arXiv:2603.01174v1 [cs.CV] 1 Mar 2026

VP-Hype：一种采用视觉-文本提示进行高光谱图像分类的混合 Mamba-Transformer 框架

Abdellah Zakaria Sellam[1,2],Fadi Abdeladhim Zidi[3], Salah Eddine Bekhouche[4],Ihssen Houhou[5],Marouane Tliba[7],Cosimo Distante[1,2], 和 Abdenour Hadid[61]应用科学与智能系统研究所（ISASI），CNR，意大利莱切 73100[2]萨伦托大学创新工程系，意大利莱切 73100[3]比斯克拉大学电子与自动化系 VSC实验室，阿尔及利亚比斯克拉[4]巴斯克地区大学（UPV/EHU）计算机科学与人工智能系，西班牙[5]比斯克拉大学电子与自动化系 VSC实验室，阿尔及利亚[6]阿布扎比索邦大学索邦人工智能中心，阿联酋阿布扎比[7]索邦巴黎北部大学伽利略研究所，法国维勒塔讷兹 F-93430

2026年3月3日

摘要

高光谱图像（HSI）的精准分类往往受制于高维光谱数据与标注训练样本极度稀缺之间的矛盾。LoLA-SpecViT 等分层模型已展现出局部窗口注意力和参数高效微调的优势，但标准 Transformer 的二次复杂度仍然阻碍着模型扩展。我们提出 VP-Hype，这一框架通过一种新颖的混合架构，将状态空间模型（SSM）的线性时间效率与 Transformer 的关系建模能力相结合，重新构思高光谱图像分类。在稳健的 3D-CNN光谱前端基础上，VP-Hype 以混合Mamba-Transformer骨干网络取代传统注意力模块，从而捕捉

长程依赖，同时显著降低计算开销。此外，我们通过集成双模态视觉与文本提示，为特征提取过程提供上下文感知引导，从而解决标签稀缺问题。实验评估表明，VP-Hype 在低数据量场景下达到了新的先进水平。具体而言，在训练样本仅占 2% 的情况下，该模型在 Salinas数据集和龙口数据集上的总体精度（OA）分别达到 99.69% 和 99.45%。这些结果表明，混合序列建模与多模态提示的结合，为实现高性能、样本高效的遥感提供了可靠的发展路径。

关键词：高光谱图像分类，Mamba，视觉Transformer，提示学习，多模态学习，Remote Sensing。

1 引言

高光谱成像（HSI）可获取数百个连续波段上的密集光谱测量数据，生成高维数据立方体，从而支持精细的材质识别和细粒度土地覆盖判别，适用于精准农业、环境监测和城市制图等应用 [1]。这些光谱特征能够捕捉细微的生化和结构特性，例如早期胁迫标志和营养缺乏，而这些特性往往无法通过多光谱或RGB传感器观测到 [2]。与此同时，高光谱图像分类面临着波段间冗余显著、维数灾难以及地面真值获取成本高昂所导致的标签严重稀缺等挑战。

因此，高光谱图像深度学习方法必须平衡三个相互作用的目标：（i）保留光谱-空间归纳偏置，以捕捉细粒度纹理和波段级线索；（ii）建模长程光谱依赖，以区分光谱相似类别；（iii）确保对高分辨率数据立方体的计算具有可行性。卷积网络能够有效捕捉局部光谱–空间结构 [3], ，而视觉变换器则为全局交互提供了灵活有效的机制 [4]；然而，标准自注意力的计算复杂度随序列长度呈平方增长，这对于高维高光谱序列而言可能难以承受 [5]。近期研究通过分层令牌化和局部窗口注意力，以及采用注意力机制的替代方案（例如能够实现线性时间序列建模的状态空间模型（SSMs）），来应对这一权衡 [6]。此外，基于提示的条件引导（视觉和文本提示）已成为一种轻量级机制，能够以极少的参数更新引导预训练模型适应下游任务 [7], ；提示范式在复原任务中也展现出特别大的潜力 [8]。

受这些互补趋势的启发，我们提出了**VP-Hype**，一种用于高光谱图像分类的混合Mamba–Transformer架构，它融合了高效的基于SSM的光谱处理、分层窗口注意力和双模态提示条件引导。如图1所示，VP-

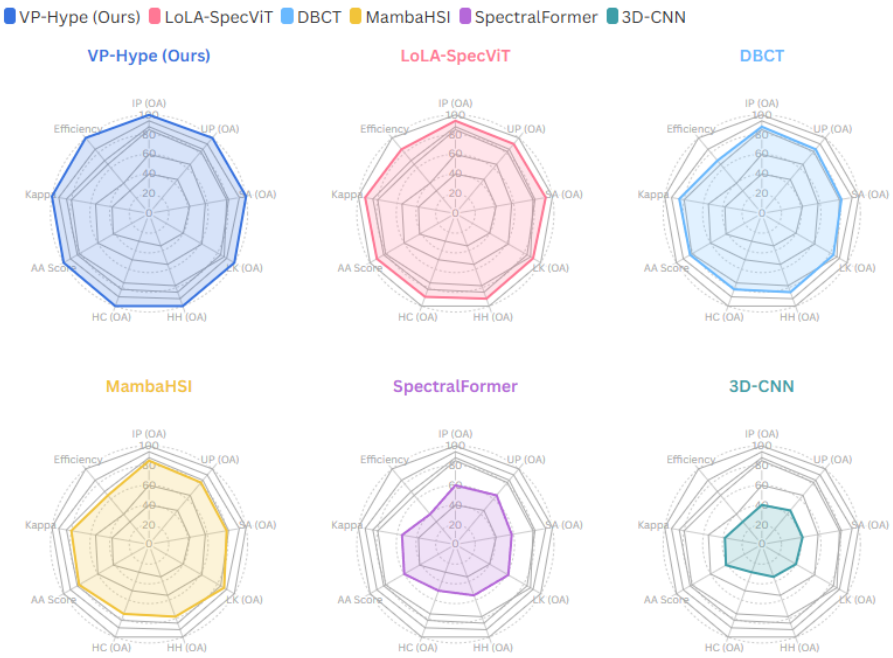

图1：九种高光谱图像（HSI）分类模型的性能指标对比。所有分数均相对于所提出的VP-Hype进行归一化，VP-Hype在所有数据集（IP、UP、SA、LK、HH、HC）、总体精度指标（平均精度、卡帕系数）以及计算效率方面均表现出更优性能。

与六个基准数据集上的九种有竞争力的基线方法相比，Hype 在总体精度、平均准确率和卡帕系数方面表现更优，同时保持了良好的计算效率。VP-Hype 采用紧凑的三维卷积光谱前端，以保留局部归纳偏置；在分层骨干网络中交替使用状态空间（Mamba）和窗口化自注意力模块，以兼顾长程建模能力与线性规模的计算效率；并注入可学习的视觉–文本提示，在中间特征层级提供语义和空间指导。本研究的主要贡献如下：

- 我们设计了**VP-Hype**，这是一种混合 Mamba–Transformer 分类器，将 3D-CNN 光谱前端与分层骨干网络相结合，并在其中交替使用状态空间模型混合器和窗口化注意力，以实现效率与表达能力之间的权衡 y 。
- 我们提出了一种视觉–文本提示融合模块，将 CLIP 风格的文本描述符与可学习的空间视觉提示相结合，从而实现任务感知的条件调节，在监督有限的情况下提升判别能力。
- 我们在标准高光谱基准数据集上开展了全面实验，并通过消融研究分别评估混合架构中的混合器、提示模态和提示注入策略的贡献。

本文其余部分的结构如下：第 2 节回顾相关文献；第 3 节详细介绍 VP-Hype 架构；第 4 节描述实验设置；第 5 节展示并分析研究结果；第 6 节开展系统的消融研究，以分别评估各架构组件的贡献并验证关键设计选择；第 7 节解读实验结果，探讨所提框架的局限性，并结合更广泛的高光谱图像分类研究对结果进行分析；第 8 节总结本研究并提出未来研究方向。

2 相关工作

高光谱图像（HSI）分析已从手工设计的描述符发展到学习得到的光谱–空间表示，旨在为大型高光谱图像数据立方体 [9]兼顾判别能力与计算可行性。与这一目标相关的文献可归纳为四个相辅相成的方向：局部光谱–空间建模、全局上下文建模及其高效替代方案、架构效率与参数适配，以及基于提示的多模态条件化。每个方向都各具优势，也暴露出相应局限，推动了面向判别性高光谱图像任务的混合式提示感知分类器的发展。

2.1 局部光谱—空间建模

早期的高光谱图像分类深度学习方法以卷积架构为主，原因在于卷积天然具有面向局部光谱–空间结构的归纳偏置。Li 等人 [3] 的开创性工作表明，三维卷积能够同时沿光谱维度和空间维度进行滤波，为联合特征提取奠定了基础范式。在此基础上，Roy 等人 [10]提出了**HybridSN**，通过级联三维卷积阶段和二维卷积阶段，逐步降低波段间冗余，同时保留紧凑的联合表征；这种设计在不牺牲判别能力的前提下显著提升了效率。后续研究致力于进一步增强局部感受野：Yang 等人[11] 引入多尺度感受野模块，在参数量没有相应增长的情况下拓展空间上下文；Cui 等人 [12] 则在卷积块中加入通道注意力，以强化局部聚合；Ding 等人 [13] 通过 LANet 的局部性感知特征细化，解决了与之相辅相成的边界歧义问题。

总体而言，这些研究显著推动了卷积范式的发展；然而，它们存在一个根本性的架构限制。由于卷积算子的感受野本质上是局部的，要捕捉长程光谱依赖——这对于区分高光谱图像中光谱相似类别至关重要——就必须进行显式的感受野设计，或采用越来越深的级联结构 [14]。这两种方法都会带来不可忽视的计算开销，并增加优化难度，对于高光谱数据特有的长光谱序列而言尤其如此 [5]。这一内在局限促使研究转向基于注意力的架构和序列建模架构，下一小节将对此展开讨论。

