MST++: Multi-stage Spectral-wise Transformer for Efficient Spectral Reconstruction

Yuanhao Cai

摘要

现有的光谱重建（SR）领先方法专注于设计更深或更宽的卷积神经网络（CNN），以学习从 RGB 图像到其高光谱图像（HSI）的端到端映射。这些基于神经网络的方法实现了令人印象深刻的恢复性能，但在捕获长距离相关性和自相似性先验方面存在局限性。为了解决这个问题，我们提出了一种新的基于 transformer 的方法，即 Multi-stage Spectral-wise Transformer（MST++），用于有效的光谱重建。特别地，我们使用基于 HSI 空间稀疏而频谱自相似性质的 spectral-wise Multi head self-attention（S-MSA）来组成基本单元，spectral-wise attention block（SAB）。然后，SAB 建立了单阶段 Transformer（SST），该 transformer 利用 U 形结构来提取多分辨率上下文信息。

关键词：光谱重建；超分辨率

1 引言

高光谱图像（Hyperspectral Image, HSI）指的是光谱分辨率在 λ 数量级范围内的光谱图像。相较于常规的 RGB 图像而言，高光谱图像有着更多的波段（即通道数更多如 31，28）来更加准确全面的描述被捕获场景的特性。那么既然高光谱图像那么有用，我们应该如何获取它呢？传统的成像设备采用光谱仪对成像场景进行空间域通道维度的扫描，费时费力，不适用于运动场景。近些年，科学家们专门设计了快照压缩成像（Snapshot Compressive Imaging，SCI）系统来解决这一问题，但这种方法成本很高，为了降低成本，有一种方法是使用 RGB 来恢复高光谱图像，即光谱重建。传统的 SR 方法主要基于稀疏编码或相对较浅的学习模型。然而，这些基于模型的方法具有有限的表示能力和较差的泛化能力。最近，随着深度学习的发展，SR 取得了重大进展。深度卷积神经网络（CNN）已被应用于学习从 RGB 图像到 HSI 立方体的端到端映射函数。尽管已经取得了令人印象深刻的性能，但这些基于 CNN 的方法在捕获长距离相关性和光谱间自相似性方面存在局限性。近年来，自然语言处理（NLP）模型 Transformer[70] 已应用于计算机视觉，并取得了巨大成功。Transformer 中的多头自关注（MSA）机制在建模长距离依赖性和非局部自相似性方面比 CNN 做得更好，这可以缓解基于 CNN 的 SR 算法的局限性。我们提出了第一个基于 Transformer 的框架，即 Multi-stage Spectral-wise Transformer，用于从 RGB 图像进行有效的光谱重建。该工作是为光谱压缩成像恢复而定制的。首先，我们注意到 HSI 信号在空间上稀疏，而在频谱上自相似。基于这一性质，我们采用 spectral-wise Multi head self-attention（S-MSA）来构成基本单元，光谱式注意块（SAB）。S-MSA 将光谱特征图视为沿光谱维度计算自注意力特征图。第二，我们的 SABS 构建了我们提出的 singel state transformer（SST），该 transformer 利用 U 形结构来提取对 HSI 恢复至关重要的多分辨率谱上下文信息。最后，我们的 MST++ 在几个 SST 的支持下，开发了一个多阶段学习方案，以逐步提高从粗到细的重建质量，这显著提高了性能。

2 相关工作

2.1 传统高光谱图像的成像方式

用于收集 HSI 的传统成像系统通常采用光谱仪来沿着空间或光谱维度扫描场景。通常使用三种主要类型的扫描仪，包括扫帚扫描仪、推扫室扫描仪和带序列扫描仪来捕获人机接口。几十年来，这些传感器已广泛应用于探测、遥感、医学成像和环境监测。例如，推室扫描仪和扫帚扫描仪已用于摄影测量和遥感的卫星传感器[1]。然而，扫描过程通常需要很长时间，这使得它不适合测量动态场景。此外，成像设备的物理尺寸通常太大，无法插入便携式平台。为了解决这些限制，研究人员开发了 SCI 系统[2]来捕获 HSI，其中 3D HSI 立方体被压缩为单个 2D 测量[3]。在这些 SCI 系统中，编码孔径快照光谱成像（CASSI）[4]脱颖而出，并形成了一个有前途的研究方向。尽管如此，迄今为止，SCI 系统对于消费级应用而言仍然非常昂贵。即使是“低成本”的 SCI 系统也往往在 10 万至 10 万美元之间。因此，本课题具有重要的研究价值和实用价值。

