Bidirectional Dilation Transformer for Multispectral and Hyperspectral Image Fusion

作者:Shangqi Deng ,Liang-Jian Deng ,Xiao Wu, Ran Ran, Rui Wen

复现者：Pan Liu

摘要

从当前的各项研究来看，基于 Transformer 的方法可以有效地实现远距离建模、捕捉空间和光谱信息，并在各种计算机视觉任务中表现出很强的归纳偏差。Transformer 模型一般包括两种常见的多头自注意（MSA）模式：空间多头自注意力（Spa-MSA）和光谱多头自注意力（Spe-MSA）。但是，Spa-MSA 计算效率高，但却将全局空间响应限制在了局部窗口内。Spe-MSA 可以计算通道自注意力，以适应高分辨率图像，但它忽略了对低级视觉任务至关重要的局部信息。在这个研究中，作者提出了一种用于多光谱和高光谱图像融合（MHIF）的 Bidirectional Transformer（BDT），旨在利用 MSA 的优势和 MHIF 任务特有的潜在多尺度信息。BDT 由两个设计模块组成：扩张空间多头自注意力（D-Spa）和分组光谱多头自注意力（G-Spe），前者通过给定的空心策略动态扩展空间感受野，后者提取特征图中的潜在特征并学习局部数据特征。此外，为了充分利用来自不同空间分辨率输入的多尺度信息，作者在 BDT 中采用了双向分层策略，从而提高了模型性能。

关键词： Transformer；空间多头自注意力；光谱多头自注意力

1 引言

高光谱成像（HSI）是一项广泛应用于包括农业 [1,2]、食品安全 [3]、生物医学诊 [4] 断和大气环境探测 [5] 在内的各个领域的技术。具有高光谱分辨率的恒星成像仪能产生精确的光谱特性曲线，而且波段丰富，便于进行相互波段校正。然而，由于目前物理成像技术的限制，自然成像过程中空间分辨率和光谱分辨率之间存在权衡。因此，不可能同时生成高空间分辨率和高光谱分辨率的图像。因此，多光谱和高光谱图像融合（MHIF）已成为生成必要的高分辨率高光谱图像（HR-HSI）的一种有前途的方法。针对 MHIF 开发了许多方法，大致可分为两类：传统方法 [5–7] 和基于深度学习 (DL) 的技术 [8–10]。

近年来，基于深度学习（DL）的技术越来越受欢迎，其中 CNN 模块因其空间无关性和特定信道卷积特性而成为目前 MHIF 问题的最先进技术 [11]。研究人员设计了特定的卷积模块，并将它们堆叠在一起以构建一个通用网络结构，从而有效地从数据库中提取潜在行为。然而，CNN 的局部感受野限制了长程依赖性，可能会妨碍图像的内部建模。Vision Transformer（ViT） [12] 在各种计算机视觉任务中表现出令人印象深刻的性能。ViT 基于一种自我关注机制，通过研究标记之间的联系来有效捕捉全局交互。为了将 Transformer 应用于视觉任务，出

现了许多解决方案，如基于空间窗口的 MSA、Spe-MSA 、线性复杂度自注意等等。基于空间窗口的 MSA 设置了一个合适的窗口大小，并将图像的空间大小划分为若干个斑块。为简明起见，这种方法也被称为 Spa-MSA。

考虑到 MHIF 任务的特性，作者提出了一种融合架构，它整合了空间和光谱信息，并充分利用 MSA 对高光谱图像中的相似斑块进行建模。Spa-MSA 缺乏对长距离信息的建模，而 Spe-MSA 则没有充分利用数据内部的信息。作者提出的架构包括扩张 Spa-MSA 和分组 Spe-MSA 模块，实现了更广泛的相关性，具有一定的复现意义。

然而从原文提供的代码中可以发现，在空间信息的提取过程中只是单一的对图像进行了一个窗口的扩张自注意力操作，使得在光谱维度的感受野不足。针对这一问题，本次课程的的论文复现工作在保持原论文网络结构主干不变的条件下，提升了模型在光谱维度的感受野。

