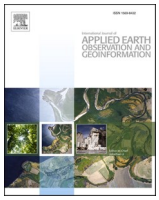
International Journal of Applied Earth Observations and Geoinformation 109 (2022) 102773

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



DsTer: A dense spectral transformer for remote sensing spectral super-resolution

Jiang He [a], Qiangqiang Yuan [a], Jie Li [a,*], Yi Xiao [a], Xinxin Liu [b,c], Yun Zou [d]

[a] School of Geodesy and Geomatics, Wuhan University, Hubei, China
[b] College of Electrical and Information Engineering, Hunan University, Hunan, China
[c] Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan, China
[d] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Hubei, China

ARTICLE INFO

Keywords:
Spectral super-resolution
Transformer
Multispectral imaging
Hyperspectral imaging

ABSTRACT

To obtain high-resolution hyperspectral data, spectral super-resolution is a popular computational imaging technique directly from high-resolution multispectral images. Besides sparse recovery, deep learning-based methods perform well in the past years for their powerful nonlinear mapping from multispectral to hyperspectral domains. However, convolutions in deep learning only focus on local information and have been blamed for the neglect of long-range relationships. Nowadays, transformer has been attracting great interest for its global attention to long-range interaction. In this study, we propose a dense spectral transformer with ResNet to achieve spectral super-resolution for multispectral remote sensing images. Combining transformer with ResNet meets the need for 3D data handling to remote sensing images as well as learning long-range relationships. Dense connection helps model exploit features from multi-level transformers. Moreover, spectral recovery results on natural data and three remote sensing data sets all prove the advantage of the proposed model. Furthermore, we also carry out classification experiments on real data to verify the dependability of the reconstructed spectra.

1. Introduction

Hyperspectral images contain much more spectral information than the traditional RGB or multispectral images, which are usually acquired by hyperspectral sensors with numerous bands in a fine spectral resolution (Melgani and Bruzzone, 2004; Sun et al., 2020). Benefiting from such rich spectral details, hyperspectral images can easily reflect different radiance properties of the same scene and have been used in numerous applications, such as remote sensing (Bioucas-Dias et al., 2013; Wang et al., 2021), food safety (Gowen et al., 2007), atmosphere monitoring (Barnsley et al., 2004; Wang et al., 2022), and medical imaging (Lu and Fei, 2014; Shao et al., 2021).

However, remote sensing imagery with rich spectral information also comes with high acquisition costs and suffers low spatial resolution, which limits further development in fine applications (Xiao et al., 2021a; Dian and Li, 2019; Hong et al., 2020). In contrast, multispectral sensors always capture high-spatial-resolution images which are usually available for free but with only several channels, in other words, with low spectral resolution(Jin et al., 2022; Hu et al., 2021). To obtain hyperspectral images with high spatial resolution, many researchers (Wei et al., 2015; Dalponte et al., 2008; Yokoya et al., 2012; Dian et al., 2020) utilize high-resolution multispectral images as auxiliary data to achieve data fusion. However, the hyperspectral data corresponding to the high-resolution multispectral images are not easily available(Xiao et al., 2021b). Even if we could acquire the corresponding multispectral and hyperspectral images, the registration and preprocessing are also knotty and would further affect the algorithm accuracy. To overcome the problems mentioned above, spectral super-resolution is proposed by enhancing the spectral resolution of high-resolution multispectral data without extra hyperspectral data.

There have been many researches on spectral super-resolution in the past few decades (Fu et al., 2020; White et al., 2018; Song et al., 2021; Fu et al., 2021; Yang et al., 2021; Li et al., 2022; Liu et al., 2021). In the early days, many researchers recover hyperspectral images using sparse representation. Nguyen et al. achieve scene spectral reconstruction from RGB images through a training methods (Nguyen et al., 2014). Then, Robles-Kelly utilized constrained sparse coding to exploit color and appearance information and achieved spectral super-resolution with the

* Corresponding author.
E-mail addresses: hej96.work@gmail.com (J. He), yqiang86@gmail.com (Q. Yuan), jli89@sgg.whu.edu.cn (J. Li), xiao_yi@whu.edu.cn (Y. Xiao), liuxinxin@hnu.edu.cn (X. Liu), yunzou@hust.edu.cn (Y. Zou).

https://doi.org/10.1016/j.jag.2022.102773
Received 1 January 2022; Received in revised form 2 March 2022; Accepted 4 April 2022
Available online 30 April 2022

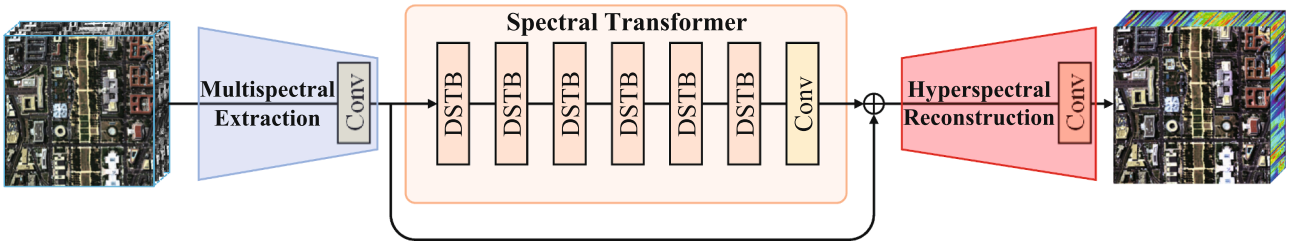


Fig. 1. Framework of the proposed DsTer.

help of a prototype training set (Robles-Kelly, 2015). Arad and Ben-Shahar built a large hyperspectral dataset about natural scenes and employed K-means Singular Value Decomposition to extract hyperspectral dictionary (Arad and Ben-Shahar, 2016). Jia et al. changed the three-to-many mapping problem in spectral super-resolution to a three-to-three one and recover hyperspectral vector using manifold method (Jia et al., 2017). Aeschbacher and Wu et al. transformed from their proposed A + in spatial super-resolution, proposed a similar method to increase the bands of RGB data (Wu et al., 2017). Akhtar et al. employed Gaussian processes to optimize the hyperspectral dictionary and further model natural spectra (Akhtar and Mian, 2020). The main idea of these types of methods involves three steps. Firstly, extract dictionary from a hyperspectral data set. Secondly, calculate the sparse coefficients on the test high-resolution multispectral images. Finally, utilize multispectral coefficients and hyperspectral dictionary to recover high-resolution hyperspectral images. The framework is similar to spectral unmixing. Nevertheless, spectra of pure pixels can hardly be learned adequately through a finite hyperspectral image set (Hang et al., 2021).

Deep learning is recently becoming more and more popular for its huge advantage in feature extraction and nonlinear mapping and has shown great advantages compared with most traditional approaches in areas, such as classification (Wambugu et al., 2021; Kussul et al., 2017), segmentation (Luo et al., 2021; Kampffmeyer et al., 2016), denoising (Yu and Chen, 2014; Yuan et al., 2019), missing reconstruction (Li et al., 2020; Zhang et al., 2020b, 2021b,a), and spatial super-resolution (Muad and Foody, 2012; Chen et al., 2020). Improving the semantic segmentation architecture Tiramisu (Jegou et al., 2017), Galliani et al. proposed a deep dense Unet (Galliani et al., 2017). Further, Rangnekar et al. employed generative learning trained on their own aerial hyperspectral dataset, *AeroCampus*, to learn the mapping from a RGB image to 31 spectral bands (Rangnekar et al., 2017). Xiong et al. explored different network frameworks and utilized a 20-layer deep convolutional neural network (CNN) with residual connection to achieve good recovery fidelity (Xiong et al., 2017). Replacing the residual block with the dense block and a fusion scheme, Shi et al. proposed HSCNN + and achieved better performance (Shi et al., 2018). Fu et al. presented a spatial-spectral CNN to simultaneously select spectral response function and recover hyperspectral image from a single RGB image (Fu et al., 2018). Meanwhile, Nie et al. used a 1×1 convolution to simulate spectral response functions and recover hyperspectral images from RGB images using Unet (Nie et al., 2018). Can et al. designed a 9-layer residual CNN to prove that moderately deep learning can also perform well in spectral super-resolution (Can and Timofte, 2018). Currently, Zhang et al. used multi-scale kernels and proposed a multi-column network to adaptively fuse features on each pixel (Zhang et al., 2020a). In our previous works (He et al., 2021), we proposed a deep optimization-driven network with spectral response functions as a guide for spectral super-resolution combining the physical model with CNN, which achieves ideal performance than CNN-based models.

Although the performance is significantly improved compared with traditional model-based methods (He et al., 2022), they generally suffer from several problems. First, using the same convolution kernel to restore different ground objects may not be the best choice. Second, convolution can hardly consider the influence of long-range pixels. Nowadays, *Transformer* (Vaswani et al., 2017), with a self-attention mechanism, shows a strong ability for global information capture and long-range interaction on similar spectra. Therefore, transformer has shown promising performance in many remote sensing image processing, such as segmentation (Dosovitskiy et al., 2020; Hong et al., 2021) and image restoration (Liang et al., 2021). However, there are two new problems that transformer brings. The first is that the previous works still keep the Multi-Layer Perceptron (MLP) in transformer, which is originally designed for machine translation and not suitable to processing 3D images. Second, the efficiency of image restoration, as a low-level vision task, will be slowed down by the huge structure of transformers.

In this paper, especially for remote sensing multispectral images, we propose a **D**ense **s**pectral **T**ransform**er** (**DsTer**) to boost the spectral super-resolution. Focusing on the spatial-spectral feature in multispectral images, we replace MLP with ResNet to deal with 3D remote sensing data. Furthermore, the proposed dense transformer with bottlenecks similar to DenseNet (Huang et al., 2017) can fully exploit features from multi-level transformers and does not acquire more parameters. The main contributions of this paper are as follows:

- This paper is an early attempt that utilizes transformer to recover hyperspectral images from multispectral images. Moreover, embedding convolutions through the proposed transformer-based model ensures DsTer to capture long-range interactions as well as consider local information.
- Aiming at the spatial-spectral features in multispectral remote sensing images, we use ResNet rather than MLP in transformer to extract 3D features and achieve good performance. ResNet can easily explore features from the spectral dimension in remote sensing data while MLP designed for discrete data can hardly capture spatial-spectral features.
- Features from multi-level transformers are equally important in spectral super-resolution. Improved by a dense strategy, multi-level features can be exploited in the proposed DsTer and the model would achieve good spectral recovery. Also with the bottleneck layer, it doesn't take a longer time to run DsTer.

