arXiv:2305.02086v2 [cs.CV] 8 Sep 2023

重新审视卫星图像时间序列的编码

蔡鑫、毕雅欣、彼得·尼科尔和罗伊·斯特里特

阿尔斯特大学计算机学院，英国贝尔法斯特{cai-x, y.bi, p.nicholl, r.sterritt}@ulster.ac.uk

摘要。 卫星图像时间序列（SITS）表征学习十分复杂，这是由于其具有高时空分辨率、不规则的采集时间以及复杂的时空交互。这些挑战促使人们为卫星图像时间序列分析设计专用的神经网络架构。该领域的先驱研究者已取得令人鼓舞的成果，但由于现有表征学习框架仍不够理想，将计算机视觉（CV）领域的最新进展或成熟范式迁移至卫星图像时间序列仍然极具挑战。受近期采用基于查询的 Transformer 解码器来简化目标检测或图像分割流程这一趋势的启发，本文提出一种全新的视角，将卫星图像时间序列处理视为直接集合预测问题。我们进一步提出将卫星图像时间序列的表征学习过程分解为三个明确的步骤：收集–更新–分发。这种方法计算效率高，适用于不规则采样且异步的卫星时间观测。得益于这一独特的重新表述，我们提出的卫星图像时间序列时间学习骨干网络，先在资源高效的像素集合格式上进行预训练，再针对下游密集预测任务进行微调，已在 PASTIS基准数据集上取得新的最先进结果。具体而言，在卫星图像时间序列的语义/全景分割流程中，时间组件与空间组件的明确分离使我们能够利用计算机视觉领域的最新进展，例如通用图像分割架构；与迄今为止报告的最佳成绩相比，平均交并比分别提升了 2.5 个百分点，PQ 提升了 8.8 个百分点。

1 引言

近年来，人们对利用重访频率较高的卫星（如欧洲航天局（ESA）的 Sentinel 卫星）自动监测地表的兴趣日益浓厚。尤其是，大规模自动化作物类型制图最能受益于对卫星图像时间序列（SITS）中复杂时间动态的利用，这有助于公平分配农业补贴，并规范农民所采用的最佳耕作实践。然而，利用深度学习模型从 SITS 中提取具有代表性的特征并非易事。例如，一些仅简单拼接空间编码器和时间

本文已被 BMVC 2023 接收。

编码器的模型，甚至难以超越随机森林分类器 [16], 的性能，这迫使研究人员投入大量精力开发专门的神经网络架构。

开创性的 PSE+TAE [9]/PSE+L-TAE [7] 为卫星图像时间序列提出了一种颇具前景的学习范式：首先，利用多层感知机（MLP），对农田地块空间范围内的光谱值进行统计汇总。多层感知机独立处理无序像素集合。随后，将汇总后的空间特征输入采用自注意力机制的时间编码器，以揭示潜在的时间模式，遵循先空间后时间的分解顺序。然而，近期研究 TSViT[34], 提供了实证依据，并指出，对于卫星图像时间序列分析而言，先时间后空间的分解顺序是更直观的设计选择，因为与高分辨率光学影像相比，中分辨率卫星影像中的空间上下文信息缺乏判别力，尤其是在植被监测或作物类型制图任务中。这一研究方向揭示了设计卫星图像时间序列深度学习模型时的一个重要方面：将学习框架解耦为在空间和时间上彼此分离的组件。然而，由于缺乏对不同输入格式（即像素集合或图像序列格式）的灵活处理能力，PSE+TAE 或 TSViT 的应用受到限制。因此，计算机视觉领域的经典预训练-微调范式，即在大型数据集（例如 ImageNet [3]）上通过全监督/自监督学习 [5,10]对分类模型进行预训练，再针对目标检测 [27] 或语义分割 [19], 等下游任务进行微调，尚未成功应用于卫星图像时间序列分析。

与此同时，正如先前研究所指出的 [7, 9], ，有效学习卫星图像时间序列的表征面临的另一项重大挑战，是捕捉作物物候中的复杂时间动态，即植物事件发生的精确时间对于区分不同作物类型至关重要 [25]。然而，近期有关卫星图像时间序列分析的研究 [7–9, 25] 主张采用自注意力 [36] 作为核心计算单元，却未质疑其用于时间建模的有效性，尤其是考虑到其置换不变性。基于时间序列预测领域的最新发现 [41, 45],，由于数值时间序列数据缺乏丰富的语义，自注意力操作对复杂时间关系进行建模的能力被夸大了。对时间顺序具有强内置先验或归纳偏置的模块，例如经典的指数平滑 [41]或频率分析方法 [46]，已被证明在时间模式提取方面优于标准的自注意力机制。但卫星图像序列中普遍存在的时间轴不规则性——例如云层遮挡导致的光学影像采集缺失——使问题更加复杂，通常需要将插补或插值作为预处理步骤 [16] ，或者开发一种端到端学习框架，以协调插值与分类之间可能相互冲突的优化目标 [32] 。除了近期受到质疑的自注意力用于时间建模的有效性之外，相对于所处理序列长度而言，其二次方的空间和时间复杂度还会给模型设计带来额外的计算方面的顾虑，并限制其在卫星图像时间序列密集预测任务中的应用 [8, 34]。

这两点观察促使我们重新思考现有的卫星图像时间序列编码方案：我们真的需要为卫星图像时间序列开发专用的神经网络架构吗？能否通过一种简单而通用的表征学习框架，将成熟的计算机视觉范式适配到卫星图像时间序列中？具体而言，我们提出将卫星图像时间序列表示为观测集合，这一思路受到 [11] 提出的用于分类不规则采样的异步时间序列的表述方式启发，其中每个元素由其光谱特征以及静态或动态协变量表示，例如日历时间或热量时间 [25]。借助这一独特视角，我们提出了一种简单而有效的表征学习框架，称为 Exchanger，通过将编码过程分解为三个步骤——收集–更新–分发——来处理卫星图像时间序列，从而避免使用自注意力并规避其局限性。我们只需将提出的 Exchanger 与计算机视觉领域常用的分割模型相连接，便首次展示了：在像素集合格式数据集上对分类模型进行预训练，再将其微调到以图像序列为输入的下游密集预测任务上，即可在 PASTIS [8] 上取得相较于高度专用网络架构的全新最先进性能。此外，我们还可以直接将最新的通用图像分割架构 Mask2Former [2] 引入卫星图像时间序列的语义分割/全景分割，而无需进行任何修改，只需让其接收由 Exchanger 编码的输出特征图，即可显著超越此前的最先进模型。总而言之，本文的贡献包括：

- 重新定义卫星图像时间序列表示，将其视为实例集合，消除模型设计上的限制，以适应卫星图像时间序列的不同输入数据格式。这使我们能够利用资源高效的像素集合格式进行预训练，随后在下游密集预测任务上进行微调。我们认为，这是将计算机视觉中的预训练-微调范式引入卫星图像时间序列的一种更理想的方式。
- 明确地将卫星图像时间序列的表征学习过程分解为三个步骤：收集–更新–分发，由此构建出一个概念清晰、计算高效的学习框架，称为 Exchanger，用于卫星图像时间序列的通用特征提取。
- 与现有工作在密集预测流程中将时间组件和空间组件错综复杂地交织在一起不同，我们认为，明确分离时间编码器和空间编码器能够大幅降低模型设计的复杂度，并有助于利用计算机视觉领域的最新进展，从而缩小计算机视觉与卫星图像时间序列之间的差距。
- 通过开展大量实验，我们验证了所提出模型的有效性。该模型在 PASTIS基准数据集的语义分割和全景分割任务中均大幅超越了此前的最先进模型。

