# Hyperspectral Pansharpening with Transformer-based Spectral Diffusion Priors

Hongcheng Jiang and ZhiQiang Chen
University of Missouri Kansas City
hjq44@mail.umkc.edu, chenzhiq@umkc.edu

## Abstract

*Hyperspectral pansharpening aims to fuse a spatially high-resolution panchromatic image (HR-PCI) with a low-resolution hyperspectral image (LR-HSI) to generate a high-resolution hyperspectral image (HR-HSI). Latest deep learning-based methods have achieved notable success in addressing this task, but they often rely on supervised or pre-trained models with high-quality labeled datasets, leading to limitations in practical applications. This paper introduces an unsupervised model that leverages transformer-based diffusion and spectral priors (uTDSP). In this model, spectral priors are learned from LR-HSI images, for which we assume that the stochastic distribution of spectral profiles in an LR-HSI is similar to that in the target HR-HSI. The learned prior is then used to optimize the fusion process by incorporating it as a regularization term, which is estimated simultaneously to adjust the contribution of the diffusion and the learned priors in reconstructing the target HR-HSI. Experimental results on benchmark datasets highlight the proposed method outperforms the state-of-the-art methods. Developed code will be available at this repository.*

## 1 Introduction

Hyperspectral image (HSI) is a data cube comprising spatially distributed spectral profiles, where each profile represents the reflectance or radiance values in a vector at a pixel within a specific wavelength range. The spectral information provided by HSIs offers insights into the physical properties of materials, making them invaluable for various remote sensing application tasks in the areas of mining, agriculture, environmental monitoring, and urban planning [28]. To deal with high spectral dimensionality in HSIs, specific image processing and understanding methods have been well studied at the pixel level, including those for denoising (e.g., [27]), unmixing (e.g., [6]), classification (e.g., [12]), and segmentation (e.g., [13]). To recognize or detection objects semantically in HSIs, it demands high spatial resolution as commonly required when conducting object-based understanding in color images. However, hyperspectral imaging systems often face limitations in capturing HSIs at the desired spatial resolution due to economic, technical, or physical constraints. These limitations typically result in hyperspectral images with a low spatial resolution. In contrast, low-cost imaging hardware, producing panchromatic images (PCIs) with high spatial resolution (HR-PCIs), can be readily integrated with a hyperspectral imaging system. Therefore, a promising solution is to reconstruct HR-HSIs with both high spatial and spectral resolution by fusing LR-HSIs and the corresponding HR-PCIs, a process known as hyperspectral pansharpening [26].

Existing hyperspectral pansharpening methods can be mainly categorized into two categories: model-based approaches [2, 5], which are mostly unsupervised estimation methods; and deep-learning-based approaches [9, 10], most of which are supervised and particularly rely on latest deep learning architectures. Model-based methods formulate the reconstruction process as an inverse problem and employ optimization methods. Under this category, four types of methods are found: [26]: (1) component substitution (CS), (2) multi-resolution analysis (MRA), (3) Bayesian estimation, and (4) variational. These methods are relatively computationally efficient but often introduce spectral distortions when reconstructing HR-HSIs from HR-PCIs [38].

In contrast, deep-learning-based methods, most of which are supervised, typically involve two key phases: reduced-resolution training and full-resolution pansharpening [15, 16]. During training, both LR-HSI and HR-PCI are degraded to generate resolution-reduced data, which are used to train a learning architecture, for example, a convolutional neural network (CNN). Subsequently, the trained CNN model can leverage scale-invariance principles to integrate incoming LR-HSIs and HR-PCIs, producing HR-HSIs [14]. While full-resolution training approaches are also feasible [3, 25], they remain a less common consideration. As reviewed in [33], the latest deep-learning-based methods have achieved notable success, outperforming statistical model-based methods. However, their dependence on the use of high-quality labeled datasets renders it a highly costly process involving human efforts, leading to signifi-

cant limitations in practical applications. On the contrary, unsupervised methods retain their benefits of being more operationally ready for practical applications. As indicated in [4], however, few unsupervised pansharpening methods with deep learning architectures are found successfully to this end.

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [18] have become prominent in image restoration tasks beyond image generation, due to their non-Markovian sampling processes that accelerate diffusion model sampling. Their versatility and efficiency are evident in applications such as super-resolution [34] and segmentation [11]. As a powerful unsupervised representation learner, DDPMs have inspired advancements in hyperspectral pansharpening. For instance, Rui et al. [32] introduced an unsupervised low-rank diffusion method for hyperspectral pansharpening, leveraging a low-rank subspace representation technique alongside a pre-trained unconditional diffusion model. Xing et al. [39] proposed CrossDiff, a cross-predictive diffusion model that utilizes DDPM's forward diffusion and reverse denoising processes for self-supervised pansharpening. However, these methods overlook the probabilistic distribution of spectral profiles (in short, spectral distribution), which is crucial for accurately preserving spectral fidelity across bands. In the recent work of [24], the authors employed spectral distribution knowledge, termed *spectral priors*, to implement a super-resolution process for HSI images.

In this work, we adopt the notion of spectral priors, and propose an unsupervised hyperspectral pansharpening framework utilizing transformer-based diffusion and spectral priors (uTDSP). We start from the premise that the stochastic distribution of spectral profiles in an LR-HSI is similar to that of the target HR-HSI. As such, a LR-HSI profile can be treated as a derived one through linear blurring and downsampling operations applied to the HR-HSI counterpart. Leveraging this premise, we construct the uTDSP model built on the DDPM architecture, transferring its learned spectral distribution into the fusion process. This transfer is achieved by preserving the inverse transition states of the uTDSP model, which introduces a regularization term within the maximum a posteriori (MAP) framework. To solve the resulting optimization problem, which involves multiple sub-problems corresponding to the timesteps, we treat the target HR-HSI as trainable parameters. Using the Adam optimization scheme, we iteratively update these parameters by following the spectral reverse generative sequence of the uTDSP model.

The main contributions of this paper are summarized as follows:

- We propose an unsupervised model that leverages transformer-based diffusion and spectral priors (uTDSP), designed to reconstruct HR-HSI from LR-HSI and HR-PCI. Specifically, the clean HR-HSIs are generated from Gaussian noise of the same dimensions as the HR-HSIs, following the reverse generative process defined by the uTDSP model.
- We propose a spectral diffusion prior for hyperspectral pansharpening, grounded in the observation that the stochastic distribution of spectral profiles in an LR-HSI closely aligns with that of the target HR-HSI. These learned spectral profiles are integrated as a regularization term within the maximum *a posteriori* (MAP) estimation framework.

