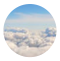
Hyperspectral and Panchromatic images Fusion Based on The Dual Conditional Diffusion Models

Article *in* IEEE Transactions on Geoscience and Remote Sensing · September 2023

DOI: 10.1109/TGRS.2023.3321318

3 authors:

Li Shuangliang
Wuhan University
7 PUBLICATIONS **31** CITATIONS
SEE PROFILE



Siwei Li
Wuhan University
72 PUBLICATIONS **694** CITATIONS
SEE PROFILE

Lihao Zhang
Beijing Normal University
12 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Hyperspectral and Panchromatic images Fusion Based on The Dual Conditional Diffusion Models

Shuangliang Li, Siwei Li*, Lihao Zhang

***Abstract*—The fusion between the low resolution hyperspectral image (LRHSI) and the panchromatic (PAN) image could obtain the high-resolution hyperspectral image (HRHSI). Recently, deep learning (DL)-based fusion methods have been explored widely due to their powerful feature learning ability. However, most DL-based methods that use the one-step fusion manner can suffer from the blurring effect. In addition, they have not fully utilized the spatial and spectral feature information of two input images, which hinders the improvement of the resulting image quality. Therefore, to fully mitigate the blurring effect and utilize two input images, we propose a dual conditional diffusion models-based fusion network (DCDMF) to obtain the fused HRHSI. The conditional diffusion model (CDM) can generate the high quality image with realistic details in an iterative denoising manner (in the inference sampling stage) other than the one-step fusion manner, which could mitigate the blurring effect greatly. To improve the spatial and spectral fidelity of the fused HRHSI, we propose the dual spatial and spectral CDM (two noise prediction networks with different conditional input) to respectively extract the image feature from the LRHSI and PAN images with different image characteristics and reconstruct the corresponding HRHSI feature and fuse them. In addition, considering the high-dimensional property of the HSI, we pre-train an auto-encoder to encode the HSI into the low-dimensional latent space with more discriminate features to reduce the computational cost. Based on the auto-encoder, we also perform the image generation process in the residual latent space to focus on learning the residual latent spatial details. Extensive experimental results on three datasets show the superiority of our method over several state-of-the-art (SOTA) methods. (The ziyuan dataset and codes could be available at https://github.com/rs-lsl/DCDMF)**

***Index Terms*—hyperspectral image fusion; diffusion model; spectral and spatial feature; auto-encoder; residual space;**

I. Introduction

THE hyperspectral image (HSI) with hundreds of spectral bands provides rich spectral information for different materials and objects. And this has made it used in many real-world applications, such as classification [1][2], spectral unmixing [3][4], mineral exploitation [5] and so on.

However, due to the limitations of satellite imaging sensors, HSI always suffers from coarse spatial resolution, which leads to the loss of detailed information including tiny spatial textures and details. For example, the MODIS satellite captures the HSI at 500m resolution [6] and the Ziyuan (ZY) 1-02D satellite obtains the HSI at 30m resolution [7]. To effectively improve the spatial resolution of the HSI, fusion methods between LRHSI and PAN images have been explored greatly in recent years [8][9][10][11][12], which aim to inject the PAN details into the LRHSI. And this fusion process is called 'hyperpansharpening' for the remote sensing image.

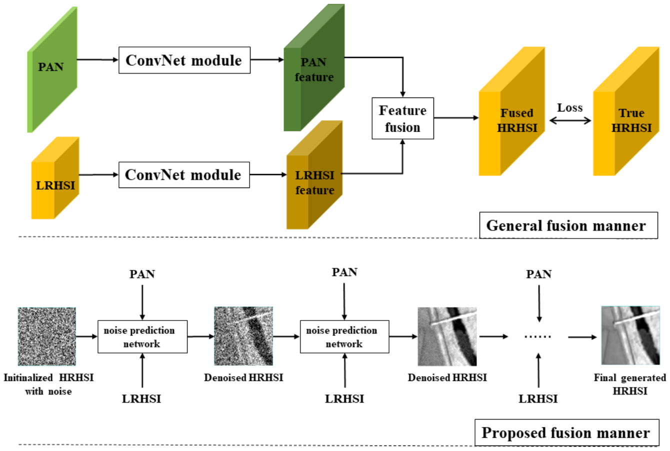


Fig. 1. Comparison between the fusion manner of the general fusion network and the proposed fusion network. 'ConvNet module' represents the prevalent CNN network structure for extracting image features from input images.

Most traditional hyperpansharpening methods are actually borrowed from the field of multi-spectral image pansharpening, which may degrade the resulting HRHSI quality due to the different image characteristics of these two fusion tasks (number of spectral bands and spatial resolution ratio). And they can be classified into three main classes: component substitute (CS)-based, multi-resolution analysis (MRA)-based and variational optimization (VO)-based. However, due to the stationary prior adopted by these methods, CS-based and MRA-based methods generally suffer from the spectral and spatial distortion. And VO-based methods usually take a long time to compute due to the iterative parameter optimization process.

Recently, deep learning (DL)-based fusion methods have been developed greatly. This could be attributed to their superiority over traditional methods in terms of accuracy and efficiency, especially in the fields of image classification and fusion fields [2][13][14][15][16][17]. In particular, the convolutional neural network (CNN) and the residual network with powerful feature learning ability are becoming increasingly popular in the fusion field. For example, He et al. [18][19] proposed the Hyperpnn and HSpeSet fusion networks. These two networks all concatenate the PAN image and LRHSI features as input and produce fused HRHSI through the

S. Li and S. Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, China (e-mail: whu_lsl@whu.edu.cn; siwei.li@whu.edu.cn;)*(Corresponding author: Siwei Li.)*

L. Zhang is with the Faculty of Geographical Science, Beijing Normal University, Beijing, 100088, China (e-mail: zhanglihaocug@gmail.com)

cascade convolution layers and reconstruction module. Qu et al. [20] proposed a residual hyper-dense network that uses the DenseNet to solve the fusion problem between the LRHSI and the PAN image. He et al. [21] designed a hierarchical pyramid sub-pixel mapping network with the high-frequency-aware differential architecture, and an HP architecture to achieve explicit multi-scale feature map supervision.

Actually, the hyperpansharpening task is a very underdetermined problem. For example, the ZY-1 02D satellite obtains the LRHSI of pixel size 30m with 166 bands and the paired single-band PAN image with a spatial resolution of 2.5m [7]. The differences in spatial and spectral resolution between these two images are enormous. And the infinite possible HRHSI fusion results can be generated from the available LRHSI and PAN images. However, as shown in Fig. 1, the prevalent DL-based fusion networks are mostly trained by pixel-wise loss functions (e.g. mean square error (MSE)) and produce only one fused HRHSI by the one-step fusion manner, which is the average of many possible generated HRHSIs with different spatial and spectral details. This leads to the blurring effect in the fused HRHSI, which lacks the necessary details [22][23][24].

Recently, the Denoising Diffusion Probabilistic Model (DDPM) [25] has exploded in the field of image generation, especially the CDM with the conditional information. Its generative capacity is better than the general network (e.g. GAN) in avoiding the model collapse and achieving the trade-off between fidelity and diversity. DDPM can closely learn the distribution of the training data through the interpretable VLB loss function and provides excellent generative performance in image generation [26][27][28], super-resolution [22][23][29][30][31], repair [32], and so on. Specifically, as shown in Fig. 2, in the reverse denoising process (inference sampling stage), DDPM learns to model the posterior distribution at each time step and gradually transfers the samples in the Gaussian distribution to the target data distribution. The final generated samples are following the target image distribution and are full of realistic detail. In addition, it has been verified that the DDPM can generalize well to the heavily out-of-distribution test image feature [23][33]. Therefore, it is desirable to apply the CDM to the hyperpansharpening task which could generate the high quality HRHSI.

Several works have applied the CDM to image super-resolution and fusion tasks. For example, Li et al. [29] designed a novel SISR diffusion probabilistic model with the residual prediction to super-resolution the natural image. Liu et al. [30] adopted the generative diffusion model with detail complement for the remote sensing image super-resolution and designed the detail supplement module to improve its recovery ability. In addition, Wu et al. [34] proposed an HSI fusion approach with a conditional diffusion model and employed a progressive learning strategy to exploit the global information. Shi et al. [35] proposed a deep fusion method based on the conditional denoising diffusion probabilistic model. However, these methods have not designed the specific module to extract the unique feature from each of the two input images, which may degrade the spatial and spectral fidelity of the resulting image.

To effectively improve the detail restoration and fidelity of the fused HRHSI, we design a dual CDM-based fusion network (DCDMF). To improve the spatial and spectral fidelity of the resulting HRHSI, we propose the dual spatial and spectral CDM (DCDM) to sufficiently exploit the conditional information. Considering the high-dimensional property of the HSI, we construct an auto-encoder to project the HSI into the low-dimensional latent feature. We also adopt the residual space design to perform the diffusion process in the residual latent space. This could make the network focus on learning the spatial details and ease the training process.

The main contributions of this study can be summarized as follows:

(1) We propose a dual CDM-based fusion network (DCDMF) to fuse the LRHSI and PAN image, which could benefit from its powerful image generation capability to generate realistic and fruitful details. We also perform the image generation process in the residual space to focus on learning the spatial details.

(2) To improve the fused image fidelity, we propose the dual spatial and spectral CDM (DCDM) to sufficiently extract the unique feature information from two conditional images. And we pre-trained an auto-encoder to learn the low-dimensional latent feature of HRHSI, which could reduce the computational cost in the inference stage greatly.

(3) Unlike the prevalent DDPM-based model, which is trained on the large-scale dataset, the proposed DCDMF can be trained on the small-scale dataset and shows excellent performance on the test dataset, which could be attributed to the designed auto-encoder and residual space that reduce the dependency on the large-scale dataset. This demonstrates the feasibility and effectiveness of training the DDPM-based fusion network on the small-scale dataset.

The remaining part of this paper is organized as follows. Section II introduces the related work with the research issues, including hyperpansharpening methods and denoising diffusion probabilistic models. Section III and IV describe the based theory and detailed network structure. Experimental results are given in section V. Finally, the conclusion is made in section VI.

