Mask-guided Spectral-wise Transformer for Efficient Hyperspectral Image Reconstruction

摘要

本文所复现的高光谱图像（HSI）重建算法是在编码孔径快照光谱成像（CASSI）系统中，从二维测量中恢复三维空间光谱信号。高光谱图像在光谱维度上是高度相似和相关的，因此光谱间相关性的利用有利于 HSI 的重建。然而，现有的基于卷积神经网络的方法在捕获谱间相似性和远距离相似性方面存在局限性。此外，在 CASSI 中，高光谱图像信息通过编码孔径（物理掩码）进行调制。然而，目前的重建算法并没有充分挖掘物理掩码对 HSI 恢复的引导作用。在复现的文献中，作者提出了一种用于 HSI 重建的新框架，基于掩码指导的逐光谱 Transformer（MST）高光谱图像重建算法。具体来说，作者提出了一种逐光谱多头自注意力机制（S-MSA），它将每个波段的光谱特征视为一个令牌，并沿光谱维度应用自注意力机制。此外，作者还提出了一个掩模引导机制（MM），指导 SMSA 关注具有高保真光谱表示的空间区域。大量实验表明，MST 在 HSI 数据集上明显优于最先进的（SOTA）方法，同时只需要更低的计算和内存成本。

关键词：高光谱图像重建算法；多头自注意力；卷积神经网络

1 引言

高光谱图像数据是一个三维数据，既包含了空间场景，也包含了场景内物体的光谱信息。高光谱图像的三个维度分别对应于空间场景的高、宽以及光谱维度。相比于传统的 RGB 图像，高光谱拥有更多的光谱维度，包含更多的物体信息，因此高光谱图像常常用在遥感成像、农业检测、物体识别与分类等领域。高光谱成像技术可以分为两种。第一种是扫描式成像。扫描式成像借助于机械结构和光学元件，每扫描一次获取部分数据，虽然可以获得准确的光谱数据，但是成像速度慢。为了解决上述问题，进一步衍生出了快照式高光谱成像技术，快照式成像仅仅曝光一次，极大地提高了成像速度。传统的快照式成像依靠光学物理器件，因此相机体积大、光谱分辨率低，无法应用到广泛的现实场景中。伴随着压缩感知的理论提出，基于压缩感知的高光谱图像重建算法应运而生。压缩感知理论指出，对于一个稀疏信号或者在某种变换域下是稀疏的信号，可以使用较低的采样频率进行采样，使用重建算法对测量值进行重建。CASSI 是一种基于编码孔径的高光谱图像采样方法，它使用编码孔径和光栅对入射光进行调制，图像传感器进一步将调制过的三维数据压缩成了二维数据，重建算法对该二维数据进行处理，将其重建为三维数据，也就是高光谱图像。传统的重建算法是基于数学建模的迭代优化算法，这类算法重建结果精确，但计算时间长，无法满足快照式成像的应用成像；卷积神

经网络可以较好地解决这个问题，但如何提高重建准确率是限制其应用的主要问题。近年来，自然语言处理（NLP）模型 Transformer [19] 已被引入计算机视觉，并在许多任务中优于 CNN 方法。Transformer 中的多头自注意力（MSA）机制擅长捕获非局部相似性和远距离依赖性。但是，直接应用原始 Transformer 可能不适合进行 HSI 恢复。首先，原始的 Transformer 学习捕捉空间维度上的远距离依赖性，但 HSI 在光谱维度上是高度相似的。在这种情况下，光谱间的相似性和相关性没有被很好地建模及利用。同时，光谱信息在空间维度上具有稀疏性。捕获空间维度上的相关特征的对于 HSI 重建可能收效甚微。其次，在 CASSI 系统中，HSI 经过物理掩码片调制。原始的 Transformer 在使用自注意力机制时，没有充分的利用物理掩码片所能提供的信息，很容易注意到许多低保真度和信息量较少的图像区域，这可能会降低重建的效率。第三，当使用原始全局 Transformer [7] 时，计算复杂度是空间分辨率大小的二次方。这一负担非同小可，有时甚至难以承受。当使用基于局部窗口的 Transformer [15] 时，MSA 模块的感受野被限制在特定位置的窗口内，因此一些远距离的、高度相关的特征就会被忽略。为了克服一般 Vision Transformer 的局限性，作者提出了一种新的，基于物理掩码片指导的光谱间 Transformer 技术（MST）的高光谱图像重建算法。首先，在图 1(a) 中可以观察到，由于特定波长的限制，HSI 的每个光谱通道包含着同一场景的不完整的部分。这表明 HSI 表示沿着光谱维度存在相似性的并且部分波段之间是互补的。因此，作者提出了一种光谱间的多头自注意力机制（S-MSA）来捕捉远距离光谱间的依赖性。其次，在图 1（b）中，在 CASSI 系统中使用物理掩码片来调制 HSI。掩码片上不同位置的透光率变化很大。这表明调制光谱信息的保真度是位置敏感的。因此，作者利用掩码片作为一个关键线索，并提出了一种新的掩码指导机制（MM），指导 S-MSA 模块注意到具有高保真光谱表示的区域。同时，MM 算法也缓解了 S-MSA 在 HSI 表示空间相关性建模方面的局限性。

