用于高光谱图像重建的从粗到细的稀疏变换器

摘要

人们已经开发了许多算法来解决编码孔径快照光谱成像 (CASSI) 的逆问题，即从 2D 压缩测量中恢复 3D 高光谱图像 (HSI)。近年来，基于学习的方法表现出了良好的性能并主导了主流研究方向。然而，现有的基于 CNN 的方法在捕获长程依赖性和非局部自相似性方面表现出局限性。之前基于 Transformer 的方法对输入进行密集采样，而因为高光谱信号空间稀疏，其中一些输入信息较少，没有进行计算多头自注意力 (MSA) 的必要。

在本文中，提出了一种新的基于 Transformer 的方法，从粗补丁到精细块的 Transformer (CST) 模型，方法首先将 HSI 稀疏性嵌入深度学习中以进行高光谱重建。同时，模型使用提出的光谱感知筛选机制 (SASM) 进行粗略补丁选择，再将选定的补丁输入到定制的频谱聚合散列多头自注意力（SAH-MSA）模块中，以进行精细像素聚类和自相似性捕获。

关键词： Transformer；光谱重建；编码孔径快照压缩成像

1 引言

人眼的视觉系统能够将不同波长的光赋予不同的颜色，从而区分自然界中的物体。为了模拟人类的视觉系统，光学成像系统得以应运而生。通常，人眼中具有三种视锥细胞，它们对各个波长的光有不同的响应。在受到光刺激时，这三种细胞会产生不同的响应，然后通过人的神经系统进行处理，最终赋予相应的颜色，从而实现对物体的感知。类似地，我们常用的 RGB 三色成像就是基于这一原理。它通过滤光片与传感器来实现对各个波长的光的不同响应，并通过图像处理算法获取目标的彩色图像。

然而，由于 RGB 图像的三个通道分别是有效光谱范围内的积分，这导致无法获取物体在不同波长下的响应，从而损失了很多光谱信息。相比之下，光谱成像不仅能获得目标场景的完整空间信息，还可以获取更多谱段的光强响应。因此，光谱成像在遥感、医学成像、环境监测、农业产业等领域得到了广泛应用。这种技术的优势在于它能够提供更多关于被测目标的特征和细节信息。

传统的光谱成像技术通常采用基于扫描的方法，例如点扫描、线扫描等。然而，这些成像方式通常在时间分辨率上有所牺牲，以换取更高的光谱分辨率，并且无法捕捉动态场景。随着光谱成像应用领域的扩展，传统的光谱成像已经难以满足多样化应用场景的需求。为了克服传统光谱成像技术的缺陷，一些研究者通过将计算和成像相结合的方式来获取目标的光谱数据。

与传统的光谱成像技术直接获取光谱信息不同，计算成像技术通常得到的是经过调制后的光谱信息，目标光谱数据需要通过计算来获取。计算成像技术克服了传统方法中时间分辨率较低的问题，实现了在高空间、时间和光谱分辨率下获取光谱图像。这为光谱成像技术的发展提供了新的方向。其中，编码孔径压缩光谱成像是一种典型的计算成像方式，它通过编码模板和棱镜对光进行调制，获取编码图像，然后通过计算重建出目标的光谱图像，具有获取光谱视频的能力。随着深度学习和人工智能的发展，深度网络在目标识别检测 [17]、图像识别分类 [18] 和机器翻译 [19] 等许多领域表现出优越的性能。凭借深度网络出色的学习能力，学者们开始使用深度网络以端到端的方式直接学习快照测量到高光谱图像的非线性映射，这种端到端的学习方法显著减少了重构时间。

