复现论文：Spectral Enhanced Rectangle Transformer for Hyperspectral Image Denoising

摘要

去噪是高光谱图像（HSI）应用的一个关键步骤。尽管深度学习在 HSI 去噪上的表现优越，但现有的 HSI 去噪方法在捕获非局部自相似性方面存在局限性。Transformer 在捕获长程依赖关系方面显示出了不俗的潜力，但很少有人尝试使用专门设计的 Transformer 来模拟 HSI 中的空间和频谱相关性。论文通过提出一种光谱增强矩形 Transformer 来解决这些问题，驱动它探索 HSI 的非局部空间相似性和全局光谱低秩特性。对于前者，论文利用水平和垂直的矩形自注意力来捕捉空间域中的非局部相似性。对于后者，论文设计了一个频谱增强模块，该模块能够提取空间光谱立方体的全局底层低秩属性以抑制噪声，同时实现不重叠的空间矩形之间的交互。在模拟噪声 HSI 和真实噪声 HSI 上进行了广泛的实验，结果显示了论文提出的方法在客观度量和主观视觉质量方面的有效性。

关键词：去噪；高光谱图像 (HSI)；Transformer；自注意力

1 引言

高光谱图像（HSI）拥有丰富的光谱信息，能够提供比 RGB 图像更详细的特征，从而区分不同的物质。因此，HSI 已被广泛应用于环境监测、资源勘探、医学诊断 [10] 等领域。但由于传感器噪声和传输过程中的干扰，原始的 HSI 数据常常伴随着噪声，除了较差的视觉效果，这种不期望的退化还对下游应用产生了负面影响。为了在 HSI 视觉任务中获得更好的视觉效果和性能，去噪是 HSI 分析和处理的一个基本但关键的步骤。

与 RGB 图像类似，HSI 在空间域具有自相似性，表明相似的像素可以被分组并一起去噪。此外，由于高光谱成像系统能够在标称光谱分辨率下获取图像，HSI 在光谱域具有内在的相关性，因此，在设计 HSI 去噪方法时，考虑空间和光谱域是很重要的。传统的基于模型的 HSI 去噪方法 [4] 通过迭代求解优化问题，利用手工设计的先验来探索空间和光谱相关性。在这些工作中，总变差先验、非局部相似性、低秩特性和稀疏性正则化经常被利用，但这些方法的性能依赖于手工设计先验的准确性。在实际的 HSI 去噪中，基于模型的方法通常耗时且在不同场景下的泛化能力有限。

为了获得鲁棒的去噪学习，深度学习方法被应用于 HSI 去噪，并取得了令人印象深刻的表现。然而，这些方法 [1] 大多数利用卷积神经网络进行特征提取，并依赖于局部滤波器响应在有限的感受野内分离噪声和信号。

近年来，视觉 Transformer 在高级任务和低级任务中都取得了具有竞争力的结果，显示出其在图像区域建模长距离依赖关系的强大能力。为了降低图像大小的二次计算成本，许多工作研究了空间注意力的高效设计。Swin Transformer [9] 将特征图分割成可移位的正方形窗口。CSWin Transformer [6] 提出了跨特征图的条带窗口以扩大注意力区域。由于 HSI 通常具有较大的特征图，探索噪声像素之外的相似性可能会造成不必要的计算负担，因此，如何高效地建模非局部空间相似性仍然是 HSI 去噪 Transformer 面临的挑战。

HSI 通常位于光谱低秩子空间 [3]，可以保持区分信息并抑制噪声。这表明非局部空间相似性和低秩光谱统计量应联合用于 HSI 去噪。然而，现有的 HSI 去噪方法主要通过矩阵分解利用低秩特性，这基于单个 HSI 且需要较长时间求解。大型数据集中的全局低秩特性很少被考虑。

论文提出了一种光谱增强矩形 Transformer（SERT）用于 HSI 去噪。为了以合理的成本增强模型能力，论文开发了一种多形状矩形自注意力模块（RA），全面探索非局部空间相似性。此外，论文在光谱增强模块（SE）中聚合最有信息量的光谱统计量以抑制噪声，该模块在全局光谱记忆单元的辅助下将空间-光谱立方体投影到低秩向量中。光谱增强模块还提供了非重叠空间矩形之间的交互。通过论文提出的 Transformer，空间非局部相似性和全局光谱低秩特性被联合考虑以改善去噪过程。实验结果表明，论文的方法在模拟噪声和真实噪声 HSI 中都有优异表现。