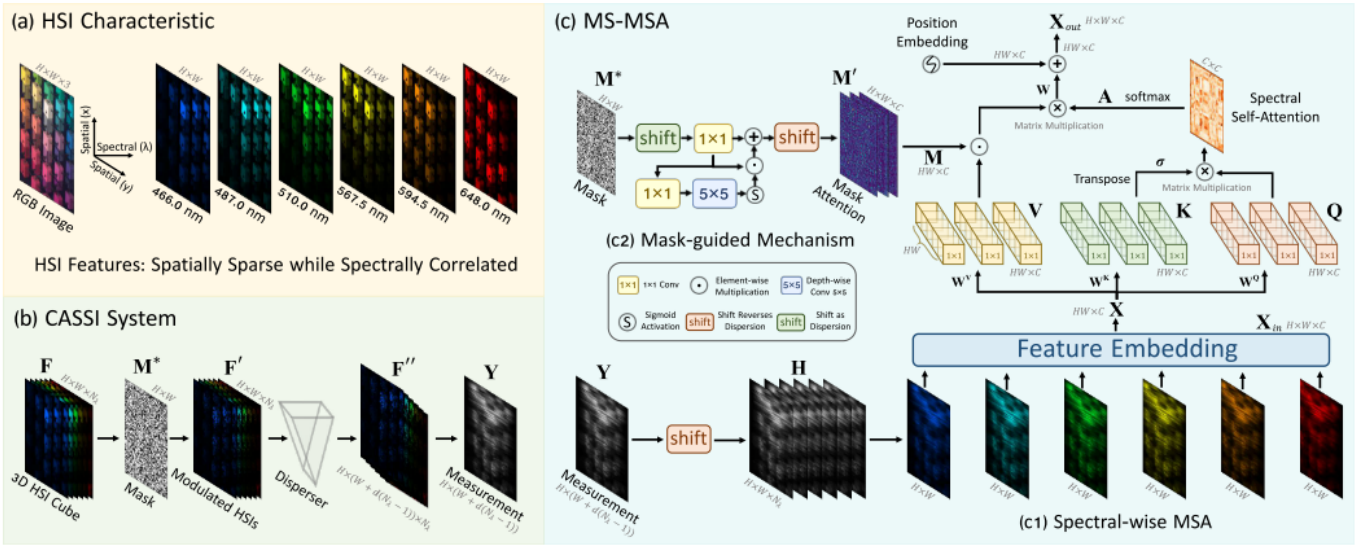

图 1. 方法示意图 (a) 高光谱图像表示；(b)CASSI 采样系统示意图；(c) 本文提出的多头自注意力机制示意图

在文章中，作者提出了一种新的方法 MST，用于 HSI 重建。这是第一次尝试挖掘 Transformer 在 HSI 重建任务中的潜力。具体而言，作者提出了一种新的自注意力机制，S-MSA，用于捕捉光谱间的相似性和依赖性。作者设计了一个掩码片指导机制 MM，指导 S-MSA 注意高保真度的区域。MST 方案显着优于 SOTA 方法在模拟的所有场景，同时只需要更少的

参数和运算。此外，MST 在真实世界的 HSI 重建中产生了更好的视觉效果。

2 相关工作

2.1 高光谱图像重建

传统的高光谱图像重建算法，使用了人工定义的先验特征 [8, 11, 13, 14, 18, 20, 22–24]，对这些先验特征进行数学建模，使用一些数学算法进行迭代求解。例如，GAP-TV [23] 方法就引入了全变分的先验知识，DeSCI [14] 则探索了高光谱图像的低秩特性和非局部相似性。然而，这些基于数学建模的方法由于表达能力差而不能达到令人满意的性能和通用性。最近，深度卷积神经网络已经被应用于学习 HSI 重建的端到端映射函数 [9, 10, 16, 17, 21]，并且有能力带来更理想的效果的性能。TSA-Net [16] 使用三个空间-光谱自注意力模块来捕获空间或光谱维度中的依赖性。引入较多的空间-光谱自注意力模块使得算法的计算开销更大，但是效果提升却很有限。DGSMP [10] 提出了一种可解释的 HSI 重建方法，学习的混合高斯尺度（GSM）先验知识。这些基于 CNN 的方法产生了优秀的性能表现，但在挖掘光谱间的相似性和相关性方面存在局限性。此外，目前的基于 CNN 的重建方法并没有研究物理掩码片的指导效果。

2.2 Vision Transformer

Transformer 的首次提出是用于机器翻译。目前，Transformer 已经广泛用于图像处理领域，并且取得了不俗的效果，尤其是在图像分类 [1, 2, 7, 15]、物体检测 [4–6, 25]、图像分割 [3] 和人类姿态估计 [12] 领域。例如，SwinIR 使用 Swin Transformer 块来建立残差网络，并在图像重建中实现优越的重建效果。然而，这些 Transformer 主要旨在捕获空间维度上的长范围依赖性。对于光谱相似和经过物理掩码片调制的 HSI，直接应用先前的 Transformer 在捕获光谱相关性方面可能不太有效。此外，多头自注意力机制可能会关注信息较少的空间区域，无法关注到包含信息量多的区域。