II. RELATED WORKS

A. Hyperpansharpening Methods

Most of the hyperpansharpening methods can be divided into traditional and DL-based methods. Traditional methods include CS-based, MRA-based and VO-based pansharpening methods.

In detail, CS-based methods use the statistics feature-matched PAN image to substitute the generated intensity band from the LRHSI and then project it back into the original HSI space. This class includes Gram-Schmidt adaptive (GSA) [36] and intensity-hue-saturation (IHS) [37]. MRA-based methods generally generate a multi-scale image pyramid from the PAN image and inject multi-scale detail information into the LRHSI. Common methods in this class include smoothing filter-based intensity modulation (SFIM) [38], wavelet transform (Wavelet) [39], and modulation transfer function with

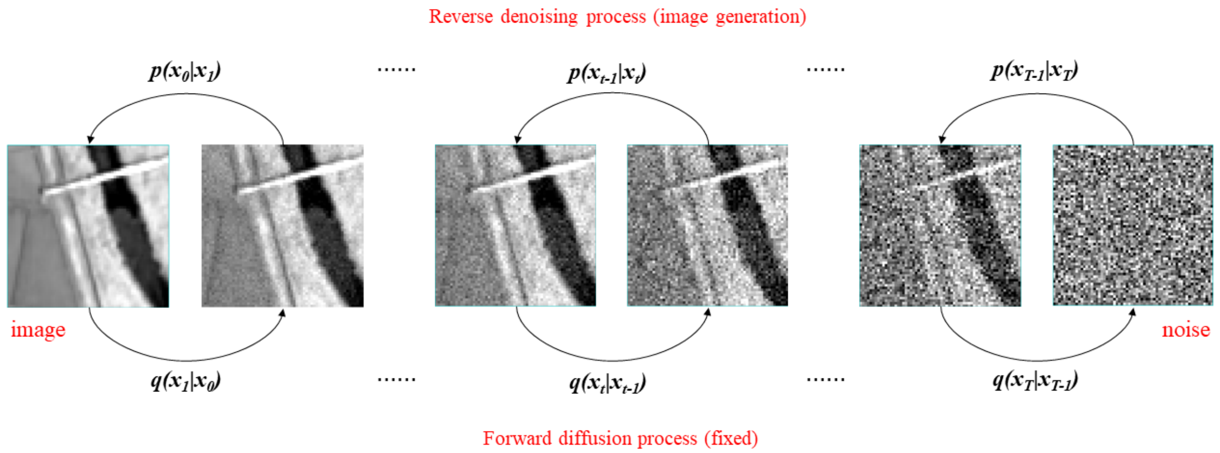


Fig. 2. The demonstration of the forward and reverse diffusion process in DDPM. Generally, the $q()$ is pre-defined and $p()$ is parameterized by the learnable neural network. In addition, the 'forward noise addition process' from x_0 to x_T is finished in one step by Eq. 3.

a generalized Laplacian pyramid with high-pass modulation (MTF_GLP_HPM) [40]. Despite the fast fusion speed, these two classes of methods generally introduce spectral and spatial distortions into the fused HRHSI.

VO-based methods usually formulate the fusion process as a constrained optimization function under the specific prior (e.g. sparse and low-rank priors) and alternatively optimize the desired variables until convergence. Popular methods in this category include coupled nonnegative matrix factorization (CNMF) [41], Bayesian sparsity promoted Gaussian prior (Bayesian Sparse) [8], HySure [9] and Bayesian naive Gaussian prior (Bayesian Naive) [42]. In addition, some tensor-based fusion methods are proposed to complete the hyperpansharpening task [43][44][45][46][47]. For example, based on the coupled sparse tensor factorization, Dian et al. [44] proposed an HSI fusion method to exploit the spatial-spectral structures of the input images. Even with the satisfactory fusion performance they could achieve, most of these methods suffer from the high computational cost due to the iterative parameter optimization process [13].

DL-based methods have seen explosive development in recent years due to their powerful feature learning ability [48]. Many hyperspectral and PAN image fusion networks have been designed and achieved great performance. For example, Zheng et al. [10] exploited the deep hyperspectral prior (DIP) to upsample the LRHSI and proposed the spatial and spectral attention network to inject the PAN details into the LRHSI. Then, Wele et al. [49] used the improved DIP and residual structure to obtain the reconstructed HRHSI with a learnable spectral response function. Qu et al. [11] designed a dual-branch detail extraction network that could sharpen the LRHSI with any number of spectral bands by the pre-trained model. Recently, based on the transformer module [50], Wele et al. [12] proposed a HyperTransformer fusion network including two separate feature extractors, a multi-head attention module, and a spectral-spatial feature fusion module to reconstruct the HRHSI. However, as shown in Fig. 1, these methods use the one-step fusion manner and lack the ability to generate realistic details.

B. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPM) [25][51] have recently attracted much attention in the image generation tasks, including text-to-image, super-resolution, deblurring and inpainting. This model is a generative model, which could be separated into the forward noise addition and the backward denoising process. And the images are generated in the denoising process, which gradually denoises a Gaussian noise to an image that obeys the target image distribution.

The DDPMs have been explored in the image super-resolution field recently. For example, Chitwan et al. [22] adapted the DDPM to perform the super-resolution conditioned on a low-resolution image under the training process with multi-scale noise. Chung et al. [23] proposed a new super-resolution method based on score-based reverse diffusion sampling. Sahak et al. [33] introduced SR3+, a diffusion-based model for blind super-resolution that combines the self-supervised training with composite and parameterized degradation. Niu et al. [52] proposed a simple but non-trivial diffusion model-based super-resolution post-processing framework.

DDPM has also been applied to the image fusion field. For example, Wu et al. [34] proposed an HSI fusion approach with a conditional diffusion model and used a progressive learning strategy to exploit the global information. Rui et al. [53] proposed an unsupervised pansharpening network by combining the diffusion model with the low-rank matrix factorization technique. Meng et al. [54] proposed a DDPM-based pansharpening method called PanDiff to learn the data distribution of the difference maps between the upsampled LRMSI (LR multi-spectral image) and HRMSI. However, for the fusion task, few studies propose the specific module to fully extract the respective feature information from two conditional images.

III. Preliminaries: Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Model (DDPM) is adopted as the generative model for the hyperspectral image

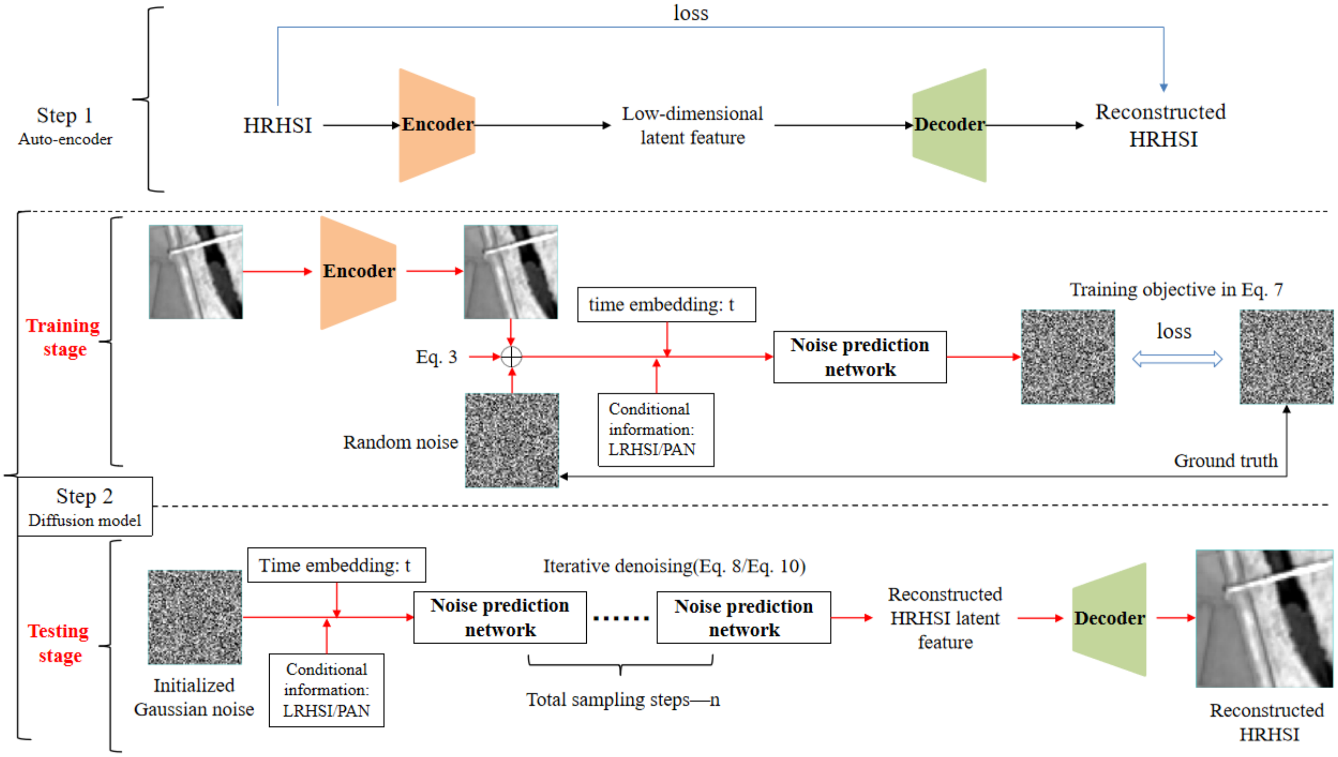


Fig. 3. The overall framework of the proposed DDPM-based fusion method. In step 1, the auto-encoder is trained firstly where the HRHSI is encoded into the latent feature by the 'encoder'. And then it is reconstructed into HRHSI through the 'Decoder'. In the training stage of step 2, the 'noise prediction network' is optimized by the loss function in Eq. 7. Then it is used to sample iteratively to get the reconstructed HRHSI latent feature in the testing stage. Note the 'residual design' is not depicted in this figure.

fusion task. This model could learn the target data distribution through the iterative denoising manner, as shown in Fig. 2. It starts from the white Gaussian noise $x_T \sim \mathcal{N}(0,1)$ and gradually transforms it into the image space x_0. T means the iterative step of the denoising process. This backward denoising process is the reverse process of the forward noise addition process and they are related by the Bayes theorem. Therefore, we first define the forward noise addition as:

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

where $\mathcal{N}()$ means the Gaussian distribution. Each x_t in the noise-adding process is Obey the Gaussian distribution with the mean of $\sqrt{1-\beta_t} x_{t-1}$ and variance of $\beta_t \mathbf{I}$. β_t is the noise variance in time step t and it is generally pre-fixed (it is set to gradually increase from 1e-4 to 2e-2).