2 相关工作

卫星图像时间序列的编码 卫星的高频重访周期使得人们能够利用丰富的时间动态信息，以进行作物类型制图或植被监测。传统机器学习方法 [37] 依赖人工设计的特征，尽管这需要大量领域专业知识，但特征编码问题却一直未得到妥善解决。近年来，微分神经网络架构已在该领域占据主导地位。具体而言，卷积神经网络（CNNs） [26]和循环神经网络（RNNs） [30] 分别被采用为空间特征和时间特征编码的事实标准选择。此外，研究人员还提出了卷积循环混合模型 [29] ，将卫星图像时间序列视为时空信号进行处理。尽管这些早期方法取得了令人鼓舞的成果，但它们忽略了自然图像/视频与卫星图像时间序列之间的重要差异。开创性的工作PSE+TAE [9] 提出使用多层感知机来汇总空间统计量，以弥补中分辨率哨兵-2号图像缺乏丰富空间语义的问题，并使用自注意力对时间模式进行编码。随后提出的PSE+L-TAE [7]则采用轻量级Transformer解码器来提取时间特征。像素集编码器（PSE）在处理地块几何形状的不规则性方面尤为有效，它将地块表示从 $T \times C \times H \times W$ 简化为 $T \times C \times N$，其中T表示时间序列的长度，C表示通道数，H/W分别表示高度/宽度，N表示像素数量，因此与图块格式相比，该方法所需的存储空间显著更少 [9] 。然而，对于下游密集预测任务，如先前最先进的模型U-TAE [8], 所示，TAE需要以复杂的方式集成到空间编码器中，这阻碍了成功的预训练-微调范式的复现。TSViT [34]是通过在视觉Transformer（ViT） [4], 中引入一种独特的归纳偏置来弥合卫星图像时间序列分析与计算机视觉（CV）之间差距的首次尝试。该归纳偏置采用先时间后空间的分解方式，其依据是空间上下文为作物类型识别提供的边际信息有限。然而，ViT中的图块标记化方案本质上是为图像设计的，因此TSViT无法直接处理无序的像素集合格式，而这种格式对于卫星图像时间序列分类和预训练而言更加高效。此外，由于整个时间学习过程中始终保留空间维度，自注意力所需的大量计算开销进一步加剧，使得TSViT难以应用于密集预测任务。

3 所提方法

本节首先将卫星图像时间序列（SITS）的表示重新表述为观测集合，以区别于传统的时空信号。随后，我们简化当前的 SITS 编码流程，不再需要专门考虑空间维度，并进一步将时间特征学习过程分解为三个明确的步骤：收集–更新–分发。具体的网络实现留待补充材料介绍。

定义 1. 我们将特定地理参考位置、具有一定空间范围的卫星图像序列描述为一组 $\mathfrak{S}_i$ 实例/集合$\mathfrak{S}_i = \left\{\boldsymbol{S}^1, \ldots, \boldsymbol{S}^n\right\}$，其中每个实例/集合 $\boldsymbol{S}^j$ 由一组时间采集数据 $\boldsymbol{S}^j = \left\{\boldsymbol{s}^j_{t_1}, \ldots, \boldsymbol{s}^j_{t_m}\right\}$构成。我们假设每个观测 $\boldsymbol{s}^j_{t_k}$ 由 $\left[\boldsymbol{v}^j_{t_k}, \boldsymbol{p}^j_{t_k}, \odot\right]$表示，其中 $\boldsymbol{v}^j_{t_k}$是传感器测量值的特征嵌入，$\boldsymbol{p}^j_{t_k}$ 是特定采集时间的时间位置嵌入，而 $[\cdot]$ 则充当其他静态或动态协变量的占位符，例如几何边界或模态信息，从而为实现 *SITS* 的通用表示创造了可能。 $[\cdot]$ 表示一个任意算子，用于混合其中包含的特征，例如求和或拼接。请注意， $\boldsymbol{s}^j_{t_k}$ 的上标和下标分别表示空间标识符和时间标识符；为避免符号过于繁杂，我们省略用于区分地块的索引 i 。

与通常将卫星观测表示为时空信号$\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{T \times C \times H \times W}$的做法不同，我们放宽了规则网格对空间维度施加的约束，因为空间结构先验对于卫星图像时间序列处理并非不可或缺 [1]，而且还会进一步限制模型设计的灵活性。我们认为，应当更加重视时间维度，而空间信息的聚合则可根据不同任务的输出要求灵活处理。通过这种更具普适性的重新表述，卫星图像时间序列的分类问题与多实例学习（MIL） [12] 密切相关：在多实例学习中，单个类别标签被赋予一组实例，这些实例之间没有顺序关系或强依赖关系。也就是说，将从地块内不同子位置采样的每条观测时间序列视为独立实例，并为其赋予不同权重，以共同决定最终的包级分类结果。对于密集预测问题，保留规则网格排列仅仅是为了匹配所需的输出格式，而非用于挖掘高层空间语义。我们在实验中还观察到，先汇总卫星图像时间序列的时间信息，再接入成熟的语义分割模型（如 U-Net [28] ），其性能优于专为卫星图像时间序列设计的分割网络（如 U-TAE [8], ）。这表明，对卫星图像时间序列进行时间处理后会涌现丰富的语义信息，也印证了 TSViT [34]所倡导的先时间、后空间的因子分解顺序。

3.1 时间上下文簇

得益于我们重新构建的卫星图像时间序列表示，由于空间依赖关系较弱，空间建模未被纳入卫星图像时间序列表征学习流程。

[1] 请注意，我们将该假设限定于基于卫星图像时间序列进行作物类型制图或植被监测。如 [13], 所示，可利用空间邻近性对卫星影像进行对比式表征学习。此外，特定地表覆盖识别（如建筑物轮廓提取）主要依赖单时相但高分辨率的影像 [6]。

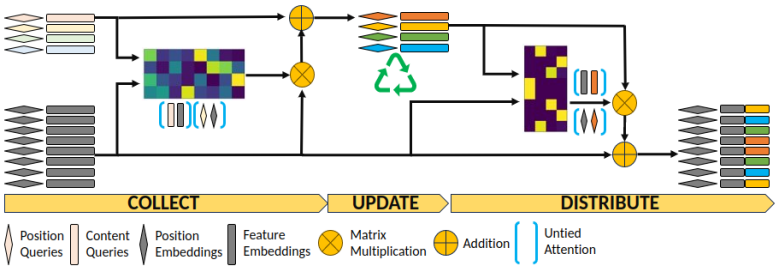


图1：所提出的用于卫星图像时间序列通用表征学习的收集–更新–分发过程示意图。

对于密集预测任务，在对 SITS 进行时间特征提取后，接入语义分割模型即可挖掘高层语义。这大大简化了现有 SITS 密集预测模型的设计；在这些模型中，时间编码组件与空间编码组件错综复杂地交织在一起。受时间序列分析中以其他时间建模模块替代自注意力所取得的成功启发 [41, 46], ，考虑到 SITS 获取时间不规则性带来的额外复杂性，我们提出使用一组可学习查询作为外部记忆模块，与输入交换时间信息，并因此将我们的模型命名为“Exchanger”。