2.2 从 RGB 重建高光谱图像

传统的 SR 方法[5]主要基于手工制作的高光谱先验。例如，Paramar 等人[6]提出了一种用于 HSI 重建的数据稀疏性扩展方法。Arad 等人[7]提出了创建 HSI 信号及其 RGB 投影字典的天冬氨酸编码方法。Aeschbacher 等人[5]建议在实现光谱超分辨率之前，使用来自特定光谱的相对较浅的学习模型。然而，这些基于模型的方法具有有限的表示能力和较差的泛化能力。最近，受深度学习在自然图像恢复中的巨大成功[8]的启发，已经利用神经网络来学习从 RGB 到 HSI 的底层映射函数[9]。例如，Xiong 等人[10]提出了一个统一的 HSCNN 框架，用于从 RGB 图像和压缩测量中重建 HSI。Shi 等人[11]使用自适应残差块为 SR 建立深度残差网络 HSCNN-R。Zhang 等人[12]自定义像素感知深度函数混合网络，该网络由 RGB 到 HSI 映射建模组成。然而，这些基于 CNN 的 SR 方法取得了令人印象深刻的结果，但在捕获非本地自相似性和长距离互依赖性方面存在局限性。

2.3 Vision Transformer

NLP 模型 Transformer[13]用于机器翻译。近年来，它已被引入计算机视觉，并因其在捕捉空间区域之间的长距离相关性方面的优势而受到广泛欢迎。在高级视觉中，Transformer 已广泛应用于图像分类[14]、对象检测[15]、语义分割[16]、人体姿态估计[17]等。此外，视觉 Transformer 也已用于低级视觉[18]。例如，Cai 等人[18]提出了第一个基于 Transformer 的端到端框架 MST，用于从压缩测量重建 HSI。Lin 等人[19]将 HSI 稀疏性嵌入 Transformer 中，以建立光谱压缩成像的粗到细学习方案。优先工作 Uformer[20]采用了 Swin Transformer[21]块构建的 U 形结构，用于自然图像恢复。然而，据我们所知，尚未探索 Transformer 在光谱超分辨率方面的潜力。这项工作旨在填补这一研究空白。

3 本文方法

3.1 本文方法概述

如图 1所示，（a）描述了所提出的 Multi-stage Spectral-wise Transformer，其由 Ns 个 singel state transformer（SST）级联。我们的 MST++ 将 RGB 图像作为输入，并重建 HSI 对应图像。利用长身份映射来简化训练过程。图 1（b）显示了由编码器、瓶颈和解码器组成的 U-shaped SST。嵌入和映射块是单个 conv3×3 层。编码器中的特征图依次进行下采样操作（跨步凸 4×4 层）、N1 个 SAB、下采样操

作和 N2 个 SAB。瓶颈由 N3 个 SAB 组成。解码器采用对称结构。上采样操作是跨步去卷积 2×2 层。为了避免下采样中的信息丢失，在编码器和解码器之间使用跳跃连接。