By using the independence property of the Gaussian noise added at each step in Eq. 1, we can calculate the total noise variance as $\bar{\alpha}_t=\prod_{s=1}^t \alpha_s$ and $\alpha_t=1-\beta_t$. And the one-step forward noise addition process from x_0 to x_t can thus be rewritten as:

$$q\left(x_t \mid x_0\right)=\mathcal{N}\left(x_t ; \sqrt{\bar{\alpha}_t} x_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \tag{2}$$

This one-step noise addition formulation could be derived by combining all of Eq. 1 from time step $0-t$. In this case, when t=T in Eq. 2, the distribution of x_T would be almost the white Gaussian noise.

Then we could sample the x_t from x_0 in one step as:

$$\mathbf{x_t}=\sqrt{\bar{\alpha}_\mathbf{t}} \mathbf{x_0}+\sqrt{\left(\mathbf{1}-\bar{\alpha}_\mathbf{t}\right)} \epsilon_\mathbf{0} \tag{3}$$

Where ϵ_0 is the added noise to x_0 that follows the standard Gaussian distribution ($\mathcal{N}(0, \mathrm{I})$).

To reverse the noise addition process and sample the x_0 from x_T, we need to iteratively sample the distribution of $q\left(x_{t-1} \mid x_t\right)$ from $t=0 \rightarrow T$ (T means the number of reverse iterations). And this distribution could be learned by the neural network. Note that according to the Bayesian theorem, this reverse distribution $q\left(x_{t-1} \mid x_t\right)$ also obeys the Gaussian distribution with the mean of $\mu_\theta\left(x_t, t\right)$ and the variance of $\Sigma_\theta\left(x_t, t\right)$:

$$p_\theta\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{4}$$

The above parameters of the Gaussian distribution are learned by considering the variational lower bound (vlb) of the $\log p_\theta\left(\mathbf{x}_0\right)$.

$$\begin{aligned} \mathbb{E}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0: T}\right)}{q\left(\mathbf{x}_{1: T} \mid \mathbf{x}_0\right)}\right] \\ &=\mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right)-\sum_{t \geq 1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right]=L_{v l b} \end{aligned} \tag{5}$$

As demonstrated by Ho et al. [25], by introducing $\mathbf{x}_0$ as another condition, this **vlb** loss (L_{vlb}) can be further decomposed as:

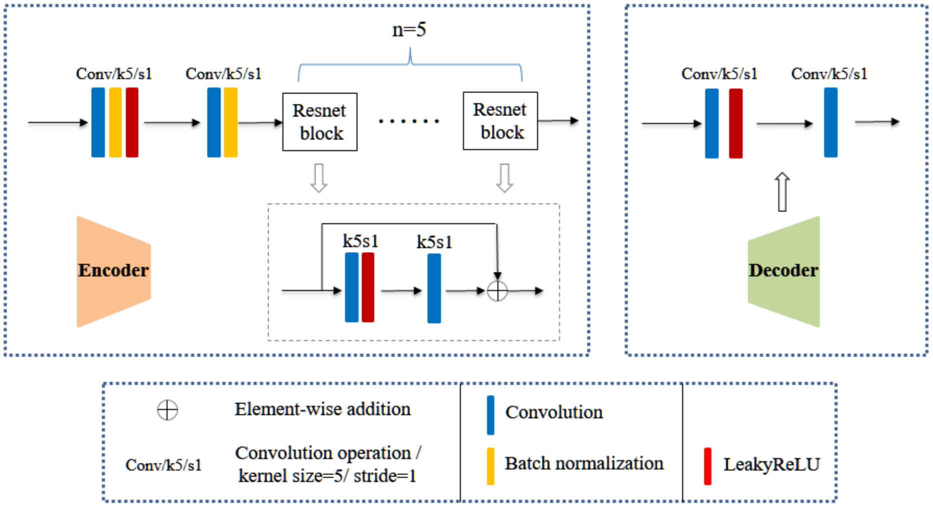


Fig. 4. The detailed network structure of the constructed auto-encoder.

$$L_{vlb}=\mathbb{E}_q\Bigg[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T\mid\mathbf{x}_0\right)\|p\left(\mathbf{x}_T\right)\right)}_{L_T} \tag{6}$$

$$+\sum_{t>1}\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0\right)\|p_\theta\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t\right)\right)}_{L_{t-1}}\underbrace{-\log p_\theta\left(\mathbf{x}_0\mid\mathbf{x}_1\right)}_{L_0}\Bigg]$$

The critical term L_{t-1} trains the neural network in Eq. 4 to perform the reverse denoising step. Furthermore, after introducing x_0, the distribution of $q(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0)$ is also Gaussian according to the Bayesian theorem.

As reported by Ho et al. [25], the best way to parameterize the model is to predict the cumulative noise ϵ_0 that is added to the current intermediate image x_t, as in Eq. 3. And from the L_{t-1} in Eq. 6, Ho et al. [25] derived the following simplified training objective:

$$L_{\text{simple}}=E_{t,x_0,\epsilon_0}\left[\left|\epsilon_0-\epsilon_\theta\left(x_t,t\right)\right|^2\right] \tag{7}$$

We could efficiently obtain pairs of training data (t, ϵ_0, x_t) to train a reverse transition step based on Eq. 3, as shown in the 'training stage' of Fig. 3.

Then, we could get the following parameterization of the predicted mean $\mu_\theta\left(x_t,t\right)$ in Eq. 4 as:

$$\mu_\theta\left(x_t,t\right)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t,t\right)\right) \tag{8}$$

And the $\Sigma_\theta\left(x_t,t\right)$ is usually predefined as:

$$\Sigma_\theta\left(x_t,t\right)=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{9}$$

Then, based on Eq. 4, we could sample the x_{t-1} condition on the x_t:

$$x_{t-1}=\mu_\theta\left(x_t,t\right)+\Sigma_\theta\left(x_t,t\right)\epsilon_t \tag{10}$$

After iteratively running the above denoising process from $t=0\rightarrow T$, the x_0 is obtained, as shown in the testing stage of Fig. 3.

IV. Method

A. Auto-encoder

We design the auto-encoder network to perform the diffusion process in the encoded low-dimensional latent space [28]. On the one hand, the sampling stage of the diffusion model is to gradually transfer the noise to the HRHSI, and the intermediate noise in each step all have the same dimension as the resulting image—HRHSI. Thus, encoding the high-dimensional HSI into the low-dimensional latent feature could reduce the memory and computational resources in the training and sampling process of the diffusion model. On the other hand, the original HSI feature is redundant in the spectral dimension (it could be decomposed into the endmember and abundance features, reflecting its redundancy). The encoded latent feature could reserve more discriminative features, which could improve training stability by regularizing the input and output variable space [55].

Therefore, we construct an auto-encoder network structure to encode the HSI feature into the low-spectral-dimensional latent feature. Then, the subsequent training and sampling processes of the diffusion model are performed on the encoded low-dimensional feature space. The final generated latent feature would be decoded to the HRHSI by the decoder, as shown in the testing stage of Fig. 3. Note that in order to preserve the spatial detail information, we only reduce the spectral dimension of the HRHSI in the encoding operation:

$$H_e=\varepsilon(H) \tag{11}$$

where H is the HRHSI and H_e is the encoded latent feature of the HRHSI. ε represents the constructed encoder.

The network structures of the encoder include two 'conv' blocks and cascading 'resnet' blocks, as shown in Fig. 4. The first 'conv' block includes Conv+Batchnorm+LeakyReLU

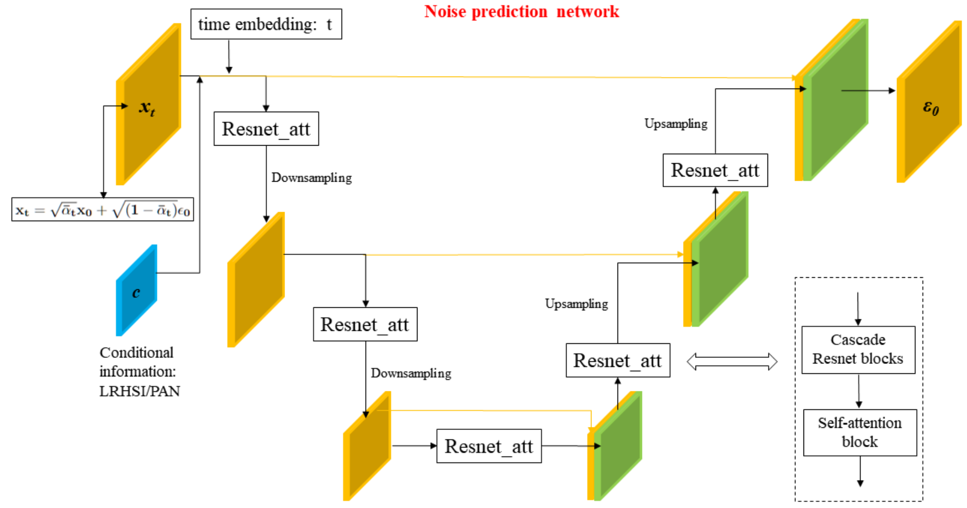


Fig. 5. The ‘unet’ structure of the proposed noise prediction network. The conditional information includes the LRHSI and PAN images. Each of the ‘Resnet_att’ blocks consists of the cascade ‘Resnet’ block as in Fig. 4 and the self-attention block as in Fig. 6. The ‘Upsampling’ is implemented by the ‘nearest’ interpolate followed by the convolution operation. The ‘Downsampling’ only has one stride convolution layer. The yellow arrows mean the short connections.

three layers, while the second contains only Conv+Batchnorm. These two ‘conv’ blocks are used to extract the shallow features of the input image. Then the cascaded ‘resnet’ blocks are used to extract the deep features. Each ‘resnet’ block consists of two convolutional layers and a LeakyReLU function between them. And the short connection from the input to the output is added in the ‘resnet’ block.