在本文中，我们提出了一种用于高光谱重建的新方法，即从粗到细的稀疏 Transformer 模型 (CST)。模型中包含两项关键技术。首先，由于空间区域的 HSI 信息量变化很大，所以提出了一种用于粗补丁选择的光谱感知筛选机制（SASM）。具体来说，在图 1中，SASM 模块将图像划分为不重叠的补丁，然后检测提供 HSI 表示信息的补丁。随后，只有检测到的补丁（黄色）被输入到自注意力机制中，以减少无信息区域（绿色）的低效计算并提高模型的成本效益。其次，我们的目标是计算内容密切相关的令牌的自注意力，而不是像以前的变形金刚那样一次使用所有投影令牌。为此，我们定制了用于精细像素聚类的谱聚合散列多头自注意力（SAH-MSA）。SAH-MSA 通过搜索产生最大内积的相似元素来学习将标记聚类到不同的组。每个桶内的令牌被认为在内容上密切相关。然后在每个桶内应 MSA 操作。最后，利用所提出的技术，我们实现了一种从粗到精的学习方案，将 HSI 空间稀疏性嵌入到基于学习的方法中。我们建立了一系列从小到大的 CST 系列，它们的性能优于最先进的方法，同时需要便宜得多的计算成本。

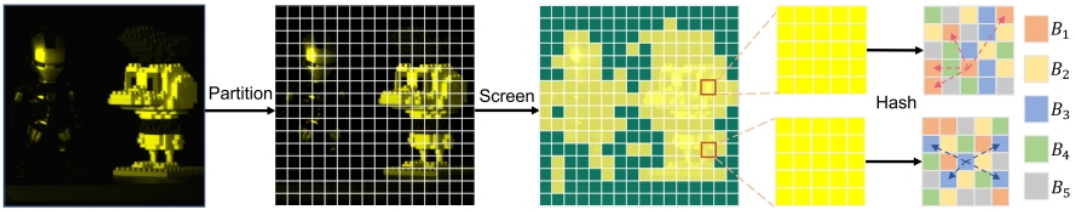

图 1. 模型方案示意图

2 相关工作

2.1 压缩光谱成像

早期，奈奎斯特 (Nyquist) 定理是信号进行采样需要满足的基本条件。奈奎斯特采样定理指出，如果采样率是信号最大频率分量的两倍或更大，则可以从样本中精确地恢复原始信号。这种采样方法在采样之前必须将频带不受限制的信号转换为频带受限的信号，需要消耗较多的计算资源、传输带宽和存储内存，而且容易产生大量的重建错误。近年来，压缩感知理论 (Compressed Sensing, CS) [9] 由 Cande’ s 和 Donoho 提出，即大多数可用信号具有稀疏性，如果用适当的变换基表示，则可以从相对较少的稀疏数据样本中重构原始信息。这项研究表明，信号的稀疏性和不一致性是数据样本可以从非常有限的采样信号中精确重建的两

个关键因素，利用压缩感知测量矩阵的有限等距性质 (Restricted Isometric Property, RIP) [7] 在采样端进行参数微调，可以确保随机信号样本中包含所需数量的信息内容，从而重建算法在恢复原始信息时可以获得更好的性能。与奈奎斯特采样定理相比，压缩感知理论进一步增强了从一组不完整数据中恢复信号的能力，在各种应用领域中都具有较好的适用性。

编码孔径快照光谱成像 (CASSI) [2] 系统是一种典型的快照成像方式，学者们使用分散器/棱镜和编码孔径建立了各种编码孔径快照光谱成像系统。双分散器编码孔径快照光谱成像 (DD-CASSI) 系统是 CASSI 系统的先驱，它使用一个编码孔径和两个分散器来实现光谱调制，在空间域和光谱域中进行编码。后来又开发了单分散器编码孔径快照光谱成像 (SD-CASSI) 系统，包含一个编码孔径和一个分散器，通过移动分散器实现光谱调制，并且只在空间域中进行编码。最近，考虑到其他介质对光谱变化的响应，基于磨玻璃的光场调制器、散射器和空间光调制器也被用于构建 CASSI 系统。如图 2所示，SD-CASSI 系统由编码孔径、单分散器、探测器和物镜组成。入射光首先通过物镜汇聚到达编码孔径。然后，使用单分散器对掩码调制后的空间频谱信息进行分散。最终，在探测器上获得二维快照测量图像。