2 相关工作

2.1 高光谱图像去噪

高光谱图像（HSI）去噪是计算机视觉和遥感中一个发展良好的研究领域。主流的 HSI 去噪方法可以分为基于模型的方法和深度学习方法。

传统的基于模型的方法将去噪视为一个带有手工设计先验的迭代优化问题。[7] 提出了自适应空间-光谱字典方法。[8] 将空间非局部相似性和全局光谱低秩特性集成用于去噪。此外，其他传统的空间正则化器和低秩正则化也被引入来建模噪声 HSI 的空间和光谱特性。

深度学习方法在 HSI 去噪中具有自动学习和表示特征的巨大潜力。[2] 提出了一种深度空间-光谱全局推理网络，以考虑 HSI 去噪的局部和全局信息。与那些具有有限感受野和固定特征提取范式的基于卷积的网络不同，论文提出的方法利用 Transformer 更好地建模空间和光谱域中的内在相似性。

2.2 视觉 Transformer

用于 RGB 图像的 Transformer。由于 Transformer 在建模长距离依赖关系方面的强大能力，其已被积极应用于视觉任务。自注意力机制在前人的工作中已被证明是有效的。当应用于空间区域时，Transformer 需要考虑计算成本和模型容量之间的权衡。为了减少图像大小的二次计算增长，Dosovitskiy 等人首次将 Transformer 应用于图像识别，将图像分割成小块。Swin Transformer 提出了带有移位窗口的空间域自注意力。为了进一步扩大自注意力的感受野，[5] 引入了下采样注意力。Dong 等人在不丢失空间信息的情况下，采用水平和垂直条带计算自注意力。然而，对于 HSI 去噪，这些 Transformer 在有限窗口中进行空间自注意力或引入不必

要的计算成本，未能有效探索非局部空间相似性。此外，很少有研究同时考虑空间和光谱域。

用于 HSI 的 Transformer。近几年，使用 Transformer 进行 HSI 修复和分类的趋势正在兴起，但这些工作没有考虑空间和光谱域中的相似性。论文引入了光谱增强矩形 Transformer 用于 HSI 去噪，探索 HSI 最重要的两个特性，包括空间非局部相似性和全局低秩特性。

3 本文方法

3.1 本文方法概述

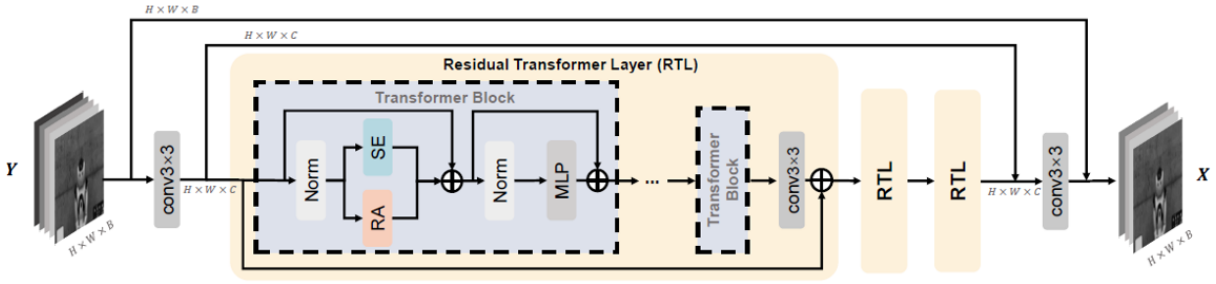

图 1. SERT 总体框架

假设降级的噪声为 $Y \in \mathbb{R}^{H\times W\times B}$，其中 H、W 和 B 分别表示 HSI 的高度、宽度和波段，则噪声衰减可以如公式1表示：

$$Y = X + n \tag{1}$$

其中，X $n \in \mathbb{R}^{H\times W\times B}$，$X$ 表示所需的干净 HSI，n 表示添加的随机噪声。在实际的 HSI 降级情况下，HSI 会被各种类型的噪声破坏，例如高斯噪声、条带噪声、截止时间噪声、脉冲噪声或以上噪声的混合。

