Bidirectional Dilation Transformer for Multispectral and Hyperspectral Image Fusion

Shangqi Deng, **Liang-Jian Deng**[*], **Xiao Wu**, **Ran Ran** and **Rui Wen**
University of Electronic Science and Technology of China
shangqideng0124@gmail.com, liangjian.deng@uestc.edu.cn, wxwsx1997@gmail.com,
ranran@std.uestc.edu.cn wenrui202102@163.com

Abstract

Transformer-based methods have proven to be effective in achieving long-distance modeling, capturing the spatial and spectral information, and exhibiting strong inductive bias in various computer vision tasks. Generally, the Transformer model includes two common modes of multi-head self-attention (MSA): spatial MSA (Spa-MSA) and spectral MSA (Spe-MSA). However, Spa-MSA is computationally efficient but limits the global spatial response within a local window. On the other hand, Spe-MSA can calculate channel self-attention to accommodate high-resolution images, but it disregards the crucial local information that is essential for low-level vision tasks. In this study, we propose a bidirectional dilation Transformer (BDT) for multispectral and hyperspectral image fusion (MHIF), which aims to leverage the advantages of both MSA and the latent multiscale information specific to MHIF tasks. The BDT consists of two designed modules: the dilation Spa-MSA (D-Spa), which dynamically expands the spatial receptive field through a given hollow strategy, and the grouped Spe-MSA (G-Spe), which extracts latent features within the feature map and learns local data behavior. Additionally, to fully exploit the multiscale information from both inputs with different spatial resolutions, we employ a bidirectional hierarchy strategy in the BDT, resulting in improved performance. Finally, extensive experiments on two commonly used datasets, CAVE and Harvard, demonstrate the superiority of BDT both visually and quantitatively. Furthermore, the related code will be available at the GitHub page of the authors.

1 Introduction

Hyperspectral imaging (HSI) is a widely used technology in various fields, including agriculture [Lu *et al.*, 2020; Wu *et al.*, 2011], food safety [Feng and Sun, 2012], biomedical diagnostics [Piqueras *et al.*, 2011], and atmospheric environment detection [Gao *et al.*, 2006]. HSIs with high spectral resolution produce precise spectral characteristic curves, and the abundance of bands makes it convenient for mutual band correction. However, due to the current physical imaging technology's constraints, there is a trade-off between the spatial and spectral resolution of the natural imaging process. Therefore, it is impossible to produce an image with high spatial and spectral resolution simultaneously. As a result, multispectral and hyperspectral image fusion (MHIF) has emerged as a promising method to generate the necessary high-resolution hyperspectral images (HR-HSI). Numerous approaches have been developed for MHIF and can be broadly categorized into two categories: traditional methods [Guo *et al.*, 2020; Yang *et al.*, 2020b; Yang *et al.*, 2020a]

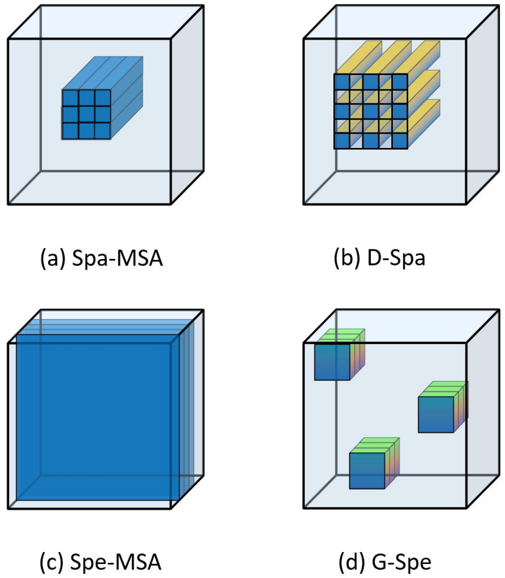


Figure 1: The comparison of (a) Spa-MSA [Liu *et al.*, 2021], (b) the proposed D-Spa based on Spa-MSA, (c) Spe-MSA [Zamir *et al.*, 2022], and (d) the proposed G-Spe based on Spe-MSA. The blue clusters indicate the image tokens in (a) and (b). Utilizing the dilation operation, the proposed D-Spa can expand the receptive field of Spa-MSA. In (c) and (d), the blue slices denote the image tokens, and we design G-Spe to allow the model to learn more data behavior inside the feature map.

[*] Corresponding author

and deep learning (DL)-based techniques [Yan *et al.*, 2022; Zhou *et al.*, 2022; Cao *et al.*, 2020].

In recent years, deep learning (DL)-based techniques have become increasingly popular, with CNN modules being the current state-of-the-art for MHIF problems due to their spatial-agnostic and channel-specific convolutional properties [Li *et al.*, 2021]. Researchers have designed specific convolution modules and stacked them to construct a general network structure that effectively extracts potential behavior from databases. However, the local receptive field in CNNs limits long-range dependencies and may hinder the internal modeling of the image. Recently, the Vision Transformer (ViT)[Kolesnikov *et al.*, 2021] has demonstrated impressive performance on various computer vision tasks[Hu *et al.*, 2022]. For concision, this method is also referred to as Spa-MSA.

We propose a fusion architecture that integrates spatial and spectral information and fully exploits MSA to model similar patches in a hyperspectral image, considering the properties of the MHIF task. While Spa-MSA lacks the modeling of longer-distance information, Spe-MSA does not make full use of the information inside the data. To achieve a more wide-range correlation, our proposed architecture includes dilation Spa-MSA and grouped Spe-MSA modules. The contributions of this paper are listed as follows (also find more details in Fig. 1):

- We present a novel bidirectional dilation Transformer (BDT) architecture that utilizes both dilation Spa-MSA (D-Spa) and grouped Spe-MSA (G-Spe) modules for MHIF. Our experimental results on benchmark datasets demonstrate that our method achieves state-of-the-art (SOTA) performance. We also conduct additional experiments to evaluate the efficiency of D-Spa and G-Spe modules, the bidirectional structures, and the impact of dilation rates on the overall performance.
- To improve the receptive field of Spa-MSA, we design the D-Spa to extract a broader range of local information for the MHIF task. Specifically, D-Spa does not require additional parameters and calculations, which can be viewed as a plug-and-play module for all Spa-MSA based approaches. Various experiments in Sect. 3 demonstrate the effectiveness of the proposed dilation strategy.
- To fully exploit the spatial information along channel dimension, we design a so-called G-Spe to extract latent features inside the feature map and learn local data behavior.