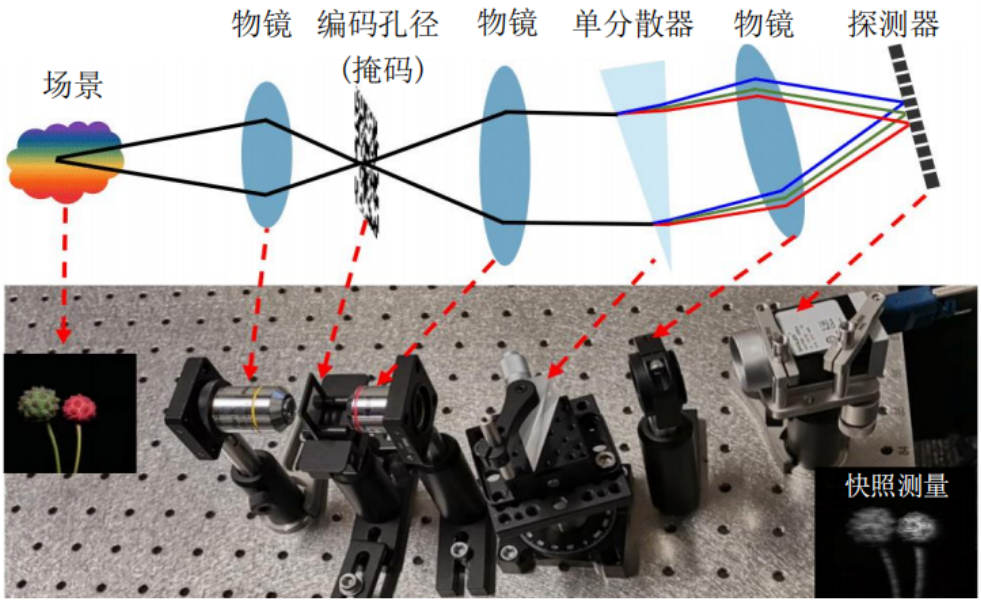

图 2. SD-CASSI 系统示意图

2.2 传统光谱重建

传统的高光谱图像重建方法基于稀疏编码，利用了一些容易获得的先验知识（例如稀疏性和低秩）来实现高光谱重建。例如，梯度投影 [6] 算法被用来处理 HSI 稀疏性，GAP-TV [15] 使用了全变分正则化器，而 DeSCI 模型 [11] 使用了低秩属性和非局部自相似性。然而，这些传统的基于模型的方法存在重建速度慢、泛化能力差的问题。最近，CNN 已被用来解决光谱 SCI 的逆问题。这些基于 CNN 的算法可以分为三类，即端到端（E2E）方法、深度展开方法和即插即用（PnP）方法。E2E 算法应用深度 CNN 作为强大的模型来学习 HSI 恢复的 E2E 映射函数。深度展开方法采用经过训练的多级 CNN，将测量结果映射到所需信号。每个阶段包含两个部分，即线性投影和将信号传递给充当降噪器的 CNN。PnP 方法将预先训练的 CNN 降噪器插入到基于模型的方法中来解决 HSI 重建问题。尽管如此，这些基于 CNN 的算法在

捕获远程空间依赖性和建模非局部自相似性方面表现出局限性。此外，HSI 表示的稀疏性没有得到很好的解决，给 HSI 重建模型带来了低效率的问题。

2.3 Transformer 模型

人类对事物的观察通常更侧重于关键部分，而忽略其他相对不重要的信息，这被称为“注意焦点”。深度学习中的注意力机制（Attention Mechanism）受到人类视觉特点的启发，通过为输入特征的每个部分分配不同的权重，迅速捕捉当前任务中的重要信息，并抑制低价值或无关信息。这种机制可以被视为一种启发式方法，模仿人类处理信息时关注关键细节的方式，从而提高模型在处理复杂任务时的效率和准确性。通过注意力机制，深度学习模型能够更有效地处理大量输入数据，更好地适应不同任务的复杂性，实现更智能的学习和推理。

