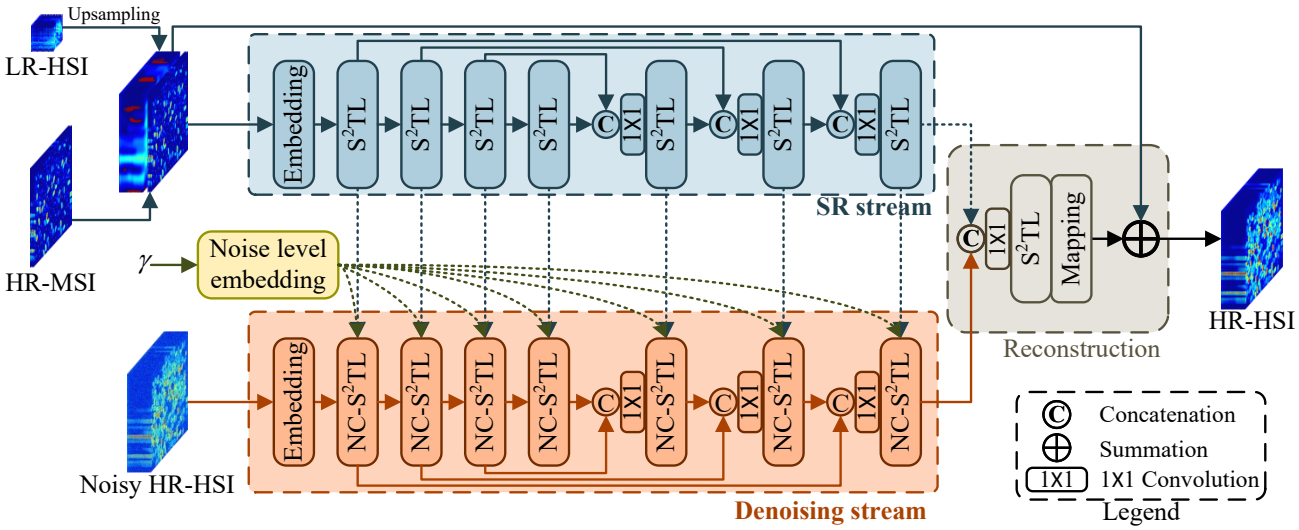


Figure 3: Architecture of Conditional Denoising Transformer.

be directly sampled in closed form. This implies that

$$q\left(\mathbf{Z}_t \mid \mathbf{Z}_0\right)=\mathcal{N}\left(\mathbf{Z}_t ; \sqrt{\gamma_t} \mathbf{Z}_0,\left(1-\gamma_t\right) \mathbf{I}\right), \tag{3}$$

where $\gamma_t=\prod_{i=1}^{t} \alpha_i$. In addition, the posterior distribution of $\mathbf{Z}_{t-1}$ given $\mathbf{Z}_0$ and $\mathbf{Z}_t$ can be derived by

$$\begin{aligned} q\left(\mathbf{Z}_{t-1} \mid \mathbf{Z}_0, \mathbf{Z}_t\right) &= \mathcal{N}\left(\mathbf{Z}_{t-1} ; \boldsymbol{\mu}, \sigma^2 \mathbf{I}\right) \\ \boldsymbol{\mu} &= \frac{\sqrt{\gamma_{t-1}}\left(1-\alpha_t\right)}{1-\gamma_t} \mathbf{Z}_0+\frac{\sqrt{\alpha_t}\left(1-\gamma_{t-1}\right)}{1-\gamma_t} \mathbf{Z}_t \\ \sigma^2 &= \frac{\left(1-\gamma_{t-1}\right)\left(1-\alpha_t\right)}{1-\gamma_t}. \end{aligned} \tag{4}$$

As the CDFormer is tasked with predicting $\mathbf{Z}_0$, the posterior is useful in the reverse process.

3.2.2 Reverse Markovian Process

The reverse process inferences $\mathbf{Z}_0$ via iterative refinement. It starts from a pure Gaussian noise $\mathbf{Z}_T$ and goes in the opposite direction of the forward process:

$$\begin{aligned} p_\theta\left(\mathbf{Z}_{0: T} \mid \mathbf{X}, \mathbf{Y}\right) &= p\left(\mathbf{Z}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{Z}_{t-1} \mid \mathbf{Z}_t, \mathbf{X}, \mathbf{Y}\right) \\ p\left(\mathbf{Z}_T\right) &= \mathcal{N}\left(\mathbf{Z}_T ; \mathbf{0}, \mathbf{I}\right) \\ p_\theta\left(\mathbf{Z}_{t-1} \mid \mathbf{Z}_t, \mathbf{X}, \mathbf{Y}\right) &= \mathcal{N}\left(\mathbf{Z}_{t-1} ; \mu_\theta\left(\mathbf{X}, \mathbf{Y}, \mathbf{Z}_t, \gamma_t\right), \sigma_t^2 \mathbf{I}\right), \end{aligned} \tag{5}$$

