MST++：用于高效光谱重建的多级光谱 Transformer

陈运来

摘要

高光谱图像 (Hyperspectral image HSI) 记录了窄带的真实场景光谱，其中每个波段捕获特定光谱波长的信息。与正常的 RGB 图像相比，HSI 具有更多的光谱波段，可以存储更丰富的信息，描绘出更多捕获场景的细节。由于这一优势，HSI 具有广泛的应用，如医学图像处理、遥感、目标跟踪，缺点在于: 高光谱相机通常比较昂贵，成像质量的不确定性大。鉴于当前 RGB 相机的普及程度，本次工作的目标旨在设计一个能从 RGB 有效重建出高光谱图像的模型

关键词：transformer；光谱；重建

1 引言

高光谱图像指的是光谱分辨率在 10^{-2} 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。在很多时候，从 RGB 图像中无法观测出异常，但是从高光谱图像的某一个波段中却能一眼看出问题所在。这么说可能不太好理解，举个例子，比如在深夜，如果直接看 RGB 图像的话，可能是一片漆黑，但是如果通过红外夜视仪的话，就能很清晰看到发热的活物。这个红外夜视仪捕获的就是红外光谱图像。也正因为光谱图像有着这样的特性，它被广泛地应用于目标检测与追踪，图像识别，遥感，医疗影像等领域。

传统的高光谱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，快照压缩成像 (Snapshot Compressive Imaging，SCI) 系统来解决这一问题。在诸多 SCI 系统当中，编码孔径快照光谱成像（Coded Aperture Snapshot Spectral Imaging）CASSI 系统脱颖而出，成为捕获获取光谱图像的重要手段。但是 CASSI 的设备很贵，价格大约几万美元。

本质上 RGB 和 HSI 都是同一场景的不同光谱通道成像，既然深度学习模型如 CNN，Transformer 在计算机视觉展现出强大的潜力，而且伴随着 RGB 相机的普及,RGB 图像遍地都是，我们可以尝试学习一个从 RGB 到 HSI 的映射。

2 相关工作

硬件成像设备过于昂贵，带动了光谱重建 (Spectral Reconstruction SR) 算法的发展，也就是通过算法由几个通道重建出几十个甚至更多的通道数，鉴于 RGB 图片比较容易获取，所以目前的重建算法大部分都是用 RGB 图片来重建。而重建算法的发展也经历了三个阶段:

2.1 传统方法

传统的 SR 方法[1][2]主要是主要是基于手工制作的高光谱先验，适用性比较差。

2.2 浅层学习

Aeschbacher 等人[3][4][5]建议使用相对较浅的学习模型，从一个特定的光谱先验来实现光谱重建。然而，这些基于模型的方法存在表征能力有限和泛化能力差的问题。

2.3 深度学习

最近，受到深度学习在自然图像修复中取得巨大成功的启发[6],[7]。CNN 已经被用来学习从 RGB 到 HSI 的基本映射函数. 然而，这些基于 CNN 的重建方法取得了出色的结果，但在捕捉非局部的自相似性和长距离的相互依赖性方面显示出局限性。

3 本文方法

3.1 网络结构

此部分对本文将要复现的工作进行概述，如图 2 所示，（a）描述了所提出的多级频谱 Transformer（MST++），由 N_s 个单级频谱 Transformer（SST）级联。MST++ 将 RGB 图像作为输入，并重建其对应的 HSI 图像，利用长恒等 (long identity) 映射来简化训练过程。图 2（b）显示了由编码器、瓶颈层（bottleneck) 和解码器组成的 U 形 SST。嵌入和映射块为单个 conv3×3 层。编码器中的特征图依次经历下采样操作（跳步 4×4 卷积层）、N_1 个基于光谱的注意力模块（SAB）、下采样操作和 N_2 个 SAB。瓶颈由 N_3 个 SAB 组成。解码器采用对称结构。上采样操作是一个跳步的 2×2 反卷积层，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 2（c）说明了 SAB 的组成部分，即前馈网络（如图 2（d）所示的 FFN）、光谱多头自注意（S-MSA）和两层正则化。S-MSA 的详细信息如图 2（e）所示。