形式上，我们将卫星图像时间序列的表征学习过程提炼为三个步骤：收集–更新–分发，如图1所示；借助一组时间上下文簇，该过程进一步拆分为两个组件：内容查询和位置查询： $\boldsymbol{C}^v \in \mathbb{R}^{N\times d}$， $\boldsymbol{C}^p \in \mathbb{R}^{N\times d}$ 以避免彼此污染，其中 N 是簇的数量。

▷ **收集** 给定输入特征嵌入 $\boldsymbol{V} \in \mathbb{R}^{T\times d}$ 和时间位置嵌入 $\boldsymbol{P} \in \mathbb{R}^{T\times d}$，时间簇 $\boldsymbol{C}^v$ 首先通过计算成对相似度，从特征嵌入 $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T]$ 中收集信息，随后使用选择函数 $\mathcal{S}$ 过滤掉最不重要的信息，其形式化表示如下：

$$\begin{aligned} \boldsymbol{A}_1 &= \text{cal_simlarity}\left([\boldsymbol{C}^v, \boldsymbol{V}], [\boldsymbol{C}^p, \boldsymbol{P}]\right) \\ \boldsymbol{W} &= \mathcal{S}\left(\boldsymbol{A}_1\right) \\ \boldsymbol{C}^v &= \boldsymbol{C}^v + \boldsymbol{W}\boldsymbol{V} \end{aligned} \tag{1}$$

其中， $\boldsymbol{A}_1 \in \mathbb{R}^{N\times T}$ 是亲和矩阵，并通过选择函数 $\mathcal{S}$ 进一步细化，以获得 $\boldsymbol{W}$ ，再与 $\boldsymbol{V}$相乘，从而完成收集过程。

▷ **更新** 随后，仅依靠 $\boldsymbol{C}^v, \boldsymbol{C}^p$更新时间簇，以便在不同时间片段之间实现全局信息交换，公式如下：

$$\boldsymbol{C}^v = \text{Update}\left(\boldsymbol{C}^v, \boldsymbol{C}^p\right) \tag{2}$$

▷ **分配** 更新簇后，通过以硬或软的方式将每个时间元素 $\boldsymbol{v}_i$ 分配给 $\boldsymbol{C}_j^v$，将更稳健且更具代表性的时间上下文簇特征分配回去，公式如下：

$$\begin{aligned} \boldsymbol{A}_2 &= \text{cal_simlarity}\left([\boldsymbol{V}, \boldsymbol{C}^v], [\boldsymbol{P}, \boldsymbol{C}^p]\right) \\ \boldsymbol{I} &= \text{assign}\left(\boldsymbol{A}_2\right) \\ \boldsymbol{V} &= \boldsymbol{V} + \boldsymbol{I}\boldsymbol{C}^v \end{aligned} \tag{3}$$

其中，$\boldsymbol{A}_2 \in \mathbb{R}^{T\times N}$是亲和矩阵，$\boldsymbol{I} \in \mathbb{R}^{T\times N}$的每一行都包含一个硬索引或软概率向量，用于指示每个时间元素 $\boldsymbol{v}_i$ 所属的时间上下文簇。

所提出的时间表示学习范式“收集–更新–分发”在处理时间序列数据的不规则性和不同步问题时尤为有效，因为它不需要任何先验假设，例如按顺序处理时间观测。每个时间元素的特征都可以通过与时间上下文簇交互来更新，而不同时间片段之间的信息流则通过上下文簇之间的通信实现，这是一种计算效率更高的信息交换方式。与自注意力$\mathcal{O}\left(T^2d\right)$的计算复杂度相比，该方法只需 $\mathcal{O}($ $NTd$$)$，其中 $N \ll T$，因此随着时间令牌数量的增加，其扩展性要好得多。更重要的是，所提出的卫星图像时间序列表征学习框架可视为当前基于自注意力的模型（如 L-TAE [7]或 TSViT [34]）的泛化。具体而言，L-TAE [7]是一种轻量级 Transformer解码器，它使用一组可学习查询从空间编码器的输出中提取关键特征，这与我们提出的收集步骤相对应。由于缺少更新和分发步骤，L-TAE 在编码方面的效果较差，因为它没有实现特征更新机制。TSViT [34]的时间编码器会在输入时间元素前添加一组类别令牌，并依靠自注意力进行特征学习；这可以看作是我们所提框架的一个特例，其中收集–更新–分发步骤通过自注意力隐式实现。添加的外部令牌与输入时间元素彼此同步通信，相较于我们提出的分解方案，这种方式计算开销更大，且概念上也较为模糊。

3.2 网络实例化

鉴于卫星图像时间序列重构方式的灵活性，以及所提出的收集—更新—分发学习流程的多样性，我们选择借鉴计算机视觉领域的最新进展：在这些工作中，Transformer解码器中的对象查询被重新诠释为簇中心，交叉注意力则被重新表述为一种聚类操作 [2,33,42,44], ，由此重拾了经典观点，即将图像分割视为像素分组过程，而非逐像素分类。由于聚类本质上是一种量化过程，其中冗余信息会逐步

被滤除，因而抽象概念或高层语义可能由此涌现，因此它有望用于通用表征学习，而不局限于图像分割任务，近期的开创性工作 [22, 43]已证明了这一点。本文的主要目标是为卫星图像时间序列建立有效的表征学习框架，因此我们决定借用GPViT中的核心构建单元Group Propagation Block（GP块） [43] 来实现这一构想，将架构创新留待未来工作。为完整起见，我们仅简要介绍GP块的构造如下，并请读者参阅原作 [43]以了解具体细节：

$$\boldsymbol{C}^v = \mathrm{Concat}_h\left(\mathrm{Softmax}\left(\frac{1}{\sqrt{2d}}\boldsymbol{C}^v\boldsymbol{W}_h^Q\left(\boldsymbol{V}\boldsymbol{W}_h^K\right)^T + \frac{1}{\sqrt{2d}}\boldsymbol{C}^p\boldsymbol{U}_h^Q\left(\boldsymbol{P}\boldsymbol{U}_h^K\right)^T\right)\boldsymbol{V}\boldsymbol{W}_h^V\right) \tag{4}$$

其中，$\boldsymbol{W}^{Q,K,V,h}$ 和 $\boldsymbol{U}^{Q,K,h}$ 分别是内容和位置嵌入的投影矩阵。式（4）通过交叉注意力实现收集过程：亲和矩阵通过缩放点积计算，并使用Softmax函数选择最相关的时间元素。

$$\begin{aligned}\boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_1\left(\mathrm{LayerNorm}\left(\boldsymbol{C}^v\right)^T\right)^T \\ \boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_2\left(\mathrm{LayerNorm}\left(\boldsymbol{C}^v\right)\right)\end{aligned} \tag{5}$$

式（5）通过使用 MLPMixer [35] 实现上下文簇更新，其中一个多层感知机沿 token 维度运行，另一个多层感知机沿通道维度运行。

$$\begin{aligned}\boldsymbol{Z} &= \mathrm{Concat}_h\left(\mathrm{Softmax}\left(\frac{1}{\sqrt{2d}}\boldsymbol{V}\tilde{\boldsymbol{W}}_h^Q\left(\boldsymbol{C}^v\tilde{\boldsymbol{W}}_h^K\right)^T + \frac{1}{\sqrt{2d}}\boldsymbol{P}\tilde{\boldsymbol{U}}_h^Q\left(\boldsymbol{C}^p\tilde{\boldsymbol{U}}_h^K\right)^T\right)\boldsymbol{C}^v\tilde{\boldsymbol{W}}_h^V\right) \\ \boldsymbol{Z}' &= \mathrm{Concat}\left(\boldsymbol{Z}, \boldsymbol{V}\right)\tilde{\boldsymbol{W}}_{proj} \\ \boldsymbol{V}' &= \boldsymbol{Z}' + \mathrm{FFN}\left(\boldsymbol{Z}'\right)\end{aligned} \tag{6}$$