2.2 全局上下文建模与高效替代方案

卷积感受野的根本局限促使研究范式转向能够建模长程光谱依赖关系的基于序列的架构。Dosovitskiy 等人 [4]提出了视觉Transformer，奠定了将图像块视为令牌序列的理论基础；这一原则对于高光谱图像尤为重要，因为波段间的相关性横跨整个光谱轴。鉴于通用预训练可能无法忠实迁移至高光谱领域，Ibanez 等人 [15] 提出了掩码光谱预训练，使自监督目标与光谱结构相契合；Bai 等人 [16] 则开发了跨传感器自监督方法，进一步提升模型在异构采集条件下的迁移能力。然而，标准自注意力的计算量随序列长度呈二次增长，这对于高光谱图像立方体固有的高维令牌序列而言，代价高得难以承受。为此，Sun 等人 [17]提出记忆增强注意力以降低资源需求，Zhou 等人 [18] 引入令牌剪枝以消除冗余光谱令牌，Xu 等人 [19] 则提出分层令牌化，以较低成本捕捉多分辨率结构；与此同时，Arshad 等人

[20] 探索了混合 Transformer-CNN 设计，在保留全局建模能力的同时恢复局部归纳偏置。另一种结构上截然不同的方案则完全摒弃注意力，以应对可扩展性问题。Gu 等人 [6] 提出了 Mamba，这是一种能够以线性时间进行序列建模且表达能力具有竞争力的状态空间模型；Ahmad 等人 [21]则证明，基于状态空间模型的架构能够有效迁移至高光谱序列。总而言之，这些研究表明，全局光谱上下文对于类别判别不可或缺，但如何高效计算仍是一个尚待解决的架构挑战 [22]。

2.3 架构效率与参数适配

即使全局建模十分高效，在目标高光谱图像数据集上部署大型预训练模型仍会遇到第二个瓶颈：在带标签数据有限的情况下进行全量微调的成本。为此，研究逐渐形成了两个互补方向：降低架构成本，以及在不修改骨干网络的情况下实现轻量级适配。在架构方面，Khan 等人 [23] 提出基于分组的注意力，以直接降低二次复杂度；Fu 等人[24] 开发了动态剪枝，以在运行时移除冗余的空间与通道计算；Zhang 等人 [22]则提出卷积感知型 MLP 变体，以更低的参数预算达到有竞争力的准确率。在适配方面，Houlsby 等人 [25] 引入适配器模块，将小型可训练瓶颈模块插入冻结的骨干网络，为高效迁移确立了范式。基于这一原则，Jia 等人 [26] 证明，只需在冻结的骨干网络前置可学习视觉词元，而无需修改内部权重，就足以使其适应新任务；BenZaken 等人 [27]则表明，仅选择性更新偏置参数也能实现有竞争力的适配；Hu 等人 [28] 提出将低秩分解作为一种有原则的方法，用以约束参数更新空间。尽管这些技术在自然图像场景中十分有效，但并不能轻易泛化至高光谱图像，因为光谱保真度和对分布偏移的敏感性带来了额外约束，而通用适配策略并未考虑这些约束 [5]。

2.4 提示学习与多模态条件化

上述参数高效适配方法主要作用于模型权重；一种更灵活的替代方案是通过输入空间提示来调节模型行为，从而在不修改模型权重的情况下实现任务感知引导。Jia 等人 [7] 为视觉骨干网络确立了这一原则，表明在输入序列前添加一小组可学习的提示词元，便能有效引导冻结模型适应下游任务。鉴于单独的视觉提示所承载的语义内容有限，后续图像复原研究引入了更丰富的条件信号：PromptIR [29]和 PIP [30] 提出了能够编码损坏类型的退化感知视觉提示，而 InstructIR [31]则证明，自然语言指令可以作为灵活且人类可理解的条件信号。Wu 等人 [8], 探索了这两种模态的融合，

他们提出 MP-HSIR，将文本提示与视觉提示融合，用于多退化高光谱图像复原；PromptHSI [32], 则专门将基于文本的提示应用于光谱复原。总体而言，这些研究表明，双模态提示能够比任一单独模态提供更丰富的任务表征；然而，它们仅限于生成式复原场景。将其扩展到判别式高光谱图像分类仍鲜有探索。在这种任务中，提示不仅要编码退化先验，还必须在标签严重稀缺的情况下表征语义与光谱-空间类别结构。VP-Hype 正是为解决这一空白而设计的。

综述的这些工作沿着四个互补方向推进了高光谱图像分类：局部细节保真度、长程光谱建模、计算可处理性以及灵活的任务条件化。然而，每个方向上的进展都暴露出了一些局限性，现有的单一方法尚无法完全解决这些问题。卷积混合架构受感受野限制；基于变换器的模型会产生随高光谱序列长度扩展且表现欠佳的二次注意力开销；状态空间模型具备线性时间效率，但需要谨慎地融入光谱先验；而基于提示的框架几乎仅在生成式复原场景中得到验证，尚未充分应用于标签稀缺条件下的判别式分类。

这些观察结果共同指向一个明确的架构要求：一种有原则的解决方案必须同时保留局部光谱-空间归纳偏置，以可处理的计算开销建模全局上下文，并在有限监督下利用语义条件化。VP-Hype通过在统一且计算高效的框架内，将基于状态空间模型的光谱骨干网络与分层窗口自注意力以及双模态视觉-文本提示融合机制相结合，满足了这一三重要求。

3 提出的方法

本节介绍所提出的**VP-Hype**，一种用于高光谱图像分类的、带有视觉-文本提示的混合Mamba-Transformer框架。该架构将状态空间模型（Mamba）与自注意力机制相结合，以实现高效的光谱-空间特征处理，并通过任务感知的文本提示和视觉提示加以增强。如图2所示，该架构包含四个主要组成部分：（1）图块嵌入与光谱-空间特征提取，（2）采用分层处理的混合Mamba-Transformer骨干网络，（3）用于任务感知适配的视觉-文本提示系统，以及（4）带有全局特征聚合的分类头。

3.1 问题形式化与输入预处理

令$\mathbf{I} \in \mathbb{R}^{B\times C\times H\times W}$ 表示一批高光谱图像，其中 B 为批量大小， C 为光谱波段数，$H \times W$ 为空间分辨率。令$t \in \{0,\ldots,T-1\}$ 为可选的任务标识符， N为目标的数量

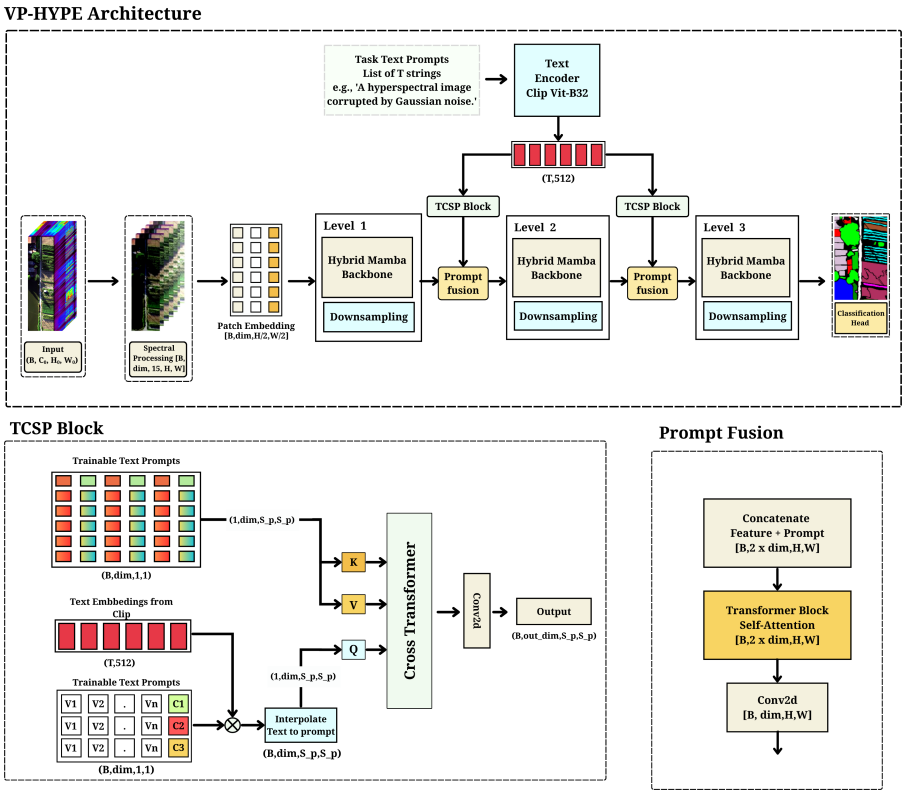

图 2：所提出的 VP-Hype 框架架构。该模型将混合 Mamba–Transformer 骨干网络与用于高光谱图像分类的视觉-文本提示相结合。由 CLIP 编码的任务特定文本提示和可学习视觉提示通过文本条件空间提示（TCSP）模块进行融合，并注入网络的多个阶段。经提示增强的特征逐步下采样，最终送入分类头进行预测。

类别。模型学习一个映射

$$f : \mathbb{R}^{B\times C\times H\times W} \times \mathbb{Z}^{B} \to \mathbb{R}^{B\times N}, \tag{1}$$

生成逻辑值 $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times N}$。

3.2 光谱-空间前端

紧凑型卷积前端在提取光谱-空间词元的同时保留局部归纳偏置。具体而言，该前端采用两个步幅为3×3 的卷积，生成一个嵌入张量

$$\mathbf{X}_0 = \mathrm{PatchEmbed}(\mathbf{I}) \in \mathbb{R}^{B\times d\times H'\times W'}, \tag{2}$$

其中，$H' = H/4$、$W' = W/4$和 d 为基础嵌入维度。该模块降低空间分辨率，并生成紧凑的特征词元，以供后续序列建模使用。

3.3 分层混合骨干网络

骨干网络采用层次化结构，包含 L 个阶段。在阶段 l 中，特征张量记为

$$\mathbf{X}_l \in \mathbb{R}^{B\times d_l\times H_l\times W_l}, \quad d_l = d\cdot 2^l, \quad (H_l, W_l) = \frac{(H', W')}{2^l}. \tag{3}$$

