MAESTRO：用于多模态、多时态和多光谱地球观测数据的掩码自动编码器

Antoine Labatie [1] Michael Vaccaro [1] Nina Lardiere [1] Anatol Garioud [1] Nicolas Gonthier [1,2]

[1] Institut national de l'information géographique et forestière (IGN), France

[2] Univ Gustave Eiffel, ENSG, IGN, LASTIG, France

{ firstname.lastname } @ign.fr

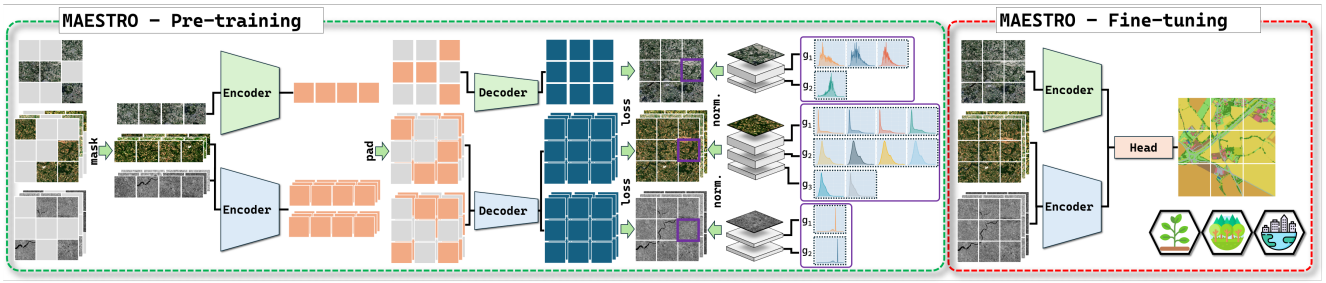


Figure 1. Overview of MAESTRO . MAESTRO extends the Masked Autoencoder to orchestrate the complex interplay of multimodal, multitemporal, and multispectral Earth Observation data. It employs token-based early fusion across time steps and similar modalities, and token-based late fusion across dissimilar modalities. It uses joint-token fusion for multispectrality, but still relies on a novel normalization of reconstruction targets—namely, patch-group-wise within groups of highly correlated bands—to inject a useful spectral prior during pretraining. Best viewed in color.

自监督学习（SSL）是近年来自然语言处理 [16, 63] 和计算机视觉 [34, 94] 突破的核心。它是一种有效的预训练策略，在对多种下游任务进行微调时，能够提高模型的多功能性和数据效率，尤其是在预训练大量无标签数据的基础模型的情况下。

自监督学习（SSL）范式在地球观测（EO）应用中极具前景，因在该领域中未标记数据丰富，但标记数据仍然稀缺且成本高昂。然而，为了充分利用 SSL 的潜力，可能有必要将现成的方法适应于 EO 数据的独特特征。EO 数据的一个关键特征在于其在分辨率、尺度、模式和光谱波段上具有异质性 [65] 。这种异质性常常在多模态、多时态和多光谱的 EO 数据集中体现，通过集成各种传感器可有效解决空间-时间-光谱问题 “resolution dilemma” [2] 。

我们在这项工作中的目标是高效地将掩码自编码的经典 SSL 框架适配到 EO 数据的特定特性，以学习有用且多功能的表示。我们适配了在自然图像上表现良好且计算效率高的掩码自编码器（MAE）[34] 。然而，将 MAE 扩展到多模态、多时序和多光谱的 EO 数据带来了若干挑战。首先，集成多种模态和时序观测需要一个有效的融合策略。一种方法是联合 token 融合，将所有数据维度投射到一个共享的 token 空间中。另一种方法是基于 token 的融合，首先对数据维度进行聚类，分别嵌入它们，然后融合所得的 tokens。

基于令牌的融合似乎更适合处理多模态性和多时间性，原因有几个：(i) 它更好地捕捉了不同模态和时间步之间的异质性，(ii) 它允许整合特定于模态和时间性的编码，(iii) 它支持使用跨模态和跨时间的自监督信号。然而，仍然不清楚在模型中应该在早期还是晚期进行融合。

其次，处理多光谱数据增加了更多的复杂性。在这里，基于 token 的融合和联合 token 融合之间的选择并不明确。虽然联合 token 融合提供了计算效率，但它限制了整合跨光谱自监督信号的能力。一个关键问题是，是否可以在保持联合 token 融合的计算优势的同时，通过精心设计的重构目标归一化重新引入对多光谱结构的有效先验。

我们在 Sec. 3.1 中研究了这些问题，并基于我们的发现，在 Sec. 3.2 中引入了我们提出的方法 MAESTRO。MAESTRO 通过应用基于 token 的早期融合来扩展 MAE 框架，适用于时间步和相似模态，以及基于 token 的后期融合，适用于不相似的模态。对于多光谱性，MAESTRO 采用了联合 token 的早期融合，结合了一种针对重建目标的新颖归一化方案。

我们的主要贡献如下：

- 我们广泛基准测试了各种融合策略和目标归一化方案，用于地球观测中的多模态、多时态和多光谱自监督学习。
- 我们介绍了一种新颖的分组依赖的补丁归一化方法，该方法在高度相关的光谱频带组内逐个补丁地对重建目标进行归一化。此方法在几乎没有计算成本的情况下将有用的光谱先验注入了自监督信号中。

- 在上述基础上，我们提出了 MAESTRO，一种量身定制的 MAE 框架的适应，将多模态、多时态和多光谱地球观测数据的使用有效地协调起来——无论是在性能还是计算成本方面。
- 我们在四个不同的地球观测基准上验证了 MAESTRO，证明了它在与多时态动态紧密相关的任务中取得了最先进的结果，同时在其他任务中也保持了高度竞争力。

1. 相关工作

由于遥感数据资源丰富，

用于地球观测的 SSL 方法和模型。 SSL 已成为一种对 EO 颇具前景的方法。最初，EO 的 SSL 方法的发展借鉴了自然图像领域的方法。早期的工作主要集中在对比 SSL [6, 14, 39, 42, 52, 54, 61, 73, 86, 88, 89, 92]，通常使用卷积神经网络 [6, 39, 52, 54, 61, 73, 88, 92]。正样本对通常来源于相同位置的不同模态、时间或者增强视图。

最近的工作越来越多地转向使用适合于掩码自编码的变压器的生成式自监督学习（SSL）。这些方法包括受到 MAE [9, 11, 13, 35, 37, 40, 48, 49, 58, 64, 71, 72, 76, 78, 81–84, 90, 91, 95, 97] 或 SimMIM [17, 33, 36, 55, 68, 70] 启发的方法，或者自定义生成方法 [20, 21, 67, 85, 87]。一些最近的工作也在探索混合对比/生成方法 [18, 57] 或潜在空间生成式 SSL [4, 5, 12, 45, 77, 80]。

遥感的多模态和多时态自监督学习。 最近对 EO 的研究已经开始解决多模态 [8, 9, 14, 24, 33, 36, 37, 39, 41, 58, 61, 62, 80, 86, 88, 91, 95] 或多时态的 SSL [13, 21, 40, 59, 71]，但很少同时解决两者。当包含多模态时，通常通过参数共享或后期融合处理 [7, 33, 37, 58, 62, 86, 91, 95]。迄今为止，大多数发表的 EO 基础模型仅通过后期融合支持多模态或多时态输入，这限制了表示学习和下游性能（见 Sec. 3.1）。

值得注意的例外包括 Presto [76]、Galileo [77]、OmniSat [4]、AnySat [5]、SeaMo [46]、EarthMAE [78]、SkySense v1/v2 [32, 98] 和 SkySense++ [93]。在这些例子中，Presto 在没有空间上下文的情况下逐像素处理数据；OmniSat 和 AnySat 在时间步长之间使用联合令牌的早期融合（通过 LTAE [29]）；Galileo 在所有时间步长和模态之间应用基于令牌的早期融合。相比之下，MAESTRO 仅在相似模态的时间步长之间采用基于令牌的早期融合，避免了在不同模态间共享编码器的低效。尽管融合设计在多模态和多时态自监督学习 (SSL) 中非常重要（见 Sec. 3.1），但据我们所知，仅有一项之前的研究 [46] 对不同的融合策略进行了基准测试——即便如此，多时态融合也仅限于 SSL 预训练中，且未包含微调评估。

在地球观测领域的多种自监督学习研究中，已有研究整合了 SAR 数据 [36, 42, 45, 49]、多光谱数据 [13, 15, 21, 37, 40, 54, 55, 62, 67, 68, 71, 73, 92] 或两者兼有 [4, 5, 7–9, 14, 24, 33, 39, 41, 43, 47, 58, 61, 76–78, 80, 86, 88–91, 95]。然而，仅有少数研究在模型架构或自监督信号中显式编码了光谱先验。一些先前的工作使用基于 token 的多光谱融合，通过为每个光谱波段 [9, 62] 或光谱波段组 [13, 37, 76, 77] 分配单独的 token。虽然这些设计可以帮助缓解前期信息瓶颈、增强建模能力及实现跨光谱自监督信号，但代价是更高的计算需求。

一些研究探讨了多光谱数据重建目标的归一化，但它们的方法局限于不考虑底层光谱结构的逐块归一化 [58, 78]。相反，我们通过对高度相关的光谱带群进行归一化，有效地将光谱先验注入自监督信号中。这种方法在不增加标记数的情况下增强了光谱表示学习，并通过类似的带分组策略自然地扩展到 SAR 输入。

2. 方法

2.1. 架构

在本节中，我们描述了一种专为地球观测数据量身定制的 MAE 衍生方法。我们的方法适应地球观测数据的异质性，特别是其多模态、多时态和多光谱特征。

我们在一个固定的数据集 $\mathcal{D}$ 的背景下描述我们的方法，该数据集由一组模态 $m \in \mathcal{M}$ 组成。对于每个数据集的瓦片，每个模态 m 关联一个形状为 $I_m \times I_m \times T_m \times C_m$ 的输入张量，其中 I_m 表示空间大小，$T_m \geq 1$ 表示原始时间步数，C_m 表示通道数量。我们假设 C_m 和 I_m 在所有瓦片中保持不变，而 T_m 可能会变化。这种变化反映了卫星时间序列的性质，其可能跨越固定的持续时间但由于位置的不同而具有不同的重访间隔——如哨兵-1 或哨兵-2 所示。

我们的预处理流程从确保模型输入的张量形状固定开始。对于每种模态 m，我们定义了离散时间步长的目标数量 D_m。对于模态 m，原始输入张量形状 $I_m \times I_m \times T_m \times C_m$ 通过一个两步过程被缩减到张量形状 $I_m \times I_m \times D_m \times C_m$：(i) 时间分箱，其中序列被重塑为 D_m 个箱，每个箱覆盖 T_m/D_m 个时间步长（如果需要，经过随机截断）；以及 (ii) 每个箱内的时间步选择。

在训练过程中，每个区间内的时间步选择是随机的，以引入数据增强。在验证和测试中，选择旨在最大化每个区间的代表性。有关更多信息，请参见 SM Sec. 6.2.1。

打补丁/取消补丁。 一旦来自不同模态的输入被缩减到固定形状，它们便被传递给模态特定的分词器。这些分词器在给定模态内的各个时间步之间是共享的，但在不同模态之间是不同的。分词在每个时间步内独立进行。

我们采用标准的 Vision Transformer (ViT) 切块策略 [19]，对于任何模态 m 和时间步，首先将输入在空间上划分为大小为 P_m 的不重叠的块。然后将每个块展平为形状为 $P_m^2 C_m$ 的向量，并在编码器之前投影为维度为 C_e 的块嵌入。所有光谱波段共同投影到相同的 token 中，实现联合 token 多光谱融合。在 SSL 预训练期间，我们应用转置反切块操作，将解码器解码出的形状为 C_d 的 token 转换回形状为 $P_m^2 C_m$ 的重构展平块。

对于每个模态 m，补丁位置集 $p \in \mathcal{P}_m$ 的基数为 $|\mathcal{P}_m| = (I_m/P_m)^2$，而时间窗口集 $t \in \mathcal{T}_m$ 的基数为 $|\mathcal{T}_m| = D_m$。重要的是，我们不假设所有模态共享相同数量的空间位置或时间窗口——也就是说，它们的空间和时间分辨率可能不同。例如，像 Sentinel-1 和 Sentinel-2 这样的模态可能使用较粗的空间网格但较细的时间网格，而类似航空和 SPOT 6-7 的模态可能使用较细的空间网格但较粗的时间分辨率。在整个标记化过程中保持这些粗或细的分辨率有助于防止标记器中的信息瓶颈——即模型的入口模块。

在分词之后，我们添加 (i) 基于地面取样距离的二维空间位置编码，类似于 ScaleMAE [64] 中的做法，以及 (ii) 类似于 SatMAE [13] 中使用的正弦-余弦时间编码。我们不包含显式模式编码，因为特定模式的分词器和可学习的特定模式的 `[mask]` 代币隐含地编码了与每个代币相关联的源模式或目标模式。更多信息请参考 SM Sec. 6.2.2 。

在这个阶段，我们为每种模态 m 获取到形状为 $I_m/P_m \times I_m/P_m \times D_m \times C_e$ 的张量形式的嵌入输入。与原始 MAE [34] 中一样，这些张量经过两个阶段的处理：Transformer 编码器仅处理可见的 token，而 Transformer 解码器则处理编码后的 token，并与 `[mask]` token 级联处理。

然而，原始的 MAE 工作流程是为单模态、单时间数据设计的，如何扩展以支持多模态和多时间数据仍然存在模糊性。这提出了关键问题：

- 来自不同模态和时间步的信息应该通过早期融合还是晚期融合进行融合？
- 编码器和解码器参数应该在不同模态之间共享还是保持独立？

为了解答这些问题，我们探索了五种不同的融合模式，如 Fig. 2 所示：

- 模式共享：跨模态和时间步的后期融合。参数在所有模态中共享。
- 模式单一温度：与共享相同，但每种形态保持参数独立。
- 模式 mod：在不同模态之间进行后期融合，但在时间步之间进行早期融合。每种模态的参数保持独立。
- 模式组：在预定义的模态组之间进行后期融合，但在时间步和每个组内进行早期融合。各组之间的参数保持独立。
- 组间模式：类似于组模式，但将最后三个编码器块替换为融合块，从而实现跨组 token 交互。来自不同组的模态采用中间融合，而不是后期融合。

请注意，严格来说，术语 “late fusion” 仅适用于微调，在微调中，池化发生在预测头之前。在 SSL 预训练中，解码器之前不会发生这种池化。

我们的分词器将每个模态的所有光谱波段联合编码为共享的标记，实现联合标记的多光谱融合。虽然这种设计提高了效率，但我们仍然旨在将有意义的光谱先验嵌入到重建任务中。为此，我们扩展了最初在 MAE 中提出的补丁目标归一化策略，并在多个专注于地球观测的工作中采用，通过补丁组目标归一化策略来实现。具体来说，对于每个模态 m，我们考虑将光谱波段分成一组波段组 $g \in \mathcal{G}_m$，并为每个补丁和每个波段组独立地归一化目标。