其中，$\tilde{\boldsymbol{W}}^{Q,K,V,h}$ 和 $\tilde{\boldsymbol{U}}^{Q,K,h}$ 是另一组用于内容~和位置嵌入的投影矩阵；$\boldsymbol{W}_{proj}$ 用于将拼接后的特征线性投影到与输入相同的维度，而 FFN 是前馈神经网络。式（6）通过将输入时间元素用作查询，从更新后的上下文簇中收集信息，并沿相反方向执行交叉注意力，从而实现分配过程。

4 实验

在本节中，我们通过大量消融研究验证所提出的卫星图像时间序列表征学习框架的有效性，并

将其与此前最先进模型在语义分割和全景分割任务上的表现进行比较。请注意，具体实现细节见补充材料。代码已公开发布于 https: // github.com/TotalVariation/Exchanger4SITS。

4.1 数据集

我们选择 PASTIS（Panoptic Agricultural Satellite TIme Series）[2] 基准数据集 [8]来评估所提出模型的性能，并与此前的最先进模型进行比较。该数据集包含 2433 个形状为 $10 \times 128 \times 128$ 的多光谱图像序列，每个序列包含采集于 2018 年 9 月至 2019 年 11 月期间的时间观测，序列长度各不相同，介于 38 至 61 之间，总计超过 20 亿个像素。此外，PASTIS 覆盖法国四个气候和作物分布各异的地区，跨度超过 4000 km^2 ，包括 18 种作物类型以及一个背景类别。除具有高质量语义和全景标注的时空格式 $T \times C \times H \times W$、超过 120，000 个边界框和像素级精确掩码外，该数据集还附带一个用于基于地块的作物类型分类的像素集合格式 $T \times C \times N$数据集 [9]。我们主要使用 PASTIS 官方提供的 5 折划分，以开展广泛的消融研究和模型性能评估，并另外报告在另一个数据集 MTLCC [30]上的语义分割结果。MTLCC 数据集覆盖慕尼黑以北 102 km $\times 42$km 的德国大面积感兴趣区域（AOI），包含 17 个不同的作物类别，以及在 2016 年和 2017 年两个生长季中采集的、长度分别为 46 和 52 的两组时间观测[3]。

4.2 实现细节

4.3 分类

我们在 PASTISpixelset 格式的数据集上训练和验证分类模型。基于 [31,39] 的观察结果，即额外的多层感知机投影器有助于缩小无监督预训练与监督式预训练之间的迁移差距，我们在特征提取器 Exchanger 后添加 t-ReX [31] 提出的投影器，并使用余弦 softmax 交叉熵损失。我们使用 AdamW [20]优化器，批量大小为 128，权重衰减为 0.005，初始学习率为 0.0002，并采用阶梯式学习率调度器，在总训练步数的 0.7 和 0.9 处将学习率衰减 10 倍，在 4 张 V100 GPU 上训练模型 50 个 epoch。作为一种数据增强策略，我们从 0.2 到 0.4 的区间内均匀采样，随机丢弃时间观测，以抵消云层遮挡带来的不利影响；语义分割和全景分割模型的训练也采用了这一策略。

[2] https://github.com/VSainteuf/pastis-benchmark[3] 请注意，MTLCC 中的单个样本具有有限的空间分辨率 24×24。

4.4 语义与全景分割

随后，我们使用预训练模型初始化 Exchanger，除非另有说明，否则它在语义/全景分割流程中充当时间编码器。对于用作空间编码器的 Unet [28]，我们使用 AdamW [20]优化器，批量大小为 4，权重衰减为 0.005，初始学习率为 0.0002，并采用多项式衰减学习率调度器，在 4 张 V100 GPU 上训练模型 100 个 epoch；语义分割使用 Focal 交叉熵损失 [18]，全景分割则使用 Parcels-as-Points (PaPs) 预测头和 PaPs 损失 [8]。由于空间分辨率为 128×128 且时间长度超过 30 的单个卫星图像时间序列样本无法装入配备 16G 内存的 V100 GPU，我们在训练时进行随机裁剪，裁剪尺寸为 32×32，并在 A100 GPU 上以完整分辨率测试模型性能。将 Exchanger 与 Mask2Former [2] 框架结合时，我们主要遵循 [2] 中的设置，仅将学习率改为 2×10^{-5}。模型训练 100 个 epoch，随机裁剪尺寸增大为 64×64，并在 8 张 V100 GPU 上以 1 的批量大小进行训练。请注意，在评估用于全景分割的 Exchanger+Mask2Former 时，我们将输入切分为四个 64×64 图像块，并将预测结果拼接在一起 [4]。

4.5 消融研究

表1：在 PASTIS 验证数据集上通过 5 折交叉验证进行的 Exchanger 核心设计选择消融研究。括号中的数字表示使用的内容/位置查询数量。

	精确率%	召回率%	F1 分数%	#参数(M)	浮点运算次数
无位置查询（4）	80.0+0.8	77.0+1.0	78.3+0.9	0.50	117 G
使用位置查询 (4)	83.5+0.6	80.9+0.7	82.0+0.5	0.52	125 G
非绑定内容与位置注意力 (4)	83.6+0.6	81.1+0.7	82.2+0.5	0.52	125 G
非绑定内容与位置注意力 (8)	83.9+0.5	81.7+1.0	82.6+0.7	0.52	138 G
非绑定内容与位置注意力（16）	83.4+0.4	81.3+0.9	82.2+0.6	0.52	164 G
两阶段 (8)	84.3+0.4	82.3+0.4	83.1+0.3	0.94	283 G
临时自注意力 (8)	83.8+0.6	81.9+1.0	82.6+0.6	0.55	277 G
时间与空间自注意力 (8)	**84.5+0.6**	**82.7+1.0**	**83.4+0.8**	0.95	332 G

我们首先在 PASTIS 验证数据集上研究 Exchanger 中若干关键设计选择的影响，并将其与一个强基线模型进行比较。该基线模型采用自注意力处理时空特征，做法与 TSViT 中相同

