高光谱图像下游任务中的基础模型微调

摘要

基础模型（FMs）正在深刻改变遥感（RS）场景的分析与理解，涵盖了航空 RGB、多光谱和 SAR 影像等。然而，基于高光谱图像（HSIs）的基础模型应用仍相对稀少，现有方法多集中于特定任务，缺乏广泛适用性。为填补这一空白，诞生了一种高光谱图像解译基础模型 HyperSIGMA，支持扩展至超过十亿参数。该模型为应对高光谱图像中的光谱和空间冗余问题，设计了一种创新的稀疏采样注意力（SSA）机制，有效促进多样化上下文特征的学习，并成为 HyperSIGMA 的核心模块。HyperSIGMA 结合了专为高光谱图像设计的光谱增强模块，能够融合空间与光谱特征，并在大规模高光谱数据集 HyperGlobal-450K 上进行了预训练。大量实验结果表明，HyperSIGMA 在多种高层与低层高光谱任务中，展现出优异的通用性和表征能力，超越了当前最先进的方法。此外，HyperSIGMA 在可扩展性、鲁棒性、跨模态迁移能力以及实际应用方面均具显著优势。

关键词：遥感；高光谱图像；基础模型；注意力机制；下游任务

1 引言

随着航空工程、传感器技术和计算机科学的迅猛发展，获取高光谱分辨率的海量遥感图像（HSIs）成为可能 [6]。高光谱图像涵盖了从可见光到近红外、短波和中红外的广泛光谱范围，通过连续且精细的波段捕捉目标的光谱特征。这种技术产生了近乎连续的光谱曲线，提供了详细的地表信息，从而能够识别物质之间的微小光谱差异，用于精确的土地覆盖分析 [12]。高光谱图像极大地提升了地球观测和监测的全面性、准确性和时效性 [16]。

基础模型（Foundation Models，FMs）作为一种新兴的技术，通过对普适性知识的编码，为解决这一问题提供了一个具有潜力的途径。近期，FMs 因其在多种任务中的卓越表现以及推动通用人工智能发展的潜力，成为人工智能领域的突破性进展，吸引了广泛关注 [1]。大多数 FMs 采用 Transformer 架构 [18]，并最早在自然语言处理（NLP）领域取得了显著成就，涌现了 XLNet [26]、GPT 系列 [14] 等一系列重要模型，包括 ChatGPT。该领域的成功进一步激发了基于视觉转换器（Vision Transformer，ViT）[15] 的大规模视觉基础模型的研发，并在多个视觉任务中展现出令人瞩目的性能。

遥感图像作为一种独特的视觉数据，传统上通过对现有模型进行直接微调来进行解译 [25]，通常是在自然图像数据集（如 ImageNet [5]）上进行预训练。然而，由于遥感图像与自然图像之间存在领域差异，这种方法可能不是最优选择。此外，遥感图像具有丰富的不同分辨率、波长和模态，且可以通过持续的卫星监测在广泛区域内进行获取。因此，近期的研究开始引入基于 ViT 的遥感数据预训练的基础模型。现有的遥感基础模型大多针对高分辨率的

航空 RGB 图像 [20]、多光谱图像 [3]、SAR 图像 [23] 或多模态图像 [7] 进行设计。然而，专门为高光谱图像解译量身定制的基础模型尚属稀缺。这个差距可能源于高光谱图像在数据采集、大规模预训练和模型架构设计等方面所面临的独特挑战。

尽管现有的遥感基础模型能够应用于高光谱图像解译任务，但它们仍受到领域差异的制约。自然场景的 RGB 图像通常由红、绿、蓝三个波段组成，而由无人机或卫星搭载的 SAR 相机获取的 SAR 图像一般具有 2 至 4 个波段。无人机或卫星上的多光谱相机所采集的图像则通常包括红、绿、蓝和近红外等波段。常见的航空 RGB 图像多来自 Google Earth 等在线资源，并通过相应通道提取自多光谱遥感图像。与这些图像类型不同，高光谱图像由无人机或卫星上的高光谱相机获取，包含数百个波段，能够在更细致的波长范围内捕捉地面物体的特征，显著优于航空 RGB、SAR 和多光谱图像在波长解析度上的能力。