与之前相同的符号表示下，令 $\boldsymbol{x}_m, \boldsymbol{x}_m^{\text{rec}} \in \mathbb{R}^{I_m/P_m \times I_m/P_m \times D_m \times P_m^2 C_m}$ 分别表示未归一化和重建目标在模态 m 下的补丁级别表示。基于 L^1 损失的重建损失取决于所选择的归一化策略：

- 不进行归一化：目标 $\boldsymbol{x}_m$ 按原样使用：

$$\mathcal{L} = \frac{1}{\sum_m |\mathcal{P}_m||\mathcal{T}_m|} \sum_{m,p,t} \left\| \boldsymbol{x}_{m,p,t} - \boldsymbol{x}^{\text{rec}}_{m,p,t} \right\|_1 ;$$

- 逐块归一化：目标在所有波段上独立地对每个块进行归一化 [34, 35, 58, 78, 81]：

$$\hat{\mathcal{L}} = \frac{1}{\sum_m |\mathcal{P}_m||\mathcal{T}_m|} \sum_{m,p,t} \left\| \hat{\boldsymbol{x}}_{m,p,t} - \boldsymbol{x}^{\text{rec}}_{m,p,t} \right\|_1 ,$$

$$\forall m,p,t: \quad \hat{\boldsymbol{x}}_{m,p,t} = \frac{\boldsymbol{x}_{m,p,t} - \mu(\boldsymbol{x}_{m,p,t})}{\sigma(\boldsymbol{x}_{m,p,t})};$$

- 按补丁组归一化（见 Fig. 1 ）：目标是根据每个补丁和频带组独立归一化：

$$\hat{\mathcal{L}}^{\text{grp}} = \frac{1}{\sum_m |\mathcal{P}_m||\mathcal{T}_m||\mathcal{G}_m|} \sum_{m,p,t,g} \left\| \hat{\boldsymbol{x}}^{\text{grp}}_{m,p,t,g} - \boldsymbol{x}^{\text{rec}}_{m,p,t,g} \right\|_1 ,$$

$$\forall m,p,t,g: \quad \hat{\boldsymbol{x}}^{\text{grp}}_{m,p,t,g} = \frac{\boldsymbol{x}_{m,p,t,g} - \mu(\boldsymbol{x}_{m,p,t,g})}{\sigma(\boldsymbol{x}_{m,p,t,g})}.$$

我们的补丁组目标归一化概括了原始的补丁归一化；具体来说，当所有频带被分成一个单一的集合时，它就简化为 [34, 35, 58, 78, 81] 中使用的原始策略。

我们发现，应用具有精心选择的频谱组 $g \in \mathcal{G}_m$ 进行的补丁组正则化，可以获得较强的性能。这种目标正则化，如 Sec. 3.1 中所示，使得联合标记融合能够达到——甚至有时超过——基于标记的融合的性能，同时在计算效率上显著提高。

另一个必须适应异构数据的方面是掩码策略。我们的目标是在每个编码器和解码器处理的标记集中保持固定的掩码比例。然而，在异构模态的情况下，如何在不同的标记集之间应用这种掩码仍然不明确。

我们选择了一种双阶段方法：(i) 在模态、空间和时间维度上应用结构化掩蔽，(ii) 执行非结构化调整以满足总体掩蔽比例为 75 % 的要求。关于详细信息和与之前工作的掩蔽策略的比较，请参见 SM Sec. 6.3.1 。

2.2. 下游任务

根据 MAE 框架，我们仅在 SSL 预训练后传输编码器，以减轻与 `[mask]` 标记 [16, 50] 相关的领域偏移效应。

因此，编码器的输出是每种模态 m 的形状为 $I_m/P_m \times I_m/P_m \times D_m \times C_e$ 的特征张量。

然后，我们附加分类和分割头来处理这些编码特征（见 Fig. 1 ）：

- 分类头的操作如下：

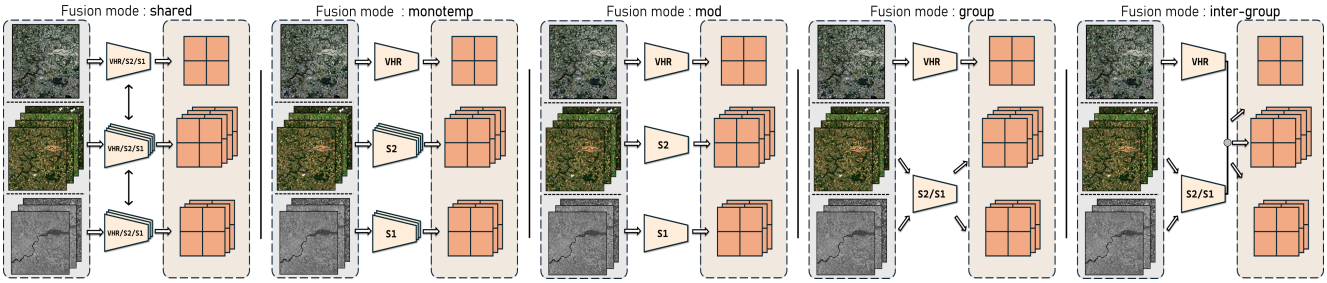


Figure 2. 基于令牌的融合模式用于处理多模态和多时间性。模式 shared 和 monotemp 涉及跨模态和时间步的晚期融合，其中参数可以在模态间共享（shared）或者为每个模态独立保持（monotemp）。模式 group 涉及跨预定义的模态组的晚期融合，但在时间步和每个组内部进行早期融合。模式 inter-group 通过用融合块替代最终的编码器块来扩展 group，从而实现跨组的令牌交互。模式 mod 是 group 的一个特例，包含跨所有模态的晚期融合，但在时间步上的早期融合。

(i) 连接所有模态 m 、空间位置 p 和时间步骤 t 的标记；
(ii) 应用注意力池化 [22] 来聚合连接的标记；
(iii) 应用一个输出维度等于目标类别数量的全连接层。

- 分割头首先将所有特定模态的编码张量对齐到一个共用的空间标记参考网格。然后，对于每个空间位置 p ，它们的操作如下：
  (i) 将所有模态 m 和时间步 t 的标记串联起来；
  (ii) 应用注意力池化以聚合连接后的标记；
  (iii) 应用一个输出维度等于目标类别数量的密集层。

3. 实验

我们的工作流程根据评估的模型而有所不同。

对于 MAEs，我们在同一个数据集上按顺序遵循三阶段工作流程：

1. 自监督预训练：模型从无标签数据中学习表示而无需监督。
2. 探测：在冻结的预训练骨干网络之上训练一个特定任务的头部，以评估学到的表示的质量。在此阶段，仅更新特定任务头部的参数。
3. 微调：整个模型——包括骨干网络和任务特定的头部——以完全监督的方式进行端到端训练。

对于基线基础模型（FMs）和监督 ViTs，我们仅关注监督微调阶段，这更接近于操作设置。

请注意，我们的 MAE 工作流程仅涉及数据集内的迁移学习；本文中不涉及跨数据集的迁移。重要的是，我们的 SSL 预训练仅在训练集和验证集的联合上进行，严格排除测试集。这确保了 MAE 模型不会通过任何形式的先前曝光于微调测试域来获得不公平的优势，包括通过输入模态间接获得的优势。

我们将我们的工作流程应用于四个涉及多模态性、多时态性和多光谱性的大中型数据集：

- TreeSatAI-TS [1, 4] –树种识别，具有 15 个多标签类别。数据集包括 50,381 块 60 m × 60 m 的图块，分布在德国，其中包含分辨率为 0.2 m 的航空影像（RGB + NIR），以及覆盖全年最接近航空获取时间的 Sentinel-1 和 Sentinel-2 时间序列。
- PASTIS-HD [4, 30] –农业作物分割包含 19 个语义类别。该数据集包括法国 433 块 1280 m × 1280 m 的区域，具有非常高分辨率（VHR）的卫星图像（SPOT 6–7）重采样到 1 米，以及覆盖大约 70 次在上升和下降轨道上的获取数据的 Sentinel-1 时间序列和大约一年时间跨度的 Sentinel-2 时间序列。
- FLAIR # 2 - 包含 12 种语义类别的土地覆盖分割。该数据集包括 77,762 个位于法国的 102.4 m × 102.4 m 的图块，具有 0.2 m 分辨率的航空和彩色高程图像（RGB + NIR + DSM），以及跨越整年的哨兵-2 时间序列。与 [5] 中的方法类似，我们使用的是没有超级补丁的 FLAIR # 2 版本，将哨兵-2 时间序列裁剪到 VHR 图像范围内（丢弃了 93.5 % 的像素）。
- FLAIR-HUB [28] –具有 15 个语义类别的土地覆盖分割。FLAIR # 2 的这个扩展版本包括 241,100 块每块面积为 102.4 m × 102.4 m 的地块，这些地块位于法国，包含分辨率为 0.2 米的航空和高程影像（RGB + NIR + DEM + DSM），以及覆盖一整年的 Sentinel-1 和 Sentinel-2 时间序列。

请注意，对于某些数据集，我们不使用完整标签集或输入模式（详见附录 Sec. 6.1 ）。为了研究预训练和微调数据集规模的缩放定律，我们还创建了每个数据集的过滤版本，其中包含原始样本的 20 % 和 5 %。这些子集是通过基于地理坐标的最远点采样获得的。

3.1. 消融研究

在本节中，我们评估了各种 SSL 策略和模型组件的选择对我们最终方法设计的影响。我们在 TreeSatAI-TS、PASTIS-HD（折 I）和 FLAIR-HUB（在 20 % 处过滤，使用分割 1）上进行了实验。我们比较包括 MAE 和 ViT 模型 [19]，以及一些基准 FM—DINO-v2 [60]，DINO-v2 卫星. [74]，DOFA [95]，和 CROMA [24] —其超参数在经过广泛调整后被仔细设置。

默认情况下，我们使用组融合模式，将 Sentinel-1 的升轨和降轨模式组合在一起。对于多光谱数据，我们基于 Sentinel-2、Sentinel-1 和航拍方式中表现出强烈组内相关性和较弱组间相关性的光谱波段组，通过在重建过程中应用联合令牌融合和分组补丁目标归一化。完整细节见 SM Secs. 6.3 and 6.4 。

我们在 Fig. 3 和 SM Tab. 11 中评估了 Fig. 2 中概述的不同多模态和多时态融合模式。

在多模态融合方面，我们观察到在相似的模态之间进行早期融合有小幅的益处，但在对不相似的模态进行融合时性能显著下降（所有模态组合在一起的 “group” 在三个数据集上表现最差）。此外，当使用晚期融合时，共享模态之间的编码器参数并没有一致地改善效果（“monotemp” 与 “shared” 相比）。

总体而言，这些结果表明，跨模态协同学习的潜在好处可能无法弥补模态特定专业化减少的缺点。这一发现——很可能反映了 EO 模态的强烈异质性——对旨在构建通用且与模态无关的 EO FMs 的策略表示怀疑。

关于多时态性，我们发现早期多时态融合（“mod” 、“group” 和 “inter-group” ）在 TreeSatAI-TS（加权 F1）和 PASTIS-HD（mIoU）上稳定地比晚期多时态融合（“shared” 和 “monotemp” ）表现出色 2-3 %，在 FLAIR-HUB（mIoU）上提高 1 %。这种性能提升在时间动态起关键作用的任务中尤为显著（见 SM Sec. 7.5 ）。这种优势对于 MAEs 来说更加明显，而不是监督的 ViTs，这表明有效利用时间动态既重要又不容易。

这些发现强调了一个在多时间 SSL 中可能被忽视的机会，其在之前的工作中受到的关注少于多模态 SSL。值得注意的是，大多数现有 FMs 天生是单时间的，因此仅兼容后期多时间融合，导致与涉及早期多时间融合的模型相比存在显著的性能差距。

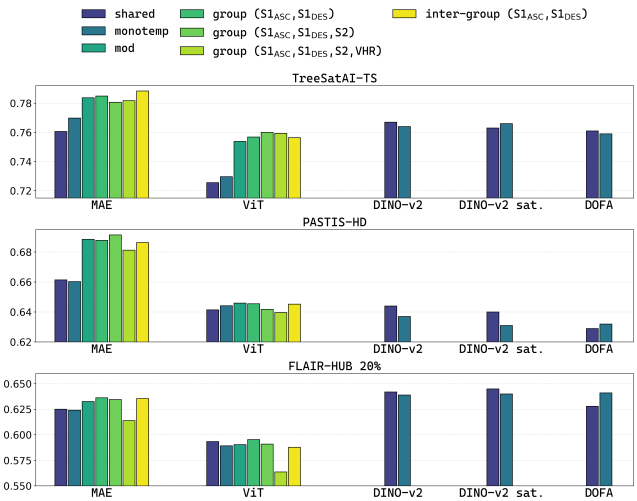


Figure 3. 比较不同多模态和多时态融合模式在 MAEs、ViTs 和基线 FMs 上的表现。我们报告了 TreeSatAI-TS 上的加权 F1 得分（%）以及 PASTIS-HD 和 FLAIR-HUB 20 上的 mIoU（%）。具体数值和使用 CROMA 的附加结果请参见 SM Tab. 11 。

多光谱融合与目标归一化。 在 Fig. 4 和 SM Tab. 12 中，我们评估了多光谱融合和目标归一化的不同选择。在这里，我们将评估范围限定在 TreeSatAI-TS 和 PASTIS-HD（使用折叠 I），因为 FLAIR-HUB 的性能主要受航空模态驱动（见 SM Sec. 7.6 ），而航空模态仅表现出较弱的多光谱特性。

我们首先评估了联合标记融合中不同目标归一化选择的影响，其中所有光谱波段组被组合为相同的标记。与之前的工作 [34, 35] 一致，逐块归一化提高了性能，但我们提出的逐块组归一化方法显著提升了结果。这种归一化支持有效的迁移学习，即使只使用 20 % 或 5 % 的预训练数据，如 Fig. 4 所示。总体来看，我们的研究结果强调了目标归一化对于充分利用更大型预训练数据集是必不可少的。

令人惊讶的是，基于补丁组的规范化结合联合 token 融合能够达到，甚至有时超过基于补丁规范化结合基于 token 融合的性能，其中每个光谱波段组形成一个单独的 token [9, 13, 35, 37, 62, 76, 77] 。在基于 token 的融合中，由于每个组对应一个不同的 token，因此补丁规范化自然适用于每个光谱组，使这两种方法在目标规范化方面等价。

这一结果尤其值得注意，因为它在计算上的影响是显著的：虽然基于标记的融合带来的开销随着光谱组数量的线性增长而增加，但结合了补丁组逐步归一化和联合标记融合的方法在几乎没有额外开销的情况下实现了相似的性能。