As a final comment, our proposed method eliminates the generation of HR-HSI data as labels during learning, ensuring economic operation. Moreover, if the panchromatic pixel values are considered as target values during learning, the process can be categorized as weakly-supervised learning.

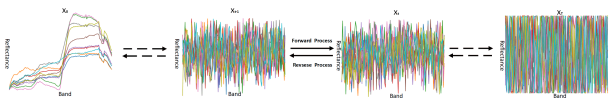


Figure 1. Illustration of forward spectral diffusion and reverse spectral generation in uTDSP model, featuring ten spectra.

## 2 Preliminaries

### 2.1 Diffusion Models

Diffusion models are probabilistic generative models designed to capture the dynamics of data evolution [18, 35]. They have demonstrated remarkable performance in various tasks, including audio and text generation [19, 21]. A diffusion model consists of two main processes: the *Forward Process* and the *Reverse Process*.

#### 2.1.1 Forward Process

The forward process progressively adds Gaussian noise to a clean sample $\boldsymbol{x}_0$ over $T$ timesteps according to a noise schedule $\{\beta_t\}_{t=1}^{T}$. At each timestep $t$, the noisy sample $\boldsymbol{x}_t$ is obtained as:

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

where $\beta_t$ is the noise variance and $\boldsymbol{I}$ is the identity matrix. Marginalizing over all timesteps gives:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. Alternatively, $\boldsymbol{x}_t$ can be directly expressed as:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{3}$$

#### 2.1.2 Reverse Process

The reverse process removes noise iteratively to recover the original data $\boldsymbol{x}_0$. It is modeled as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I}), \tag{4}$$

where $\mu_\theta$ and $\sigma_t^2$ are parameters, with $\mu_\theta$ being learned and $\sigma_t^2$ pre-defined. The predicted mean $\mu_\theta$ is computed as:

$$\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( \boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \right), \tag{5}$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is the neural network's prediction of the added noise $\boldsymbol{\epsilon}$.

#### 2.1.3 Loss Function and Sampling

The denoising network $\epsilon_\theta(\boldsymbol{x}_t, t)$ is trained to predict the noise $\boldsymbol{\epsilon}$ added during the forward process. The training objective is:

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}, t} \left[ \| \boldsymbol{\epsilon} - \epsilon_\theta \left( \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \|^2 \right], \tag{6}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $t$ is randomly sampled.

After training, the learned noise $\boldsymbol{\epsilon_\theta}$ is used to compute the mean $\mu_\theta(\boldsymbol{x}_t, t)$. The reverse sampling process generates data iteratively as:

$$\boldsymbol{x}_{t-1} = \mu_\theta(\boldsymbol{x}_t, t) + \sigma_t \boldsymbol{\epsilon}, \tag{7}$$

#### 2.1.4 Transformer-Based Diffusion Model

In this work, we develop the uTDSP model, illustrated in Fig. 1, as an extension of the DDPM. By integrating transformer architectures, the uTDSP model captures long-range dependencies, significantly enhancing its performance. The structure of the uTDSP model, outlined in Tab. 1, comprises an MLP block and $N$ stages transformer block, followed by a linear layer. Each stage generates $N_B$-dimensional feature representations. Like other diffusion models, uTDSP model is conditioned on the timestep $t$. The timestep is encoded using a sinusoidal embedding $\phi$ of dimension $N_\phi$, which is further processed through an MLP block to generate features with the same spectral dimension, $N_{\text{out}}$, as the HSI, thereby refining the time-dependent features.

| Component | Input Dim. | Output Dim. |
|---|---|---|
| **Time embedding $\phi$** | 1 | $N_\phi$ |
| **MLP Block** | | |
| Linear | $N_\phi$ | $N_{\text{out}}$ |
| Activation | $N_{\text{out}}$ | $N_{\text{out}}$ |
| Linear | $N_{\text{out}}$ | $N_B$ |
| **N Stages Transformer Block** | | |
| Linear | $N_B$ | $N_B$ |
| Activation | $N_B$ | $N_B$ |
| Linear | $N_B$ | $N_B$ |
| Multihead Self-Attention | $N_B$ | $N_B$ |
| Fully Connected Layer | $N_B$ | $N_B$ |
| Norm | $N_B$ | $N_B$ |
| Dropout | $N_B$ | $N_B$ |
| **MLP Layer** | | |
| Linear | $N_B$ | $N_{\text{out}}$ |

Table 1. uTDSP Model Architecture Details: Input and Output Dimensions.

### 2.2 Hyperspectral Pansharpening

Hyperspectral pansharpening is an image reconstruction task that aims to generate a high-resolution HSI (HR-HSI) $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{H} \times \boldsymbol{W} \times \boldsymbol{S}}$ from an observed low-resolution HSI (LR-HSI) $\boldsymbol{Y} \in \mathbb{R}^{\boldsymbol{h} \times \boldsymbol{w} \times \boldsymbol{S}}$ and a high-resolution panchromatic image (HR-PCI) $\boldsymbol{C} \in \mathbb{R}^{\boldsymbol{H} \times \boldsymbol{W} \times 1}$. The LR-HSI $\boldsymbol{Y}$ has a reduced spatial resolution compared to the spatially mapped panchromatic image $\boldsymbol{C}$, though not pixel-by-pixel spatially registered. To realize the pansharpening process, the preceding assumptions are: (1) given a spatial location, the panchromatic value in a HR-PCI provides an univariate spatial mean statistic that summarizes the spectral profile across the measuring bandwidth at the location in the HR-HSI; and (2) at the spatial location, the hyperspectral value in the LR-HSI provides a vectored spectral mean statistic that summarizes the spectral profiles at the location.