本文主要复现并对 HyperSIGMA（HyperSpectral IntelliGence coMprehension foundAtion model）模型进行了微调，这是首个专门针对高光谱图像解译的基础模型。HyperSIGMA 通过专门设计的光谱增强模块融合了空间和光谱特征。为了解决高光谱图像中的光谱与空间冗余问题，该模型引入了一种创新的稀疏采样注意力（SSA）机制，能够有效促进多样化上下文特征的学习，并成为 HyperSIGMA 的核心组件。此外，模型还在大规模高光谱数据集 HyperGlobal-450K 上进行了预训练。该数据集包含约 45 万张高光谱图像，等效于超过 2000 万张非重叠通道的三光谱图像。在 HyperGlobal-450K 数据集上进行预训练，使得 HyperSIGMA 的参数规模扩展至超过十亿。

2 相关工作

2.1 高光谱图像处理

早期的高光谱图像特征提取方法主要依赖于原始像元数据或降维后的光谱信息，常见的算法包括 k 近邻 [17]、多项式逻辑回归 [10] 等多种传统机器学习方法。然而，这些方法在很大程度上依赖于专家知识，且性能存在局限性。近年来，基于深度学习的特征提取器，特别是卷积神经网络（CNN），已成为主流解决方案，能够在不同层次上捕捉光谱 [9]、空间 [13] 以及光谱-空间特征 [2]。此外，注意力机制通过为各通道或空间位置分配权重，进一步提升了特征提取的能力，尤其在远距离信息感知方面表现突出。最近，基于 Transformer 的网络 [8] 因其卓越的上下文建模能力，并通过多头自注意力机制（MHSA）获得了广泛应用。然而，由于这些模型通常是在每个场景中独立训练的，现有的特征提取器缺乏跨场景的泛化能力。因此，通过 HyperSIGMA 在大规模高光谱数据上进行预训练，从而实现更为广泛和通用的特征提取。

2.2 注意力机制

多头自注意力机制（MHSA）是 Transformer 架构的核心，能够自适应地捕捉全局上下文信息。然而，这一机制在处理大规模数据时效率低下。为了解决这一问题，Swin-Transformer [11] 引入了窗口化多头自注意力（WMHSA），通过将注意力限制在非重叠的局部窗口内，从而实现线性复杂度。然而，固定大小的窗口在处理真实场景中不同尺度和分布的物体时存在局限。基于此，RVSA [21] 在 VSA 的基础上进行了扩展，引入了可学习的旋转机制，以适应不同方

向物体的上下文提取。除了基于窗口的解决方案，近年来可变形卷积也逐渐获得关注。可变形卷积 [4] 通过自适应地采样卷积核位置进行处理。HyperSIGMA 提出了一种全新的稀疏采样注意力（SSA）机制，通过灵活地捕捉可变形区域内的上下文信息，增强了模型的适应性。该机制在每个区域内只采用少量采样点，从而有效处理高光谱图像中的冗余信息。

3 本文方法

3.1 本文方法概述

HyperSIGMA 的构建过程主要分为三个关键步骤：首先，通过预训练初始化模型权重；其次，引入稀疏采样注意力（SSA）机制以增强模型结构；最后，融合空间与光谱特征。为提取空间和光谱特征，模型采用了两个并行的子预训练网络，在 HyperGlobal-450K 数据集上使用 MAE [24] 进行预训练以获得模型的初始权重。需要强调的是，空间网络和光谱网络是独立进行预训练的。接下来，通过引入 SSA 机制，进一步增强了模型的结构性能。最后，空间特征和光谱特征相结合，提升了特征提取的表达能力，从而最终构建出 HyperSIGMA 模型。

3.2 数据来源与处理

3.2.1 数据来源

HyperGlobal-450K 数据集包含来自两种传感器的图像：地球观测一号（EO-1）和高分五号（GF-5B）。EO-1 是由 NASA 于 2000 年 11 月发射的卫星，轨道高度大约为 705 公里，执行地球表面的 242 个光谱波段扫描，覆盖从可见光到短波红外的范围，光谱分辨率为 10nm，空间分辨率为 30m。高分五号（GF-5B）是中国于 2018 年 5 月发射的卫星，运行轨道同样位于约 705 公里高空。GF-5B 涵盖了从紫外到长波红外（0.4-2.5μ m）的光谱范围，采集了 330 个高光谱波段的数据，光谱分辨率高达 5nm，空间分辨率同样为 30m。EO-1 卫星已于 2017 年退役，而 GF-5B 卫星仍在服役中。

3.2.2 数据处理

数据主要通过官方网站下载。对于 EO-1 数据，选择了 2011 年至 2017 年间的图像。对于 GF-5B 数据，选取了来自中国内蒙古自治区和河北省的 152 幅影像，河南省的 24 幅影像，以及内蒙古自治区、吉林省和辽宁省的 39 幅影像，如图 1所示。所有收集到的高光谱图像被切割成 64×64 大小的非重叠小块，用于模型的预训练。图中展示了来自不同地区的典型景观样本，包括森林、草地、裸地和农田，清晰地呈现了各自区域的独特地理特征。

