Dual-Window Multiscale Transformer for Hyperspectral Snapshot Compressive Imaging

Fulin Luo[1], **Xi Chen**[1], **Xiuwen Gong**[2], **Weiwen Wu**[3], **Tan Guo**[4*]
[1]College of Computer Science, Chongqing University
[2]Faculty of Engineering, The University of Sydney
[3]Department of Biomedical Engineering, Sun-Yat-sen University
[4]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications
luoflyn@163.com, chenxi2000cd@gmail.com, gongxiuwen@gmail.com,
wuweiw7@mail.sysu.edu.cn, guot@cqupt.edu.cn

Abstract

Coded aperture snapshot spectral imaging (CASSI) system is an effective manner for hyperspectral snapshot compressive imaging. The core issue of CASSI is to solve the inverse problem for the reconstruction of hyperspectral image (HSI). In recent years, Transformer-based methods achieve promising performance in HSI reconstruction. However, capturing both long-range dependencies and local information while ensuring reasonable computational costs remains a challenging problem. In this paper, we propose a Transformer-based HSI reconstruction method called dual-window multiscale Transformer (DWMT), which is a coarse-to-fine process, reconstructing the global properties of HSI with the long-range dependencies. In our method, we propose a novel U-Net architecture using a dual-branch encoder to refine pixel information and full-scale skip connections to fuse different features, enhancing the extraction of fine-grained features. Meanwhile, we design a novel self-attention mechanism called dual-window multiscale multi-head self-attention (DWM-MSA), which utilizes two different-sized windows to compute self-attention, which can capture the long-range dependencies in a local region at different scales to improve the reconstruction performance. We also propose a novel position embedding method for Transformer, named con-abs position embedding (CAPE), which effectively enhances positional information of the HSIs. Extensive experiments on both the simulated and the real data are conducted to demonstrate the superior performance, stability, and generalization ability of our DWMT. Code of this project is at https://github.com/chenx2000/DWMT.

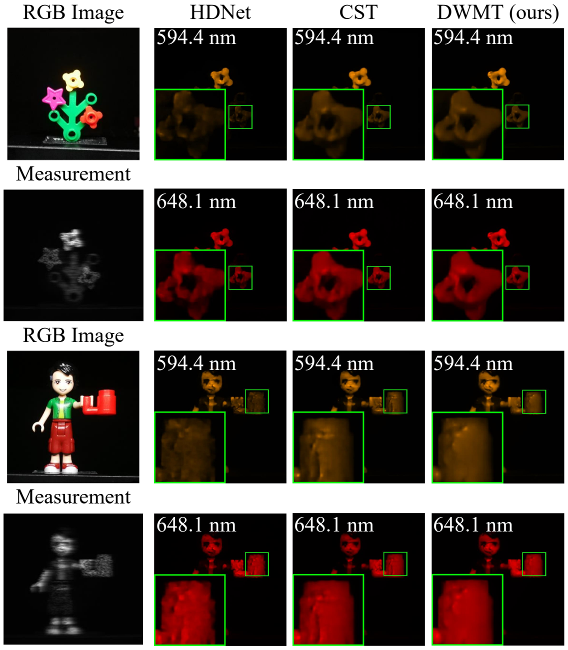

Figure 1: Real HSI reconstructed results of our DWMT compared with three end-to-end trained methods on two (out of 28) spectral channels of *Scene* 1 and *Scene* 3.

Introduction

Hyperspectral images (HSI) consist of multiple contiguous spectral bands, providing richer spatial and spectral information compared with RGB images. Since HSI can reveal the spectral properties of object, it has been widely used in computer vision (CV) tasks such as image classification (Guo et al. 2023), change detection (Luo et al. 2023), and medical image processing (Meng et al. 2020). Therefore, it is highly valuable to explore efficient and cost-effective hyperspectral imaging techniques. The traditional hyperspectral image (Schechner and Nayar 2002) typically is captured with hyperspectral imager that is a very expensive equipment. To reduce the imaging costs, the snapshot compressive imaging (SCI) systems are developed to capture HSIs, using a 2D measurement to reconstruct a 3D HSI with a computational spectral imaging algorithm (Yuan, Brady, and Katsaggelos 2021). This significantly improves imaging efficiency and enables capturing dynamic scenes. Among these systems, coded aperture snapshot spectral imaging (CASSI) (Gehm et al. 2007) has shown impressive performance. CASSI modulates the HSI signals of different wavelengths through the coded aperture and disperser, and then combines all the modulated signals to achieve 2D measurement. The core task of CASSI is to address the ill-posed inverse problem, recovering a 3D HSI from the 2D measurement.

The traditional methods (Yuan 2016; Liu et al. 2018;

[*]Corresponding author.

Zhang et al. 2019) for HSI reconstruction utilize handcrafted priors to regularize the reconstruction process, generally resulting in poor generalization and limited reconstruction quality. Recently, CNN-based methods (Miao et al. 2019; Meng, Ma, and Yuan 2020; Hu et al. 2022) have been extensively employed to construct the 3D HSI from the 2D measurement. However, CNN-based methods still have limitations in capturing long-range dependencies to achieve global properties for HSI reconstruction.

Transformer (Vaswani et al. 2017) has achieved tremendous success in the field of natural language processing (NLP), which has motivated researchers to explore the application of Transformer in computer vision (Dosovitskiy et al. 2021). The vision Transformer (ViT) was first proposed in CV to achieve the long-range dependencies between tokens, overcoming the limitations of CNNs. For HSI reconstruction, Transformers also generate impressive results (Cai et al. 2022a,b). However, the global Transformers (Dosovitskiy et al. 2021) have high computational cost, while the local Transformers (Liu et al. 2021) have limitations in capturing long-range dependencies. Due to HSIs containing a large number of channels, a challenging problem is to simultaneously capture the long-range dependencies and the local information with reasonable computational costs.