2.3 CASSI 采样系统

图 1(b) 简单的展示了 CASSI 采样的原理。当给定一个三维的高光谱图像数据立方体 $F \in R^{H\times W\times N_\lambda}$，$H$，$W$，$N_\lambda$ 分别代表高光谱图像的高、宽和波段数。图像首先经过物理掩码片 $M^* \in R^{H\times W}$ 的调制，调制过程如下：

$$F'(:,:,n_\lambda) = F(:,:,n_\lambda) \odot M^*$$

F' 即为调制后的数据，$\odot$ 代表矩阵的哈达玛积。经过一个色散器后，每个波段的图像数据在同一个轴上产生位移，令发生位移的轴为 y 轴，令 $F'' \in R^{H\times(W+d(N_\lambda-1))\times N_\lambda}$ 表示经过色散器后的图像数据，d 为移动的步长，假设 λ_c 为参照波长，该波长的光经过色散器不会发生位移，那么色散器的调制作用如下：

$$F''(u,v,n_\lambda) = F'(x, y+d(\lambda_n-\lambda_c), n_\lambda)$$

(u,v) 代表传感器平面上的二维空间坐标点，λ_n 代表第 n 个光谱波段的中心波长，$d(\lambda_n-\lambda_c)$ 代表第 n 个波段产生的位移，最终图像传感器产生的二维图像 $Y\in R^{H\times d(N_\lambda-1)}$ 为

$$Y=\sum_{n_\lambda}^{N_\lambda}F''(:,:,n_\lambda)+G$$

G 为获取图像时产生的噪声。

3 本文方法

3.1 总体架构

文章所提出的算法整体架构如图 2(a) 所示。MST 的整体架构是一个 U 型结构，类似于 U-net，包含一个编码器和解码器以及中间过程 Bottleneck。具体来说，MST 是由物理掩码片指导的光谱自注意力机制 (MSAB) 构成的。首先，算法对分散器的调制过程进行反向操作将测量值移回以获得一个初始化信号 $S\in R^{H\times W\times N_\lambda}$，具体过程如下：

$$S(x,y,n_\lambda)=Y(x,y-d(\lambda_n-\lambda_c))$$

将该初始化信号 S 作为输入传入的神经网络中。首先，MST 利用一个卷积核大小为 3×3 的卷积层从初始化信号 S 中提取出特征图 $X_0\in R^{H\times W\times C}$。接下来，$X_0$ 先后经过 N_1 个 MSAB 块的处理，一个降采样卷积层，N_2 个 MSAB 块的处理和一个降采样卷积层后生成层次特征。编码器中的降采样层使用 4×4 的卷积核，将图像的高宽减半并将通道数加倍。接下来，编码器的输出 X_2 进入到由 N_3 个 MSAB 构成的 Bottleneck 层中。参照 U-net 中编解码器的思想，MST 同样包含一个与编码器对称的解码器。编码器中的降采样层对应一个解码器中的上采样层，因此上采样层使用大小为 2×2 的卷积核将输入的特征图的高宽加倍，通道数减半。跳跃连接则将编码器与解码器对应阶段的特征图汇聚，避免降采样导致的信息丢失。最后，通过一个 3×3 的卷积层生成一个残差图像，与初始值信号相加则得到最终的高光谱图像重建结果。