where the distribution $p_\theta\left(\mathbf{Z}_{t-1} \mid \mathbf{Z}_t, \mathbf{X}, \mathbf{Y}\right)$ is parameterized with θ. Note that the CDFormer provides a prediction of $\hat{\mathbf{Z}}_0$. Thus, according to (4), each refinement step takes the following form:

$$\begin{aligned} \mathbf{Z}_{t-1}=& \frac{\sqrt{\gamma_{t-1}}\left(1-\alpha_t\right)}{1-\gamma_t} f_\theta\left(\mathbf{X}, \mathbf{Y}, \mathbf{Z}_t, \gamma_t\right) \\ &+\frac{\sqrt{\alpha_t}\left(1-\gamma_{t-1}\right)}{1-\gamma_t} \mathbf{Z}_t+\sqrt{1-\alpha_t} \boldsymbol{\epsilon}, \end{aligned} \tag{6}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$ and f_θ is the CDFormer.

3.2.3 Noise Schedule

Inspired by the research reported in [4], we sample γ with two steps during training. In particular, we first sample a time step $t \sim U\{1, T\}$ and then randomly select $\gamma \sim U\left(\gamma_{t-1}, \gamma_t\right)$. As such, $\gamma \sim p(\gamma)=\sum_{t=1}^{T} \frac{1}{T} U\left(\gamma_{t-1}, \gamma_t\right)$. Normally, the model with a large T can achieve better results. However, we find (through empirical investigations) that the performance is not very sensitive to the exact values of T. Therefore, no hyper-parameter search about T is conducted and we set $T=2000$ for simplicity. As for the inference process, we set the maximum generation iterations to 100, employing a linear noise schedule.

3.3. Conditional Denoising Transformer

The property of non-local self-similarity of HSIs is often exploited in denoising tasks but is usually not well captured by CNN-based models. Due to the effectiveness of Transformer layer in capturing non-local long-range dependencies, the potential of Transformer is explored in conditional denoising of HSI. Unfortunately, the vanilla Transformer focuses only on spatial relationships between pixels while neglecting the spectral dimension. Besides, denoising networks in conditional diffusion models normally concatenate

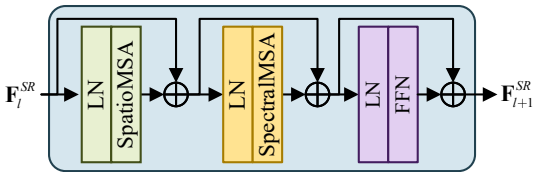


Figure 4: Spatio-Spectral Transformer Layer, where "LN" denotes layer normalization.

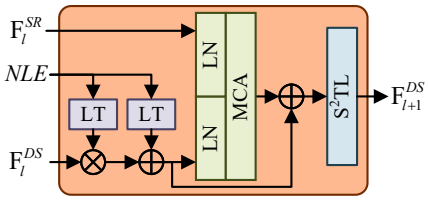


Figure 5: Noise-Aware Conditional Spatio-Spectral Transformer Layer, where "LT" represents linear transform, "LN" denotes layer normalization, and "MCA" is the abbreviation of multi-head cross-attention.

all images together as input, which may hinder the extraction of useful spatio-spectral information in LR-HSIs and HR-MSIs. Hence, the CDFormer adopts a two-stream architecture and is constructed with stacked Spatio-Spectral Transformer Layers (S^{2}TLs).

The architecture of the CDFormer is shown in Figure 3. The SR stream first utilizes a 3×3 convolution to generate low-level feature embeddings $\mathbf{F}_0^{SR}$ and then transforms it into deep features $\mathbf{F}_l^{SR}$ with stacked-S^{2}TLs. Instead of adopting t as done in the existing work [4], our method is conditioned on γ directly to achieve efficient generation. The denoising stream contains multiple noise-aware conditional S^{2}TLs (NC-S^{2}TLs) that take as input the embedded noise level and the image representation $\mathbf{F}^{SR}$. The Reconstruction module is set to produce a noise-free HR-HSI, by employing residual learning to alleviate the difficulty of HR-HSI generation while mapping the features onto HR-HSI via a 3×3 convolution and addition operation. It should be noted that no downsampling operation is employed.

3.3.1 Noise Level Embedding

The noise level offers essential information for denoising models. Inspired by the work of [4], we embed noise level within the models with sinusoidal positional encoding. The process of noise level embedding (NLE) can be formulated as follows:

$$\begin{aligned} NLE_{\gamma,2i} &= \sin\left(\gamma/10000^{2i/C}\right) \\ NLE_{\gamma,2i+1} &= \cos\left(\gamma/10000^{2i/C}\right), \end{aligned} \tag{7}$$

where C is the number of channels of S^{2}TLs; $i \in [1, C/2]$.