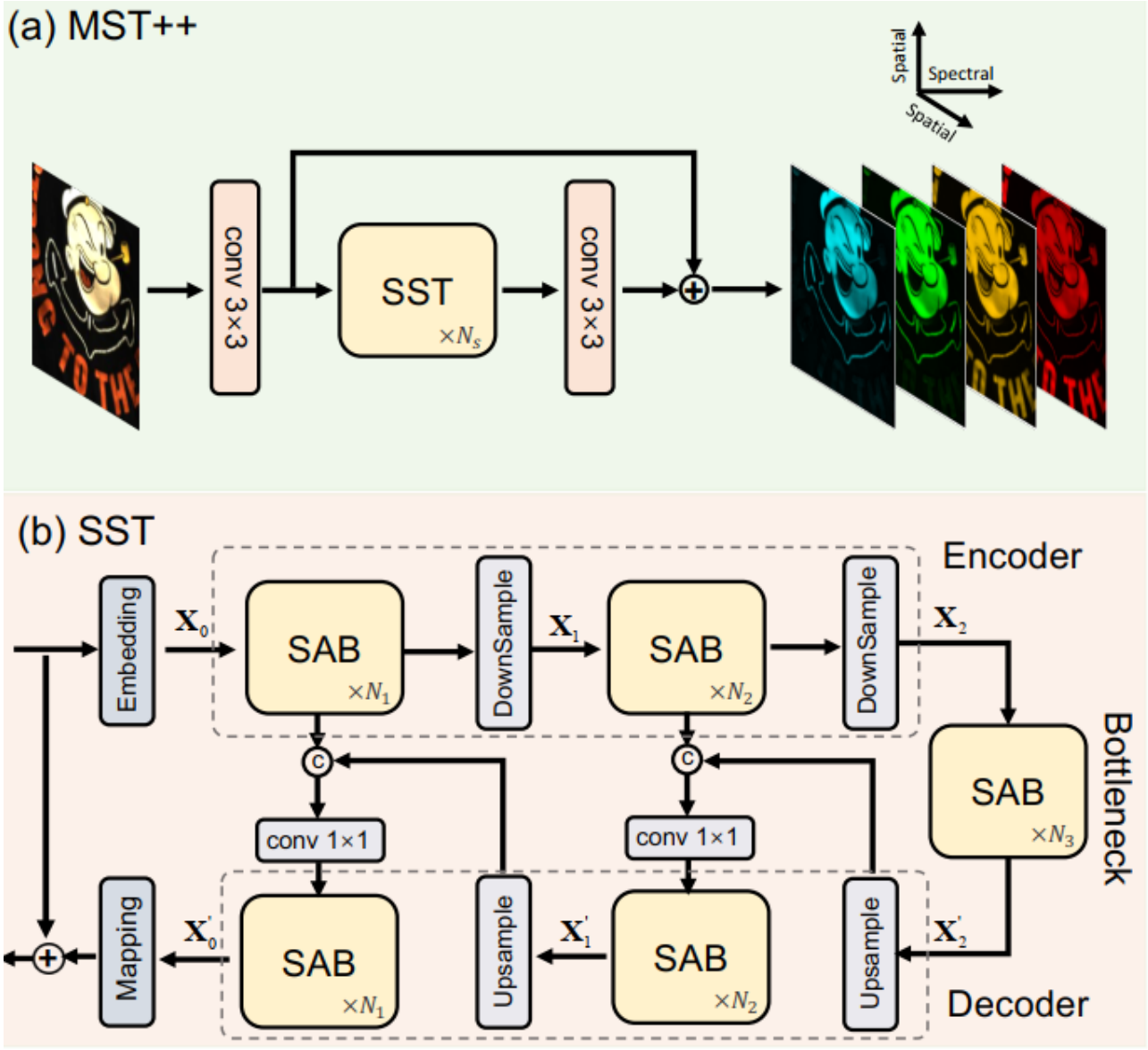

图 1: 整体结构图

3.2 特征提取模块

图 2（c）说明了 SAB 的组成部分，即前馈网络（如图 2（d）所示的 FFN）、频谱式多头自关注（S-MSA）和两层归一化。S-MSA 的详细信息如图 2（e）所示。