To balance the long-range dependencies and the local representation, we propose a dual-window multiscale Transformer (DWMT) for HSI reconstruction. In this method, we introduce a novel attention mechanism, namely dual-window multiscale mult-head self-attention (DWM-MSA). Based on the divide-and-conquer approach, our DWM-MSA is divided into four branches. Two branches perform self-attention within the window to achieve the local information, while the other two branches shuffle the tokens to perform the long-range dependencies. We implement the long-range dependencies in two local regions with different window sizes, which can balance the global reconstruction properties and the computational cost with a local manner. By using diverse window sizes, the varying scales of features and details can be captured. In addition, we introduce a novel position embedding method, con-abs position embedding (CAPE), to enhance the positional information of image. As shown in Figure 2, our method produces the reconstructed images on the real dataset that are clearer compared with the other methods (Hu et al. 2022; Cai et al. 2022a), and our method recovers more fine details in the images. The main contributions of this paper include the following:

- An end-to-end Transformer algorithm, DWMT, is proposed for HSI reconstruction, consisting of two stages, *i.e.*, coarse feature extraction and fine pixel refinement.
- We design a novel attention mechanism, DWM-MSA, which can simultaneously capture local information, long-range dependencies and multi-scale information.
- A novel position embedding method, CAPE, is developed to enhance the accuracy of positional information within the image efficiently.
- We develop a fine pixel refinement branch to enhance the representation of pixels, which improves the reconstruction performance of our method.

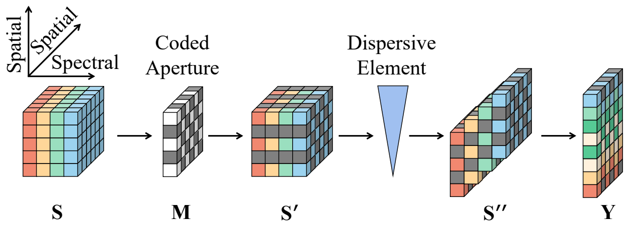

Figure 2: Schematic diagram of CASSI.

Related Work

Coded Aperture Snapshot Spectral Imaging

Coded aperture snapshot spectral imaging (CASSI) system utilizes a coded aperture (physical mask) and one or more dispersive elements to modulate HSI signals at different wavelengths, which can capture a 2D projection of the 3D HSI (Gehm et al. 2007; Wagadarikar et al. 2008). The schematic diagram of CASSI is shown in Figure 2.

Specifically, we denote the 3D HSI as $\mathbf{S} \in \mathbb{R}^{H\times W\times N}$, where *H*, *W*, and *N* represent the height, width, and number of spectral channels, respectively. In the CASSI system, the first step is to modulate the initial signal $\mathbf{S}$ using the coded aperture $\mathbf{M} \in \mathbb{R}^{H\times W}$ (Cai et al. 2022a,b). $\mathbf{S}' \in \mathbb{R}^{H\times W\times N}$ denotes the modulated HSI, then the n^{th} wavelength of $\mathbf{S}'$ can be represented as:

$$\mathbf{S}'_n = \mathbf{M} \odot \mathbf{S}_n \tag{1}$$

where $n \in [1, ..., N]$ is the n^{th} spectral channel, and $\odot$ denotes the element-wise multiplication. The dispersive element contains a linear dispersion α and a center wavelength λ_c (Gehm et al. 2007). After passing through the dispersive element, $\mathbf{S}'$ undergoes a shear along the y-axis, resulting in the sheared HSI $\mathbf{S}'' \in \mathbb{R}^{H\times(W+\alpha(N-1))\times N}$. α denotes the linear dispersion which can be understood as the shifting step. Then, $\mathbf{S}''$ can be formulated as:

$$\mathbf{S}''(x', y', n) = \mathbf{S}'(x, y + \alpha(\lambda_n - \lambda_c), n) \tag{2}$$

where (x', y') denotes the coordinates on the detector plane, and λ_n represents the wavelength of n^{th} spectral channel. Therefore, $\alpha(\lambda_n - \lambda_c)$ denotes the spatial displacement of the n^{th} channel. Consequently, the 2D measurement $\mathbf{Y} \in \mathbb{R}^{H\times(W+\alpha(N-1))}$ captured by CASSI can be expressed as:

$$\mathbf{Y} = \sum_{n=1}^{N} \mathbf{S}''(:, :, N) + E \tag{3}$$

where $\mathbf{E} \in \mathbb{R}^{H\times(W+\alpha(N-1))}$ is the measurement noise. In summary, the core issue of HSI reconstruction is to recover the HSI $\mathbf{S}$ with the 2D measurement $\mathbf{Y}$.

HSI Reconstruction

HSI reconstruction is a crucial step in hyperspectral SCI. However, the reconstruction problem is ill-posed, necessitating an appropriate prior representation of HSI. Traditional HSI reconstruction methods (Wagadarikar et al. 2008; Yuan 2016; Liu et al. 2018; Zhang et al. 2019) rely on predefined