3.3.2 Spatio-Spectral Transformer Layers

Figure 4 illustrates the architecture of one S^{2}TL, which consists of a Spatial Multi-head Self-Attention (SpatioMSA), a Spectral Multi-head Self-Attention (SpectralMSA), and a Feed Forward Network (FFN). SpatioMSA and SpectralMSA learn the interactions of spatial regions and inter-spectra relationships, respectively. To alleviate the computational burden, we adopt the transposed attention [41] in SpectralMSA. SpatioMSA applies the popular window partitioning strategy [20] to reduce the computational complexity. In addition, the gating mechanism [41] is employed in the implementation of FFN.

3.3.3 Noise-Aware Conditional S^{2}TLs

To condition the overall model on the hierarchical features of SR stream, we feed $\mathbf{F}_l^{SR}$ to the Noise-Aware Conditional S^{2}TLs (NC-S^{2}TLs), each of which is a key building block of the denoising stream. Figure 5 depicts the structure of an NC-S^{2}TL, which takes as input the NLE (a vector), $\mathbf{F}_l^{SR}$ and $\mathbf{F}_l^{DS}$, where $\mathbf{F}_l^{SR}$ and $\mathbf{F}_l^{DS}$ have the same spatial resolution. NLE is first transformed and merged with $\mathbf{F}_l^{DS}$ with the result subsequently processed with the means of multi-head cross-attention (MCA) [3] in order to condition the model on $\mathbf{F}_l^{SR}$. As a result, each S^{2}TL learns the spatio-spectral dependencies.

3.4. Progressive Learning Strategy

Deep learning-based HSI-SR models are normally trained on small cropped patches. However, on one hand, training the proposed CDFormer on fixed-size image patches may not appropriately reflect the global relationships, which may result in suboptimal performance on full-resolution images when used. On the other hand, during the training phase, CDFormer can not be uploaded to Graphics Processing Unit (GPU) because of the limited memory. With this objective in mind, we adopt a strategy of progressive learning. In the initial epochs, the network is trained using smaller image patches. Subsequently, as training progresses, we transition towards utilizing larger patches, and ultimately, full-resolution images in the concluding training epochs. The resulting model, cultivated through the integration of varying sizes via progressive learning, exhibits notable enhancements in performance during testing. This is particularly evident when handling images of diverse resolutions, a scenario frequently encountered in the realm of the HSI-SR task.

To reduce the pressure on the demand of GPU memory, we only train the second half of CDFormer on full-

resolution images. The loss function used for such training is defined as follows:

$$\begin{aligned} \mathcal{L} &= \|\mathbf{X} - \mathbf{R}\hat{\mathbf{Z}}_0\|_1 + \|\mathbf{Y} - \hat{\mathbf{Z}}_0\mathbf{D}\|_1 + \|\mathbf{Z}_0 - \hat{\mathbf{Z}}_0\|_1 \\ \hat{\mathbf{Z}}_0 &= f_\theta(\sqrt{\gamma}\mathbf{Z}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \mathbf{X}, \mathbf{Y}), \end{aligned} \tag{8}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is sampled from the training set, and the noise schedule about γ has been discussed above. The first two terms are designed according to the observation models, while the last one is based on the assumption of Laplace distribution.

4. Experiments

Systematic experiments are herein conducted on four commonly-used public-available HSI-SR datasets to demonstrate the effectiveness of the proposed approach.

4.1. Datasets

Four datasets including CAVE [38], PaviaU [21], Chikusei [39], and HypSen [37] are used in our experiments, with the following details on each.

CAVE: There are 32 scenes with a spatial size of 512×512 in the CAVE dataset, where we select the first 20 HSIs for training, with the remaining 12 images used for testing. We generate LR-HSIs by Gaussian blur and down-sampling using a factor of 32 as done in [31]. HR-MSIs are acquired by integrating all HR-HSI bands according to the spectral response function of Nikon D700. The original HR-HSIs are treated as ground truth.

PaviaU: Collected by the University of Pavia, Italy, the original HSI dataset consists of 610×340 pixels in which the top-left 128×128 area is extracted as the test data, with the remaining used for training. Except for water absorption bands, all other 103 bands are chosen for the experiments. Note that the down-sampling factor for the generation of LR-HSIs is four, and the spectral response function is the same as that of the WorldView-3 satellite.

Chikusei: This dataset consists of 128 bands with a spectral range of 363 nm to 1018 nm and a spatial resolution of 2517×2335. The original HSI data was taken by an airborne visible and near-infrared imaging sensor over Chikusei, Japan. To alleviate the impact of the back boundary and noise, we crop the center area and remove noise bands. The processed image has a size of $2048 \times 2048 \times 110$. The top half $1024 \times 2048 \times 110$ area is selected as the training data, while the rest half is split into eight testing $512 \times 512 \times 110$ patches. For the production of LR-HSIs and HR-MSIs, this dataset adopts the same processing as with PaviaU.