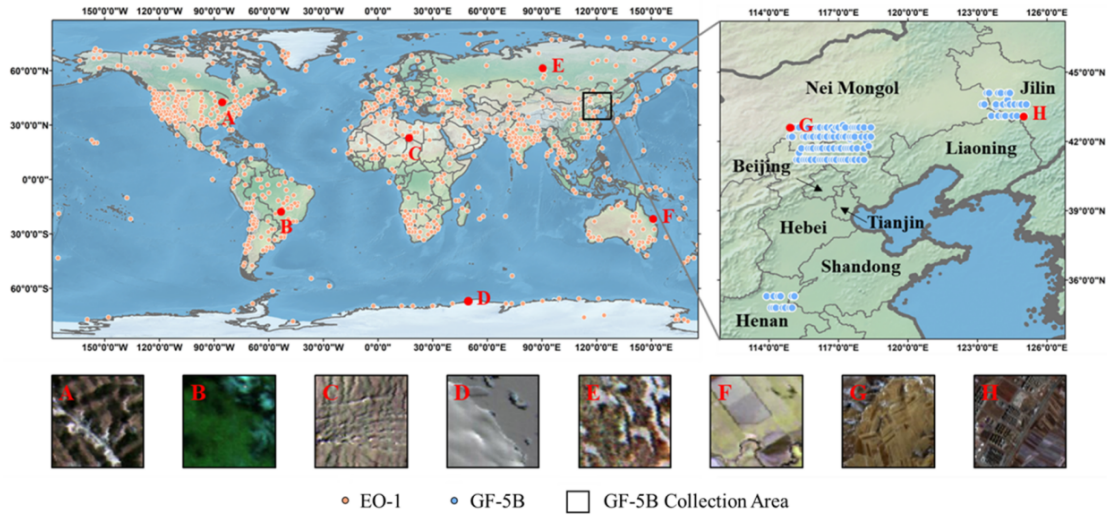

图 1. 数据获取与采集

3.3 模型预训练

MAE 是一种广泛应用的自监督学习方法。在 MAE 中，图像被划分为互不重叠的图像块，并且部分被遮掩。然后，模型利用可见的图像块来预测并恢复被遮掩的部分。损失函数通过对比模型预测值与实际被遮掩区域的真实值来计算。在大规模无标签数据集上，MAE 特别适用于 ViT 的预训练，主要用于在 HyperGlobal-450K 数据集上对空间子网络和光谱子网络进行预训练。空间子网络以 ViT 为基础结构，采用 MAE 进行预训练。在自然图像上的原始实现中，唯一的调整是修改块嵌入层的输入通道，以适应输入 HSI 中的通道数量。

3.4 模型结构

3.4.1 稀疏采样注意力机制

给定 $Q = \{q_1, \ldots, q_N\}$，$K = \{k_1, \ldots, k_N\}$，$V = \{v_1, \ldots, v_N\}$，对于坐标 (c_x, c_y) 处的每个查询向量 q，使用线性层 $W_p \in R^{D^{'} \times 2N_p}$ 预测 N_p 个偏移。使用双线性插值从这些位置的原始密钥和值矩阵 K 和 V 中采样新的 $k^{'}$ 和值 $v^{'}$。这一过程可以表述为:

$$k_j^{'} = \sum_{(o_x, 0_y)} max(0, 1 - |o_x - (c_x + \Delta x_j)|) max(0, 1 - |o_y - (c_y + \Delta y_j)|) K[o_x, o_y, :] \tag{1}$$

$$v_j^{'} = \sum_{(o_x, 0_y)} max(0, 1 - |o_x - (c_x + \Delta x_j)|) max(0, 1 - |o_y - (c_y + \Delta y_j)|) V[o_x, o_y, :] \tag{2}$$

这里，$j = 1, \ldots, N_p, K[o_x, o_y, :]$ 是从 K 中 (o_x, o_y) 处抽取的向量，其中 (o_x, o_y) 表示所有的坐标。对 $N \cdot N_p$ 点进行全采样，得到 $K^{'}, V^{'} \in R^{N \times N_p \times D^{'}}$。因此，SSA 可以表示为:

$$SSA(U) = softmax(\frac{Q \otimes K^{'}}{\sqrt{D^{'}}}) \otimes V^{'} \tag{3}$$

3.4.2 模型结构

在获取 SpatViT 和 SpecViT 的预训练权重后，HyperSIGMA 模型通过将原始的全自注意力（SA）替换为稀疏自注意力（SSA）并融合空间与光谱特征进行构建，整体架构如图2所示。该模型将从空间子网络和光谱子网络中提取的特征进行有效结合。特别需要注意的是，在 SpecViT 中，每个通道被转化为一个一维 token 向量，这导致了空间结构的破坏。为了保留光谱网络中的一维特征，采用了专门设计的光谱增强模块（Spectral Enhanced Module，SEM），该模块通过利用光谱信息来增强空间特征，从而实现空间与光谱特征的融合。