From a sampling perspective, the LR-HSI can be treated as a derived image (or tensor) from the HR-HSI through a spatial downsampling process. This process primarily degrades the spatial details of the HR-HSI while preserving its spectral content, potentially including a blurring process due to sampling noises. Conceptually, we can express this process as

$$\boldsymbol{Y} = \boldsymbol{D}(\boldsymbol{B}(\boldsymbol{X})) \tag{8}$$

where $\boldsymbol{D}$ and $\boldsymbol{B}$ represent the spatial downsampling and blurring operations, respectively. The relationship between the HR-HSI and the HR-PCI is commonly expressed through a linear operation, given by:

$$\boldsymbol{C} = \boldsymbol{X} \times_3 \boldsymbol{R} \tag{9}$$

where $\boldsymbol{R} \in \mathbb{R}^{1 \times \boldsymbol{S}}$ is the spectral response vector as an operator, and $\boldsymbol{X} \times_3 \boldsymbol{R}$ denotes mode-3 tensor multiplication between $\boldsymbol{X}$ and $\boldsymbol{R}$. The pansharpening process is to learn a mapping that reversely estimates these operations in Eqs. 8 and 9, hence reconstructing $\boldsymbol{X}$ from $\boldsymbol{Y}$ and $\boldsymbol{C}$.

## 3 Proposed Method

The goal of hyperspectral pansharpening is to generate the HR-HSI $\boldsymbol{X}$ by fusing the LR-HSI $\boldsymbol{Y}$ and HR-PCI $\boldsymbol{C}$. This problem can be formulated in a maximum a posteriori (MAP) framework as:

$$\max_{\boldsymbol{X}} p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{C}). \tag{10}$$

Using Bayes' theorem, we can rewrite the posterior $p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{C})$ as:

$$p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{C}) \propto p(\boldsymbol{Y}, \boldsymbol{C}|\boldsymbol{X})p(\boldsymbol{X}), \tag{11}$$

leading to the equivalent optimization problem:

$$\min_{\boldsymbol{X}} -\log p(\boldsymbol{Y}, \boldsymbol{C}|\boldsymbol{X}) - \log p(\boldsymbol{X}), \tag{12}$$

where $-\log p(\boldsymbol{Y}, \boldsymbol{C}|\boldsymbol{X})$ represents the data fidelity term, and $-\log p(\boldsymbol{X})$ serves as a regularization term to promote desired properties in $\boldsymbol{X}$.

The data fidelity term can be expressed as:

$$\mathcal{L}_X(\boldsymbol{X}) = \lambda\|\boldsymbol{Y} - \boldsymbol{D}(\boldsymbol{B}(\boldsymbol{X})\|_F^2 + \|\boldsymbol{C} - \boldsymbol{R}\boldsymbol{X}\|_F^2, \tag{13}$$

where $\lambda > 0$ is a balance parameter, and $\|\cdot\|_F$ denotes the Frobenius norm.

For the regularization term, we propose a spectral diffusion prior. We assume the spectral vectors $\boldsymbol{x} \in \mathbb{R}^N$ in $\boldsymbol{X}$ are independently distributed, i.e.,

$$-\log p(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} \log p(\boldsymbol{x}). \tag{14}$$

To model $p(\boldsymbol{x})$, we use a spectral diffusion process, where the spectral vector $\boldsymbol{x} = \boldsymbol{x}_0$ follows a given distribution $q(\boldsymbol{x}_0)$, with the joint distribution of its states in a Markov chain given by:

$$\log q(\boldsymbol{x}_{0:T}) = \log q(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t). \tag{15}$$

The transition $\log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is approximated using the uTDSP model $\epsilon_\theta(\boldsymbol{x}_t, t)$. The optimization problem is thus reformulated as:

$$\min_{\boldsymbol{X}} \mathcal{L}_X(\boldsymbol{X}) + \gamma \sum_{t=1}^{T} \|\boldsymbol{x}_t - \epsilon_\theta(\boldsymbol{x}_t, t)\|_F^2, \tag{16}$$

where $\gamma$ controls the regularization strength.

The final objective consists of an ordinary fidelity term and a spectral diffusion regularization term. We solve the optimization by sequentially updating $\boldsymbol{X}$ from $t = T$ to $t = 1$, using the Adam optimizer and performing gradient updates for each subproblem. The complete process is summarized in Algorithm 1.

**Algorithm 1** uTDSP Method

**Input:** Gaussian HR-HSI $\boldsymbol{X}$, LR-HSI $\boldsymbol{Y}$, HR-PCI $\boldsymbol{C}$, Blurring Operator $\boldsymbol{B}$, Spectral Response Operator $\boldsymbol{R}$, Spatial Downsampling Operator $\boldsymbol{D}$, Regularization Parameters $\lambda$ and $\gamma$, Learning Rate $\mu$.
Train the loss function defined in Equation (6) of the uTDSP model $\epsilon_\theta(\boldsymbol{x}_t, t)$ by randomly sampling spectra from $\boldsymbol{Y}$.
Initialize parameters $\boldsymbol{X}$ and fix network parameters $\theta$.
**for** time step $t = T$ **to** $1$ **do**
  **for** iteration $k = 1$ **to** $K$ **do**
    Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ in $\mathbb{R}^{H\times W\times S}$
    Compute $\boldsymbol{X}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{X} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
    Evaluate $\epsilon_\theta \in \mathbb{R}^{H\times W\times S}$ using $\boldsymbol{X}_t$ and $t$
    Update $\boldsymbol{X} \leftarrow \boldsymbol{X} - \mu\nabla_{\boldsymbol{X}}\left[L_{\boldsymbol{X}}(\boldsymbol{X}) + \gamma\sum_{\boldsymbol{t}}\|\boldsymbol{x}_t - \epsilon_\theta(\boldsymbol{x}_t, t)\|_F^2\right]$
  **end for**
**end for**
**Output:** Enhanced HR-HSI $\boldsymbol{X}$.

## 4 Experiment

### 4.1 Datasets and Evaluation Metrics

We evaluate the hyperspectral pansharpening performance on four publicly available datasets: the Botswana dataset (size: $1476 \times 256 \times 145$) [36], the Chikusei dataset (size: $2517 \times 2335 \times 128$) [41], the Pavia Center (PaviaC) dataset (size: $1096 \times 715 \times 102$) [29], and the Pavia University (PaviaU) dataset (size: $610 \times 340 \times 102$) [7]. From each dataset, we crop the central region to obtain HR-HSI of size $256 \times 256$. The LR-HSI and HR-PCI are generated according to Wald's protocol [31, 37]. Specifically, the LR-HSI is created by applying a Gaussian filter of size $4\times4$ to spatially blur the HR-HSI, followed by downsampling the result by a factor of 8. The HR-PCI is generated by averaging the visible bands of the HR-HSI.