HypSen: This dataset concerns a real scenario consisting of a 30m-resolution HSI and a 10m-resolution MSI. The Hyperion sensor on the Earth Observing-1 satellite provided the HSI with 242 spectral bands in the spectral range of 400 nm to 2500 nm, and the MSI with 13 bands was captured by the Sentinel-2A satellite. The blue, green, red, and near-infrared bands of MSI in our experiments are selected due to their high spatial resolution. To eliminate the impact of noise and water absorption, we remove those relevant bands, with 84 bands remaining in the HSI. We crop sub-images of size 250×330 and 750×990 from the Hyperion HSI and Sentinel-2A MSI respectively, in our study, with the pairs of sub-image patches spatially registered.

4.2. Methods Compared and Evaluation Metrics Used

Six state-of-the-art HSI-SR approaches are taken for comparison, including: UTV-TD [34], UAL [43], BRRes-Net [16], Fusformer [15], CMHF-Net [31], and UAL-DMI [28]. UTV-TD is a tensor-based technique; UAL, BRRes-Net, Fusformer, and CMHF-Net fall into the category of the DL-based methods; and UAL-DMI can be regarded as an upgraded version of UAL.

Four quantitative quality metrics are employed for performance evaluation, including Peak Signal-to-Noise Ratio (PSNR), Spectral Angle Mapper (SAM), Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS, namely error relative global dimensionless synthesis), and Structure Similarity (SSIM). The smaller ERGAS and SAM are, the larger PSNR and SSIM are, the better the fusion result is.

4.3. Implementation Specification

All DL-based methods are trained on the same datasets. For those compared methods, we use the publicly available source codes with default hyper-parameters as given in the corresponding research papers. Our HSR-Diff is implemented on the PyTorch framework. The learnable parameters of the CDFormer are initialized with Kaiming initialization [12] and trained on 2 NVIDIA GeForce GTX 3090s. The dimension of the embedding of F_0^{SR} and F_0^{DS} is set to 256, and the number of parameters of the CDFormer is 31.56M. We utilize the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize the CDFormer. With limited GPU memory, the batch size is set to 4 and 2 for 128^2 and 512^2 images, respectively. It costs 20000 epochs on the CAVE and PaviaU datasets while consuming 5000 epochs on the Chikusei dataset. The learning rate is initialized as 1×10^{-4}.

4.4. Comparisons with State-of-the-art Methods

In this set of experiments, the evaluations are carried out using the first three datasets listed above without involving the real-world dataset, HypSen (which will be dealt with in the next section).

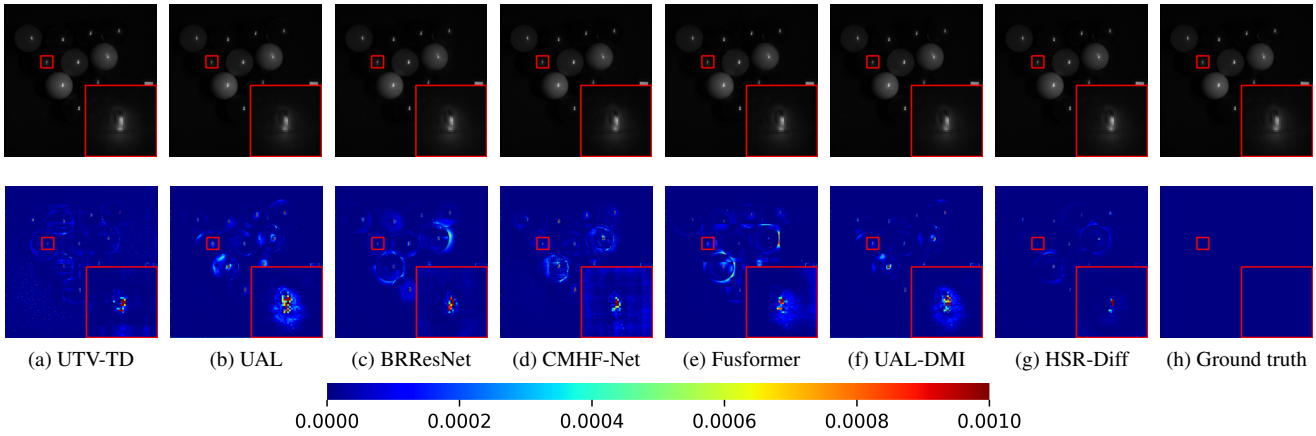


Figure 6: Visual quality comparison for fused HSIs of all competing methods on CAVE, where first and second rows show fourth band and corresponding heatmaps (mean squared error), respectively.