2 Related Works

2.1 Transformer in MHIF

The Transformer architecture has demonstrated strong performance in various vision tasks, and many researchers are attempting to leverage it for the MHIF problem with promising results. For instance, Hu $et\ al.$[Hu *et al.*, 2022] were the first to use Transformer for MHIF and achieved powerful performance with a lightweight network. Additionally, Meng $et\ al.$[Meng *et al.*, 2022] proposed an advanced transformer-based model for remote sensing pansharpening. Ma $et\ al.$[Ma *et al.*, 2021] utilized Transformer instead of CNN to learn the prior of hyperspectral images (HSIs) and then used an unfolding network to simulate iterative solution processes for HSI super-resolution. Furthermore, Zhou $et\ al.$ [Zhou *et al.*, 2021] proposed a customized Transformer that facilitates collaborative feature learning across two modalities for remote sensing pansharpening.

2.2 Motivation

Despite the promising outcomes of the aforementioned methods, which largely rely on the powerful self-attention module, they often adopt the self-attention or Transformer structure for various image fusion tasks without fully considering their deficiencies, especially for the specific MHIF problem. For instance, Spa-MSA can restore image details and reduce computational complexity by correlating local pixels, but its receptive field is significantly restricted by the window size. Similarly, previous Spe-MSA treats channels as tokens and uses the information of the entire space for self-attention, but this does not fully utilize the information inside the image. To address the issue of Spa-MSA, we are inspired by the concept of dilation convolution [Li *et al.*, 2018b] to design a new 2D dilation structure specifically for Spa-MSA called D-Spa. D-Spa can effectively enlarge the receptive field without introducing additional parameters or computational complexity. To address the issue of Spe-MSA, we propose a Grouped G-Spe that groups the space and then performs Spe-MSA in small groups, which may extract information within the feature and better learn local data behavior. Additionally, we design a bidirectional hierarchy structure for better exploiting multiscale information of the two inputs, which have different spatial resolutions, for the specific application of MHIF.

3 Methodology

In this section, we present our BDT designed for the MHIF task. We first introduce the overall architecture of our BDT in Sec. 3.1. Subsequently, we analyse the function of D-Spa in Sec. 3.2. Finally, we describe the design of G-Spe in Sec. 3.3.

3.1 The Overall Architecture

Our BDT is outlined in Fig. 2, which is a hierarchical bidirectional input architecture with two stages, *i.e.,* Bimodal Feature Extraction (BFE) and Bimodal Feature Fusion (BFF). In order to extract spatial information, we concatenate the bicubic interpolated LR-HSI $\mathcal{X}^U \in \mathbb{R}^{H\times W\times S}$ and HR-MSI $\mathcal{Y} \in \mathbb{R}^{H\times W\times s}$ as the input of the spatial branch. Besides, D-Spa in BFE is designed to learn the spatial information, where output feature maps are $\mathcal{D}_i$, $i = 1, 2, 3$. In detail, the process of BFE is as follows:

$$\mathcal{D}_i = \mathrm{SpatialBranch}\left(\mathrm{Conv}_1\left(\mathrm{Cat}\left(\mathcal{Y}, \mathcal{X}^{\mathrm{U}}\right)\right)\right), \qquad (1)$$

where Conv_1 is a convolutional structure. Using HR-HSI $\mathcal{X} \in \mathbb{R}^{h\times w\times S}$ as the input of the spectral branch, the information on the spectrum is dynamically learned through G-Spe, and outputs feature maps $\mathcal{G}_i$ $(i = 1, 2, 3)$ as shown in the following formula:

$$\mathcal{G}_i = \mathrm{SpectralBranch}\left(\mathrm{Conv}_2\left(\mathcal{X}\right)\right), \qquad (2)$$

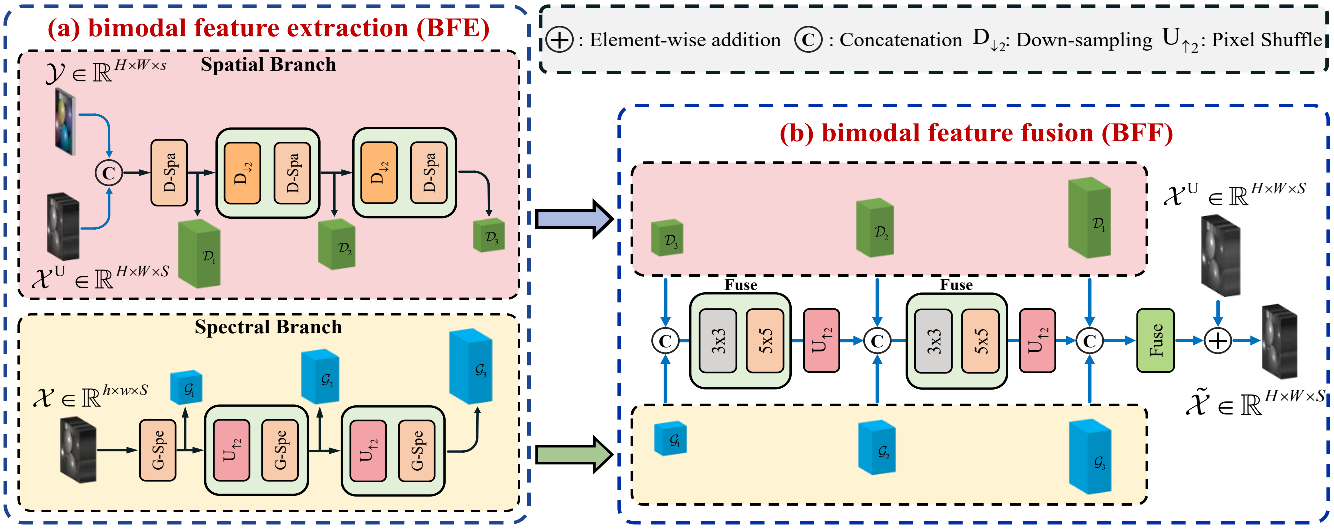

(a) bimodal feature extraction (BFE)
Spatial Branch
$\mathcal{Y} \in \mathbb{R}^{H\times W\times s}$
$\mathcal{X}^{\mathrm{U}} \in \mathbb{R}^{H\times W\times S}$
C
D-Spa
$\mathrm{D}_{\downarrow 2}$
D-Spa
$\mathrm{D}_{\downarrow 2}$
D-Spa
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_3$
Spectral Branch
$\mathcal{X} \in \mathbb{R}^{h\times w\times S}$
G-Spe
$\mathrm{U}_{\uparrow 2}$
G-Spe
$\mathrm{U}_{\uparrow 2}$
G-Spe
$\mathcal{G}_1$
$\mathcal{G}_2$
$\mathcal{G}_3$
⊕ : Element-wise addition
C : Concatenation
$\mathrm{D}_{\downarrow 2}$: Down-sampling
$\mathrm{U}_{\uparrow 2}$: Pixel Shuffle
(b) bimodal feature fusion (BFF)
$\mathcal{D}_3$
$\mathcal{D}_2$
$\mathcal{D}_1$
$\mathcal{X}^{\mathrm{U}} \in \mathbb{R}^{H\times W\times S}$
Fuse
3x3
5x5
$\mathrm{U}_{\uparrow 2}$
Fuse
3x3
5x5
$\mathrm{U}_{\uparrow 2}$
C
C
C
Fuse
⊕
$\tilde{\mathcal{X}} \in \mathbb{R}^{H\times W\times S}$
$\mathcal{G}_1$
$\mathcal{G}_2$
$\mathcal{G}_3$