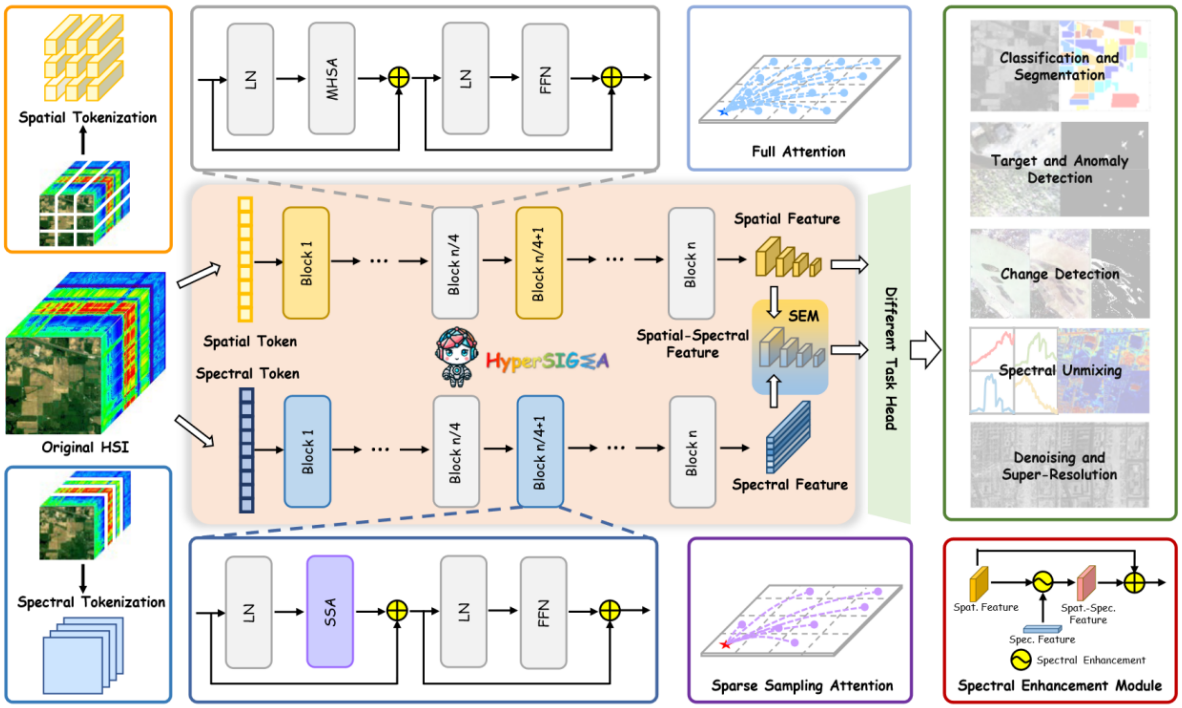

图 2. 模型结构

具体地，给定一个空间特征 $F_{spat} \in R^{H^{'} \times W^{'} \times D}$ 和一个谱特征 $F_{spec} \in R^{N_{spec} \times D}$，首先使用线性层压缩它们的维数，得到 $F_{spat} \in R^{H^{'} \times W^{'} \times D_1}$ 和 $F_{spec} \in R^{N_{spec} \times D_1}$。减少空间维度的 F_{spec} 通过平均聚合创建一个 1D 向量 $V \in R^{N_{spec}}$。最后，我们使用另一个线性层将 V 的维度与 F_{spat} 对齐。光谱增强过程表述如下：

$$F^* = (1 + V) * F_{spat} \tag{4}$$

使用跳跃连接来保留原始的空间信息。通过这种方式，利用光谱信息对提取的空间特征进行校准。

4 复现细节

4.1 与已有开源代码对比

本文复现参考代码来源于文献 [19] 公开的 HyperSIGMA。与原文开源代码相比，本文主要添加了数据集、增加了可视化展示代码和卷积块对性能影响的探索实验。

4.2 实验设置

对光谱子网络进行了预训练，使用了不同的掩模比率，范围从 0.15-0.9，步长为 0.15，共训练了 400 个 epochs，batch size 设置为 2048。训练过程中采用 AdamW 优化器，学习率为 0.00015，权重衰减系数为 0.05，其他参数遵循 MAE 的默认配置。训练时使用了交叉熵损失函数，并通过总体精度 (OA) 来评估模型性能。在确定了最佳掩模比后，采用之前提到的默认配置重新训练了空间和光谱子网络。为确保训练充分，将预训练过程延长至 1600 个 epochs。对于给定的图像大小 X_0，空间 ViT 的图像块大小设置为 8，从而生成 64 个 tokens。空间子网络的训练参数，包括批次大小、学习率和权重衰减，与光谱子网络的训练配置保持一致。