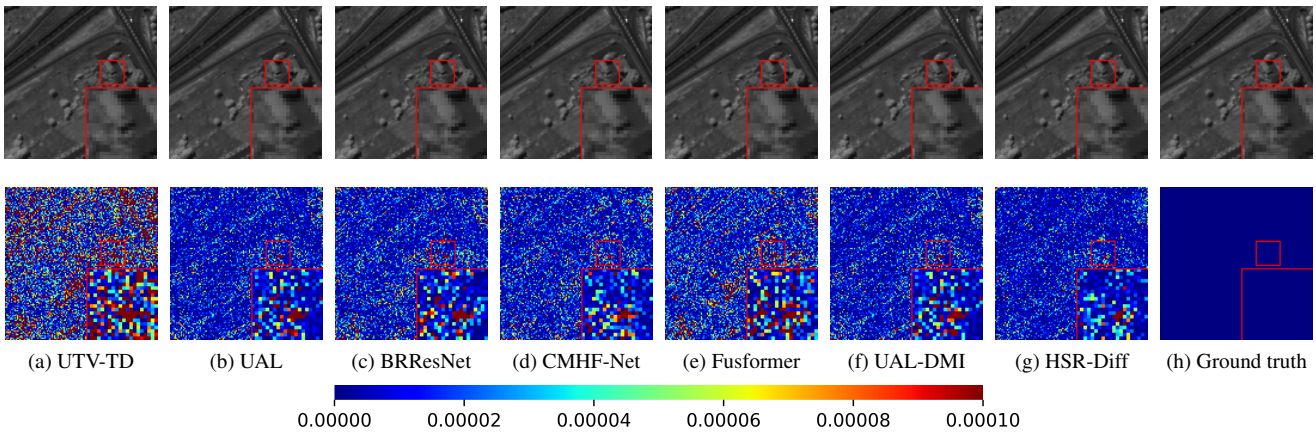


Figure 7: Visual quality comparison for fused HSIs of all competing methods on PaviaU, where first and second rows show 81st band and corresponding heatmaps (mean squared error), respectively.

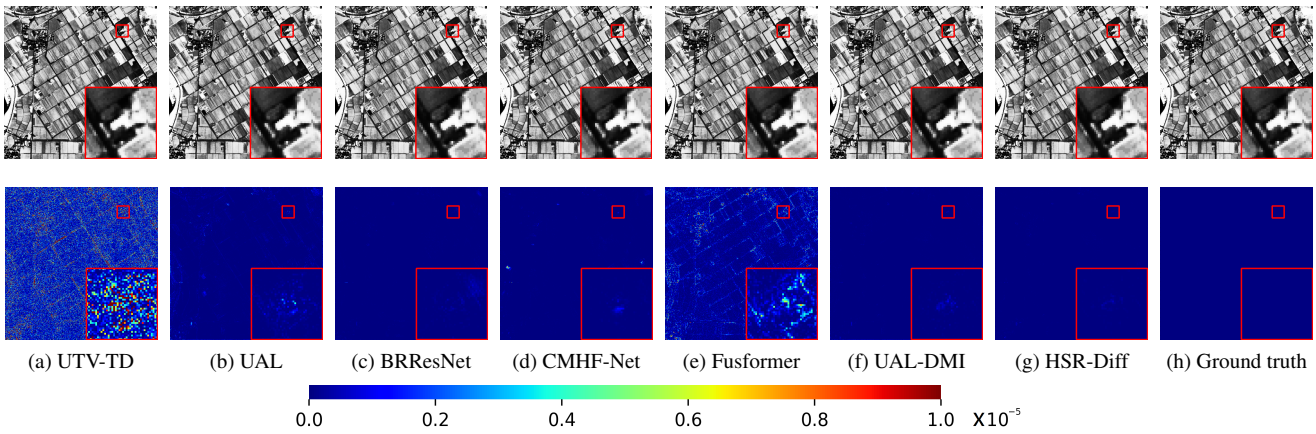


Figure 8: Visual quality comparison for fused HSIs of all competing methods on Chikusei, where first and second rows show 67th band and corresponding heatmaps (mean squared error), respectively.

Dataset	Methods	PSNR↑	SSIM↑	SAM↓	ERGAS↓
CAVE	UTV-TD [34]	38.66	0.9799	7.98	0.329
	UAL [43]	40.55	0.9933	4.33	0.271
	BRResNet [16]	41.36	0.9929	4.70	0.250
	CMHF-Net [31]	42.54	0.9939	4.69	0.216
	Fusformer [15]	41.52	0.9934	4.71	0.243
	UAL-DMI [28]	42.74	0.9950	3.79	0.213
	HSR-Diff	**44.33**	**0.9951**	**3.71**	**0.179**
PaviaU	UTV-TD [34]	44.46	0.9952	1.80	1.236
	UAL [43]	45.42	0.9964	1.54	1.148
	BRResNet [16]	45.53	0.9965	1.53	1.111
	CMHF-Net [31]	45.77	0.9965	1.50	1.096
	Fusformer [15]	45.66	0.9965	1.52	1.109
	UAL-DMI [28]	45.68	0.9966	1.49	1.113
	HSR-Diff	**46.47**	**0.9977**	**1.45**	**1.053**
Chikusei	UTV-TD [34]	48.38	0.9989	0.99	1.303
	UAL [43]	56.18	0.9998	0.49	0.421
	BRResNet [16]	56.79	0.9998	0.46	0.366
	CMHF-Net [31]	55.99	0.9998	0.50	0.483
	Fusformer [15]	55.92	0.9998	0.52	0.492
	UAL-DMI [28]	56.57	0.9998	0.47	0.387
	HSR-Diff	**57.34**	**0.9999**	**0.43**	**0.324**