Figure 2: The overall architecture of the proposed BDT approach. (a) The diagram of proposed BFE consisted of spatial and spectral branches. (b) The inputs of proposed BFF are the output of the spectral branch and the spatial branch in the BFE, respectively. Please note that $\mathcal{X}$ is the LR-HSI, $\mathcal{Y}$ is the HR-MSI, and $\mathcal{X}^U$ is the bicubic interpolation LR-HSI. $\mathcal{D}_i$ and $\mathcal{G}_i$ respectively represent spatial information and spectral information extracted from bimodal feature extraction (BFE), *i.e.,* the subgraph on the left. Then, $\mathcal{D}_i$ and $\mathcal{G}_i$ are paired into the bimodal feature fusion (BFF) to generate the final output, *i.e.,* the subgraph on the right.

where Conv_2 is a multi-layer convolution structure used to increase the channels. To fuse the feature maps, *i.e.,* $\mathcal{D}_i$ and $\mathcal{G}_i$, we design the BFF model, which is an efficient two-layer convolutional structure. In detail, we concatenate $\mathcal{D}_3$ and $\mathcal{G}_1$ first, and send the concatenated one to the fusion module which involves a 3×3 kernel and a 5×5 kernel, and then upsample through PixelShuffle, as shown in the following formula:

$$\mathcal{F}_1 = \mathrm{PixelShuffle}\left(\mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{D}_3, \mathcal{G}_1\right)\right)\right). \tag{3}$$

Then, we concatenate $\mathcal{F}_1$, $\mathcal{D}_2$ and $\mathcal{G}_2$ together, and upsample the concatenated result. After that, we fuse the upsampled result as the following formula:

$$\mathcal{F}_2 = \mathrm{PixelShuffle}\left(\mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{F}_1, \mathcal{D}_2, \mathcal{G}_2\right)\right)\right). \tag{4}$$

Finally, we add the fusion results of $\mathcal{F}_2$, $\mathcal{D}_3$ and $\mathcal{G}_1$ to the Bicubic interpolated LR-HSI $\mathcal{X}^U$, and the final output $\tilde{\mathcal{X}} \in \mathbb{R}^{H\times W\times S}$ is expressed by the following formula:

$$\tilde{\mathcal{X}} = \mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{F}_2, \mathcal{D}_3, \mathcal{G}_1\right)\right) + \mathcal{X}^U. \tag{5}$$

3.2 D-Spa

Vanilla convolution is a fundamental building block of convolutional neural networks (CNNs) which have seen tremendous success in several computer vision tasks, *e.g.,* image classification [Hong *et al.*, 2021], image super-resolution [Liang *et al.*, 2021], and image segmentation[Liu *et al.*, 2021]. Dilation convolution increases the receptive field of the convolution kernel without adding additional parameters, retains the internal structure of data and avoids using a pooling layer to downsample the feature map. The dilation convolution operation with elements $k \times k$ in the kernel and a dilation rate d at the (i, j)th pixel position can be expressed as a linear combination of input $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ around (i, j)th pixel position, which can be expressed as follows:

$$\mathbf{F}^{'}_{(:,i,j)} = \sum_{(x,y)\in\Omega(i,j)} \mathbf{W}\left[\mathrm{P}_{(i,j)} - \mathrm{P}_{(x,y)}\right]\mathbf{F}_{(:,x,y)}, \tag{6}$$

where $\mathbf{F}_{(:,x,y)} \in \mathbb{R}^C$ indicates the vector of the (x, y)th pixel position in the input feature map $\mathbf{F}$; $\Omega\,(i, j)$ represents the coordinate set of the dilation area centered on the (i, j)th pixel position; $\mathbf{F}^{'}_{(:,i,j)} \in \mathbb{R}^{C^{'}}$ indicates the vector of the (i, j)th pixel position in the output feature map $\mathbf{F}^{'} \in \mathbb{R}^{C^{'}\times H\times W}$ and $\mathbf{W} \in \mathbb{R}^{C^{'}\times C\times k\times k}$ is the convolution kernel of $k \times k$, where $\mathbf{W}\left[\mathrm{P}_{(i,j)} - \mathrm{P}_{(x,y)}\right] \in \mathbb{R}^{C^{'}\times C}$ means the convolution kernel weight which contains coordinate offset $\left[\mathrm{P}_{(i,j)} - \mathrm{P}_{(x,y)}\right] \in \left\{\left(-\frac{k+1}{2}d, -\frac{k+1}{2}d\right), \left(-\frac{k-1}{2}d, -\frac{k-1}{2}d\right), ..., \left(\frac{k-1}{2}d, \frac{k-1}{2}d\right)\right\}$ with dilation rate d. Jiao *et al.* [Jiao *et al.*, 2023] used the unfold operation to implement the expansion of the window and designed a sliding mode. However, our D-Spa expands the window in fixed position, instead of sliding it pixel by pixel, and expands windows by index values. Han *et al.* [Han *et al.*, 2021] present a novel point of view, which regards Spa-MSA as a variant of convolution, with the properties of sparse connectivity, weight sharing, depth separation, and dynamic weight. To this end, we can represent D-Spa in the form of convolution.

We operate three 1×1 convolutions on the input feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ to generate three tensors, *i.e.,* $\mathbf{Q} \in \mathbb{R}^{C\times H\times W}$, $\mathbf{K} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$, respectively. Taking only one head in D-Spa as an example, given a window size k and a dilation rate d of 2, the output $\mathbf{V}^{'} \in \mathbb{R}^{C\times H\times W}$ of D-Spa operation at the (i, j)th pixel position can be expressed as a linear aggregation of corresponding

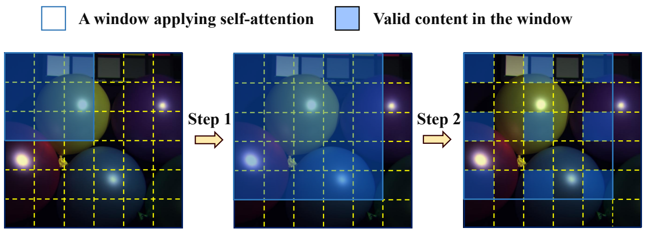


Figure 3: The dilation in D-Spa (dilation rate = 2) consists of two steps, *i.e.,* expanding and hollowing. Step 1 expands the 3×3 window to 5×5, and step 2 hollows out part of the window.