5 实验结果分析

本文主要评估了 HyperSIGMA 在具有代表性的 HSI 高层解译任务：图像分类和变化检测，以及低层任务：光谱解混上的性能。

5.1 高光谱图像分类

本研究采用了三个广泛使用的高光谱图像数据集，包括 Indian Pines、Pavia University 和 WHU-Hi 数据集中 HongHu 场景的数据。分类任务主要针对高光谱图像中的每个像素进行分类，类似于自然或航空 RGB 图像中的语义分割任务。该任务可以通过图斑级分类或图像级分类来实现。图斑级分类选择以每个像素为中心的图斑，并根据图斑的特征对中心像素进行分类，而图像级分类则直接对应于语义分割任务。在实验中，分别采用 33 × 33 和 128 × 128 大小的图斑进行斑块级分类和分割，空间图斑大小设为 8。对于 Indian Pines (IP)、Pavia University (PU) 以及 HongHu 数据集，训练集和测试集的划分比例为 9:1。模型的训练使用了 0.00006 的学习率，共进行了 500 个 epoch，并将光谱分支的 token 数设置为 100。分类性能的评估通过总体精度（OA）、平均精度（AA）和 Kappa 系数（Kappa）进行。

表 1. 地物分类评价结果

Dataset	AA	OA	Kappa
Indian pines	0.9993	0.9980	0.9977
Pavia University	0.9936	0.9974	0.9966
Honghu	0.9986	0.9994	0.9992

表1展示了模型在不同评估指标下的分类精度，结果表明，在这三个典型的地物分类数据集上，模型在各个指标上均取得了显著的表现，特别是在 HongHu 数据集上，OA 达到了 0.9992，光谱信息的引入显著提升了精度。然而，AA 指标的表现略显不足，表明由于图像块尺寸较大，模型在捕捉小型物体时面临一定的困难。

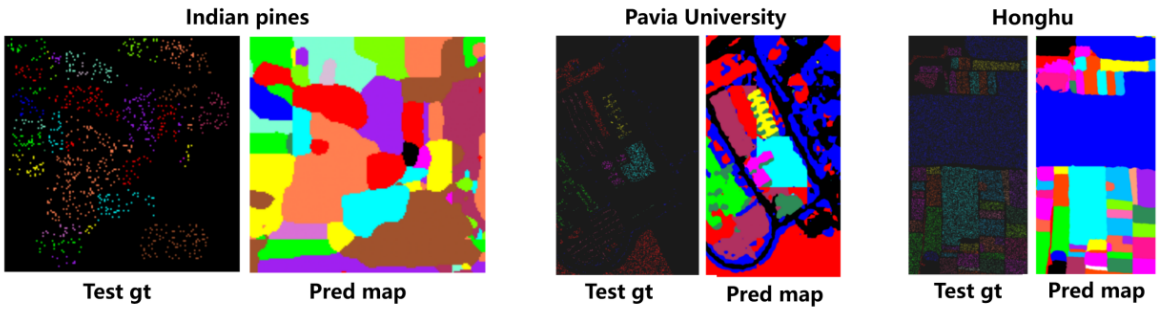

图 3. 地物分类图

图3展示了不同数据集上的分类结果可视化。通过对三个数据集的分类结果进行定性分析，结果表明，模型有效地缓解了传统方法中常见的诸如椒盐噪声、过度平滑以及误分类等问题。

5.2 高光谱变化检测

主要对用于高光谱变化检测的预训练模型进行了微调，该模型与常规的航空 RGB 图像遥感任务相结合，重点处理经典的双时态场景。该模型在四个常用的基准数据集 Hermiston、Farmland、Barbara 和 BayArea 上进行了评估。借鉴 SST-Former [22] 在高光谱变化检测任务中的成功经验，这是一种基于 Transformer 的变化检测方法，采用了分类网络。在训练时，Hermiston 和 Farmland 数据集使用了 5×5 大小的图像块作为输入，而 BayArea 和 Barbara 数据集则使用了 15×15 的图像块。为了适应较大的变异性，空间子网络的嵌入层图斑大小设置为 1（针对 Hermiston 和 Farmland）和 2（针对 BayArea 和 Barbara）。每个数据集随机选取 500 个不变像素和 500 个变化像素用于训练。模型训练采用 0.00006 的学习率和 32 的批次大小，进行了 50 个 epoch 的训练。光谱分支使用了 144 个 token 来捕获细致的光谱信息。我们通过总体精度（OA）、Kappa 系数、F1 值、精确率和召回率等指标评估模型性能。表2展示了在 Hermiston、Farmland、BayArea 和 Santa Barbara 数据集上的定量结果。