据我们所知，分块归一化提高性能的根本原因仍然不清楚，即使对于那些重构 RGB 图像或视频的自监督学习方法也是如此。一个假设是，它通过确保各个块之间的难度和损失贡献的一致性来帮助平衡前置任务。

然而，分块归一化可能无法在多光谱地球观测数据上提供这样的平衡。例如，如果不同波段的直方图窄但不重叠，那么分块归一化往往等同于应用一个不依赖块的恒定归一化。分组分块归一化通过适应归一化到块和光谱组来解决这个问题。

我们假设这一平衡需求是像素空间生成 SSL 所特有的。相比之下，潜在空间生成 SSL 方法通常通过在损失 [3] 之前的层（如 LayerNorm）进行归一化，或在损失内通过 softmax [60, 80, 99] 或余弦相似度 [77] 来隐式实现平衡。

总体而言，我们的研究结果表明，对于多光谱数据，确保预训练任务的平衡可能比使用复杂的融合策略更为重要。

我们对遮盖策略的评估在 SM Sec. 7.4 中给出。简而言之，遮盖策略的影响较小，但沿每个轴的结构化遮盖在探测过程中可以产生显著的好处。

3.2. MAESTRO 的设计

根据 Sec. 3.1 中的分析，我们确定了新的方法的 SSL 策略和模型组件的选择。我们采用组间或组内融合模式，将 Sentinel-1 上升和下降方式组合在一起。这一选择使得能够在时间步和 Sentinel-1 模态对之间进行早期融合，同时允许在其他模态对之间进行中间或后期融合。在多光谱方面，我们在重建过程中使用联合标记融合以及补丁分组目标归一化。

我们将我们的方法命名为 MAESTRO，以反映它如何在 MAE 框架内调控 EO 数据中的多模态、多时间和

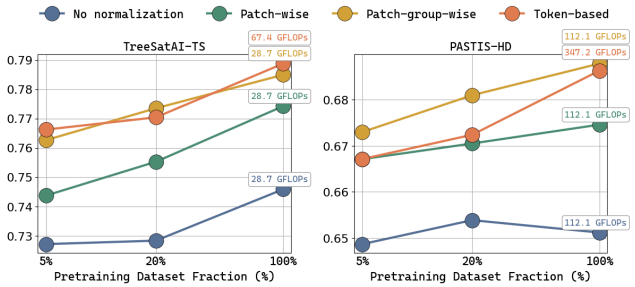


Figure 4. 比较不同的多光谱融合选择和目标归一化对 MAE-B 模型的影响。我们报告了在 TreeSatAI-TS 上的加权 F1 分数（%）以及在 PASTIS-HD 上的 mIoU（%），这些分数是随着预训练数据集比例的变化而变化的。对于每个数据集和多光谱融合的选择，我们还指出了每次前向传递（单个批元素）的 GFLOPs 预训练成本。具体数值请参阅附件 Tab. 12 和 Tab. 20 。

多光谱组件之间的复杂相互作用，如 Fig. 1 所示。

3.3. 性能

我们在 Tab. 1 中评估 MAESTRO 与基准 FMs 和监督 ViTs 在四个评估数据集上的性能。

对于监督 ViTs，我们采用组内或组间融合模式，将 Sentinel-1 升轨和降轨模式进行分组，就像在 MAESTRO 中所做的那样。对于基准 FM，受 Fig. 3 和 SM Tab. 11 结果的指导，我们选择共享融合模式，除了 CROMA，我们使用跨 CROMA 模式（见 SM Sec. 6.4.3 ）。如前所述，基准 FM 已通过超参数搜索进行了广泛微调以确保公平比较（见 SM Sec. 6.4 ）。

MAESTRO 与 SOTA 对比。 MAESTRO 在两个评估数据集上设定了新的最先进水平（SOTA），并且在剩下的两个数据集上几乎与之前的 SOTA 或表现最佳的基线 FMs 相匹配。具体来说，它在 TreeSatAI-TS 上以 +2.7 %（加权 F1）超越了之前的 SOTA，在 PASTIS-HD 上以 +2.5 %（mIoU）超越，而在 FLAIR # 2 上仅以-0.2 %（mIoU）和在 FLAIR-HUB 上仅以-0.1 %（mIoU）落后于我们对 DINO-v2 的改编（参见 SM Sec. 6.4.1 ），该改编在这两个数据集上设定了新的 SOTA。

MAESTRO 相比监督学习的 ViTs 表现出更大的优势：在 TreeSatAI-TS 上加权 F1 提高了 3.7%，在 PASTIS-HD 上的 mIoU 提高了 4.4%，在 FLAIR 2 上的 mIoU 提高了 5.7%，在 FLAIR-HUB 上的 mIoU 提高了 3.8%。值得注意的是，即使在非常大规模的 FLAIR-HUB 数据集上，MAESTRO 对 ViTs 仍保持着显著的优势，这是其他 SSL 方法不一定能观察到的。总体而言，这些结果表明，从 SSL 预训练中进行迁移学习对于实现良好的微调性能是至关重要的。

在 TreeSatAI-TS 和 PASTIS-HD 上，

MAESTRO 与基线 FMs 的比较。 MAESTRO 显著优于基线 FMs，这主要是由于这些基线 FMs 仅支持的后期多时相融合策略的局限性对这些数据集特别不理想（见 SM Sec. 7.5 ）。然而，这种优势并没有延伸到 FLAIR # 2 和 FLAIR-HUB，在这些数据集中，MAESTRO 的性能与表现最佳的基线 FMs 持平或略低。这很可能是因为这些数据集主要依赖于单时间点的航空影像（见 SM Sec. 7.6 ），因此多时相融合策略的选择影响减弱。

影响这种比较的其他因素可能会相互抵消。例如，基线 FMs 通常在更大、更平衡的数据集 [60] 上使用计算量更大的方法进行预训练。然而，相对于微调，预训练模态或预处理流程中的差异可能会限制迁移的有效性。

有趣的是，可能也是不出所料，当微调数据集中占主导地位的模态与预训练期间使用的模态紧密匹配时，基线 FMs 表现最佳（见 Tab. 1 和 SM Tab. 3 ）。

按数据集大小缩放。 在 Fig. 5 和 SM Tab. 14 中，我们最终评估了 MAESTRO 和监督型 ViTs 在不同的预训练和微调数据集比例下的表现。MAESTRO 始终优于监督型 ViTs，尤其是在微调数据有限的情况下。比较完整和缩减的预训练数据集，我们发现更多的无标签数据系统地提高了下游性能，不论标签微调数据的数量多少。

这些结果表明，MAESTRO 可能会从大规模无标签的预训练中获得进一步的好处。然而，充分利用如此大规模的数据集可能需要额外的平衡策略——除了我们的目标归一化外——以适应大的预训练任务 [42, 79] 。

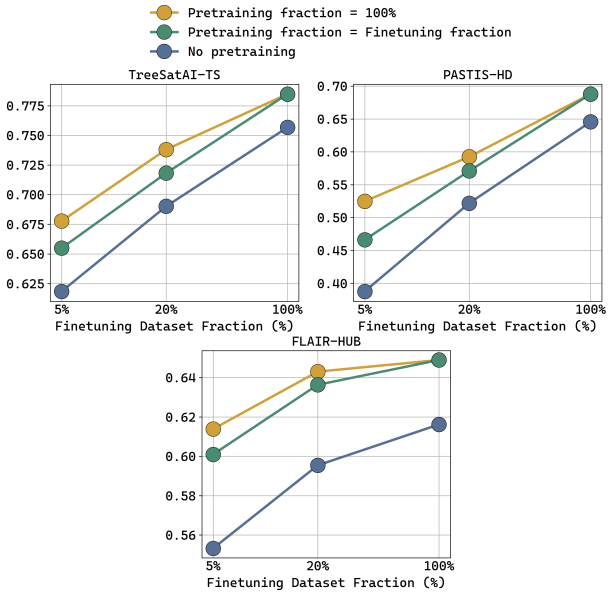


Figure 5. 对 MAESTRO-B 和 ViT-B 模型进行不同预训练/微调数据集比例的扩展。我们报告在 TreeSatAI-TS 上的加权 F1 分数（%）以及在 PASTIS-HD 和 FLAIR-HUB 上的 mIoU（%），对于三种微调数据集比例：5 %，20 %，和 100 %。对于每种微调比例，我们比较三种预训练设置：在 100 % 的数据上进行预训练、在与微调相同的比例上进行预训练、以及没有预训练。具体数值请参阅 SM Tab. 14 。

Table 1. 在四个评估数据集上对 MAESTRO、基础 FMs 和监督式 ViTs 的性能进行比较。我们报告了 TreeSatAI-TS 上的加权 F1 分数（%）和 PASTIS-HD、FLAIR # 2 以及 FLAIR-HUB 上的 mIoU（%）。箭头（ / ）表示与最佳的其他 FM/ViT/SOTA 相比的收益/损失。除 AnySat 和先前的 SOTA（ ⋆ ）外，我们已经训练了此表中列出的所有模型。MAESTRO † 模型已进行两倍纪元的预训练。

Model	Model size	Fusion mode	TreeSatAI-TS	PASTIS-HD fold I	FLAIR # 2 split 1	FLAIR-HUB split 1
MAESTRO (ours)	Base	group	78.5	68.8	63.8	64.9
MAESTRO (ours)	Base	inter-group	78.8	68.6	62.6	65.9 ↓ 0.1
MAESTRO (ours) †	Base	group	79.1	68.8	64.0 ↓ 0.2	64.8
MAESTRO (ours) †	Base	inter-group	79.4 ↑ 2.7	69.0 ↑ 2.5	63.3	65.8
DINO-v2 [60]	Base	shared	76.7	64.4	64.2	66.0
DINO-v2 sat. [74]	Large	shared	76.3	64.0	63.5	66.0
DOFA [95]	Base	shared	76.0	62.9	62.3	65.1
CROMA [24]	Base	inter-croma	70.5	65.0	39.0	44.3
AnySat ⋆ [5]	Base		75.1	66.5	55.1	
ViT	Base	group	75.7	64.6	58.3	61.6
ViT	Base	inter-group	75.6	64.5	58.2	62.1
Previous SOTA ⋆			75.1 [5]	66.5 [5]	64.1 [69]	64.3 [28]

4. 结论

在这项工作中，我们探讨了如何将 SSL 方法适应 EO 数据的独特特征，重点关注异质性的三个方面：多模态性、多时间性和多光谱性。

(i) 多模态性——当模态相似时，早期融合略有优势，但在模态显著不同的情况下可能损害性能，对通用的无关模态的 EO 基础模型的可行性提出了疑问。(ii) 多时态性——对于强烈依赖多时态动态的任务，早期融合始终优于后期融合，揭示了多时态 SSL 中的一个未充分探索的机会。目前的基础模型由于本质上是单时态的，在微调时仅限于晚期的多时态融合，这可能导致显著的性能下降。(iii) 多光谱性——结合适当的目标归一化进行重建的联合标记早期融合可以达到基于标记的融合的性能，同时效率显著更高。

基于这些发现，我们提出了 MAESTRO——一种适应 MAE 的方法，能够有效地协调在地球观测中使用多模态、多时序和多光谱数据。通过在四个地球观测基准上进行评估，MAESTRO 在强烈依赖时间动态的任务中实现了 SOTA 结果，同时在其他任务中仍然保持高度竞争力。

我们希望我们的工作能够为设计专门考虑到地球观测数据独特特征的 SSL 策略做出贡献。

5.

致谢 本工作获得了由 GENCI-IDRIS 提供的 HPC/AI 资源的访问权限（授权编号：A0181013803, A0161013803, 和 AD010114597R1）。我们感谢 IGN 的 SIMV/SDM 同事 Guillaume Astruc、Loïc Landrieu、Alexandre Tuel 和 Thomas Kerdreux 的有益讨论。

References

[1] S. Ahlswede, C. Schulz, C. Gava, P. Helber, B. Bischke, M. Förster, F. Arias, J. Hees, B. Demir, and B. Kleinschmit. TreeSatAI benchmark archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, 15(2):681–695, 2023.

[2] Firouz A Al-Wassai and NV Kalyankar. Major limitations of satellite images. *Journal of Global Research in Computer Science*, 4(5):51–59, 2013.

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[4] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision (ECCV)*, pages 409–427. Springer, 2024.

[5] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An earth observation model for any resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF international conference on computer vision (CVPR)*, 2025.

[6] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.

[7] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.

[8] Hanbo Bi, Yingchao Feng, Boyuan Tong, Mengyu Wang, Haichen Yu, Yongqiang Mao, Hao Chang, Wenhui Diao, Peijin Wang, Yue Yu, et al. RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation. *arXiv preprint arXiv:2504.03166*, 2025.

[9] Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. FoMo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 27858–27868, 2025.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[11] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023.

[12] Shabnam Choudhury, Yash Salunkhe, Sarthak Mehrotra, and Biplab Banerjee. REJEPA: A novel joint-embedding predictive architecture for efficient remote sensing image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2373–2382, 2025.

[13] 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.

[14] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. TerraFM: A scalable foundation model for unified multisensor earth observation. *arXiv preprint arXiv:2506.06281*, 2025.

[15] Rangel Daroya, Elijah Cole, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. WildSAT: Learning satellite image representations from wildlife observations. *arXiv preprint arXiv:2412.14428*, 2024.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[17] Wenhui Diao, Haichen Yu, Kaiyue Kang, Tong Ling, Di Liu, Yingchao Feng, Hanbo Bi, Libo Ren, Xuexue Li, Yongqiang Mao, et al. RingMo-Aerial: An aerial remote sensing foundation model with a affine transformation contrastive learning. *arXiv preprint arXiv:2409.13366*, 2024.

[18] Philipe Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, H Lexie Yang, and Dalton Lunga. OReole-FM: successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 597–600, 2024.

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

[20] Chuc Man Duc and Hiromichi Fukui. SatMamba: Development of foundation models for remote sensing imagery using state space models. *arXiv preprint arXiv:2502.00435*, 2025.

[21] Iris Dumeur, Silvia Valero, and Jordi Inglada. Self-supervised spatio-temporal representation learning of satellite image time series. *IEEE Journal of Selected Topics in*

Applied Earth Observations and Remote Sensing, PP:1–18, 2024.

[22] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12371–12384, 2024.

[23] Fajwel Fogel, Yohann Perron, Nikola Besic, Laurent Saint-André, Agnès Pellissier-Tanon, Martin Schwartz, Thomas Boudras, Ibrahim Fayad, Alexandre d'Aspremont, Loic Landrieu, et al. Open-Canopy: Towards very high resolution forest monitoring. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1395–1406, 2025.

[24] Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[25] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. FLAIR# 1: semantic segmentation and domain adaptation dataset. *arXiv preprint arXiv:2211.12979*, 2022.

[26] Anatol Garioud, Apolline De Wit, Marc Poupée, Marion Valette, Sébastien Giordano, and Boris Wattrelos. FLAIR# 2: textural and temporal information for semantic segmentation from multi-source optical imagery. *arXiv preprint arXiv:2305.14467*, 2023.