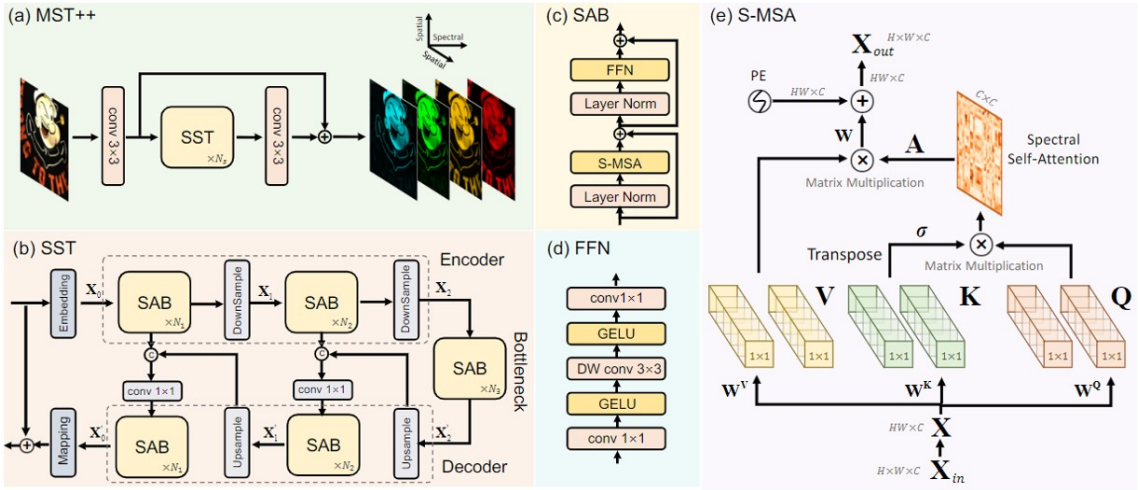

图 1: 方法示意图

3.2 基于光谱的多头自注意力 (S-MSA)

假设 $\mathbf{X}_{in} \in \mathbb{R}^{H\times W\times C}$ 作为 S-MSA 的输入, 将其形状重塑 (reshape) 为 tokens $\mathbf{X} \in \mathbb{R}^{HW\times C}$. 随后 $\mathbf{X}$ 被线性映射为 *query* $\mathbf{Q} \in \mathbb{R}^{HW\times C}$, *key* $\mathbf{K} \in \mathbb{R}^{HW\times C}$, and *value* $\mathbf{V} \in \mathbb{R}^{HW\times C}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}, \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}, \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}, \tag{1}$$

其中 $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, 和 $\mathbf{W}^{\mathbf{V}} \in \mathbb{R}^{C\times C}$ 可学习参数; 简单起见，忽略偏置项 $biases$. 随后, 沿着光谱维分解 $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ 成为 N 个 *heads* : $\mathbf{Q} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_N]$, $\mathbf{K} = [\mathbf{K}_1, \ldots, \mathbf{K}_N]$, and $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_N]$。每个 head 的维度是 $d_h = \dfrac{C}{N}$，与原始 MSA 不同，我们的 S-MSA 将每个频谱表征视为一个 token，并计算 $head_j$:

$$\mathbf{A}_j = \text{softmax}(\sigma_j \mathbf{K}_j^{\mathrm{T}} \mathbf{Q}_j), \quad head_j = \mathbf{V}_j \mathbf{A}_j, \tag{2}$$

其中 $\mathbf{K}_j^{\mathrm{T}}$ 代表 $\mathbf{K}_j$ 的转置矩阵. 因为光谱密度相对于波长变化很大, 我们使用可学习的参数 $\sigma_j \in \mathbb{R}^1$

通过重新加权 $head_j$ 里的矩阵乘法 $\mathbf{K}_j^{\mathrm{T}}\mathbf{Q}_j$ 来适应自注意力结果 $\mathbf{A}_j$. 随后，将 N 个 *heads* 的输出连接起来进行线性投影，然后添加位置嵌入：

$$\text{S-MSA}(\mathbf{X}) = \big(\underset{j=1}{\overset{N}{\text{Concat}}}(head_j)\big)\mathbf{W} + f_p(\mathbf{V}), \tag{3}$$

其中 $\mathbf{W} \in \mathbb{R}^{C\times C}$ 是可学习参数, $f_p(\cdot)$ 是生成位置编码的函数. 由两个深度可分离卷积 *conv*3×3 , 一个 GELU 激活函数, 和重塑操作，HSI 按光谱维度的波长排序。因此，我们利用这种嵌入来编码不同光谱通道的位置信息。最后，我们重塑方程 Eq. (3) 的结果，得到输出的特征图 $\mathbf{X}_{out} \in \mathbb{R}^{H\times W\times C}$。

4 复现细节

4.1 尝试改进

4.1.1 Piexl attention

注意力机制在提高计算机视觉任务深度模型的性能方面表现出极大的优势, Roy 等人[8]使用了 1×1 卷积层以生成空间注意力特征,CBAM[9]使用核大小的 2D 卷积层计算空间注意力. 显然，上述大多数方法都专注于开发复杂的注意力模块以获得更好的性能，此处，此处的 PA 旨在以较低的计算复杂度学习有效的像素注意力。