2 相关工作

2.1 Transformer 在 MHIF 中的应用

Transformer 架构在各种视觉任务中表现出了强大的性能，许多研究人员正试图利用它来解决 MHIF 问题并取得了不错的成果。例如，Hu 等人 [13] 率先将 Transformer 用于 MHIF，并利用轻量级网络实现了强大的性能。Ma 等人 [14] 利用 Transformer 代替 CNN 学习高光谱图像（HSI）的先验值，然后使用展开网络模拟 HSI 超分辨率的迭代求解过程。此外 Zhou 等人 [15] 提出了一种定制的 Transformer，可促进两种模态的协作特征学习，用于遥感泛锐化。

2.2 动机

尽管上述方法在很大程度上依赖于强大的自注意模块，取得了比较好的成果，但这些方法往往采用自注意或 Transformer 结构来完成各种图像融合任务，而没有充分考虑其不足之处，尤其是对于特定的 MHIF 问题。作者为 Spa-MSA 设计了一种新的二维扩张结构，称为 D-Spa。D-Spa 可以在不引入额外参数或计算复杂度的情况下有效扩大感受野。提出了分组 G-Spe，将空间分组，然后在小组中执行 Spe-MSA，这样可以提取特征内的信息，更好地学习局部数据行为。此外针对 MHIF 的具体应用，作者还设计了一种双向层次结构，以更好地利用具有不同空间分辨率的两个输入的多尺度信息。

3 本文方法

3.1 整体架构

BDT [16] 结构如图 1所示，它是一个分层双向输入结构，包括两个阶段，即双模特征提取（BFE）和双模特征融合（BFF）。为了提取空间信息，将双三次插值的 LR-HSI $\mathcal{X}^U \in \mathbb{R}^{H\times W\times S}$ 和 HR-MSI $\mathcal{Y} \in \mathbb{R}^{H\times W\times s}$ 拼接起来，作为空间分支的输入。此外，BFE 中的 D-Spa 是为学习空间信息而设计的，其输出特征图为 $\mathcal{D}_i$,i=1,2,3。具体来说，BFE 的过程如下：

$$\mathcal{D}_i = \text{SpatialBranch}\left(\text{Conv}_1\left(\text{Cat}\left(\mathcal{Y}, \mathcal{X}^{\text{U}}\right)\right)\right), \quad (1)$$

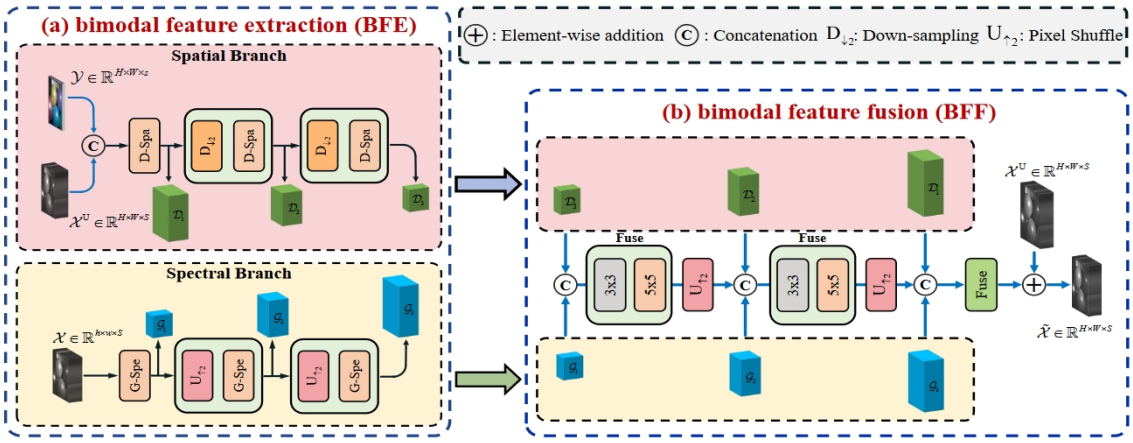

图 1. 整体架构图

Conv_1 是卷积结构。以 HR-HSI $\mathcal{X} \in \mathbb{R}^{h \times w \times S}$ 作为频谱分支的输入，通过 G-Spe 动态学习频谱信息，输出特征图 $\}_i$（i= 1、2、3），如下式所示：

$$\mathcal{G}_i = \mathrm{SpectralBranch}\left(\mathrm{Conv}_2\left(\mathcal{X}\right)\right), \qquad (2)$$