values $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$ in the local window containing the (i, j)th pixel position.

$$\mathbf{V}'_{(:,i,j)} = \sum_{(x,y)\in\Omega(i,j)} \mathbf{W}_{(i,j\to x,y)}\mathbf{V}_{(:,x,y)}, \tag{7}$$

where $\mathbf{V}_{(:,x,y)} \in \mathbb{R}^{C}$ indicates the value of the (x, y)th pixel position in the values map $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$; $\Omega\,(i, j)$ indicates the coordinate set of a dilation window which contains $k \times k$ pixel positions. In Fig. 3, the solid blue box represents the window applied self-attention. Taking the window size of 3×3 and dilation rate of 2 as an example, the window shape becomes 5×5 after dilating, and the blue patches in the window indicate the tokens that validly participates in the self-attention computation. The area $\Omega\,(i, j)$ is generated by two steps, i.e., the first step is to expand the original window, and the second is to prohibit some tokens from participating in the calculation of Spa-MSA. In Eq. 7, D is a constant variable; $\mathbf{V}'_{(:,i,j)} \in \mathbb{R}^{C}$ indicates the vector of the (i, j)th pixel position in the output feature map $\mathbf{V}' \in \mathbb{R}^{C\times H\times W}$; $\mathbf{W}_{(i,j\to x,y)} \in \mathbb{R}$ indicates an element in the attention matrix which is computed as the softmax normalization of the dot-product between the query $\mathbf{Q}_{(i,j)} \in \mathbb{R}^{C}$ and the key $\mathbf{K}_{(x,y)} \in \mathbb{R}^{C}$:

$$\mathbf{W}_{(i,j\to x,y)} = \frac{e^{\frac{1}{\sqrt{D}}\mathbf{Q}^{\mathrm{T}}_{(i,j)}\mathbf{K}_{(x,y)}}}{\mathrm{S}_i}, \tag{8}$$

where

$$\mathrm{S}_i = \sum_{x=1,y=1}^{k,k} e^{\frac{1}{\sqrt{D}}\mathbf{Q}^{\mathrm{T}}_{(i,j)}\mathbf{K}_{(x,y)}}. \tag{9}$$

By observing the generation of $\mathbf{W} \in \mathbb{R}^{k\times k}$ in the Eq. 8, the D-Spa is a convolution operation with the content-aware characteristic. In other words, it dynamically generates weights at each position. Fig. 1 above shows the properties of Spa-MSA and D-Spa. It can find that D-Spa can expand receptive fields like dilation convolution and learn the local information simultaneously. Furthermore, the D-Spa is pre-fixed, has no sliding characteristic, and adopts a multi-head attention mechanism, which groups the channels first, and each group shares a learned parameter.

3.3 G-Spe

Fully connected layer (FC) [Gardner and Dorling, 1998] is a basic linear unit in the CNNs, which connects the two hidden layers with the learnable parameters. Given input is $\mathbf{F} \in \mathbb{R}^{HW\times C}$, and the parameters of FC can be expressed as a matrix $\mathbf{W} \in \mathbb{R}^{C\times C'}$, the FC can be expressed in the form of matrix multiplication:

$$\mathbf{F}' = \mathbf{F}\mathbf{W}, \tag{10}$$

where $\mathbf{F}' \in \mathbb{R}^{HW\times C'}$ is the output of FC, and W is updated by the backpropagating gradient. However, the weight of FC is as spatial-agnostic as the vanilla convolution kernel, which does not build a relationship with the input. In order to better express the channel-wise relationship with the input, Hu *et al.* [Hu *et al.*, 2018] propose the idea of channel-attention (CA), which can be represented by:

$$\mathbf{F}' = \mathbf{F} \odot \mathbf{W}, \tag{11}$$

where $\mathbf{F}' \in \mathbb{R}^{HW\times C}$ is the output of CA, $\odot$ represents dot product operation and $\mathbf{W} \in \mathbb{R}^{C}$ is learned from the following formula:

$$\mathbf{W} = \mathit{\Phi}\,(\mathbf{F}), \tag{12}$$

where $\mathbf{W}$ is a weight learned by the network $\mathit{\Phi}$ from the input $\mathbf{F}$, whose value is content-aware with the input. From this view, the weights in Spe-MSA are also content-aware, *i.e.,* Spe-MSA generates a weight matrix using spatial similarity.

In the Spe-MSA, the weight contains spatially related information, and the matrix multiplication operation can be regarded as a dynamic FC operation on one head of Spe-MSA. Given the Spe-MSA with one head, the process can be demonstrated as follows:

$$\mathbf{V}' = \mathbf{V}\mathbf{W}, \tag{13}$$

where $\mathbf{V}' \in \mathbb{R}^{HW\times C}$ indicates the output of Spe-MSA, $\mathbf{V} \in \mathbb{R}^{HW\times C}$ means the value of Spe-MSA, and $\mathbf{W} \in \mathbb{R}^{C\times C}$ is generated by the following formula:

$$\mathbf{W}_{(i,j)} = \frac{e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}_{(:,j)}}}{\mathrm{S}_j}, \tag{14}$$

in which

$$\mathrm{S}_j = \sum_{i=1}^{C} e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}_{(:,j)}}, \tag{15}$$

where $\mathbf{Q} \in \mathbb{R}^{HW\times C}$ means the query of input; $\mathbf{K} \in \mathbb{R}^{HW\times C}$ means the key of input; $\mathbf{W}_{(i,j)}$ indicates the (i, j)th position of weight matrix $\mathbf{W} \in \mathbb{R}^{C\times C}$, which is generated by softmax normalization of the dot product between query $\mathbf{Q}_{(:,j)} \in \mathbb{R}^{HW}$ and key $\mathbf{K}_{(:,i)} \in \mathbb{R}^{HW}$; S_j is the result of summing the jth column in the matrix generated by the numerator in Eq. 14 and D is a constant variable. By comparing the weight generation in Eq. 10, Eq. 11, and Eq. 13, we can find that Spe-MSA has the dense connection properties of FC and the content-aware ability of CA, which means that Spe-MSA dynamically establishes the connection between channels. To make full use of high-resolution spatial information and local content in HR-MSI, we envisage the G-Spe as a grouped design for space. In detail, we subdivide the value $\mathbf{V} \in \mathbb{R}^{HW\times C}$, $\mathbf{Q} \in \mathbb{R}^{HW\times C}$, and $\mathbf{K} \in \mathbb{R}^{HW\times C}$ into g^2 groups, and in the kth group we get the corresponding