论文提出的光谱增强矩形 Transformer（Spectral Enhanced Rectangle Transformer，以下简称 SERT）网络总体框架如图 1所示，在具体实现中，每个 Residual Transformer Layer（RTL）由 6 个 Transformer Block 组成。而 Transformer Block 主要包含两个基本组件，即矩形自注意力（RA）模块和光谱增强（SE）模块，而将 RA 和 SE 的输出相加，可以实现全面特征嵌入并用于噪声去除。

3.2 矩形自注意力模块

论文提出的 RA 模块的详细信息如图 2所示。通过将特征图分割成若干个不重叠的矩形，RA 关注于信息丰富的邻近像素，并在非局部区域获得更全面的信息。为了获得全面的特征，在光谱分割操作后，RA 分别在垂直和水平方向上进行，其中 W-RMSA 表示水平矩形多头自注意力，H-RMSA 表示垂直矩形多头自注意力。论文还增加了一个光谱洗牌操作来交换两个分支的信息，由于垂直和水平矩形自注意力关注不同的区域且具有不同的感受野，洗牌操作也扩大了整个模块的感受野。

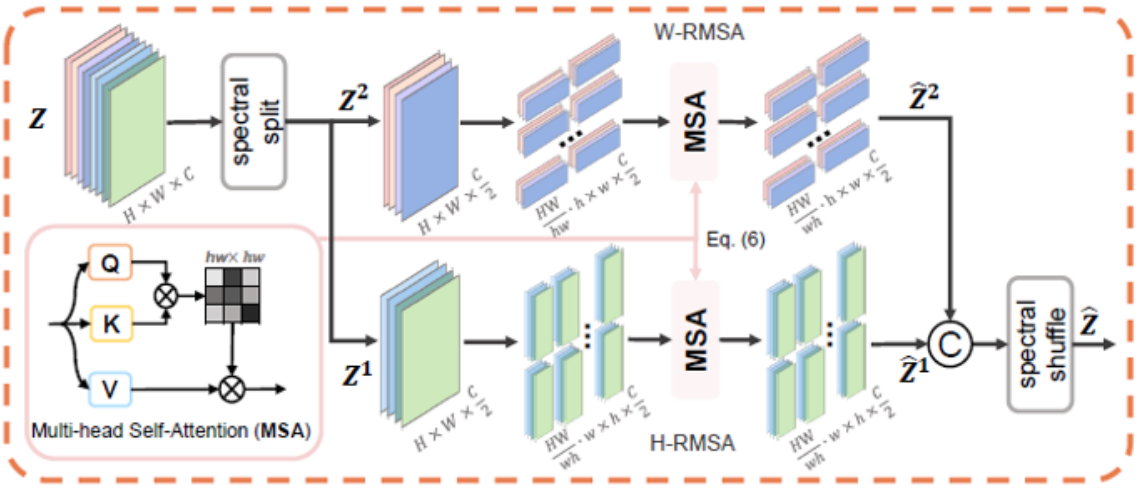

图 2. 矩形自注意力（RA）模块

3.3 光谱增强模块

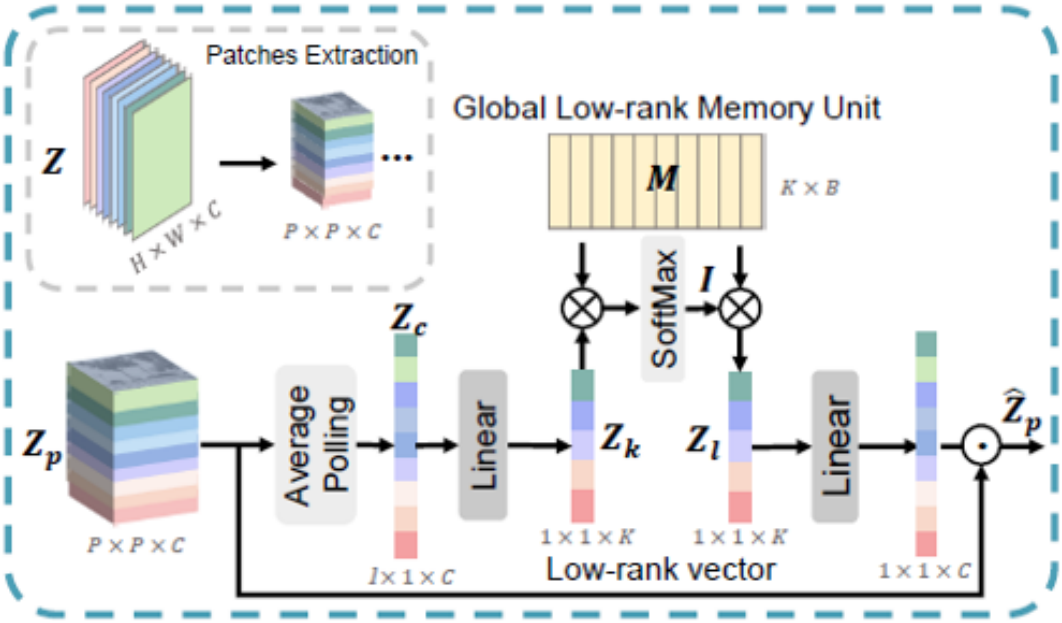

图 3. 光谱增强（SE）模块

论文提出的 SE 模块的详细信息如图 3所示。特征首先被划分为多个大小为 $P\times P\times C$ 的立方体块 Z_p 以探索光谱-空间相关性，在实现中，P 被设置为 RA 模块中矩形的长边。为了在子空间中获得区分性的光谱信息，采用挤压操作并聚合立方体块 Z_p 中的特征以产生大小为 $1\times1\times K$ 的投影光谱向量 Z_k。具体来说，首先在空间域进行下采样操作以获得聚合的光谱向量 Z_c，然后将其投影以获得 Z_k，Z_k 位于秩为 K 的子空间中。为了探索当前 HSI 立方体之外的光谱-空间相关性并增强低秩光谱向量的表达能力，论文引入了一个记忆单元（MU）来存储光谱信息，MU 模块维护一个全局记忆库 $M\in\mathbb{R}^{K\times B}$，作为网络的参数进行学习。对于光谱向量 Z_k，论文在 MU 中寻找最相关的光谱低秩向量 $I\in\mathbb{R}^{1\times B}$，并使用 I 来协助调整 Z_k。论文不是对整个图像进行全局聚合，而是关注立方体内的信息，因为邻近像素倾向于共享相似的光谱统计信息。