[4] 我们通过实证发现，由于 ViT [1,4] 使用了空间位置编码外推和图块标记化层，全景评估指标对空间分辨率尤其敏感。

[34]。如表1所示，不使用位置查询会导致性能最差，相较于其他所有模型，绝对性能约下降 4%，这表明日期特定的时间嵌入对于捕捉作物物候特征至关重要。不同于在注意力计算中混合内容信息与位置信息，采用 TUPE 提出的非绑定内容与位置注意力 [14]可使 F1 分数略微提升 0.2%，因此除非另有说明，后续实验均默认采用此设置。随后，我们通过将内容与位置令牌的数量从 4 增至 8，再增至 16，评估 Exchanger 的性能。如表 1 所示，令牌数为 8 时，Exchanger 在精确率、召回率和 F1 指标上均取得最佳成绩。与自然语言处理中仅在输入序列前添加 1 个类别令牌不同，我们推测，作物类型识别需要稍多令牌，是由于显著的类内差异以及多模态特性；我们将在补充材料中展示其潜在嵌入。与 TSViT 中将令牌数量固定为待识别类别数的做法相反 [34], ，我们发现，持续增加内容/位置查询并未带来预期的性能提升，反而显著增加了计算成本。比较非绑定内容与位置注意力（8）和对应的自注意力方法（临时自注意力（8））可知，Exchanger 在参数量相近的情况下取得了几乎相同的结果，同时大幅降低了计算成本（十亿次浮点运算次数减少近 50%）。最后，通过堆叠两个相同的 Exchanger 模块（两阶段（8）），其 F1 分数达到 83.1，与时间与空间自注意力（8）（一种改进的 TSViT） [34] 的结果相当，同时计算开销更低（十亿次浮点运算次数减少约 15%）。此外，后者（时间与空间自注意力（8））可视为在时间自注意力模块之后添加注意力多实例学习池化组件 [12] ，以识别关键空间实例。然而，我们证明，仅增加 Exchanger 的深度也能带来类似的性能提升；并且与 TSViT 中用于密集预测后便被弃用的模块不同，Exchanger 还可复用于下游任务 [34]。

4.6 收敛性分析

我们证明了预训练-微调范式能够成功地从计算机视觉迁移到卫星图像时间序列分析，这得益于重新构建的卫星图像时间序列表示方式——从时空信号转变为实例集合。这使骨干网络能够在高效的像素集合格式上进行预训练，然后在标准时空网格上进行微调，以完成下游密集预测任务。具体而言，如图2所示，将预训练的 Exchanger 作为骨干网络，并接上一个权重随机初始化的常用分割模型 Unet，与完全从头训练相比，能够更快收敛、训练更稳定，并取得更高的验证准确率。

4.7 与最先进方法的比较

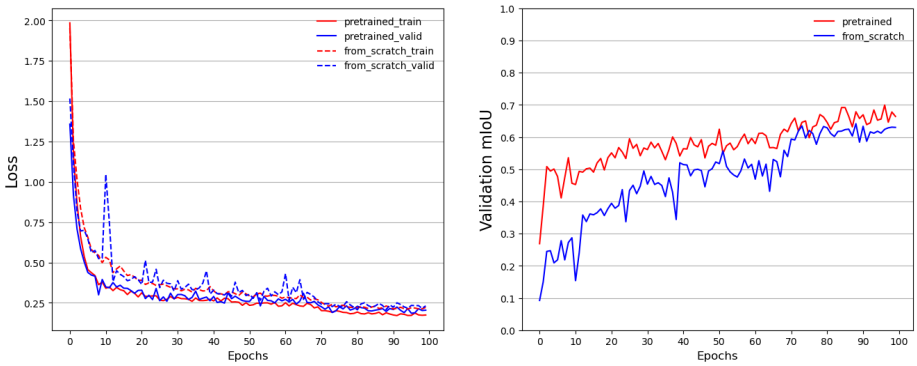


图2：使用预训练骨干网络或从头训练的 Exchanger+Unet 在 PASTIS 验证数据集（第1折）上的收敛性分析。左图显示训练损失和验证损失。右图显示验证数据集上的评估指标平均交并比。

表2：在 PASTIS 和 MTLCC 测试数据集上与最先进模型的比较。括号中的数值表示 PASTIS 官方 5 折划分中的标准差。浮点运算次数基于单个卫星图像时间序列样本和 $T \times C \times H \times W = 30 \times 10 \times 128 \times 128$ 计算。

	平均交并比（%）		#参数量（M）	浮点运算次数
	PASTIS	MTLCC		
FPN +ConvLSTM [24]	57.1	73.7	1.45	714 G
Unet +ConvLSTM [21]	57.8	76.2	2.33	55 G
Unet-3D [21]	58.4	75.2	1.55	92G
U-TAE [8]	63.1	77.1	1.09	47 G
TSViT [34]	65.4	84.8	2.16	558 G
Exchanger+Unet	66.8(+1.2)	**90.7**	8.08	300 G
Exchanger+Mask2Former	**67.9(+1.2)**	90.5	24.59	329 G

语义分割 如表2所示，将作为纯时间编码器的 Exchanger 与仅专注于空间语义挖掘的普通 Unet [28] 结合，在 PASTIS 和 MTLCC 上分别轻松取得 66.8% 和 90.7% 的平均交并比，超过 TSViT [34]此前取得的最先进结果，分别高出 1.4 和 5.9 个百分点，而浮点运算次数仅为其 53%。时间组件与空间组件的解耦进一步使我们能够探索采用近期提出的强大通用图像分割框架 Mask2Former [2] 的潜力，该框架以 PVT2 [38]为骨干网络、以 FPN [17] 为像素解码器。与文献报道的最佳结果相比，该方法绝对提升约 2.5%；与 Exchanger+Unet 相比，提升约 1.1%，但仅需不到

计算成本增加了 10%。值得注意的是，除 TSViT 外，所有以往用于卫星图像时间序列的语义分割模型都由复杂的空间组件和时间组件构成，这阻碍了它们利用计算机视觉领域的最新进展。尽管 TSViT 是首个用于卫星图像时间序列处理的全注意力神经网络架构，但由于图块标记化层使其无法直接处理像素集合格式，因此在采用预训练-微调范式时会面临额外障碍；此外，自注意力操作在密集预测任务中可能会带来高得难以承受的计算成本。另一个值得注意的事实是，时间优先、空间后续的处理顺序已被证明是卫星图像时间序列分析中更理想的归纳偏置 [34]，但这会导致时间编码器消耗所需计算量中的很大一部分。例如，在 Exchanger Unet 中，Exchanger 占总计算成本的近 96%。还应指出，我们提出的模型相对于输入序列长度仅具有线性计算复杂度 $\mathcal{O}(NTd)$。

表 3：与 PASTIS 测试数据集上的最先进模型进行比较。括号中的数值表示 PASTIS 官方五折划分中的标准差。浮点运算次数基于单个卫星图像时间序列样本计算，样本包含$T \times C \times H \times W = 30 \times 10 \times 128 \times 128$。推理时间（IT）在第1折上计算，约包含 490 个序列，并使用单块 A100 GPU。

	SQ	RQ	PQ	#参数(M)	浮点运算次数	IT(s)
Unet+ConvLSTM+PaPs [8]	80.2	43.9	35.6	2.50	55 G	660
U-TAE+PaPs [8]	81.5	53.2	43.8	1.26	47 G	207
Exchanger+Unet+PaPs	80.3(+0.1)	58.9(+0.6)	47.8(+0.4)	9.99	301 G	252
Exchanger+Mask2Former	**84.6(+0.9)**	**61.6(+1.6)**	**52.6(+1.8)**	24.63	332 G	**154**

全景分割 为进一步证明我们的表征学习框架的有效性，我们在 PASTIS 上的全景分割任务 [15] 中测试了其性能。该任务将语义分割和实例分割统一为一个联合任务，从而提供完整的场景理解视觉系统。尽管 [8] 中所做的开创性工作已将单阶段实例分割网络 CenterMask [40] 改造为名为 Parcels-as-Points（PaPs）的全景分割模块，但由于针对自然图像或视频提出的大多数现有全景分割网络在直接处理卫星图像时间序列时并不特别有效，因此该任务仍然极其困难。我们认为，强大的时间编码器是从卫星图像时间序列中提取高层语义的关键，它能够将低信噪比的 4-D 卫星数据 $T \times C \times H \times W$ 转换为丰富的语义 3-D 特征图 $C \times H \times W$，然后将其输入现成的全景分割模型。我们报告类别平均的分割质量（SQ）、识别质量（RQ）和