$\mathbf{V}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$, $\mathbf{Q}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$ and $\mathbf{K}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$, where $k \in \left\{1, 2, 3, \cdots, \frac{HW}{g^2}\right\}$. Then we calculate the weight matrix $\mathbf{W}^k \in \mathbb{R}^{C\times C}$ in the kth group independently as follows:

$$\mathbf{W}^k_{(i,j)} = \frac{e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}^k_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}^k_{(:,j)}}}{\mathrm{S}^k_j}, \tag{16}$$

where the S^k_j is calculated by the following formula:

$$\mathrm{S}^k_j = \sum_{i=1}^{C} e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}^k_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}^k_{(:,j)}}. \tag{17}$$

We will perform matrix multiplication between W^k and V^k, as shown in the following formula:

$$\mathbf{V}^{k'} = \mathbf{V}^k\mathbf{W}^k. \tag{18}$$

Each group of G-Spe realizes a kind of dynamic FC operation, *i.e.,* a content-aware weight generator. We merge together the calculated $\mathbf{V}^{k'} \in \mathbb{R}^{\frac{HW}{g^2}\times C}$ according to the spatial dimension to get the output $\mathbf{V}^{'} \in \mathbb{R}^{HW\times C}$, where $k \in \left\{1, 2, 3, \cdots, \frac{HW}{g^2}\right\}$. In this way, G-Spe realizes the grouped design along the spatial dimension through the regular space subdivision so that the model has a rich information expression capability.

Overall Loss Function: We optimize the parameters of the network in a unified and end-to-end manner. The overall loss function consists of the weighted sum of two losses:

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda_{ssim}\mathcal{L}_{ssim}, \tag{19}$$

where $\mathcal{L}_1$ means Sum of Absolute Difference, the loss $\mathcal{L}_{ssim}$ is expressed as:

$$\mathcal{L}_{ssim} = 1 - \mathrm{SSIM}(\bar{\mathcal{X}}, \tilde{\mathcal{X}}), \tag{20}$$

where the SSIM[1] means Structural SIMilarity, $\bar{\mathcal{X}}$ represents the reference, $\tilde{\mathcal{X}}$ denotes the output of our network, and λ_{ssim} is a positive hyperparameter fixed to 0.1 in our experiments.

4 Experiments

Datasets: To test the performance of our model, we conduct experiments on the CAVE [2] and Harvard[3] datasets. CAVE dataset contains 32 HSIs, including 31 spectral bands ranging from 400 nm to 700 nm at 10 nm steps. We randomly select 20 images for training the network, and the remaining 11 images constitute the testing dataset. In addition, Harvard dataset contains 77 HSIs of indoor and outdoor scenes, and each HSI has a size of 1392 × 1040 × 31, covering the spectral range from 420 nm to 720 nm. We crop the upper left part (1000 × 1000) of the 20 Harvard images, 10 of which have been used for training, and the rest has been exploited for testing.

Data Simulation: The proposed network takes LR-HSI and HR-MSI $(\mathcal{X}, \mathcal{Y})$ as input pairs, while the ground-truth (GT) for training is HR-HSI $\bar{\mathcal{X}}$. However, since HR-HSI is not available as a reference, a simulation stage is required. In our experiments using the CAVE dataset, we produce 3920 overlapping patches with a size of $64 \times 64 \times 31$ by cropping 20 chosen training images. These patches serve as the HR-HSI (ground-truth) $\bar{\mathcal{X}}$ patches. To simulate appropriate LR-HSIs, we apply a 3×3 Gaussian blur kernel with a standard deviation of 0.5 to the original HR-HSIs. We then downsample the blurred patches with a scaling factor of 4. The HR-MSI patches are generated using the common spectral response function of the Nikon D700[4] camera. Therefore, the input pairs $(\mathcal{X}, \mathcal{Y})$ consist of 3920 LR-HSI patches with a size of $16\times16\times31$ and RGB image patches with a size of $64\times64\times3$. The pairs and their related GTs are randomly divided into training data (80%) and validation data (20%). The same procedure is used to simulate the input LR-HSI and HR-MSI products and GTs for the Harvard dataset.

Benchmark: To assess the performance of our approach, we compare it with various state-of-the-art methods for MHIF. The upsampled LR-HSI in Fig. 2 is the bicubic-interpolated result, which is added to the experiment as a baseline. Model-based techniques include the MTF-GLP-HS [Selva *et al.*, 2015], the CSTF-FUS [Li *et al.*, 2018a], the LTTR[Dian *et al.*, 2019], the LTMR[Dian and Li, 2019], and the IR-TenSR[Xu *et al.*, 2022] approaches. In addition, we perform a comparison with other deep learning methods, such as the DBIN [Wang *et al.*, 2019], the SSRNet [Zhang *et al.*, 2020], the ResTFNet [Liu *et al.*, 2020], the HSRNet [Hu *et al.*, 2021], the MoG-DCN [Dong *et al.*, 2021], the Fusformer [Hu *et al.*, 2022] and the DHIF [Huang *et al.*, 2022] network. All the deep learning approaches are trained with the same input pairs for a fair comparison. Moreover, the related hyperparameters are selected consistent with the original papers.

Implementation Details: The proposed network implements in PyTorch 1.11.0 and Python 3.7.0 using AdamW optimizer with a learning rate of 0.0001 to minimize $\mathcal{L}_{total}$ by 2000 epochs and Linux operating system with a NVIDIA RTX3090 GPU.

Results on CAVE Dataset: We test our model on the CAVE dataset. Fig. 4 presents the 11 testing images in an RGB color composition. From Tab. 1, we can see that the proposed approach overcomes the other methods in 4 quality indexes (QIs), *i.e.,* PSNR, SAM, ERGAS, and SSIM. Specifically, we observe an improvement of ∼1.30/4.93/8.11/0.028% in PSNR/SAM/ERGAS/SSIM compared to the second best method, *i.e.,* MoG-DCN [Dong *et al.*, 2021]. Compared with the third best method, DHIF [Huang *et al.*, 2022], our approach gets the gains ∼2.41/3.98/16/0.09% in PSNR/SAM/ERGAS/SSIM. In terms of visual assessments (see Fig. 5), we present the pseudo-color representations of the fused products and some error maps to aid the visual inspection. Compared to the benchmark, our approach has better details and visual effects. Having a look at the error maps, the reconstruction of BDT is closest to the all zero map, and significantly lower values than compared approaches.