表 2. 不同数据集上的变化检测精度及对比结果

Dataset	Method	F1-score	Recall	Precision	OA	Kappa
BayArea	SpatSIGMA	0.9894	0.9928	0.9861	0.9886	0.9772
	HyperSIGMA	0.9901	0.9921	0.9881	0.9894	0.9787
Barbara	SpatSIGMA	0.9882	0.9849	0.9916	0.9908	0.9806
	HyperSIGMA	0.9916	0.9890	0.9942	0.9934	0.9861
Hermiston	SpatSIGMA	0.9156	0.9636	0.8720	0.9600	0.8895
	HyperSIGMA	0.9158	0.9614	0.8744	0.9602	0.8899
Farmland	SpatSIGMA	0.9551	0.9840	0.9279	0.9732	0.9361
	HyperSIGMA	0.9563	0.9836	0.9304	0.9739	0.9377

与 SpatSIGMA 模型相比，HyperSIGMA 模型在所有测试数据集上均达到了最高的 OA、Kappa 和 F1 分数，体现了其显著的优势。然而，模型的分类精度仍然存在改进空间，尤其是在区分不变区域和变化区域时，模型可能会出现误分类的情况。

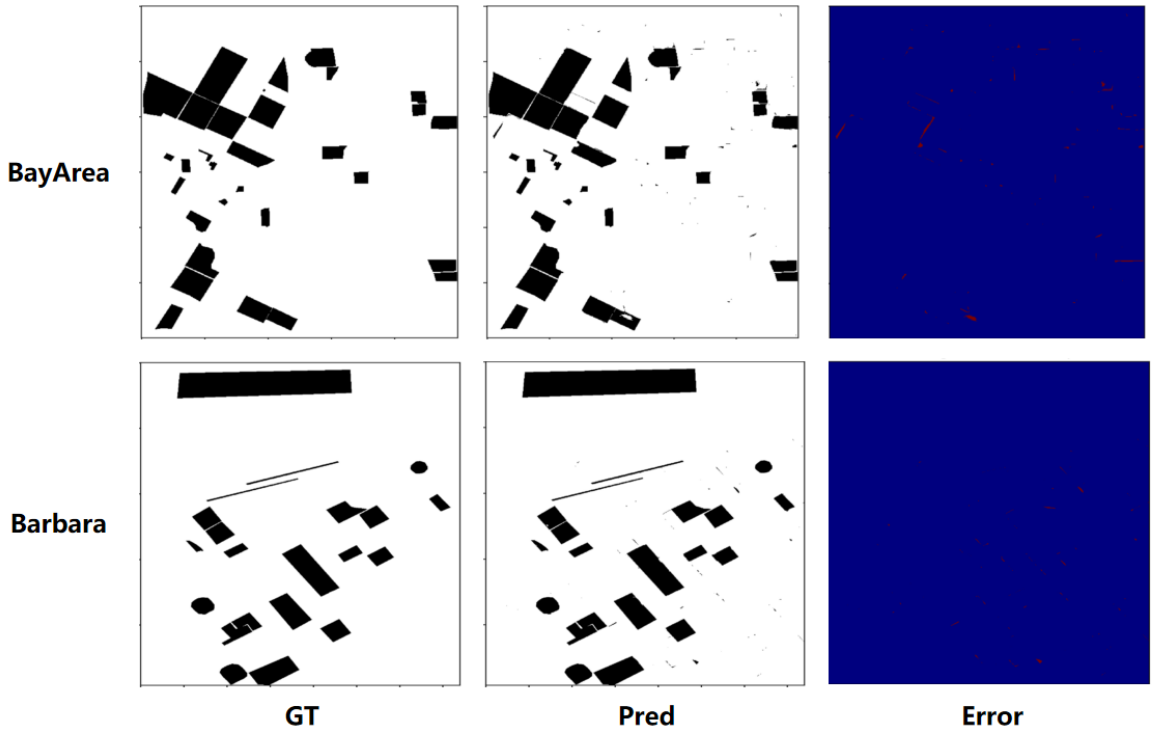

图 4. 在 BayArea 数据集上的变化检测图

图4呈现了 BayArea 数据集上的变化检测结果，模型能够精确识别大部分变化区域，显著提高了召回率，提供了更加细致和全面的检测结果。这进一步验证了 HyperSIGMA 在高光谱变化检测任务中卓越的特征表达能力。