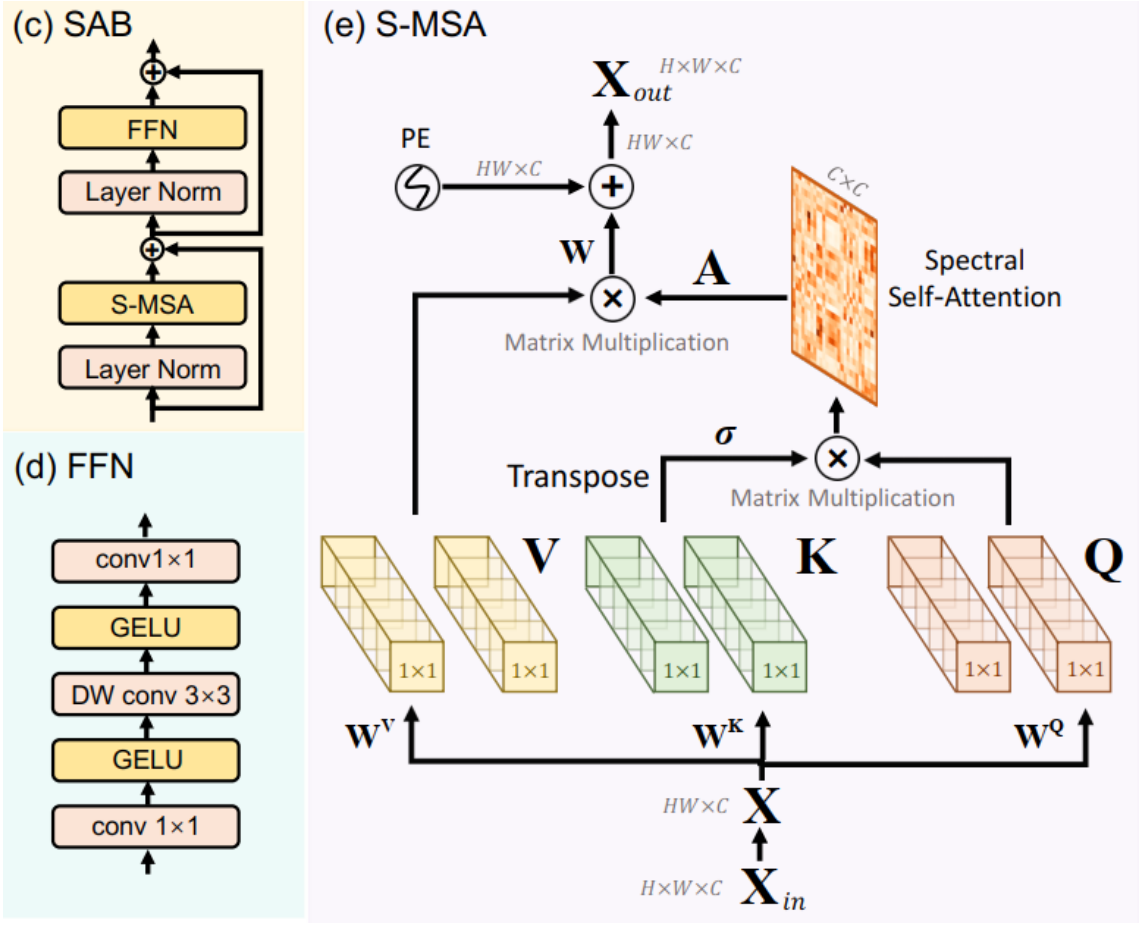

图 2: 特征提取模块

3.3 损失函数定义

采用的是 MRAE 损失函数：$Loss = |Y + R|/Y$

4 复现细节

4.1 与已有开源代码对比

将该模型应用到遥感领域高光谱重建，即将数据集更换为 150 个通道的高光谱图像，重新训练并验证模型。

4.2 实验环境配置

本地使用 11th Gen Intel(R) Core(TM) i7-11700 以及 32G 内存的个人电脑调试代码，使用服务器 P100 训练和测试模型

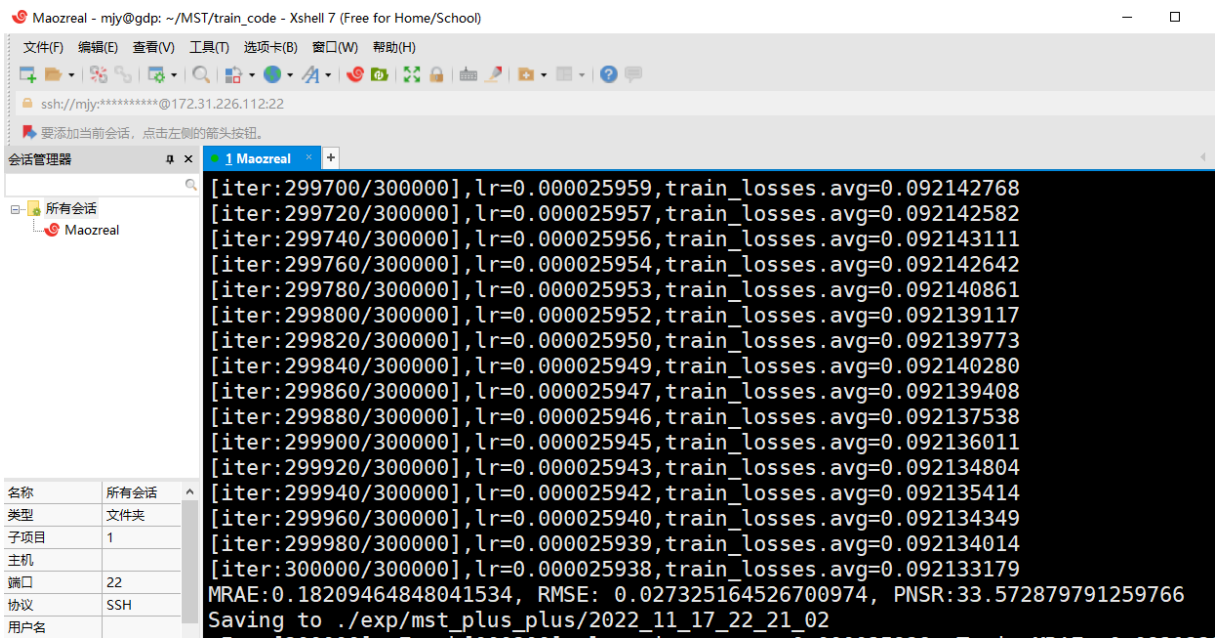

图 3: 训练过程

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

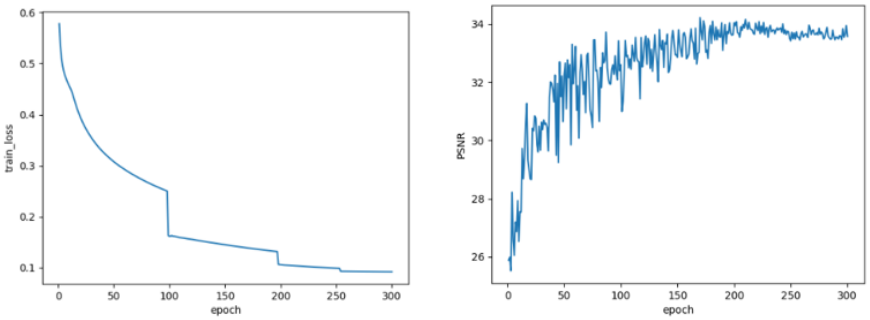

图 4: 训练损失和 psnr 随着训练轮次变化并最后收敛

图 5: 左为输入 RGB，右为输出 HSI（选取红 (30)，绿 (14)，蓝 (4) 波段）

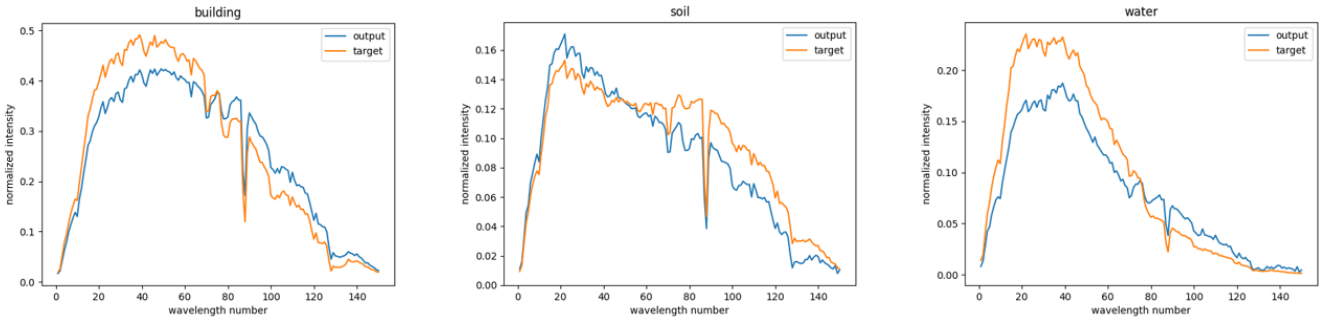

图 6: 光谱曲线对比，橙色为标签，蓝色为模型输出

参考文献

[1] BREUER M, ALBERTZ J. Geometric correction of airborne whiskbroom scanner imagery using hybrid auxiliary data[J]. International Archives of Photogrammetry and Remote Sensing, 2000, 33(B3/1; PART 3): 93-100.

[2] CAO X, YUE T, LIN X, et al. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world[J]. IEEE Signal Processing Magazine, 2016, 33(5): 95-108.

[3] YUAN X, BRADY D J, KATSAGGELOS A K. Snapshot compressive imaging: Theory, algorithms, and applications[J]. IEEE Signal Processing Magazine, 2021, 38(2): 65-88.

[4] MENG Z, MA J, YUAN X. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. 2020: 187-204.