[27] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. FLAIR: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36:16456–16482, 2023.

[28] Anatol Garioud, Sébastien Giordano, Nicolas David, and Nicolas Gonthier. FLAIR-HUB: Large-scale multimodal dataset for land cover and crop mapping. *arXiv preprint arXiv:2506.07080*, 2025.

[29] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pages 171–181. Springer, 2020.

[30] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[31] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305, 2022.

[32] 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.

[33] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27852–27862, 2024.

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021.

[35] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] Huiyang Hu, Peijin Wang, Hanbo Bi, Boyuan Tong, Zhaozhi Wang, Wenhui Diao, Hao Chang, Yingchao Feng, Ziqi Zhang, Yaowei Wang, et al. RS-vHeat: Heat conduction guided efficient remote sensing foundation model. *arXiv preprint arXiv:2411.17984*, 2024.

[37] Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y Ng. USat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199*, 2023.

[38] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[39] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Self-supervised learning for invariant representations from multi-spectral and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:7797–7808, 2022.

[40] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[41] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. TerraMind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

[42] Thomas Kerdreux, Alexandre Tuel, Quentin Febvre, Alexis Mouche, and Bertrand Chapron. Efficient self-supervised learning for earth observation via dynamic dataset curation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 3017–3027, 2025.

[43] Jingtao Li, Yingyi Liu, Xinyu Wang, Yunning Peng, Chen Sun, Shaoyu Wang, Zhendong Sun, Tian Ke, Xiao Jiang, Tangwei Lu, et al. HyperFree: A channel-adaptive and tuning-free foundation model for hyperspectral remote sensing imagery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23048–23058, 2025.

[44] Shuaipeng Li, Penghao Zhao, Hailin Zhang, Xingwu Sun, Hao Wu, Dian Jiao, Weiyan Wang, Chengjun Liu, Zheng Fang, Jinbao Xue, et al. Surge phenomenon in optimal learning rate and batch size scaling. *Advances in Neural Information Processing Systems*, 37:132722–132746, 2024.

[45] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024.

[46] Xuyang Li, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. SeaMo: A multi-seasonal and multimodal remote sensing foundation model. *arXiv preprint arXiv:2412.19237*, 2024.

[47] Xuyang Li, Chenyu Li, Pedram Ghamisi, and Danfeng Hong. FlexiMo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*, 2025.

[48] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *European Conference on Computer Vision*, pages 260–278. Springer, 2024.

[49] Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[50] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6252–6261, 2023.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[52] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5270, 2023.

[53] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35:7697–7711, 2022.

[54] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.

[55] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.

[56] Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024.

[57] Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

[58] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024.

[59] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multispectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024.

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[61] Jonathan Prexl and Michael Schmitt. Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2136–2144, 2023.

[62] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *DAGM German Conference on Pattern Recognition*, pages 317–331. Springer, 2024.

[63] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[64] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[65] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: mission critical - satellite data is a distinct modality in machine learning. 2024.

[66] Xian Shuai, Yiding Wang, Yimeng Wu, Xin Jiang, and Xiaozhe Ren. Scaling law for language models training considering batch size. *arXiv preprint arXiv:2412.01505*, 2024.

[67] Michael J Smith, Luke Fleming, and James E Geach. EarthPT: a time series foundation model for earth observation. In *NeurIPS CCAI workshop*, 2023.

[68] Caleb S Spradlin, Jordan A Caraballo-Vega, Jian Li, Mark L Carroll, Jie Gong, and Paul M Montesano. SatVision-TOA: A geospatial foundation model for coarse-resolution all-sky remote sensing imagery. *arXiv preprint arXiv:2411.17000*, 2024.

[69] Jakub Straka and Ivan Gruber. Modernized training of u-net for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 776–784, 2024.

[70] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinfglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo a remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023.

[71] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024.

[72] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-Scale MAE: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023.

[73] Jiayuan Tian, Jie Lei, Jiaqing Zhang, Weiying Xie, and Yunsong Li. SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[74] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, et al. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 300:113888, 2024.

[75] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[76] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

[77] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Forty-second International Conference on Machine Learning*, 2025.

[78] Diego Velazquez, Pau Rodriguez, Sergio Alonso, Josep M Gonfaus, Jordi Gonzalez, Gerardo Richarte, Javier Marin, Yoshua Bengio, and Alexandre Lacoste. EarthView: A large scale remote sensing dataset for self-supervision. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1228–1237, 2025.

[79] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

[80] Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2204–2214, 2025.

[81] 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.

[82] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. MTP: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[83] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[84] F Wang, H Wang, D Wang, Z Guo, Z Zhong, L Lan, J Zhang, Z Liu, and M Sun. Scaling efficient masked image modeling on large remote sensing dataset. *arXiv preprint arXiv:2406.11933*, 2024.

[85] Fengxiang Wang, Hongzhen Wang, Yulin Wang, Di Wang, Mingshuo Chen, Haiyan Zhao, Yangang Sun, Shuo Wang, Long Lan, Wenjing Yang, et al. RoMA: Scaling up mamba-based foundation models for remote sensing. *arXiv preprint arXiv:2503.10392*, 2025.

[86] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142. IEEE, 2022.

[87] Yuelei Wang, Ting Zhang, Liangjin Zhao, Lin Hu, Zhechao Wang, Ziqing Niu, Peirui Cheng, Kaiqiang Chen, Xuan Zeng, Zhirui Wang, et al. RingMo-lite: A remote sensing multi-task lightweight network with cnn-transformer hybrid framework. *arXiv preprint arXiv:2309.09003*, 2023.

[88] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer, 2024.

[89] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Multilabel guided soft contrastive learning for efficient earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[90] 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.

[91] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified Copernicus foundation model for earth vision. *arXiv preprint arXiv:2503.11849*, 2025.

[92] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Extending global-local view alignment for self-supervised learning with remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2443–2453, 2024.

[93] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Luo Junwei, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, pages 1–15, 2025.

[94] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: a simple framework for masked image modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2021.

[95] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

[96] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[97] Lixian Zhang, Yi Zhao, Runmin Dong, Jinxiao Zhang, Shuai Yuan, Shilei Cao, Mengxuan Chen, Juepeng Zheng, Weijia Li, Wei Liu, et al. A2-MAE: A spatial-temporal-spectral unified remote sensing pre-training method based on anchor-aware masked autoencoder. *arXiv preprint arXiv:2406.08079*, 2024.

[98] Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. SkySense V2: A unified foundation model for multi-modal remote sensing. *arXiv preprint arXiv:2507.13812*, 2025.

[99] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

MAESTRO：用于多模态、多时态和多光谱地球观测数据的掩码自动编码器

Supplementary Material

附录的结构如下：Sec. 6 提供了我们完整的实验细节；Sec. 7 报告了在 Figs. 3 to 5 中展示的实验的详细结果以及附加的消融研究；Sec. 8 对推理结果进行了定性分析；Sec. 9 报告了计算成本。

6. 实验细节

在本节中，我们提供完整的实验细节。我们提供关于四个评估数据集（Sec. 6.1 ）的细节，各模型共享的细节（Sec. 6.2 ），特定于 MAEs/MAESTRO 的细节（Sec. 6.3 ），以及特定于基线 FMs 的细节（Sec. 6.4 ）。此外，我们还提供了报告超参数值详尽列表的表格（Sec. 6.5 ）。

6.1. 数据集的实验细节

在本小节中，我们提供了关于四个评估数据集（ Secs. 6.1.1 to 6.1.4 ）中输入模态和标签选择的附加细节。我们还报告了应用于这些数据集的一些输入模态的原始预处理步骤的详细信息（**??** ）。

6.1.1. TreeSatAI-TS

TreeSat 于 [1] 年引入，其特点是具有 0.2 米分辨率的航空影像（RGB + NIR），以及单时相的 Sentinel-1 和 Sentinel-2 数据（Sentinel-2 以季节性中值提供）。它后来在 [4] 年被扩展为 TreeSatAI-TS，增加了覆盖最接近航空拍摄全年的 Sentinel-1 和 Sentinel-2 时间序列。在这项工作中，我们保留了四种不同的模态：航空影像、两条轨道（升轨和降轨）的 Sentinel-1 时间序列，以及 Sentinel-2 时间序列。我们忽略了原始的单时相 Sentinel-1 和 Sentinel-2 数据。

[1] 中引入的原始标签被制定为回归目标，表示每个树种在一个区域内的空间比例。根据 [4, 5] ，我们将其重新表述为一个多标签分类任务：如果某个物种的空间比例超过 0.07，则认为该物种类存在。

6.1.2. PASTIS-HD

PASTIS 首次在 [30] 中提出，该版本使用了历时约一年的 Sentinel-2 时间序列。之后在 [31] 中，PASTIS-R 通过加入覆盖大约 70 个日期的 Sentinel-1 时间序列（包括升轨和降轨）得以扩展，随后在 [4] 中，PASTIS-HD 通过将 SPOT 6-7 影像重采样至 1 米分辨率进一步扩展。我们保留了四种不同的模态：SPOT 6-7 影像、包括升轨和降轨的 Sentinel-1 时间序列，以及 Sentinel-2 时间序列。

[30] 中的作物分割标签涵盖了语义分割和全景分割。在这里，我们仅保留语义分割标签，省略全景分割。限制为语义分割使我们能够引入一个空间化微调任务，而不会添加与专用头或损失函数相关的额外复杂性。我们假设任务的空间化特性（例如，语义分割与分类）比具体的分割类型（例如，全景与语义）在标记不同的 SSL 方法时起更为决定性的作用。

6.1.3. Flair # 2

在 [25] 中引入了 FLAIR，具有 77,762 张分辨率为 0.2 米的航空和高程影像（RGB + NIR + DSM）。它在 [26, 27] 扩展为 FLAIR # 2，包括覆盖全年时间序列形式的 Sentinel-2 超补丁，其空间覆盖范围比航空影像更大（400 米 × 400 米对比 102.4 米 × 102.4 米），以提供额外的空间上下文。根据 [5] ，我们裁剪 Sentinel-2 时间序列以匹配 VHR 影像的范围，弃用 93.5 % 的像素。我们还包括从 FLAIR-HUB 扩展 [28] 中提取的分辨率为 0.2 米的高程影像，用于 77,762 张 FLAIR # 2 瓦片。总共，我们保留三种不同的数据模式：航空影像（RGB + NIR）、高程影像（DEM + DSM）和 Sentinel-2 时间序列。

我们使用土地覆盖语义分割标签，根据 [25–27] 中描述的过滤过程：我们仅保留 18 个原始类别中的 12 个，将剩下的 6 个类别从损失和指标计算中排除。

6.1.4. FLAIR-HUB

FLAIR-HUB [28] 是 FLAIR # 2 的扩展，包含 241,100 块面积为 102.4 m × 102.4 m 的图块。与 FLAIR # 2 相比，FLAIR-HUB 丰富了高程影像（DEM + DSM），并将其视为单独的模态；将 Sentinel-2 影像裁剪至航空影像的范围（如我们在 FLAIR # 2 版本中所做的）；此外还添加了一整年的 Sentinel-1 时间序列，以及 SPOT 6–7 影像和历史航空影像。我们保留了五种不同的模态：航空影像（RGB + NIR）、高程影像（DEM + DSM）、两轨道的 Sentinel-1 时间序列和 Sentinel-2 时间序列。我们忽略了 SPOT 6–7 影像，这并未显著改善微调性能——这可能是因为与航空模态的冗余——以及历史影像，由于其与标签的不同步性。

我们按照 [28] 中的过滤步骤使用土地覆盖语义分割标签：我们保留 18 个原始类中的 15 个，将其余 3 个从损失和度量计算中排除。我们省略了作物语义分割标签。

在 TreeSatAI-TS、FLAIR # 2 和 FLAIR-HUB 中，哨兵-1 的后向散射系数数据以线性尺度提供。我们应用对数变换将其表达为分贝（dB）尺度（乘以一个常数）。对于 PASTIS-HD，哨兵-1 数据已经是 dB 尺度；然而，我们只保留与垂直（VV）和水平（VH）极化相对应的前两个通道，并舍弃表示它们比率（VV/VH）的第三个通道。

在 FLAIR # 2 和 FLAIR-HUB 中，我们还对高程影像（DEM + DSM）应用了一个简单的预处理步骤。我们将其重新定义为 DEM 和重新缩放的高程 $10^3 \times$ (DSM − DEM)，以确保两个通道具有相似的数值范围。

TreeSatAI-TS 航拍影像图块的空间范围略大于其他模态（304 × 304 像素，对应 60.8 米，而不是 300 × 300 像素）。因此，我们应用居中裁剪来使其在空间上与其他输入对齐。

6.2. 实验细节在模型之间共享

在这一小节中，我们提供了关于预处理管道（Sec. 6.2.1 ）、位置/时间编码（Sec. 6.2.2 ）、数据增强（Sec. 6.2.3 ）、正则化（**??** ）和优化器（Sec. 6.2.5 ）的详细信息。

6.2.1. 预处理

这里我们详细介绍在 Sec. 2.1 中介绍的预处理流程。完整的实现可以在我们代码中的 `ssl\_models/dataset/dataset.py` 找到。

数据集特定的裁剪超参数 每个数据集 $\mathcal{D}$ 都有一个特定的超参数来控制原始图块内作物的空间范围。对于分类任务，标签适用于整个图块，我们将作物范围设置为等于原始图块范围。对于分割任务，作物范围可以更小，对分割标签应用相同的作物。选择作物大小需要平衡两个相互竞争的效果：

- 较小的裁剪会在固定的令牌预算下增加令牌网格的空间分辨率。对于固定的令牌预算 $(I_m/P_m)^2 D_m$ ，减小图像大小 I_m 会减小补丁大小 P_m ，从而降低分词器（模型的入口模块）中信息瓶颈的风险。
- 更大的裁剪为模型提供了更多的空间上下文。

当裁剪至比原始瓦片更小的范围时，我们为每个 epoch 应用一个重复因子 $R_{\mathcal{D}}$ ，该因子等于原始瓦片面积与裁剪后面积的比率。这确保了模型在每个 epoch 中看到的总空间面积与完整数据集相匹配。实际上，这意味着数据集的长度按 $R_{\mathcal{D}}$ 进行缩放：每个瓦片，原本映射到一个索引，现在与 $R_{\mathcal{D}}$ 个索引关联，有效地以不同的裁剪方式重复这么多次。

作物采样策略在训练和验证/测试之间有所不同：

- 训练：从瓦片内所有有效的裁剪中随机采样裁剪，唯一的限制是裁剪边界必须与每种模态的整数像素索引对齐 m 。
- 验证与测试：图块被划分为不重叠的区域，并被穷尽地迭代。在测试阶段应用相同的重复因子，确保测试指标覆盖与原始图块完全相同的空间范围。