Conv_2 是用于增加通道的多层卷积结构。为了融合特征图，即 $\mathcal{D}_i$ 和 $\}_i$，作者设计了 BFF 模型，这是一种高效的两层卷积结构。先将 $\mathcal{D}_3$ 和 $\}_1$ 合并，然后将合并后的 $\mathcal{F}_1$ 送入融合模块，该模块涉及一个 3×3 内核和一个 5×5 内核，然后通过 PixelShuffle 进行上采样，如下式所示：

$$\mathcal{F}_1 = \mathrm{PixelShuffle}\left(\mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{D}_3, \mathcal{G}_1\right)\right)\right). \quad (3)$$

然后将 F_1、D_2 和 g_2 合并，并对合并结果进行升采样。按以下公式对上采样结果进行融合：

$$\mathcal{F}_2 = \mathrm{PixelShuffle}\left(\mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{F}_1, \mathcal{D}_2, \mathcal{G}_2\right)\right)\right). \quad (4)$$

最后，将 F_2、D_1 和 g_3 的融合结果加入到双三次插值的 LR-HSI $\mathcal{X}^U$ 中，最终输出 $\overline{\tilde{\mathcal{X}}} \in \mathbb{R}^{H \times W \times S}$ 用下式表示：

$$\mathcal{F}_2 = \mathrm{PixelShuffle}\left(\mathrm{Fuse}\left(\mathrm{Cat}\left(\mathcal{F}_1, \mathcal{D}_2, \mathcal{G}_3\right)\right)\right). \quad (4)$$

3.2 D-Spa

作者对输入特征 $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ 进行三次 1×1 卷积，分别生成三个张量，即 $\mathbf{Q} \in \mathbb{R}^{C \times H \times W}$、$\mathbf{K} \in \mathbb{R}^{C \times H \times W}$ 和 $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$。仅以 D-Spa 中的一个头为例，给定窗口大小 k 和扩张率 d 为 2，D-Spa 运算在 (i, j) 像素位置的输出 $\mathbf{V}' \in \mathbb{R}^{C \times H \times W}$ 可以表示为包含 (i, j) 像素位置的本地窗口中相应值 $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ 的线性聚合。

$$\mathbf{V}'_{(:,i,j)} = \sum_{(x,y) \in \Omega(i,j)} \mathbf{W}_{(i,j \to x,y)} \mathbf{V}_{(:,x,y)}, \quad (6)$$

其中 $\mathbf{V}_{(:,x,y)} in \mathbb{R}^C$ 表示值图 $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ 中第 (x,y) 个像素位置的值，$\Omega\left(i,j\right)$ 表示包含 k×k 个像素位置的扩张窗口坐标集。