在自然语言处理领域中，Vaswani 等人 [13] 提出了一种自注意力机制 (Self-Attention)，并进一步将多个自注意力机制拼接得到多头注意力机制 (Multi-Head Self-Attention)。多头注意力机制在 transformer 网络 [8] 中表现出优越的性能。Beyer 等人 [5] 提出了一种用于图像分类任务的 Vision Transformer 网络，该网络主要包括三部分：图像块的线性投影操作、Transformer 编码器和多层感知机分类头 (MLP Head)。MST [3] 是第一个提出将 Transformer 模型应用在高光谱重建工作上的工作，它将每一波段的光谱作为输入，并计算沿谱维度的自注意力。王震东等人 [14] 提出了一种由 Swin Transformer [10] 的基本块构建的用于自然图像恢复的 UFormer 模型。然而，现有的 Transformer 对输入进行密集采样，其中一些输入对应于信息有限的区域，并计算一些内容不相关的信息之间的多头注意力。如何将高光谱空间稀疏性嵌入到 Transformer 模型中以提高模型效率仍有待研究。本文的工作旨在填补这一研究空白。

3 本文方法

3.1 网络结构

CST 总体框架图如图 3所示。CST 由两个关键组件组成，即用于粗补丁选择的频谱感知筛选机制（SASM）和频谱聚合散列多重用于精细像素聚类的头部自注意力（SAH-MSA）。图 3(a) 描述了 SASM 和 CST 的网络架构。图 3(b) 显示了 CST 的基本单元，频谱感知散列注意块（SAHAB）。图 3(c) 说明了 SAHAB 最重要的组成部分的 SAH-MSA。

给定 2D 测量 $\mathbf{Y} \in \mathbb{R}^{H\times(W+d(N_\lambda-1))}$，反转方程中的色散向后移 $\mathbf{Y}$ 得到初始化输入信号 $\mathbf{H} \in \mathbb{R}^{H\times(w)\times N_\lambda}$ 为

$$\mathbf{H}(x, y, n_\lambda) = \mathbf{Y}(x, y - d(\lambda_n - \lambda_c)). \tag{1}$$

然后 $\mathbf{H}$ 与 **3D** 物理掩码 $\mathbf{M} \in \mathbb{R}^{H\times W\times N_\lambda}$ 连接，通过内核大小为 1×1 的卷积层生成初始化特征 $\mathbf{X} \in \mathbb{R}^{H\times W\times N_\lambda}$。

首先，利用稀疏度估计器，将 $\mathbf{X}$ 处理为稀疏掩模 $\mathrm{M}_s \in \mathbb{R}^{H\times W}$ 和浅层特征 $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$。其次，浅层特征 $\mathbf{X}_0$ 经过三级对称编码器-解码器并嵌入到深层特征 $\mathbf{X}_d \in \mathbb{R}^{H\times W\times C}$ 中。编码器或解码器的第 i 级包含 $\mathbf{N}_i$ 个 SAHAB 模块。如图 3(b) 所示，SAHAB 由两层归一化层、SAH-MSA 和前馈网络组成。为了减轻下采样操作造成的信息损失，编码器特征通过恒等连接与解码器特征聚合。最后，将通过内核大小为 3×3 的卷积层应用于 $\mathbf{X}_d$ 以生成残差 $\mathbf{R} \in \mathbb{R}^{H\times W\times N_\lambda}$。

然后将 $\mathbf{R}$ 与 $\mathbf{X}$ 相加即可得到重建的高光谱图像为 $\mathbf{X}'$，即 $\mathbf{X}' = \mathbf{X} + \mathbf{R}$。在实验中，设置基本通道 $C = N_\lambda = 28$ 来存储 HSI 信息，并更改图 3(a) 中的组合 (N1,N2,N3)，以建立具有小型、中型和大型模型的 CST 系列大小和计算复杂性，它们分别是 CST-S(1,1,2)、CST-M (2,2,2) 和 CST-L(2,4,6)。