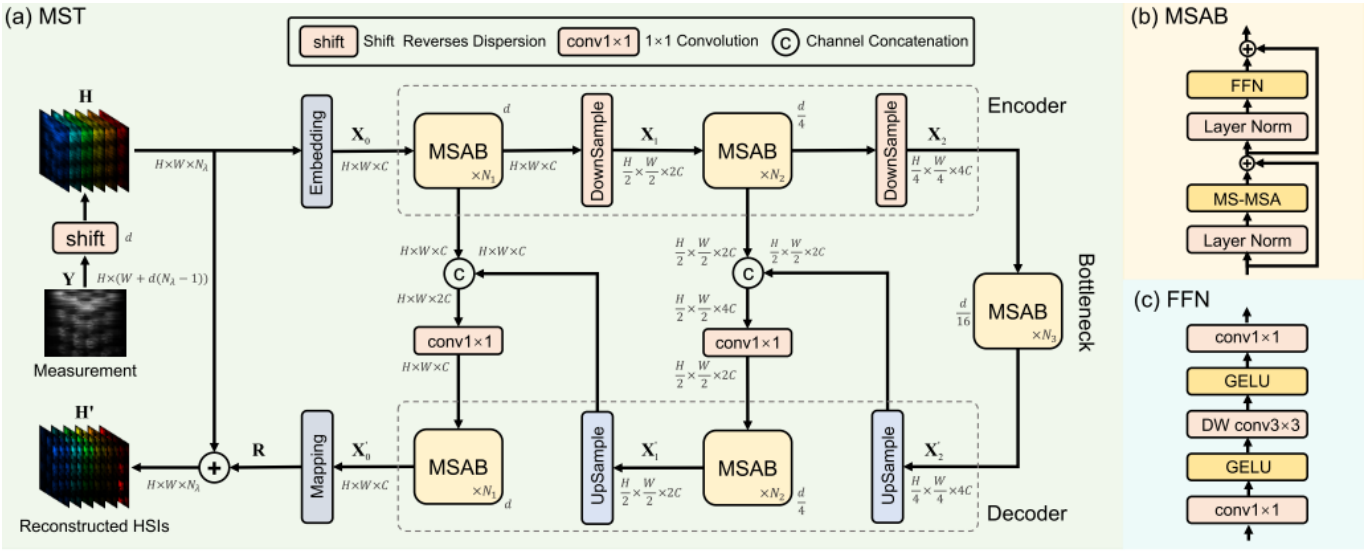

图 2. MST 整体架构图。(a)MST 网络架构；(b)MSAB 块结构；(c)FFN 块结构

3.2 光谱间多头自注意力

非局部相似性是 HSI 重建中常用的一种方法，但基于 CNN 神经网络的方法通常不能很好地挖掘出 HSI 的非局部相似性。由于 Transformer 在捕获非局部远距离依赖关系方面的有效性以及在其他视觉任务中带来的优越的性能，作者希望将 Transformer 引入到高光谱图像重建算法中，以探究 Tranformer 在 HSI 重建方面的能力。然而，当直接将 Transformer 应用于 HSI 重建时，存在两个主要问题。第一个问题是，原始的 Transformers 在空间维度上对远距离依赖性进行建模。但 HSI 图像在空间上是稀疏的并且在光谱上是相关的，如图 1(a) 所示。因此直接应用 Transformer 捕捉空间维度上的远距离依赖性是性价比较低的策略。因此，作者提出光谱间多头自注意力机制，将每个光谱特征图作为一个令牌，并沿着其光谱维度进行自注意力运算。以图 2(a) 中 MST 整体框架中的第一个 MSAB 块为例，图 1(c1) 展示了其 S-MSA 的运算流程。首先，将输入的特征图 $X_{in} \in R^{H\times W\times C}$ 变换为一个令牌向量矩阵 $X \in R^{HW\times C}$，然后按照 Transformer 的想法将 X 投影成 query 矩阵 $Q \in R^{HW\times C}$，key 矩阵 $K \in R^{HW\times C}$ 和 value 矩阵 $V \in R^{HW\times C}$，具体的投影流程如下：

$$Q = XW^Q, K = XW^K, V = XW^V$$

W^Q,W^K，$W^V \in R^{C\times C}$ 均是神经网络中的可学习变量。之后，按照多头注意力的思想将 Q，K，V 沿着光谱维度分为 N 个部分 $Q = [Q_1, \cdot, \cdot, \cdot, Q_n]$，$K = [K_1, \cdot, \cdot, \cdot, K_n]$，$V = [V_1, \cdot, \cdot, \cdot, V_n]$，每个头的通道数为 $d_h = \frac{C}{N}$，图 1(c1) 中展示的是当 $N = 1$ 时的情况。与一般的 MSA 不同的是，S-MSA 对 Q，K，V 中每个头的运算如下：

$$A_j = softmax(\sigma_j K_j^T Q_j), head_j = V_j A_j$$

由于光谱强度随波长变化很大，因此 S-MSA 使用一个可学习的参数 σ_j 来调整 A_j 的结果。随后，将 N 个头的输出在光谱维度上连接以进行线性投影，然后添加一个位置嵌入特征：

$$S - MSA(X) = \text{Concat}(head_j)W + f_p(V)$$

$W \in R^{C\times C}$ 是一个可学习参数，$f_p(\cdot)$ 是一个用于产生位置嵌入特征图的函数。在具体的实现中，$f_p(\cdot)$ 使用一个卷积块实现，具体包含两个通道分离的 3×3 的卷积层，一个 GeLU 非线性激活函数和一个变形操作。HSI 沿着光谱维度按波长排列。因此，S-MSA 利用这种特点进行提取位置嵌入特征以编码不同光谱通道的位置信息。将最终的结果变形成与输入特征图同样的形状。因为 S-MSA 将整个光谱特征图视为令牌，所以我们的 S-MSA 的感受野是全局的，并不限于特定位置的窗口。

3.3 掩码片指导机制

直接使用 Transformer 进行 HSI 恢复的第二个问题是，原始 Transformer 可能会处理一些具有低保真 HSI 表示的信息量较少的空间区域。在 CASSI 系统中，使用物理掩码片来对入射光进行调制。并且，物理掩码片上不同位置的透光率不同。因此，光谱信息的调制过程是位置敏感的。这一操作可以说明物理掩码片应该被用作一个 HSI 重建的线索，来指导重建算法着重关注高保真度的 HSI 的区域。先前的基于 CNN 的方法使用掩码片作为指导线索主要在初始化的 HSI 图像 S 和物理掩码片 M^* 之间进行内积，以生成调制输入。该方案引入了