The restoration results are evaluated using three quantitative indices to assess the quality and accuracy of the reconstructed HSI. The Peak Signal-to-Noise Ratio (PSNR) is used to quantify perceptual image quality, providing a measure of the fidelity of the reconstructed image relative to the original. The Spectral Angle Mapper (SAM) evaluates spectral similarity by calculating the angular difference between the original and reconstructed spectra, offering a measure of spectral consistency. The Error Relative Global Dimension Synthesis (ERGAS) assesses reconstruction quality by computing the normalized average error across all spectral bands, providing a holistic evaluation of reconstruction performance.

To complement these quantitative evaluations, visual results are generated by creating visible images from the re-

| Parameter | Value |
|---|---|
| Batch size | 1024 |
| Dropout rate | 0.001 |
| Epochs | 30000 |
| Optimizer | Adam |
| Time embedding dimension $N_\phi$ | 64 |
| $\beta$ scheduler | Linear |
| $\beta_1$ | 0.0001 |
| $\beta_T$ | 0.02 |
| Regularization Parameters $\lambda$ | 0.1 |
| Regularization Parameters $\gamma$ | 0.001 |
| Learning rate $\mu$ | 0.01 |
| Learning rate scheduler | $0.001 \times \max(1000 - \text{epoch}/10, 1)$ |

Table 2. Implementation Details of the uTDSP Model

constructed HR-HSI of each method and comparing them to the ground truth. For the visualization, three bands are randomly selected from each dataset to generate the visible images. Specifically, bands 20, 50, and 80 are chosen for the Botswana dataset; bands 15, 30, and 56 for the Chikusei dataset; bands 10, 40, and 100 for the PaviaC dataset; and bands 40, 65, and 80 for the PaviaU dataset. This selection allows for a representative visual comparison across different datasets and highlights the ability of each method to accurately reconstruct the spectral and spatial details.

### 4.2 Implementation Details

The uTDSP model is implemented in PyTorch to ensure efficient and scalable performance. Table 2 outlines key implementation details—covering the model architecture, hyperparameters, and training setups—offering a comprehensive overview of the model's structure and configurations. For further information on Blind Point Spread Function (PSF) and Spectral Response Function (SRF) estimation, refer to [24, 32].

### 4.3 Results for Hyperspectral Pansharpening

This comprehensive set of experiments evaluates the performance of the proposed uTDSP method by conducting comparisons against eleven methods, including DBDENet [30], DDLPS [22], DHP-DARN [43], DIP-HyperKite [8], DMLD-Net [42], GPPNN [40], GSA [1], HyperPNN [17], Indusion [20], PLRDiff [32], and SFIM [23]. Both quantitative and visual evaluation results are provided in Table 3 and Fig. 2, illustrating the comparative performance of these methods on the testing datasets. The results highlight that the proposed uTDSP method consistently delivers superior performance, outperforming all competing methods across multiple evaluation metrics. This demonstrates its robustness and effectiveness in addressing the challenges of spectral and spatial reconstruction, making it a reliable choice for hyperspectral pansharpening task.

| Dataset | Method | PSNR | SAM | ERGAS |
|---|---|---|---|---|
| Botswana | DBDENet [30] | 22.84 | 8.5207 | 11.3979 |
| | DHP-DARN [43] | 28.85 | 4.9084 | 2.8164 |
| | DIP-HyperKite [8] | 30.24 | 4.8305 | 2.1305 |
| | DMLD-Net [42] | 26.87 | 6.5379 | 3.7552 |
| | GPPNN [40] | 26.44 | 8.6439 | 3.8965 |
| | HyperPNN [17] | 29.83 | 4.9803 | 2.2254 |
| | DDLPS ∗ [22] | 22.27 | 6.9539 | 17.5198 |
| | GSA ∗ [1] | 23.80 | 6.2035 | 11.6626 |
| | Indusion ∗ [20] | 15.30 | 5.4225 | 9.7633 |
| | PLRDiff ∗ [32] | 17.84 | 15.1475 | 9.0164 |
| | SFIM ∗ [23] | 26.81 | 5.4225 | 2.7995 |
| | uTDSP ∗ | **31.61** | **3.7777** | **1.9155** |
| Chikusei | DBDENet [30] | 25.02 | 6.5243 | 4.2316 |
| | DHP-DARN [43] | 25.24 | 6.0044 | 3.9208 |
| | DIP-HyperKite [8] | 25.63 | 5.4180 | 3.7059 |
| | DMLD-Net [42] | 25.28 | 6.9856 | 4.1170 |
| | GPPNN [40] | 25.17 | 6.5704 | 4.1423 |
| | HyperPNN [17] | 25.34 | 5.7174 | 3.8096 |
| | DDLPS ∗ [22] | 26.85 | 5.3557 | 3.7616 |
| | GSA ∗ [1] | 24.21 | 6.2670 | 5.3903 |
| | Indusion ∗ [20] | 22.29 | 5.4171 | 5.0367 |
| | PLRDiff ∗ [32] | 26.18 | 6.3831 | 3.5699 |
| | SFIM ∗ [23] | 25.54 | **5.4171** | 4.0868 |
| | uTDSP ∗ | **26.86** | 5.9152 | **3.3178** |
| PaviaC | DBDENet [30] | 23.63 | 18.5981 | 6.7944 |
| | DHP-DARN [43] | 26.70 | 12.4018 | 4.7917 |
| | DIP-HyperKite [8] | 26.48 | 12.5846 | 4.9198 |
| | DMLD-Net [42] | 26.03 | 16.8061 | 5.1832 |
| | GPPNN [40] | 27.37 | 11.2643 | 4.4916 |
| | HyperPNN [17] | 26.10 | 17.5919 | 5.1762 |
| | DDLPS ∗ [22] | 27.52 | **10.1478** | 4.3781 |
| | GSA ∗ [1] | 25.30 | 10.4678 | 5.6518 |
| | Indusion ∗ [20] | 25.84 | 10.4645 | 5.8683 |
| | PLRDiff ∗ [32] | 27.39 | 11.6904 | 4.6245 |
| | SFIM ∗ [23] | 24.99 | 10.4488 | 5.9011 |
| | uTDSP ∗ | **28.53** | 10.4176 | **4.2664** |
| PaviaU | DBDENet [30] | 28.84 | 6.5032 | 2.7593 |
| | DHP-DARN [43] | 29.27 | 6.8826 | 2.5350 |
| | DIP-HyperKite [8] | 29.30 | 6.0972 | 2.5114 |
| | DMLD-Net [42] | 28.66 | 6.8624 | 2.7985 |
| | GPPNN [40] | 29.86 | **6.0788** | 2.4812 |
| | HyperPNN [17] | 28.96 | 6.7555 | 2.6472 |
| | DDLPS ∗ [22] | 27.81 | 7.1405 | 4.3781 |
| | GSA ∗ [1] | 26.47 | 7.2522 | 2.9323 |
| | Indusion ∗ [20] | 25.82 | 7.8229 | 4.1539 |
| | PLRDiff ∗ [32] | 28.57 | 7.5217 | 2.9453 |
| | SFIM ∗ [23] | 25.66 | 7.8229 | 3.8125 |
| | uTDSP ∗ | **30.68** | 6.8019 | **2.4660** |