After obtaining the deep latent feature of the input image, the decoder is then used to transform the latent feature into the original image:

$$\widehat{H} = De(H_e) \tag{12}$$

where De is the decoder and $\widehat{H}$ is the reconstructed HRHSI. As shown in the right part of Fig. 4, the decoder contains only two convolution layers and a LeakyReLU function.

Note that we first train the auto-encoder until it converges. Then we train the diffusion model with the parameter-fixed auto-encoder structure. The auto-encoder is trained using the L_2 distance loss function:

$$L_2 = \|H - \widehat{H}\|_2 \tag{13}$$

B. Noise prediction network

As discussed in section III, to reverse the noise addition process and gradually sample the HRHSI from the initialized noise, we need to learn the mean—$\mu_\theta\left(x_t, t\right)$ of the distribution $q(x_{t-1}|x_t)$ as in Eq. 4.

In fact, the training objective has been transferred to predict the noise x_0 that is involved in the input noisy image x_t [25] as in Eq. 7:

$$\mu_\theta\left(x_t, t\right) \rightarrow \epsilon_\theta\left(x_t, t\right) \tag{14}$$

where x_t is obtained by Eq. 3. ‘t’ is the time step embedding. $\epsilon_\theta\left(x_t, t\right)$ represents the noise prediction network. Note that the time embedding is the cumulative noise $\sqrt{(1-\bar{\alpha}_t)}$. After obtaining the $\epsilon_\theta\left(x_t, t\right)$, the $\mu_\theta\left(x_t, t\right)$ could be calculated by Eq. 8.

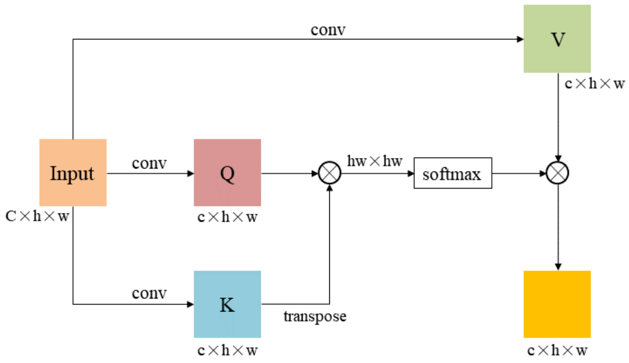


Fig. 6. The network structure of the self-attention block in the ‘Resnet_att’ block of Fig. 5. ‘conv’ means the convolution operation. ‘transpose’ means the transpose operation. ‘softmax’ means the softmax operation to normalize the matrix.

The noise prediction network is used to predict the noise ϵ_0 involved in the input image x_t. So we need to learn the global noise distribution for different noisy images x_t and time embeddings t. And the ‘unet’ structure is adopted to learn this noise distribution [25]. This structure could perfectly learn the local texture feature of the image at the high-resolution scale and the global structure feature at the down-sampled low-resolution scale.

The overall network structure of the ‘unet’ is shown in Fig. 5. Its input includes the noisy image x_t, the time embedding t and the conditional information c. Note that the conditional information includes the ‘bilinear’ upsampled LRHSI or PAN image, which could improve the spectral and spatial fidelity of the resulting image. This network then outputs the predicted noise:

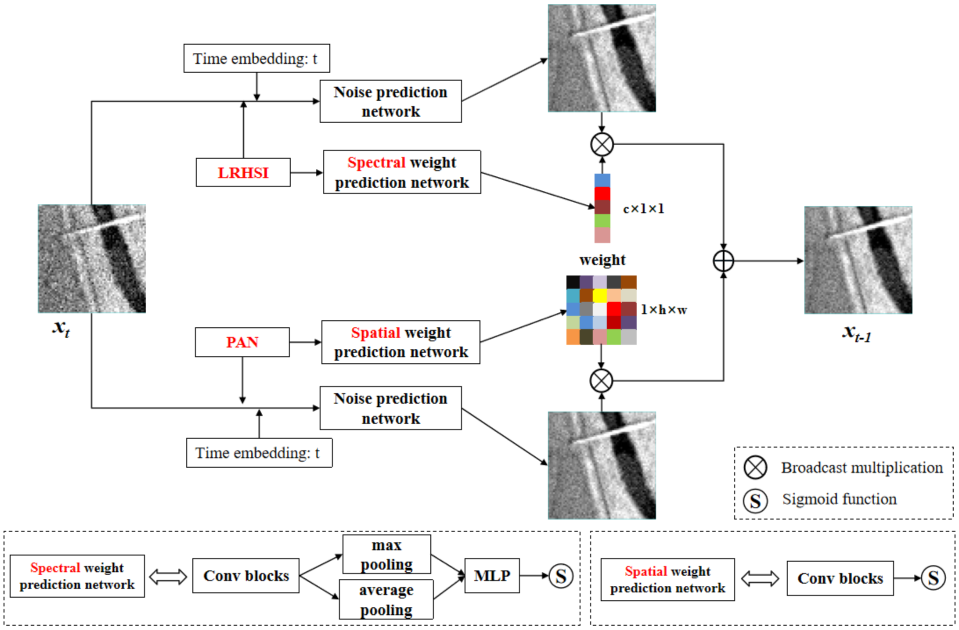


Fig. 7. The network structure of the proposed dual spatial and spectral CDM (DCDM). Note that the above two 'Noise prediction networks' are not sharing the parameters, which is the same as the 'Weight prediction network'. The 'max pooling' means the global max pooling operation, while the 'average pooling' means the global average pooling. The 'MLP' means the multi-layer perceptron network.

$$\epsilon_\theta\left(x_t, t, c\right) \overset{predict}{\longrightarrow} \widehat{\epsilon_0} \tag{15}$$

where 'c' means the conditional information. $\widehat{\epsilon_0}$ is the predicted noise involved in the x_t (x_t is derived from the true ϵ_0 based on Eq. 3). The time embedding layer is added to accurately predict the noise [56].

Thus, the proposed 'noise prediction network' is optimized by the loss function in Eq. 7:

$$L_1 = \|\epsilon_0 - \widehat{\epsilon_0}\|_1 \tag{16}$$

In detail, the 'unet' structure in Fig. 5 has three resolution scales with a depth of 3. And the actual depth is determined by the ablation study. We also add the skip connection in each resolution scale to alleviate the gradient vanishing and exploding problem. The main backbone of the 'unet' is the 'Resnet_att'ention block as shown in the right part of Fig. 5. Each of these blocks consists of two parts: the cascade 'Resnet' block as shown in Fig. 4, and the self-attention block in Fig. 6. As shown in Fig. 4, the 'Resnet' block contains several convolution layers and an activation function.

As shown in Fig. 6, the self-attention layer is adopted to enhance the input image feature by using the computed global attention map. And we add this attention block at different resolutions to fully learn the multi-scale global attention information.

With the noise prediction network, we could iteratively sample the low-dimensional latent feature of the HRHSI from the Gaussian noise by Eq. 8 and 10, as shown at the bottom of Fig. 3. The decoder is then used to reconstruct the HRHSI from the denoised latent feature.

C. Dual Spatial and Spectral CDM

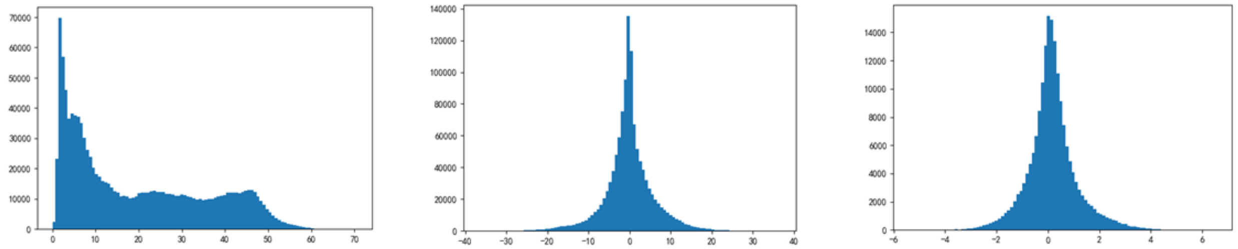
According to the use of conditional information (including LRHSI and PAN images), the diffusion model could be classified into the conditioned and unconditioned (for the unconditioned model, the input to the noise prediction network did not include the conditional information). And the addition of conditional information could improve the content consistency (fidelity) between the resulting image and the conditional information. Therefore, in order to improve the spatial and spectral fidelity of the resulting image by fully exploiting the conditional information, we propose the dual spatial and spectral CDM (DCDM), as shown in Fig. 7. This could sufficiently extract the spatial texture information involved in the PAN image and the spectral feature in the LRHSI, and fuse them into the HRHSI.

Note that we base the proposed DCDM's derivation on the score-based formulation of a diffusion model, and it is equivalent to the training objective in Eq. 6, which was verified in [57]. Therefore, our goal is to learn $\nabla_{x_t} \log p(x_t|c)$ (simplified as $\nabla \log p(x_t|c)$), which is the score of the conditional model at time t. ∇ means to compute the gradient. c is the conditional information including the LRHSI and PAN images.