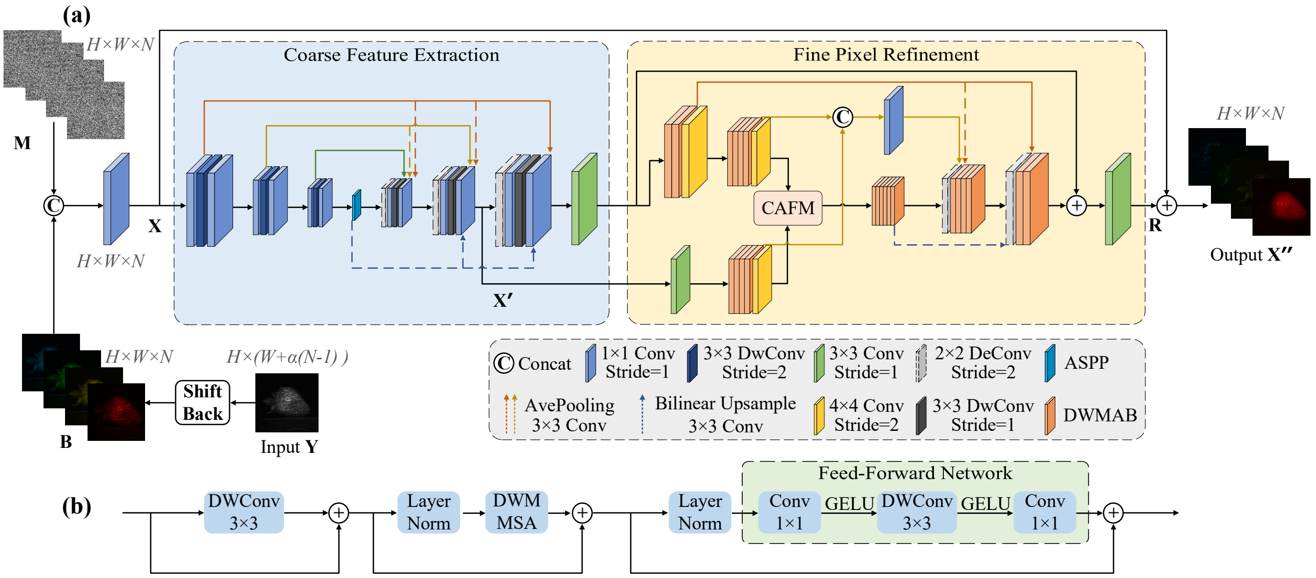

Figure 3: (a) Overall architecture of our DWMT. (b) Dual-window multiscale attention block (DWMAB).

hand-crafted priors. For example, GAP-TV (Yuan 2016) adopts total variation (TV) regularization as the prior for each spectral band. However, these approaches suffer from poor generalization and limited performance. Recently, deep learning-based methods (Miao et al. 2019; Meng, Ma, and Yuan 2020; Hu et al. 2022; Cai et al. 2022a,b) have shown the capability to directly learn image priors by training on large-scale datasets. For example, TSA-Net (Meng, Ma, and Yuan 2020) introduces a spatial-spectral self-attention, which sequentially processes each dimension in an order-independent manner. However, CNN-based methods have certain limitations in capturing long-range dependencies. While Transformer-based methods (Cai et al. 2022a,b) are effective to learn the long-range dependencies. However, it is very key to balance the long-range dependencies and the local representation for reasonable computational cost.

Transformer

Vision Transformer Transformer (Vaswani et al. 2017) is a network architecture based on the attention mechanism, completely abandoning recurrence and convolution. It was initially introduced in the field of NLP for machine translation tasks, and achieved outstanding performance. Inspired by the success of Transformer in NLP, researchers developed Vision Transformer (ViT) (Dosovitskiy et al. 2021). ViT has shown promising performance in various CV tasks, such as semantic segmentation (Dai et al. 2021a; Zhu et al. 2021), object detection (Dai et al. 2021b; Gao et al. 2021), and image classification (Bhojanapalli et al. 2021; Lanchantin et al. 2021), *etc*. However, the high computational demand of the global Transformer cannot be ignored, as its computational complexity is a multiple of the quadratic power of image size.

ViT Variants To reduce the computational cost of global Transformers, some works (Liu et al. 2021; Wang et al. 2022) have adopted local Transformers. Swin-Transformer (Liu et al. 2021) utilizes non-overlap window and performs self-attention in the window, exhibiting linear computational complexity with respect to the image size. To achieve a larger receptive field while ensuring reasonable computational cost, some works adopt the dilated window (Tu et al. 2022; Wang et al. 2022). Furthermore, some works (Xia et al. 2022; Ren et al. 2022; Zhu et al. 2023) utilize the sparse adaptive patterns to compute the attention. However, for HSI, the multiple channels result in the computational cost of global Transformer impractically, while the local Transformer struggles with capturing the long-range dependencies of HSI. Therefore, we combine the advantages of the global and local Transformers, effectively balancing the long-range dependencies and the computational cost.

Dual-Window Multiscale Transformer

The overall structure of DWMT is illustrated in Figure 3. We develop an end-to-end framework based on Transformer for HSI reconstruction, consisting of two parts, *i.e.*, coarse feature extraction and fine pixel refinement. Inspired by (Cao et al. 2022; Cai et al. 2022a), we design two U-Net (Ronneberger, Fischer, and Brox 2015) architectures separately for the two parts, and full-scale skip connections (Huang et al. 2020) are utilized. Among them, we employ an enhanced U-Net architecture with a dual-branch encoder for fine pixel refinement, which includes the most crucial part in the proposed DWMT network, *i.e.*, dual-window multiscale attention block (DWMAB).

Overall Architecture

Firstly, the 2D measurement $\mathbf{Y} \in \mathbb{R}^{H\times(W+\alpha(N-1))}$ is shifted back to a 3D image. Thus, the n^{th} wavelength of the shifted signal $\mathbf{B} \in \mathbb{R}^{H\times W\times N}$ can be represented as:

$$\mathbf{B}_n(x,y) = \mathbf{Y}(x, y - \alpha((\lambda_n - \lambda_c)) \qquad (4)$$