∗ denotes an unsupervised method

Table 3. Comparison of methods on various datasets.

- **Botswana Dataset:** In the Botswana dataset, the proposed uTDSP achieves the best overall performance, with the highest PSNR (31.61 dB) and the lowest SAM (3.7777) and ERGAS (1.9155), demonstrating its exceptional ability to reconstruct high-quality spectral and spatial features. Visually, uTDSP pro-

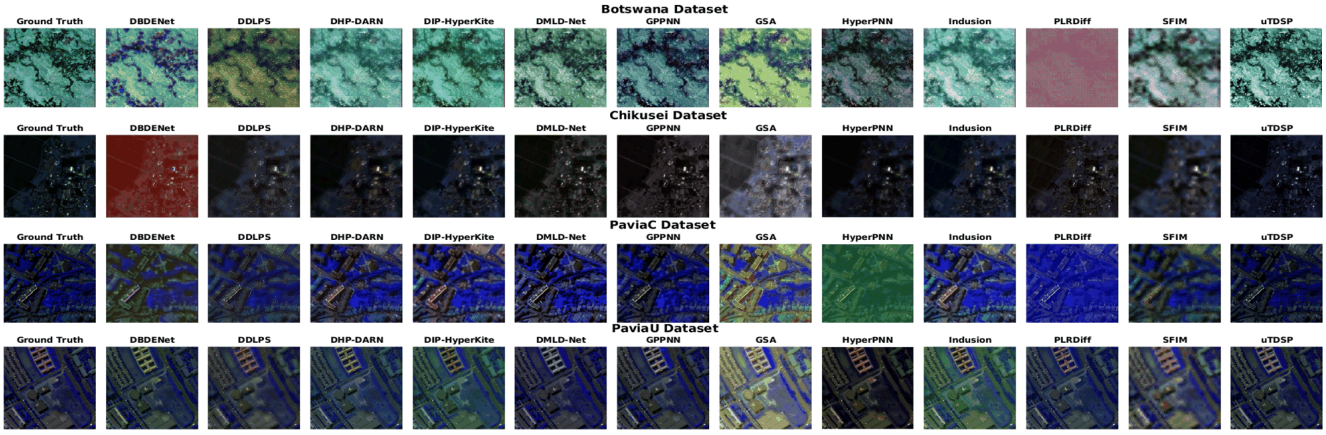


Figure 2. Visual results across various datasets.

| Diffusion Steps | 800 | | | 1300 | | | 1800 | | | 2300 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS |
| Botswana | 31.08 | 4.1781 | 2.0822 | **31.61** | **3.7777** | **1.9155** | 31.29 | 3.8198 | 1.9620 | 30.76 | 3.9183 | 2.0484 |
| Chikusei | 24.90 | 7.6737 | 4.0223 | 26.27 | 6.2801 | 3.5414 | **26.86** | **5.9152** | **3.3178** | 26.06 | 6.3808 | 3.6651 |
| PaviaC | **28.53** | **10.4176** | **4.2664** | 28.05 | 12.0213 | 4.6686 | 27.30 | 14.4649 | 5.2019 | 27.05 | 15.7526 | 5.4399 |
| PaviaU | 29.94 | 7.1776 | 2.6621 | **30.68** | **6.8019** | **2.4660** | 29.05 | 6.9999 | 2.8546 | 29.21 | 7.5859 | 2.8467 |

Table 4. Comparison of different diffusion steps $T$ across various datasets.

duces sharp, artifact-free results that closely resemble the ground truth. Among the competing methods, DIP-HyperKite and HyperPNN achieve strong performances with PSNRs of 30.24 dB and 29.83 dB, respectively, but their outputs lack the fine structural clarity of uTDSP. DHP-DARN delivers a solid PSNR of 28.85 dB but struggles to preserve spectral details as effectively. Traditional methods like PLRDiff and Indusion perform poorly, with PSNRs of 17.84 dB and 15.30 dB, respectively, and produce blurred, distorted reconstructions. GPPNN and DMLD-Net yield moderate results, with PSNRs of 26.44 dB and 26.87 dB, but their outputs suffer from spectral noise and oversmoothing. SFIM, while achieving a relatively high PSNR of 26.81 dB, fails to maintain detailed textures. DBDENet and DDLPS produce subpar results, with DDLPS displaying significant spectral distortions due to its simplistic assumptions.

- **Chikusei Dataset:** In the Chikusei dataset, the proposed uTDSP achieves the highest PSNR (26.86 dB) and the lowest ERGAS (3.3178), demonstrating its superior ability to preserve spectral and spatial information, although its SAM (5.9152) is slightly higher than some competitors. Visually, uTDSP generates the most accurate reconstruction, closely resembling the ground truth with sharp and detailed outputs. Among other methods, DDLPS also performs well, achieving a PSNR of 26.85 dB and the lowest SAM (5.3557), but its results lack the precision and clarity of uTDSP. DIP-HyperKite and HyperPNN show competitive performances, with PSNRs of 25.63 dB and 25.34 dB, respectively, but exhibit less spatial sharpness. PLRDiff achieves a strong ERGAS (3.5699), but its visual outputs remain blurred. Methods such as DBDENet, DHP-DARN, and DMLD-Net deliver moderate results, with PSNRs between 25.02 dB and 25.28 dB, yet fail to capture fine details, while GPPNN, SFIM, and Indusion exhibit higher distortions and noise. Traditional methods like GSA and Indusion perform poorly, with GSA achieving the lowest PSNR (24.21 dB) and Indusion generating oversmoothed and artifact-prone outputs.