每个阶段由一组混合块构成。对于阶段 l，我们用 depth_l 表示块深度，并将各块的索引记为 $i \in \{0, \ldots, \mathsf{depth}_l - 1\}$。混合器的选择遵循简单的混合规则

$$\mathrm{Mixer}_{l,i} = \begin{cases} \mathrm{MambaVisionMixer}, & i < \lfloor \mathrm{depth}_l/2 \rfloor, \\ \mathrm{WindowedAttention}, & \text{otherwise.} \end{cases} \tag{4}$$

窗口化注意力作用于大小为 w_l 的不重叠空间窗口，每个窗口生成长度为 $L_l = w_l^2$ 的序列。窗口划分以及将其重塑为序列的过程表示为

$$\mathbf{X}_{l,\mathrm{win}} = \mathrm{WindowPartition}(\mathbf{X}_l; w_l), \qquad \mathbf{s}_l = \mathrm{Reshape}(\mathbf{X}_{l,\mathrm{win}}) \in \mathbb{R}^{(B\cdot N_{w,l})\times L_l\times d_l}, \tag{5}$$

其中 $N_{w,\,l} = \lceil H_l/w_l \rceil \cdot \lceil W_l/w_l \rceil$ 表示阶段 l 中的窗口数量。每个混合块均采用 Transformer 式残差连接，并包含归一化和多层感知机。将通用混合器记为 $\mathcal{M}$，则单个块的更新为

$$\mathbf{u} \leftarrow \mathbf{u} + \gamma_1 \odot \mathrm{DropPath}(\mathcal{M}(\mathrm{LayerNorm}(\mathbf{u}))), \quad \mathbf{u} \leftarrow \mathbf{u} + \gamma_2 \odot \mathrm{DropPath}(\mathrm{MLP}(\mathrm{LayerNorm}(\mathbf{u}))). \tag{6}$$

这种构造利用 Mamba 的线性时间序列建模，在早期获取全局光谱上下文，并利用注意力的二次表达能力进行局部空间细化。

3.4 MambaVisionMixer

对于指定为 MambaVisionMixer 的模块，我们先进行投影，再执行选择性状态空间操作。给定一个序列 $\mathbf{s} \in \mathbb{R}^{M \times L_l \times d_l}$ （其中 $M = B \cdot N_{w,\ l}$），混合器执行：

$$\mathbf{z} = \mathbf{s}\mathbf{W}_{in} \in \mathbb{R}^{M \times L_l \times d_{in}}, \qquad [\mathbf{z}_1, \mathbf{z}_2] = \mathrm{Split}(\mathbf{z}), \tag{7}$$

其中，$\mathbf{W}_{in} \in \mathbb{R}^{d_l \times d_{in}}$，且 d_{in} 为扩展维度。每个分支经过处理（可选地使用深度卷积），随后传入选择性扫描算子 $\mathcal{S}$，以实现结构化状态空间递归：

$$\mathbf{y} = \mathcal{S}(\mathbf{z}_1; \Theta), \qquad \mathbf{h}_{\mathrm{mamba}} = \mathrm{Concat}(\mathbf{y}, \mathbf{z}_2)\mathbf{W}_{out}, \tag{8}$$

其中，Θ 表示学习得到的状态空间模型参数，$\mathbf{W}_{out} \in \mathbb{R}^{d_{in} \times d_l}$ 将其投影回模型维度。公式（8）抽象了实现细节，同时保留了混合器的作用：先进行线性时间的谱递归，再通过拼接实现门控融合。

3.5 窗口多头自注意力

当 Mixer$_{l,\ i}$ 为窗口化注意力时，我们在每个窗口序列内计算标准的缩放点积注意力。给定一个窗口序列，$\mathbf{s}$我们将查询、键和值构造为

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{W}_{qkv}\,\mathbf{s}, \tag{9}$$

应用多头注意力，并使用$\mathbf{W}_{proj}$ 对拼接后的各个头进行投影，以获得$\mathbf{h}_{\mathrm{attn}}$，随后通过 WindowReverse(·)将其重新整合到阶段空间布局中。

3.6 视觉-文本提示

为注入语义与空间先验，我们采用双模态提示，并将其融合后插入中间阶段。

文本提示源自冻结的 CLIP编码器。设$\mathbf{E}_{clip} \in \mathbb{R}^{T \times 512}$为预先计算的任务嵌入，$\mathbf{w} \in \{0,1\}^{B \times T}$ 为独热任务指示向量。任务条件文本向量为

$$\mathbf{e}_t = \sum_{i=0}^{T-1} \mathbf{w}_{:,i} \odot \mathbf{E}_{clip}[i,:] \in \mathbb{R}^{B \times 512}, \tag{10}$$

再通过$\mathbf{W}_{clip}$将其投影至提示维度 d_p。

可学习视觉提示被参数化为紧凑的空间张量$\mathbf{P}_v \in \mathbb{R}^{1 \times d_p \times S_p \times S_p}$。文本条件空间提示（TCSP）模块通过一个小型交叉注意力块融合$\mathbf{e}_t$ 和$\mathbf{P}_v$，生成空间提示

$$\mathbf{P} = \mathrm{TCSP}(\mathbf{e}_t, \mathbf{P}_v) \in \mathbb{R}^{B \times C_f \times H_f \times W_f}, \tag{11}$$

其中通过插值和轻量投影，确保$\mathbf{P}$与骨干网络特征几何结构（C_f，H_f，W_f）相匹配。

3.7 提示融合与分类

生成的提示通过通道拼接和紧凑的基于 Transformer 的融合与骨干网络特征合并：

$$\mathbf{F}_{concat} = [\mathbf{F}, \mathbf{P}], \qquad \mathbf{F}_{fused} = \mathrm{Proj}_{1\times1}\big(\mathrm{TransformerBlock}(\mathbf{F}_{concat})\big), \tag{12}$$

其中，$\mathrm{Proj}_{1\times1}$ 恢复原始通道维度。最终的骨干网络输出$\mathbf{X}_{L-1}$ 经过归一化和全局平均池化：

$$\mathbf{f}_{pool} = \frac{1}{H_{L-1}W_{L-1}} \sum_{h=1}^{H_{L-1}} \sum_{w=1}^{W_{L-1}} \mathbf{X}_{L-1}[:,:,h,w] \in \mathbb{R}^{B\times d_{L-1}}. \tag{13}$$

逻辑值由线性分类器生成，

$$\hat{\mathbf{Y}} = \mathbf{W}_{cls}\,\mathbf{f}_{pool} + \mathbf{b}_{cls} \in \mathbb{R}^{B\times N}, \tag{14}$$

类别概率则通过 Softmax 得出。该架构采用模块化设计：卷积前端保留光谱-空间局部性，混合骨干网络结合线性时间递归以捕获广泛的光谱上下文，并通过注意力进行空间细化；TCSP 机制则在无需大量针对特定任务重新训练的情况下注入语义与空间先验。第 6 节的消融研究量化了 Mamba 与注意力模块之间的权衡，并评估了提示注入层级和提示模态贡献。

4 实验

4.1 数据集

在本研究中，我们采用三个公开可用的高光谱数据集——WHU- Hi-LongKou、WHU-Hi-HongHu 和 Salinas，全面评估分类框架在多种农业场景下的表现。这些数据集在空间分辨率、光谱覆盖范围、场景复杂度和土地覆盖类别数量方面各不相同，为评估模型的泛化能力提供了坚实基础。

WHU-Hi-LongKou 数据集于2018年7月17日在龙口镇采集，使用安装在DJI Matrice 600 Pro无人机上的Headwall Nano-Hyperspec传感器。该影像覆盖一个 550×400像素的区域，空间分辨率约为0.463米，包含从400 nm到1000 nm的270个光谱波段。它描绘了一个简单的农业场景，包含六种主要作物：玉米、棉花、芝麻、阔叶大豆、窄叶大豆和水稻。所提供的地面真值定义了九个土地覆盖类别，包括六种作物、水体、建设用地和混合杂草 [33]。**WHU-Hi-HongHu** 数据集于2017年11月20日在洪湖市采集，使用同一台安装在DJI Matrice 600 Pro无人机上的Nano-Hyperspec传感器。该图像大小为 940×475 像素，具有270个光谱波段，空间分辨率为0.043米。该场景描绘了一个复杂的农业试验场，包含多种作物类型，以及道路、屋顶和裸土等非作物地物。地面真值包含22个土地覆盖

类别，涵盖植被和基础设施类别 [34]。**Salinas** 场景是一幅美国加利福尼亚州萨利纳斯谷农业用地的机载高光谱图像。它由美国国家航空航天局的AVIRIS仪器以3.7米地面采样距离获取。该场景为 512×217 像素，包含224个原始波段。它描述了蔬菜田、葡萄园和裸土，并标注为16个土地覆盖类别 [35]。

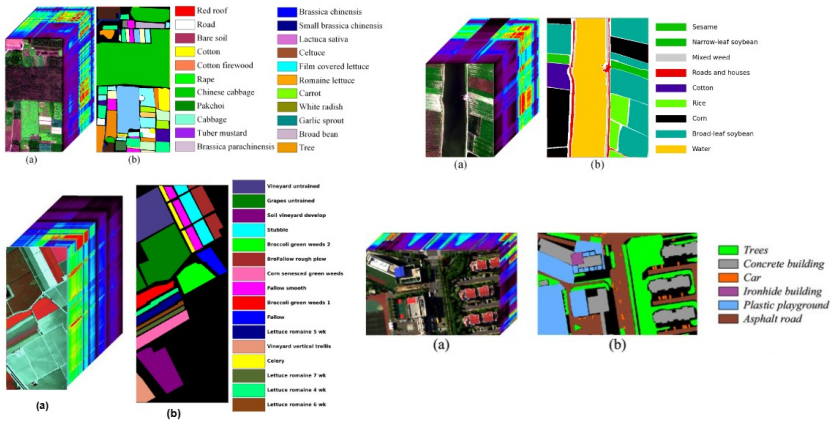

图 3：WHU-Hi-HongH、WHU-Hi-Longkou 和 Salinas 数据集的高光谱数据立方体及其对应真实分类图的可视化：(a) 高光谱图像立方体；(b) 真实标签图。