Then, the score of the conditional model could be represented as:

$$\nabla \log p(x_t|c) = \nabla \log p(x_t|\mathrm{H_L}, \mathrm{P}) \tag{17}$$

where $\mathrm{H_L}$ and P are LRHSI and PAN images. Actually, the LRHSI and PAN images could be assumed to be independent of each other, especially their contributions to the x_t. So the $\nabla \log p(x_t|\mathrm{H_L}, \mathrm{P})$ could be decomposed into:



Original HRHSI

Residual details

Encoded residual details

Fig. 8. The transformation of the data distribution. The 'residual details' is the difference between the HRHSI and upsampled LRHSI. Then these details are encoded by the encoder to get the encoded residual details.

$$\nabla \log p(x_t|\mathrm{H_L}, \mathrm{P}) \propto \alpha \nabla \log \mathrm{p}(\mathrm{x_t}|\mathrm{H_L}) + \beta \nabla \log \mathrm{p}(\mathrm{x_t}|\mathrm{P}) \tag{18}$$

where α is the spectral weight learned from the LRHSI by the proposed 'spectral weight prediction network' as shown in Fig. 7. And the β is the spatial weight learned from the PAN image.

Based on Eq. 18, as shown in Fig. 7, we design the DCDM, which include dual noise prediction networks with the LRHSI and PAN images as the conditional information, respectively. This could sufficiently extract the spectral and spatial information from the LRHSI and PAN image and fuse into the final HRHSI feature. And the conditional images are also used to re-weight the reconstructed HRHSI feature.

As shown in the upper part of Fig. 7, the x_t, LRHSI and the time embedding t are fed into the noise prediction network to predict the denoised result. This could be represented as:

$$\epsilon_\theta^l(x_t, \mathrm{H_L}, \mathrm{t}) \xrightarrow{\text{predict}} \epsilon_0^{\mathrm{l}} \tag{19}$$

$$\epsilon_0^l \xrightarrow{Eq.8,10} x_{t-1}^l \tag{20}$$

where the ϵ_0^l is the predicted noise and it is transferred to the denoised result x_{t-1}^l by Eq. 8 and 10. And the weights are learned by the weight prediction network from the LRHSI:

$$W^l = WPN(\mathrm{H_L}) \tag{21}$$

$$x_{t-1}^l = W^l \cdot x_{t-1}^l \tag{22}$$

where $WPN()$ is the weight prediction network and W^l is the predicted band-wise weight from the LRHSI—$\mathrm{H_L}$. Then the W^l is multiplied by the x_{t-1}^l to get the reweighted x_{t-1}^l. Similarly, we get the reweighted x_{t-1}^p by using the PAN image. Note that the learned weight from the LRHSI is band-wise with a shape of c×1×1, while the learned weight from the PAN image is the spatial-wise with a shape of 1×h×w. This could sufficiently benefit from two conditional images with the different characteristics.

Finally, these two results are summed to get the final denoised x_{t-1}:

$$x_{t-1} = x_{t-1}^l + x_{t-1}^p \tag{23}$$

D. Perform the diffusion process in the residual latent space

Since each step in the reverse sampling process is based on the Gaussian assumption as in Eq. 10, the complex image features are difficult to simulate and sample unless a large number of steps are used in the reverse sampling process [58]. Therefore, making the target image follow the Gaussian distribution could ease the training process and reduce the number of sampling steps.

As shown in Fig. 8, the first statistical histogram is computed from the original HRHSI, which is much more irregular than the Gaussian distribution. The residual design is then applied, which means that the residual image details are used in the training and sampling stages of the auto-encoder and the 'noise prediction network'. Actually, performing the image generation process in the residual space could make the network focus on learning the spatial details and improve the generated image quality [59]. In detail, the auto-encoder is trained as follows:

$$H_e^r = \varepsilon(H - up(H_L)) \tag{24}$$

$$\widehat{H} = De(H_e^r) + up(H_L) \tag{25}$$

where H_L is the LRHSI, up means the 'bilinear' upsampling function. As shown in the middle of Fig. 8, obviously, the residual details obviously obey the Gaussian distribution and their data range is approximately between [-20, 20].

However, since the reverse sampling is initialized from the standard Gaussian distribution, it still requires many reverse sampling iterations to converge to the above data range. And considering the designed auto-encoder structure, we choose to feed the auto-encoder with the residual details of the HRHSI. And then, as shown in the right part of Fig. 8, the encoded low-dimensional residual details also follow the Gaussian distribution and are within the reduced data range. This eases the training process and reduces the number of reverse sampling iterations of the diffusion model.

Therefore, the input to the 'noise prediction network' is the encoded residual detail,

$$\epsilon_\theta\left(x_t, t, c\right) \rightarrow \epsilon_\theta\left(H_e^r, t, c\right) \tag{26}$$

where H_e^r is the encoded residual detail. And the iteratively reconstructed $\widehat{x}_0$ by Eq. 23 is actually the reconstructed residual latent feature and would be decoded and added to the upsampled LRHSI to obtain the final fused HRHSI:

$$\widehat{H} = De(\widehat{x}_0) + up(H_L) \tag{27}$$

where $\widehat{x}_0$ is the reconstructed residual latent feature according to Eq. 23 in the last time step 0. And $\widehat{H}$ represents the final generated HRHSI.

V. Experimental Results

This section presents the experimental results on three datasets. First, we describe these datasets and the experimental details, including hyperparameter settings, compared methods and quality indices. Then, we present the results of the ablation study on the overall fusion framework and the detailed hyperparameters. Finally, the fusion results are presented visually and quantitatively.

A. Datasets

1). ZY dataset: This dataset was acquired by the Ziyuan-1 02D satellite and includes the LRHSI and PAN images. Their spatial resolutions are 30 and 2.5 meters, respectively. The LRHSI has 166 bands and its spectral wavelength ranges from 395 to 2501 nm. Due to the low SNR in some long wavelength bands, we select a total of 76 spectral bands for the fusion experiments. Due to the lack of ground truth HRHSI, following the Wald protocol [60], we generate the downsampled LRHSI and PAN images as the input conditional images of the network, and the original HSI is regarded as the reference image.

2). Chikusei dataset [61]: The Chikusei scene was captured in Chikusei, Japan. The original image consists of 128 spectral bands from 363 to 1018 nm. The spatial size is 2517×2335 with a resolution of 2.5m. Following the Wald protocol [60], we generate the simulated LRHSI by the blurring and downsampling operation with a ratio of 12. And the PAN image is from the linear combination of HRHSI spectral bands weighted by the SRF of the worldview2 imaging sensor. The original HRHSI is regarded as the reference image.

3). XiongAn dataset [62]: This hyperspectral image was captured at Matiwan Village in XiongAn New Area, China. It mainly consists of different types of crops and grasses. This dataset includes 256 spectral bands with a spatial resolution of 0.5m. Its wavelength ranges from 391 to 1002nm. The height and width of this dataset are 1580 and 3750. And following the same operation with the Chikusei dataset, we generate the simulated LRHSI and PAN images. And the original HSI is seemed as the reference image.

The training patch sizes are cropped to 96×96, 96×96 and 8×8 for HRHSI, PAN image and LRHSI on these two datasets. We select almost 90% of these datasets for training the diffusion model, and the rest is for performance testing without overlapping patches. This demonstrates the feasibility and effectiveness of training the proposed DDPM-based fusion model on small-scale datasets, which is the advantage of the proposed method.

TABLE I
AVERAGE QUANTITATIVE RESULTS ON ZY DATASET WITH DIFFERENT NETWORK STRUCTURES. 'Auto-encoder' MEANS THE PRE-TRAINED 'Auto-encoder'. 'residual space' MEANS THAT TAKE THE TRAINING PROCESS IN THE 'residual space'. 'Baseline' MEANS HAS ALL THESE THREE MODULES.

Modules	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
w/o Auto-encoder	40.022	0.905	0.035	1.006	0.659
w/o DCDM	41.510	0.927	0.032	0.574	0.816
w/o residual space	26.798	0.875	0.069	1.858	0.144
Baseline	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**

B. Training Details

The designed auto-encoder is trained using the Adam optimizer with the proposed loss functions in Eq. 13, 24 and 25. The number of training epochs is 600 with a batch size of 32. The learning rate is initialized to 1e-4 and decays at a rate of 0.2 in 200 and 400 epochs.

The designed 'noise prediction network' is trained using the Adam optimizer with the proposed loss functions in Eq. 7 and Eq. 26. The number of training iterations is 10000 with a batch size of 32. The learning rate is set to 1e-4 in the training stage. And we use the exponential moving average (EMA) to stabilize the training process.

All experiments are run under the Paddle 2.4.0 framework and Python 3.7 environment on a single V100 Graphical Processing Unit (GPU). Using this device, the proposed fusion network took about 10000 seconds to train on the ZY dataset.

C. Comparison Methods And Quality Measures Metrics

We compare the proposed fusion method with several popular methods, including CS, MRA, VO and DL-based approaches. The first class includes GSA [36]. And MRA-based methods consist of SFIM [38], Wavelet [39] and MTF_GLP_HPM [40]. The VO-based has CNMF [41], while HSpeSet2 [19], Pgnet [63] and Srdiff [29] (DDPM-based) are all belong to DL-based methods. For a fair comparison, the training epochs and batch sizes of these methods are all modified to achieve their best fusion performance. In detail, the training epochs of HSpeSet2 and Pgnet are set to 500, while the training iteration of the Srdiff is 10000. And the batch sizes of these methods are set to 16, 16, 32, respectively.

The quality metrics used to measure and compare the different methods include Peak Signal-to-Noise Ratio (PSNR) [41], structure similarity (SSIM) [64], Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [13], Spectral Angle Mapper (SAM) [13], and spatial consistency coefficient (SCC) [13]. PSNR and ERGAS compute the absolute errors between the fused HRHSI and the reference image. SAM indicates the angular separation, while SSIM and SCC represent the spectral and spatial similarity. Note that the first rank is highlighted by the bold font, while the underlined results represent the second rank among all methods.