- **Pavia Center Dataset:** In the PaviaC dataset, the proposed uTDSP achieves the best overall performance, with the highest PSNR (28.53 dB), the lowest ERGAS (4.2664), and a competitive SAM (10.4176), demonstrating its superior ability to reconstruct high-quality spectral-spatial details. Visually, uTDSP closely resembles the ground truth, offering sharp and clear structures without distortions or artifacts. DDLPS follows closely with a PSNR of 27.52 dB and the lowest SAM (10.1478), though its reconstructions exhibit less spatial clarity compared to uTDSP. GPPNN and PLRDiff also perform well, with PSNRs of 27.37 dB and 27.39 dB, respectively, but both fail to preserve finer details. HyperPNN and DMLD-Net yield mod-

| Stages | 3 | | | 4 | | | 5 | | | 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS |
| Botswana | 31.39 | 3.8501 | 1.9664 | 31.57 | **3.7382** | 1.9479 | **31.61** | 3.7777 | **1.9155** | 31.08 | 3.9791 | 2.0565 |
| Chikusei | 26.42 | 6.1475 | 3.5083 | **26.86** | **5.9152** | **3.3178** | 26.67 | 6.1444 | 3.4277 | 26.20 | 6.2178 | 3.5943 |
| PaviaC | 28.14 | 10.6808 | 4.4047 | 27.98 | 10.5842 | 4.4408 | **28.53** | **10.4176** | **4.2664** | 28.26 | 10.6045 | 4.3668 |
| PaviaU | 29.96 | 6.8580 | 2.6363 | 30.00 | 6.9284 | 2.6345 | **30.68** | **6.8019** | **2.4660** | 30.17 | 7.0138 | 2.6002 |

Table 5. Comparison of different $N$ Stages transformer block across various datasets.

| Batch Sizes | 256 | | | 512 | | | 1024 | | | 2048 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS |
| Botswana | 30.93 | 4.5930 | 2.0791 | 31.51 | 3.7788 | 1.9712 | **31.61** | 3.7777 | **1.9155** | 31.43 | **3.7232** | 1.9759 |
| Chikusei | 26.32 | 5.9681 | 3.4547 | 26.72 | 6.0297 | 3.4166 | **26.86** | **5.9152** | **3.3178** | 26.73 | 5.9884 | 3.3524 |
| PaviaC | 28.33 | 10.5321 | 4.3340 | 28.18 | 10.6451 | 4.3953 | **28.53** | **10.4176** | **4.2664** | 28.39 | 10.7603 | 4.3064 |
| PaviaU | 30.03 | 6.8901 | 2.6213 | 30.55 | 6.8938 | 2.5116 | **30.68** | **6.8019** | **2.4660** | 29.78 | 7.0839 | 2.6969 |

Table 6. Comparison of different batch sizes across various datasets.

erate results, with PSNRs around 26 dB, but their outputs are affected by blurring and spatial inconsistencies. Methods such as DIP-HyperKite, DHP-DARN, and Indusion demonstrate similar PSNRs in the range of 25–26 dB but suffer from higher SAM and ERGAS values, indicating less robust spectral reconstruction. Traditional methods like GSA and SFIM perform poorly, with lower PSNRs and visible spectral distortions. Overall, uTDSP establishes itself as the most effective method, consistently outperforming all competitors in both quantitative metrics and visual quality for the PaviaC dataset.

- **Pavia University Dataset:** In the PaviaU dataset, the proposed uTDSP achieves the best overall performance, with the highest PSNR (30.68 dB) and the lowest ERGAS (2.4660), while maintaining a competitive SAM (6.8019). Visually, uTDSP closely resembles the ground truth, producing clear and detailed reconstructions with minimal artifacts. GPPNN follows as a strong competitor with a PSNR of 29.86 dB and the lowest SAM (6.0788), though it falls slightly short in terms of ERGAS and visual sharpness compared to uTDSP. DIP-HyperKite and DHP-DARN deliver competitive results, achieving PSNRs of 29.30 dB and 29.27 dB, respectively, but their outputs exhibit less clarity and precision. HyperPNN and DMLD-Net yield moderate performance, with PSNRs of 28.96 dB and 28.66 dB, but their outputs suffer from oversmoothing and spectral inconsistencies. Traditional methods like GSA, Indusion, and SFIM perform poorly, with significantly lower PSNR values and noticeable spectral distortions. DDLPS, despite achieving a PSNR of 27.81 dB, produces blurred and less accurate outputs due to its limitations in reconstruction.

### 4.4 Ablation Study

#### 4.4.1 Analysis of Diffusion Step Variations

The total number of diffusion steps, denoted as $T$, significantly influences the reverse spectral generation in the uTDSP model. Table 4 presents the PSNR, SAM, and ERGAS metrics of the proposed method under different total diffusion steps $T$, ranging from 800 to 2300 in increments of 500, across various datasets. It can be observed that for all datasets, as $T$ increases, the PSNR initially rises but then tends to saturate or decrease. Similarly, SAM and ERGAS values first decrease and then tend to either saturate or increase. This behavior suggests that there is an optimal range of diffusion steps for achieving the best performance, which varies depending on the dataset. Specifically, the uTDSP model achieves the best results on the Botswana and PaviaU datasets when the diffusion steps are set to 1300. In contrast, for the PaviaC dataset, the best performance is observed at 800 diffusion steps, while the Chikusei dataset shows optimal results at 1800 diffusion steps. These observations emphasize the importance of selecting dataset-specific diffusion steps to maximize the effectiveness of the uTDSP for spectral reconstruction.