在实际操作中，我们对 TreeSatAI-TS、FLAIR # 2 和 FLAIR-HUB 使用全瓦片范围。然而，对于 PASTIS-HD，我们发现大幅减少裁剪尺寸是有益的。这允许我们为 Sentinel-2 和 Sentinel-1 分别设置 $P_m = 2$ ，连同 $D_m = 16$ 和 $D_m = 4$ 时间段。如果不进行裁剪，要保持相同的标记预算和时间段数量，对于这些模式将需要 $P_m = 16$ 。

数据集中每种模式 m 都有一组可配置的预处理超参数：

- 图像大小 I_m ，默认情况下对应于覆盖裁剪空间范围的像素数量（然而，可以为基线 FM 调整该大小，以匹配特定的传递补丁大小 P_m ）；
- 目标时间区间的数量 D_m ；
- 是否使用雪/云掩膜来指导在时间段内选择有效时间步长，以及用于忽视高于该阈值的掩膜值的概率阈值；
- 应用于模式的常数乘法归一化因子。

预处理步骤。 我们数据集的 `\_\_getitem\_\_` 方法分三步操作：

(i) 根据数据集索引确定采样的数据集切片；
(ii) 对所有模态共享的空间裁剪进行采样 $m \in \mathcal{D}$ ；
(iii) 根据采样的裁剪处理每种模态 m 。

对于每个模式 m ，处理步骤 (iii) 如下：

(a) 如果原始时间步长的数量 T_m 不是 D_m 的倍数，则随机截断该序列至 $D_m \times \lceil T_m/D_m \rceil$ 个连续的时间步长。
(b) 读取由作物的空间范围和（可能被截断的）时间范围定义的相应瓦片部分。为了最小化 I/O 开销，我们使用惰性加载：
  - 对于 `.tif` 文件：使用窗口读取通过 `rasterio` 进行读取。
  - 对于 `.npy` 文件：使用参数 `mmap\_mode="r"` 通过 `numpy.load` 读取。
  - 对于 `.h5` 文件：以读取模式使用 `h5py.File` 进行读取，并索引所需的数组部分。
(c) 将数组重塑为 D_m 个时间区间。
(d) 对于每个时间槽，从槽内随机抽取一个时间步：
  - 如果使用雪/云掩膜，则过滤掉包含任何高于配置阈值的掩膜值的时间步骤。当没有有效的时间步骤剩余时，

出现一个例外，此时保留箱中所有的时间步骤。
  - 从剩余的时间步中选择一个。在训练过程中，选择是随机的，以便作为数据增强。在验证和测试过程中，我们通过以下步骤来最大化代表性：计算时间区间内所有有效时间步的像素中位数；计算每个时间步相对于该中位数的平均绝对偏差；选择平均绝对偏差最小的时间步。

(e) 如果图像大小 I_m 不匹配覆盖裁剪空间范围的像素数，则使用 `torch.nn.functional.interpolate` 并使用 `mode="nearest"` 重新插值该数组的空间。

(f) 用配置好的常数归一化因子对数组进行缩放。

6.2.2. 位置/时间编码

如 Sec. 2.1 中所述，我们不包含显式模态编码，因为我们使用特定于模态的分词器和可学习的特定于模态的 `[mask]` 令牌，这些令牌隐式编码每个令牌的源或目标模态。然而，我们确实包含了空间和时间位置编码，具体如下。

为了编码空间信息，我们遵循通用做法，使用二维正弦-余弦位置编码。如在 Scale-MAE 中，我们根据每种模式的地面采样距离来缩放这些位置编码。具体来说，我们计算跨模式的标记网格大小的最小公倍数（LCM），即 $\text{lcm}_{m\in\mathcal{D}}\{I_m/P_m\}$，并在由该 LCM 定义的高分辨率网格上生成位置编码。对于每个模式 m，我们通过对对应于每个低分辨率网格单元的所有高分辨率网格单元进行平均，来获取其位置编码，通过对高分辨率网格进行下采样。

为了编码时间信息，我们首先在每个时间区间中提取所选择时间步的年中天和一天中的时间（见 Sec. 6.2.1）。年中天通过 365.25 标准化，一天中的时间通过 24 标准化。对这些标准化值应用正弦和余弦变换为每个时间步生成四个时间特征。

此外，对于每个区块，我们定义一个在所有模态 m 中共享的参考日期，以捕捉跨年度的时间差异，而不仅仅是在单一年份内（正如使用正弦/余弦编码所受到的限制）。在每个时间步中，我们计算其获取日期与这个参考日期之间的差异，从而获得第五个时间特征。这个值随后被复制四次，导致每个时间步总共拥有八个时间特征。

聚合。 空间和时间编码通过串联进行聚合。在具有潜在维度 C_e 的编码器中，空间编码占据 $C_e - 8$ 维度，而时间编码使用 8 维度，串联后导致 C_e 维度。同样的方法应用于具有潜在维度 C_d 的解码器中。

6.2.3. 数据增强

我们使用多达三种类型的数据增强：

- 随机空间裁剪：在训练期间，在预处理开始时，会从原始图块中随机采样一个配置范围的裁剪区域。这在验证和测试阶段是禁用的，其中图块被划分为不重叠的裁剪区域，并在每个周期中完全处理。当裁剪范围与原始图块范围匹配时，即使在训练期间，这种增强也会被隐式禁用。
- 随机时间步选择：在训练过程中，从每个时间区间内的有效步中随机采样时间步。在验证和测试中，我们选择在有效步中具有最大代表性的时间步。
- D4 增强: 在训练、验证和测试过程中，经过预处理管道后，在所有模态和语义分割标签上应用同步的 D4 变换。我们在验证和测试中保留这一增强，因为假设模型在训练期间已经学习了对此类变换的等变性。

6.2.4. 正则化

作为正则化，我们在微调期间应用模型权重的指数移动平均（EMA）[56]，类似于随机权重平均 [38]。具体来说，权重的 EMA 在每个时期都会更新，其平滑窗口等于总微调时期的百分之二十%。设模型权重为 θ，其 EMA 为 θ_{EMA}，EMA 权重初始化为 θ，并在每个时期更新为：

$$\theta_{\text{EMA}} = \alpha\theta_{\text{EMA}} + (1-\alpha)\theta, \qquad \alpha = 1 - (0.2 \times N_{\text{epochs}})^{-1}.$$

在验证和测试期间，我们使用 EMA 权重而不是常规模型权重。

6.2.5. 优化器

在所有实验中，我们使用 AdamW 优化器 [51]。学习率随着批大小的平方根进行缩放 [44, 53, 66, 96]；也就是说，我们将学习率设置为基础学习率乘以批大小的平方根。然而，对于给定的数据集和训练阶段（即 SSL 预训练、探测或微调），我们在所有实验中保持批大小和学习率不变，以避免任何混杂因素。

在所有阶段——SSL 预训练、探测和微调中，我们使用余弦衰减学习率调度器，具有单一周期和预热期。这种方法在最近使用 Transformer 的计算机视觉 SSL 研究中很常见 [10, 34, 60, 94]。我们使用 `torch.optim.lr\_scheduler.OneCycleLR` 实现这一点，采用其默认的退火策略，并在总轮数的 20 % 中进行预热。

在自监督学习预训练和探测过程中，学习率逐步衰减到其最大值的 1×10^{-4} 倍。在微调过程中，学习率则仅衰减到最大学习率的一半。这有助于充分利用 EMA 策略作为权重平均，因为 (i) 它自然地降低了噪声 [56]，并且 (ii) 当学习率保持足够高时，可以从训练轨迹中的更大模型多样性中获益。

6.3. 特定于 MAEs/MAESTRO 的实验细节

在本小节中，我们提供了关于 MAEs/MAESTRO 的实验细节。我们详述了掩码策略（Sec. 6.3.1 ）和用于块组归一化（Sec. 6.3.2 ）的频带组的选择。

6.3.1. 掩蔽策略

我们采用一个分为两个阶段的屏蔽策略：

(i) 结构化掩码：

(a) 模态结构：以固定概率 0.25 遮蔽每种模态；

(b) 空间结构：在每种模态中，以固定概率对每个空间位置进行遮罩 0.25 ；

(c) 时间结构：在每个模态中，以固定概率 0.25 遮掩每个时间位置。

(ii) 非结构化屏蔽：调整步骤 (i) 中的屏蔽以匹配总体 75 % 的屏蔽比例：

(a) 如果步骤 (i) 导致被掩盖的标记数过少，则随机掩盖其他未掩盖的标记；

(b) 如果步骤 (i) 导致被掩码的标记过多，则随机取消一些掩码。

我们的掩码策略与前人工作中提出的几种方法有关，概述如下：

- 步骤 (i-a) 引入模态结构化掩码，从概念上讲与 Fomo-Net [9] 中的带状掩码策略相似，尽管若要直接等同则需要将每个光谱波段视为一种独特的模态。
- 步骤 (i-b)，引入了空间结构化的掩码，与 SatMAE [13] 中使用的一致掩码和 VideoMAE [75] 中的管状掩码是一致的。这也类似于 SeaMo [46] 中的方法，但有一个关键区别：SeaMo 在单个时间步的跨模态中实施一致性，而我们在单个模态的时间步内部实施一致性。
- 引入时间结构化掩码的步骤 (i-c) 对应于 Presto 中的 “Timesteps” 策略 [76] 。
- 我们结合的掩蔽策略与 AnySat [5] 、EarthMAE [78] 和 Galileo [77] 中的策略有相似之处，但仍然有重要的区别。AnySat 和 Galileo 仅应用空间和时间结构的掩蔽，而 EarthMAE 只使用空间结构的掩蔽。此外，我们整合结构化和非结构化阶段的方法与这些方法不同。

6.3.2. 频带分组

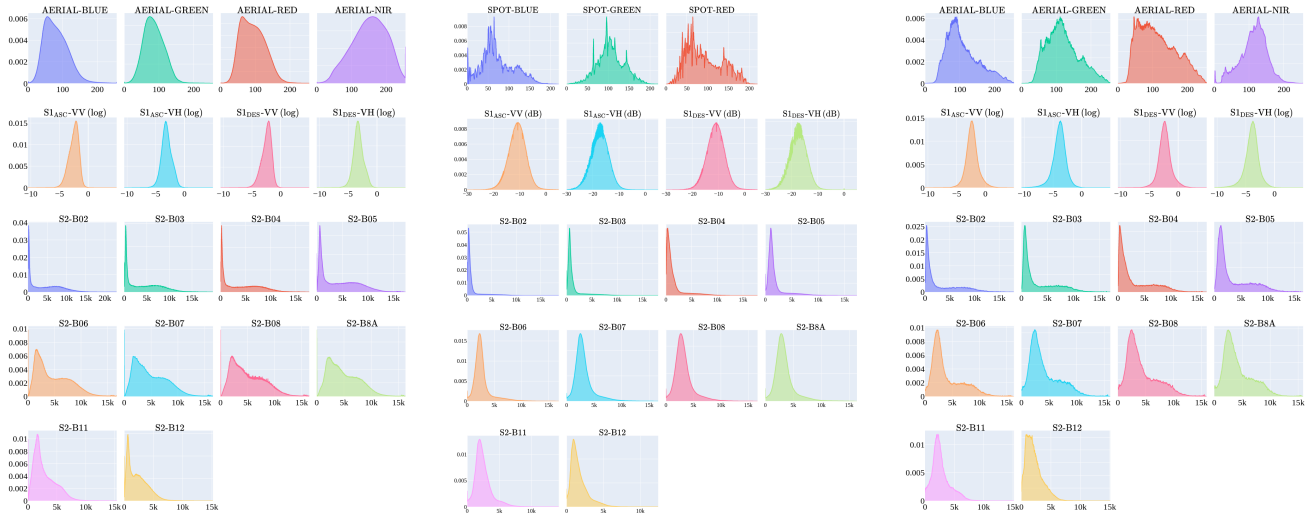
在此，我们提供了关于我们为补丁组归一化选择的光谱带组的详细信息。

我们首先在 TreeSatAI-TS、PASTIS-HD 和 FLAIR-HUB 上计算 VHR、Sentinel-1 和 Sentinel-2 模式的每波段直方图。为了降低计算成本，使用了包含 250 个瓦片的 FLAIR-HUB 玩具版本进行直方图计算。所得的直方图如 Fig. 6a 、Fig. 6b 和 Fig. 6c 所示。

对于超高分辨率航空影像，红、绿、蓝波段的直方图相对类似，而近红外波段则显示出不同的分布。

对于 Sentinel-1，VV 和 VH 的直方图明显不同，这与它们不同的极化响应一致；VH 的后向散射通常低于 VV。

对于 Sentinel-2，这十个波段聚集成三个自然的组，每组内部具有很强的相似性，但组间差异较大。



(a) TreeSatAI-TS (b) PASTIS-HD (c) FLAIR-HUB

Figure 6. TreeSatAI-TS、PASTIS-HD 和 FLAIR-HUB 数据集上的 VHR、Sentinel-1 和 Sentinel-2 模态的每波段直方图。

基于这些见解，我们定义用于分块组归一化的光谱频带群，详见 Tab. 2 。

Table 2. 为补丁组归一化选择的光谱波段组。

Modality	Group 1	Group 2	Group 3
Aerial	RED, GREEN, BLUE	NIR	
SPOT	RED, GREEN, BLUE		
$S1_{ASC}$ / $S1_{DES}$	VV	VH	
S2	B02, B03, B04, B05	B06, B07, B08, B8A	B11, B12

请注意，哨兵-2 波段组按照自然波长顺序排列，这与空间分辨率的顺序并不完全一致（例如，波段 B05 的分辨率为 20 米，而 B08 的分辨率为 10 米）。这与之前一些对 S2 波段进行分组的方法有所不同 [13, 37] 。

6.4. 基线 FMs 的实验细节

在本小节中，我们描述了在 MAESTRO 的比较评估中用作基准的 FMs，以及它们在我们背景下的具体调整。选定的模型包括 DINO-v2 sat. [74] ，CROMA [24] ，和 DOFA [95] ，它们在 EO 数据上进行了预训练，以及在自然图像上进行预训练的 DINO-v2 [60] 。

为了在我们的实验设置中优化这些基线模型的性能，我们引入了几个关键的调整：

(i) 输入调整大小：图像被调整大小以匹配 MAESTRO 的标记网格，同时保留原始的补丁大小（参见 Tabs. 5 to 7 ）。

(ii) 多模态标记化：多种模态通过并行的、模态特定的标记器同时处理。对于 DINO-v2 和 DINO-v2 sat.，标记器被调整以传递 RGB 权重，同时仍允许处理非 RGB 输入。

(iii) 后期融合：对于 DINO-v2、DINO-v2 sat. 和 DOFA，我们采用共享和单时间步融合模式在跨模态和时间步进行后期融合（参见 Sec. 2.1 ）。对于 CROMA，使用特定的多模态和多时间步融合模式。