5 实验结果

5.1 HongHu 数据集（10% 训练样本）

下表报告了使用 10% 标注训练样本时在 HongHu 数据集上的分类性能。比较指标包括总体精度（OA）、平均准确率（AA）、科恩卡帕系数以及各类别的准确率。

表 1：HongHu 数据集上的总体性能比较（10% 训练样本）。

指标	**Ours**	HSIMAE	HybridSN	ViT	MASSFormer	DBCT	LSGA
总体准确率 (%)	**99.64**	96.24	96.76	96.76	98.14	99.14	99.02
平均精度 (%)	**99.23**	92.99	92.50	92.50	96.25	98.07	97.54
卡帕系数 (%)	**99.55**	96.18	96.77	96.77	97.65	98.91	98.75

VP-Hype在所有评估指标上均达到了最先进的性能。具体而言，我们取得了 **99.64%** 的总体精度（OA），相当于 $+$**0.50%**

相较于次优方法（DBCT：99.14%）的提升，以及相较于排名第三的方法（LSGA：99.02%）$+\mathbf{1.50\%}$ 的提升。平均准确率（AA）达到$\mathbf{99.23\%}$，比DBCT（98.07%）高出 $+\mathbf{1.16\%}$，比LSGA（97.54%）高出 $+\mathbf{1.69\%}$。平均准确率上的显著差距表明，其在所有类别上均具有更优性能，而不仅仅是主导类别。VP-Hype的卡帕系数达到$\mathbf{99.55\%}$，反映出预测结果与真实分类之间高度一致，较DBCT（98.91%）提升了 $+\mathbf{0.64\%}$ 。

鉴于以下几点，这些性能提升尤为显著：(1) 训练数据有限（仅占样本的10%），因此泛化颇具挑战；(2) 竞争方法的基线性能较高，改进难度较大；(3) 在全部三项汇总指标上均持续占优。

5.2 Salinas数据集（10% 训练样本）

为评估所提出模型在 Salinas 数据集上的性能，我们将总体精度、平均精度、卡帕系数和各类别精度与多种先进方法进行了比较，包括 MADANet [12], A^2MFE [11], HybridSN [36], SSFTT [37], HiT [38], LANet [13], 和 LoLA [5]。结果表明，我们的模型在大多数指标和类别上均取得了最高精度。

表 2：Salinas 数据集上的总体性能比较（10% 训练样本）。

指标	**Ours**	LoLA	MADANet	A^2MFE	HybridSN	SSFTT	HiT	LANet
总体准确率 (%)	**99.99**	99.87	99.17	98.56	97.05	98.61	97.83	96.67
平均精度 (%)	**99.99**	99.71	99.12	98.72	97.52	98.97	98.87	97.12
卡帕系数 (%)	**99.99**	99.49	98.34	97.35	96.72	98.46	97.58	96.39

VP-Hype 在所有评估指标上均展现出最先进的卓越性能，分类精度近乎完美。具体而言，我们的总体精度（OA）达到 $\mathbf{99.99\%}$ ，较排名第二的方法（LoLA：99.87%）提高了 $+\mathbf{0.12\%}$ ，较排名第三的方法（MADANet：99.17%）提高了 $+\mathbf{0.82\%}$ 。平均准确率（AA）达到 $\mathbf{99.99\%}$，比 LoLA（99.71%）高出 $+\mathbf{0.28\%}$，比 MADANet（99.12%）高出 $+\mathbf{0.87\%}$。VP-Hype 的卡帕系数达到 $\mathbf{99.99\%}$ ，表明预测分类与真实分类几乎完全一致；较 LoLA（99.49%）提高了$+\mathbf{0.50\%}$ ，较 MADANet（98.34%）提高了 $+\mathbf{1.65\%}$ 。

鉴于以下因素，这些性能提升尤为显著：(1) 训练数据有限（仅占样本的10%），使泛化颇具挑战；(2) 竞争方法的基线性能已经很高（LoLA 的准确率达到>99%），进一步提升极其困难；(3) 在全部三项汇总指标上同时实现近乎完美的准确率（99.99%），这在高光谱图像分类中实属罕见。

5.3 龙口数据集（10% 训练样本）

在龙口数据集上对所提出的模型进行了评估，并将其与近期方法进行比较，包括 Cross [16], Hybrid-ViT [20], Hir-Transformer [39],SSMamba [40], MorpMamba [40], E-SR-SSIM [41], 和 LOLA [5]。我们的模型在所有指标上均取得最高准确率。

表 3：龙口数据集（10% 训练样本）的总体准确率、平均准确率和卡帕系数（%）。

指标	我们的方法	Cross	Hybrid-ViT	Hir-Trans.	SSMamba	MorpMamba	E-SR-SSIM	LOLA
总体准确率 (%)	**99.95**	99.78	99.75	99.68	99.51	99.70	94.43	99.92
平均精度 (%)	**99.83**	99.45	99.36	99.14	98.45	99.25	81.78	99.73
卡帕系数 (%)	99.81	99.78	99.68	99.59	99.36	99.61	92.63	**99.89**

VP-Hype 在主要评估指标上均取得了最先进性能。具体而言，我们的总体精度（OA）达到 **99.95%** ，比排名第二的方法（Cross：99.78%）高出 **+0.17%**，比排名第三的方法（Hybrid-ViT：99.75%）高出 **+0.20%** 。平均准确率（AA）达到 **99.83%**，比 LOLA（99.73%）高出 **+0.10%** ，比 Cross（99.45%）高出 **+0.38%**。VP-Hype 的卡帕系数为 99.81%，表现颇具竞争力；LOLA 的 Kappa 系数最高，为 99.89%，两者相差 -0.08%。不过，VP-Hype 在两个主要指标（OA 和 AA）上表现最佳，而这两个指标通常被视为衡量分类性能最重要的指标。

鉴于以下因素，这些性能提升尤为显著：(1) 训练数据有限（仅占样本的 10%），使得泛化颇具挑战；(2) 竞争方法的基线性能较高，因此难以进一步提升；(3) 竞争格局激烈，存在多种强劲方法，包括近期基于 Mamba 的方法（SSMamba、MorpMamba）和 Transformer 变体（Hybrid-ViT、Hir-Transformer）。

5.4 龙口数据集上的比较分析

VP-Hype 在所有评估指标上均取得最先进性能，即使训练数据极少，也展现出显著的有效性。具体而言，我们取得了**99.45%** 的总体精度（OA），比排名第二的方法（AMHFN：97.1%）高出 **+2.35%** ，比排名第三的方法（GAHT：96.8%）高出**+2.65%** 。平均精度（AA）达到**99.2%**，比 AMHFN（96.1%）高出 **+3.1%**，比 GAHT（95.8%）高出 **+3.4%**。VP-Hype 的卡帕系数达到 **99.28%** ，反映出预测结果与真实分类高度一致；相比 AMHFN（96.17%）提高了**+3.11%**，相比 GAHT（95.86%）提高了 **+3.42%** 。

鉴于以下因素，这些性能提升尤为显著：（1）训练数据极其有限（仅占样本的 2%），使泛化极具挑战性；

表 4：龙口数据集上的分类精度（%）（2% 训练样本）。最佳结果以粗体标示。

类别	ViT	PiT	HiT	GAHT	AMHFN	**Ours**
1	89.7	89.4	89.8	95.9	95.9	**99.99**
2	63.6	87.4	90.2	94.8	95.3	**98.82**
3	75.1	48.2	88.0	94.4	96.2	**98.99**
4	90.1	89.6	90.6	95.6	96.3	**99.84**
5	79.0	76.8	91.6	95.8	94.5	**98.48**
6	96.4	95.5	96.1	97.9	98.1	**99.82**
7	97.6	97.6	97.6	99.0	99.0	**99.89**
8	94.1	91.8	93.3	96.9	99.1	94.6
9	89.7	90.6	92.3	92.0	92.8	**93.35**
AA	86.2	85.2	92.2	95.8	96.1	**99.2**
OA	91.5	91.7	93.2	96.8	97.1	**99.45**
卡帕系数	88.93	89.20	91.20	95.86	96.17	**99.28**

（2）竞争方法的基线性能已经很高，进一步提升十分困难；（3）三项汇总指标均取得了显著提升，充分体现了全面优势。

5.5 有限训练样本下的性能评估

表 5：MambaVision-T 在三个数据集上的有限训练样本性能。

数据集	训练比例（%）	总体准确率 (%)	平均精度 (%)	卡帕系数（%）
龙口	2	99.45	99.20	99.28
Salinas	2	99.69	99.78	99.65
QUH-Qingyun	10	99.34	99.18	99.13

VP-Hype 在所有三个数据集上均取得了卓越且稳定的性能 ，展现出强大的泛化能力。在训练样本占比为 2% 的**龙口** 数据集上，VP-Hype 的**总体精度（OA）**为**99.45%**，**平均精度（AA）**为**99.20%**，**Kappa 系数**为**99.28%** 。在训练样本占比为 2% 的**Salinas** 数据集上，VP-Hype 的**OA**为**99.69%**，**AA**为**99.78%**，**Kappa 系数**为**99.65%**。在训练样本占比为 10% 的**QUH-Qingyun**数据集上，VP-Hype 的**OA**为**99.34%**，**AA**为**99.18%**，**Kappa 系数**为**99.13%** 。

这些结果表明，在多种数据集上均展现出**卓越的一致性**，三项基准测试中的所有性能指标均超过99%的阈值。此外，该模型表现出**稳健的性能**，在

稀疏数据条件下，即使仅使用2%的训练样本，准确率仍保持在99%以上。这种稳定性进一步体现在各数据集上观察到的**极小方差**，总体精度（OA）介于99.34%至99.69%之间，仅0.35%的可忽略差距凸显了该模型的可靠性。最后，研究结果证实了所提方法在广泛的土地覆盖类别、农业场景以及不同地理位置上的**全面有效性**。

6 消融研究

表 6：使用 2% 训练样本的消融实验结果

数据集配置		视觉	文本提示	总体准确率 (%)	平均精度 (%)	卡帕系数 (%)	
龙口	仅视觉	✓	✗	✓	99.14	97.23	98.88
	仅文本	✗	✓	✓	98.30	99.00	98.10
	无提示	✓	✓	✗	99.26	97.55	99.02
	完整模型（我们的方法）	✓	✓	✓	**99.45**	**99.20**	**99.28**
Salinas	仅视觉	✓	✗	✓	98.35	99.03	98.16
	仅文本	✗	✓	✓	98.30	99.00	98.10
	无提示	✓	✓	✗	98.28	99.05	98.08
	完整模型（我们的方法）	✓	✓	✓	**99.69**	**99.78**	**99.65**