2. Methodology

2.1. Problem formulation

Given a high-resolution multispectral image $\boldsymbol{M} \in R^{W \times H \times c}$ (where W and H are the image width and height of multiepectral image, and c denotes the channel number), spectral super-resolution is to recover the desired high-resolution hyperspectral image $\boldsymbol{H} \in R^{W \times H \times C}$, where $C > c$ is the channel number of hyperspectral images. Given a specific satellite sensor, $\boldsymbol{\Phi} \in R^{c \times C}$ that represents the spectral transformation between multispectral and hyperspectral imaging is also fixed:

$$\boldsymbol{M} = \boldsymbol{\Phi H} \tag{1}$$

Rather than to build a complete hyperspectral dictionary, the aim of deep learning-based spectral super-resolution is to find a mapping function $f(:, \theta)$ to recover $\boldsymbol{H}$ from $\boldsymbol{M}$. To achieve this goal is equal to

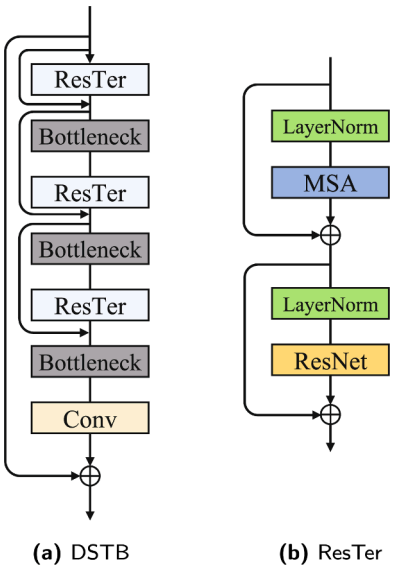


Fig. 2. Dense Spectral Transformer Block (DSTB) and ResNet-based Transformer Layer (ResTer) in spectral transformer.

minimize the following objective function:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathscr{L}(f(\boldsymbol{M}, \theta), \boldsymbol{H}) \tag{2}$$

where $\mathscr{L}(:,:)$ presents the employed loss function in training, such as L1 loss, L2 loss, or perceptual loss.

2.2. Network architecture

In this paper, we proposed a ResNet-based transformer with dense strategy to deal with spectral super-resolution, which consists of three main parts, including multispectral feature extraction, spectral transformer, and hyperspectral reconstruction, as shown in Fig. 1. Convolution layer provides a simple and convenient way to increase the feature dimension, which has shown stable performance in many tasks. To extract multispectral feature $F_M \in R^{W\times H\times C_{mid}}$ for subsequent spectral transformation, we use a 3×3 convolutional layer $f_{ME}(\cdot)$ as the first module in DsTer:

$$F_M = f_{ME}(\boldsymbol{M}) \tag{3}$$

where C_{mid} is the feature number. After the multispectral feature extraction is the proposed spectral transformer f_{ST} with residual learning to map multispectral feature to hyperspectral feature $F_H \in R^{W\times H\times C_{mid}}$:

$$F_H = f_{ST}(F_M) + F_M \tag{4}$$

where $f_{ST}(\cdot)$ denotes the proposed spectral transformer which contains n dense spectral transformer blocks (DSTB) and a convolution. Specifically, the relationship between intermediate features $F_1,\ F_2, \cdots,\ F_n$ and the hyperspectral feature F_H can be shown as:

$$\begin{aligned} F_1 &= \mathrm{DSTB}_1(F_M) \\ F_i &= \mathrm{DSTB}_i(F_{i-1}),\quad i = 1, 2, \cdots, n \\ F_H &= \mathrm{Conv}(F_n) + F_M \end{aligned} \tag{5}$$

where $\mathrm{DSTB}_i(\cdot)$ denotes the i-th dense spectral transformer block and $\mathrm{Conv}(\cdot)$ presents the final convolution. With residual learning, the proposed spectral transformer can focus on the difference between multispectral features and hyperspectral features. Moreover, using a convolutional layer at the end of feature mapping can bring the local information of the convolution into the transformer-based algorithm, and lays a better foundation for the integration of shallow and deep features. Obtaining transformed hyperspectral features, the proposed DsTer subsequently employs a spectral convolution to reconstruct hyperspectral images $\widehat{\boldsymbol{H}}$.

$$\widehat{\boldsymbol{H}} = f_{HR}(F_H) \tag{6}$$

Only considering feature extraction, feature transformation, and image reconstruction, the proposed DsTer can be written as:

$$f(\boldsymbol{M}, \theta) = f_{HR}(f_{ME}(\boldsymbol{M}) + f_{ST}(f_{ME}(\boldsymbol{M}))) \tag{7}$$

Employing L1 loss function to train model could effectively enhance the extraction of spatial details, and the objective function of optimizing the proposed DsTer is

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \| f_{HR}(f_{ME}(\boldsymbol{M}) + f_{ST}(f_{ME}(\boldsymbol{M}))) - \boldsymbol{H} \|_1 \tag{8}$$

2D convolution extract multispectral features preliminarily, and spectral transformer transform the low-dimensional feature to hyperspectral features with dense spectral transformer blocks. In the end, a spectral convolution with kernel size of 1 is employed to fuse features and recover hyperspectral images. The whole network is full of the combination between CNN and transformer.

2.3. Dense spectral transformer block

In this study, with a strong ability of feature mapping, transformer is employed to transform multispectral features to hyperspectral features. To helps model make full use of multi-level features from different transformer layers and boost computational speed, we proposed a spectral transformer consisting of several dense spectral transformer blocks and a bottleneck convolution.

As shown in Fig. 2a, a dense spectral transformer block contains ResNet-based transformers and a convolution with dense strategy. Given the input feature F_i^0 of the i-th dense spectral transformer block with k ResNet-based transformers,

$$\begin{aligned} F_i^1 &= \mathrm{ResTer}_1(F_i^0) \\ F_i^j &= \mathrm{ResTer}_j(Bottle(Cat(F_i^0, \cdots, F_i^{j-1})),\quad j = 2, 3, \cdots, k \\ F_{i+1}^0 &= \mathrm{Conv}(Bottle(Cat(F_i^0, \cdots, F_i^k)) + F_i^0 \end{aligned} \tag{9}$$

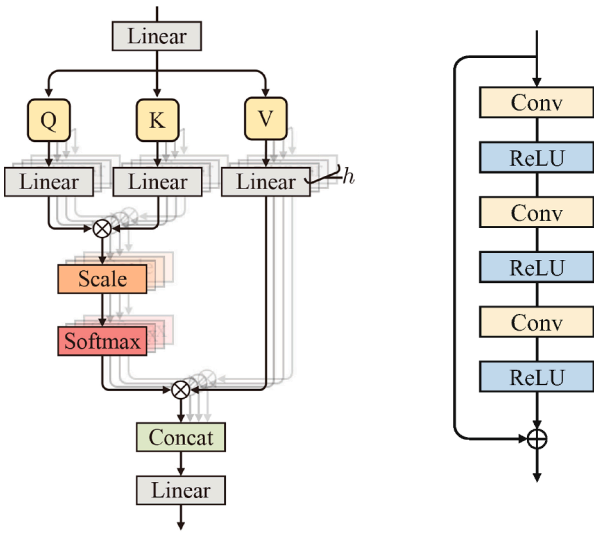
where $\mathrm{ResTer}_j(\cdot)$ denotes the j-th ResNet-based transformer, $Bottle(\cdot)$ and $Cat(\cdot)$ present the bottleneck and concatenation in dense strategy, respectively. Concatenation helps DsTer competitively exploit features from multi-level transformer layers, meanwhile, bottleneck keeps the high efficiency of transformer by feature compression.

At the end of every DSTB, we employ a convolution layer to achieve that features are alternately fed into CNN and transformer in the whole DsTer, which combines the local information extraction of CNN with the long-range dependency modeling of transformer.

With dense strategy, the proposed dense spectral transformer block can not only exploit features from multi-level transformer layers but also keeps a good computational speed. Furthermore, the convolution-based bottleneck between ResNet-based transformer layers can further strengthen the cooperation between CNN and transformer (Dosovitskiy et al., 2020; Touvron et al., 2021; Xiao et al., 2022).

2.4. ResNet-based transformer layer

Transformer is famous for the self-attention mechanism to capture global interactions between contexts, which always contains self-attention, layer normalization, and feed-forward modules as shown in Fig. 2b. However, traditional transformers utilize MLP as feed-forward module, which is designed for natural language processing.To extract 3D spatial-spectral features in remote sensing images, this paper employs ResNet as feed-forward module. Moreover, layer normalization is employed to keep the stability of feature distribution for quick training speed. Self-attention (Bello et al., 2019; Zhao et al., 2020; Sinha and Dolz, 2020) is an attention mechanism to learn the global relationship



(a) Multi-head self-attention (MSA) **(b)** ResNet

Fig. 3. MSA and ResNet in ResNet-based transformer layer.

Table 1
Information about training samples and test samples. Specifically, $n \times c \times w \times h$ denotes n image patches with c channels and the size is $w \times h$.

	Input MSI	Output HSI	Training	Test
CAVE	512 × 3 × 128 × 128	512 × 31 × 128 × 128	416	96
Chikusei	1016 × 4 × 128 × 128	1016 × 32 × 128 × 128	888	128
Xiong'an	5184 × 4 × 64 × 64	5184 × 32 × 64 × 64	4672	512
DC Mall	2424 × 4 × 64 × 64	2424 × 32 × 64 × 64	2168	256

Table 2
Quantitative results on natural data set. Results in bold are best and the underlined are second best.