空间保真度信息，但存在以下局限性：(1) 该操作破坏了输入的 HSI，破坏了其空间上的连续性，导致信息丢失；(2) 该方案仅在输入端起作用。掩码片在引导网络关注具有高保真 HSI 表示的区域方面的指导作用尚未得到充分探索。(3) 该方案不利用可学习的参数来对空间相关性进行建模。与以前的方法不同，作者提出的掩码片指导机制 (MM) 保留了所有的输入 HSI 表示，并学习指导 S-MSA 关注高保真光谱表示的空间区域。具体来说，当给定图 1(c1) 所示的掩码片 $M^* \in R^{H\times W}$，因为经过掩码片调制的 HSI 被 CASSI 系统中的色散器剪切移位，MM 首先像色散器的调制过程一样调制 M^*，调制结果记为 $M_s \in R^{H\times(W+d(N_\lambda-1))\times N_\lambda}$。在 M^* 上位移到 y 轴范围之外的区域设置为 0。为了匹配 MST 的阶段 i 中的特征图 X_i 的形状，M_s 需要经过图 2(a) 中相同的降采样操作。接下来，M_s 先经过一个卷积核大小为 1×1 的卷积层，然后将结果作为输入进入两条独立路径。如图 1(c2) 中，其中一条路径对输入不做任何处理，该路径用来保证保真度信息不会丢失；进入另一条路径的输入数据需要先经过一个 1×1 的卷积层，一个通道分离的 5×5 的卷积层，和一个 $sigmoid$ 非线性激活函数，最后与第一条路径的输出做内积。S-MSA 在捕获光谱间依赖性方面是有效的，但在挖掘 HSI 的空间相互作用方面则显示出了局限性。因此，第二条路径被设计为捕获空间方面的相关性。具体来说，上述操作的运算公式为

$$M_s' = (W_1 M_s) \odot (1 + \delta(f_{dw}(W_2 W_1 M_s)))$$

W_1 和 W_2 分别代表两个 1×1 的卷积层的可学习参数，$f_{dw}(\cdot)$ 代表通道分离 5×5 卷积层的映射函数，$\delta(\cdot)$ 代表 $sigmoid$ 非线性激活函数，M_s' 代表上述操作得到的结果。为了与 MST 的输入图像维度一致，要对 M_s' 进行分散器调制的逆调制过程，最终得到与输入的初始信号 S 维度一致的掩码片特征 M'。接下来，将其应用到多头自注意力机制中。首先，先将 M' 变形为 $M \in R^{HW\times C}$ 来匹配矩阵 V 的维度。之后在光谱维度上将 M 分离成 N 个头，即 $M = [M_1, \cdot, \cdot, \cdot, M_n]$，对于每个头 $head_j$，MM 将 M_j 用来调整 V_j 的权重。因此，当使用 MM 来指导 S-MSA 时，S-MSA 模块仅需要通过对 $head_j$ 的计算公式进行简单的调整即可，即

$$head_j = (M_j \odot V_j)A_j$$

S-MSA 的后续步骤保持不变。通过使用 MM，S-MSA 可以提取 HSI 特征而不破坏其结构，接受位置敏感保真度信息的指导，并自适应地对空间相关性进行建模。

4 复现细节

4.1 与已有开源代码对比

本次复现共包含三个部分。第一部分是本文方法 MST 的复现。在该部分的复现工作中，我通过仔细阅读文章中网络架构和数据预处理部分的内容，完成了代码的设计与实现。第二部分是对用于对照试验的算法的复现，本次对照试验共设有 3 个对照组，分别是 ADMM-Net，TSA-Net 和 lambda-Net，这些算法的复现则使用现有的开源代码与其对应的神经网络权重。对照组的代码是由本文作者发布于 https://github.com/caiyuanhao1998/MST/。第三部分是对评价指标的计算，本次复现主要以两个评价指标为参考，分别是峰值信噪比和结构相似性。

4.2 实验环境搭建

本次算法的实现是基于 PyTorch 框架的，由于文章内对网络架构介绍的十分详细，因此针对网络架构的复现结果较为明确和统一。实验使用时 CPU 是 Intel Core i7-12700H，GPU 是 Tesla v100s。实验时使用的训练与测试代码均为独立编写。训练中的超参数、优化函数均使用文章中提到的信息。神经网络的训练使用 Adam 优化器，其相关的超参数设置为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$，训练轮次设定为 300 轮。学习率使用动态更新策略，学习率在开始时设置为 4×10^{-4}，在训练过程中使用余弦退火调度器。当使用高光谱图像数据进行实验时，将从高光谱 3D 立方体中随机裁剪出的空间维度大小为 256×256 的图像块作为网络的输入，因为实验时使用的物理掩码片数据是 TSA-Net 提供的数据，该数据空间维度大小为 256×256。算法模拟的色散器对波段产生的位移步长 d 设定为 2。MST 的每个阶段都使用了 MM 机制，因此都需要对物理掩码片执行一个逆色散器调制的过程，但每个阶段的通道数不一致，因此逆调制的的位移步长也不一致，每一阶段的逆调制位移步长设定为 $\frac{d}{4^i}$ $(i = 0\ 1\ 2)$。训练的目标是最小化重建结果和真实 HSI 之间的均方根误差（RMSE）。均方根误差（RMSE）是在均方误差（MSE）基础上发展而来的一种衡量预测值与真实值差异的指标。其计算公式为