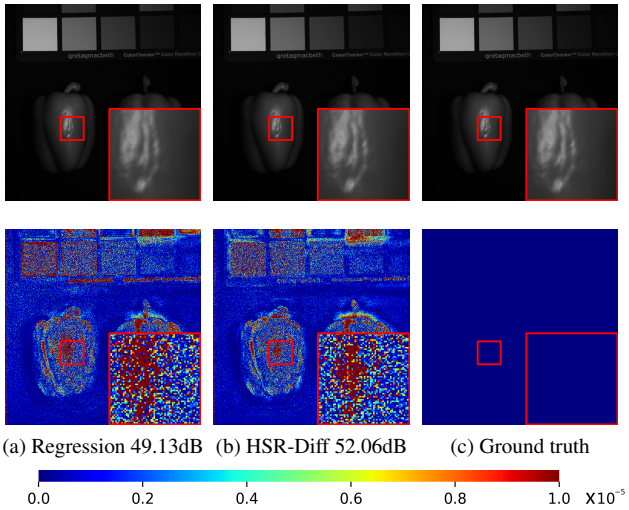
Table 1: Averaged PSNR, SSIM, SAM, and ERGAS of compared methods on CAVE, PaviaU, and Chikusei datasets. The ↑ or ↓ indicates higher or lower values corresponding to better results.

Qualitative Comparison. To assess the performance of HSR-Diff qualitatively, we visualize example bands of HSIs in Figures 6, 7, and 8. It can be seen from these visual results that all compared methods produce satisfactory outcomes. In particular, HSR-Diff generates gives the best result with minor errors since the corresponding MSE (mean squared error) images are much clearer than the others.

Quantitative Comparison. To further verify the superior performance of the proposed HSR-Diff, quantitative results are presented in Table 1. Note that the performance indices on the CAVE and Chikusei datasets are averaged over all testing samples (12 samples for CAVE and eight samples for Chikusei), respectively. It can be inferred from the results that the proposed HSR-Diff surpasses all competitors with a clear margin on all evaluation metrics.

4.5. Ablation Study

Effect of conditional diffusion models. Much of the early work on HSI-SR was based on the use of the regression model. To compare the effects of the diffusion and the regression model, we train the regression model containing the CDFormer. Note that the loss function, optimizer, and hyper-parameters are all the same as the conditional diffusion models. Figure 9 presents the fused results and corresponding error maps of utilising HSR-Diff and the regression model. As can be seen from the error maps, the HSIs produced by HSR-Diff have less distortion than those by the regression model. This is because HSR-Diff works with a series of iterative refinement steps, facilitating the capture of richer information on data distributions of HR-HSIs. Furthermore, Table 2 showcases the quantitative results and effectively highlights the significant contribution of iterative refinement.



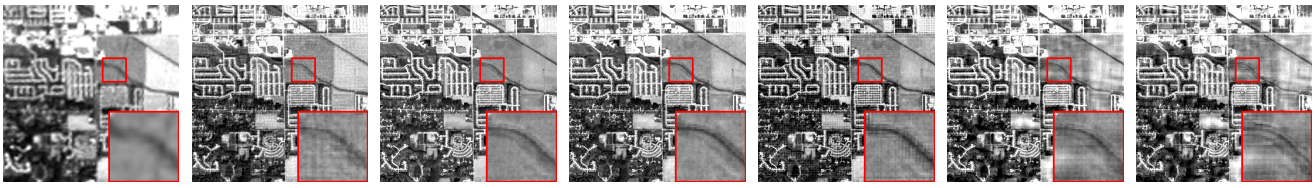
(a) Regression 49.13dB (b) HSR-Diff 52.06dB (c) Ground truth

Figure 9: Fusion results for HSR-Diff and Regression on the *real and fake peppers* image of CAVE. The PSNR values (dB) are given in the subtitles.

Dataset	Model	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
CAVE	Regression	43.26	0.9942	4.03	0.193
	CDFormer	**44.33**	**0.9951**	**3.71**	**0.179**
PaviaU	Regression	46.03	0.9968	1.52	1.178
	CDFormer	**46.47**	**0.9977**	**1.45**	**1.053**
Chikusei	Regression	56.71	**0.9999**	0.47	0.372
	CDFormer	**57.34**	**0.9999**	**0.43**	**0.324**

Table 2: Ablation study on conditional diffusion models. The ↑ or ↓ indicates higher or lower values corresponding to better results.