3.4 均方误差损失函数

论文采用了均方误差（Mean Square Error，MSE）作为损失函数来计算预测值与真实值之间的差距，以便评估学习效果。MSE 计算公式2如下：

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2}$$

其中，x_i 和 y_i 分别表示对应的预测值和真实值。

4 复现细节

4.1 与已有开源代码对比

原论文的源代码公开在Github上。本复现工作是通过对论文源代码进行修改、调试而展开的，修改了源代码并添加了一些注释方便理解，特别是其中生成模拟噪声训练数据的部分代码。但由于原有 ICVL 数据集链接失效，在下花费了许多时间在各处寻找有效的ICVL 数据集。

4.2 实验环境搭建

软件环境：python>=3.8，torch>=2.0，torchaudio>=2.0，torchvision>=0.19，lmdb，tqdm，源代码必需的其它库，Ubuntu 20.04 操作系统。

硬件环境：16G 的 Tesla P100 显卡，Intel(R) Xeon(R) CPU。

5 实验结果分析

5.1 模拟噪声

为了模拟 HSI 噪声效果，复现工作利用公式1向 ICVL 数据集中干净的 HSI 分别添加 10-70 级的高斯噪声和复杂噪声（包括非 i.i.d 高斯噪声、高斯 + 条带噪声、高斯 + 截止时间噪声、高斯 + 脉冲噪声、混合噪声）来模拟噪声 HSI。

复现工作采用原论文提供的训练参数训练 10-70 级随机高斯噪声水平的网络，在不同噪声水平下的测试结果与原论文结果对比如表1所示；训练混合噪声的网络，在不同复杂噪声下的测试结果与原论文结果对比如表2所示。

表 1. ICVL 数据集上不同高斯噪声水平下的平均结果

	10		30		50		70		10-70	
SERT	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**
PSNR/dB	**47.72**	47.34	**43.56**	43.30	**41.33**	41.09	**39.82**	39.59	**42.82**	42.47
SSIM	**0.9988**	0.9986	**0.9969**	0.9967	**0.9949**	0.9946	**0.9929**	0.9924	**0.9957**	0.9956
SAM	**1.36**	1.42	**1.77**	1.85	**2.05**	2.16	**2.30**	2.41	**1.88**	1.98