仅视觉配置（启用视觉提示，禁用文本提示）凸显了可学习空间提示的重要性。在**龙口**数据集上，仅视觉配置的总体准确率（OA）为 99.14%，平均精度（AA）为 97.23%，卡帕系数为 98.88% ；与无提示基线相比，OA 提高了 $+0.88\%$，AA 提高了 $+1.77\%$，卡帕系数提高了 $+0.40\%$。在 **Salinas**数据集上，仅视觉配置的 OA 为 98.35%，AA 为 99.03%，卡帕系数为 98.16%；与无提示基线相比，OA 提高了 $+0.07\%$，AA 降低了 0.02%，卡帕系数提高了 $+0.08\%$。

这些结果表明，视觉提示带来了显著优势，尤其是在总体精度和 Kappa 系数方面。可学习空间提示使模型能够捕获任务特定空间模式，以补充基础特征表示，从而提升分类性能。

仅视觉配置展示了可学习空间提示在全部三个数据集上的重要性。在**龙口** 数据集上，仅视觉取得了99.52%的总体准确率、99.47%的平均精度和99.46%的卡帕系数；与无提示基线相比，总体准确率、平均精度和卡帕系数分别提升了 $+0.28\%$、 $+0.29\%$和 $+0.37\%$。在**Salinas** 数据集上，仅视觉取得了99.84%的总体准确率、99.82%的平均精度和99.82%的卡帕系数；与无提示基线相比，总体准确率、平均精度和卡帕系数分别提升了 $+0.03\%$、-0.03%和 $+0.03\%$。在**洪湖**数据集上，仅视觉取得了99.47%的总体准确率、98.94%的平均精度和99.33%的卡帕系数；与无提示基线相比，总体准确率、平均精度和卡帕系数分别提升了 $+0.04\%$、 $+0.17\%$和 $+0.05\%$。

这些结果表明，视觉提示在不同数据集上均能带来稳定的收益，尤其是在总体精度和 Kappa 系数方面。可学习空间

表 7：10% 训练样本下的消融实验结果

数据集配置		视觉	文本	提示	总体准确率（%）	平均精度（%）	卡帕系数（%）
龙口	仅视觉	✓	✗	✓	99.52	99.47	99.46
	仅文本	✗	✓	✓	99.59	99.37	99.13
	无提示	✓	✓	✗	99.24	99.18	99.09
	完整模型（我们的方法）	✓	✓	✓	**99.72**	**99.59**	**99.66**
Salinas	仅视觉	✓	✗	✓	99.84	99.82	99.82
	仅文本	✗	✓	✓	99.75	99.76	99.72
	无提示	✓	✓	✗	99.81	99.85	99.79
	完整模型（我们的方法）	✓	✓	✓	**99.91**	**99.88**	**99.89**
洪湖	仅视觉	✓	✗	✓	99.47	98.94	99.33
	仅文本	✗	✓	✓	99.43	98.80	99.28
	无提示	✓	✓	✗	99.43	98.77	99.28
	完整模型（我们的方法）	✓	✓	✓	**99.81**	**99.53**	**99.29**

提示使模型能够捕捉任务特定空间模式，从而补充基础特征表示。

7 讨论

7.1 架构设计分析

实验结果验证了 VP-Hype 的 Mamba-Transformer 混合架构。始终保持最先进性能（洪湖上的总体准确率为 99.64%，Salinas 上为 99.99%，龙口上为 99.95%），证明了结合状态空间模型与自注意力的互补优势。该混合策略利用 Mamba 的 O(n) 复杂度进行高效的早期处理，并利用 Transformer 的 O(n2) 注意力进行判别性细化；相较于纯基线方法的性能提升（比视觉变换器高 $+2.88\%$ 总体准确率，比 SS-Mamba 高 $+0.44\%$ 总体准确率）验证了这一点。通过逐步下采样进行分层处理（在图块嵌入处为 $4\times$，每个层级为 $2\times$），可捕捉从 $d \times H/4 \times W/4$ 到 $8d \times H/32 \times W/32$ 的多尺度特征，从而区分光谱相似类别。基于窗口的注意力将每个窗口的复杂度从 O($H \times W$) 降至 O(w_l^2)，同时确保适用于高分辨率高光谱图像的可扩展性。

文本-视觉提示系统通过三种互补机制实现任务感知的特征适配：预计算的 CLIP 文本嵌入、可学习的空间视觉提示以及交叉注意力融合。文本提示通路利用针对特定任务描述的预计算嵌入，发挥 CLIP 的语义知识，从而实现高效适配且不会增加推理开销。任务嵌入的加权组合使特征提取能够适应任务上下文，这一点可从相较于无提示基线的持续提升中得到验证（OA 提高了 $+0.38\%$ 至 $+1.41\%$）。消融研究表明，文本提示尤其能提升平均准确率（洪湖数据集在训练数据占比为 10% 时，AA 达到 $+0.76\%$），这说明语义引导有助于区分光谱相似的土地覆盖类型，从而实现更均衡的性能。

视觉提示通路采用随机初始化的可学习空间参数，并采用较为保守的缩放系数（$0.001\times$），以确保训练稳定性。视觉提示在交叉注意力中充当键和值，提供空间结构，以补充基于文本的语义引导。消融结果表明，视觉提示在提升总体精度和 Kappa 系数方面表现出色（龙口数据集在训练数据占比为 10% 时，OA 达到 $+0.28\%$，Kappa 达到 $+0.37\%$），这表明它们能够捕捉空间模式，从而增强判别性特征学习。

文本-视觉空间提示（TCSP）模块通过交叉注意力融合整合两条路径，并采用可学习温度参数 τ 自适应控制注意力锐度。该自适应机制根据任务需求平衡空间信息与语义信息，这对于多样化的高光谱成像场景至关重要。消融研究验证了这种协同效应：完整模型始终优于各个单独组件（在龙口上的总体准确率为 $+0.48\%$，在 Salinas 上以 10% 训练数据训练时为 $+1.41\%$）。

在多个层次级别（通常为级别 1 和 2）进行提示注入，可在不同规模上适配特征。早期级别的注入提供粗粒度的任务感知引导，而后期级别的注入则细化判别性特征。这种多层次策略确保任务感知适配贯穿整个特征学习过程，而不仅限于分类阶段。提示融合模块将特征与提示拼接，应用自注意力，并投影回原始维度，从而实现无缝集成，同时不影响基础架构的特征流。

7.2 视觉对比

图 4 展示了 VP-Hype 与具有代表性的先进方法在三个具有不同空间特征和语义复杂度的高光谱基准场景上的定性比较：Salinas、WHU-Hi-LongKou 和 WHU-Hi-HongHu。这些视觉评估结果印证了表 1、2、4 中的定量结果：在中等（10%）和极其稀疏（2%）的训练方案下，VP-Hype 均取得了接近满分的表现。尤为重要的是，该方法在仅使用 10% 监督数据时，仍能在 Salinas 上保持近乎完美的准确率；即使在 2% 的训练数据下也表现稳健，并在更具挑战性的 LongKou 和 HongHu 场景中取得同样出色的结果。除了汇总指标外，预测的分类图还显示出显著提升的结构保真度、边界精度和区域一致性；这些改进直接源自各架构组件的协同整合。

图 4 的面板 I 中的 Salinas 场景展现了一个根本性挑战：如何区分光谱特征高度重叠、且分布在大面积连续地块中的 16 个农业类别。VP-Hype 生成的分类图几乎与地面真值无异，田块内部均匀，过渡区域定位精准，几乎没有虚假预测；相比之下，其他方法在有限监督下会产生零散的田内误差和模糊的类别间边界。这种性能差异源于三个相互作用的组成部分：由 CLIP嵌入生成的文本提示提供高层语义锚点，使模型能够依据学得的语义关系，而非光谱邻近度，区分光谱特征相似的作物

；视觉提示通过编码田块边界几何先验和内部均一性的可学习模板，强化地块级空间结构；Hybrid-Mamba骨干网络则通过交替的状态空间模块和注意力模块，将这些多模态线索传播至整个区域。其中，Mamba 的线性时间递归建立全局光谱上下文，窗口化注意力则细化局部空间边界。提示模块与骨干网络之间的双向信息流确保语义引导能够在多个尺度上指导空间特征提取，从而避免纯局部模型中常见的块级不一致性。

图版 II 中的龙口场景带来了独特挑战：狭长条带、不规则田块几何形状以及强烈的空间各向异性，容易使感受野有限的模型将细长结构割裂。即使在 2 %训练设置下，VP-Hype 仍能保持狭长地块的连贯性，并准确贴合复杂边界；而基线方法则会产生不连续片段、边界溢出和椒盐伪影。在此，架构优势尤为明显：视觉提示通过可学习的空间模板突出边缘、脊线和细长区域等具有判别性的空间原语，以强调几何连续性；文本提示则通过将特征表示锚定到语义原型，减少视觉相似覆盖类型之间的混淆；Hybrid-Mamba骨干网络通过分层处理整合这些多尺度信号，其中早期 Mamba 模块沿条带方向捕获长程依赖，而后期注意力模块则细化局部边界细节。这种渐进式细化结合多个阶段的提示注入，使模型能够沿延展结构进行连贯标注，同时不牺牲局部精度；对于农业监测而言，这是一项关键能力，因为田块边界决定了管理分区。

图版 III 中的洪湖场景具有 22 个土地覆盖类别，呈现出最具挑战性的配置：这些类别具有显著的类内差异，且异质背景中嵌有大量小目标。VP-Hype 在实现高精度的同时，保持了分类别识别的平衡，这一点从总体精度与平均准确率之间的差距极小可见；在视觉上，它还能保留稀有类别和精细结构，而竞争方法往往会过度平滑这些类别和结构，或将其误分类为主导类别——大面积区域仍保持连贯，目标边界则清晰锐利。在高维类别条件下，双提示机制通过互补协作实现了这种平衡的性能：文本提示充当语义锚点，有助于提升众多类别之间的区分度，尤其使少数类受益，避免它们被与主导类别的光谱相似性所淹没；冻结的 CLIP 嵌入则提供稳定的语义先验，避免对稀有类别的稀疏样本过拟合。视觉提示通过空间模式强化维持目标级完整性，确保小目标保留其结构身份，而非被周围上下文吸收；Hybrid-Mamba 骨干网络中的分层处理则逐步细化这些多尺度特征：浅层捕捉精细的光谱—空间细节，中间层整合局部上下文，深层融合全局关系——所有层级均由适当规模注入的提示信息引导