#### 4.4.2 Analysis of Different Stages

As shown in Fig. 1, the structure of the uTDSP model can be controlled by adjusting the number of $N$ Stages transformer blocks, which significantly influences the complexity of the uTDSP model. Table 5 presents the PSNR, SAM, and ERGAS metrics of the proposed method under different values of $N$, ranging from 3 to 4 in increments of 1, across various datasets. It can be observed that for all datasets, as $N$ increases, the PSNR initially rises but then tends to saturate or decrease. Similarly, SAM and ERGAS values first decrease and then tend to either saturate or increase. Par-

| $N_B$ | 256 | | | 512 | | | 1024 | | | 2048 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS |
| Botswana | 30.53 | 4.9518 | 2.1681 | **31.61** | 3.7777 | **1.9155** | 31.52 | **3.6886** | 1.9345 | 31.40 | 3.7104 | 1.9494 |
| Chikusei | 26.38 | 6.0499 | 3.4846 | **26.86** | **5.9152** | **3.3178** | 26.41 | 6.0497 | 3.5018 | 25.77 | 6.5076 | 3.7732 |
| PaviaC | 28.50 | 11.0266 | 4.4170 | **28.53** | 10.4176 | **4.2664** | 28.07 | **10.3042** | 4.3718 | 28.35 | 10.5473 | 4.3079 |
| PaviaU | 30.01 | 7.3412 | 2.6898 | **30.68** | **6.8019** | **2.4660** | 29.86 | 7.1344 | 2.6913 | 29.86 | 7.1187 | 2.6687 |

Table 7. Comparison of different $N_B$ values of transformer blocks across multiple datasets.

| $N_\phi$ | 16 | | | 32 | | | 64 | | | 128 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS | PSNR | SAM | ERGAS |
| Botswana | 31.29 | 3.9995 | 1.9732 | 31.57 | **3.7767** | 1.9368 | **31.61** | 3.7777 | **1.9155** | 31.42 | 3.8903 | 1.9585 |
| Chikusei | 26.65 | 6.0661 | 3.4145 | 26.47 | 6.0296 | 3.4689 | **26.86** | **5.9152** | **3.3178** | 26.52 | 6.0488 | 3.4747 |
| PaviaC | 28.24 | 10.4577 | 4.3417 | 28.10 | 10.6125 | 4.3940 | **28.53** | **10.4176** | **4.2664** | 28.16 | 10.5399 | 4.3724 |
| PaviaU | 30.03 | 6.9324 | 2.6131 | 29.87 | 6.9149 | 2.6427 | **30.68** | **6.8019** | **2.4660** | 29.88 | 6.8828 | 2.6436 |

Table 8. Comparison of different time embedding dimension $N_\phi$ across multiple datasets.

ticularly, the uTDSP model on all datasets does not achieve good results with $N$ equal to 3 or 6, indicating that both lower and higher complexities lead to suboptimal performance. For the Botswana, PaviaC, and PaviaU datasets, the uTDSP model achieves the best results when $N$ equals 5. However, for the Chikusei dataset, the best performance is observed when $N$ equals 4. These findings highlight the importance of selecting an optimal number of stages to balance model complexity and performance.

#### 4.4.3 Analysis of Batch Size Effects

The batch size function determines how many samples are processed together in one forward process and reverse process through the uTDSP model. The batch size significantly influences the reverse spectral generation in the uTDSP model. Table 6 presents the PSNR, SAM, and ERGAS metrics of the proposed method under different batch sizes, ranging from 256 to 2048 in increments of doubling, across various datasets. The uTDSP model achieves the best results when the batch size equals 1024.

#### 4.4.4 Analysis of $N_B$ Values in Transformer Blocks

As shown in Fig. 1, the structure of the uTDSP model can be controlled by adjusting the number of $N_B$ values in the transformer blocks, which significantly influences the complexity of the uTDSP model. Table 7 presents the PSNR, SAM, and ERGAS metrics of the proposed method under different values of $N_B$, ranging from 256 to 2048 in increments of doubling, across various datasets. It can be observed that for all datasets, as $N_B$ increases, the PSNR initially rises but then tends to saturate or decrease. Similarly, SAM and ERGAS values first decrease and then tend to either saturate or increase. The uTDSP model achieves the best results when $N_B$ equals 512 among all datasets.

#### 4.4.5 Analysis of time embedding dimension $N_\phi$

The sinusoidal timestep embedding plays a vital role in diffusion models by effectively representing temporal information during training and sampling. As shown in Fig. 1, the structure of the uTDSP model mirrors other diffusion models and is conditioned on the timestep $t$ through the sinusoidal timestep embedding $N_\phi$. Table 8 presents the PSNR, SAM, and ERGAS metrics for the proposed method under varying $N_\phi$ values, ranging from 16 to 128 with incremental doublings, across multiple datasets. The results reveal that PSNR initially improves with increasing $N_\phi$ but eventually saturates or declines. Similarly, SAM and ERGAS values decrease initially and later stabilize or rise. Among all datasets, the uTDSP model achieves optimal performance when $N_\phi = 64$.

## 5 Conclusion

This paper introduces the unsupervised transformer-based diffusion and spectral priors (uTDSP) model for hyperspectral pansharpening, which leverages transformer-based diffusion and spectral priors to enhance performance. By learning spectral priors from low-resolution hyperspectral images (LR-HSI) and incorporating them into the fusion process, uTDSP adaptively balances diffusion and prior information to reconstruct high-resolution hyperspectral images (HR-HSI). Experimental results demonstrate that uTDSP outperforms state-of-the-art methods.

## 6 Acknowledgement

The first author appreciates the support from the University of Missouri-Kansas City SGS/SSE. The second author thanks to the support from the University of Alabama in Huntsville under a contract with the National Aeronautics and Space Administration (80NSSC22K0014).