表3中的全景质量 [5] (PQ)。可以看出，与 U-TAE+PaPs 相比，采用 Unet [28] 作为空间编码器并使用 PaPs 模块 [8] 进行全景预测的 Exchanger，其 RQ 和 PQ 分别显著提升了 5.7% 和 4.0%。此外，值得注意的是，Exchanger 与 Mask2Former [2] 结合后，SQ、RQ 和 PQ 分别比 Ex- changer+Unet+PaPs 高出 4.3、2.7 和 4.8 个点，刷新了当前最佳水平。此外，由于流程精简且并行性高，Exchanger+Mask2Former 在 A100 GPU 上所需的推理时间显著更短。

4.8 定性结果



图3：定性比较。我们从 PASTIS 第1折验证数据集中随机抽取了 4 个卫星图像时间序列样本，并展示了 U-TAE+PaPs、Exchanger+Unet+PaPs 和 Exchanger+Mask2Former 的全景预测结果。请注意，最后一列中的伪影是由将 64×64 predic- tions 拼接到 128×128所导致的。

[5] 请注意，我们遵循 [8] 中的评估协议，其中 PQ 的计算仅涉及事物类别。

本节对先前的最先进模型 U-TAE+PaPs、Exchanger+Unet+PaPs，以及首个通用卫星图像时间序列分割架构 Exchanger+Mask2Former 进行定性比较。后者是将 Exchanger 作为时间编码器，与近期提出的通用自然图像分割框架 Mask2Former [2]相结合的成果。如图 3 所示，U-TAE+PaPs 检出的农田地块数量几乎与 Ex- changer+PaPs 相同，但更容易产生错误预测，这表明 U-TAE 的表征学习能力较弱。将 Exchanger 与性能更强的分割架构 Mask2Former [2], 结合后，全景预测质量显著提升，具体体现在作物类型识别准确率和作物形状预测方面，这与表 3 报告的 SQ 和 RQ 指标一致。

5 结论

总而言之，本文首先将卫星图像时间序列的表征独特地重新定义为实例集合，从而放宽了传统时空网格带来的限制，并进一步支持设计能够灵活处理卫星图像时间序列的像素集合格式和图像序列格式的模型。随后，我们提出将卫星图像时间序列的表征学习过程明确分解为三个步骤：收集–更新–分发，由此构建出概念清晰且计算高效的特征学习框架 Exchanger。得益于前述两项创新，我们首次展示了将预训练-微调范式从计算机视觉迁移至卫星图像时间序列的成功实践，形成了简化的语义分割与全景分割流程，并且性能显著优于先前的最先进模型。

致谢

本研究得到阿尔斯特大学（UU）经济部（DfE）国际学生奖学金的支持。论文中所展示的所有实验均在阿尔斯特大学的高性能计算（HPC）中心完成。我们感谢审稿人提出的建设性且富有见地的意见。

参考文献

1. Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., Pavetic, F.：Flexivit：适用于所有图块尺寸的单一模型。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第14496–14506页（2023） 2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.：用于通用图像分割的掩码注意力掩码Transformer。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第1290–1299页（2022）

3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.：ImageNet：大型分层图像数据库。载于：2009年IEEE计算机视觉与模式识别会议。第248–255页。IEEE（2009） 4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等：一张图像相当于16×16个词：用于大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929（2020） 5. Feng, Y., Jiang, J., Tang, M., Jin, R., Gao, Y.：重新思考监督式预训练，以实现更好的下游迁移。arXiv预印本 arXiv:2110.06014（2021） 6. Garioud, A., Peillet, S., Bookjans, E., Giordano, S., Wattrelos, B.：Flair：来自航空航天影像的法国土地覆盖。（2022） 7. Garnot, V.S.F., Landrieu, L.：用于卫星图像时间序列分类的轻量级时间自注意力。载于：时间数据高级分析与学习：第5届ECML PKDD研讨会，AALTD 2020，比利时根特，2020年9月18日，修订精选论文6。第171–181页。施普林格（2020） 8. Garnot, V.S.F., Landrieu, L.：利用卷积时间注意力网络对卫星图像时间序列进行全景分割。载于：IEEE/CVF国际计算机视觉会议论文集。第4872–4881页（2021） 9. Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N.：利用像素集编码器和时间自注意力进行卫星图像时间序列分类。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第12325–12334页（2020） 10. He, K., Chen, X., Xie, S., Li, Y., Doll´ar, P., Girshick, R.：掩码自编码器是可扩展的视觉学习器。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第16000–16009页（2022） 11. Horn, M., Moor, M., Bock, C., Rieck, B., Borgwardt, K.：用于时间序列的集合函数。载于：国际机器学习会议。第4353–4363页。PMLR（2020） 12. Ilse, M., Tomczak, J., Welling, M.：基于注意力的深度多实例学习。载于：国际机器学习会议。第2127–2136页。PMLR（2018） 13. Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S.：Tile2vec：针对空间分布数据的无监督表征学习。载于：AAAI人工智能会议论文集。第33卷，第3967–3974页（2019） 14. Ke, G., He, D., Liu, T.Y.：重新思考语言预训练中的位置编码。arXiv预印本 arXiv:2006.15595（2020） 15. Kirillov, A., He, K., Girshick, R., Rother, C., Doll´ar, P.：全景分割。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9404–9413页（2019） 16. Kondmann, L., Toker, A., Rußwurm, M., Camero Unzueta, A., Peressuti, D., Mil- cinski, G., Long´ep´e, N., Mathieu, P.P., Davis, T., Marchisio, G., 等：Denethor：DynamicEarthNet数据集——用于从太空进行协调一致、互操作、可供分析的每日作物监测。载于：第35届神经信息处理系统会议数据集与基准赛道。第1–13页（2021） 17. Lin, T.Y., Doll´ar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.：用于目标检测的特征金字塔网络。载于：IEEE计算机视觉与模式识别会议论文集。第2117–2125页（2017） 18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Doll´ar, P.：用于密集目标检测的焦点损失。载于：IEEE国际计算机视觉会议论文集。第2980–2988页（2017）