Effect of CDFormer. Recall that CDFormer is conditioned on the hierarchical representations of HR-MSI and LR-HSI via a two-stream architecture. However, alternative outstanding diffusion models [14, 23] that are also excellent for conditional image generation are equipped with CNN-based U-Nets, where degenerated images are concatenated with noisy high-resolution output images. To show the effectiveness of hierarchical representations, we remove the



(a) HSI (b) UTV-TD (c) UAL (d) BRResNet (e) CMHF-Net (f) UAL-DMI (g) HSR-Diff

Figure 10: Visual fusion results of all competing methods for HypSen.

Dataset	Network	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
CAVE	CD-CNN	38.84	0.9797	7.32	0.318
	C-w/o-SR	43.74	0.9942	3.94	0.188
	CDFormer	**44.33**	**0.9951**	**3.71**	**0.179**
PaviaU	CD-CNN	42.75	0.9962	1.75	1.362
	C-w/o-SR	46.08	0.9976	1.47	1.080
	CDFormer	**46.47**	**0.9977**	**1.45**	**1.053**
Chikusei	CD-CNN	47.63	0.9980	1.20	1.794
	C-w/o-SR	56.68	**0.9999**	0.46	0.425
	CDFormer	**57.34**	**0.9999**	**0.43**	**0.324**

Table 3: Ablation study on CDFormer. The ↑ or ↓ indicates higher or lower values corresponding to better results.

Dataset	Methods	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
CAVE	Fixed	43.17	0.9927	4.99	0.203
	Progressive	**44.33**	**0.9951**	**3.71**	**0.179**
PaviaU	Fixed	45.06	0.9970	1.63	1.173
	Progressive	**46.47**	**0.9977**	**1.45**	**1.053**
Chikusei	Fixed	55.92	**0.9999**	0.50	0.453
	Progressive	**57.34**	**0.9999**	**0.43**	**0.324**

Table 4: Ablation study on progressive learning. The ↑ or ↓ indicates higher or lower values corresponding to better results.

SR stream of CDFormer and name the resulting network "C-w/o-SR". In addition, we compare CDFormer with a CNN version of CDFormer that replaces all S^2TL with convolutional layers and show its results as "CD-CNN" in Table 3. The quantitative results show the use of CDFormer performs better than CD-CNN, demonstrating the effectiveness of global statistics. Indeed, with the two-stream architecture, CDFormer offers the best results thanks to the use of hierarchical features.

Effect of progressive learning. Progressive learning helps CDFormer to capture long-range dependencies of spatio-spectral information in HR-HSIs. To illustrate the effect of progressive learning, we train the CDFormer with fixed patches (128^2 for CAVE and Chikusei; 64^2 for PaviaU) with the results shown under the heading of "Fixed" in Table 4. As can be seen, progressive learning (from 128^2 to 512^2 for CAVE and Chikusei; from 64^2 to 128^2 for PaviaU) provides better results than training with fixed patches.

4.6. Generalization Analysis on Real Dataset

To examine the generalization ability of the implementations following the proposed approach, we test the performance of competitors on the real-world HypSen dataset [37]. Due to the lack of an ideal HR-HSI to train deep neural networks, we utilize the networks trained on the PaviaU dataset to merge observed LR-HSI and the corresponding HR-MSI. In addition, interpolation is applied to addressing the problem of an inconsistent number of bands between datasets. The fusion results of all compared methods are visualized in Figure 10, from which it can be seen that our method generates rich details, attaining satisfactory quality.

5. Conclusion

In this paper, we have presented the novel HSR-Diff approach that initializes an HR-HSI with pure Gaussian noise and then, iteratively refines it subject to the condition of the LR-HSIs and HR-MSIs of interest. At each step, the noise is removed with CDFormer which exploits the hierarchical representations of HR-MSIs and LR-HSIs rather than the original images. In addition, we employ a progressive learning strategy to maximize the use of the global information of full-resolution images, where CDFormer is trained on small patches in the early epochs with high efficiency while on the global images in the later epochs to obtain the global statistics. Systematic experimental investigations have been conducted, on four public datasets to validate the superior performance of the proposed approach, in comparison with state-of-the-art methods. However, diffusion models have low inference efficiency due to their sequential nature, computational complexity, time-consuming sampling process, and large model size. For future work, we will try to resolve the challenging issue of the relatively low image-generation efficiency of HSR-Diff.

References

[1] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Bayesian sparse representation for hyperspectral image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3631–3640, 2015. 1

[2] Ricardo Augusto Borsoi, Tales Imbiriba, and José Carlos Moreira Bermudez. Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability. *IEEE Transactions on Image Processing*, 29:116–127, 2019. 1

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 5

[4] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *Proceedings of the International Conference on Learning Represent*, 2021. 4, 5

[5] Zhao Chen, Hanye Pu, Bin Wang, and Geng-Ming Jiang. Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods. *IEEE Geoscience and Remote Sensing Letters*, 11(8):1418–1422, 2014. 1

[6] Phuong D Dao, Kiran Mantripragada, Yuhong He, and Faisal Z Qureshi. Improving hyperspectral image segmentation by applying inverse noise weighting and outlier removal for optimal scale selection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:348–366, 2021. 1