这种多分辨率方法能够同时保留小目标并维持大面积区域的连贯性，而纯Transformer 或卷积神经网络通常难以实现这种权衡。

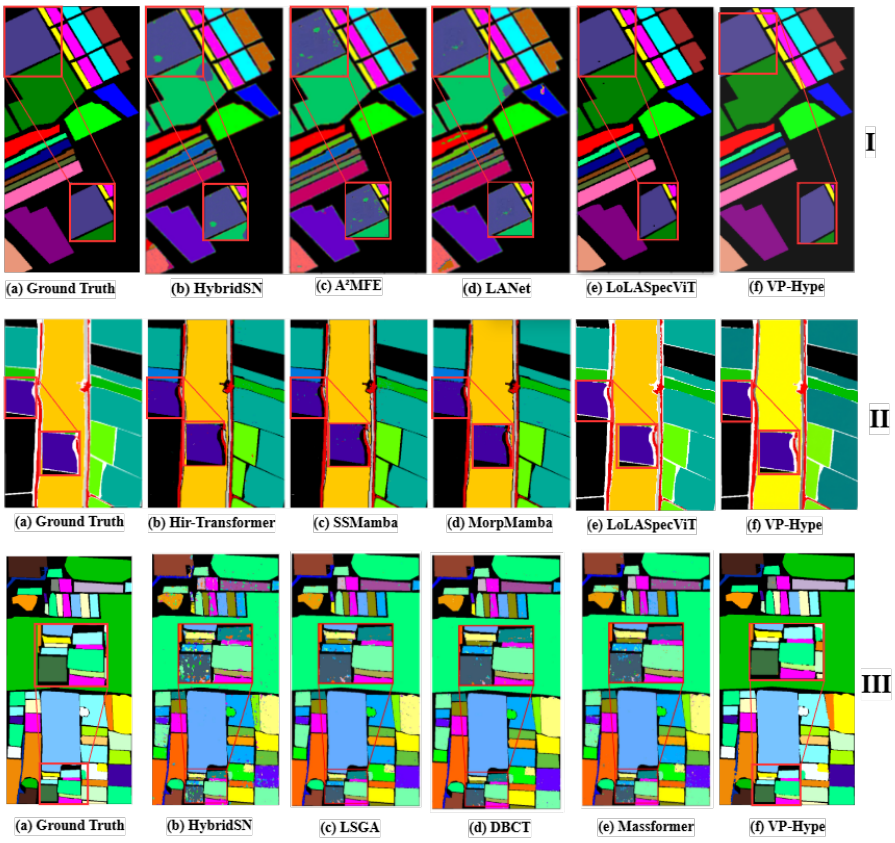

图 4：在标准高光谱基准数据集上的分类性能对比：(I) Salinas、(II) WHU-Hi-LongKou 和 (III) WHU-Hi-HongHu 数据集。图中展示了地面真值和各方法的预测结果，并突出显示了具有挑战性的区域（红框）。

在各数据集上，VP-Hype能够产生更清晰的边界、更强的空间连贯性，并更好地保留细长区域与小区域。这源于紧密的架构耦合：卷积前端保留局部光谱空间先验；Hybrid-Mamba骨干网络交替使用线性时间状态空间模块以获取全局上下文，并使用窗口自注意力进行局部细化；冻结的CLIP文本提示提供语义先验，而可学习视觉提示编码空间模板；TCSP交叉注意力结合可学习温度以及提示融合模块，将经过校准的多尺度信号注入骨干网络。简而言之，文本提示指导关注什么，视觉提示指示

在何处观察，而混合骨干网络决定如何整合这些线索，从而在极端标签稀疏条件下生成样本高效、结构可靠的图。

7.3 提示系统有效性

在 Salinas 数据集上使用 2% 训练样本进行的消融研究，系统地分离了 VP-Hype 框架中各提示模态的贡献。如表 8 所示，同时移除两种提示类型会导致性能最低（总体准确率 $= 98.28\%$、平均精度 $= 99.05\%$、卡帕系数 $= 98.08\%$），这证实提示系统是整体架构中的关键组成部分。

仅引入文本提示（4 个提示）即可显著提升性能：总体准确率提高 $+1.03\%$（达到 99.31%）、平均精度提高 $+0.49\%$（达到 99.54%）、卡帕系数提高 $+1.15\%$（达到 99.23% ）。这表明，通过基于 CLIP 的文本嵌入编码的语义引导，对特征判别具有很强的正则化作用。仅使用视觉提示（4 个提示）也能优于基线（总体准确率提高 $+0.54\%$、平均精度提高 $+0.26\%$ 、卡帕系数提高 $+0.60\%$），但提升幅度较小，这说明空间模式编码提供了互补但次要的判别线索。值得注意的是，与视觉提示相比，文本提示对总体精度和类别可分性的影响更大，这表明在此设置中，语义上下文是占主导地位的归纳偏置。

表8：文本提示与视觉提示数量对VP-Hype性能影响的消融研究（Salinas数据集，2%训练样本）。

文本提示	视觉提示	总体准确率 (%)	平均精度 (%)	卡帕系数（%）
无（基线）	None	98.28	99.05	98.08
4 个提示	None	99.31	99.54	99.23
None	4 个提示	98.82	99.31	98.68
2 个提示	2 个提示	99.43	99.60	99.36
4 个提示	**4 个提示**	**99.69**	**99.78**	**99.65**

结合两种模态（2 个文本 $+$ 2 视觉提示）可将总体准确率提升至 99.43%（比基线高 $+1.15\%$），已超越单模态配置，充分证明了空间与语义线索之间存在明显的协同作用。扩展至完整配置（4 个文本 $+$ 4 视觉提示）后，各项指标均达到最佳表现（总体准确率 $= 99.69\%$、平均精度 $=99.78\%$、卡帕系数 $= 99.65\%$），总体准确率、平均精度和卡帕系数分别提升了 $+1.41\%$、$+0.73\%$ 和 $+1.57\%$。随着提示数量增加，性能持续单调提升，这证实了更丰富的提示表示能够带来信息量更高的任务感知条件化。

这些结果验证了 TCSP 模块的设计。该模块通过带有可学习温度 τ 的交叉注意力机制，自适应地调节视觉与文本线索之间的融合强度，使模型能够利用互补信息，同时避免过度依赖任一模态。使用预先计算的 CLIP 嵌入可确保这种任务感知适配不会带来额外的

推理开销，在保持计算效率的同时，在不同训练设置下持续提升准确率。

8 结论

我们提出了 VP-Hype，这是一种采用视觉-文本提示增强的混合 Mamba-Transformer架构，用于高光谱图像分类。该方法将状态空间模型的线性复杂度与自注意力的表征能力协同结合，并通过任务感知的多模态提示系统加以增强，该系统融合了可学习空间模式与基于 CLIP 的语义嵌入。

这种混合架构解决了若干根本性挑战：Mamba 的线性复杂度能够高效处理高维光谱数据，同时保留长程依赖；注意力则捕捉细粒度的空间-光谱关系。视觉-文本提示系统结合可学习的视觉模板与语义文本引导，实现任务感知适配，在少样本场景中尤为有效。实验表明，视觉和文本组件能够带来互补优势，而 Cross 注意力融合可取得更优性能。该架构在多种高光谱数据集上展现出强大的泛化能力，在标准和极端数据稀缺条件下均持续达到最先进性能。

其局限性包括：文本描述固定，可能无法充分涵盖不同采集条件下的多样性；计算效率仍有优化空间，难以满足实时应用需求；目前侧重于分类任务，尚可扩展至异常检测、目标检测和光谱解混。该方法假设任务类别已知，因此在开放集场景中的有效性受限。

未来的研究方向包括开发自适应提示生成机制，探索适用于资源受限环境的高效架构变体，研究时间信息等其他模态，扩展至开放集分类，并加深对多模态提示融合机制的理论认识。

A VP-Hype总体框架

VP-Hype框架详见算法1，采用混合式Mamba–Transformer架构，旨在捕捉高光谱图像中固有的复杂空间-光谱依赖。该流程首先将输入数据立方体转换为图块嵌入，再通过基于CLIP的文本提示对其进行增强，以提供任务特定语义上下文。在每个层级阶段，模型交替使用MambaVisionMixer模块和窗口化自注意力，在全局建模效率与局部特征细化之间取得平衡。该方法的核心创新之一是文本条件软提示（TCSP）模块，它能够将文本先验动态融合到

视觉特征图中，以增强类别可区分性。最后，通过全局聚合和线性分类层生成最终预测结果。

B 模型规模比较

参数数量能够直观反映模型规模和复杂度。表 B.1 将 VP-Hype 与具有代表性的基线方法进行比较，涵盖了仅有数十万个参数的轻量级设计，以及拥有数亿个参数的大型 Transformer 模型。VP-Hype 处于经过审慎选择的中间位置：它比 ViT-Huge 等超大型 Transformer 紧凑得多，同时又比 Madanet 或 LANet 等轻量级架构具有更强的表达能力。值得注意的是，VP-Hype 的额外容量用于混合Mamba-Transformer骨干网络和双模态提示融合模块；这两者都直接提升了其分类性能，而非造成冗余的过度参数化。

表 B.1：代表性模型的参数数量比较，包括紧凑型 CNN、大规模 Transformer，以及所提出的 VP-Hype。VP-Hype 在模型表达能力与计算开销之间实现了良好的平衡。

模型	**数据集／设置**	**参数量（百万）**
Madanet [12]	Salinas	0.16
LANet [13]	Salinas	5.3
PiT-B [42]	龙口 (2%)	73.8
ViT-Huge [43]	龙口 (2%)	632.0
VP-Hype（我们的方法）	所有数据集	**77.30**