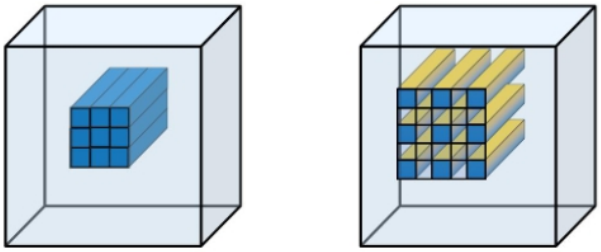

图 3. Spa-MSA D-spa

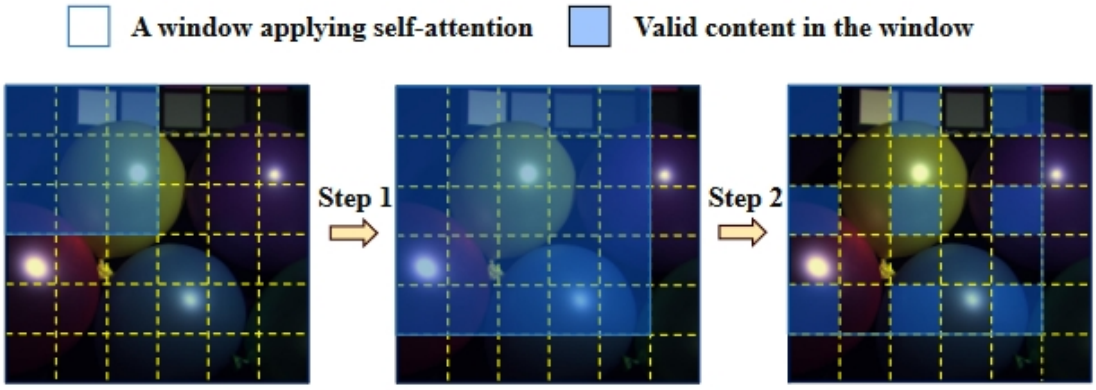

图 2. D-spa

在图 3 中，区域 $\Omega(i,j)$ 由两步产生，第一步是扩展原始窗口，第二步是禁止某些标记参与 Spa-MSA 计算。在公式 7 中，D 是一个常量变量；$\mathbf{V}' \in \mathbb{R}^C$ 表示输出特征图 $\mathbf{V}' \in \mathbb{R}^{C\times H\times W}$ 中第（i,j）个像素位置的向量；$\mathrm{W}_{(i,j\to x,y)} \in \mathbb{R}$ 表示注意力矩阵中的一个元素，该矩阵是根据查询 $\mathbb{Q}_{(i,j)} \in \mathbb{R}^C$ 与关键 $\mathbb{K}_{(x,y)} \in \mathbb{R}^C$ 之间点积的软最大归一化计算得出的：

$$\mathbf{W}_{(i,j\to x,y)} = \frac{e^{\frac{1}{\sqrt{D}}\mathbf{Q}^{\mathrm{T}}_{(i,j)}\mathbf{K}_{(x,y)}}}{\mathrm{S}_i}, \qquad (7)$$

其中

$$\mathrm{S}_i = \sum_{x=1,y=1}^{k,k} e^{\frac{1}{\sqrt{D}}\mathbf{Q}^{\mathrm{T}}_{(i,j)}\mathbf{K}_{(x,y)}}. \qquad (8)$$

图 2 显示了 SpaMSA 和 D-Spa 的特性。可以发现，D-Spa 可以像扩张卷积一样扩展感受野，并同时学习局部信息。此外，D-Spa 是预先固定的，没有滑动特性，并采用多头注意机制，先将通道分组，每组共享一个学习参数。

3.3 G-Spe

为了充分利用 HR-MSI 中的高分辨率空间信息和局部内容，我们将 G-Spe 设想为空间分组设计。具体来说，作者将值 $\mathbf{V} \in \mathbb{R}^{HW\times C}$、$\mathbf{Q} \in \mathbb{R}^{HW\times C}$ 和 $\mathbf{K} \in \mathbb{R}^{HW\times C}$ 细分为 g^2 组，在第 k 组中得到相应的 $\mathbf{V}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$、$\mathbf{Q}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$ 和 $\mathbf{K}^k \in \mathbb{R}^{\frac{HW}{g^2}\times C}$，其中 $k \in \left\{1,2,3,\cdots,\frac{HW}{g^2}\right\}$。然后，按如下方法独立计算第 k 组的权重矩阵 $\mathrm{W}^k \in \mathbb{R}^{C\times C}$：

$$\mathbf{W}^k_{(i,j)} = \frac{e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}^k_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}^k_{(:,j)}}}{\mathrm{S}^k_j}, \qquad (9)$$