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

其中，$\hat{y}_i$ 是重建图像中一个像素点对应的光谱值，y_i 则是同一位置的真实值，n 为高光谱图像的空间维度大小。

4.3 算法评价指标

算法的评价指标主要包含两个值。一个是峰值信噪比（PSNR），另一个是结构相似性(SSIM)。峰值信噪比（PSNR）是一种用于衡量图像或信号质量的指标。它主要用于评估经过处理（如压缩、传输等）后的图像或信号与原始图像或信号之间的差异程度。PSNR 的值越高，表示处理后的图像或信号与原始版本越接近，质量越好。其计算公式为

$$PSNR = 10 \times \log_{10}\left(\frac{MAX^2}{MSE}\right)$$

MAX 是图像信号中的最大值。结构相似性（SSIM）是一种用于衡量两幅图像之间相似性的指标。与传统的基于像素差异的指标（如均方误差 MSE 或峰值信噪比 PSNR）不同，SSIM 考虑了图像的结构信息，包括亮度、对比度和结构。亮度是指图像的平均灰度值，对比度与图像的标准差有关，结构则体现了图像内容的排列情况。SSIM 通过综合评估这些因素来衡量两幅图像的相似程度，其值范围在-1 到 1 之间，值越接近 1 表示两幅图像越相似。其相关公式为

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

$$SSIM(x,y) = l(x,y)c(x,y)s(x,y)$$

x 和 y 是两幅图像信号，μ_x 和 μ_y 代表两幅图像的均值，δ_x 和 δ_y 代表两幅图像的方差。$l(x,y)$ 是一个亮度比较函数，其中 $C_1 = (k_1 L)^2$ L 是像素值的动态范围，k_1 是一个很小的常数，$C(x,y)$ 是一个对比度比较函数，其中 $C_2 = (k_2 L)^2$ k_2 是一个很小的常数，通常取 0.03，$s(x,y)$ 是一个结构比较函数，其中 $C_3 = \frac{C_2}{2}$。

5 实验结果分析

MST 模型训练与测试使用的数据集是公开的高光谱图像数据集 CAVE 和 KAIST。CAVE 数据集共包含 205 个高光谱图像数据立方体，空间维度大小为 1024×1024，KAIST 数据集共包含 30 个高光谱图像数据立方体，空间维度大小为 2704×3376。为了一致性，参考 TSA-Net 的训练策略，训练集使用 CAVE 数据集，测试集则从 KAIST 数据集中随机选取 10 张图像。需要注意的是，数据集中图像的波段数均为 28 个波段。对照组分别为 ADMM-Net，TSA-Net 和 lambda-Net，分别使用这些神经网络算法对同一场景进行重建，这些方法对同一场景的结果如图 3所示。当我们仔细观察各方法的重建效果，发现其余方法要么产生过于平滑的结果，牺牲细粒度的结构内容和纹理细节，要么引入彩色伪影和斑点纹理。相比之下，作者的 MST 更能够产生主观上更优秀的清晰图像，并保持均匀区域的空间平滑度。这主要是因为 MST 加入了调制过程的指导，并且能够捕获不同光谱间的远距离依赖性。客观上，我使用 10 张 KAIST 数据集中的高光谱图像对各方法进行评价指标上的度量，最终结果如图 4和图 5所示。可以看出，文章提出的 MST 方法在 PSNR 和 SSIM 指标上拥有最优秀的结果。需要注意的是，根据网络架构中各阶段 MASB 块的数量 $[N_1, N_2, N_3]$，将网络可分为 S，M，L 三个尺寸，复现及后续实验均在 S 级尺寸的 MST 上进行。