D. Ablation study

1) Network Modules and Designs

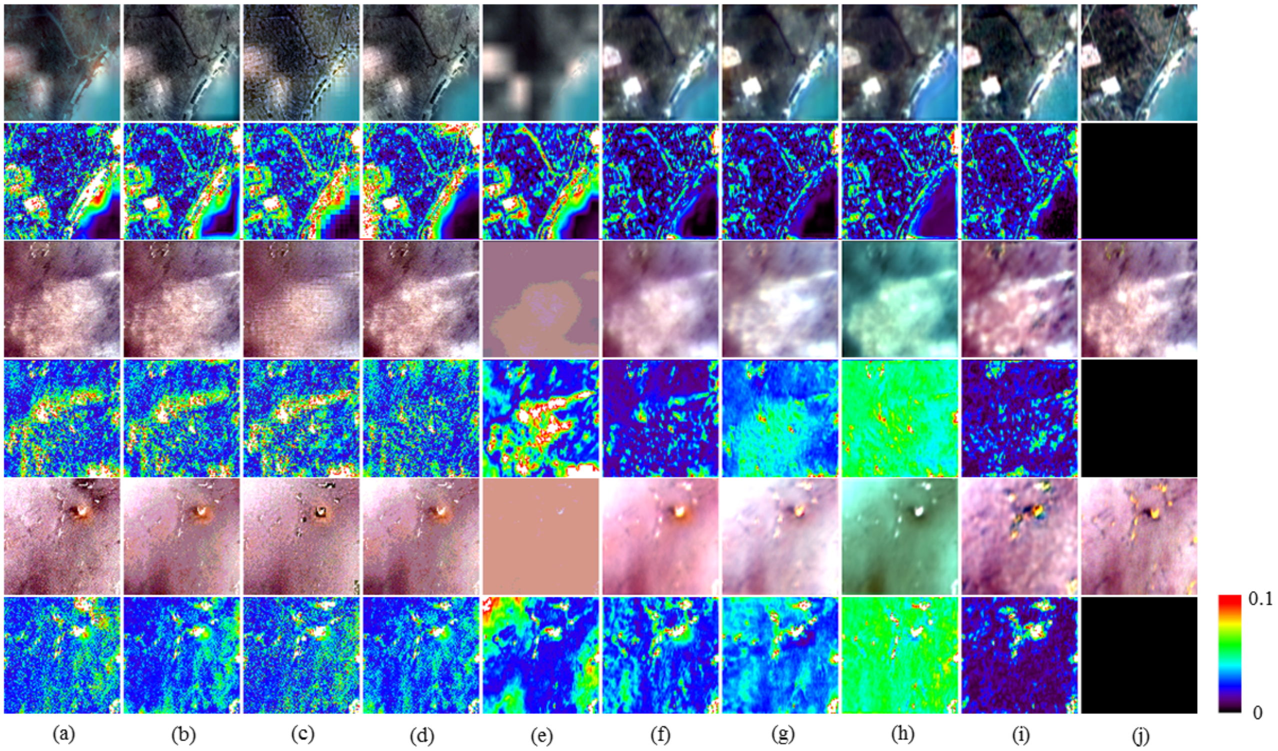


Fig. 9. Fused results (odd row) and error maps (even row) of different methods on the reduced ZY dataset. (a) GSA. (b) SFIM. (c) Wavelet. (d) MTF_GLP_HPM. (e) CNMF. (f) HSpeSet2. (g) Pgnet. (h) Srdiff. (i) Ours. (j) ground truth. The RGB image is shown by utilizing the 7-th, 16-th, and 35-th bands of the original HSI. Note that the error map ranging from 0 to 0.1 is calculated from the average of all bands' absolute difference between the reference and the fused result.

To demonstrate the effectiveness of all modules in the proposed DDPM-based fusion method, we remove each of them separately. The quantitative results of these ablation studies are listed in Table I. Obviously, all these modules contribute to the final excellent fusion performance, especially the implementation of the diffusion process in the residual space, which greatly improves the image quality. The improvement of using the residual space verifies the data distribution transformation pattern in Fig. 8, where indicates that the residual space design could transform the target HRHSI data into the Gaussian distribution. And the DDPM is relies on the Gaussian distribution, which means that the initialized noise in the denoising step is the standard Gaussian distribution noise and the generated image in each denoising step follows the Gaussian distribution. Therefore, by adopting the residual design to make the target image follow the Gaussian distribution could make the image generation process (denoising step) easier to fit to the target HRHSI distribution and improve the resulting image quality.

Note that 'w/o DCDM' means that there's only one noise prediction network with the concatenation of the LRHSI and PAN images as the conditional information. And it could be seen from Table I that the proposed 'DCDM' improves the spatial and spectral fidelity of the resulting HRHSI. In addition, the 'auto-encoder' module not only improves the quality of the sampled image but also reduces the computational cost in the sampling stage. Therefore, the above ablation studies demonstrate the effectiveness of all the proposed modules.

TABLE II
AVERAGE QUANTITATIVE RESULT ON **ZY** DATASET WITH DIFFERENT NUMBER OF LATENT FEATURE BANDS IN THE AUTO-ENCODER.

number of latent feature bands	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
10	40.612	0.934	0.033	0.581	0.820
20	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**
30	39.469	0.930	0.030	0.685	0.817
40	40.622	0.936	0.032	0.608	0.818

TABLE III
AVERAGE QUANTITATIVE RESULT ON **ZY** DATASET WITH DIFFERENT NETWORK DEPTHS IN THE 'unet' STRUCTURE.

'unet' network depth	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
3	38.987	0.929	0.032	0.631	0.793
4	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**
5	40.875	0.932	0.032	0.643	0.817

2) Auto-encoder and 'unet' Structure

To improve the accuracy of the latent feature extracted by the auto-encoder and to reduce the computational cost, we test the different band numbers of the latent feature. As shown in Table II, the band number of 20 achieves the best performance and maintains a low computational cost. Therefore, we set the

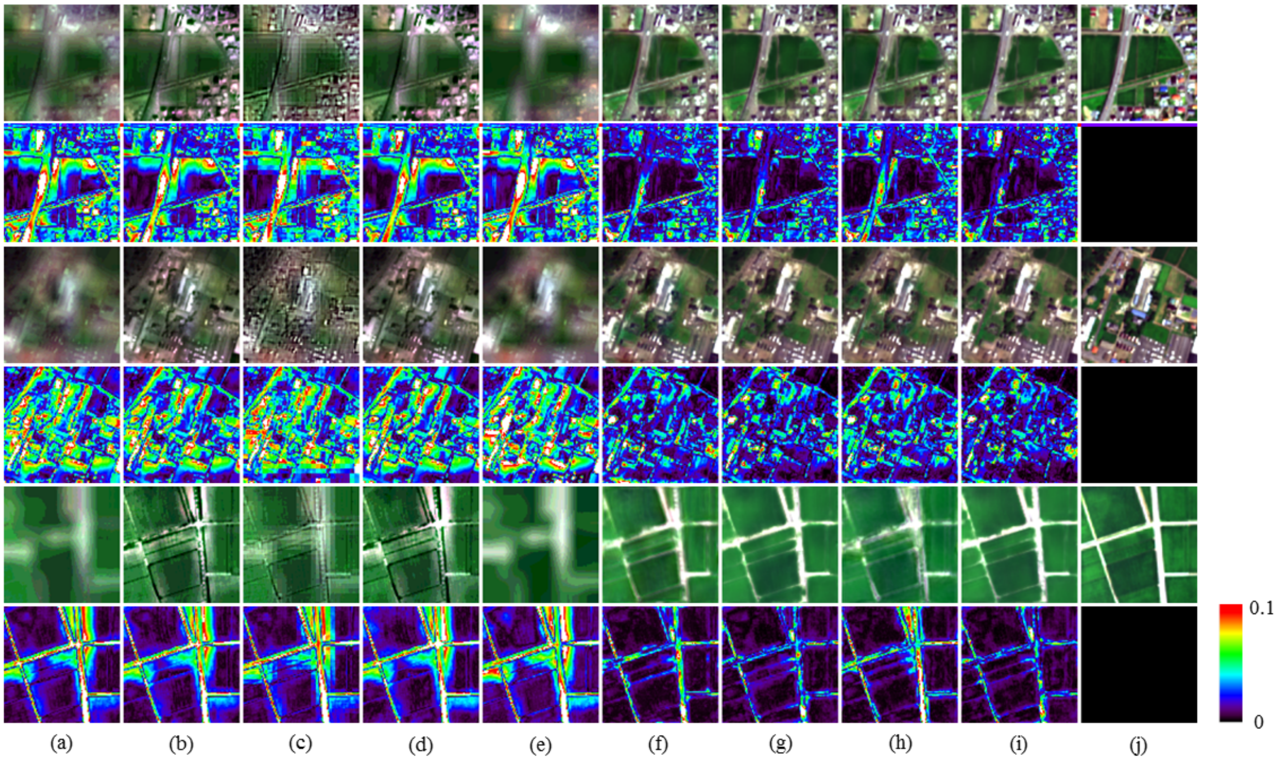


Fig. 10. Fused results (odd row) and error maps (even row) of different methods on the Chikusei dataset. (a) GSA. (b) SFIM. (c) Wavelet. (d) MTF_GLP_HPM. (e) CNMF. (f) HSpeSet2. (g) Pgnet. (h) Srdiff. (i) Ours. (j) ground truth. The RGB image is shown by utilizing the 20-th, 40-th, and 60-th bands of the original HSI.

TABLE IV
AVERAGE QUANTITATIVE RESULT ON **ZY** DATASET WITH THE SELF-ATTENTION BLOCK USED ON DIFFERENT RESOLUTIONS

12	24	48	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
✗	✗	✗	41.700	0.934	0.030	0.629	0.811
✓	✗	✗	41.556	0.933	0.030	0.621	0.812
✓	✓	✗	40.361	0.935	0.031	0.621	0.810
✓	✓	✓	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**

band number of the baseline to 20.

To accurately predict the noise involved in the x_t with the different scales as in Eq. 3, the noise prediction network—'unet' structure should be adjusted to fully adapt to the input image feature. First, we test the different network depths as in Fig. 5. This figure has a depth of 3, but it is adjustable to achieve better noise prediction performance. In general, increasing the network depth would require more computational resources. As shown in Table III, depth 4 achieves the best fusion performance, demonstrating its superiority. And the number of feature bands in each depth is set to 32, 64, 128 and 128.