(iv) 时间编码：由于这些基线 FMs 缺乏本地时间处理，Sec. 6.2.2 的时间编码在编码器之后被注入到嵌入令牌中。这个选择避免了干扰转移的表示。

所有基线 FMs 本身都使用 ViT 主干 [19] ，如同在 MAESTRO 中一样。我们还使用与 MAESTRO 中相同的分类和分割头。

关于预训练模态和选择用于微调的模态的详细信息在 Tab. 3 中提供。

Table 3. 不同基线 FMs 的预训练模态和选择用于微调的模态。

Model	Pre-training modalities	Fine-tuning modalities			
		VHR	$S1_{ASC}$ / $S1_{DES}$	S2	DEM/DSM
DINO-v2 [60]	Natural RGB images (LVD-142M dataset)	✓	✗	✓	✗
DINO-v2 sat. [74]	Maxar Vivid2 mosaic imagery	✓	✗	✓	✗
DOFA [95]	Sentinel-1, Sentinel-2, Gaofen, NAIP, EnMaP	✓	✓	✓	✗
CROMA [24]	Sentinel-1, Sentinel-2 (SSL4EO dataset)	✗	✓	✓	✗

6.4.1. DINO-v2 和 DINO-v2 sat。

DINO-v2 [60] 是一个在自然 RGB 图像上使用 SSL 教师-学生蒸馏框架预训练的 FM，它结合了全局（图像级）和局部（补丁级）目标。由于自然图像和地球观测图像之间的领域差距，将 DINO-v2 转移到 EO 任务上并不简单。

为了适应我们的多模态设置，我们修改了其 patchification 操作，以支持通过两个并行 patchification 层同时处理 VHR 和 Sentinel-2 输入。对于 PASTIS-HD、FLAIR # 2 和 FLAIR-HUB，我们保留了预训练的 RGB 通道权重，而额外通道则被初始化为接近于零的值以尽量减少干扰，遵循 [23] 。相比之下，对于 TreeSatAI-TS，我们发现将预训练的 RGB 权重映射到下游任务中的红外颜色（IRC）通道 NIR、RED 和 GREEN 是有益的，同时将 BLUE 通道权重初始化为接近于零的值。我们将这种有益效果归因于两个因素：（i）自然 RGB 图像上学习到的表征可以很好地转移到 IRC 航拍图像上，和（ii）在 TreeSatAI-TS 中 NIR 通道起着决定性作用，从中获益于被映射到转移的权重。

我们尝试了两种融合模式来处理多模态性和多时态性：(i) 共享模式，其中一个单独的编码器（用 DINO-v2 的权重初始化）独立处理所有模态和时间步长，跨它们的权重共享；(ii) 单时态模式，其中特定模态的编码器（同样用 DINO-v2 初始化）独立处理每个模态和时间步长。由于 DINO-v2 缺乏原生的时间处理，Sec. 6.2.2 的时间编码被注入到后编码器的 token 嵌入中。类 token 被丢弃。

DINO-v2 sat. [74] 基于与 DINO-v2 相同的架构，但在 Maxar 的超高分辨率 RGB 卫星影像上进行预训练。我们对该模型应用了与 DINO-v2 相同的适应协议。

6.4.2. DOFA

DOFA [95] 是一个在多源 EO 数据上使用 MAE 框架预训练的 FM，并通过一个超网络进行动态通道处理，该超网络根据输入波长生成补丁嵌入权重（见表 4）。虽然 DOFA 本质上是灵活的，但在我们的背景下需要进行适应以支持多模态和多时态。

我们实现了四个并行的补丁化层（从原始补丁化层初始化），以处理 VHR、Sentinel-2，以及升轨和降轨的 Sentinel-1。与 DINO-v2 和 DINO-v2 sat. 一样，我们探索了共享和单模态融合模式以处理多模态和多时态。由于原始模型缺乏本地的时间处理，Sec. 6.2.2 的时间编码再次被注入到后编码器的 token 嵌入中。

Table 4. DOFA 的每种模态波长。

Modality	Bands	Wavelengths
Aerial	RED, GREEN, BLUE, NIR	0.64, 0.56, 0.48, 0.81
SPOT	RED, GREEN, BLUE	0.66, 0.56, 0.48
$S1_{ASC}$	VV, VH	5.405, 5.405
$S1_{DES}$	VV, VH	5.405, 5.405
S2	B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12	0.490, 0.560, 0.665, 0.705, 0.740, 0.783, 0.842, 0.865, 1.610, 2.190

6.4.3. CROMA

CROMA [24] 是一个专为 Sentinel 图像设计的 FM，具有分别针对 Sentinel-1 和 Sentinel-2 的独立编码器，并通过一个交叉编码器融合它们的中间表示。由于我们只在其原始模态上对 CROMA 进行微调，因此只需进行最小的适应。Sentinel-1 在上升和下降轨道上的时间序列被连接并作为一个单一序列传递给 Sentinel-1 编码器。与其他基线 FM 一样，时间编码在编码器后注入。

我们考虑了两种处理多模态和多时间性的融合模式：(i) late-croma，其中单时间 Sentinel-1 和 Sentinel-2 编码器输出在分类/分割头之前在模态和时间步骤上进行拼接（禁用跨编码器）；(ii) inter-croma（默认），其中单时间 Sentinel-1 和 Sentinel-2 编码器输出成对分组，通过跨编码器融合，最后在分类/分割头之前在各对间进行拼接。

注意，CROMA 本质上仍然是单时间的：late-croma 仅进行晚期多时间融合，而 inter-croma 在各个 Sentinel-1/Sentinel-2 对内进行中期多时间融合，但在对之间进行晚期融合。

6.5. 超参数表

在本小节中，我们提供了一些表格，这些表格报告了我们超参数值的详尽列表。

6.5.1. 数据集特定超参数

Table 5. TreeSatAI-TS 的超参数。

	Spatial extent of original tiles: 60 m Spatial extent of crops: 60 m			
	Aerial	$S1_{ASC}$	$S1_{DES}$	S2
Dataset's original resolution (m)	0.2	10	10	10
Image size (I_m)				
MAE/MAESTRO/ViT	300	6	6	6
DINO-v2	210	✗	✗	42
DINO-v2 sat.	240	✗	✗	48
DOFA	240	48	48	48
CROMA	✗	24	24	24
Patch size (P_m)				
MAE/MAESTRO/ViT	20	2	2	2
DINO-v2	14	✗	✗	14
DINO-v2 sat.	16	✗	✗	16
DOFA	16	16	16	16
CROMA	✗	8	8	8
Number of temporal bins (D_m)	1	4	4	16
Number of channels (C_m)	4	2	2	10
Multiplicative normalization factor	255	5	5	5×10^3
Use cloud/snow mask				✓ (mask proba. > 0)
MAE/MAESTRO's band groups ($\mathcal{G}_m$)	RED, GREEN, BLUE NIR	VV VH	VV VH	B02, B03, B04, B05 B06, B07, B08, B8A B11, B12

Table 6. PASTIS-HD 的超参数。

	Spatial extent of original tiles: 1280 m Spatial extent of crops: 160 m Spatial resolution of token grid of reference: 20 m			
	SPOT	$S1_{ASC}$	$S1_{DES}$	S2
Dataset's orig. resolution (m)	1	10	10	10
Image size (I_m)				
MAE/MAESTRO/ViT	160	16	16	16
DINO-v2	140	✗	✗	112
DINO-v2 sat.	160	✗	✗	128
DOFA	160	128	128	128
CROMA	✗	64	64	64
Patch size (P_m)				
MAE/MAESTRO/ViT	16	2	2	2
DINO-v2	14	✗	✗	14
DINO-v2 sat.	16	✗	✗	16
DOFA	16	16	16	16
CROMA	✗	8	8	8
Number of temporal bins (D_m)	1	4	4	16
Number of channels (C_m)	3	2	2	10
Multiplicative normalization factor	255	20	20	1×10^4
Use cloud/snow mask				✗ (not avail.)
MAE/MAESTRO's band groups ($\mathcal{G}_m$)	RED, GREEN, BLUE	VV VH	VV VH	B02, B03, B04, B05 B06, B07, B08, B8A B11, B12

Table 7. FLAIR # 2 和 FLAIR-HUB 的超参数。

	Spatial extent of original tiles: 102.4 m Spatial extent of crops: 102.4 m Spatial resolution of token grid of reference: 3.2 m				
	Aerial	DEM/DSM	$S1_{ASC}$	$S1_{DES}$	S2
Present in FLAIR # 2	✓	✓	✗	✗	✓
Dataset's original resolution (m)	0.2	0.2	10.24	10.24	10.24
Image size (I_m)					
MAE/MAESTRO/ViT	512	512	10	10	10
DINO-v2	448	✗	✗	✗	70
DINO-v2 sat.	512	✗	✗	✗	80
DOFA	512	✗	80	80	80
CROMA	✗	✗	40	40	40
Patch size (P_m)					
MAE/MAESTRO/ViT	16	32	2	2	2
DINO-v2	14	✗	✗	✗	14
DINO-v2 sat.	16	✗	✗	✗	16
DOFA	16	✗	16	16	16
CROMA	✗	✗	8	8	8
Number of temporal bins (D_m)	1	1	4	4	16
Number of channels (C_m)	4	2	2	2	10
Multiplicative normalization factor	255	1×10^3	5	5	5×10^3
Use cloud/snow mask					✓ (mask proba. > 0)
MAE/MAESTRO's band groups ($\mathcal{G}_m$)	RED, GREEN, BLUE NIR	DEM, DSM	VV VH	VV VH	B02, B03, B04, B05 B06, B07, B08, B8A B11, B12

6.5.2. 数据增强/正则化超参数

Table 8. 数据增强/正则化超参数。

Augmentation/Regularization	Phase used	Hyperparameter
Random spatial cropping	Pre-train/Probe/Fine-tune	
Random time step selection	Pre-train/Probe/Fine-tune	
D4 augmentation	Pre-train/Probe/Fine-tune	
EMA [56]	Fine-tune	$\alpha = 1 - (0.2 \times N_{\text{epochs}})^{-1}$

6.5.3. 优化器超参数

Table 9. 优化器超参数。

Phase	Hyperparameter	TreeSatAI-TS	PASTIS-HD	FLAIR # 2	FLAIR-HUB
	Optimizer	AdamW	AdamW	AdamW	AdamW
	LR schedule	cosine decay	cosine decay	cosine decay	cosine decay
	Warmup fraction	20 %	20 %	20 %	20 %
	Weight decay	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}
	β_1	0.9	0.9	0.9	0.9
	β_2	0.99	0.99	0.99	0.99
Pre-train	Base learning rate	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}
	Final div. factor	1×10^{4}	1×10^{4}	1×10^{4}	1×10^{4}
	Batch size				
	Dataset frac. 100 %	96	72	72	144
	Dataset frac. 20 %	96	72	–	72
	Dataset frac. 5 %	96	72	–	72
	Epochs				
	Dataset frac. 100 %	100	100	100	100
	Dataset frac. 20 %	200	200	–	100
	Dataset frac. 5 %	400	400	–	200
Probe	Base learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
	Final div. factor	1×10^{4}	1×10^{4}	1×10^{4}	1×10^{4}
	Batch size				
	Dataset frac. 100 %	96	48	48	96
	Dataset frac. 20 %	96	48	–	48
	Dataset frac. 5 %	96	48	–	48
	Epochs				
	Dataset frac. 100 %	10	10	15	15
	Dataset frac. 20 %	20	20	–	15
	Dataset frac. 5 %	40	40	–	30
Fine-tune	Base learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
	Final div. factor	2	2	2	2
	Batch size				
	Dataset frac. 100 %	96	48	48	96
	Dataset frac. 20 %	96	48	–	48
	Dataset frac. 5 %	96	48	–	48
	Epochs				
	Dataset frac. 100 %	50	50	100	100
	Dataset frac. 20 %	100	100	–	100
	Dataset frac. 5 %	200	200	–	200

6.5.4. SSL 超参数

Table 10. SSL 超参数。

Hyperparameter	
Reconstruction loss	L_1 w/ patch-group-wise normalization (for some ablations: L_1 w/o normalization or L_1 w/ patch-wise normalization)
Masking ratio	75 %
Probability modality-structured masking	0.25 (if not disabled)
Probability temporally-structured masking	0.25 (if not disabled)
Probability spatially-structured masking	0.25 (if not disabled)

7. 附加结果

在本节中，我们报告了多模态和多时态融合（Sec. 7.1 ）、多光谱融合和目标归一化（Sec. 7.2 ）、以及与预训练和微调数据集大小相关的缩放（Sec. 7.3 ）的详细结果。我们还报告了关于我们的遮罩策略（Sec. 7.4 ）的额外消融实验、多时态组件的影响（Sec. 7.5 ）、以及每个模态在数据集中的重要性（Sec. 7.6 ）。

7.1. 多模态/多时相融合

在 Tab. 11 中可以找到不同多模态和多时态融合模式的详细评估数据，包括 MAEs、基线 FMs 和 ViTs，此外还有 Fig. 3 。

Table 11. 评估不同的数据集上的多模态和多时相融合模式用于 MAEs、基线 FMs 和 ViTs。我们报道了 TreeSatAI-TS 数据集上的加权 F1 得分（%）以及 PASTIS-HD 和 FLAIR-HUB 20 数据集上的 mIoU（%）。

Model	Model size	Fusion mode	Modality groups	TreeSatAI-TS	PASTIS-HD fold I	FLAIR-HUB filt. 20 % / split 1
MAE	Base	shared		76.1	66.1	62.5
MAE	Base	monotemp		77.0	66.0	62.4
MAE	Base	mod		78.4	68.9	63.3
MAE	Base	group	$S1_{ASC}$, $S1_{DES}$	78.5	68.8	**63.6**
MAE	Base	group	$S1_{ASC}$, $S1_{DES}$, S2	78.1	**69.1**	63.5
MAE	Base	group	$S1_{ASC}$, $S1_{DES}$, S2, VHR	78.2	68.1	61.4
MAE	Base	inter-group	$S1_{ASC}$, $S1_{DES}$	**78.8**	68.6	63.5
DINO-v2 [60]	Base	shared		76.7	64.4	64.2
DINO-v2 [60]	Base	monotemp		76.4	63.7	63.9
DINO-v2 sat. [74]	Large	shared		76.3	64.0	64.5
DINO-v2 sat. [74]	Large	monotemp		76.6	63.1	64.0
DOFA [95]	Base	shared		76.1	62.9	62.8
DOFA [95]	Base	monotemp		75.9	63.2	64.1
CROMA [24]	Base	late-croma		69.8	64.9	41.9
CROMA [24]	Base	inter-croma		70.5	65.0	42.5
ViT	Base	shared		72.6	64.1	59.3
ViT	Base	monotemp		73.0	64.4	58.9
ViT	Base	mod		75.4	**64.6**	59.0
ViT	Base	group	$S1_{ASC}$, $S1_{DES}$	75.7	64.6	**59.5**
ViT	Base	group	$S1_{ASC}$, $S1_{DES}$, S2	**76.0**	64.2	59.1
ViT	Base	group	$S1_{ASC}$, $S1_{DES}$, S2, VHR	75.9	64.0	56.4
ViT	Base	inter-group	$S1_{ASC}$, $S1_{DES}$	75.6	64.5	58.8