表 2. ICVL 数据集上不同复杂噪声的平均结果

SERT	Non-i.i.d Gaussian		Gaussian+Deadline		Gaussian+Impulse		Gaussian+Stripe		Gaussian+Mixture	
	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**	Paper	**Mine**
PSNR/dB	**44.20**	43.87	**43.66**	43.42	**42.67**	42.74	**43.68**	43.60	**40.00**	41.03
SSIM	**0.9971**	0.9971	**0.9969**	0.9968	**0.9959**	0.9962	**0.9969**	0.9969	**0.9937**	0.9948
SAM	**1.69**	1.94	**1.99**	2.03	**2.30**	2.26	**1.97**	2.03	**2.84**	2.68

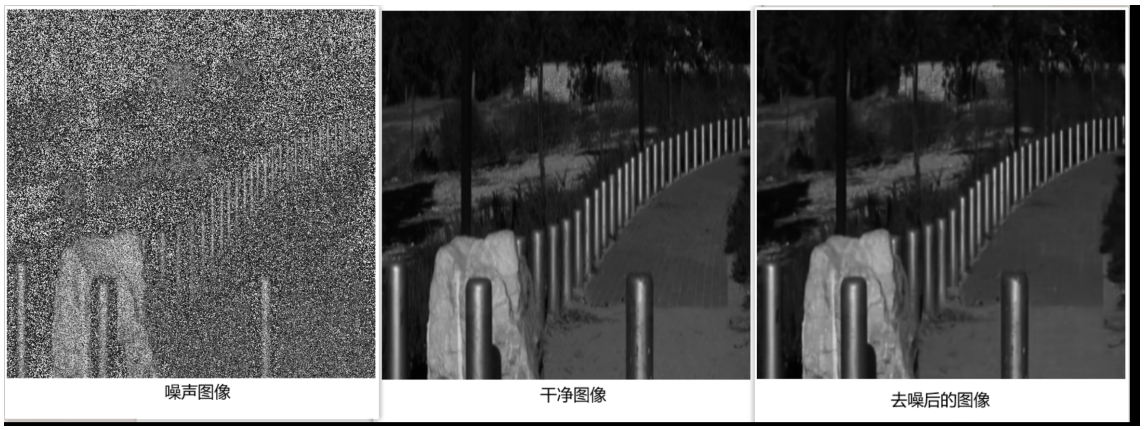

图 4. 噪声级别为 50 的高斯噪声下，波段为 28 的 ICVL 的视觉效果

图 5. 混合复杂噪声下，波段为 28 的 ICVL 的视觉效果

表1和表2展示了复现结果与论文结果的对比，两表中 Paper 列表示论文结果，Mine 列表示复现结果，从两者结果的对比可以看出复现结果与论文结果存在差异但差异不大，这可能是因为对数据集的预处理、添加的模拟噪声或硬件环境不完全相同导致的。总的来说，复现结果基本和论文结果相同。

图4展示了波段为 28 的 ICVL 数据集在噪声级别为 50 的高斯噪声下的去噪效果与对应干净图像的对比效果，从中可以看出，复现训练出的模型很好地去除了噪声，基本与干净图像无异，只有仔细查看下才能发现地砖部分的纹理信息被平滑处理掉了。图5展示了波段为 28 的 ICVL 数据集在混合的复杂噪声下的去噪效果与对应干净图像的对比效果，从中可以看出，模型在去除复杂噪声上的表现也十分不错，但在一些细节上处理的不是很好。

5.2 真实噪声

复现工作同论文一样采用 Realistic 数据集 [11] 进行真实噪声的实验，选取论文源代码提供的 Realistic 数据集中的 44 个场景作为训练集，余下的作为测试集。实验结果与论文结果对比如表3所示。