Results on Harvard Dataset: Besides, we evaluate the performance of our BDT on another hyperspectral image dataset

[1] https://en.wikipedia.org/wiki/Structural_similarity

[2] https://www.cs.columbia.edu/CAVE/databases/multispectral/

[3] http://vision.seas.harvard.edu/hyperspec/index.html

[4] https://www.maxmax.com/nikon_d700_study.htm

Methods	CAVE					Harvard				
	PSNR	SAM	ERGAS	SSIM	#params	PSNR	SAM	ERGAS	SSIM	#params
Bicubic	34.33±3.88	4.45±1.62	7.21±4.90	0.944±0.0291	–	38.71±4.33	2.53±0.67	4.45±41.81	0.948±0.0268	–
MTF-GLP-HS [Selva *et al.*, 2015]	37.69±3.85	5.33±1.91	4.57±2.66	0.973±0.0158	–	33.81±3.50	6.25±2.42	3.47±1.82	0.952±0.0321	–
CSTF-FUS [Li *et al.*, 2018a]	34.46±4.28	14.37±5.30	8.29±5.29	0.866±0.0747	–	39.13±3.50	6.91±2.66	4.64±1.80	0.913±0.0487	–
LTTR[Dian *et al.*, 2019]	35.85±3.49	6.99±2.55	5.99±2.92	0.956±0.0288	–	37.91±3.58	5.35±1.94	2.44±1.06	0.972±0.0183	–
LTMR[Dian and Li, 2019]	36.54±3.30	6.71±2.19	5.39±2.53	0.963±0.0208	–	38.41±3.58	5.05±1.70	2.24±0.97	0.970±0.0166	–
IR-TenSR[Xu *et al.*, 2022]	35.61±3.45	12.30±4.68	5.90±3.05	0.945±0.0267	–	40.47±3.04	4.36±1.52	5.57±1.57	0.962±0.0140	–
DBIN [Wang *et al.*, 2019]	50.83±4.29	2.21±0.63	1.24±1.06	0.996±0.0026	0.469M	47.88±3.87	2.31±0.46	1.95±0.81	0.988±0.0066	0.469M
ResTFNet [Liu *et al.*, 2020]	45.58±5.47	2.82±0.70	2.36±2.59	0.993±0.0056	2.387M	45.93±4.35	2.61±0.69	2.56±1.32	0.985±0.0082	2.387M
SSRNet [Zhang *et al.*, 2020]	48.62±3.92	2.54±0.84	1.63±1.21	0.995±0.0023	**0.027M**	47.95±3.37	2.31±0.60	2.30±1.42	0.987±0.0070	**0.027M**
HSRNet [Hu *et al.*, 2021]	50.38±3.38	2.23±0.66	1.20±0.75	0.996±0.0014	0.633M	48.29±3.03	2.26±0.56	1.87±0.81	0.988±0.0064	0.633M
MoG-DCN [Dong *et al.*, 2021]	51.63±4.10	2.03±0.62	1.11±0.82	0.997±0.0018	6.840M	47.89±4.09	2.11±0.52	1.89±0.82	0.988±0.0073	6.840M
Fusformer [Hu *et al.*, 2022]	49.98±8.10	2.20±0.85	2.50±5.21	0.994±0.0111	0.504M	47.87±5.13	2.84±2.07	2.04±0.99	0.986±0.0101	0.467M
DHIF [Huang *et al.*, 2022]	51.07±4.17	2.01±0.63	1.22±0.97	0.997±0.0016	22.462M	47.68±3.85	2.32±0.53	1.95±0.92	0.988±0.0074	22.462M
BDT (ours)	**52.30±3.98**	**1.93±0.55**	**1.02±0.77**	**0.997±0.0014**	2.668 M	**48.83±3.45**	**2.07±0.49**	**1.83±0.81**	**0.989±0.0067**	2.668 M
Ideal value	∞	**0**	**0**	**1**	-	∞	**0**	**0**	**1**	-

Table 1: Average quantitative comparisons on 11 CAVE examples and 10 Harvard examples simulating a scaling factor of 4. The best values are highlighted in bold, and the second best values are underlined. M refers to millions.

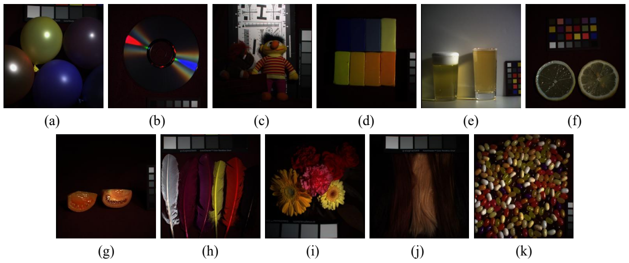


Figure 4: The testing images from the CAVE dataset: (a) *balloons*, (b) *cd*, (c) *chart and stuffed toy*, (d) *clay*, (e) *fake and real beers*, (f) *fake and real lemon slices*, (g) *fake and real tomatoes*, (h) *feathers*, (i) *flowers*, (j) *hairs*, and (k) *jelly beans*. An RGB color representation is used to depict the images.

(*i.e.,* Harvard). We consider the original HSI as ground-truth, and simulate the LR-HSI in the same way as the CAVE dataset. From Tab. 1, the results show that deep learning approaches outperform traditional ones. Our method gets the best results (outperforms high-performance approaches such as DHIF and Fusformer). The proposed approach shows an excellent trade-off between performance and computational costs on the Harvard dataset.

4.1 Ablation Study

In this section, we provide an in-depth discussion of D-Spa and G-Spe in the BDT to demonstrate their effectiveness and rationale. We compare their performance with ablation on self-structure and other existing networks. To maintain generality and conciseness, we present our analysis based on the CAVE dataset.

1) D-Spa and G-Spe: To verify the effectiveness, in Tab. 2, results show that replacing D-Spa with Spa-MSA will bring the performance gain, and replacing G-Spe with Spe-MSA will also boost performance. And our BDT utilizes both D-Spa and G-Spe obtaining the best results. It proves that the designed modules boost performance of networks. Please note that Spa-MSA and Spe-MSA indicates the dilation 1 of D-Spa and the group 1 of G-Spe in BDT, respectively.

D-Spa	G-Spe	PSNR	SAM	ERGAS	SSIM
✓	✓	**52.30±3.98**	**1.93±0.55**	**1.02±0.77**	**0.997±0.0014**
✓	✗	52.03±3.79	2.02±0.59	1.04±0.75	0.997±0.0014
✗	✓	51.96±3.72	2.03±0.59	1.04±0.74	0.997±0.0013
✗	✗	51.91±3.77	2.02±0.59	1.05±0.76	0.997±0.0014