7.2. 多光谱融合/目标归一化

有关多光谱融合和目标归一化选择的评估详细数值，以及与 MAEs 的关系，可以在 Tab. 12 和 Fig. 4 中找到。我们还在 Tab. 13 中报告了探测评估后的结果。

我们观察到，与使用补丁组归一化的联合-令牌融合相比，使用基于令牌的融合与逐补丁归一化时，性能略有下降。这可能归因于两个因素的结合：(i) 某些 Sentinel-2 波段组可能与下游任务关联性较弱，(ii) 我们的分割头通过单个注意力池化层聚合多模态和多时态特征，层次较浅，可能难以过滤掉不相关的令牌，从而在最终预测中引入噪声。

Table 12. 在两个数据集上评估多光谱融合和目标归一化的不同选择对于 MAE-B 模型的效果。我们报告了在 TreeSatAI-TS 上的加权 F1 分数（%）以及在 PASTIS-HD 上的 mIoU（%），其中预训练数据集的比例是可变的。

Multispectral fusion	Target normalization	TreeSatAI-TS			PASTIS-HD fold I		
		Pre-training dataset frac.			Pre-training dataset frac.		
		5 %	20 %	100 %	5 %	20 %	100 %
Joint-token	No normalization	72.7	72.8	74.6	64.9	65.4	65.1
Joint-token	Patch-wise	74.4	75.5	77.4	66.7	67.1	67.5
Joint-token	Patch-group-wise	**76.3**	**77.4**	**78.5**	**67.3**	**68.1**	**68.8**
Token-based	Patch-wise	**76.6**	**77.0**	**78.9**	**66.7**	**67.2**	**68.6**

7.3. 根据数据集规模进行扩展

除了 Fig. 5 之外，还在 Tab. 14 中提供了不同预训练/微调数据集比例的性能详细数据。

Table 13. 探究对不同选择的多光谱融合和目标归一化的 MAE-B 模型在两个数据集上的评估。我们报告了 TreeSatAI-TS 上的加权 F1 分数（%）和 PASTIS-HD 上的 mIoU（%），并改变了预训练数据集的比例。

Multispectral fusion	Target normalization	TreeSatAI-TS			PASTIS-HD fold I		
		Pre-training dataset frac.			Pre-training dataset frac.		
		5 %	20 %	100 %	5 %	20 %	100 %
Joint-token	No normalization	57.2	61.1	63.2	52.5	56.9	57.9
Joint-token	Patch-wise	62.9	65.1	67.7	58.2	59.7	61.2
Joint-token	Patch-group-wise	**64.8**	**68.6**	**69.3**	**59.0**	**60.2**	**61.2**
Token-based	Patch-wise	**61.9**	**66.2**	**69.3**	**55.9**	**57.5**	**61.0**

在所有设置中，使用 MAESTRO 进行 SSL 预训练的表现优于监督 ViTs，尤其是在微调数据稀缺的情况下，效果提升尤为显著。表中的 “Pre-training fraction = 100 % ” 和 “Pre-training fraction = Fine-tuning fraction” 对比显示，额外的预训练数据提高了下游性能，而不论可用于微调的标注数据量如何。这表明，MAESTRO 可能会从更大规模的无标签预训练中进一步受益。

Table 14. 在三个数据集上，不同比例的预训练/微调数据集情况下，MAESTRO-B 和 ViT-B 模型的扩展效果。我们报告了在 TreeSatAI-TS 上的加权 F1 分数（%），以及在 PASTIS-HD 和 FLAIR-HUB 上的 mIoU（%）对于三种微调数据集比例：5 %，20 % 和 100 %。对于每种微调比例，我们比较了三种预训练设置：没有预训练，在与微调相同比例上的预训练，以及在 100 % 数据上预训练。请注意，当在 100 % 的数据上微调时，与微调相同比例的预训练和在 100 % 数据上的预训练效果相同，因此表现一致。

	TreeSatAI-TS			PASTIS-HD fold I			FLAIR-HUB split 1		
	Fine-tuning dataset frac.			Fine-tuning dataset frac.			Fine-tuning dataset frac.		
	5 %	20 %	100 %	5 %	20 %	100 %	5 %	20 %	100 %
No pre-training	61.8	69.0	75.7	38.8	52.2	64.6	55.3	59.5	61.6
Pre-training fraction = Fine-tuning fraction	65.5	71.8	78.5	46.6	57.1	68.8	60.1	63.6	64.9
Pre-training fraction = 100 %	67.8	73.8	78.5	52.5	59.2	68.8	61.5	64.6	64.9

7.4. 遮罩策略

我们研究了各种屏蔽策略组成部分对下游任务性能的影响。特别是，我们评估在自监督学习（SSL）伪任务中包含不同结构化屏蔽步骤的效果——即 Sec. 6.3.1 中详细的步骤 (i-a)、(i-b) 和 (i-c)。实验结果在 Tab. 16 和 Tab. 15 中报告，分别对应于微调后的性能和探测后的性能。在这些实验中，我们使用组融合模式，将 Sentinel-1 升轨和降轨模式组合在一起。对于多光谱数据，我们结合了修补组别目标归一化在重建过程中使用联合令牌融合。

我们发现，结构化掩码在微调性能上仅有轻微的提升。然而，时间结构化掩码在 PASTIS-HD 的探测性能上带来了显著的提升。这表明，当直接评估时（即不进行微调），结构化掩码可能会产生更有用的表示，尽管这些优势在模型微调后未必能够完全转移。

Table 15. 在三个数据集上评估 MAE-B 模型不同遮盖选择的影响。我们报告了 TreeSatAI-TS 上的加权 F1 分数（%）以及 PASTIS-HD 和 FLAIR-HUB 20 上的 mIoU（%）。

Masking structure			TreeSatAI-TS	PASTIS-HD fold I	FLAIR-HUB filt. 20 % / split 1
Modality	Spatial	Temporal			
✓	✓	✓	78.5	68.8	63.6
✗	✓	✓	78.5	68.4	63.3
✓	✗	✓	78.5	68.6	63.0
✓	✓	✗	78.3	68.5	63.4
✗	✗	✗	78.2	68.4	63.6

Table 16. 在三个数据集上探讨评估 MAE-B 模型不同掩码选择的方法。我们报告了 TreeSatAI-TS 上的加权 F1 得分（%）以及在 PASTIS-HD 和 FLAIR-HUB 20 上的 mIoU（%）。

Masking structure			TreeSatAI-TS	PASTIS-HD fold I	FLAIR-HUB filt. 20 % / split 1
Modality	Spatial	Temporal			
✓	✓	✓	69.3	61.2	56.2
✗	✓	✓	69.8	61.4	56.3
✓	✗	✓	69.6	61.5	56.3
✓	✓	✗	69.2	57.7	56.2
✗	✗	✗	69.9	57.6	56.3

7.5. 多时态成分的重要性

我们研究了不同多时态组件对 MAE-B 和 ViT-B 模型的下游任务性能的影响。在这些实验中，我们使用了组融合模式，将 Sentinel-1 的升轨和降轨模态组合在一起。对于多光谱数据，我们使用了联合代币融合，并在重建过程中结合了补丁组的目标归一化。

我们在 Tab. 17 中报告了对于 Sentinel-2 和 Sentinel-1 的上升和下降模式，改变时间箱数量 D_m 的效果。此外，我们在 Tab. 18 中展示了去除日期编码和禁用与时间箱内随机时间步选择相关的数据增强的影响。

在 Tab. 17 中，我们观察到当 Sentinel-2 作为驱动性能的模式时，其 D_m 的选择至关重要，例如在 TreeSatAI-TS 和 PASTIS-HD 中（见 Sec. 7.6 ）。值得注意的是，Sentinel-2 的 $D_m = 16$ 和 $D_m = 1$ 之间的性能差距远大于 $D_m = 1$ 和 $D_m = 0$ 之间的差距。这表明 Sentinel-2 的多时态动态性而不是其单时态组件是关键。相比之下，改变 Sentinel-1 的 D_m 仅有微小的影响，反映了其较小的作用（见 Sec. 7.6 ）。

在 Tab. 18 中，我们发现，删除日期编码会导致性能明显下降，特别是在多时态动态性重要的情况下。虽然其益处的大小在不同数据集之间有所变化，但训练期间随机时间步长选择通常能提高性能。

Table 17. MAE-B 和 ViT-B 模型在三个数据集上的时间分段数量的重要性。我们报告了 TreeSatAI-TS 上的加权 F1 得分 (%) 和 PASTIS-HD 及 FLAIR-HUB 20 上的 mIoU (%)。箭头（ / ）表示与使用哨兵-2 模式的 16 个时间分段以及哨兵-1 上行和下行模式的 4 个时间分段相比的差异。

Model	Number of temporal bins			TreeSatAI-TS	PASTIS-HD fold I	FLAIR-HUB filt. 20 % / split 1
	$S1_{ASC}$	$S1_{DES}$	S2			
MAE	4	4	16	78.5	68.8	63.6
MAE	4	4	4	75.3 ↓ 3.2	64.3 ↓ 4.5	62.4 ↓ 1.2
MAE	4	4	1	73.1 ↓ 5.4	49.3 ↓ 19.5	61.8 ↓ 1.8
MAE	4	4	0	72.6 ↓ **5.9**	44.9 ↓ **23.9**	61.6 ↓ **2.0**
MAE	1	1	16	78.3 ↓ 0.2	68.8 ↑ 0.0	63.5 ↓ 0.1
MAE	0	0	16	78.2 ↓ 0.3	69.2 ↑ 0.4	63.1 ↓ 0.5
ViT	4	4	16	75.7	64.6	59.5
ViT	4	4	4	70.6 ↓ 5.1	61.2 ↓ 3.4	58.4 ↓ 1.1
ViT	4	4	1	66.1 ↓ 9.6	46.5 ↓ 18.1	57.6 ↓ 1.9
ViT	4	4	0	65.8 ↓ **9.9**	41.5 ↓ **23.1**	57.4 ↓ **2.1**
ViT	1	1	16	75.8 ↑ 0.1	64.6 ↑ 0.0	59.8 ↑ 0.3
ViT	0	0	16	75.7 ↑ 0.0	64.4 ↓ 0.2	59.2 ↓ 0.3

7.6. 每个数据集模态的重要性

我们对 MAE-B 和 ViT-B 模型进行模态移除的消融研究。结果在 Tab. 19 中报告。在这些实验中，我们使用组融合模式，将 Sentinel-1 的上升和下降模态组合在一起。对于多光谱数据，我们在重建过程中使用联合标记融合以及补丁组内的目标归一化。

我们观察到，无论是 MAEs 还是 ViTs，通常在包含所有模态时表现最佳。Sentinel-2 模态对与多时态动态紧密相关的任务具有强烈影响，例如树种识别和农作物分割。此外，源自真值的模态显著影响性能——例如，来自 FLAIR-HUB 的航拍图像和来自 PASTIS-HD 的 Sentinel-2。

在 PASTIS-HD 上省略 Sentinel-1 时观察到的轻微性能下降可能源于两个因素的结合：(i) 与 Sentinel-2 相比，Sentinel-1 在任务相关性方面较低，以及 (ii) 相对浅的分割头，它通过单一的注意力池化聚合多模态和多时序特征，可能难以抑制无关特征，从而在预测中引入噪声。

Table 18. 移除日期编码和禁用随机时间步选择对 MAE-B 和 ViT-B 模型在三个数据集上的影响。我们报告在 TreeSatAI-TS 上的加权 F1 得分（%），以及在 PASTIS-HD 和 FLAIR-HUB 20 % 上的 mIoU。箭头（ / ）表示与包括日期编码和启用随机时间步选择相比的差异。

Model	Number of temporal bins			Date encodings	Random time step selection	TreeSatAI-TS	PASTIS-HD fold I	FLAIR-HUB filt. 20 % / split 1
	$S1_{ASC}$	$S1_{DES}$	S2					
MAE	4	4	16	✓	✓	78.5	68.8	63.6
MAE	4	4	16	✗	✓	77.5 ↓ 1.0	67.8 ↓ 1.0	63.6 ↑ 0.0
MAE	4	4	16	✓	✗	78.8 ↑ 0.3	68.2 ↓ 0.6	63.7 ↑ 0.1
ViT	4	4	16	✓	✓	75.7	64.6	59.5
ViT	4	4	16	✗	✓	74.8 ↓ 0.9	63.8 ↓ 0.8	59.4 ↓ 0.1
ViT	4	4	16	✓	✗	75.4 ↓ 0.3	63.9 ↓ 0.7	58.7 ↓ 0.8

Table 19. 去除模态对三种数据集上的 MAE-B 和 ViT-B 模型的影响。我们报告 TreeSatAI-TS 上的加权 F1 分数（%）以及 PASTIS-HD 和 FLAIR-HUB 20 上的 mIoU（%）。箭头（ / ）表示与包括所有模态相比的差异。

Model	TreeSatAI-TS				PASTIS-HD fold I				FLAIR-HUB filt. 20 % / split 1				
	Aerial	S1	S2	wF1 (%)	Spot	S1	S2	mIoU (%)	Aerial	S1	S2	DEM/DSM	mIoU (%)
MAE	✓	✓	✓	78.5	✓	✓	✓	68.8	✓	✓	✓	✓	63.6
MAE	✗	✓	✓	76.8 ↓ 1.7	✗	✓	✓	68.6 ↓ 0.2	✗	✓	✓	✓	52.2 ↓ 11.4
MAE	✓	✗	✓	78.2 ↓ 0.3	✓	✗	✓	69.2 ↑ 0.4	✓	✗	✓	✓	63.1 ↓ 0.5
MAE	✓	✓	✗	72.6 ↓ 5.9	✓	✓	✗	44.9 ↓ 23.9	✓	✓	✗	✓	61.6 ↓ 2.0
MAE									✓	✓	✓	✗	62.7 ↓ 0.9
ViT	✓	✓	✓	75.7	✓	✓	✓	64.6	✓	✓	✓	✓	59.5
ViT	✗	✓	✓	75.3 ↓ 0.4	✗	✓	✓	64.5 ↓ 0.1	✗	✓	✓	✓	47.9 ↓ 11.6
ViT	✓	✗	✓	75.7 ↑ 0.0	✓	✗	✓	64.4 ↓ 0.2	✓	✗	✓	✓	59.2 ↓ 0.3
ViT	✓	✓	✗	65.8 ↓ 9.9	✓	✓	✗	41.5 ↓ 23.1	✓	✓	✗	✓	57.4 ↓ 2.1
ViT									✓	✓	✓	✗	58.3 ↓ 1.2