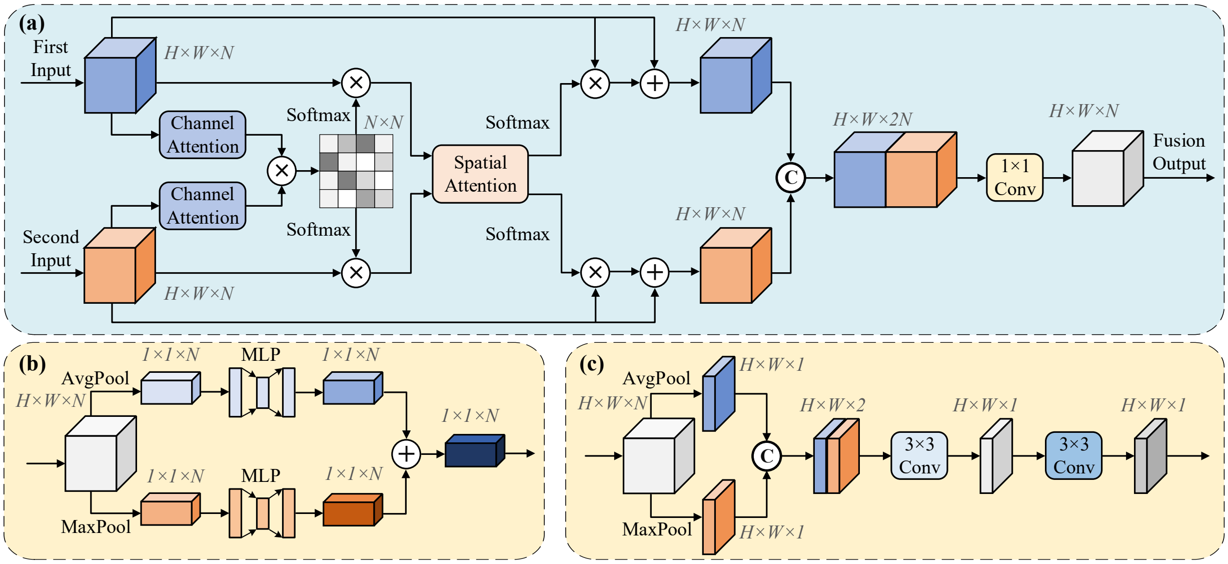

Figure 4: (a) Architecture of cross attention fusion module (CAFM). (b) Channel attention module. (c) Spatial attention module.

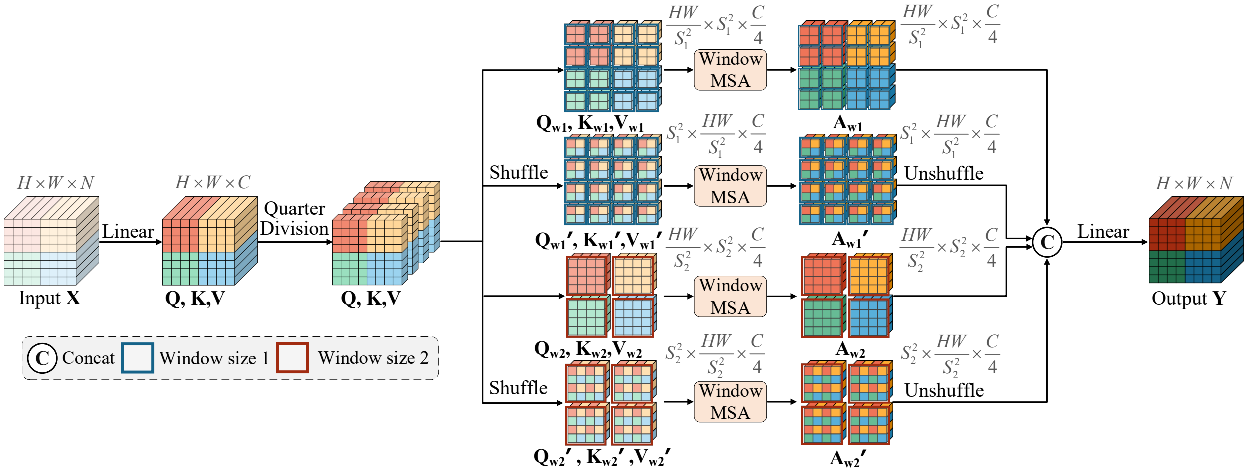

Figure 5: Schematic diagram of dual-window multiscale multi-head self-attention (DWM-MSA).

where $n \in [1, ..., N]$ is the n^{th} channel. Then, we copy the coded aperture $\mathbf{M}$ for N (the number of channels) times and concatenate them with $\mathbf{B}$. The obtained result is subsequently fed into a 1×1 convolution (Conv) layer for further processing, and its output $\mathbf{X} \in \mathbb{R}^{H\times W\times N}$ is passed to the coarse feature extraction and fine pixel refinement. Next, we add the result of fine pixel refinement $\mathbf{R} \in \mathbb{R}^{H\times W\times N}$ to X, resulting in the reconstructed HSI $\mathbf{X}'' \in \mathbb{R}^{H\times W\times N}$.

If we denote the ground truth HSI as $\mathbf{X}^* \in \mathbb{R}^{H\times W\times N}$, then the overall network loss function can be represented as:

$$\mathcal{L} = \|\mathbf{X}'' - \mathbf{X}^*\|_2 \tag{5}$$

Coarse Feature Extraction The coarse feature extraction part is a U-Net architecture comprising of a four-stage encoder-decoder. In the encoder, we use a 3×3 depth-wise convolution (DwConv) layer for downsampling. The bottleneck layer utilizes an atrous spatial pyramid pooling (ASPP) module (Chen et al. 2017, 2018). In the decoder, we use a 2×2 deconvolution (DeConv) layer for upsampling.

Fine Pixel Refinement The fine pixel refinement part consists of a three-stage encoder-decoder architecture, where the encoder is a dual-branch structure. In the first branch of the encoder, each stage employs a 4×4 Conv layer for downsampling. The second branch takes the features from a smaller scale and processes them through DWMABs. These processed features are fused with the results obtained from the first branch of the encoder for feature enhancement. According to (Woo et al. 2018; Zhou et al. 2023), we design a cross attention fusion module (CAFM) to effectively integrate the features from the two encoder branches. The schematic diagram of CAFM is shown in Figure 4. In the decoder, we use a 2×2 DeConv layer for upsampling.