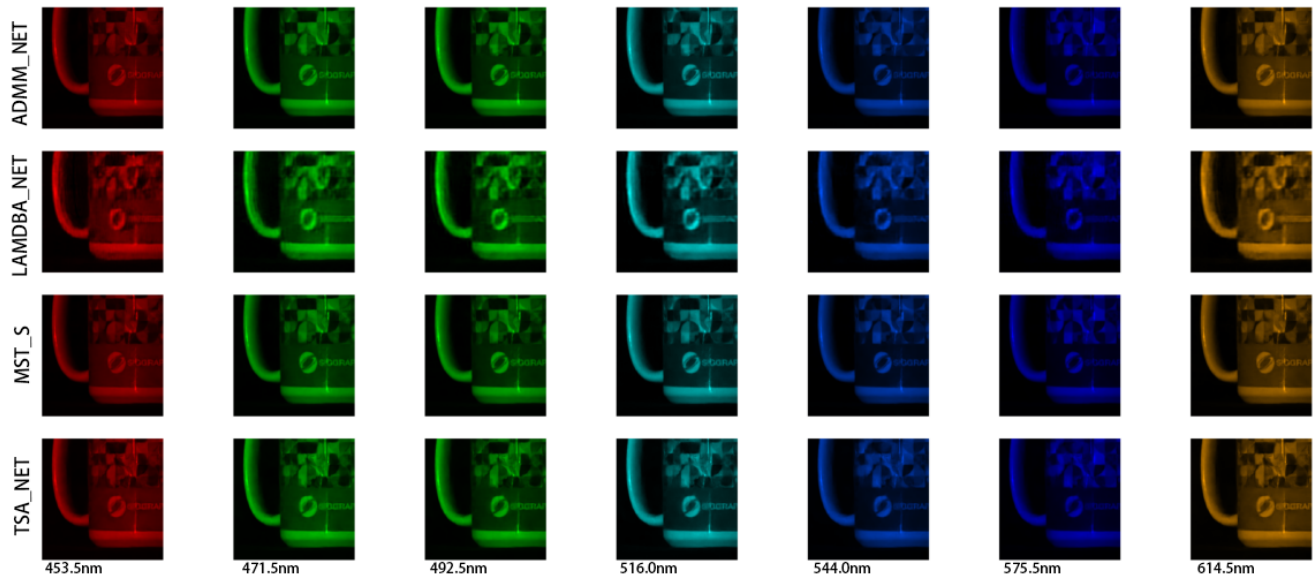

图 3. 各方法重建结果展示（28 个波段中选取了 7 个波段）

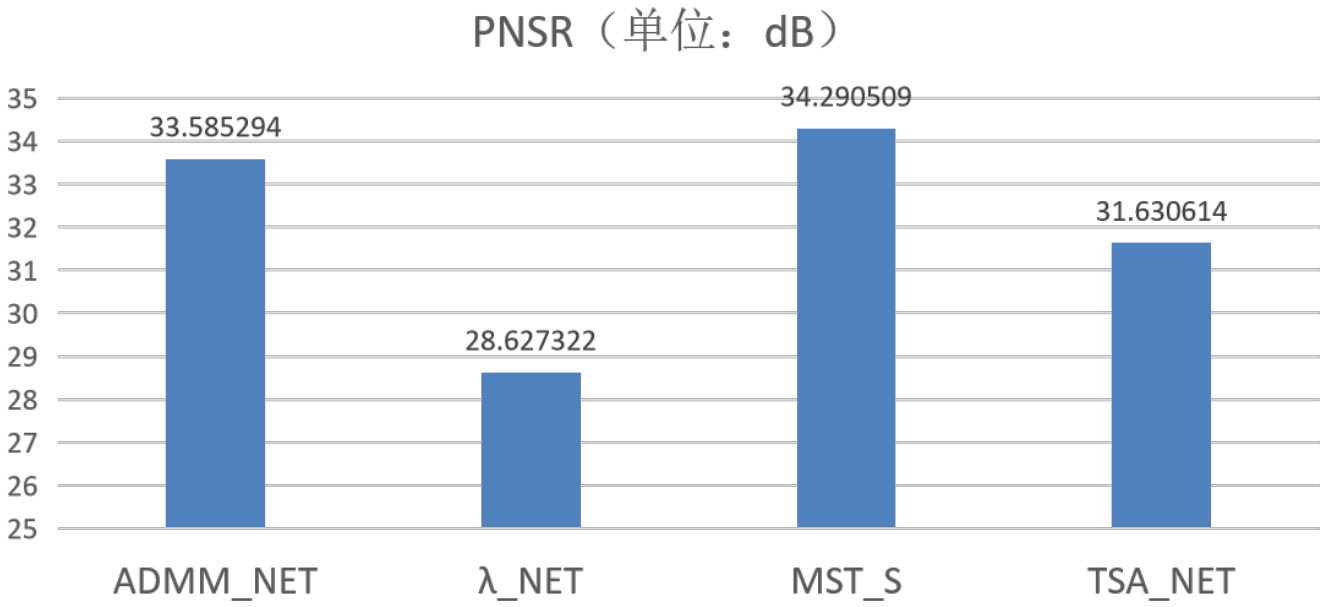

图 4. PSNR 结果比较

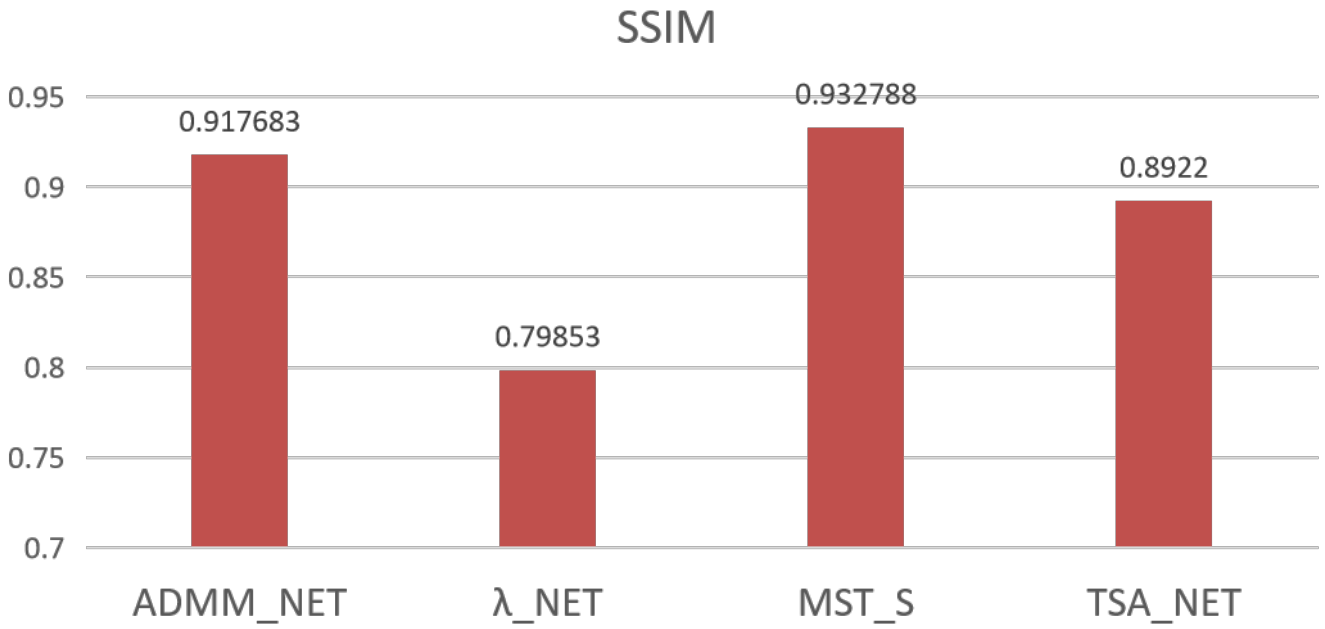

图 5. SSIM 结果比较

6 总结与展望

文章提出了一种全新的 Transformer 机制的使用方法 MST，并在 CAVE 和 KAIST 数据集上取得优异的成果。同时，将成像系统对高光谱图像的调制过程引入到了后续的重建算法中，这样一来重建算法可以更好挖掘图像中特征以重建出更优良的效果。但是，文章的成像系统是基于 CASSI 系统，CASSI 系统中包含有大量的光学器件以及复杂的光路设计，所以 CASSI 的成像系统通常具有较大的体积，不适用于许多要求便携性的场景。除 CASSI 外，高

光谱图像的压缩采集过程还可以使用彩色滤光片阵列来完成，这种方式极大程度的降低了成像系统的体积，并且还可以使用压缩感知理论对滤光片参数的设计进行指导以获得更好的重建效果。除此以外，文章提出的算法需要较高的计算开销，常规的计算单元可能无法负担起高昂的计算开销，但是高性能计算平台又会造成系统体积庞大、功耗过高等问题。因此，目前高光谱图像重建算法需要一个较为简单的算法，不需要超高的运算性能，真正做到芯片级的高光谱成像。

参考文献

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Herve Jegou. Xcit: Cross-covariance image transformers. *NeurIPS*, pages 20014–20027, 2021.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *Proc. Int. Conf. on Computer Vision*, pages 6816–6826, 2021.

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proc. Euro. Conf. on Computer Vision*, pages 205–218, 2023.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Proc. Euro. Conf. on Computer Vision*, page 213–229, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Euro. Conf. on Computer Vision*, pages 213–229, 2020.

[6] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. pages 2968–2977, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. Int. Conf. on Learning Representations*, pages 1–21, 2021.

[8] Ying Fu, Yinqiang Zheng, Imari Sato, and Yoichi Sato. Exploiting spectral-spatial correlation for coded hyperspectral image restoration. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 3727–3736, 2016.

[9] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral

compressive imaging. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 17521–17530, 2022.

[10] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 16211–16220, 2021.