[7] Shang-Qi Deng, Liang-Jian Deng, Xiao Wu, Ran Ran, Danfeng Hong, and Gemine Vivone. Psrt: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[8] Renwei Dian, Shutao Li, Anjing Guo, and Leyuan Fang. Deep hyperspectral image sharpening. *IEEE transactions on neural networks and learning systems*, 29(11):5345–5355, 2018. 2

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[10] Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. Hyperspectral image super-resolution with optimized rgb guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11661–11670, 2019. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 8

[15] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Hong-Xia Dou, Danfeng Hong, and Gemine Vivone. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2, 3, 6, 8

[16] Zi-Rong Jin, Liang-Jian Deng, Tian-Jing Zhang, and Xiao-Xu Jin. Bam: Bilateral activation mechanism for image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4315–4323, 2021. 3, 6, 8

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[18] Shutao Li, Renwei Dian, Leyuan Fang, and José M Bioucas-Dias. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Transactions on Image Processing*, 27(8):4118–4130, 2018. 1

[19] Ying Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1):67, 2017. 2

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[21] Gamba Paolo. Pavia centre and university. https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes, 2011. 6

[22] Ying Qu, Hairong Qi, and Chiman Kwan. Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2511–2520, 2018. 2

[23] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 8

[24] Yue Shi, Liangxiu Han, Lianghao Han, Sheng Chang, Tongle Hu, and Darren Dancey. 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, 2022. 2

[25] Dong Wang, Yunpeng Bai, Bendu Bai, Chanyue Wu, and Ying Li. Heterogeneous two-stream network with hierarchical feature prefusion for multispectral pan-sharpening. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1845–1849. IEEE, 2021. 1

[26] Dong Wang, Yunpeng Bai, Chanyue Wu, Ying Li, Changjing Shang, and Qiang Shen. Convolutional lstm-based hierarchical feature fusion for multispectral pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 1

[27] Dong Wang, Pei Zhang, Yunpeng Bai, and Ying Li. Metapan: Unsupervised adaptation with meta-learning for multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[28] Xiuheng Wang, Jie Chen, and Cédric Richard. Hyperspectral image super-resolution with deep priors and degradation model inversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2814–2818. IEEE, 2022. 6, 8

[29] Qi Wei, José Bioucas-Dias, Nicolas Dobigeon, and Jean-Yves Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3658–3668, 2015. 1

[30] Zhenyu Wu, Lin Wang, Wei Wang, Tengfei Shi, Chenglizhao Chen, Aimin Hao, and Shuo Li. Synthetic data supervised salient object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5557–5565, 2022. 2

[31] Qi Xie, Minghao Zhou, Qian Zhao, Zongben Xu, and Deyu Meng. Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3, 6, 8

[32] Fengchao Xiong, Jun Zhou, and Yuntao Qian. Material based object tracking in hyperspectral videos. *IEEE Transactions on Image Processing*, 29:3719–3733, 2020. 1

[33] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[34] Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, Xi-Le Zhao, and Jie Huang. Hyperspectral image superresolution using unidirectional total variation with tucker decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4381–4398, 2020. 6, 8

[35] Xizhe Xue, Haokui Zhang, Bei Fang, Zongwen Bai, and Ying Li. Grafting transformer on automatically designed convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 1

[36] Jing Yang, Chanyue Wu, Tengfei You, Dong Wang, Ying Li, Changjing Shang, and Qiang Shen. Hierarchical spatio-spectral fusion for hyperspectral image super resolution via sparse representation and pre-trained deep model. *Knowledge-Based Systems*, 260:110170, 2023. 1

[37] Jingxiang Yang, Yong-Qiang Zhao, and Jonathan Cheung-Wai Chan. Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sensing*, 10(5):800, 2018. 6, 9

[38] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010. 6

[39] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5, 2016. 6

[40] Tengfei You, Chanyue Wu, Yunpeng Bai, Dong Wang, Huibin Ge, and Ying Li. Hmf-former: Spatio-spectral transformer for hyperspectral and multispectral image fusion. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2022. 2, 3

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 5

[42] Haokui Zhang, Chengrong Gong, Yunpeng Bai, Zongwen Bai, and Ying Li. 3-d-anas: 3-d asymmetric neural architecture search for fast hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2021. 1

[43] Lei Zhang, Jiangtao Nie, Wei Wei, Yanning Zhang, Shengcai Liao, and Ling Shao. Unsupervised adaptation learning for hyperspectral imagery super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2020. 2, 3, 6, 8

[44] Meilin Zhang, Xiongli Sun, Qiqi Zhu, and Guizhou Zheng. A survey of hyperspectral image super-resolution technology. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4476–4479. IEEE, 2021. 1

[45] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2020. 3

[46] Ke Zheng, Lianru Gao, Wenzhi Liao, Danfeng Hong, Bing Zhang, Ximin Cui, and Jocelyn Chanussot. Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2487–2502, 2020. 2