Models	CC	mPSNR	mSSIM	SAM
Arad	0.9486	24.4613	0.7913	21.3129
A+	0.9873	32.8830	0.9297	20.5403
DenseU	0.9907	32.5510	0.9642	8.1915
CanNet	0.9925	33.5975	0.9685	8.6435
HSCNN+	0.9934	34.4354	0.9766	7.8048
sRCNN	0.9916	34.3669	0.9731	9.0175
HSRnet	<u>0.9935</u>	<u>34.4903</u>	<u>0.9771</u>	<u>7.6208</u>
DsTer	**0.9938**	**34.6626**	**0.9799**	**7.3832**

within an input sequence.

In this paper, rather than using a single self-attention, we introduce multi-head self-attention into ResNet-based transformer as shown in Fig. 3a. Given an input feature Z, firstly, a linear layer is employed to embed features into query Q, key K, and value V, which is important to compute self-attention.

$$Q = P^Q(Z), K = P^K(Z), V = P^V(Z) \tag{10}$$

where P^Q, P^K, and P^V are projection weights in linear layer. Then, we further project the queries, keys, and values linearly h times and perform self-attention in parallel to produce multiple heads.

$$\begin{aligned} head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \\ Attention(Q,K,V) &= Softmax(\frac{QK^T}{\sqrt{d}})V \end{aligned} \tag{11}$$

where W_i^Q, W_i^K, and W_i^V present weights in different projections for the i-th head, and $\sqrt{d}$ is a scaling factor. The results are integrated through a fusion convolution for the final multi-head self-attention:

$$MSA(Q,K,V) = Cat(head_1, \cdots, head_h)W^F \tag{12}$$

where W^F denotes the weights of the final Linear layer, which is replaced with a 1×1 convolution in this work. After obtaining the results of multi-head self-attention, traditional transformer often uses multi-layer perceptron (MLP) for further feature transformation. However, the traditional transformer is designed for natural language translation, where the features are in vectors and with weak local information.

In remote sensing, especially for remote sensing image processing, pixels in images always present ground objectives. According to *Tobler's First Law*: everything is related to everything else, but near things are more related to each other (Tobler, 1970). In other words, the local information in remote sensing multispectral images is also important. For this reason, after calculating multi-head self-attention, DsTer explores local information by ResNet with 2D convolution, details are shown in Fig. 3b.

With multi-head self-attention, layer normalization, and ResNet, the proposed ResNet-based transformer layer designed for remote sensing images can be formulated as

$$\begin{aligned} F_{msa} &= MSA(LN(F_{in})) + F_{in} \\ F_{out} &= ResNet(LN(F_{msa})) + F_{msa} \end{aligned} \tag{13}$$

where $LN(\bullet)$ is the layer normalization. In the proposed ResNet-based transformer layer, multi-head self-attention learns the global relationship between long-range interactions, ResNet explores detailed features from local range. Moreover, layer normalization keeps the stability of feature distribution and achieve more quick convergence.

3. Results and discussions

3.1. Experimental setting

To verify the superiority of DsTer, we select six algorithms as comparison methods, including Arad (Arad and Ben-Shahar, 2016), A+ (Wu et al., 2017), DenseUnet (Galliani et al., 2017), CanNet (Can and Timofte, 2018), HSCNN+ (Shi et al., 2018), sRCNN (Gewali et al., 2019), and HSRnet (He et al., 2021). Arad and A + are two classical methods based on spare representation and only compared on natural data set. The others are based on deep learning. DenseUnet is a very early attempt using deep learning to solve spectral super-resolution, and CanNet is famous for its lightweight model and good performance. Moreover, HSCNN + was the winner of the *New Trends in Image Restoration and Enhancement* (NTIRE 2018) challenge on spectral reconstruction from RGB images (Arad et al., 2022). At last, HSRnet is an early work combining model-driven with deep learning.

To quantitatively compare method performance in spectral super-resolution, we calculate three indexes to assess spatial fidelity and one for spectral distortion. Evaluation indexes in spatial domain involve correlation coefficient (CC), mean peak signal-to-noise ratio (mPSNR), and mean structural similarity (mSSIM) (Wang et al., 2004). Meanwhile, spectral angle mapper (SAM) (Kruse et al., May 1993) is used to compare the spectral distortion generated by all comparison methods. Note that, mPSNR is in decibel units and SAM is in degree. The CC, mPSNR, and mSSIM indicate better spatial fidelity. Moreover, the lower SAM means the better spectral recovery.

To keep a balance between accuracy and computation time, in the proposed DsTer, we set $C_{mid} = 90$, the number of DSTB $n = 4$, and employ $k = 6$ in a DSTB. Moreover, the multi-head number h is set to 6 for all data. To train our DsTer, we used the adaptive moment estimation (Adam) with 0.001 learning rate. For CNN-based algorithms, models are all coded in the Pytorch framework. The models are trained on a deep learning server equipped with a Nvidia RTX A5000 GPU and the memory size of RAM is 32 GB. Moreover, the comparison models are all

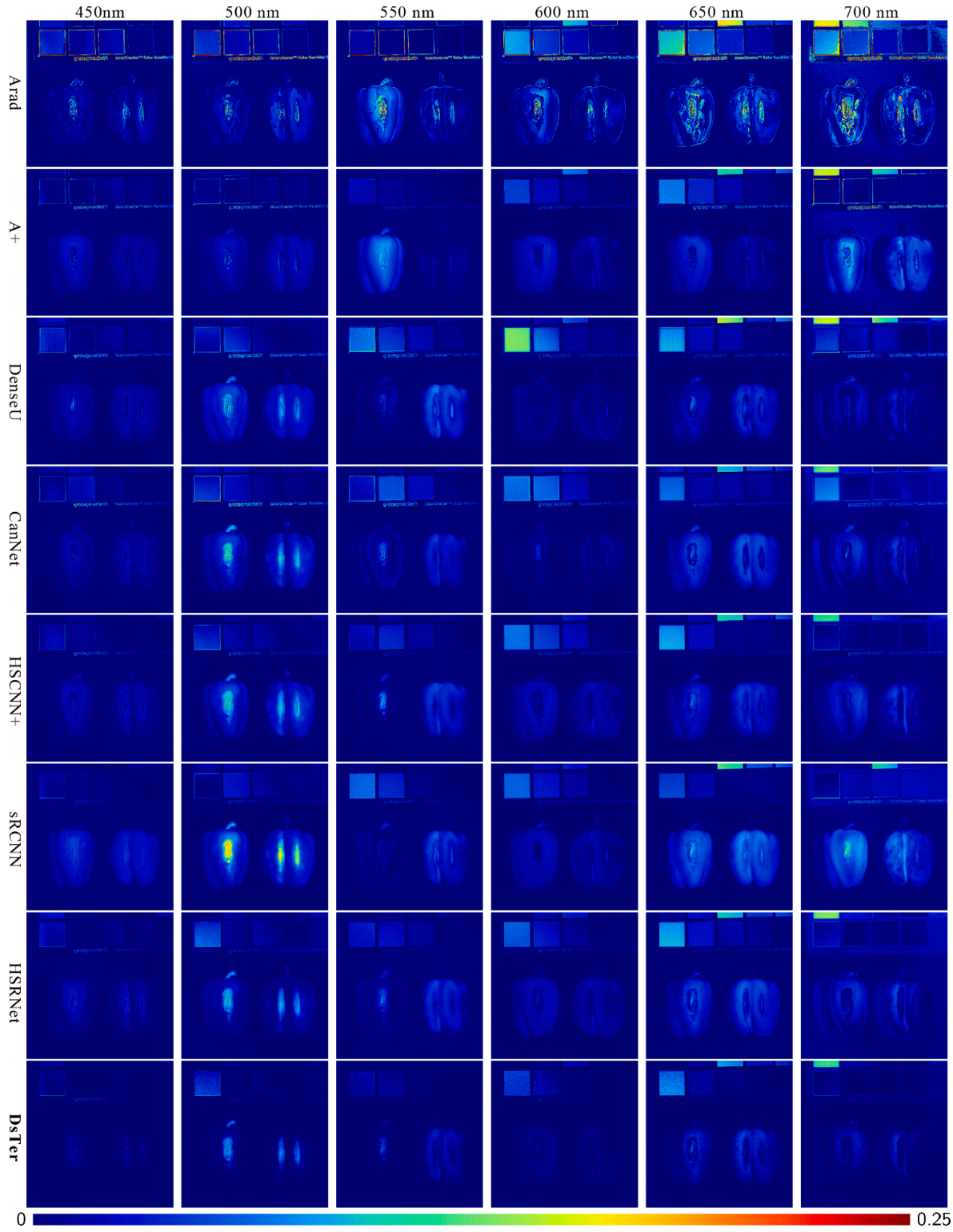


Fig. 4. Visual results of error maps in the CAVE data set.

retrained as optimally as possible by adjusting hyperparameters. For dictionary learning-based methods, we use a program coded by Wu et al. (Wu et al., 2017) to reproduce the models of A + and Arad.

3.2. Dataset description

3.2.1. Natural data set

CAVE dataset (Yasuma et al., 2010) is a popular natural hyperspectral image dataset and contains 32 pairs hyperspectral images with corresponding RGB images. Images in CAVE are all with a seize of 512 × 512 × 31 and cover the spectra from 400 nm to 700 nm. The objects photographed include food, paints, toys, cloths, and human face. We randomly select 26 pairs images in CAVE for model training and the rest is for test. Besides, the batchsize of CAVE training samples is set to 128 × 128. Note that, we cut each image with overlap area and get 25 patches.

Table 3
Quantitative results on Xiong'an, Washington DC Mall, and Chikusei data sets. Results in bold are best and the underlined are second best.