[11] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J. Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Appl. Opt.*, pages 6824–6833, 2010.

[12] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 1944–1953, 2021.

[13] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Trans. on Graphics*, pages 233:1–233:11, 2014.

[14] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Trans. Pattern Analysis & Machine Intelligence*, pages 2990–3006, 2019.

[15] 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.

[16] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. *Proc. Euro. Conf. on Computer Vision*, pages 187–204, 2020.

[17] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. lambda-net: Reconstruct hyperspectral images from a snapshot measurement. *Proc. Int. Conf. on Computer Vision*, pages 4058–4068, 2019.

[18] Jin Tan, Yanting Ma, Hoover Rueda, Dror Baron, and Gonzalo R. Arce. Compressive hyperspectral imaging via approximate message passing. *IEEE Journal of Selected Topics in Signal Processing*, pages 389–401, 2016.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. Conf. on Neural Information Processing Systems*, pages 6000—6010, 2017.

[20] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.*, pages B44–B51, 2008.

[21] Lizhi Wang, Chen Sun, Ying Fu, Min H. Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 8024–8033, 2019.

[22] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE Trans. Pattern Analysis & Machine Intelligence*, pages 2104–2111, 2017.

[23] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *ICIP*, pages 2539–2543, 2016.

[24] Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, and Hua Huang. In *Computational Hyperspectral Imaging Based on Dimension-Discriminative Low-Rank Tensor Recovery*, pages 10182–10191, 2019.

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proc. Int. Conf. on Learning Representations*, pages 1–16, 2021.