5.3 高光谱解混

在将 HyperSIGMA 应用于高层次的下游任务后，进一步将其拓展到低层任务，特别是高光谱解混。该任务旨在处理高光谱数据中的光谱混合问题，通过将每个像素的光谱信息分解为纯光谱特征（端元）以及它们对应的丰度（比例）。这种方法有助于识别和定量分析每个像素中的不同成分。

表 3. 在 Urban 数据集上定量比较端元预测性能

Abundance	SpatSIGMA	HyperSIGMA
Asphalt	0.0164	0.0156
Grass	0.0320	0.0302
Tree	0.0104	0.0094
Roof	0.0054	0.0050
Avg	0.0160	0.0150

为了评估模型在光谱解混任务中的表现，采用了 Urban 数据集。在实施过程中，HyperSIGMA 被用作特征提取器，类似于分类任务，为了全面理解高光谱图像场景，融合了空间分支的四个特征图和光谱分支的最终输出层，使用不同的光谱增强模块（SEM）来进行信息融合与聚合。输入数据为 64×64 大小的图像块，在空间 ViT 中使用了 2 大小的图像块。解码

器的卷积核大小设为 1。为对结果进行定量评估，采用了两种常见的指标：平均光谱角度距离（mSAD），用来比较学习到的端元与参考端元之间的相似度；以及均方误差（MSE），用以评估生成的丰度图的质量。表3中的定量结果表明，HyperSIGMA 在端元提取上表现优异。

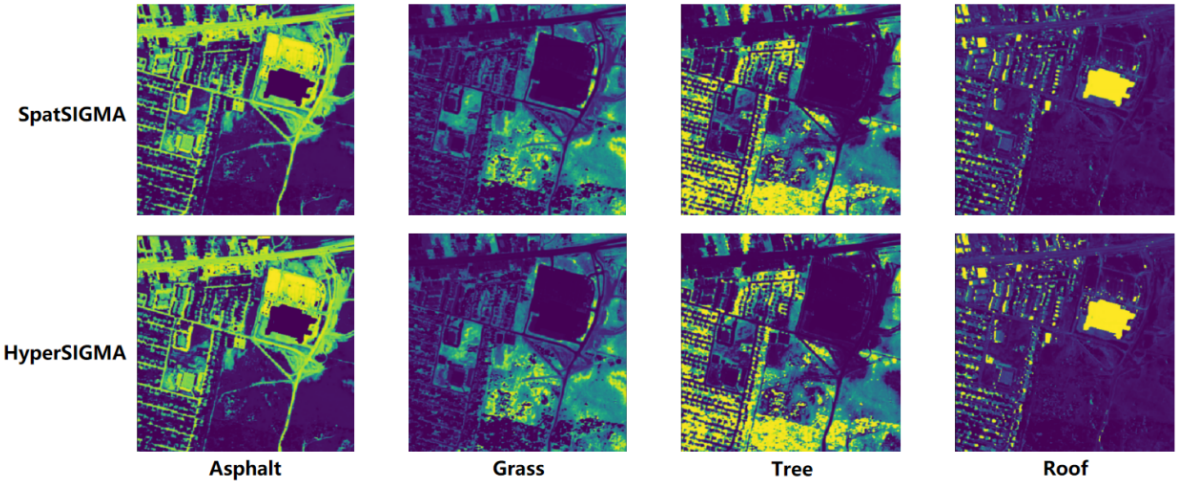

图 5. HyperSIGMA 在 Urban 数据集上的高光谱解混丰度预测结果

模型的定性评估进一步验证了这一点。图5展示了 HyperSIGMA 在高光谱图像中准确地描绘了混合地物的实际分布情况。图6则表明，HyperSIGMA 成功地捕捉到了纯端元特征，并且与地面真值非常接近，特别是在树状端元的提取上。这些结果强调了该模型在增强高光谱数据低层次任务中的潜力。