As shown in Fig. 5, the 'Resnet_att' is composed of the cascade 'resnet blocks' and the self-attention block. The detailed structure of the self-attention block is shown in Fig. 6, which could use the global feature dependency to improve the image feature. And the quantitative results with the self-attention blocks used in different resolutions are shown in Table IV. It could be concluded that the addition of the self-attention block at three scales improves the final performance.

TABLE V
DIFFERENT FORWARD NOISE SCHEDULES AND ITS QUANTITATIVE RESULTS ON ZY DATASET

Noise schedule	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
Linear	39.247	0.911	0.036	0.725	0.726
Cosine	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**
Quad	41.110	0.935	0.030	0.566	0.821

TABLE VI
DIFFERENT SAMPLING TIME STEPS AND ITS QUANTITATIVE RESULTS ON ZY DATASET

Time step	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
500	40.913	0.931	0.032	0.635	0.813
1000	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**
1500	31.425	0.869	0.064	1.483	0.561

3) Noise Schedules and Number of sampling time step

We test the different forward noise schedules including Linear, Cosine and Quad (they are set to gradually increase from 1e-4 to 2e-2). The quantitative results are listed in Table V. It can be concluded that the 'Cosine' noise schedule achieves the best performance. And the Quad schedule is

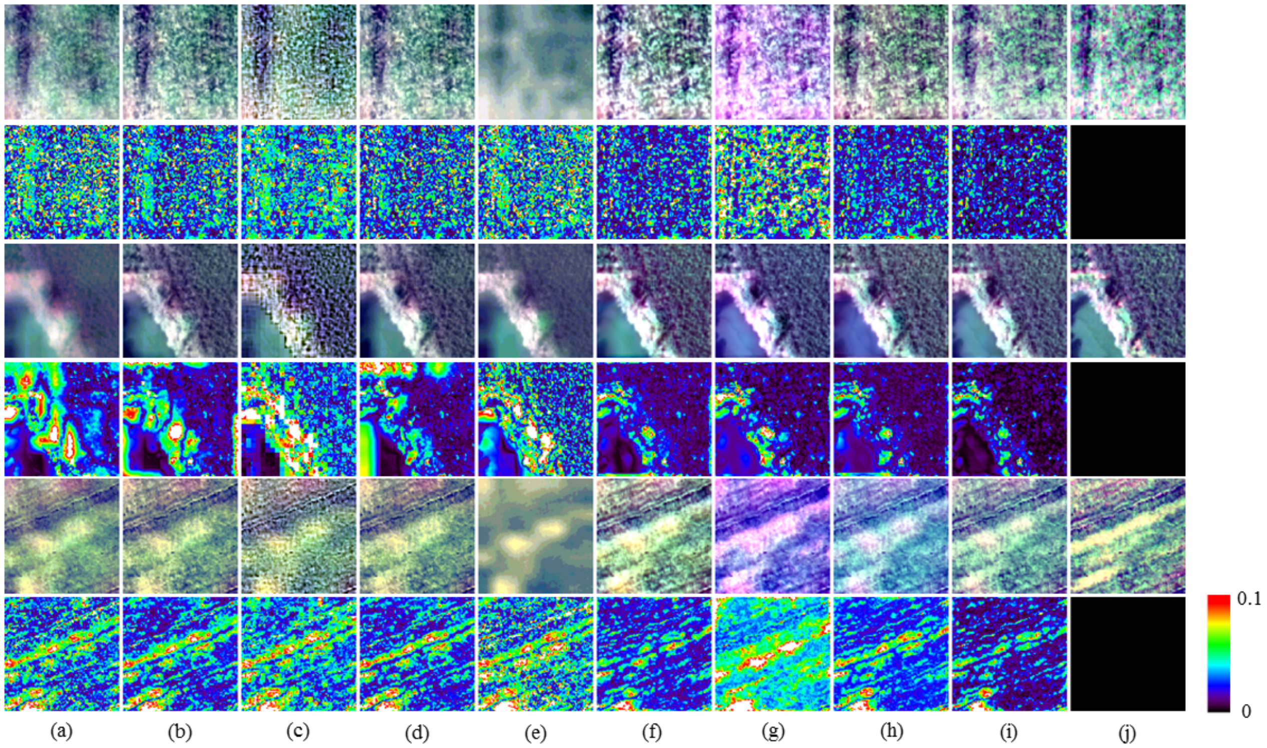


Fig. 11. Fused results (odd row) and error maps (even row) of different methods on the Xiongan dataset. (a) GSA. (b) SFIM. (c) Wavelet. (d) MTF_GLP_HPM. (e) CNMF. (f) HSpeSet2. (g) Pgnet. (h) Srdiff. (i) Ours. (j) ground truth. The RGB image is shown by utilizing the 36-th, 72-th, and 120-th bands of the original HSI.

TABLE VII
AVERAGE QUANTITATIVE RESULT ON ZY DATASET

Method	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
GSA	37.272	0.883	0.039	1.255	0.588
SFIM	37.555	0.879	0.038	1.242	0.625
Wavelet	37.360	0.832	0.041	2.270	0.503
MTF_GLP_HPM	38.038	0.884	0.037	1.170	0.669
CNMF	37.125	0.889	0.038	1.257	0.617
HSpeSet2	40.721	0.933	0.034	0.618	0.805
Pgnet	38.677	0.936	0.038	0.854	0.823
Srdiff	37.333	0.917	0.043	0.831	0.782
Ours	**42.090**	**0.939**	**0.029**	**0.542**	**0.830**

TABLE VIII
AVERAGE QUANTITATIVE RESULT ON CHIKUSEI DATASET

Method	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
GSA	32.255	0.827	0.125	3.563	0.620
SFIM	32.663	0.865	0.130	4.326	0.683
Wavelet	32.168	0.834	0.131	3.908	0.621
MTF_GLP_HPM	33.171	0.871	0.130	4.345	0.720
CNMF	32.129	0.823	0.128	3.520	0.621
HSpeSet2	36.242	0.915	0.108	2.494	0.867
Pgnet	35.717	0.926	0.103	2.336	0.862
Srdiff	36.955	0.929	0.097	2.162	0.864
Ours	**37.258**	**0.933**	**0.092**	**2.071**	**0.877**

slightly inferior to the Cosine schedule.

We also test the different sampling time steps, and the quantitative results are shown in Table VI. It shows that the time step of 1000 achieves the best performance. And with the time step increasing to 1500, the final performance drops severally. This may be since the data range in the sampling process is out of range of the target image when the sampling steps are continuously increased.

E. Comparative Experimental Results on three Datasets

1) ZY Dataset

The comparative experimental results of different methods on the down-sampled ZY dataset are shown in Fig. 9. Due to the severe ill-posedness of the LRHSI and PAN image fusion with the spatial resolution ratio of 12, the traditional methods all suffer from spatial blurring or spectral artifacts. For example, the results of the Wavelet method show the spatial pseudo-detail, while the CNMF method suffers from the spatial blurring effect, as shown in Fig. 9(e). Although DL-based methods all have the improved image quality, some methods such as Pgnet and Srdiff still show spectral distortion, as shown in the third and fifth rows of Fig. 9(g) and (h). And our proposed method achieves the best spatial and spectral fidelity, as shown by the fused results and error map in Fig. 9(i).

The quantitative results in Table VII show that our method ranks first in all five quality indices, demonstrating the

TABLE IX
AVERAGE QUANTITATIVE RESULT ON XIONGAN DATASET

Method	PSNR(↑)	SSIM(↑)	SAM(↓)	ERGAS(↓)	SCC(↑)
GSA	33.649	0.891	0.055	0.825	0.654
SFIM	34.926	0.932	0.050	0.724	0.767
Wavelet	33.410	0.872	0.061	0.925	0.656
MTF_GLP_HPM	35.890	0.937	0.049	0.642	0.816
CNMF	33.302	0.874	0.052	0.846	0.643
HSpeSet2	38.435	0.961	0.048	0.489	0.887
Pgnet	37.654	0.964	0.049	0.518	0.891
Srdiff	39.424	0.966	0.041	0.435	0.899
Ours	**40.630**	**0.972**	**0.035**	**0.365**	**0.918**

TABLE X
AVERAGE INFERENCE TIME ON ZY DATASET

Method	Time(s)(↓)
GSA	0.042
SFIM	0.330
Wavelet	0.049
MTF_GLP_HPM	0.166
CNMF	4.081
HSpeSet2	0.004
Pgnet	0.016
Srdiff	6.276
Ours	5.442

absolute superiority of our method over other methods. For example, the PSNR and SAM indices of our method are better than the sub-optimal results by 3.36% and 14.71%, respectively. These results verify the significance of our method in restoring spatial and spectral detail.

2) Chikusei Dataset

The fusion results on the simulated Chikusei dataset are shown in Fig. 10. This dataset mainly consists of buildings and cultivated land. Traditional methods generally suffer from the spatial blurring effect, especially the GSA and CNMF methods, as shown in Fig. 10(a) and (e). The edges of buildings and roads are very blurred. While the DL-based methods show improved image quality, some methods still have spatial distortions. For example, as shown in Fig. 10(f) and (h), which are the fused results of HSpeSet2 and Srdiff methods, the roads suffer from the spatial distortion problem. In contrast, the proposed method achieves the best spatial detail restoration.

The quantitative results in Table VIII also show the superiority of our method over other methods in five indices. For example, the SAM and ERGAS indices of our method are greater than the sub-optimal results by 5.15% and 4.21%, respectively. Therefore, the qualitative and quantitative results all confirm the absolute superiority of our method.

3) Xiongan Dataset

The fusion results on the simulated dataset—Xiongan dataset are shown in Fig. 11. This dataset mainly consists of cultivated land and its spectral and spatial features are different from the above two datasets. As shown in Fig. 11, the traditional methods suffer from the severe distortion that can be seen from the error maps. And the CNMF method still suffers from the spatial blurring effect. Compared to the traditional methods, the fusion results of the DL-based methods all have improved image quality, which could be clearly inferred from the error maps in the even rows of Fig. 11. And our method achieves the best spatial and spectral fidelity, especially the latter, which is quite significant for practical applications such as crop yield estimation.