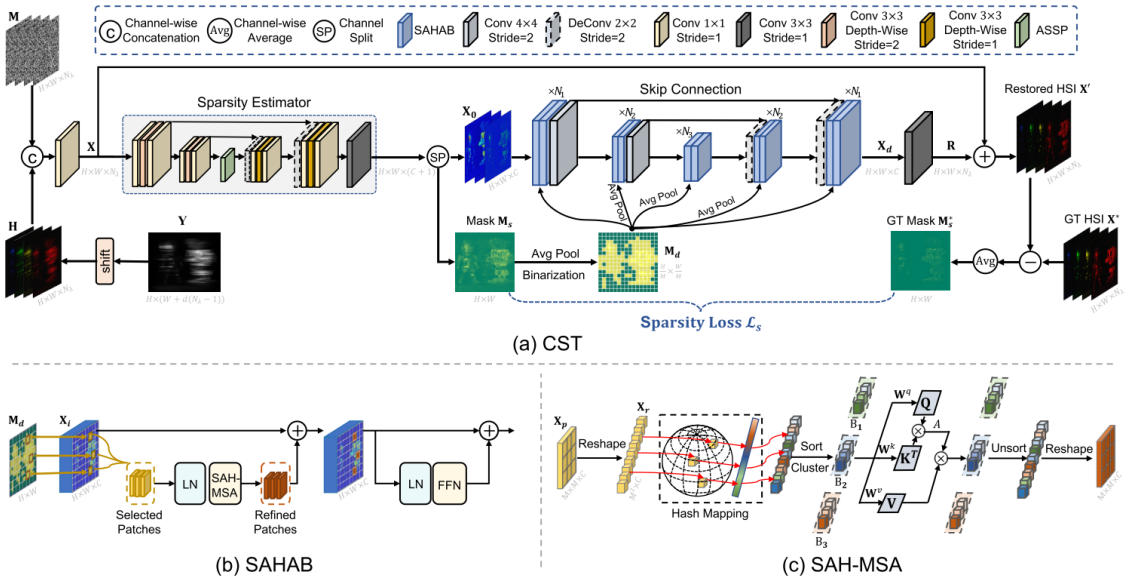

图 3. 整体框架图

3.2 光谱感知筛选机制

高光谱信号在空间维度上呈现出明显的高度稀疏性。当前的 Transformer 模型通常需要对许多缺乏信息的区域进行采样以计算多头注意力，这可能会导致模型效率下降。为了解决这一问题，本文提出了一种称为 SASM（Sparse Attention for Spectral Mapping）的方法，通过进行粗略的补丁选择，即筛选出具有密集高光谱信息的区域来生成令牌。SASM 的提出旨在有效处理高光谱信号的稀疏性，通过有选择地关注具有丰富信息的区域，减少对无信息区域的采样。这一方法有望提高模型效率，加速计算过程，同时更准确地捕捉高光谱图像中的重要特征。通过引入 SASM，本文尝试优化 Transformer 模型以更好地适应高光谱数据的特殊性。

首先，介绍稀疏度估计器的结构。如图 3（a）所示，稀疏估计器采用 U 形结构，包括两级编码器、ASSP 模块 [4] 和两级解码器。编码器的每个阶段由两个卷积核为 1×1 的卷积层和一个卷积核为 3×3 的卷积层组成。解码器的每个阶段都包含一个反卷积核的大小为 2x2 的卷积层、两个卷积核为 1×1 的卷积层和一个卷积核为 3×3 的卷积层组成。稀疏估计器将初始化的移位特征 $\mathbf{X}$ 作为输入，生成浅层特征 $\mathbf{X}_0$ 和稀疏掩模 $\mathbf{M}_s$，用于定位并筛选出具有 HSI 表示的信息丰富的空间区域。同时，通过最小化稀疏性损失来实现这一模块功能。不同的稀疏方案的差别如图 4所示，（i）随机稀疏，即要计算的补丁是随机选择的，（ii）均匀稀疏，即要计算的补丁是均匀分布的，以及（iii）本文所提出的 SASM 模块。