19. Long, J., Shelhamer, E., Darrell, T.：用于语义分割的全卷积网络。在：IEEE计算机视觉与模式识别会议论文集。第3431–3440页（2015） 20. Loshchilov, I., Hutter, F.：解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101（2017） 21. M Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M., Lobell, D.：非洲作物类型的语义分割：深度学习方法的新型数据集与分析。在：IEEE/CVF计算机视觉与模式识别会议研讨会论文集。第75–82页（2019） 22. Ma, X., Zhou, Y., Wang, H., Qin, C., Sun, B., Liu, C., Fu, Y.：图像作为点集。在：第十一届国际学习表征会议（2023） 23. Van der Maaten, L., Hinton, G.：使用t-SNE可视化数据。机器学习研究期刊 **9**(11)（2008） 24. Martinez, J.A.C., La Rosa, L.E.C., Feitosa, R.Q., Sanches, I.D., Happ, P.N.：从多时相图像序列中进行多时相作物识别的全卷积循环网络。ISPRS摄影测量与遥感杂志 **171**, 188–201（2021） 25. Nyborg, J., Pelletier, C., Assent, I.：利用热位置编码对卫星图像时间序列进行广义分类。在：IEEE/CVF计算机视觉与模式识别会议论文集。第1392–1402页（2022） 26. Pelletier, C., Webb, G.I., Petitjean, F.：用于卫星图像时间序列分类的时间卷积神经网络。遥感 **11**(5)，523（2019） 27. Ren, S., He, K., Girshick, R., Sun, J.：更快的区域卷积神经网络：利用区域建议网络实现实时目标检测。神经信息处理系统进展 **28** （2015） 28. Ronneberger, O., Fischer, P., Brox, T.：U-Net：用于生物医学图像分割的卷积网络。在：医学图像计算与计算机辅助干预–MICCAI 2015：第18届国际会议，德国慕尼黑，2015年10月5–9日，论文集，第三部分18。第234–241页。施普林格（2015） 29. Rußwurm, M., K¨orner, M.：用于遥感影像抗云分割的卷积LSTM。arXiv预印本 arXiv:1811.02471（2018） 30. Rußwurm, M., K¨orner, M.：利用序列循环编码器进行多时相土地覆盖分类。ISPRS国际地理信息期刊 **7**(4)，129（2018） 31. Sariyildiz, M.B., Kalantidis, Y., Alahari, K., Larlus, D.：没有理由不进行监督：监督模型中的泛化改进。在：ICLR 2023-国际学习表征会议。第1–26页（2023） 32. Shukla, S.N., Marlin, B.M.：用于不规则采样时间序列的插值预测网络。arXiv预印本 arXiv:1909.07782（2019） 33. Suzuki, T.：作为注意力的聚类：利用层次聚类实现统一图像分割。arXiv预印本 arXiv:2205.09949（2022） 34. Tarasiou, M., Chavez, E., Zafeiriou, S.：用于卫星图像时间序列的视觉变换器：用于卫星图像时间序列的视觉变换器。arXiv预印本 arXiv:2301.04944（2023） 35. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., 等：MLP-Mixer：一种用于视觉的全多层感知机架构。神经信息处理系统进展 **34**，24261–24272（2021） 36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.：注意力就是你所需要的一切。神经信息处理系统进展 **30**(2017)

37. Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.T.：多时相哨兵-2号数据在作物类型分类方面能带来多大程度的提升？《国际应用地球观测与地理信息期刊》**72**，122–130（2018） 38. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.：Pvtv2：基于金字塔视觉Transformer的改进基线。《计算视觉媒体》**8**(3)，1–10（2022） 39. Wang, Y., Tang, S., Zhu, F., Bai, L., Zhao, R., Qi, D., Ouyang, W.：重新审视监督预训练的迁移能力：多层感知机视角。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9183–9193页（2022） 40. Wang, Y., Xu, Z., Shen, H., Cheng, B., Yang, L.：Centermask：基于点表示的单阶段实例分割。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9313–9321页（2020） 41. Woo, G., Liu, C., Sahoo, D., Kumar, A., Hoi, S.：Etsformer：用于时间序列预测的指数平滑Transformer。arXiv预印本 arXiv:2202.01381（2022） 42. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.：Groupvit：语义分割通过文本监督得以实现。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第18134–18144页（2022） 43. Yang, C., Xu, J., De Mello, S., Crowley, E.J., Wang, X.：Gpvit：一种具有群组传播机制的高分辨率非层次化视觉Transformer。arXiv预印本 arXiv:2212.06795（2022） 44. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.：Cmt-deeplab：用于全景分割的聚类掩码Transformer。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第2560–2570页（2022） 45. Zeng, A., Chen, M., Zhang, L., Xu, Q.：Transformer对时间序列预测是否有效？arXiv预印本 arXiv:2205.13504（2022） 46. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.：Fedformer：用于长期序列预测的频率增强分解式Transformer。载于：国际机器学习会议。第27268–27286页。PMLR（2022）