Models	Chikusei				Xiong'an				Washington DC Mall			
	CC	mPSNR	mSSIM	SAM	CC	mPSNR	mSSIM	SAM	CC	mPSNR	mSSIM	SAM
Arad	0.8164	28.3686	0.7153	9.9773	0.8660	26.7137	0.7448	5.7314	0.7895	24.2044	0.5905	6.0803
A+	0.8314	29.9492	0.7791	9.0796	0.8879	31.7293	0.8699	3.6564	0.9945	40.8056	0.9859	1.9051
DenseU	0.9897	39.2096	0.9809	4.0650	0.9847	42.4634	0.9814	0.9217	0.9927	39.7343	0.9848	1.8808
CanNet	0.9967	44.2579	0.9933	3.6732	0.9946	48.3492	0.9950	0.8029	0.9987	47.8736	0.9971	1.1805
HSCNN+	0.9947	42.5542	0.9908	3.4254	0.9942	48.4972	0.9959	0.7888	0.9986	47.5770	0.9972	1.0983
sRCNN	0.9952	44.0295	0.9931	3.5714	0.9954	49.8814	0.9973	0.7623	0.9989	48.5363	0.9978	1.0179
HSRnet	0.9968	44.7133	0.9941	3.4528	0.9963	50.7362	0.9973	0.7196	0.9992	50.4457	0.9983	0.9395
DsTer	**0.9975**	**45.5546**	**0.9954**	**3.2791**	**0.9971**	**51.0023**	**0.9980**	**0.6986**	**0.9994**	**50.5205**	**0.9986**	**0.8553**

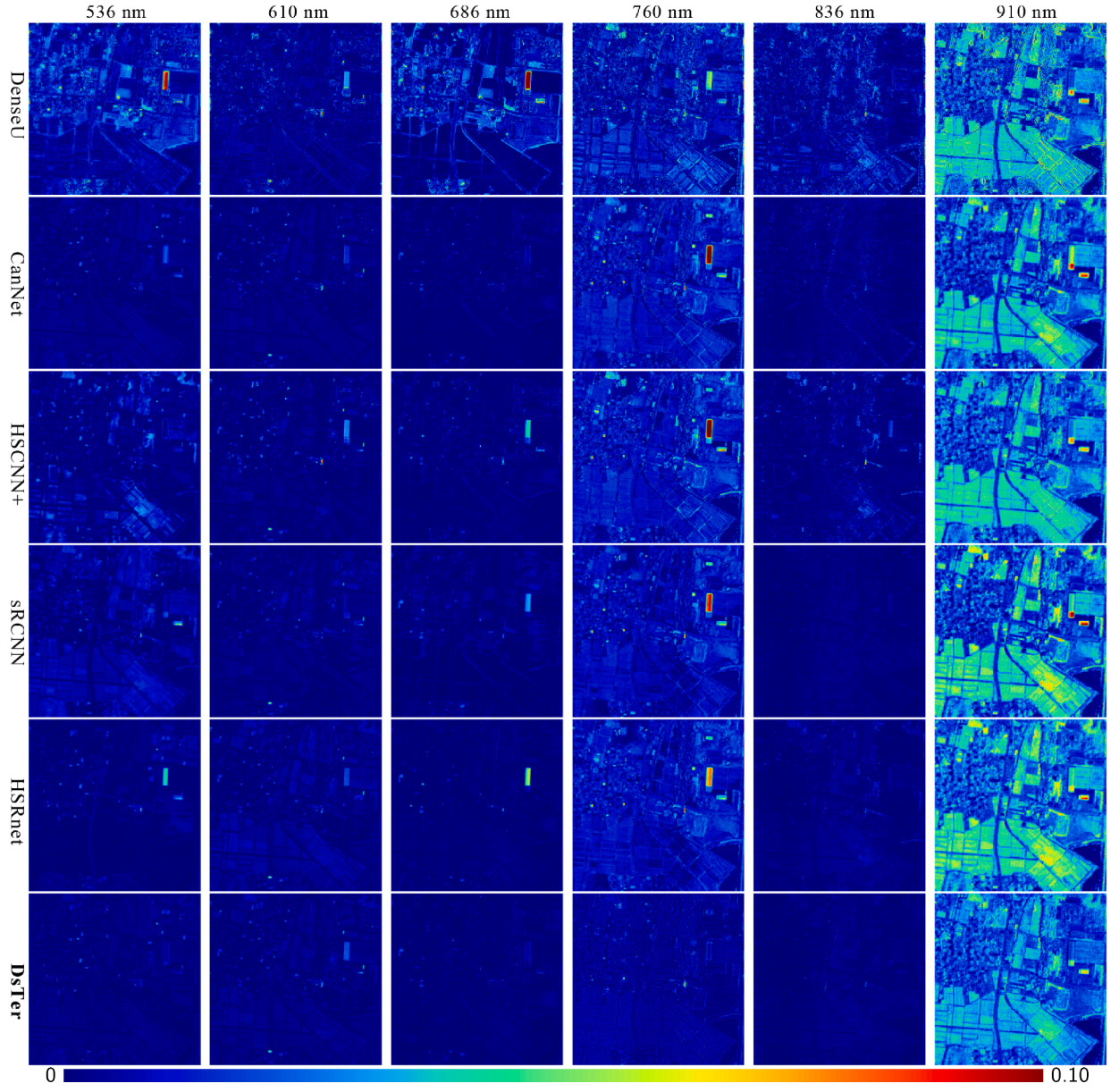


Fig. 5. Visual results of error maps on a randomly selected image in Chikusei data set. From the top to the bottom: error maps generated by DenseUnet, CanNet, HSCNN+, sRCNN, HSRnet, and by the proposed DsTer.

3.2.2. Remote sensing data set

Multispectral data: Sentinel-2 captures images covering 13 bands from 433 nm to 2280 nm with three spatial resolutions, including 10, 20, and 60 m. With two satellites, Sentinel-2 can collect data covering the same area every 5 days. Thus, competitive spatial resolution, fine spectral information, and free availability make Sentinel-2 data more and more popular. In this study, we chose 10 m Sentinel-2 channels as input spectra.

Hyperspectral data: With the same spatial resolution as Sentinel-2, spectra captured by Chinese Orbita hyperspectral satellites (OHS) are

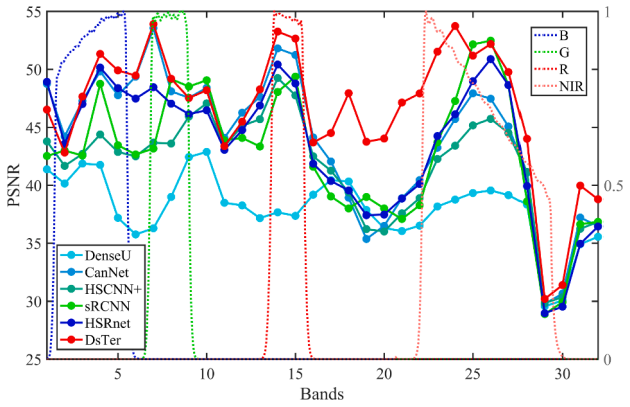


Fig. 6. PSNR values of each band generated by different CNN-based models on the selected images in Chikusei data set and the spectral response functions of four bands captured by Sentinel-2 are also displayed with dotted lines. Wavelengths in spectral response functions are mapped to the corresponding hyperspectral bands.

selected as the target in this paper, which consist of 32 bands covering spectral range from 400 nm to 1000 nm and the rich spectral information in OHS data is critically important in many research fields.

However, OHS data are only published for commercial use, so they are hardly available for free. In this paper, Simulation data sets from three free hyperspectral datasets, including Chikusei, Xiong'an, and Washington DC Mall, are used as remote sensing data set. Images in Chikusei data set is captured by the Headwall Hyperspec-VNIR-C sensor and cover Chikusei, Japan (Yokoya and Iwasaki, 2014). It contains 128 spectral bands with a size of 2517 × 2335 and spectra cover from 363 nm to 1018 nm. The second data set is Xiong'an dataset (Cen et al., 2020), which is an aerial image covering a rural area in Xiong'an New Area, China. Moreover, the spectral range of Xiong'an is similar to OHS sensors, and the image size is 3750 × 1580 with 250 bands. The main features in Xiong'an data set are various crops. The last data set, Washington DC Mall (Biehl and Landgrebe, 2002), is acquired by HYDICE airborne sensor and its spectral wavelength is 400 nm to 2500 nm. The images in Washington DC Mall data set are all with a size of 1280 × 307 × 210. In the simulation experiments, we use Hysure (Simoes et al., 2014) to downsample all channels to the same spectral resolutions as OHS and Sentinel-2 simultaneously.

In Table 1, we show the input and output size of training and test samples. Note that remote sensing images are firstly divided into training set and test set, and then they are cut into image patches with a fixed stride, respectively. Moreover, we only employ data augmentation on remote sensing data sets, which contains rotations and flips.

3.3. Results on natural images

Table 2 demonstrates the quantitative results of different models on natural images in terms of CC, mPSNR, mSSIM, and SAM, where bold fonts represent the best performance in each metric and the underlined results are the second best. Seeing this table, several conclusions can be drawn. First, in comparison with dictionary learning-based methods and CNN-based methods, CNNs not only acquire better results on spectral maintaining but also achieve good spatial fidelity, due to the sufficient spectral-spatial feature learning. Second, compared with Arad, A + achieves great improvement on spatial fidelity, because it is transferred from its namesake in spatial super-resolution, which pays more attention to the extraction of spatial details. However, the spectral enhancement is not enough, leading to the large gap contrast with CNN-based methods in terms of SAM. Third, with spectral response functions to group spectra, sRCNN and HSRnet can achieve good performance in both spatial and spectral domains. However, sRCNN recovers spectra pixel by pixel, costing much time when processing images with a big size. Moreover, HSRnet needs spectral response functions as guides, which is difficult to be acquired in practice. Lastly, combining the strong ability to global mapping of transformer with the local relationship learning of CNN, the proposed DsTer outperforms the CNN-based models significantly. The mPSNR value is increased by 0.17, and the SAM value is reduced by 0.24.