C 文本条件空间提示（TCSP）

本附录提供了主架构（第 3.1 节）中使用的文本条件空间提示（TCSP）模块的详细数学公式。TCSP 将任务条件文本线索与可学习视觉提示融合，生成空间对齐的引导图，并将其注入混合骨干网络。

C.1 文本提示投影

令$\mathbf{E}_{clip} \in \mathbb{R}^{T\times 512}$ 表示预先计算的 T 个任务描述的 CLIP 文本嵌入，并令$\mathbf{w}$ $\in \{0,1\}^{B\times T}$ 表示任务标识符的独热编码。任务条件文本嵌入的计算方式如下：

$$\mathbf{e}_t = \sum_{i=0}^{T-1} \mathbf{w}_{:,i} \odot \mathbf{E}_{clip}[i,:] \in \mathbb{R}^{B\times 512}. \tag{15}$$

算法1VP-Hype：结合视觉-文本提示

输入： 高光谱图像 $\mathbf{I} \in \mathbb{R}^{B\times C\times H\times W}$、任务标识符 t（可选），CLIP 文本嵌入$\mathbf{E}^{clip}$、可学习参数 Θ

输出： 类别逻辑值 $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times N}$

图块嵌入：
$\mathbf{X}_0 \leftarrow$补丁嵌入$(\mathbf{I})$ ▷公式（2）
文本提示编码：
$\mathbf{w} \leftarrow$独热编码(t)
$\mathbf{e}_t \leftarrow \sum_{i=0}^{T-1} \mathbf{w}_{:,\,i} \odot \mathbf{E}_{clip}[i,:]$ ▷ 公式（10）
对于 $l = 0$ **至** $L-1$，**执行**
  $\mathbf{X}_l \leftarrow$阶段输入$(\mathbf{X}_l)$
  对于 $i = 0$ **至** 深度$_l - 1$，**执行**
    如果 $i < \lfloor$深度$_l/2 \rfloor$ **则**
      $\mathbf{X}_l \leftarrow$MambaVisionMixer$(\mathbf{X}_l)$
    else
      $\mathbf{X}_l \leftarrow$窗口化自注意力$(\mathbf{X}_l)$
    结束如果
    $\mathbf{X}_l \leftarrow$多层感知机$(\mathbf{X}_l)$
  结束循环
  如果阶段 l 使用提示**则**
    TCSP 提示生成：
    $\mathbf{P}_l \leftarrow \mathrm{TCSP}(\mathbf{e}_t,\ \mathbf{P}_v)$ ▷ 公式（24）
    提示融合：
    $\mathbf{X}_l \leftarrow$ 融合$(\mathbf{X}_l,\ \mathbf{P}_l)$ ▷ 公式（12）
  结束条件
  如果 $l < L-1$ **则**
    $\mathbf{X}_{l+1} \leftarrow$下采样$(\mathbf{X}_l)$
  结束条件
结束循环
全局聚合：
$\mathbf{f}_{pool} \leftarrow$全局平均池化$(\mathbf{X}_{L-1})$ ▷ 公式（13）
分类：
$\hat{\mathbf{Y}} \leftarrow \mathbf{W}_{cls}\mathbf{f}_{pool} + \mathbf{b}_{cls}$ ▷ 公式（14）
返回 $\hat{\mathbf{Y}}$

该嵌入被投影到提示维度 d_p 并在空间维度上扩展：$\tilde{\mathbf{e}}_t = \mathbf{W}_{clip}\mathbf{e}_t \in \mathbb{R}^{B\times d_p}$, $\mathbf{T} = \mathrm{Interp}(\mathrm{Reshape}(\tilde{\mathbf{e}}_t;\ B, d_p, 1, 1), (S_p, S_p)) \in \mathbb{R}^{B\times d_p\times S_p\times S_p}$,其中 $\mathbf{W}_{clip} \in \mathbb{R}^{d_p\times 512}$ 是一个可学习的投影矩阵。 (16)

C.2 视觉提示参数化

可学习视觉提示由一个紧凑的空间张量表示

$$\mathbf{P}_v \in \mathbb{R}^{1\times d_p\times S_p\times S_p}, \tag{17}$$

该张量在训练和推理期间沿批次维度进行广播。

C.3 交叉注意力融合

文本提示 $\mathbf{T}$ 和视觉提示 $\mathbf{P}_v$ 通过交叉注意力进行融合。查询由文本提示生成，而键和值则由视觉提示生成：

$$\mathbf{Q} = \mathcal{P}_q(\mathbf{T}), \tag{18}$$
$$\mathbf{K} = \mathcal{P}_k(\mathbf{P}_v), \tag{19}$$
$$\mathbf{V} = \mathcal{P}_v(\mathbf{P}_v), \tag{20}$$

其中 $\mathcal{P}_q$、 $\mathcal{P}_k$和 $\mathcal{P}_v$ 是以轻量级卷积实现的可学习线性投影。

在将空间维度展平后，交叉注意力的计算如下

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_p}\,\tau}\right), \qquad \mathbf{U} = \mathbf{A}\mathbf{V}, \tag{21}$$

其中 $\tau > 0$ 是控制注意力锐度的可学习温度参数。

C.4 提示精炼与输出

注意力输出被重塑回空间形式，并由门控前馈网络进行精炼：

$$\mathbf{P}_{fused} = \mathrm{FFN}_g\big(\mathrm{Reshape}_{sp}(\mathbf{U})\big) \in \mathbb{R}^{B\times d_p\times S_p\times S_p}. \tag{22}$$

最后，将融合提示调整尺寸并投影，以匹配骨干网络特征几何结构：

$$\mathbf{P} = \mathbf{W}_{proj}\,\mathrm{Interp}(\mathbf{P}_{fused}, (H_f, W_f)) \in \mathbb{R}^{B\times C_f\times H_f\times W_f}, \tag{23}$$

其中，$\mathbf{W}^{proj}$ 通过轻量级 3×3 卷积实现。为便于正文中的简洁表示，上述操作概括为

$$\mathbf{P} = \mathrm{TCSP}(\mathbf{e}_t, \mathbf{P}_v). \tag{24}$$

表 B.2：Salinas数据集上的总体准确率、平均精度、卡帕系数和各类别精度（%）（10% 训练样本）。

类别	我们的方法	LoLA	MADANet	A^{2}MFE	HybridSN	SSFTT	HiT	LANet
总体准确率 (%)	**99.99**	99.87	99.17	98.56	97.05	98.61	97.83	96.67
平均精度 (%)	**99.99**	99.71	99.12	98.72	97.52	98.97	98.87	97.12
卡帕系数（%）	**99.99**	99.49	98.34	97.35	96.72	98.46	97.58	96.39
西兰花绿杂草1	**100.00**	100.00	99.64	99.46	99.30	99.34	100.00	98.32
西兰花绿杂草2	**100.00**	100.00	98.95	97.28	99.96	100.00	99.75	98.64
休耕地	**100.00**	100.00	98.90	97.99	98.16	99.89	100.00	95.09
粗犁休耕地	**100.00**	100.00	97.58	95.67	98.05	99.34	99.85	98.01
平滑休耕地	**99.96**	99.59	99.02	98.85	98.42	99.50	97.59	98.73
作物残茬	**100.00**	100.00	98.01	98.31	98.67	98.31	99.92	98.86
芹菜	**100.00**	100.00	99.34	95.76	98.96	99.91	100.00	97.04
未整枝葡萄	**100.00**	100.00	97.77	97.82	92.70	97.80	95.81	96.49
葡萄园土壤发育	**100.00**	100.00	98.23	97.56	98.53	100.00	100.00	93.11
玉米衰老期绿色杂草	**100.00**	99.80	95.91	95.23	95.44	99.81	98.19	94.27
罗马生菜（4周）	**100.00**	100.00	98.86	97.45	98.08	98.77	99.42	97.19
罗马生菜5周	**100.00**	100.00	99.23	98.34	99.33	99.89	99.84	98.42
罗马生菜6周	**100.00**	100.00	100.00	99.21	99.02	97.02	100.00	98.58
罗马生菜7周	**100.00**	100.00	99.34	98.12	98.21	98.96	99.42	98.07
未整形葡萄园	**100.00**	95.69	98.49	94.55	92.29	95.52	92.53	92.46
垂直棚架葡萄园	**100.00**	100.00	98.36	93.23	80.43	99.49	99.66	90.11

参考文献

[1] Chi Zhang、Mingjin Zhang、Yunsong Li、Xinbo Gao 和 Shi Qiu。用于高光谱图像超分辨率的差分曲率多维网络。*Remote Sensing*，13(17):3455，2021年。[2] M. Zhou、S. Wang 和 D. Li。基于高光谱成像的精准农业：综述。*Journal of Agricultural Remote Sensing*，1(1):1–15，2023。[3] Ying Li、Haokui Zhang 和 Qiang Shen。用于高光谱图像分类的三维深度学习。*IEEE*地球科学与遥感汇刊，55(11):6367–6376，2017。[4] Alexey Dosovitskiy 等人。图像胜过千言万语：用于大规模图像识别的 Transformer。发表于国际学习表征会议，2020年。[5] Fadi Abdeladhim Zidi、Djamel Eddine Boukhari、Abdellah Zakaria Sellam、Abdelkrim Ouafi、Cosimo Distante、Salah Eddine Bekhouche 和 Abdelmalik Taleb-Ahmed。LoLA-SpecViT：一种用于高光谱成像的局部注意力 SwiGLU 视觉 Transformer，结合了 LoRA。*International Journal of Applied Earth Observation and Geoinformation*，144:104924，2025。[6] Albert Gu 和 Tri Dao。Mamba：采用选择性状态空间的线性时间序列建模。*arXiv*预印本 *arXiv:2312.00752*，2023。