A.1 PASTIS 的颜色调色板

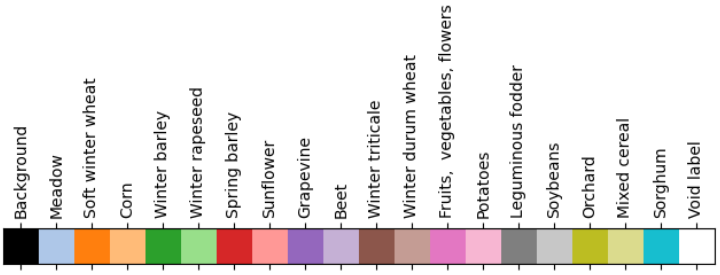


图 A.1.1：用于可视化 PASTIS 上潜在特征、语义预测和全景预测的颜色调色板。

B.2 Exchanger 中潜在特征的可视化

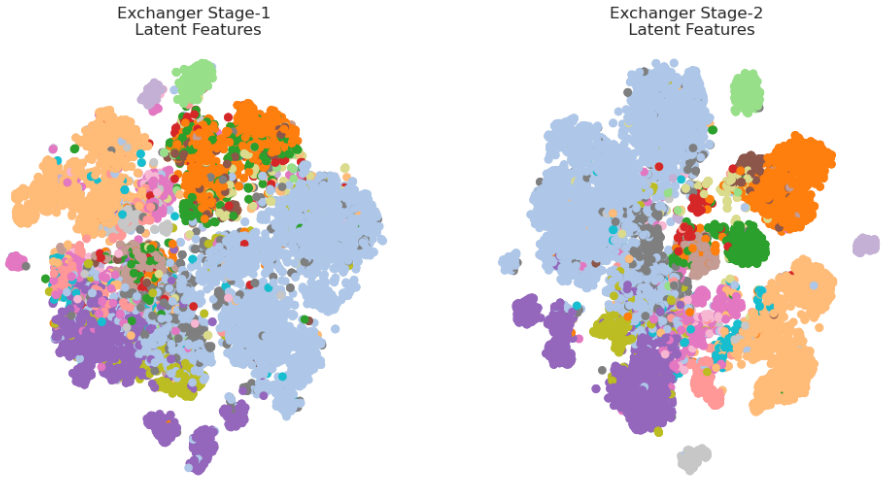


图 B.2.1：Exchanger 阶段1和阶段2潜在特征的 t-SNE [23] 可视化。

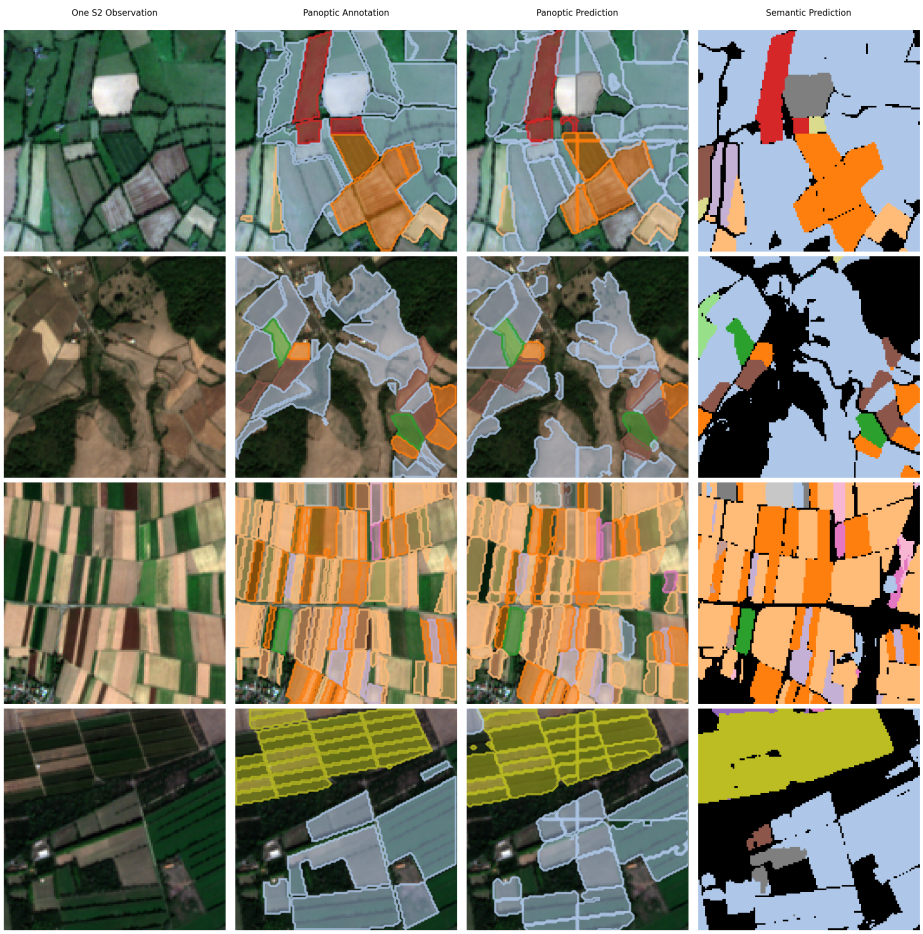
我们在图 B.2.1 中展示了 Exchanger 的阶段1和阶段2输出中、投影头之前的潜在特征。首先可以看出，类内

与阶段1的输出相比，阶段2的输出中的变化显著减少，这表明通过增加 Exchanger 的深度实现了一种层次聚类过程。此外，值得注意的是，作物类型识别所固有的多模态特性，使得 NLP 中传统的在输入序列前添加单个类别令牌的方法不再有效。

C.3 Exchanger+Mask2Former 的更多定性可视化



图 C.3.1：Exchanger+Mask2Former 的预测定性结果。请注意，分割模型与全景分割模型是分别训练的。



图C.3.2：Exchanger+Mask2Former 预测的定性结果。请注意，分割模型和全景分割模型是分别训练的。

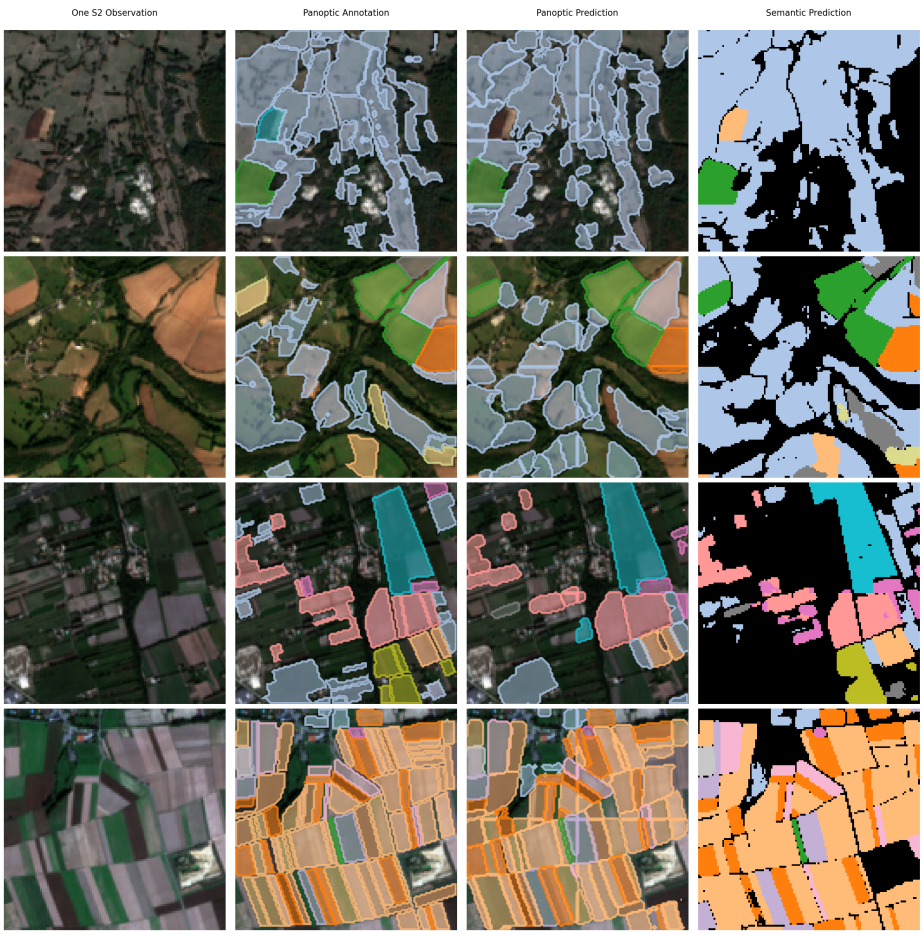


图 C.3.3：Exchanger+Mask2Former 预测结果的定性分析。请注意，分割模型和全景分割模型是分别训练的。



图 C.3.4：Exchanger+Mask2Former 的预测结果定性展示。请注意，分割模型和全景分割模型是分别训练的。

D.4 SITS 的域泛化

在本节中，我们进一步展示了在 TimeMatch 数据集 [25] 上评估的 Exchanger[2阶段（使用 8 个标记]）的结果。该数据集由来自四个不同图幅的 SITS 组成：33UVP（奥地利）、32VNH（丹麦）、30TXT（法国中西部）和 31TCJ（法国南部）。我们遵循 [25]采用的命名约定，分别将这四个哨兵-2号图幅称为 AT1、DK1、FR1 和 FR2，并采用留一地区评估协议：留出一个哨兵-2号图幅用于测试，其余三个图幅用于训练。除了用于评估作物分类器空间泛化能力的专门构建数据集外， [25]中的作者还提出使用热位置编码（TPE），以应对不同地理位置之间的时间偏移；其中，使用生长度日（GDD）替代日历时间，事实证明这有助于提升空间泛化能力。我们直接采用 [25] 提出的 TPE 方法，修改 Exchanger 中的位置编码组件。根据我们的经验观察，为获得更好的泛化性能，宜将位置嵌入的维度设置为相对较小的数值，这表明模型对正弦/余弦函数中频率分辨率的敏感性。如表 D.4.1 所示，我们提出的模型仅训练 20 个 epoch，便可取得与原始设置中训练 100 个 epoch 的 PSE+LTAE [7]相当的结果。然而，高度专门化的架构 PSE+LTAE [7] 仍展现出优于我们模型的性能，这将作为未来改进的方向。

表 D.4.1：留一区域空间泛化结果（宏平均 F1 分数）。

		AT1	DK1	FR1	FR2	平均值
像素集编码器+LTAE [7]	TPE-Fourier	84.7	79.0	77.3	80.0	80.3
	TPE-Recurrent	**86.5**	**80.3**	**86.0**	**80.5**	**83.3**
Exchanger	TPE-Fourier	84.1	77.8	84.2	77.6	80.9
	TPE-Recurrent	82.9	80.1	81.2	76.4	80.2