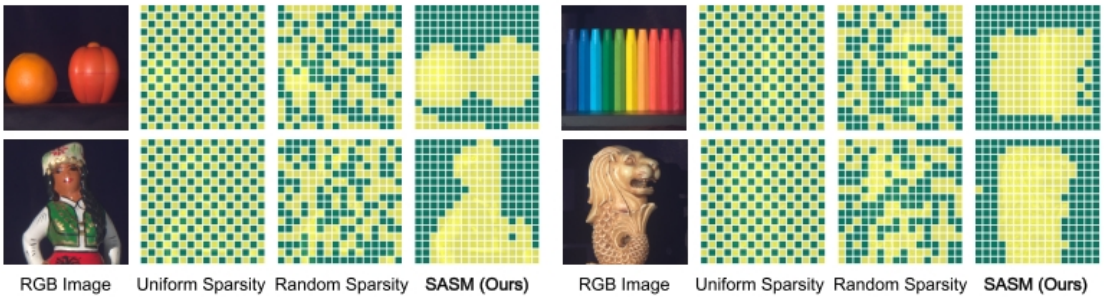

图 4. 稀疏方案可视化对比图

为了监督 $\mathbf{M}_s$，需要一个参考值来体现 HSI 上空间稀疏信息聚焦在哪些输入上。由于背景较暗且信息量不大，因此，通过沿谱维度对重建的高光谱图像 $\mathbf{X}'$ 和真实高光谱图像 $\mathbf{X}^*$ 之间的差异进行平均来设计参考信号 $\mathbf{M}_s^* \in \mathbb{R}^{H\times W}$，以避免偏差：

$$\mathbf{M}_s^* = \frac{1}{N_\lambda}\sum_{n_\lambda=1}^{N_\lambda} |\mathbf{X}'(:,:,n_\lambda) - \mathbf{X}^*(:,:,n_\lambda)|. \tag{2}$$

随后，稀疏损失 $\mathcal{L}_s$ 被构造为预测稀疏掩模 $\mathbf{M}_s$ 和参考稀疏掩模 $\mathbf{M}_s^*$ 之间的均方误差，如下:

$$\mathcal{L}_s = ||\mathbf{M}_s - \mathbf{M}_s^*||_2. \tag{3}$$

通过最小化 $\mathcal{L}_s$，稀疏性估计器检测排查前景难以重建的区域。此外，总体训练目标 $\mathcal{L}$ 是 $\mathcal{L}_s$ 和 $\mathcal{L}_2$ 损失的加权和：

$$\mathcal{L} = \mathcal{L}_2 + \lambda\cdot\mathcal{L}_s = ||\mathbf{X}' - \mathbf{X}^*||_2 + \lambda\cdot||\mathbf{M}_s - \mathbf{M}_s^*||_2. \tag{4}$$

其中 $\mathbf{X}^*$ 代表真实的高光谱图像，λ 指控制 $\mathcal{L}_2$ 和 $\mathcal{L}_s$ 之间重要性平衡的超参数。

实验时，SASM 模块将特征图划分为大小为 M×M 的不重叠的块。然后，通过预测的稀疏掩模 $\mathbf{M}_s$ 筛选出具有 HSI 表示的粗补丁，并将其输入 SAH-MSA，如图 3（b）所示。具体来说，Ms 首先通过平均池化进行下采样，然后二值化为 $\mathbf{M}_d \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}}$。接着，使用超参数稀疏比 σ 来控制二值化。更具体地说，选择下采样稀疏掩模上具有最高值的前 k 个补丁。k 由 σ 控制，即 $k = \lfloor(1-\sigma)\frac{HW}{M^2}\rfloor$。$\mathbf{M}_d$ 上的每个像素对应于特征图上的 M×M 块，表示这个补丁块是否被筛选掉。然后将 $\mathbf{M}_d$ 应用于每个 SAHAB 的 SAH-MSA。当 $\mathbf{M}_d$ 用于第 i 阶段 (i > 1) 时，利用平均池化操作将 $\mathbf{M}_d$ 下采样为 $\frac{1}{2^{i-1}}$ 大小，以匹配第 i 阶段特征图的空间分辨率。

3.3 光谱聚合散列多头自注意力

过去的 Transformer 模型在计算注意力机制时涉及对所有采样令牌之间的多头自注意力机制（MSA），其中一些令牌可能在内容上没有关联。这种全连接的计算方式可能导致计算效率低下，从而降低模型的成本效益并且容易阻碍收敛。稀疏编码方法基于这样的假设，即图像信号可以通过字典信号的稀疏线性组合来表示。受到这一观点的启发，本文提出了一种用于精细像素聚类的 SAH-MSA（Sparse Attention Head MSA）。SAH-MSA 强制在 MSA 机制中实施稀疏性约束，从而更有效地处理令牌之间的关系，减少不相关的计算，提高计算效率。这种方法有助于改善模型的性能，提高收敛速度，并在计算方面更具效益。