In addition to the quantitative comparisons, visual results on a randomly selected image, image named 'real and fake peppers' in the natural data set, are also shown in Fig. 4. Six bands are selected with 50 nm wavelength spacing. In these figures, each row presents error maps between the reconstructed results using eight spectral super-resolution models and the ground truth range from 0 to 1. The visual results are generally consistent with the quantitative results. Firstly, Arad obtains the high error and shows many spatial details on error maps. Secondly, it can be observed that most of errors are on peppers, especially for CanNet, HSCNN+, and sRCNN. Furthermore, the recovered bands around the central wavelengths of Red, Green, and Blue bands obtain lower errors because of more similar spectral information with the input spectra. Results generated by sRCNN obtain higher errors on the 650 and 700 nm bands, which indicates that sRCNN suffers more spectral distortion at the edge of the wavelength. Moreover, HSRnet presents relatively lower error on peppers while it shows a clear edge on its error maps, which illustrates that HSRnet loses some spatial details when recovering spectra. More importantly, the proposed DsTer outperforms all comparison methods on natural data set and presents lower difference compared with the ground truth whether on peppers or background, due to local–global relationship learning by integrating transformer and CNN.

3.4. Results on remote sensing images

Table 3 lists the quantitative results of comparison models on remote sensing images in terms of CC, mPSNR, mSSIM, and SAM, where bold fonts represent the best performance in each metric and the underlined results are the second best. As we can see, DsTer achieves the best performance on all three remote sensing data sets whether on spatial fidelity or spectral recovery. The maximum mPSNR gain reaches 0.84 dB on Chikusei data. Moreover, the reduction of SAM reaches 8.96% on Washington DC Mall data.

Comparing Arad and A + with DenseUnet, the gap between dictionary learning and CNN models is pronounced. Note that we retrained the hyperspectral dictionary with the code provided by Wu et al. and use the same training samples as used in CNN-based methods. On natural images, A + can achieve similar spatial fidelity compared with DenseUnet, while it gets worse when dealing with remote sensing multispectral images. Due to more abundant textures and geometry than natural images, it is difficult to represent these spatial details using a multispectral dictionary down-sampled from a learned hyperspectral dictionary with spectral response functions. Meanwhile, the SAMs of dictionary learning-based methods decrease compared with results on natural data due to many similar spectra in a remote sensing image.

Make a comparison between DenseUnet and CanNet, we can see that deeper networks do not always output better results. It is because there are many down-samplers in Unet, leading to the missing of spatial details and the large gap in CC, mPSNR, mSSIM to other CNNs. HSCNN+, sRCNN, and HSRnet perform well on spatial and spectral domains. Among these models, HSRnet benefits from physical model-based optimization flow and achieves almost the best results in all quantitative indicators. Moreover, HSCNN + performs well on Chikusei data in terms of SAM. It is noted that although sRCNN can obtain good results, it cost too much time for its pixel-by-pixel recovery. Thus, without transformer, HSRnet and HSCNN + are good for multispectral remote sensing images to recover hyperspectral information.

Combining transformer with CNN, the proposed DsTer outperforms those CNN-based algorithms on three remote sensing data sets, which indicates that simultaneously focusing on local and global information

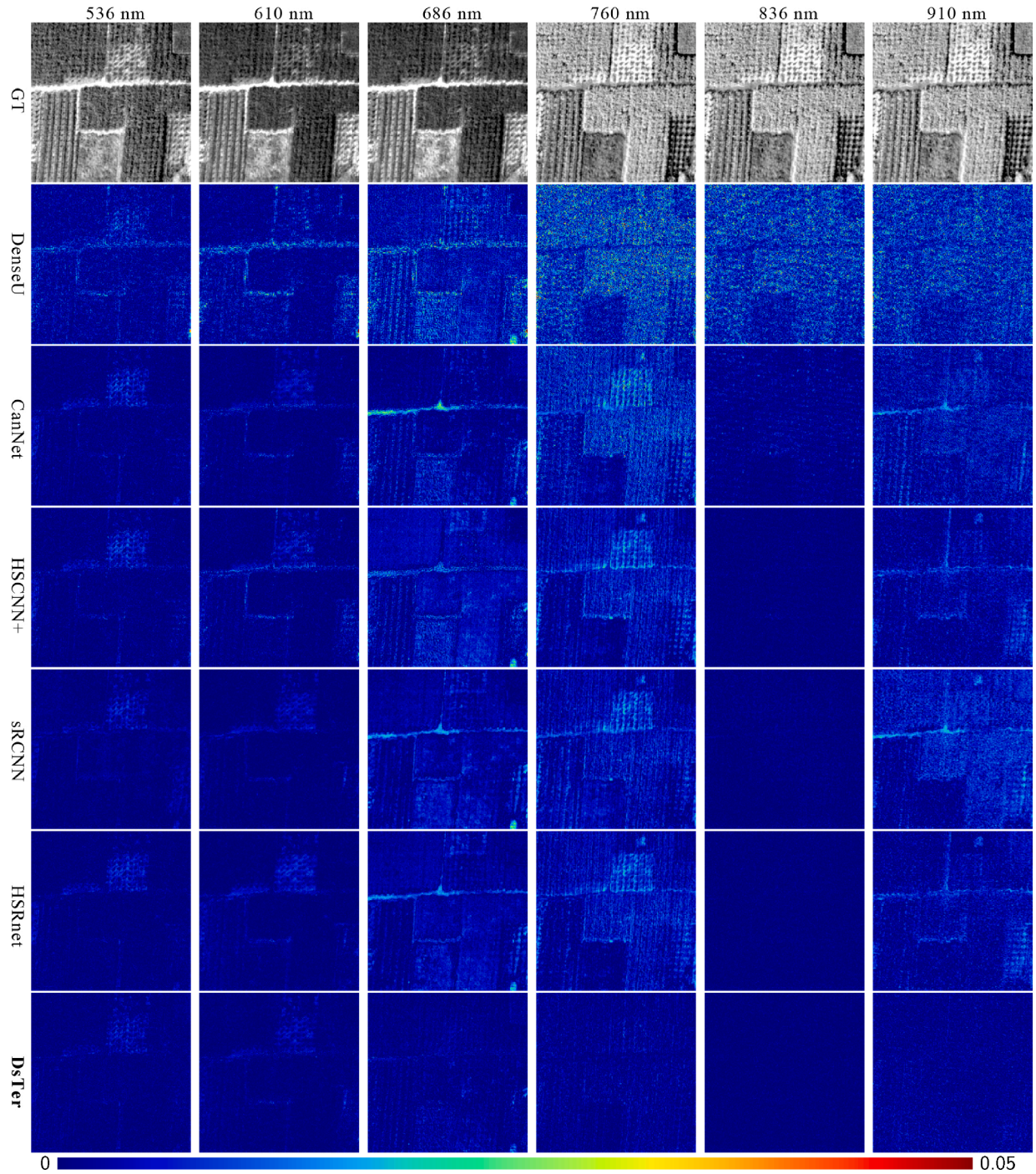


Fig. 7. Visual results of error maps on a randomly selected image in Xiong'an data set. From the top to the bottom: the ground truth, error maps generated by DenseUnet, CanNet, HSCNN+, sRCNN, HSRnet, and by the proposed DsTer.

can help recover finer-resolution spectra.

3.4.1. Chikusei data set

In Fig. 5, we randomly choose an image from Chikusei data set and display the error maps computed on the results reconstructed by CNN-based methods. Moreover, we select one band out of every five, in other words, band 5, 10, 15, 20, 25, and 30.

Seeing visualization results, we can observe that the recovered results from our DsTer are more accurate and demonstrate our method can provide higher spatial fidelity. Other methods perform badly on buildings, presenting highlighted strips in error maps. Besides, there are many spatial details including edges, textures, and geometric shapes in their error maps, which demonstrates they lose much spatial information during spectral recovery. Finally, lacking in constraint of similar input spectra, the common problem is that bands at the end of the wavelength are suffering high errors and get bad reconstruction, which can be further observed from PSNR values of each band as shown in Fig. 6. Bands around the central wavelength of input spectra are reconstructed with high PSNR while PSNR of other bands decreases. Nevertheless, the proposed DsTer obtains the highest PSNR values in most bands, especially band 16 to band 24, which further certify the superiority of our model.