Dual-Window Multiscale Attention Block

To combine the advantages of global Transformer and local Transformer, we propose a novel Transformer module called dual-window multiscale attention block (DWMAB). We apply conditional position embedding (CPE) (Chu et al. 2023) to the input $\mathbf{X}_{in} \in \mathbb{R}^{H\times W\times N}$ via a 3×3 DwConv layer:

$$\mathbf{X} = \mathrm{CPE}(\mathbf{X}_{in}) + \mathbf{X}_{in} = \mathrm{DwConv}(\mathbf{X}_{in}) + \mathbf{X}_{in} \tag{6}$$

Dual-Window Multiscale Multi-Head Self-Attention Inspired by ShuffleNet (Zhang et al. 2018) and Half-Shuffle Transformer (Cai et al. 2022c), we propose a novel self-attention mechanism called dual-window multiscale multi-head self-attention (DWM-MSA) based on the divide-and-conquer approach, which effectively captures the long-range dependencies in a local region. Additionally, it enables to capture the features and details at different scales with two different-sized windows. The schematic diagram of DWM-MSA is depicted in Figure 5.

Multiscale Window Partition First, we linearly project $\mathbf{X}$ to obtain the query, key, and value:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{q},\ \mathbf{K} = \mathbf{X}\mathbf{W}^{k},\ \mathbf{V} = \mathbf{X}\mathbf{W}^{v} \tag{7}$$

where $\mathbf{W}^{q}$, $\mathbf{W}^{k}$, $\mathbf{W}^{v}$ represent the projection weights for the query, key and value, respectively. Then, we divide the query, key, and value into four equal parts along the channel dimension: $\mathbf{Q} = [\mathbf{Q}_{w1},\ \mathbf{Q}'_{w1},\ \mathbf{Q}_{w2},\ \mathbf{Q}'_{w2}]$, $\mathbf{K} = [\mathbf{K}_{w1},\ \mathbf{K}'_{w1},\ \mathbf{K}_{w2},\ K'_{w2}]$ and $\mathbf{V} = [\mathbf{V}_{w1},\ \mathbf{V}'_{w1},\ \mathbf{V}_{w2},\ \mathbf{V}'_{w2}]$.

Next, we apply W-MSA (Liu et al. 2021) using two different window sizes to $\left[\mathbf{Q}_{w1}, \mathbf{K}_{w1}, \mathbf{V}_{w1}, \mathbf{Q}'_{w1}, \mathbf{K}'_{w1}, \mathbf{V}'_{w1}\right]$ and $\left[\mathbf{Q}_{w2}, \mathbf{K}_{w2}, \mathbf{V}_{w2}, \mathbf{Q}'_{w2}, \mathbf{K}'_{w2}, \mathbf{V}'_{w2}\right]$, respectively.

Multi-Head Self-Attention Calculation We will calculate a single window multi-head self-attention (MSA). As depicted in the top first branch of Figure 5, we employ several non-overlapping windows of size $S_1\times S_1$ to segment $\mathbf{Q}_{w1}, \mathbf{K}_{w1}, \mathbf{V}_{w1}$, which essentially reshapes them into $\mathbf{Q}_{w1}, \mathbf{K}_{w1}, \mathbf{V}_{w1} \in \mathbb{R}^{\frac{HW}{S_1^2}\times S_1^2\times\frac{C}{4}}$. We divide them into several heads along the channel dimension: $\mathbf{Q}_{w1} = \left[\mathbf{Q}_{w1}^{1}, ..., \mathbf{Q}_{w1}^{h}\right]$, $\mathbf{K}_{w1} = \left[\mathbf{K}_{w1}^{1}, \mathbf{K}_{w1}^{2}, ..., \mathbf{K}_{w1}^{h}\right]$ and $\mathbf{V}_{w1} = \left[\mathbf{V}_{w1}^{1}, \mathbf{V}_{w1}^{2}, ..., \mathbf{V}_{w1}^{h}\right]$. In addition, we incorporate positional information again by utilizing the absolute position embedding (APE) (Vaswani et al. 2017), adding a learnable parameter $\mathbf{P}_{w1}^{n} \in \mathbb{R}^{S_1^2\times S_1^2}$ for implementation. Let $d = \dfrac{C}{4h}$ denote the dimension of each head and h denote the number of heads, then the self-attention of the n^{th} head can be represented as:

$$\mathbf{A}_{w1}^{n} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{w1}^{n}{\mathbf{K}_{w1}^{n}}^{T}}{\sqrt{d}} + \mathbf{P}_{w1}^{n}\right)\mathbf{V}_{w1}^{n} \tag{8}$$

The output of the first branch can be represented as:

$$\mathbf{A}_{w1} = \mathrm{Concat}\left(\mathbf{A}_{w1}^{1}, \cdots, \mathbf{A}_{w1}^{h}\right) \tag{9}$$

Next, we will calculate the window MSA for the second branch. First, we partition $\mathbf{Q}'_{w1}, \mathbf{K}'_{w1}, \mathbf{V}'_{w1}$ into non-overlapping windows of size $S_1\times S_1$. Then, we apply the shuffle operation by reshaping them into $\mathbb{R}^{S_1^2\times\frac{HW}{S_1^2}\times\frac{C}{4}}$. The subsequent operations are the same as the first branch. After computing the attention ${\mathbf{A}_{w1}^{n}}'$, we implement an unshuffle operation, reshaped it into $\mathbb{R}^{\frac{HW}{S_1^2}\times S_1^2\times C}$. Then the output of the second branch can be represented as:

$$\mathbf{A}'_{w1} = \mathrm{Concat}\left({\mathbf{A}_{w1}^{1}}', \cdots, {\mathbf{A}_{w1}^{h}}'\right) \tag{10}$$

Subsequently, we perform the same steps to calculate the self-attention for the other two branches with a window size of $S_2\times S_2$, resulting in $\mathbf{A}_{w2}$ and $\mathbf{A}'_{w2}$. Then the outputs of the four branches are concatenated together and a linear layer is used to obtain the final output of DWM-MSA $\mathbf{X}' \in \mathbb{R}^{H\times W\times N}$:

$$\mathbf{X}' = \mathrm{Linear}(\mathrm{Concat}[\mathbf{A}_{w1}, \mathbf{A}'_{w1}, \mathbf{A}_{w2}, \mathbf{A}'_{w2}]) \tag{11}$$

Finally, $\mathbf{X}'$ is passed through a feed-forward network. The self-attention computation for each window focuses on the features at different scales, and the fusion of these features yields the ultimate self-attention result.