SAH-MSA 通过搜索产生最大内积的元素来学习将标记聚类到不同的桶中。如图 3(c) 所示，将由稀疏掩模筛选出的块特征图表示为 $\mathbf{X}_p \in \mathbb{R}^{M\times M\times C}$。将 $\mathbf{X}_p$ 重塑为 $\mathbf{X}_r \in \mathbb{R}^{N\times C}$，其中 N = M×M 是元素数量。随后，使用哈希函数以光谱方式聚合信息，并将 C 维元素（像素向量）$\mathbf{x} \in \mathbb{R}^C$ 映射为整数哈希码。我们将这个哈希映射 $h : \mathbb{R}^C \to \mathbb{Z}$ 表述为

$$h(\boldsymbol{x}) = \lfloor \frac{\boldsymbol{a}\cdot\boldsymbol{x}+b}{r} \rfloor \tag{5}$$

其中 $r \in \mathbb{R}$ 是常数，$\mathbf{a} \in \mathbb{R}^{\mathrm{C}}$ 和 $b \in \mathbb{R}$ 是满足 $\mathbf{a} = (a_1, a_2, ..., a_C)$，同时 $a_i \sim \mathcal{N}(0,1)$ 和 $b \sim \mathcal{U}(0,r)$ 的随机变量服从均匀分布。然后根据哈希码对 X_p 中的元素进行排序。第 i 个排序元素表示为 $\mathbf{x}_i \in \mathbb{R}^{\mathrm{C}}$。然后将元素分成桶，

$$\mathbf{B}_i = \{\boldsymbol{x}_j : im+1 \le j \le (i+1)m\} \tag{6}$$

其中 $\mathbf{B}_i$ 表示第 i 个桶。每个桶有 m 个元素。总共有个 $\frac{M\times M}{m}$ 桶。通过哈希聚类方案，内容密切相关的标记被分组到同一个桶中。因此，该模型可以通过仅对同一桶内的令牌应用 MSA 操作来减少内容无关元素之间的计算负担。更具体地说，对于查询元素 $q \in \mathbf{B}_i$，所以 SAH-MSA 可以表示为：

$$\text{SAH-MSA}(\boldsymbol{q}, \mathbf{B}_i) = \sum_{n=1}^{N} \mathbf{W}_n \text{head}_n(\boldsymbol{q}, \mathbf{B}_i) \tag{7}$$

其中 N 是注意力头的数量。$\mathbf{W}_n \in \mathbb{R}^{C\times d}$ 和 $\mathbf{W'}_n \in \mathbb{R}^{d\times C}$ 是可学习参数，其中 $d = \frac{C}{N}$ 表示每个头的尺寸。A_{nqk} 和 $head_n$ 指的是第 n 个头的注意力和输出，公式为

$$\mathbf{A}_{n\boldsymbol{qk}} = \underset{\boldsymbol{k}\in\mathbf{B}_i}{\text{softmax}}(\frac{\boldsymbol{q}^T\mathbf{U}_n^T\mathbf{V}_n\boldsymbol{k}}{\sqrt{d}}) \tag{8}$$

$$\text{head}_n(\boldsymbol{q}, \mathbf{B}_i) = \sum_{k\in\mathbf{B}_i} \mathbf{A}_{n\boldsymbol{qk}}\mathbf{W'}_n k \tag{9}$$

其中 $\mathbf{U}_n$ 和 $\mathbf{V}_n \in \mathbb{R}^{d\times C}$ 是可学习参数。通过散列方案，相似的元素落入不同桶的可能性很小。通过并行进行多轮散列可以进一步降低这种概率。$\mathbf{B}_i^r$ 表示第 r 轮的第 i 个桶。那么对于每个头，多轮输出是每个单轮输出的加权和，即

$$\text{head}_n(\boldsymbol{q}, \mathbf{B}_i) = \sum_{r=1}^{R} w_n^r \text{head}_n(\boldsymbol{q}, \mathbf{B}_i^r) \tag{10}$$