The quantitative results also demonstrate the superiority of the proposed method, as shown in Table IX, where our method achieves the first rank in all five indices. Same with the qualitative results in Fig. 11, the traditional methods get the worse rank than the DL-based methods. Overall, the qualitative and quantitative results demonstrate the superiority of the proposed method in improving spectral and spatial fidelity.

F. Comparison of Computational Complexity

In this part, we measure the average inference time of all fusion methods on all test images to compare their fusion efficiency. This time calculation is performed on the ZY dataset with a patch size of 76 × 96 × 96 of HRHSI.

As shown in Table X, the traditional methods cost moderate inference time, ranging from 0.042s (GSA) to 4.081s (CNMF). Due to iterative parameter updating process of the CNMF method, it took more time to fuse the LRHSI and PAN images. In comparison, the DL-based methods that based on the general CNN structure have the immediate inference time, such as HSpeSet2 and Pgnet. Despite the fast inference time of these two methods, their fusion performance is inferior to ours according to the above comparative experimental results. This is due to the iterative image generation process of our method that based on the DDPM, which could results in high quality image with realistic details. And because of the iterative process, as shown in Fig. 3, the DDPM-based fusion methods including Srdiff and our method spend more inference time, as shown in Table X. But compared to another DDPM-based method-Srdiff, our method costs less inference time with a decrease of 0.834s. This is mainly due to that the proposed method performs the image generation process in the low-dimensional residual latent space with the reduced computational cost. Therefore, it can be concluded that our method achieves the better balance between the fusion performance and model complexity than other methods.

VI. CONCLUSION

In this study, to fully restore the realistic HRHSI details and improve its spectral fidelity, we propose a diffusion model-based fusion network to complete the fusion task between the LRHSI and PAN images. To improve the spectral and spatial consistency of the fused HRHSI with the input images, we propose the dual spatial and spectral CDM to improve the spectral and spatial fidelity of the fused HRHSI. In addition, considering the high-dimensional property of the HSI, we construct an auto-encoder to encode the HSI into the low-dimensional latent space, which could save the computational cost of the diffusion model. We also perform the diffusion process on the residual space to facilitate the training and sampling process. Extensive experimental results on three datasets show the superiority of our method over several SOTA methods.

ACKNOWLEDGEMENT

The authors would like to thank Baidu AI Studio for providing computing power support. And thank other authors for their assistance in editing, refining, and finalizing the manuscript.

REFERENCES

[1] D. Manolakis and G. Shaw, “Detection algorithms for hyperspectral imaging applications,” *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[2] Q. Zhu, W. Deng, Z. Zheng, Y. Zhong, Q. Guan, W. Lin, L. Zhang, and D. Li, “A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification,” *IEEE Transactions on Cybernetics*, pp. 1–15, 2021.

[3] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, “Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing,” *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2991–3006, 2019.

[4] K. Zheng, L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui, and J. Chanussot, “Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2487–2502, 2021.

[5] N. Yokoya, J. C.-W. Chan, and K. Segl, “Potential of resolution-enhanced hyperspectral data for mineral mapping using simulated enmap and sentinel-2 images,” *Remote Sensing*, vol. 8, no. 3, 2016.

[6] V. Salomonson, W. Barnes, P. Maymon, H. Montgomery, and H. Ostrow, “Modis: advanced facility instrument for studies of the earth as a system,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 2, pp. 145–153, 1989.

[7] J. Yu, D. Liang, B. Han, and H. Gao, “Study on ground object classification based on the hyperspectral fusion images of ZY-1(02D) satellite,” *Journal of Applied Remote Sensing*, vol. 15, no. 4, p. 042603, 2021.

[8] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, “Bayesian fusion of multi-band images,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1117–1127, 2015.

[9] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, “A convex formulation for hyperspectral image superresolution via subspace-based regularization,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2015.

[10] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, “Hyperspectral pansharpening using deep prior and dual attention residual network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[11] J. Qu, S. Hou, W. Dong, S. Xiao, Q. Du, and Y. Li, “A dual-branch detail extraction network for hyperspectral pansharpening,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] W. G. C. Bandara and V. M. Patel, “Hypertransformer: A textural and spectral feature fusion transformer for pansharpening,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1767–1777.

[13] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simões, J.-Y. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya, “Hyperspectral pansharpening: A review,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 27–46, 2015.

[14] S. Li, Y. Tian, C. Wang, H. Wu, and S. Zheng, “Hyperspectral image super-resolution network based on cross-scale nonlocal attention,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[15] R. Dian, S. Li, and X. Kang, “Regularizing hyperspectral and multispectral image fusion by cnn denoiser,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1124–1135, 2021.

[16] S. Li, R. Dian, and H. Liu, “Learning the external and internal priors for multispectral and hyperspectral image fusion,” *Science China Information Sciences*, vol. 66, no. 4, p. 140303, 2023.

[17] R. Dian, A. Guo, and S. Li, “Zero-shot hyperspectral sharpening,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 650–12 666, 2023.

[18] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and B. Li, “Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3092–3100, 2019.

[19] L. He, J. Zhu, J. Li, D. Meng, J. Chanussot, and A. Plaza, “Spectral-fidelity convolutional neural networks for hyperspectral pansharpening,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5898–5914, 2020.

[20] J. Qu, Z. Xu, W. Dong, S. Xiao, Y. Li, and Q. Du, “A spatio-spectral fusion method for hyperspectral images using residual hyper-dense network,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[21] D. He and Y. Zhong, “Deep hierarchical pyramid network with high-frequency -aware differential architecture for super-resolution mapping,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[22] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022.

[23] H. Chung, E. S. Lee, and J. C. Ye, “Mr image denoising and super-resolution using regularized reverse diffusion,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 922–934, 2023.

[24] Y. Li, H. Huang, L. Jia, H. Fan, and S. Liu, “D2c-sr: A divergence to convergence approach for real-world image super-resolution,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 379–394.

[25] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

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

[30] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing," *Remote Sensing*, vol. 14, no. 19, 2022.

[31] Z. Cao, S. Cao, X. Wu, J. Hou, R. Ran, and L.-J. Deng, "Ddrf: Denoising diffusion model for remote sensing image fusion," *arXiv preprint arXiv:2304.04774*, 2023.

[32] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 461–11 471.

[33] H. Sahak, D. Watson, C. Saharia, and D. Fleet, "Denoising diffusion probabilistic models for robust image super-resolution in the wild," *arXiv preprint arXiv:2302.07864*, 2023.

[34] C. Wu, D. Wang, H. Mao, and Y. Li, "Hsr-diff:hyperspectral image super-resolution via conditional diffusion models," 2023.

[35] S. Shi, L. Zhang, and J. Chen, "Hyperspectral and multispectral image fusion using the conditional denoising diffusion probabilistic model," 2023.

[36] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms +pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, 2007.

[37] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at ihs-like image fusion methods," *Information Fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[38] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.

[39] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1204–1211, 1999.

[40] "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Engineering Remote Sensing*, vol. 72, no. 5, 2006.

[41] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2012.

[42] R. Hardie, M. Eismann, and G. Wilson, "Map estimation for hyperspectral image resolution enhancement using an auxiliary sensor," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1174–1184, 2004.

[43] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[44] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[45] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[46] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4469–4480, 2020.

[47] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5135–5146, 2019.

[48] S. Li, Y. Tian, C. Wang, H. Wu, and S. Zheng, "Cross spectral and spatial scale non-local attention-based unsupervised pansharpening network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4858–4870, 2023.

[49] W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, "Hyperspectral pansharpening based on improved deep image prior and residual reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[51] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[52] A. Niu, K. Zhang, T. X. Pham, J. Sun, Y. Zhu, I. S. Kweon, and Y. Zhang, "Cdpmsr: Conditional diffusion probabilistic models for single image super-resolution," 2023.

[53] X. Rui, X. Cao, Z. Zhu, Z. Yue, and D. Meng, "Unsupervised pansharpening via low-rank diffusion model," 2023.

[54] Q. Meng, W. Shi, S. Li, and L. Zhang, "Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[55] J. Leinonen, U. Hamann, D. Nerini, U. Germann, and G. Franch, "Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification," 2023.

[56] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8162–8171.

[57] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *CoRR*, vol. abs/2011.13456, 2020.

[58] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *CoRR*, vol. abs/2112.07804, 2021.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[60] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[61] N. Yokoya and A. Iwasaki, “Airborne hyperspectral data over chikusei,” 05 2016.

[62] Y. Cen, L. Zhang, X. Zhang, Y. Wang, W. Qi, S. Tang, and P. Zhang, “Aerial hyperspectral remote sensing classification dataset of xiongan new area (matiwan village),” *J. Remote Sens*, vol. 24, no. 11, pp. 1299–1306, 2020.

[63] S. Li, Y. Tian, H. Xia, and Q. Liu, “Unmixing-based pan-guided fusion network for hyperspectral imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, and E. P. Simoncelli, “Image quality assessment: From error measurement to structural similarity,” vol. 13, no. 4, pp. 600–612, 2003.

Shuangliang Li received the B.S. degree in Geographical Science from Hubei University, Wuhan, China, in 2019. And received the M.S. degree in Photogrammetry and Remote Sensing from China University of Geosciences, Wuhan, China, in 2023. He is now pursuing the Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University, Wuhan, China. His research interests include hyperspectral image processing, image fusion, and deep learning.

Siwei Li received Ph.D. degree in atmospheric sciences with focus on remote sensing from University at Albany, State University of New York State, Albany, NY, USA, in 2012. He is a professor of atmospheric remote sensing with school of remote sensing and information engineering, Wuhan University, Wuhan, China. He works on active and passive atmospheric remote sensing, fast atmospheric radiative transfer model and atmospheric correction, observation, simulation and investigation on air pollution and climate change.

Lihao Zhang received the B.E. degree in remote sensing science and technology and the M.S. degree in photogrammetry and remote sensing from the China University of Geosciences, Wuhan, China, in 2018 and 2021, respectively. He is now pursuing the Ph.D. degree in Cartography and Geographic Information System from Beijing Normal University, Beijing, China. His research interests include ecological remote sensing, remote sensing image processing and application.