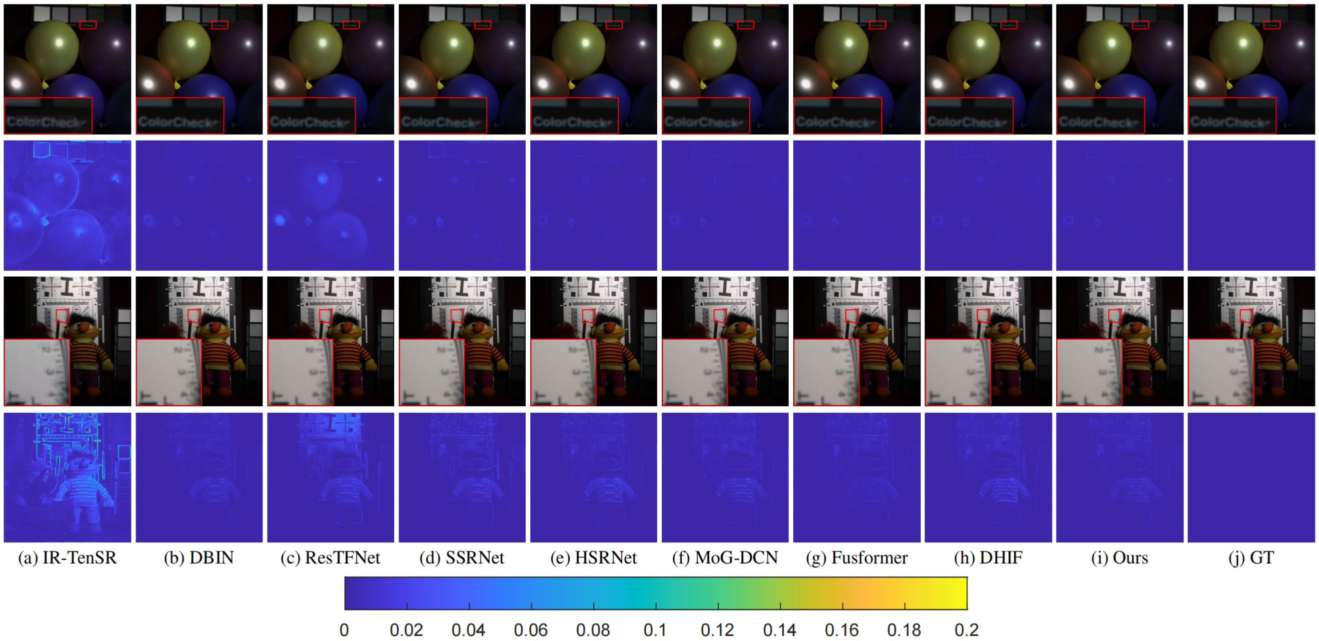
Table 2: The average four QIs and the corresponding parameters on the CAVE dataset simulating a scaling factor of 4.

Methods	PSNR	SAM	ERGAS	SSIM
Swin-Shift	51.47 ± 3.88	2.08 ± 0.60	1.09 ± 0.81	0.997 ± 0.0015
Swin-D	**51.57** ± **4.00**	**2.04** ± **0.58**	**1.10** ± **0.85**	**0.997** ± **0.0016**
Restormer-T	50.67 ± 4.36	2.34 ± 0.72	1.29 ± 1.06	0.996 ± 0.0024
Restormer-G	**51.16** ± **3.93**	**2.22** ± **0.67**	**1.15** ± **0.79**	**0.996** ± **0.0017**

Table 3: The average four QIs and the corresponding parameters on the CAVE dataset simulating a scaling factor of 4.

2) Embedding in existing networks: We test D-Spa against the Shifted Window operation in Swin Transformer and G-Spe against the Spe-MSA in Restormer. We use a Swin Transformer structure comparing the Shifted Window approach (Swin-Shift) with our D-Spa (Swin-D). And we also compare Restormer network structure with the transpose MSA (T-MSA) approach (Restormer-T) and our G-Spe (Restormer-G). After using the proposed D-Spa and G-Spe, the performance of Swin Transformer and Restormer have corresponding enhancement in Tab. 3. It proves that the proposed D-Spa and G-Spe improve the network performance for solving the MHIF task.

3) Spatial grouped design in G-Spe: In the Tab. 4, we tested the performance of Spe-MSA and G-Spe without spectral multi-head (w/o head), using BFE in BDT as the backbone. Specifically, the Spe-MSA structure is grouped in the spectral dimension, and the G-Spe structure is grouped in the



(a) IR-TenSR (b) DBIN (c) ResTFNet (d) SSRNet (e) HSRNet (f) MoG-DCN (g) Fusformer (h) DHIF (i) Ours (j) GT

Figure 5: The first and third rows show the results using the pseudo-color representation on "*balloons*" and "*chart and stuffed toy*", respectively, from the CAVE dataset. Some close-ups are depicted in the red rectangles. The second and fourth rows show the residuals between the GT and the fused products. (a) IR-TenSR [Xu *et al.*, 2022], (b) DBIN [Wang *et al.*, 2019], (c) ResTFNet [Liu *et al.*, 2020], (d) SSRNet [Zhang *et al.*, 2020], (e) HSRNet [Hu *et al.*, 2021], (f) MoG-DCN [Dong *et al.*, 2021], (g) Fusformer [Hu *et al.*, 2022], (h) DHIF [Huang *et al.*, 2022], (i) Ours, and (j) GT.

spatial dimension. The result shows that the effect of the spatial grouped design outperforms slightly than spectral dimension on the MHIF task.

Methods	PSNR	SAM	ERGAS	SSIM	#Flops
Spe-MSA	52.03 ± 3.79	2.02 ± 0.59	1.04 ± 0.75	0.997 ± 0.0014	33.52G
w/o head	**52.09** ± **3.78**	**2.00** ± **0.58**	**1.03** ± **0.75**	**0.997** ± **0.0013**	33.87G

Table 4: The average four QIs and the corresponding flops on the CAVE dataset simulating a scaling factor of 4. w/o head means G-Spe without spectral multi-head. G means gillions.

4) D-Spa with different dilations: We investigated the impact of different dilation rates on the MHIF task by designing D-Spa. The proposed D-Spa has adjustable dilations that can expand and hollow the window shown in Fig.3, thereby increasing the receptive field. As shown in Tab.5, we found that a dilation rate of 2 yields the best results. Thus, D-Spa can provide a long-range response from a flexible range, and it outperforms Spa-MSA in terms of achieving better results.

Method	PSNR	SAM	ERGAS	SSIM
$d = 1$	52.03±3.79	2.02±0.59	1.04±0.75	0.997±0.0014
$d = 2$	**52.30±3.98**	**1.93±0.55**	**1.02±0.77**	**0.997±0.0014**
$d = 3$	51.51±3.91	2.18±0.65	1.11±0.83	0.997±0.0019

Table 5: The average four QIs on the CAVE dataset simulating a scaling factor of 4. d indicates the dilation rate in D-Spa.

5) Test of multi-scaled input in bidirectional branch: We gradually reduced the participation of the spectral branch in the BFF process. The results in Tab. 6 show the spectral branch plays a vital role in the restoration of image details.