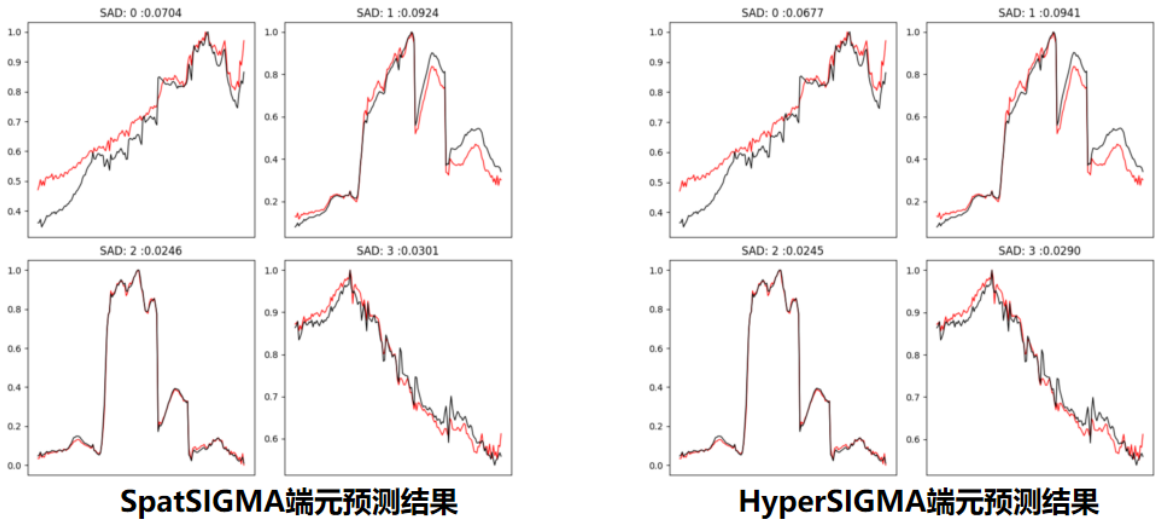

图 6. HyperSIGMA 对 Urban 数据集高光谱解混的端元预测结果

6 总结与展望

本文主要复现了第一个应用于高光谱图像解译的高光谱基础模型 HyperSIGMA，该模型具有超过 10 亿个参数。为了解决高光谱图像中的冗余问题，HyperSIGMA 提出了一种创新的稀疏采样注意力机制，能够以较少的可学习采样点自适应地感知相关上下文区域。此外，模型还设计了光谱增强模块，成功实现了高效的空谱特征融合。在高层和低层高光谱任务上的综

合评估显示，HyperSIGMA 展现了卓越的性能。进一步分析表明，HyperSIGMA 具有良好的可扩展性。与 SpatSIGMA 相比，HyperSIGMA 在性能上有所提升，尤其是在光谱子网络中。这些差异可能与光谱预训练的质量相关，尤其是在恢复完整通道时面临的挑战，以及光谱表示在训练和推理过程中对输入空间大小一致性的要求。因此，未来需要进一步的研究，以开发更高效的光谱基础模型。

参考文献

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Yushi Chen, Kaiqiang Zhu, Lin Zhu, Xin He, Pedram Ghamisi, and Jón Atli Benediktsson. Automatic design of convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):7048–7066, 2019.

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Alexander FH Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. Imaging spectrometry for earth remote sensing. *science*, 228(4704):1147–1153, 1985.

[7] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[8] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

[9] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015(1):258619, 2015.

[10] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4085–4098, 2010.

[11] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[12] Bryce Manifold, Shuaiqian Men, Ruoqian Hu, and Dan Fu. A versatile deep learning architecture for classification and label-free prediction of hyperspectral images. *Nature machine intelligence*, 3(4):306–315, 2021.

[13] Zijia Niu, Wen Liu, Jingyi Zhao, and Guoqian Jiang. Deeplab-based spatial feature extraction for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 16(2):251–255, 2018.

[14] Alec Radford. Improving language understanding by generative pre-training. 2018.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Xian Sun, Dongshuo Yin, Fei Qin, Hongfeng Yu, Wanxuan Lu, Fanglong Yao, Qibin He, Xingliang Huang, Zhiyuan Yan, Peijin Wang, et al. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. *Nature Communications*, 14(1):1444, 2023.

[17] Bing Tu, Jinping Wang, Xudong Kang, Guoyun Zhang, Xianfeng Ou, and Longyuan Guo. Knn-based representation of superpixels for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11):4032–4047, 2018.

[18] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[19] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. Hypersigma: Hyperspectral intelligence comprehension foundation model, 2024.

[20] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2022.

[21] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

[22] Yanheng Wang, Danfeng Hong, Jianjun Sha, Lianru Gao, Lian Liu, Yonggang Zhang, and Xianhui Rong. Spectral–spatial–temporal transformers for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[23] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[24] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

[25] Kejie Xu, Hong Huang, Peifang Deng, and Yuan Li. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5751–5765, 2021.

[26] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.