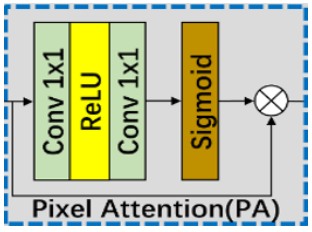

图 2: Pixel attention

4.1.2 3D convolution

鉴于高光谱图像具有谱间相似性的特征，可以考虑使用三维卷积，考虑信道间的相互关系来优化光谱数据的提取。

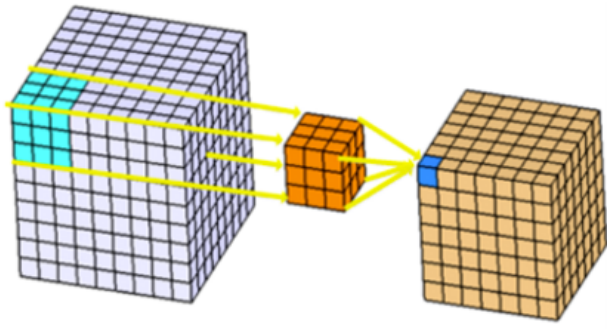

图 3: 3D 卷积

4.1.3 Sub-Pixel Convolution

亚像素卷积最先在超分辨率网络上提出[10], 主要功能是将低分辨的特征图，通过多通道间的重组得到高分辨率的特征图 (相当于上采样), 该操作不需要参数，可以应用到轻量化网络中，部署到移动

设备上。

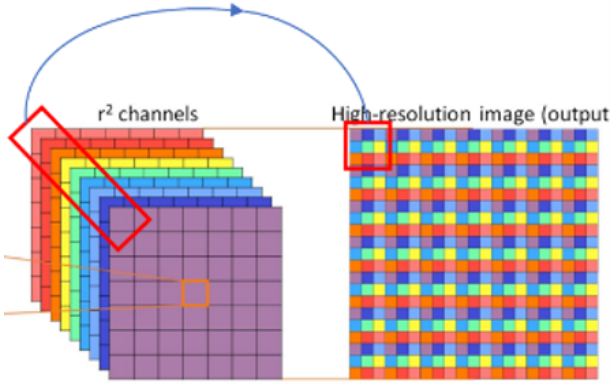

图 4: 亚像素卷积

5 实验结果分析

加入 PA，在测试集上误差 (MRAE) 降低约 1.9%，将下采样替换为 Pixelshuffle，误差虽稍微上升，但是减少了参数量，加入 3d-convolution，在测试集上误差 (MRAE) 降低约 1.2%，三者融合，最终误差降低 2.6% 可以判断三者存在一定的互补性。

6 总结与展望

在本次课程，我们对基于 Transformer 的框架 MST++ 尝试改进，用于从 RGB 进行光谱重建。基于 HSI 的空间稀疏性和光谱自相似性，采用 S-MSA 将每个光谱特征图作为一个自注意计算的令牌来组成基本单元 SAB(Self-attention block)。最后 MST++ 采用多阶段学习方案，从粗到细逐步提高重建质量，同时需要更低廉的内存和计算成本。本人尝试的改进主要有三个：一是加入了一个简单有效的 pixel-attention 机制，在模型中加入 Pixel-shuffle，以及采用三维卷积提取特征，每个模块输入输出形状相同，所以模块添加的位置也有很多种方案，经过各种尝试，最终误差降低 2.6%。将来，可以对基于通道的自注意力进一步完善，也可以针对 attention 结果的可解释性做进一步探讨。

参考文献

[1] ARAD B, BEN-SHAHAR O. Sparse Recovery of Hyperspectral Signal from Natural RGB Images[J]. Springer International Publishing eBooks, 2016.

[2] PARMAR M, LANSEL S, WANDELL B A. Spatio-spectral reconstruction of the multispectral datacube using sparse recovery[J]. International Conference on Image Processing, 2008.

[3] WU J, AESCHBACHER J, TIMOFTE R. In Defense of Shallow Learned Spectral Reconstruction from RGB Images[J]. International Conference on Computer Vision, 2017.

[4] YUAN X. Generalized Alternating Projection Based Total Variation Minimization for Compressive Sensing[J]. arXiv: Information Theory, 2015.

[5] YANG J, YUAN X, LIAO X, et al. Video compressive sensing using Gaussian mixture models.[J]. IEEE transactions on image processing, 2014.

[6] SHI Z, CHEN C, XIONG Z, et al. HSCNN+: Advanced CNN-Based Hyperspectral Recovery from RGB

Images[J]. Computer Vision and Pattern Recognition, 2018.

[7] LI J, WU C, SONG R, et al. Adaptive Weighted Attention Network With Camera Spectral Sensitivity Prior for Spectral Reconstruction From RGB Images[J]. Cornell University - arXiv, 2020.

[8] ROY A G, NAVAB N, WACHINGER C. Concurrent Spatial and Channel ‘Squeeze & Excitation’ in Fully Convolutional Networks[J]. Lecture Notes in Computer Science, 2018.

[9] ZHAO H, KONG X, HE J, et al. Efficient Image Super-Resolution Using Pixel Attention[J]. Springer International Publishing eBooks, 2020.

[10] PENG H, CHEN X, ZHAO J. Residual Pixel Attention Network for Spectral Reconstruction from RGB Images[J]. Computer Vision and Pattern Recognition, 2020.