8. 推理结果

在本节中，我们展示了来自 MAESTRO 和监督 ViTs 在 PASTIS-HD 和 FLAIR-HUB 分割任务上的推理结果示例。我们再次考虑具有组融合模式的 MAESTRO 和 ViTs，将 Sentinel-1 上升和下降模态组合在一起。对于多光谱数据，我们在重建过程中使用联合标记融合以及逐块组目标归一化。

结果分别在 Fig. 7 和 Fig. 8 中报告用于 PASTIS-HD 和 FLAIR-HUB。对于每个图块，我们展示高分辨率影像、哨兵-2 影像、MAESTRO 预测、ViT 预测以及相应的真实值。示例是从测试集随机抽取的。

在 Fig. 7 中，MAESTRO 在 PASTIS-HD 上生成精确的分割遮罩，与地块和树木覆盖区域的边界紧密匹配。它的预测在空间上比 ViTs 更具连贯性，并能够更好地区分作物类型。

在 Fig. 8 中，尽管 FLAIR-HUB 中的复杂场景引入了预测模糊性，MAESTRO 仍然提供了比 ViTs 更清晰和更准确的物体边界。它在灌木丛和水等类别上表现更好，并且更有效地区分了不透水和透水表面。

9. 计算成本

在本节中，我们量化浮点运算（FLOPs）（Sec. 9.1 ），并报告本文中所示实验的总体计算成本（Sec. 9.2 ）。

9.1. MAC/FLOPS

我们分别报告了用于 SSL 预训练（Sec. 9.1.1 ）和探测/微调（Sec. 9.1.2 ）的每次前向传递的 FLOPs。当只考虑前向 FLOPs 时，探测和微调是相同的，但在考虑反向 FLOPs 时，探测则显著更高效。

对于 MAESTRO-B、MAE-B 和 ViT-B 模型，前向 FLOPs（针对单个批次元素）被计算。我们首先计算乘加操作（MACs）的数量，然后使用关系 FLOPs $= 2 \times$ MACs 转换为 FLOPs。按照标准做法，我们只包括模型组件中具有显著 MAC 贡献的操作，排除如归一化、非线性、偏置和组件汇总等逐元素操作。

我们再次考虑群组融合模式，将 Sentinel-1 的上升和下降模式结合在一起。对于多光谱数据，我们评估联合令牌和基于令牌的多光谱融合。

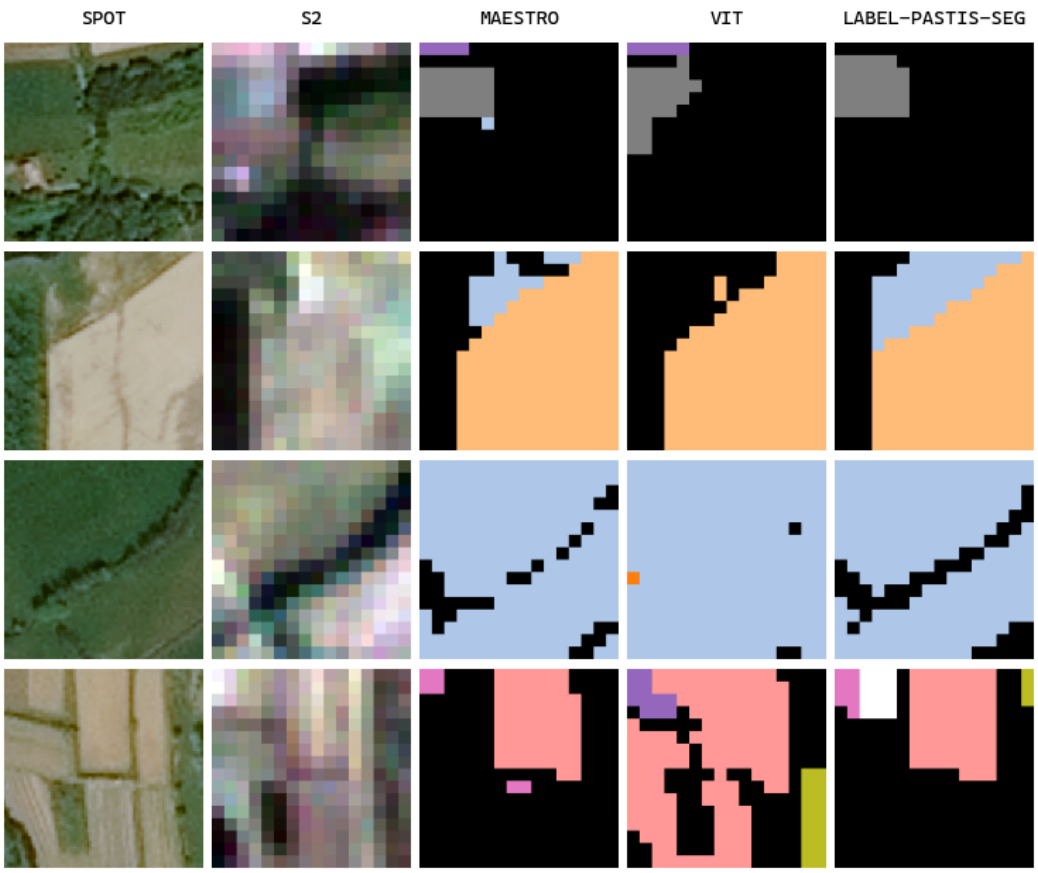


Figure 7. MAESTRO-B 和 ViT-B 模型在 PASTIS-HD 上的推理结果。对于 Sentinel-2 图像，我们报告在时间分箱中的像素中位数。白色地块对应于缺失标注的区域（空标签）。

9.1.1. 预训练 MACs/FLOPs

对于每种模态 m ，令 L_m 表示它的序列长度。通过联合-令牌多光谱融合，序列长度为 $L_m = (I_m/P_m)^2 D_m$ ，其中 I_m 是图像尺寸，P_m 是补丁大小，D_m 是时间段的数量。通过基于令牌的多光谱融合，序列长度变为 $L_m = (I_m/P_m)^2 D_m|\mathcal{G}_m|$ ，其中 $|\mathcal{G}_m|$ 是映射到不同令牌的波段组集合的基数。

对于一种没有通过早期多模态融合与其他模态组合在一起的模态 m ，编码器和解码器中的 MACs 为：

$$\begin{aligned}\text{MACs}^{\text{enc}} &= \left(12\lfloor(1-M)L_m\rceil C_e^2 + 2\lfloor(1-M)L_m\rceil^2 C_e\right) N_e, \\ \text{MACs}^{\text{dec}} &= \left(12L_m C_d^2 + 2L_m^2 C_d\right) N_d,\end{aligned}$$

，其中 M 是被屏蔽标记的分数，N_e 和 N_d 是编码器和解码器中的层数，而 C_e 和 C_d 是编码器和解码器中的潜在维度。

对于分组的 Sentinel-1 升轨和降轨模式，通过用这两种模式的序列长度之和替换 L_m 来计算 MACs^{enc} 和 MACs^{dec} 。

从编码器空间投影到解码器空间的密集层贡献：

$$\text{MACs}^{\text{enc-to-dec}} = \sum_m \lfloor(1-M)L_m\rceil C_e C_d.$$

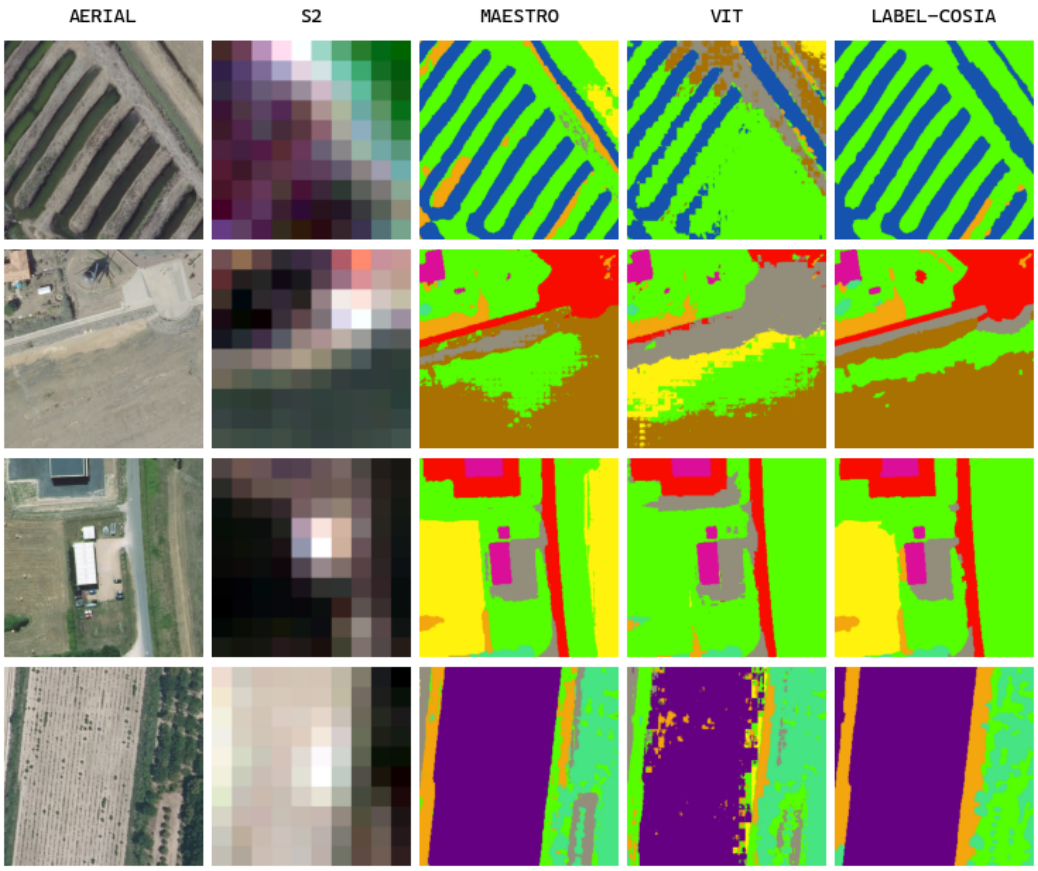


Figure 8. MAESTRO-B 和 ViT-B 模型在 FLAIR-HUB 上的推理结果。对于 Sentinel-2 影像，我们报告了时间段中的逐像素中位数。

拼补和去拼补操作贡献：

$$\text{MACs}^{\text{patchify}} = \sum_m I_m^2 D_m C_m C_e,$$
$$\text{MACs}^{\text{unpatchify}} = \sum_m I_m^2 D_m C_m C_d,$$

，其中 C_m 是每种模式 m 的通道数量。

把所有内容放在一起，使用 $M = 0.75$ 代表被掩盖标记的分数，$N_e = 12$ 、$C_e = 768$ 用于编码器，$N_d = 3$ 、$C_d = 512$ 用于解码器，我们根据 Tab. 5 、Tab. 6 和 Tab. 7 中的默认配置，计算出 MAESTRO-B 和 MAE-B 模型在预训练期间的前向 MACs/FLOPs。该计算针对四个评估数据集的联合标记和基于标记的多光谱融合进行。结果在 Tab. 20 中报告。

Table 20. 在预训练期间，MAESTRO-B 和 MAE-B 模型的 MACs/FLOPs。我们报告在经过评估的四个数据集上，单批元素的前向 MACs/FLOPs（用于单个批元素）的联合令牌和基于令牌的多光谱融合。乘数（x）表示与使用联合令牌多光谱融合相比，FLOPs 的增加。

Multispectral fusion	TreeSatAI-TS		PASTIS-HD		FLAIR # 2		FLAIR-HUB	
	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)
Joint-token	14.3	28.7	56.1	112.1	59.1	118.2	65.4	130.9
Token-based	33.7 × 2.4	67.4 × 2.4	173.6 × 3.1	347.2 × 3.1	133.9 × 2.3	267.8 × 2.3	146.9 × 2.2	293.8 × 2.2

9.1.2. 探测/微调 MACs/FLOPs

我们重用与 Sec. 9.1.1 中相同的符号。如前所述，序列长度在联合标记的多光谱融合中由 $L_m = (I_m/P_m)^2 D_m$ 给出，在基于标记的多光谱融合中由 $L_m = (I_m/P_m)^2 D_m|\mathcal{G}_m|$ 给出。

对于一个未通过早期多模态融合与其他模态组合的 m，其中编码器中的 MACs 为：

$$\text{MACs}^{\text{enc}} = \left(12L_mC_e^2 + 2L_m^2C_e\right) N_e.$$

对于分组的 Sentinel-1 升降模式，通过将 L_m 替换为两个模式的序列长度之和来计算 MACs^{enc}。

分类和分割头中的注意力池化操作贡献：

$$\text{MACs}^{\text{attn-pool}} = 2\sum_m L_mC_e.$$

在分类和分割头中的最终密集投射贡献：

$$\text{MACs}^{\text{proj-cls}} = C_eC_{\text{cls}},$$
$$\text{MACs}^{\text{proj-seg}} = L_{\text{ref}}C_eC_{\text{seg}},$$

，其中 L_{ref} 是参考空间标记网格的序列长度（见 Sec. 2.2），而 C_{cls} 和 C_{seg} 分别是分类和分割任务的语义类别数量，包括在损失和度量计算中被忽略的类别（见 Sec. 6.1）。

结合所有内容，并使用 $N_e = 12$ 和 $C_e = 768$ 作为编码器，我们在探测/微调期间基于 Tab. 5、Tab. 6 和 Tab. 7 的默认配置，计算了 MAESTRO-B、MAE-B 和 ViT-B 模型的正向 MACs/FLOPs。这一计算是在四个评估的数据集上针对联合 token 和基于 token 的多光谱融合进行的。结果展示在 Tab. 21 中。

Table 21. 在探测/微调期间，MAESTRO-B、MAE-B 和 ViT-B 模型的 MACs/FLOPs。我们报告了四个评估数据集中每个批量元素的前向 MACs/FLOPs，涉及联合代币和基于代币的多光谱融合。倍数 (x) 表示相对于使用联合代币多光谱融合的 FLOPs 增加。

Multispectral fusion	TreeSatAI-TS		PASTIS-HD		FLAIR # 2		FLAIR-HUB	
	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)	MACs ($\times 10^9$)	FLOPs ($\times 10^9$)
Joint-token	39.1	78.3	163.4	326.8	167.4	334.8	185.1	370.3
Token-based	95.0 × 2.4	190.0 × 2.4	549.9 × 3.4	1099.9 × 3.4	403.9 × 2.4	807.8 × 2.4	440.8 × 2.4	881.7 × 2.4

9.2. 总体计算成本

我们所有的实验都在配备 V100、A40、A100 或 H100 GPU 节点的 SLURM 集群上运行。训练是在混合精度下进行的。

总的来说，论文中展示的所有实验运行共需要 17,661 小时的 V100，7,736 小时的 A40，268 小时的 A100 以及 268 小时的 H100。