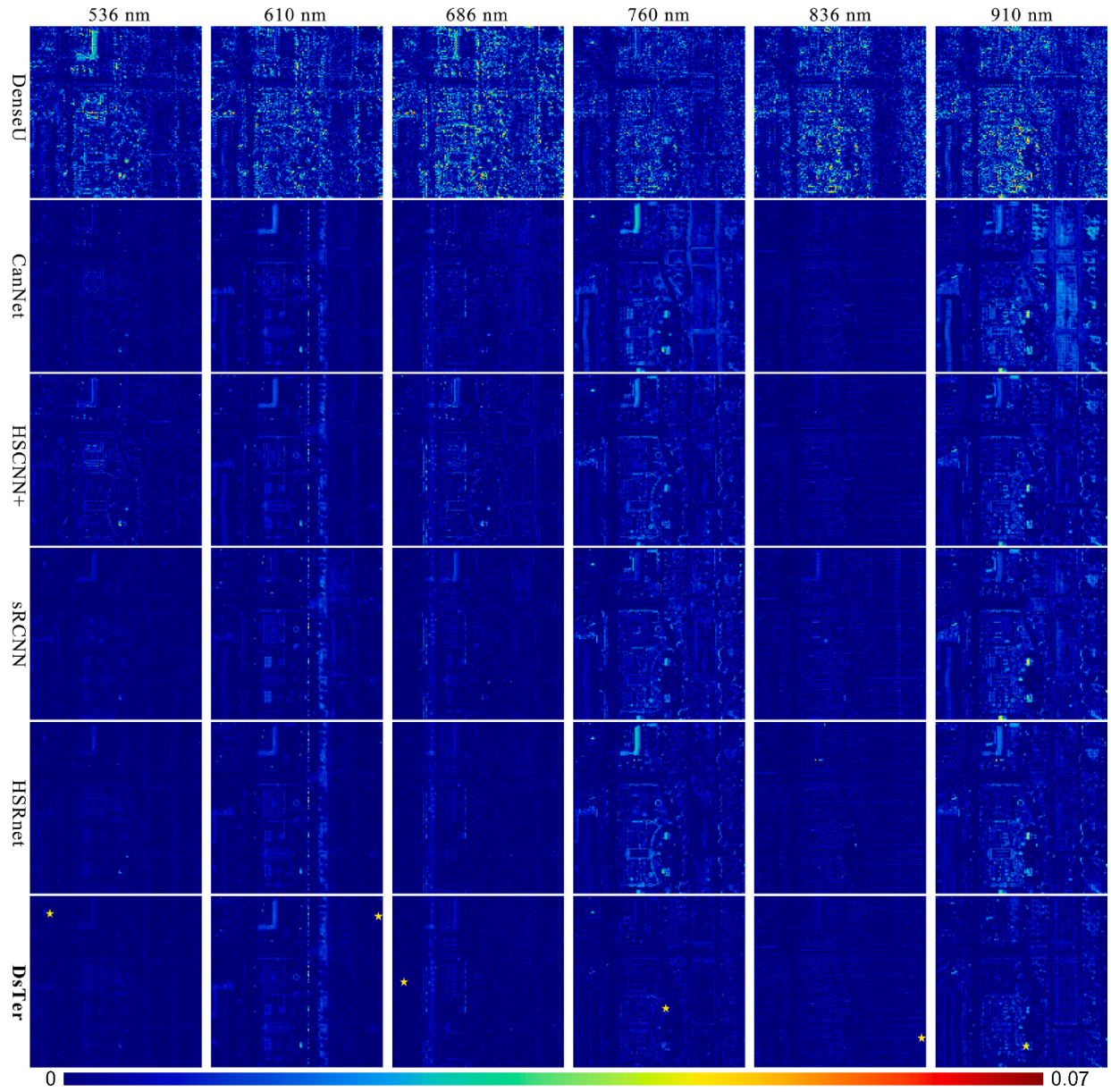


Fig. 8. Visual results of error maps on a randomly selected image in Washington DC Mall data set. From the top to the bottom: error maps generated by DenseUnet, CanNet, HSCNN+, sRCNN, HSRnet, and by the proposed DsTer. The randomly selected locations are pointed out with yellow stars on the DsTer results.

3.4.2. Xiong'an data set

Xiong'an data set covers a rural area with various crops, including grassland, elm, willow, soybean, rice, and corn, which differ in size, textures, and colors. So, we display the error maps between the reconstruction results generated by different models as well as the ground truth in Fig. 7. There are several interesting conclusions in these figures.

Firstly, seeing the visual results of the ground truth, different vegetations shows different color relationships in different bands, which demonstrates the wealth of spectra involved in Xiong'an data set. Secondly, in comparison with CanNet's results, HSCNN+, sRCNN, and HSRnet can recover better spectral information with lower error in all bands except the 686 nm band, which illustrates that they cannot perfectly learn the complete mapping from multispectral domain to hyperspectral domain. Thirdly, the vegetation coverage is much higher than bare land in Xiong'an data set, leading to the error of bare land being higher than vegetation because of insufficient learning. Moreover, seeing the results on the 760 nm band, there is an area sparsely planted with crops as shown in the ground truth. The results generated by other models all appear high errors in this sparsely-planted land, while the proposed DsTer can recover the perfectly consistent spectra with the ground truth presenting almost zero error on the 760 nm band, which further confirms the strong mapping ability of DsTer on various objectives and diverse spectra.

3.4.3. Washington DC Mall data set

The third remote sensing data set, Washington DC Mall, is a commonly used hyperspectral data covering a commercial area in Washington. On this data set, besides the error maps as shown in Fig. 8, we also randomly select several pixels and compare spectra recovered by six models with the ground truth in Fig. 9. The comparison of error maps is consistent with the quantitative results. The proposed DsTer surpasses all CNN-based models and HSRnet is the second best. Furthermore, surprisingly, CanNet performs better than HSCNN + in retaining more spatial details, showing fewer edges and spatial information in error maps. Comparing the recovered reflectance, it can be observed that the proposed DsTer can generate the most similar spectra as the ground

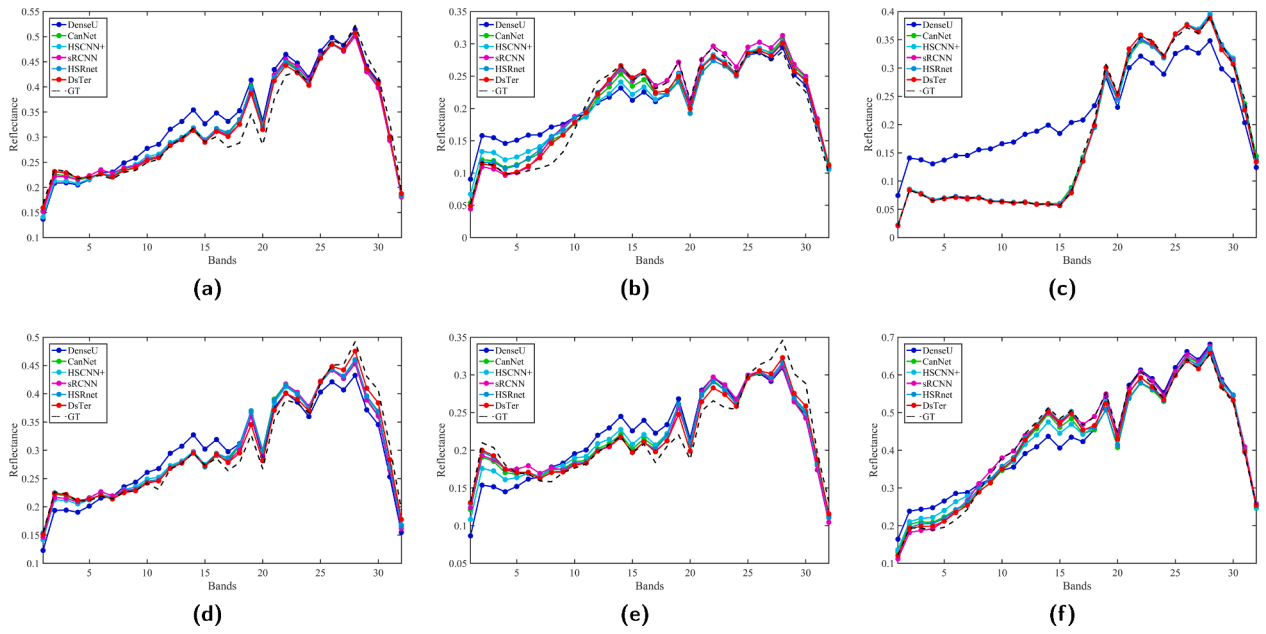


Fig. 9. Reflectance of ground truth, and results generated by six compared methods at the selected location. (a)-(f) correspond to each locations shown in Fig. 8.

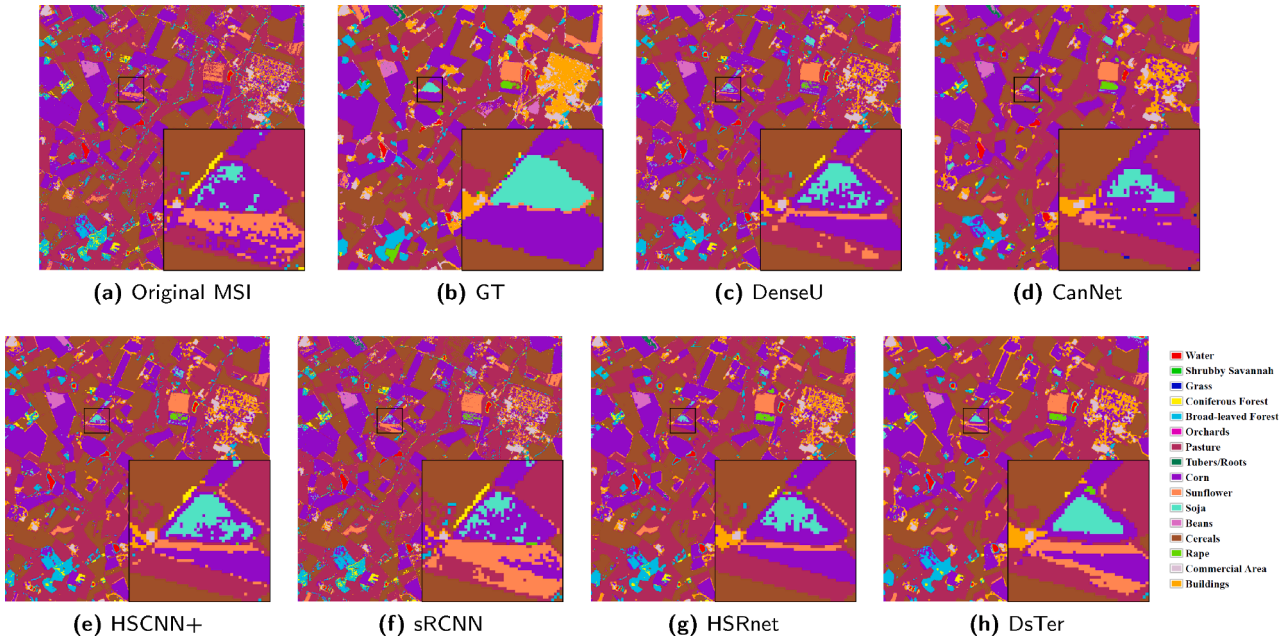


Fig. 10. Classification comparison on the hyperspectral images recovered from real Sentinel-2 data using six CNN-based models.

Table 4
Quantitative results of classification, including overall accuracy and kappa coefficient.

	OA	Kappa
Original MSI	70.7409	0.6296
DenseU	72.0516	0.6463
CanNet	72.4354	0.6511
HSCNN+	73.3791	0.6639
sRCNN	69.8376	0.6179
HSRnet	73.2212	0.6619
DsTer	74.0559	0.6732

truth. DenseUnet performs worst and the other CNN-based models can only achieve partial spectral consistency with the ground truth in several bands. Although the proposed DsTer also shows some deviation in a few bands, the spectra obtained by DsTer are the most proximal estimation of the ground truth.