其中 R 表示轮数，w_n^r 表示第 n 个头中第 r 轮的权重重要性，对查询元素 q 与属于桶 $\mathbf{B}_i^r$ 的元素之间的相似度进行评分。w_n^r 可以通过以下方式获得

$$w_n^r = \frac{\sum_{\boldsymbol{k}\in\mathbf{B}_i^r} A_{n\boldsymbol{qk}}}{\sum_{\hat{r}=1}^{R}\sum_{\boldsymbol{k}\in\mathbf{B}_i^{\hat{r}}} A_{n\boldsymbol{qk}}} \tag{11}$$

4 复现细节

4.1 与已有开源代码对比

高光谱图像是光谱分辨率在 $10^{-2}\lambda$ 数量级范围内的光谱图像。相较于 RGB 图像和多光谱图像而言，高光谱图像有着更多的波段信息来更加准确全面的描述被捕获场景的特性。然而，高光谱数据的维度较高，这可能导致维度灾难问题和计算负担。

不同数据集上数据通过不同的方式不仅且包括不同的传感器、不同的时间地点等，所以数据的差异较大。因此在本次研究中，致力于复现文章的工作，并在此过程中，尝试了应用文章中没有使用到的数据集，验证了本文的方法在不同数据集上都有一定的效果。在通过参考了本论文的开源代码作为我们复现的基础，链接如下 https://github.com/caiyuanhao1998/MST。

4.2 实验环境搭建

表 1. 实验环境

CPU	12 x Xeon Gold 6271
GPU	NVIDIA Tesla P100-16GB
内存	16GB
内存	48GB
存储	1.7TB

5 实验结果分析

实验结果如下表所示：

表 2. 实验环境

方法	数据集	峰值信噪比	结构相似性指数
CST-S（原文）	CAVE [12]	34.71	0.94
CST-S（复现）	CAVE [12]	34.72	0.94
CST-S（复现）	NTIRE2022 [1]	32.62	0.90

6 总结与展望

在本文中，我们全面探讨了基于 CST 的模型框架，并在此基础上进行了有益的扩展。主要的工作主要集中于复现了论文的实现，同时迁移到另一个数据集上进行验证。通过验证，本方法具有较好的鲁棒性，可以在不同的数据集上起到效果。未来的研究方向包括但不限于以下几点：首先，可以探索其他稀疏估计度模型，通过模型筛选信息量大的块之后再进行重建任务。其次，可以考虑在模型中引入其他的深度学习技术，以进一步提升模型的性能。此外，跨领域的迁移学习和模型的可解释性也是未来研究的潜在方向。

参考文献

[1] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, et al. Ntire 2022 spectral recovery challenge

and data set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 863–881, 2022.

[2] Carin L et al. Arce G R, Brady D J. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31(1):105–115, 2013.

[3] Lin J. Hu X. Wang H. Yuan X. Zhang Y. Timofte R. Gool L.V. Cai, Y. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 17502–17511, 2022.

[4] Schroff F et al. Chen L C, Papandreou G. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Nowak R.D. Wright S.J. Figueiredo, M.A. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 2007.

[7] Candes E J. The restricted isometry property and its implications for compressed sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 346((9-10)):589–592, 2008.

[8] Zisserman A. Jaderberg M, Simonyan K. Spatial transformer networks. *Advances in neural information processing systems*, 28:236–248, 2015.

[9] Donoho D L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[10] Lin Y. Cao Y. Hu H. Wei Y. Zhang Z. Lin S. Guo B. Liu, Z. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. Int. Conf. on Computer Vision*, pages 10012–10022, 2021.

[11] Yuan X. Suo J. Brady D. Dai Liu, Y. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[12] Jong-Il Park, Moon-Hyun Lee, Michael D Grossberg, and Shree K Nayar. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[13] Parmar N et al. Vaswani A, Shazeer N. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Bao J et al. Wang Z, Cun X. Uformer: A general u-shaped transformer for image restoration. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 17683–17693, 2022.

[15] X. Yuan. Generalized alternating projection based total variation minimization for compressive sensing. *IEEE International Conference on Image Processing*, 2016.