Con-Abs Position Embedding Transformers typically adopt a single type of position embedding method, such as absolute position embedding (APE) (Vaswani et al. 2017), relative position embedding (RPE) (Shaw, Uszkoreit, and Vaswani 2018), or conditional position embedding (CPE) (Chu et al. 2023). However, in DWMAB, we introduce a novel position embedding method called con-abs position embedding (CAPE) that combines two types of position embedding. Specifically, CAPE embeds the input with the conditional position information at the beginning of the block, and utilizes the absolute position embedding after the multiplication operation between $\mathbf{Q}$ and $\mathbf{K}$.

Experiments

Dataset Description

Similar to (Meng, Ma, and Yuan 2020; Hu et al. 2022; Cai et al. 2022a,b), we conduct the simulation and real experiments with 28 spectral channels from 450 nm to 650 nm.

Simulation Data The simulation experiments are conducted on two datasets, *i.e.*, CAVE (Park et al. 2007) and KAIST (Choi et al. 2017). CAVE consists of 32 HSIs with a spatial size of 512×512, while KAIST comprises 30 HSIs with a spatial size of 2704×3376. According to the previous works (Meng, Ma, and Yuan 2020; Hu et al. 2022; Cai et al. 2022a,b), we use CAVE as the training set and select ten scenes from KAIST for testing.

Real Data We use the real HSIs collected by the CASSI system (Meng, Ma, and Yuan 2020) to validate the real application performance of the proposed network.

Implementation Details

When conducting experiments on simulation data and real data, we cropped patches from the 3D HSI with a spatial size of 256×256 and 660×660, respectively. Data is modulated by the mask and then sheared to simulate the CASSI

Scene	GAP-TV	DeSCI	TSA-Net	HDNet	MST	CST	**DWMT**
1	26.82, 0.754	27.13, 0.748	32.03, 0.892	35.14, 0.935	35.40, 0.941	35.96, 0.949	**36.46, 0.957**
2	22.89, 0.610	23.04, 0.620	31.00, 0.858	35.67, 0.940	35.87, 0.944	36.84, 0.955	**37.75, 0.963**
3	26.31, 0.802	26.62, 0.818	32.25, 0.915	36.03, 0.943	36.51, 0.953	38.16, 0.962	**38.47, 0.965**
4	30.65, 0.852	34.96, 0.897	39.19, 0.953	42.30, 0.969	42.27, 0.973	42.44, 0.975	**44.23, 0.984**
5	23.64, 0.703	23.94, 0.706	29.39, 0.884	32.69, 0.946	32.77, 0.947	33.25, 0.955	**33.99, 0.963**
6	21.85, 0.663	22.38, 0.683	31.44, 0.908	34.46, 0.952	34.80, 0.955	35.72, 0.963	**36.17, 0.970**
7	23.76, 0.688	24.45, 0.743	30.32, 0.878	33.67, 0.926	33.66, 0.925	34.86, 0.944	**35.22, 0.949**
8	21.98, 0.655	22.03, 0.673	29.35, 0.888	32.48, 0.941	32.67, 0.948	34.34, 0.961	**34.56, 0.968**
9	22.63, 0.682	24.56, 0.732	30.01, 0.890	34.89, 0.942	35.39, 0.949	36.51, 0.957	**37.41, 0.965**
10	23.10, 0.584	23.59, 0.587	29.59, 0.874	32.38, 0.937	32.50, 0.941	33.09, 0.945	**34.00, 0.959**
Average	24.36, 0.669	25.27, 0.721	31.46, 0.894	34.97, 0.943	35.18, 0.948	36.12, 0.957	**36.82, 0.964**
Params	-	-	44.25M	2.37M	**2.03M**	3.00M	14.48M
FLOPs (G)	-	-	110.03	154.76	**28.15**	40.10	46.71

Table 1: PSNR (dB) (left in each cell) and SSIM (right in each cell) values for different methods on ten scenes in the KAIST dataset. The bold font is the best performance.