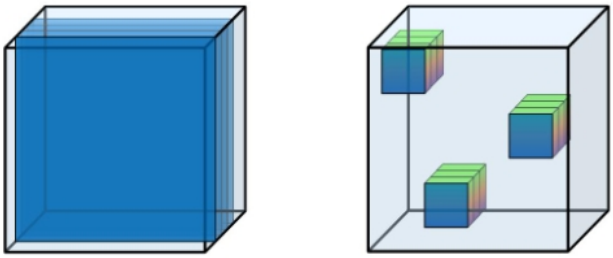

图 4. Spe-MSA G-spe

其中 S_j^k 按以下公式计算:

$$\mathrm{S}_j^k = \sum_{i=1}^{C} e^{\frac{1}{\sqrt{D}}\left(\mathbf{K}^k_{(:,i)}\right)^{\mathrm{T}}\mathbf{Q}^k_{(:,j)}}. \quad (10)$$

最后在 w^k 和 w^k 之间进行矩阵乘法运算，如下式所示:

$$\mathrm{V}^{k'} = \mathrm{V}^k\mathrm{W}^k. \quad (11)$$

在图 4 中，描述了 Spe-MSA 与 G-Spe 之间的关系。可以发现，Spe-MSA 利用整个空间的特征来获取权重，而 G-Spe 则利用部分空间信息来获取动态权重。由于 MHIF 任务的特性，局部丰富表示具有一定的优势，G-Spe 的效果优于 Spe-MSA。

3.4 损失函数定义

以端到端的统一方式优化网络参数。总体损失函数由两个损失的加权和组成:

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda_{ssim}\mathcal{L}_{ssim}, \quad (12)$$

其中 $\mathcal{L}_1$ 表示绝对差值之和，损耗 $\mathcal{L}_{ssim}$ 表示为:

$$\mathcal{L}_{ssim} = 1 - \mathrm{SSIM}(\bar{\mathcal{X}}, \tilde{\mathcal{X}}), \quad (13)$$

其中，SSIM 表示结构相似度，$\bar{\mathcal{X}}$ 表示参照，$\tilde{\mathcal{X}}$ 表示网络的输出，$\mathcal{L}_{ssim}$ 是固定为 0.1 的正超参数。

4 复现细节

4.1 与已有开源代码对比

作者在 github 上面开源了 BDT 的模型代码，在本次的复现过程中，首先将原作者的代码跑通，然后对模型的一个模块进行了一定的修改，在高光谱的空间信息读取模块 D-spa 上增加了通道分割的操作将输入的图片在光谱维度进行扩大，然后分成两块分别进行窗口注意力机制提取空间信息，再将得到的结果相加，扩大了对于光谱通道信息的感受野。在作者给出的 cave 和 harvard 数据集上面重新训练。然后由于在原论文之中作者只描述了在训练集和验证集上面进行模型的效果展示，本人编写了一个测试文件，用来在测试集上运行来观察模型效果。原文代码发表在 https://github.com/Dengshangqi/BDT。

4.2 实验环境搭建

使用 python 的 3.8.5 版本基于 pytorch 框架在 P100 显卡上面进行训练。训练时使用 AdamW 优化器更新模型参数，学习率设置为 0.0001，训练 batchsize 设置为 32，训练 2000 轮。在测试集上运行时 batchsize 设置为 1。

4.3 使用说明

将高光谱数据集放入 data 文件之中，并在 args_parser.py 文件中配置好对应的训练集、验证集和测试集的路径，运行 main.py 文件进行模型训练，训练过程中会在 checkpoints 文件加下面生成相应的最优模型参数，方便后面跑测试集的时候载入使用。运行 test.py 文件进行测试，会生成参考图像、融合结果图、结果差异图和模型光谱折线图在 data 文件夹下面。

4.4 创新点

原模型在 G-Spe 模块上面对空间信息进行了分组，实现了在空间局部信息的获取，能拥有更大的空间感受野。参考这一思想，在 D-Spa 模块对光谱维度进行加深分块分别进行窗口注意力使得在光谱维度的信息能获取地更多，增加了光谱信息的感受野。