[5] AESCHBACHER J, WU J, TIMOFTE R. In defense of shallow learned spectral reconstruction from RGB images[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 471-479.

[6] PARMAR M, LANSEL S, WANDELL B A. Spatio-spectral reconstruction of the multispectral datacube using sparse recovery[C]//2008 15th IEEE International Conference on Image Processing. 2008: 473-476.

[7] ARAD B, BEN-SHAHAR O. Sparse recovery of hyperspectral signal from natural RGB images[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. 2016: 19-34.

[8] CAI Y, HU X, WANG H, et al. Learning to generate realistic noisy images via pixel-level noise-aware adversarial training[J]. Advances in Neural Information Processing Systems, 2021, 34: 3259-3270.

[9] GALLIANI S, LANARAS C, MARMANIS D, et al. Learned spectral super-resolution[J]. arXiv preprint arXiv:1703.09470, 2017.

[10] XIONG Z, SHI Z, LI H, et al. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections[C]//Proceedings of the IEEE International Conference on Computer Vision

Workshops. 2017: 518-525.

[11] ZHAN S, CHANG C, XIONG Z, et al. Advanced CNN-Based Hyperspectral Recovery from RGB Images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA. 2018: 18-22.

[12] ZHANG L, LANG Z, WANG P, et al. Pixel-aware deep function-mixture network for spectral super-resolution[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 07. 2020: 12821-12828.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[14] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6836-6846.

[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. 2020: 213-229.

[16] CAO H, WANG Y, CHEN J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[J]. arXiv preprint arXiv:2105.05537, 2021.

[17] CAI Y, WANG Z, LUO Z, et al. Learning delicate local representations for multi-person pose estimation [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. 2020: 455-472.

[18] CAI Y, LIN J, HU X, et al. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17502-17511.

[19] CAI Y, LIN J, HU X, et al. Coarse-to-fine sparse transformer for hyperspectral image reconstruction[C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. 2022: 686-704.

[20] WANG Z, CUN X, BAO J, et al. A General U-Shaped Transformer for Image Restoration. arXiv 2021 [J]. arXiv preprint arXiv:2106.03106,

[21] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.