表 B.3：HongHu 数据集上的总体准确率、平均精度、卡帕系数和各类别精度（%）（10% 训练样本）。

类别	Ours	HSIMAE	HybridSN	ViT	MASSFormer	DBCT	LSGA	
总体准确率（%）	**99.64**	96.24	96.76	96.76	98.14	99.14	99.02	
平均精度（%）	**99.23**	92.99	92.50	92.50	96.25	98.07	97.54	
卡帕系数（%）	**99.55**	96.18	96.77	96.77	97.65	98.91	98.75	
红色屋顶	**99.84**	97.40	98.46	97.84	98.99	98.27	99.54	
Road	**99.71**	88.87	75.45	87.22	97.25	97.44	96.82	
裸土	**99.59**	95.84	95.61	94.98	98.87	99.01	98.23	
棉花	**100.00**	98.13	99.40	98.48	99.31	99.69	99.65	
棉柴	**99.62**	78.73	88.93	84.61	94.76	98.32	99.14	
Rape	**99.84**	98.81	98.41	98.60	99.11	99.47	99.34	
大白菜	**99.50**	94.10	95.77	92.50	97.15	98.81	98.01	
小白菜	**99.23**	87.53	93.72	87.50	94.16	98.38	96.59	
卷心菜	**99.82**	98.14	97.91	97.91	99.77	99.71	99.82	
根用芥菜	**99.48**	94.97	93.73	92.90	98.03	98.53	99.94	
菜心	**99.55**	91.95	91.88	88.10	95.53	98.77	97.75	
青菜	**98.46**	92.80	92.17	91.23	98.15	98.42	96.98	
小青菜	**99.56**	92.68	94.22	91.70	94.73	98.48	97.63	
莴苣	**99.24**	95.31	92.83	91.74	98.48	99.02	98.90	
莴笋	**100.00**	94.56	87.59	94.02	98.51	98.18	97.96	
覆膜生菜	**99.98**	98.14	97.30	97.69	97.03	98.59	99.16	
罗马生菜	**99.93**	96.94	92.21	95.59	98.86	99.47	99.32	
胡萝卜	**96.86**	90.37	89.72	93.51	95.90	95.16	96.56	
白萝卜	**99.63**	95.63	91.73	91.34	95.61	90.11	99.51	
蒜苗	**99.68**	88.67	89.22	86.62	93.85	99.00	97.48	
蚕豆	**96.65**	85.15	83.35	84.24	87.52	90.77	91.62	
Tree	**99.97**	91.16	95.38	88.90	88.02	96.76	98.08	

[7] Menglin Jia、Luming Tang、Bor-Chun Chen、Claire Cardie、Serge Belongie、Bharath Hariharan 和 Ser-Nam Lim。视觉提示调优。载于欧洲计算机视觉会议，第 709–727 页。施普林格，2022。[8] Zhehui Wu、Yong Chen、Naoto Yokoya 和 Wei He。MP-HSIR：一种用于通用高光谱图像复原的多提示框架。载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，第 13009–13020 页，2025。[9] H. Yue、Y. Ma 和 Q. Zhang。用于高光谱图像分类的自监督知识蒸馏。*ISPRS*摄影测量与遥感学报，183：169–183，2022。[10]Swalpa Kumar Roy、Gopal Krishna、Shiv Ram Dubey 和 Bidyut B Chaudhuri。HybridSN：探索用于高光谱图像分类的三维-二维卷积神经网络特征层次结构。*IEEE*地球科学与遥感快报，17(2)：277–281，2019。[11] J. Yang、B. Du、C. Wu 和 L. Zhang。用于高光谱图像分类的自动可调多尺度特征提取框架。载于*IEEE*国际地球科学与遥感研讨会（*IGARSS*）论文集，第 3649–3652 页，比利时布鲁塞尔，2021年。[12]Binge Cui、Jiaxiang Wen、Xiukai Song 和 Jianlong He。MADANet：一种轻量级高光谱图像分类网络，采用多尺度特征聚合和双注意力机制。*Remote Sensing*，15(21)：5222，2023。[13] L. Ding、H. Tang 和 L. Bruzzone。LANet：局部注意力嵌入，用于提升遥感图像的语义分割效果。*IEEE*地球科学与遥感汇刊，59：426–435，2021。[14] J. Ma、L. Zhao 和 H. Xu。AS2MLP：用于高光谱图像分类的自适应空间移位多层感知机。*IEEE*神经网络与学习系统汇刊，35(4)：2105–2117，2024。[15]Damian Ibanez、Ruben Fernandez-Beltran、Filiberto Pla 和 Naoto Yokoya。用于高光谱图像分类的掩码自编码光谱-空间变换器。*IEEE*地球科学与遥感汇刊，60：1–14，2022。[16] Jing Bai、Zichen Zhou、Zheng Chen、Zhu Xiao、Erlong Wei、Yihong Wen 和 焦李成。利用自监督学习进行高光谱图像分类的跨数据集模型训练。*IEEE*地球科学与遥感汇刊，2024。[17] L. Sun、H. Zhang、Y. Zheng、Z. Wu、Z. Ye 和 H. Zhao。MASSFormer：用于高光谱图像分类的记忆增强光谱-空间变换器。*IEEE*地球科学与遥感汇刊，60，2024。

[18] Z. Zhou、Y. Li 和 X. Zhang。Dict：以字典为中心的 Transformer，用于高效高光谱图像分类。《模式识别》，137：109371，2023。[19] R. Xu、X.-M. Dong、W. Li、J. Peng、W. Sun 和 Y. Xu。Dbctnet：用于高光谱图像分类的双分支卷积 Transformer 网络。*IEEE*地球科学与遥感汇刊，62，2024。[20] J. Z. T. Arshad 和 I. Ullah。用于高光谱图像分类的混合卷积 Transformer。《欧洲遥感杂志》，0(0)：2330979，2024。[21] M. Ahmad 等人。SSMamba 和 MorpMamba：用于高光谱成像的结构化状态空间模块。《遥感快报》，2025。[22]T. Zhang、R. Liu 和 Y. Wang。Lds2mlp：用于高光谱分类的可学习膨胀光谱-空间多层感知机。《应用地球观测与遥感精选主题期刊》，17：125–136，2024。[23] Musheer Khan、Qian Du、Xing Sun 和 Dong Wang。Groupformer：用于高光谱图像分类的光谱-空间特征提取组注意力 Transformer。《科学报告》，14：5261，2024。[24] Haoyang Fu、Chang Wang 和 Zhe Chen。利用基于改进 CBAM 的网络自适应选择高光谱图像分类的光谱-空间特征。《神经计算》，416：128877，2025。[25] N. Houlsby、A. Giurgiu、S. Jastrzębski、B. Morrone、Q. De Laroussilhe、A. Gesmundo、M. Attariyan 和 S. Gelly。面向自然语言处理的参数高效迁移学习。载于第*36*届国际机器学习会议论文集，第 97 卷，第 2790–2799 页，2019。[26] X. Jia、Y. Yin、K. Song、A. Golla、A. Lal、F. Huang 和 A. Anandkumar。视觉提示调优。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7092–7101 页，2022。[27] A. Ben Zaken、Y. Goldberg 和 A. Globerson。BitFit：针对基于 Transformer 的掩码语言模型的简单参数高效微调。计算语言学协会汇刊，10：496–512，2022。[28] Edward J Hu、Yushun Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。载于国际学习表征会议（*ICLR*），2022。[29] Vaishnav Potlapalli、Syed Waqas Zamir、Salman Khan 和 Fahad Shahbaz Khan。PromptIR：通过提示实现一体化盲图像复原。载于神经信息处理系统进展（*NeurIPS*），第 36 卷，2023。

[30] Bin Li、Jincheng Lu、Jun Wang、Wei Zhang 和 Li Zhao。Pip：用于通用图像复原的提示中提示学习。载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2024年。[31] Marcos V Conde、Gregor Geigle 和 Radu Timofte。Instructir：遵循人类指令进行高质量图像复原。载于欧洲计算机视觉会议（*ECCV*）论文集，2024年。[32] Chia-Ming Lee、Ching-Heng Cheng、Yu-Fan Lin、Yi-Ching Cheng、Wo-Ting Liao、Chih-Chung Hsu、Fu-En Yang 和 Yu-Chiang Frank Wang。Prompthsi：通过视觉语言调制的频率适配实现通用高光谱图像复原。*arXiv*预印本 *arXiv:2411.15922*，2024年。[33] 武汉大学 RSIDEA团队。WHU-Hi-LongKou数据集。在线获取：http://rsidea.whu.edu.cn，2018年。[34] 武汉大学 RSIDEA团队。WHU-Hi-HongHu数据集。在线获取：https://huggingface.co/datasets/rsidea/WHU-Hi-HongHu，2017年。[35] CCWintco（Elhami等人）。萨利纳斯山谷AVIRIS高光谱数据集。在线获取：https://www.ehu.eus/ccwintco/index.php/Hyperspectral_遥感__场景，2004年。[36] S.K. Roy、G. Krishna、S.R. Dubey 和 B.B. Chaudhuri。HybridSN：探索用于高光谱图像分类的3D-2D CNN特征层级。*IEEE*地球科学与遥感快报，17：277–281，2020年。[37] L. Sun、G. Zhao、Y. Zheng 和 Z. Wu。用于高光谱图像分类的光谱—空间特征词元化变换器。*IEEE*地球科学与遥感汇刊，60(1)：1–14，2022年。[38] X. Yang、W. Cao、Y. Lu 和 Y. Zhou。高光谱图像变换器分类网络。*IEEE*地球科学与遥感汇刊，60(1)：1–15，2022年。[39] Maria Ahmad、Umer Ghous、Muhammad Usama 和 Manuel Mazzara。Waveformer：用于高光谱图像分类的光谱—空间小波变换器。*IEEE*地球科学与遥感快报，21：1–5，2024年。[40] Maria Ahmad、Umer Ghous、Muhammad Usama 和 Manuel Mazzara。Memformer：用于高光谱图像分类的动态记忆增强变换器。*arXiv*预印本 *arXiv:2504.13242*，2025年。[41]Tingrui Hu、Peichao Gao、Sijing Ye 和 Shi Shen。基于波段子空间划分的改进SR-SSIM波段选择方法。*Remote Sensing*，15(14)：3596，2023年。

[42] Byeongho Heo、Sangdoo Yun、Dongyoon Han、Sanghyuk Chun、Junsuk Choe 和 Seong Joon Oh。重新思考视觉变换器的空间维度。载于*IEEE/CVF*国际计算机视觉会议论文集，第11936–11945页，2021年。[43] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby。一张图像胜过16×16个词：面向大规模图像识别的Transformer。载于国际学习表征会议（*ICLR*）论文集，2021年。