5 实验结果分析

5.1 实验设置

本实验使用了 CAVE 和 Harvard 两个数据集，CAVE 数据集有 31 个光谱波段，训练集共 3920 张高光谱图片，将大小裁剪成 64×64 其中 %80 用来训练，%20 用于验证，测试集有 11 张图片，大小为 512×512 用于测试。Harvard 数据集有 31 个波段，训练集的设置同 CAVE 数据集，测试集有 10 张图片，大小为 1000×1000，对其进行 0 填充成大小为 1024×1024 进行测试。评价指标选用峰值信噪比 PSNR 和相对无尺寸全局误差 ERGAS。

5.2 实验结果

如下表 1 是原模型和本人进行修改过后的模型在 cave 数据集和 harvard 数据集上面的 PSNR 和 ERGAS 的值。

表 1. 指标

数据集	CAVE		Harvard	
指标	PSNR	ERGAS	PSNR	ERGAS
作者	49.3732	1.1141	48.0226	1.9824
本人	48.7461	1.1720	47.7723	2.0086

图 5 从左往右分别是一张来自于 CAVE 数据集和一张来自 Harvard 数据集的高光谱图像的模型输出图和参考图像还有输出图和参考图像的差异图像。

图 5. 融合差异图

图 6 从左往右分别是原模型和本人修改过后的模型取 CAVE 数据集和 Harvard 数据集上某张图片的第十个光谱波段的折线图。

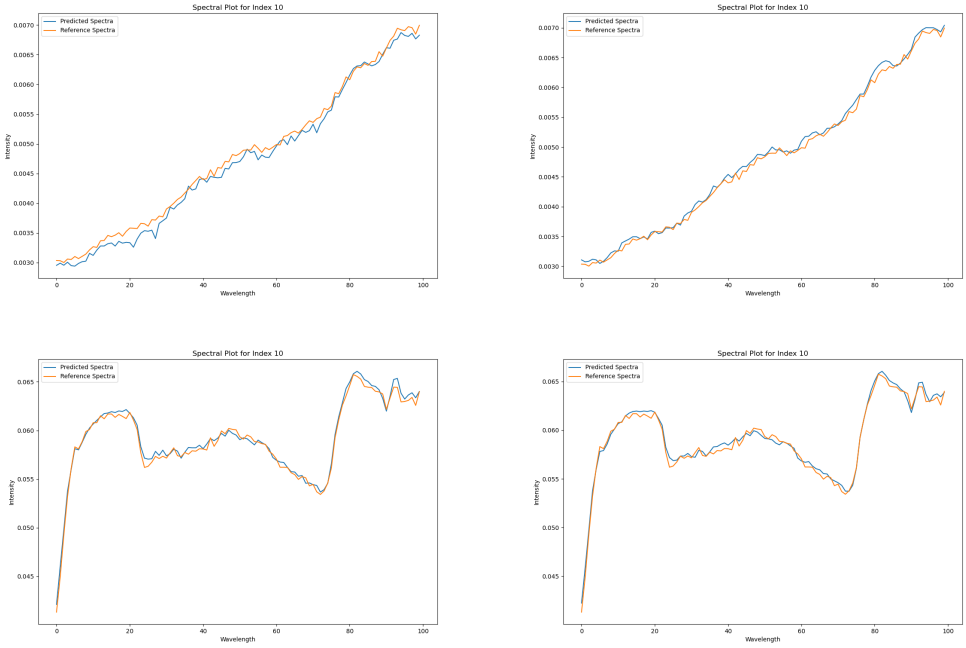

图 6. 光谱折线图

5.3 结果分析

通过上面的评价指标对比可以看出，我修改过后的模型会比原模型的输出结果差一点点，但是基本上还原出了原论文模型的融合效果，由光谱折线图可以看出，本人修改过后的模型因增加了光谱维度的感受野，因此在光谱维度上的损失相对于原论文模型会好一些。因此在同一数据集上面，不同的融合模型和预训练操作会对结果产生较大的差异。

6 总结与展望

BDT 模型在高光谱和多光谱融合空间超分辨率领域上已经有了很不错的效果，但是在光谱维度的感受野还可以进一步地进行优化，然后在不同的数据集上其实会有一定的差距，后续的研究可以在提升模型的泛化能力和光谱维度感受野上进行深入研究。

参考文献

[1] Jiangui Liu Yuhong He Bing Lu, Phuong D Dao and Jiali Shang. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens. (Basel)*, 12(16):2659, 2020.