3.4.4. Results on real Sentienl-2 data

To verify the algorithm performance on real data, we also use the comparison models trained on simulated data set to achieve spectral recovery on real Sentinel-2 images. Lack of consistent ground truth, we could hardly carry on a comparison in spatial fidelity or spectral recovery directly. As we all know, classification is an important task when users want to acquire comprehensible information from images. With the ground truth of a 10 m landcover map in the south of Nantes, France,

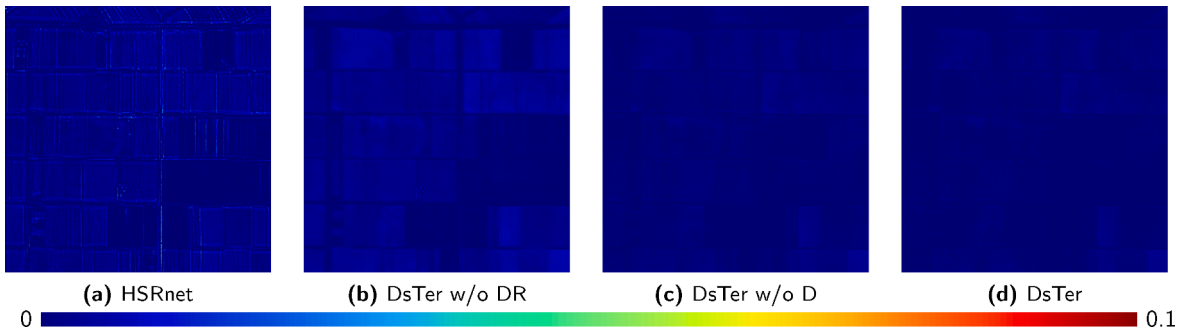


Fig. 11. Visual results in ablation study. Error maps are calculated on Chikusei data set.

Table 5
Ablation study results of the proposed elaborate designs in the proposed DsTer. Results in bold are best and the underlined are second best.

	Transformer	ResNet	Dense	CC	mPSNR	mSSIM	SAM
HSRnet	✗	✗	✗	0.9968	44.7133	0.9941	3.4528
DsTer w/o DR	✓	✗	✗	0.9970	45.1491	0.9946	3.4793
DsTer w/o D	✓	✓	✗	<u>0.9974</u>	<u>45.5397</u>	<u>0.9950</u>	<u>3.3610</u>
DsTer	✓	✓	✓	**0.9975**	**45.5546**	**0.9954**	**3.2791**

Table 6
Computational speed analysis of deep learning-based methods on CAVE data set.

	DenseUnet	sRCNN	CanNet	HSCNN+	HSRnet	DsTer w/o DR	DsTer w/o D	DsTer
Params	1360.1 K	789.3 K	163.0 K	915.1 K	769.7 K	30888.2 K	142879.8 K	9356.7 K
FLOPs	3.02×10^{10}	5.96×10^{12}	3.97×10^{10}	2.23×10^{11}	1.79×10^{11}	166×10^{12}	32.4×10^{12}	2.22×10^{12}
Training	68655s	146539s	49285s	57805s	30831s	380952s	604344s	76180s
Test	1.2598s	4.5950s	1.2387s	1.7996s	1.5364s	3.8367s	7.1648s	1.8053s

classification comparison on the hyperspectral images recovered from real Sentinel-2 data using six CNN-based models is made to verify the reliability of the recovered hyperspectral images. Sixteen classes are extracted from the reconstructed hyperspectral image using the support vector machine (SVM). Results are shown in Fig. 10.

With more spectral information in the reconstructed hyperspectral images, all models can obtain better classification results with more adjacent similar objects combined and the proposed DsTer can achieve the most similar results with the ground truth. It is noted that there are also some objectives misclassified into other classes. For example, the corn on the bottom area of the magnification is misclassified into sunflowers, which is due to the similar spectra when photographed from above. Nevertheless, the reliability of the recovered spectra generated by DsTer keeps ahead in comparison methods.

Furthermore, the quantitative evaluation also shows that combining CNN with transformer can help classification, as presented in Table 4. The quantitative results demonstrate that the improvement in OA and Kappa obtained by DsTer is superior to other CNNs. Moreover, all methods can improve the OA and Kappa except sRCNN, due to the pixel-by-pixel recovery. Because sRCNN cannot involve the global relationship between no-local similar pixels. This further indicates that DsTer can reconstruct more accurate spectra and perform well not only in image quality enhancement but also in subsequent applications.

3.5. Ablation study

In this part, we discuss the contributions of elaborate designs in the proposed DsTer on Chikusei data, including transformer structure, ResNet-based transformer layer, and dense connection. HSRnet is chosen as baseline and results are shown in Fig. 11 and Table 5.

With classical transformer structure, *DsTer w/o DR* can achieve a little better than HSRnet in the spatial domain, while leading to some spectral distortion. This may be due to the MLP used to process remote sensing images in classical transformer that cannot fully explore spatial-spectral features.

Moreover, comparing *DsTer w/o D* with *DsTer w/o DR*, ResNet-base transformer layer which is suitable to the 3D data structure of remote sensing multispectral images can further enhance the spatial fidelity of models, and more importantly, it can greatly optimize spectral information to generate hyperspectral images with lower SAM.

Furthermore, seeing the last two lines in Table 5, DsTer recovers more accurate spectra and keeps finer spatial details through dense strategy, which should be due to the features from multi-level transformers. Deep and shallow features are both important in multi-level transformer layers.

3.6. Computational Speed Analysis

In deep learning-based methods, model complexity and computational speed are also very important. In this part, we make a comparison about model parameter, floating-point operations (FLOPs), training convergence, and test time.

Results are listed in Table 6. Comparing with CNN-based models, transformer-based models commonly require more parameters. *DsTer w/o DR* denotes the classical transformer with nearly hundreds of times parameters more than CanNet, and replacing MLP with ResNet further increases the model parameter. The good news is, with ResNet, *DsTer w/o D* shows lower FLOPs, while it still costs too much time to achieve spectral super-resolution with a large image size. With dense strategy, DsTer could recycle shallow features and boost computation. Moreover, bottlenecks further improve the feature utilization and reduce model parameters. Although, compared with CNN-based models, the proposed DsTer does not show great superiority on running time, while it keeps a good balance between performance and computational speed.

4. Conclusion

In this work, we propose a dense transformer for spectral super-resolution, which mainly consists of dense spectral transformers and ResNet-based transformer layers. Firstly, dense spectral transformer can exploit different features from multi-level transformers. Secondly, ResNet, compared with MLP, is more suitable for remote sensing image processing. Thirdly, combining detailed features from the local range explored by CNN with the global relationship between long-range interactions learned by transformer, DsTer can fuse local and global information to achieve better spectral recovery. Experimental results not only on a natural data set but also on three remote sensing data sets have shown the superiority of the proposed DsTer. Furthermore, classification application on real data also verifies the reliability of the recovered spectra.

The proposed DsTer combines the advantages of CNN and transformer. Nevertheless, it is still hard for all deep learning-based methods to directly utilize the model trained on one sensor to recover spectra captured by other sensors because of varying amounts of spectral bands. Thus, in the future, how to greatly boost the training of transformer and how to deal with the model generalization considering the change of channel numbers may be the two biggest problems demanding prompt solution.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 41922008, Grant 62071341, Grant 61971319, and Grant 61901166; in part by the Hubei Science Foundation for Distinguished Young Scholars under Grant 2020CFA051; and in part by the Fundamental Research Funds for the Central Universities under Grant 531118010209.