$\mathcal{G}_1$	$\mathcal{G}_3$	$\mathcal{G}_3$	PSNR	SAM	ERGAS	SSIM	#Flops
✓	✓	✓	**52.30±3.98**	**1.93±0.55**	**1.02±0.77**	**0.997±0.0014**	33.52G
✗	✓	✓	52.04±3.84	1.99±0.57	1.03±0.76	0.997±0.0014	33.44G
✗	✗	✓	51.91±3.70	2.02±0.59	1.03±0.73	0.997±0.0012	33.10G
✗	✗	✗	50.72±3.48	4.48±1.38	3.84±1.15	0.993±0.0013	**27.74G**

Table 6: The four average QIs and the corresponding flops on the 11 testing images from the CAVE dataset simulating a scaling factor of 4. $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ indicate the output which is the result of G-Spe in spectal branch. G refers gillions.

5 Conclusions

This paper proposes the BDT, a Transformer fusion framework, to address the MHIF problem, which employs D-Spa, G-Spe, and bidirectional modules. Specifically, motivated by the MHIF problem, D-Spa and G-Spe are used for spatial and spectral information extraction, respectively. Moreover, the proposed BDT can extract and fuse multi-scale information to obtain high-quality results with moderate parameters.

Acknowledgments

This research is supported by NSFC (12271083) and Natural Science Foundation of Sichuan Province (2022NSFSC0501).

References

[Cao *et al.*, 2020] Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Trans. Geosci. Remote Sens.*, 58(7):4604–4616, 2020.

[Dian and Li, 2019] Renwei Dian and Shutao Li. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. *IEEE Trans. Image Process.*, 28(10):5135–5146, 2019.

[Dian *et al.*, 2019] Renwei Dian, Shutao Li, and Leyuan Fang. Learning a low tensor-train rank representation for hyperspectral image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(9):2672–2683, 2019.

[Dong *et al.*, 2021] Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Trans. Image Process.*, 30:5754–5768, 2021.

[Feng and Sun, 2012] Yaoze Feng and Dawen Sun. Application of hyperspectral imaging in food safety inspection and control: a review. *Crit. Rev. Food Sci. Nutr.*, 52(11):1039–1058, 2012.

[Gao *et al.*, 2006] Bocai Gao, C Davis, and A Goetz. A review of atmospheric correction techniques for hyperspectral remote sensing of land surfaces and ocean color. In *IEEE International Symposium on Geoscience and Remote Sensing*, pages 1979–1981. IEEE, 2006.

[Gardner and Dorling, 1998] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.*, 32(14-15):2627–2636, 1998.

[Guo *et al.*, 2020] Penghao Guo, Peixian Zhuang, and Yecai Guo. Bayesian pan-sharpening with multiorder gradient-based deep network constraints. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:950–962, 2020.

[Han *et al.*, 2021] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *ICLR*, 2021.

[Hong *et al.*, 2021] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2021.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, June 2018.

[Hu *et al.*, 2021] Jinfan Hu, Tingzhu Huang, Liangjian Deng, Taixiang Jiang, Gemine Vivone, and Jocelyn Chanussot. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[Hu *et al.*, 2022] Jinfan Hu, Tingzhu Huang, Liangjian Deng, Hongxia Dou, Danfeng Hong, and Gemine Vivone. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5, 2022.

[Huang *et al.*, 2022] Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, Xin Li, and Guangming Shi. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Trans. Comput Imaging.*, 8:201–214, 2022.

[Jiao *et al.*, 2023] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, pages 1–14, 2023.

[Kolesnikov *et al.*, 2021] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[Li *et al.*, 2018a] Shutao Li, Renwei Dian, Leyuan Fang, and José M Bioucas-Dias. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Trans. Image Process.*, 27(8):4118–4130, 2018.

[Li *et al.*, 2018b] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.

[Li *et al.*, 2021] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *CVPR*, pages 12321–12330, 2021.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021.

[Liu *et al.*, 2020] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion.*, 55:1–15, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[Lu *et al.*, 2020] Bing Lu, Phuong D Dao, Jiangui Liu, Yuhong He, and Jiali Shang. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens. (Basel).*, 12(16):2659, 2020.

[Ma *et al.*, 2021] Qing Ma, Junjun Jiang, Xianming Liu, and Jiayi Ma. Learning a 3d-cnn and transformer prior for hyperspectral image super-resolution. *arXiv preprint arXiv:2111.13923*, 2021.

[Meng *et al.*, 2022] Xiangchao Meng, Nan Wang, Feng Shao, and Shutao Li. Vision transformer for pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 60:1–11, 2022.

[Piqueras *et al.*, 2011] S Piqueras, L Duponchel, R Tauler, and A De Juan. Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares. *Anal. Chim. Acta.*, 705(1-2):182–192, 2011.

[Selva *et al.*, 2015] Massimo Selva, Bruno Aiazzi, Francesco Butera, Leandro Chiarantini, and Stefano Baronti. Hyper-sharpening: A first approach on sim-ga data. *IEEE J. Sel. Top Appl. Earth Obs. Remote Sens.*, 8(6):3008–3024, 2015.

[Wang *et al.*, 2019] Wu Wang, Weihong Zeng, Yue Huang, Xinghao Ding, and John Paisley. Deep blind hyperspectral image fusion. In *ICCV*, October 2019.

[Wu *et al.*, 2011] Jian Wu, Dao-Li Peng, et al. Advances in researches on hyperspectral remote sensing forestry information-extracting technology. *Spectrosc. Spect. Anal.*, 31(9):2305–2312, 2011.

[Xu *et al.*, 2022] Ting Xu, Tingzhu Huang, Liangjian Deng, and Naoto Yokoya. An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16, 2022.

[Yan *et al.*, 2022] Keyu Yan, Man Zhou, Jie Huang, Feng Zhao, Chengjun Xie, Chongyi Li, and Danfeng Hong. Panchromatic and multispectral image fusion via alternating reverse filtering network. *NeurIPS*, 2022.

[Yang *et al.*, 2020a] Yong Yang, Chenxu Wan, Shuying Huang, Hangyuan Lu, and Weiguo Wan. Pansharpening based on low-rank fuzzy fusion and detail supplement. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:5466–5479, 2020.

[Yang *et al.*, 2020b] Yong Yang, Lei Wu, Shuying Huang, Weiguo Wan, Wei Tu, and Hangyuan Lu. Multiband remote sensing image pansharpening based on dual-injection model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:1888–1904, 2020.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Minghsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[Zhang *et al.*, 2020] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Trans. Geosci. Remote Sens.*, 59(7):5953–5965, 2020.

[Zhou *et al.*, 2021] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. Effective pansharpening with transformer and invertible neural network. *IEEE Trans. Geosci. Remote Sens.*, 60, 2021.

[Zhou *et al.*, 2022] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pansharpening. In *ECCV*, pages 274–291. Springer, 2022.