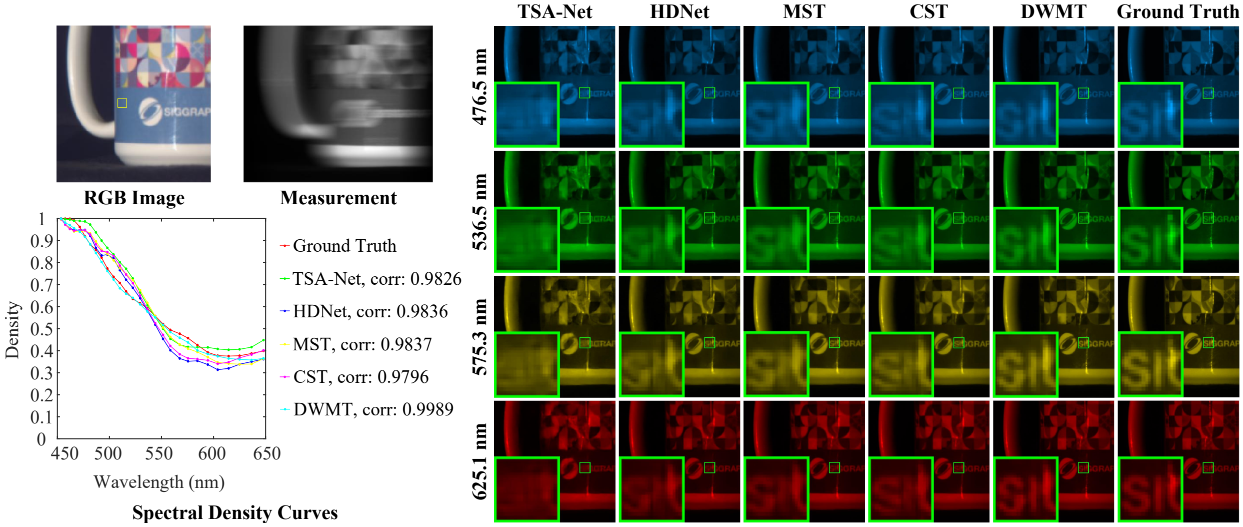

Figure 6: Reconstructed results with four (out of 28) spectral channels of *Scene* 5 on the simulation data.

system. The dispersion shifting step α is set to 2. Thus, the simulation and real data have 2D measurement sizes of 256×310 and 660×714, respectively. Data augmentation involves random rotation and flipping. In DWM-MSA, we set the sizes of the two windows as 8×8 and 16×16, respectively. The DWMT is implemented in PyTorch and trained and tested on a single RTX A6000 GPU. The model employs the Adam (Kingma and Ba 2015) optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with 500 epochs and an initial learning rate of 4×10^{-4}. The learning rate is adjusted by Cosine Annealing scheme. The batch size is set to 5, and the training objective is to minimize the Root Mean Square Error (RMSE) between the reconstructed image and the corresponding ground truth.

Simulation HSI Reconstruction

We compare DWMT with other methods in terms of PSNR, SSIM, Params, and FLOPs, including two hand-crafted prior-based methods (GAP-TV (Yuan 2016) and DeSCI (Liu et al. 2018)), two CNN-based methods (TSA-Net (Meng, Ma, and Yuan 2020), HDNet (Hu et al. 2022)), and two Transformer-based methods (MST (Cai et al. 2022b), CST (Cai et al. 2022a)). The test results on ten simulation scenes in KAIST are shown in Table 1. It can be observed that our DWMT significantly outperforms the comparative methods in terms of reconstruction quality across all ten scenes, demonstrating the effectiveness of our method. Particularly, compared with the recent Transformer-based methods (MST (Cai et al. 2022b), CST (Cai et al. 2022a)), our method achieves an average PSNR improvement of 1.64 dB and 0.70 dB, as well as SSIM improvement of 0.016 and 0.007.

As shown in Figure 6, we present the reconstruction results of our DWMT compared with four other end-to-end methods on four (out of 28) spectral channels of a simulated HSI in *Scene* 5. The bottom left portion displays the enlarged local details. It can be observed that the previous methods fail to clearly depict the letters on the cup or some letters suffer from distortions, and some reconstructed images even

Method	1	2	3	4	5
enhancement	-	✓	✓	✓	✓
CAFM	-	-	-	✓	✓
SAH-MSA	-	-	-	✓	-
DWM-MSA	-	-	✓	-	✓
PSNR	32.23	32.34	36.55	36.46	**36.82**
SSIM	0.908	0.908	0.963	0.963	**0.964**

Table 2: Ablation study on the simulated HSI datasets.

Window Size	8×8	16×16	dual-window
PSNR	36.66	36.40	**36.82**
SSIM	0.963	0.962	**0.964**
Params	16.79M	**12.18M**	14.48M
FLOPs (G)	47.94	**45.48**	46.71

Table 3: Comparison of different window sizes.

Method	APE	CPE	**CAPE**
PSNR	36.51	36.31	**36.82**
SSIM	0.963	0.961	**0.964**
Params	14.47M	**4.63M**	14.48M
FLOPs (G)	46.54	**46.71**	**46.71**

Table 4: Comparison of position embedding methods. APE is absolute position embedding, CPE is conditional position embedding, and CAPE is con-abs position embedding.

appear speckles. In contrast, our DWMT produces clearer images to identify the letter shape on the cup that is closer to the ground truth, and it is more effective to recover the fine-grained structural content and detailed information. We have also plotted the spectral density curves corresponding to the selected regions outlined in the yellow box on the RGB image. The correlation between our curves and the ground truth values is higher than the other methods, demonstrating the effectiveness of our DWMT in spectral consistency.

Real HSI Reconstruction

Following the previous works (Meng, Ma, and Yuan 2020; Cai et al. 2022a,b), we retrain our model on all scenes of the CAVE and KAIST datasets, and then conduct the restoration experiments on the real HSIs captured by the CASSI system. During training, we add 11-bit shot noise into the measurements to simulate the real imaging conditions. As shown in Figure 1, our DWMT outperforms other methods in noise suppression and detail reconstruction, demonstrating its generalizability and robustness.

Ablation Study

We conduct ablation experiments on the simulated HSI datasets (Park et al. 2007; Choi et al. 2017).

Break-Down Ablation As shown in Table 2, the absence of any module leads to a degradation in the reconstruction performance, highlighting the design rationality of our DWMT. Among them, the enhancement refers to the second branch of the encoder in fine pixel refinement stage and CAFM refers to replacing it with channel-wise concatenation and a 1×1 Conv layer. Particularly, when the DWM-MSA module is removed, the PSNR and SSIM decrease by 4.07 dB and 0.054, respectively, demonstrating the crucial role of our DWM-MSA in the HSI reconstruction.

Self-Attention Mechanism Analysis We replace our DWM-MSA with SAH-MSA (Cai et al. 2022a), another advanced attention mechanism used for HSI reconstruction. The results, as shown in Table 2, indicate that our method achieves a PSNR improvement of 0.36 dB compared with SAH-MSA. Notably, SAH-MSA has been proven to outperform some classical attention mechanisms such as G-MSA (Dosovitskiy et al. 2021) and Swin-MSA (Liu et al. 2021) in HSI reconstruction (Cai et al. 2022a). This also demonstrates the effectiveness of our DWM-MSA. Moreover, in the CST (Cai et al. 2022a) evaluation, the PSNR is 36.12 dB when SAH-MSA is employed in our framework, it increases to 36.46 dB, further validating the effectiveness of our proposed overall framework.