References

Akhtar, N., Mian, A., Jan 2020. Hyperspectral recovery from RGB images using gaussian processes. IEEE Trans. Pattern Anal. Mach. Intell. 42 (1), 100–113.
Arad, B., Ben-Shahar, O., 2016. Sparse recovery of hyperspectral signal from natural RGB images. In: Computer Vision – ECCV 2016. Springer International Publishing, pp. 19–34.
Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., et al., 2022. NTIRE 2022 spectral recovery challenge and dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
Barnsley, M., Settle, J., Cutter, M., Lobb, D., Teston, F., 2004. The proba/chris mission: a low-cost smallsat for hyperspectral multiangle observations of the earth surface and atmosphere. IEEE Trans. Geosci. Remote Sens. 42 (7), 1512–1520.
Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V., 2019. Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3286–3295.
Biehl, L., Landgrebe, D., 2002. Multispec-a tool for multispectral–hyperspectral image data analysis. Computers & Geosciences 28 (10), 1153–1159.
Bioucas-Dias, J.M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., Chanussot, J., 2013. Hyperspectral remote sensing data analysis and future challenges. IEEE Geosci. Remote Sens. Magaz. 1 (2), 6–36.
Can, Y.B., Timofte, R., 2018. An efficient cnn for spectral reconstruction from rgb images. arXiv preprint arXiv:1804.04647.
Cen, Y., Zhang, L., Zhang, X., Wang, Y., Qi, W., Tang, S., Zhang, P., 2020. Aerial hyperspectral remote sensing classification dataset of xiongan new area (matiwan village). J. Remote Sens. 24 (11), 1299–1306.
Chen, M., Ke, Y., Bai, J., Li, P., Lyu, M., Gong, Z., Zhou, D., 2020. Monitoring early stage invasion of exotic spartina alterniflora using deep-learning super-resolution techniques based on multisource high-resolution satellite imagery: A case study in the yellow river delta, china. Int. J. Appl. Earth Obs. Geoinf. 92, 102180.
Dalponte, M., Bruzzone, L., Gianelle, D., 2008. Fusion of hyperspectral and lidar remote sensing data for classification of complex forest areas. IEEE Trans. Geosci. Remote Sens. 46 (5), 1416–1427.
Dian, R., Li, S., 2019. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. IEEE Trans. Image Process. 28 (10), 5135–5146.
Dian, R., Li, S., Kang, X., 2020. Regularizing hyperspectral and multispectral image fusion by cnn denoiser. IEEE Trans. Neural Networks Learn. Syst. 32 (3), 1124–1135.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale.
Fu, Y., Zhang, T., Wang, L., Huang, H., 2021. Coded hyperspectral image reconstruction using deep external and internal learning. IEEE Trans. Pattern Anal. Mach. Intell.
Fu, Y., Zhang, T., Zheng, Y., Zhang, D., Huang, H., 2018. Joint camera spectral sensitivity selection and hyperspectral image recovery. In: Computer Vision – ECCV 2018. Springer International Publishing, pp. 812–828.
Fu, Y., Zhang, T., Zheng, Y., Zhang, D., Huang, H., 2020. Joint camera spectral response selection and hyperspectral image recovery. IEEE Tran. Pattern Anal. Mach. Intell.
Galliani, S., Lanaras, C., Marmanis, D., Baltsavias, E., Schindler, K., 2017. Learned spectral super-resolution. arXiv preprint arXiv:1703.09470.
Gewali, U.B., Monteiro, S.T., Saber, E., 2019. Spectral super-resolution with optimized bands. Remote Sensing 11 (14), 1648.
Gowen, A., O'Donnell, C., Cullen, P., Downey, G., Frias, J., 2007. Hyperspectral imaging-an emerging process analytical tool for food quality and safety control. Trends Food Sci. Technol. 18 (12), 590–598.
Hang, R., Liu, Q., Li, Z., 2021. Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery. IEEE Trans. Image Process. 30, 7256–7265.
He, J., Li, J., Yuan, Q., Shen, H., Zhang, L., 2021. Spectral response function-guided deep optimization-driven network for spectral super-resolution. IEEE Trans. Neural Networks Learn. Syst. 1–15.
He, J., Yuan, Q., Li, J., Zhang, L., 2022. Ponet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images. Information Fusion 80, 205–225.
Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., Chanussot, J., 2020. Graph convolutional networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 59 (7), 5966–5978.
Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2021. Spectralformer: Rethinking hyperspectral image classification with transformers. In: IEEE Trans. Geosci. Remote Sens..
Hu, J.-F., Huang, T.-Z., Deng, L.-J., Jiang, T.-X., Vivone, G., Chanussot, J., 2021. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. In: IEEE Trans. Neural Networks Learn. Syst..
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269.
Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., jul 2017. The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.
Jia, Y., Zheng, Y., Gu, L., Subpa-Asa, A., Lam, A., Sato, Y., Sato, I., oct 2017. From RGB to spectrum for natural scenes via manifold-based mapping. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE.
Jin, C., Deng, L.-J., Huang, T.-Z., Vivone, G., 2022. Laplacian pyramid networks: A new approach for multispectral pansharpening. Information Fusion 78, 158–170.
Kampffmeyer, M., Salberg, A.-B., Jenssen, R., June 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
Kruse, F., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., May 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44 (2–3), 145–163.
Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sens. Lett. 14 (5), 778–782.
Li, J., Du, S., Song, R., Wu, C., Li, Y., Du, Q., 2022. Hasic-net: Hybrid attentional convolutional neural network with structure information consistency for spectral super-resolution of rgb images. In: IEEE Trans. Geosci. Remote Sens..
Li, Z., Chen, H., White, J.C., Wulder, M.A., Hermosilla, T., 2020. Discriminating treed and non-treed wetlands in boreal ecosystems using time series sentinel-1 data. Int. J. Appl. Earth Obs. Geoinf. 85, 102007.
Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer. In: IEEE International Conference on Computer Vision Workshops.
Liu, Z., Ma, Q., Jiang, J., Liu, X., 2021. Improving hyperspectral super-resolution via heterogeneous knowledge distillation. In: ACM Multimedia Asia. pp. 1–7.
Lu, G., Fei, B., 2014. Medical hyperspectral imaging: a review. J. Biomed. Opt. 19 (1), 1–24.
Luo, H., Zheng, Q., Fang, L., Guo, Y., Guo, W., Wang, C., Li, J., 2021. Boundary-aware graph markov neural network for semiautomated object segmentation from point clouds. Int. J. Appl. Earth Obs. Geoinf. 104, 102564.
Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 42 (8), 1778–1790.
Muad, A.M., Foody, G.M., 2012. Super-resolution mapping of lakes from imagery with a coarse spatial and fine temporal resolution. International Journal of Applied Earth Observation and Geoinformation 15, 79–91, special Issue on Geographic Object-based Image Analysis: GEOBIA.
Nguyen, R.M.H., Prasad, D.K., Brown, M.S., 2014. Training-based spectral reconstruction from a single RGB image. In: Computer Vision – ECCV 2014. Springer International Publishing, pp. 186–201.

Nie, S., Gu, L., Zheng, Y., Lam, A., Ono, N., Sato, I., jun 2018. Deeply learned filter response functions for hyperspectral reconstruction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE.

Rangnekar, A., Mokashi, N., Ientilucci, E., Kanan, C., Hoffman, M., 2017. Aerial spectral super-resolution using conditional adversarial networks. arXiv e-prints.

Robles-Kelly, A., oct 2015. Single image spectral reconstruction for multimedia applications. In: Proceedings of the 23rd ACM international conference on Multimedia. ACM.

Shao, Z., Cheng, G., Ma, J., Wang, Z., Wang, J., Li, D., 2021. Real-time and accurate uav pedestrian detection for social distancing monitoring in covid-19 pandemic. IEEE Trans. Multimedia.

Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F., jun 2018. HSCNN+: Advanced CNN-based hyperspectral recovery from RGB images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.

Simoes, M., Bioucas-Dias, J., Almeida, L.B., Chanussot, J., 2014. Hyperspectral image superresolution: An edge-preserving convex formulation. In: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 4166–4170.

Sinha, A., Dolz, J., 2020. Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inform. 25 (1), 121–130.

Song, H., Ma, Y., Han, Y., Shen, W., Zhang, W., Li, Y., Liu, X., Peng, Y., Hao, X., 2021. Deep-learned broadband encoding stochastic filters for computational spectroscopic instruments. Adv. Theory Simul. 4 (3), 2000299.

Sun, G., Zhang, X., Jia, X., Ren, J., Zhang, A., Yao, Y., Zhao, H., 2020. Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images. Int. J. Appl. Earth Obs. Geoinf. 91, 102157.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46 (sup1), 234–240.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need.

Wambugu, N., Chen, Y., Xiao, Z., Tan, K., Wei, M., Liu, X., Li, J., 2021. Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review. Int. J. Appl. Earth Obs. Geoinf. 105, 102603.

Wang, Y., Yuan, Q., Li, T., Zhu, L., Zhang, L., 2021. Estimating daily full-coverage near surface o3, co, and no2 concentrations at a high spatial resolution over china based on s5p-tropomi and geos-fp. ISPRS J. Photogram. Remote Sens. 175, 311–325.

Wang, Y., Yuan, Q., Zhu, L., Zhang, L., 2022. Spatiotemporal estimation of hourly 2-km ground-level ozone over china based on himawari-8 using a self-adaptive geospatially local model. Geosci. Front. 13 (1), 101286.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wei, Q., Bioucas-Dias, J., Dobigeon, N., Tourneret, J.-Y., 2015. Hyperspectral and multispectral image fusion based on a sparse representation. IEEE Trans. Geosci. Remote Sens. 53 (7), 3658–3668.

White, J.C., Saarinen, N., Kankare, V., Wulder, M.A., Hermosilla, T., Coops, N.C., Pickell, P.D., Holopainen, M., Hyyppä, J., Vastaranta, M., 2018. Confirmation of post-harvest spectral recovery from landsat time series using measures of forest cover and height derived from airborne laser scanning data. Remote Sens. Environ. 216, 262–275.

Wu, J., Aeschbacher, J., Timofte, R., oct 2017. In defense of shallow learned spectral reconstruction from RGB images. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE.

Xiao, J., Li, J., Yuan, Q., Zhang, L., 2021a. A dual-unet with multistage details injection for hyperspectral image fusion. IEEE Trans. Geosci. Remote Sens. 1–13.

Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., Zhang, L., 2021b. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. IEEE Trans. Geosci. Remote Sens.

Xiao, Y., Yuan, Q., He, J., Zhang, Q., Sun, J., Su, X., Wu, J., Zhang, L., 2022. Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer. Int. J. Appl. Earth Obs. Geoinf. 108, 102731.

Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F., oct 2017. HSCNN: CNN-based hyperspectral image recovery from spectrally undersampled projections. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE.

Yang, L., Li, Z., Pei, Z., Zhang, D., 2021. Fs-net: Filter selection network for hyperspectral reconstruction. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2933–2937.

Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.K., 2010. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. IEEE Trans. Image Process. 19 (9), 2241–2253.

Yokoya, N., Iwasaki, A., 2014. Airborne unmixing-based hyperspectral super-resolution using rgb imagery. In: 2014 IEEE Geoscience and Remote Sensing Symposium. IEEE, pp. 2653–2656.

Yokoya, N., Yairi, T., Iwasaki, A., 2012. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. 50 (2), 528–537.

Yu, C., Chen, X., 2014. Remote sensing image denoising application by generalized morphological component analysis. Int. J. Appl. Earth Obs. Geoinf. 33, 83–97.

Yuan, Q., Zhang, Q., Li, J., Shen, H., Zhang, L., 2019. Hyperspectral image denoising employing a spatial-spectral deep residual convolutional neural network. IEEE Trans. Geosci. Remote Sens. 57 (2), 1205–1218.

Zhang, L., Lang, Z., Wang, P., Wei, W., Liao, S., Shao, L., Zhang, Y., 2020a. Pixel-aware deep function-mixture network for spectral super-resolution. Proc. AAAI Conf. Artif. Intell. 34 (07), 12821–12828.

Zhang, Q., Yuan, Q., Li, J., Li, Z., Shen, H., Zhang, L., 2020b. Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning. ISPRS J. Photogramm. Remote Sens. 162, 148–160.

Zhang, Q., Yuan, Q., Li, Z., Sun, F., Zhang, L., 2021a. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. ISPRS J. Photogramm. Remote Sens. 177, 161–173.

Zhang, Q., Yuan, Q., Li, J., Wang, Y., Zhang, L., 2021b. Generating seamless global daily amsr2 soil moisture (sgd-sm) long-term products for the years 2013–2019. Earth Syst. Sci. Data 13 (3), 1385–1401.

Zhao, H., Jia, J., Koltun, V., 2020. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10076–10085.