## References


[1] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007. 5

[2] Hussein A Aly and Gaurav Sharma. A regularized model-based optimization framework for pan-sharpening. *IEEE Transactions on Image Processing*, 23(6):2596–2608, 2014. 1

[3] Matteo Ciotola, Giuseppe Guarino, Antonio Mazza, Giovanni Poggi, and Giuseppe Scarpa. Pansharpening by efficient and fast unsupervised target-adaptive cnn. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5579–5582. IEEE, 2023. 1

[4] Matteo Ciotola, Giuseppe Guarino, Gemine Vivone, Giovanni Poggi, Jocelyn Chanussot, Antonio Plaza, and Giuseppe Scarpa. Hyperspectral pansharpening: Critical review, tools, and future perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 2024. 2

[5] Wenqian Dong, Song Xiao, Xiao Xue, and Jiahui Qu. An improved hyperspectral pansharpening algorithm based on optimized injection model. *IEEE Access*, 7:16718–16729, 2019. 1

[6] Yuexin Duan, Xia Xu, Tao Li, Bin Pan, and Zhenwei Shi. Undat: Double-aware transformer for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 1

[7] Mathieu Fauvel, Jón Atli Benediktsson, Jocelyn Chanussot, and Johannes R Sveinsson. Spectral and spatial classification of hyperspectral data using svms and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804–3814, 2008. 4

[8] Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M Patel. Hyperspectral pansharpening based on improved deep image prior and residual reconstruction. *arXiv e-prints*, pages arXiv–2107, 2021. 5

[9] Giuseppe Guarino, Matteo Ciotola, Giovanni Poggi, Gemine Vivone, and Giuseppe Scarpa. Hybrid gsa-cnn method for hyperspectral pansharpening. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 901–904. IEEE, 2024. 1

[10] Giuseppe Guarino, Matteo Ciotola, Gemine Vivone, Giovanni Poggi, and Giuseppe Scarpa. Pca-cnn hybrid approach for hyperspectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 2023. 1

[11] Xutao Guo, Yanwu Yang, Chenfei Ye, Shang Lu, Bo Peng, Hua Huang, Yang Xiang, and Ting Ma. Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 2

[12] Renlong Hang, Xuwei Qian, and Qingshan Liu. Cross-modality contrastive learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 1

[13] Renlong Hang, Ping Yang, Feng Zhou, and Qingshan Liu. Multiscale progressive segmentation network for high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 1

[14] Lin He, Dahan Xi, Jun Li, Honghao Lai, Antonio Plaza, and Jocelyn Chanussot. Dynamic hyperspectral pansharpening cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023. 1

[15] Lin He, Jinhua Xie, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Jiawei Zhu. Variable subpixel convolution based arbitrary-resolution hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 1

[16] Lin He, Jiawei Zhu, Jun Li, Deyu Meng, Jocelyn Chanussot, and Antonio Plaza. Spectral-fidelity convolutional neural networks for hyperspectral pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5898–5914, 2020. 1

[17] Lin He, Jiawei Zhu, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Bo Li. Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):3092–3100, 2019. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[19] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022. 2

[20] Muhammad Murtaza Khan, Jocelyn Chanussot, Laurent Condat, and Annick Montanvert. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geoscience and Remote Sensing Letters*, 5(1):98–102, 2008. 5

[21] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[22] Kaiyan Li, Weiying Xie, Qian Du, and Yunsong Li. Ddlps: Detail-based deep laplacian pansharpening for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):8011–8025, 2019. 5

[23] JG Liu. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of remote sensing*, 21(18):3461–3472, 2000. 5

[24] Jianjun Liu, Zebin Wu, and Liang Xiao. A spectral diffusion prior for unsupervised hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 5

[25] Qingjie Liu, Huanyu Zhou, Qizhi Xu, Xiangyu Liu, and Yunhong Wang. Psgan: A generative adversarial network for remote sensing image pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10227–10242, 2020. 1

[26] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon,

Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015. 1

[27] Zhaozhi Luo, Xinyu Wang, Petri Pellikka, Janne Heiskanen, and Yanfei Zhong. Unsupervised adaptation learning for real multiplatform hyperspectral image denoising. *IEEE transactions on cybernetics*, 2024. 1

[28] Dimitris G Manolakis, Ronald B Lockwood, and Thomas W Cooley. *Hyperspectral imaging remote sensing: physics, sensors, and algorithms*. Cambridge University Press, 2016. 1

[29] Antonio Plaza, Jon Atli Benediktsson, Joseph W Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, et al. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113:S110–S122, 2009. 4

[30] Jiahui Qu, Shaoxiong Hou, Wenqian Dong, Song Xiao, Qian Du, and Yunsong Li. A dual-branch detail extraction network for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 5

[31] Thierry Ranchin and Lucien Wald. Fusion of high spatial and spectral resolution images: The arsis concept and its implementation. *Photogrammetric engineering and remote sensing*, 66(1):49–61, 2000. 4

[32] Xiangyu Rui, Xiangyong Cao, Li Pang, Zeyu Zhu, Zongsheng Yue, and Deyu Meng. Unsupervised hyperspectral pansharpening via low-rank diffusion model. *Information Fusion*, 107:102325, 2024. 2, 5

[33] Ataollah Shirzadi, Himan Shahabi, Kamran Chapi, Dieu Tien Bui, Binh Thai Pham, Kaka Shahedi, and Baharin Bin Ahmad. A comparative study between popular statistical and machine learning methods for simulating volume of landslides. *Catena*, 157:213–226, 2017. 1

[34] Tao Song, Ran Wen, and Lei Zhang. Roughset-ddpm: An image super-resolution method based on rough set denoising diffusion probability model. *Tehnički vjesnik*, 31(1):162–170, 2024. 2

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[36] Stephen G Ungar, Jay S Pearlman, Jeffrey A Mendenhall, and Dennis Reuter. Overview of the earth observing one (eo-1) mission. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1149–1159, 2003. 4

[37] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997. 4

[38] Xiuheng Wang, Jie Chen, Qi Wei, and Cédric Richard. Hyperspectral image super-resolution via deep prior regularization with parameter estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1708–1723, 2021. 1

[39] Yinghui Xing, Litao Qu, Shizhou Zhang, Kai Zhang, Yanning Zhang, and Lorenzo Bruzzone. Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model. *IEEE Transactions on Image Processing*, 2024. 2

[40] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021. 5

[41] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5(5):5, 2016. 4

[42] Kai Zhang, Guishuo Yang, Feng Zhang, Wenbo Wan, Man Zhou, Jiande Sun, and Huaxiang Zhang. Learning deep multiscale local dissimilarity prior for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 5

[43] Yuxuan Zheng, Jiaojiao Li, Yunsong Li, Jie Guo, Xianyun Wu, and Jocelyn Chanussot. Hyperspectral pansharpening using deep prior and dual attention residual network. *IEEE transactions on geoscience and remote sensing*, 58(11):8059–8076, 2020. 5