Window Size Analysis We utilize a single-scale window for self-attention calculation to compare with our method (double-scale window). As shown in Table 3, when using both 8×8 and 16×16 windows simultaneously, the reconstruction performance is superior to using only an 8×8 or 16×16 window alone. This highlights that our method can learn more valuable information from the multi-scale features and achieves superior reconstruction performance.

Position Embedding Analysis We compare our proposed CAPE with two other methods, APE (Vaswani et al. 2017) and CPE (Chu et al. 2023). As shown in Table 4, for CAPE (our method), the PSNR is improved by 0.31 dB and 0.51 dB compared with APE and CPE, respectively. This verifies that our CAPE incorporates the advantages of both them, effectively enhancing positional information to improve the reconstruction performance.

Conclusion

In this paper, we address the limitations of existing Transformers for HSI reconstruction in capturing long-range dependencies and local representation with reasonable computational cost. To overcome these issues, we propose an effective Transformer-based algorithm (DWMT) for HSI reconstruction. In our DWMT, we propose a novel U-Net architecture with a dual-branch encoder to enhance the extraction of fine-grained image features. The most crucial part of DWMT is the self-attention calculation, *i.e.*, DWM-MSA, allowing to capture the long-range dependencies in a local region at different scales. Additionally, we introduce a novel position embedding method in Transformer, CAPE, which combines the advantages of absolute position embedding and conditional position embedding, enhancing the positional information of image. Through comprehensive quantitative and qualitative evaluations against the state-of-the-art algorithms, we demonstrate the effectiveness and generalizability of our proposed method. Furthermore, the ablation experiments indicate that each module in our method significantly improves the reconstruction performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62071340, 62371076 and 62201109, and the Natural Science Foundation of Chongqing under Grant CSTB2022NSCQ-MSX0452.

References

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10231–10241.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022a. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *European Conference on Computer Vision (ECCV)*, 686–704.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022b. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17502–17511.

Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Ding, H.; Zhang, Y.; Timofte, R.; and Gool, L. V. 2022c. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 37749–37761.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision (ECCV)*, 205–218.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4): 834–848.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

Choi, I.; Kim, M.; Gutierrez, D.; Jeon, D.; and Nam, G. 2017. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 36(6): 218:1–218:13.

Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; and Shen, C. 2023. Conditinal positional encodings for vision Transformers: Hierarchical vision transformer using shifted windows. In *International Conference on Learning Representations (ICLR)*, 1–19.

Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; and Zhang, L. 2021a. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7373–7382.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021b. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1601–1610.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 1–22.

Gao, P.; Zheng, M.; Wang, X.; Dai, J.; and Li, H. 2021. Fast convergence of detr with spatially modulated co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3621–3630.

Gehm, M. E.; John, R.; Brady, D. J.; Willett, R. M.; and Schulz, T. J. 2007. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics Express*, 15(21): 14013–14027.

Guo, T.; Wang, R.; Luo, F.; Gong, X.; Zhang, L.; and Gao, X. 2023. Dual-View Spectral and Global Spatial Feature Fusion Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61: 5512913.

Hu, X.; Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17542–17551.

Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; and Wu, J. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1055–1059.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 1–15.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16478–16488.

Liu, Y.; Yuan, X.; Suo, J.; Brady, D. J.; and Dai, Q. 2018. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(12): 2990–3006.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Luo, F.; Zhou, T.; Liu, J.; Guo, T.; Gong, X.; and Ren, J. 2023. Multiscale diff-changed feature fusion network for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61: 5502713.

Meng, Z.; Ma, J.; and Yuan, X. 2020. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European conference on computer vision (ECCV)*, 187–204.

Meng, Z.; Qiao, M.; Ma, J.; Yu, Z.; Xu, K.; and Yuan, X. 2020. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14): 3897–3900.

Miao, X.; Yuan, X.; Pu, Y.; and Athitsos, V. 2019. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4059–4069.

Park, J.-I.; Lee, M.-H.; Grossberg, M. D.; and Nayar, S. K. 2007. Multispectral imaging using multiplexed illumination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–8.

Ren, S.; Zhou, D.; He, S.; Feng, J.; and Wang, X. 2022. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10853–10862.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 234–241.

Schechner, Y. Y.; and Nayar, S. K. 2002. Generalized mosaicing: Wide field of view multispectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(10): 1334–1348.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*, 464–468.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision (ECCV)*, 459–479.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.

Wagadarikar, A.; John, R.; Willett, R.; and Brady, D. 2008. Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics*, 47(10): B44–B51.

Wang, W.; Yao, L.; Chen, L.; Lin, B.; Cai, D.; He, X.; and Liu, W. 2022. CrossFormer: A versatile vision transformer hinging on cross-scale attention. In *International Conference on Learning Representations (ICLR)*, 1–15.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4794–4803.

Yuan, X. 2016. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, 2539–2543.

Yuan, X.; Brady, D. J.; and Katsaggelos, A. K. 2021. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2): 65–88.

Zhang, S.; Wang, L.; Fu, Y.; Zhong, X.; and Huang, H. 2019. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10183–10192.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6848–6856.

Zhou, H.; Luo, F.; Zhuang, H.; Weng, Z.; Gong, X.; and Lin, Z. 2023. Attention Multi-hop Graph and Multi-scale Convolutional Fusion Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61: 5508614.

Zhu, F.; Zhu, Y.; Zhang, L.; Wu, C.; Fu, Y.; and Li, M. 2021. A unified efficient pyramid transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2667–2677.

Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; and Lau, R. W. 2023. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10323–10333.