[2] et al Jian Wu, Dao-Li Peng. Advances in researches on hyperspectral remote sensing forestry information-extracting technology. *Spectrosc. Spect. Anal.*, 31(9):2305–2312, 2011.

[3] Yaoze Feng and Dawen Sun. Application of hyperspectral imaging in food safety inspection and control: a review. *Crit. Rev. Food Sci. Nutr*, 52(11):1039–1058, 2012.

[4] R Tauler S Piqueras, L Duponchel and A De Juan. Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares. *Anal. Chim. Acta.*, 705(1-2):182–192, 2011.

[5] C Davis Bocai Gao and A Goetz. A review of atmospheric correction techniques for hyperspectral remote sensing of land surfaces and ocean color. *In IEEE International Symposium on Geoscience and Re499 mote Sensing*, pages 1979–1981, 2006.

[6] Peixian Zhuang Penghao Guo and Yecai Guo. Bayesian pan-sharpening with multiorder gradient based deep network constraints. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:950–962, 2020.

[7] Shuying Huang Weiguo Wan Wei Tu Yong Yang, Lei Wu and Hangyuan Lu. Multiband remote sensing image pansharpening based on dual injection model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:1888–1904, 2020.

[8] Shuying Huang Hangyuan Lu Yong Yang, Chenxu Wan and Weiguo Wan. Pansharpening based on low-rank fuzzy fusion and detail supplement. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13:5466–5479, 2020.

[9] Jie Huang Feng 606 Zhao Chengjun Xie Chongyi Li Keyu Yan, Man Zhou and Danfeng Hong. Panchromatic and multispectral image fusion via alternat608 ing reverse filtering network. *NeurIPS*, 2022.

[10] Keyu Yan Hu Yu 634 Xueyang Fu Aiping Liu Xian Wei Man Zhou, Jie Huang and Feng Zhao. Spatial-frequency domain information integration for pan- 636 sharpening. *In ECCV*, page 274–291, 2022.

[11] Zongben Xu 476 Xiangyong Cao, Jing Yao and Deyu Meng. Hyperspectral image classification with 477 convolutional neural network and active learning. *IEEE 478 Trans. Geosci. Remote Sens.*, 58(7):4604–4616, 2020.

[12] Changhu Wang Xiangtai Li Qi She Lei Zhu Tong Zhang Duo Li, Jie Hu and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. *In CVPR*, pages 12321–12330, 2021.

[13] Dirk Weissenborn Georg Heigold Jakob Uszkoreit Lucas Beyer Matthias Minderer Mostafa Dehghani Neil Houlsby Sylvain Gelly Thomas Unterthiner Alexander Kolesnikov, Alexey Dosovitskiy and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[14] Liangjian Deng Hongxia Dou Danfeng Hong Jinfan Hu, Tingzhu Huang and Gemine Vivone. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geosci. Remote 528 Sens. Lett.*, 19:1–5, 2022.

[15] Xianming Liu Qing Ma, Junjun Jiang and Jiayi Ma. Learning a 3d-cnn and transformer prior for hyperspectral image super-resolution. *IEEE Geosci. Remote Sens. Lett.*, arXiv preprint arXiv:2111.13923, 2021.

[16] Xiao Wu Shangqi Deng, Liang-Jian Deng, Ran Ran, and Rui Wen. Bidirectional dilation transformer for multispectral and hyperspectral image fusion. *the 32nd International Joint Conference on Artificial Intelligence*, 2023.