表 3. Realistic 数据集上的平均结果

	PSNR/dB	SSIM	SAM
Paper	29.36	0.9355	2.536
Mine	27.21	0.9201	2.429

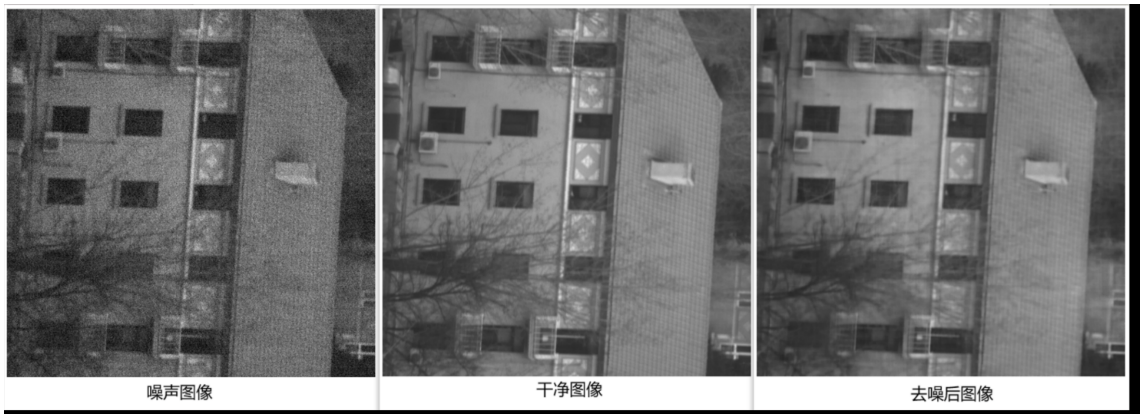

图 6. Realistic 数据集中场景 5 在波段 12 上的视觉效果

表3展示了复现结果与论文结果的对比，表中 Paper 行表示论文结果，Mine 行表示复现结果，从两者结果的对比可以看出复现结果与论文结果存在差异但差异不大。总的来说，复现结果基本和论文结果相同。

图6展示了波段为 12 的 Realistic 数据集中场景 5 的去噪效果与对应干净图像的对比效果，从中可以看出，复现训练出的模型很好地去除了噪声，保留了图像大部分的纹理细节，去噪后的图像视觉效果上比较接近干净图像，但局部颜色比干净图像稍暗。

6 总结与展望

本次复现工作让我完整体验了深度学习网络从设计构建到训练，再到测试的全部流程，也让我了解了目前高光谱图像去噪领域的研究进展和成果。复现的论文里提出了联合 HSI 空间非局部相似性和光谱低秩特性用于 HSI 去噪的方法，论文数据和复现结果都显示了该方法在去噪效果和泛化能力上的优异表现，但该方法由于网络参数较多，故而训练时间较长，对算力和硬件要求较高，不利于广泛应用。未来继续了解相关领域知识和成果，考虑应用更加优越的模块来精简优化提出方法的网络结构，在保证去噪性能的情况下缩短训练时间。

遗憾的是，由于每次训练时间长和硬件条件有限，本次复现工作未能完成对网络结构的修改优化，仅完成了论文工作成果的复现。但本次复现工作也让我学习收获了许多，相信在未来的工作和学习中这次经历也会不断提醒着我继续前进。

参考文献

[1] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng. Deep spatial-spectral global reasoning network for hyperspectral image denoising. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022.

[2] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng. Deep spatial-spectral global reasoning network for hyperspectral image denoising. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022.

[3] Yi Chang, Luxin Yan, and Sheng Zhong. Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising. Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition, pages 5901–5909, 2017.

[4] Guangyi Chen and Shen-En Qian. Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage. IEEE Transactions on Geoscience and Remote Sensing, 49(3):973–980, 2011.

[5] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. NeurIPS 2021, 2021.

[6] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition, pages 12124–12134, 2022.

[7] Ying Fu, Antony Lam, Imari Sato, and Yoichi Sato. Adaptive spatial-spectral dictionary learning for hyperspectral image denoising. Proc. Int. Conf. on Computer Vision, pages 343–351, 2015.

[8] Wei He, Quanming Yao, Chao Li, Naoto Yokoya, and Qibin Zhao. Non-local meets global: An integrated paradigm for hyperspectral denoising. Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition, pages 6861–6870, 2019.

[9] 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. Proc. Int. Conf. on Computer Vision, pages 9992–10002, 2021.

[10] Xueling Wei, Wei Li, Mengmeng Zhang, and Qingli Li. Medical hyperspectral image classification based on end-to-end fusion deep neural network. IEEE Transactions on Instrumentation and Measurement, 68(11):4481–4492, 2019.

[11] Tao Zhang, Ying Fu, and Cheng Li. Hyperspectral image denoising with realistic data. Proc. Int. Conf. on Computer Vision, pages 2248–2257, October 2021.