多模态卫星影像语义分割的位置信息预测自监督学习

JOHN WAITHAKA and MOISE BUSOGI, Carnegie Mellon University Africa, Rwanda

卫星图像的语义分割对于地球观测应用至关重要，但由于标记训练数据有限而受到限制。尽管像掩码自动编码器（MAE）这样的自监督预训练方法显示出希望，但它们侧重于重建而不是定位—定位是分割任务的基本方面。我们提出了将位置预测自监督学习方法 LOCA（位置感知）改编用于多模态卫星图像语义分割。我们的方法通过将 SatMAE 的通道分组从多光谱数据扩展到多模态数据来解决卫星数据的独特挑战，从而有效处理多种模态，并引入同组注意力掩码以在预训练期间鼓励跨模态交互。该方法使用相对补丁位置预测，鼓励空间推理用于定位而不是重建。我们在 Sen1Floods11 洪水映射数据集上评估了我们的方法，在该数据集上显著优于现有的基于重建的卫星图像自监督学习方法。我们的结果表明，相对于重建为基础的方法，当位置预测任务被正确适应于多模态卫星图像时，所学习的表示对于卫星图像语义分割更为有效。源代码可在 https://github.com/johnGachihi/scenic 上获取。

CCS Concepts: • **Computing methodologies** → **Image segmentation**.

Additional Key Words and Phrases: Earth Observation, Remote Sensing, Satellite Imagery, Multi-Modal, Self-Supervised Learning, Position Prediction, Semantic Segmentation

ACM Reference Format:
John Waithaka and Moise Busogi. 2018. 多模态卫星影像语义分割的位置信息预测自监督学习. *J. ACM* 37, 4, Article 111 (August 2018), 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 引言

卫星影像是地球观测研究的基本数据来源，其中语义分割对于分析这些影像尤为重要。语义分割可以实现，例如，提取洪水范围图、作物覆盖图和森林覆盖图，以用于灾害管理、粮食安全分析和气候研究。

虽然深度学习模型已被证明对于卫星图像的语义分割是有效的（例如在 [14, 24] 中），但语义分割仍然受到有限标注训练数据的限制。语义分割的像素级标注非常昂贵且耗时 [5, 25]，而且由于空间分辨率较低、生疏的语义类别以及需要领域专业知识，卫星图像增加了这一挑战。

预训练通常用于在标记的训练数据有限的情况下提高模型性能。自监督预训练不需要标记数据，这特别适合于卫星图像领域，虽然标记数据集稀缺，但未标记的卫星图像数据集非常庞大。

对比学习是一种突出的自监督预训练方法。它涉及对相同事物的两个不同视图进行匹配，这些视图通过独立的数据增强抽样或时间位移 [2] 生成。然而，Caron 等人 [5] 发现，用对比学习预训练的模型在语义分割任务中不能很好地迁移。他们假设这是因为对比学习鼓励全局图像级别的表示，而不需要空间推理，而语义分割是一项直观上需要空间推理的像素级任务。

掩码图像建模，特别是掩码自编码器（MAE）预训练方案 [13]，在卫星图像领域得到了广泛的探索 [1, 6, 15, 17, 18]。MAE 为自监督预训练定义了一个掩码补丁重建任务。这个任

Authors' Contact Information: John Waithaka, jwaithak@andrew.cmu.edu; Moise Busogi, mbusogi@andrew.cmu.edu, Carnegie Mellon University Africa, Kigali, Rwanda.

ACM 1557-735X/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

务通过让不同空间位置中的可见补丁预测其他位置的掩码补丁，来促进空间推理。在一些先前的工作中，基于 MAE 的方法在卫星图像语义分割中表现优于对比方法 [1, 17] 。

位置预测是一种不太显著的自监督预训练方法。特别是，LOCA（位置感知）[5] 定义了一种用于自监督学习的相对位置预测任务。更准确地说，从输入图像中采样出查询视图和参考视图，然后让查询视图中的每个块预测其在参考视图中的位置。与鼓励重建的空间推理的 MAE 不同，这个任务鼓励用于定位的空间推理。由于分割在某种程度上本质上是一个定位任务，我们假设相对位置预测可以学习到更有效用于语义分割的块表示。此外，Caron 等人 [5] 显示，在自然图像领域的各种语义分割数据集上，LOCA 优于其他自监督方法。然而，相对位置预测在卫星图像领域仍未被探索。

卫星影像与自然影像有显著的不同。自然图像通常仅由 RGB 波段组成，而卫星图像则可以由更广范围的电磁谱中的多个波段组成。此外，由于卫星图像是由不同类型的地球观测传感器捕获的，存在“多模态”图像，可以提供同一地理位置的互补视角。我们采用 LOCA 方法来有效处理卫星影像的多光谱特性，并利用其多模态性以提高卫星影像语义分割的迁移性能。

在这项工作中，我们通过扩展通道分组以处理多种模式（多光谱影像、SAR 和 DEM）并引入同组注意力掩码来鼓励预训练期间的跨模态交互，从而调整 LOCA 用于多模态卫星图像语义分割。在 Sen1Floods11 [4] 洪水制图数据集上的评估表明，我们的位置预测方法在卫星图像中优于现有的基于重建的自监督学习方法。

2 相关工作

2.1 SSL 和 LOCA 的位置预测

一种相对较不受欢迎的自监督学习（SSL）分支是补丁位置预测。补丁位置预测方法利用图像中的空间上下文来定义一个预文本任务。这些任务涉及预测图像中补丁的空间位置。Doersch 等人。[8] 从同一图像中取样两个补丁，并预测一个补丁相对于另一个的位置信息。Noroozi 和 Favaro [16] 将图像分割成不重叠的补丁，并在打乱后预测它们的真实位置。Zhai 等人。[23] 使用视觉变换器，在没有位置信息（位置编码 [21] ）的情况下预测补丁的位置。

我们的工作基于 LOCA [5] 。LOCA 定义了一个相对补丁位置预测任务。更具体地说，查询视图和参考视图是从图像中采样的，查询视图中的每个补丁预测其相对于参考视图的位置。为此，查询视图的补丁通过单个跨注意力块关注参考视图。为了控制任务的难度，参考视图的一部分对查询视图是可见的。Caron 等人 [5] 表明，LOCA 在多个自然图像语义分割数据集上优于其他 SSL 预训练方法，然而，位置预测方法在卫星图像领域仍未被充分探索。

2.2 用于密集自监督学习的补丁聚类

Ziegler 和 Asano [26] 使用聚类来生成伪标签，以监督块级分类任务。聚类分配是在在线环境中使用教师网络完成的（学生网络则用于预测聚类）。LOCA [5] 除了相对位置预测外，也使用相同的技术。

2.3 卫星图像的多模态预训练

多模态学习尝试构建能够从多种模态中提取和关联信息的 AI 模型 [3, 22] 。这受到人类感知的启发，人类收集不同模态（例如，视觉、听觉）的数据，并利用它们进行互补以获得对环境更完整的理解。模态与某种传感器相关联，该传感器捕获一种特定类型的数据 [22] 。

在地球观测领域，多种传感器捕捉地球的不同视角，每个视角都包含独特且有用的信息。这些视角之间的差异足以使得先前的研究将其视为不同的模态 [1, 11] 。在如何使用多模态卫星图像来创建更有效的地球观测解决方案方面，有着显著的研究兴趣 [10] 。

在这项工作中，我们考虑了三种模态：多光谱卫星图像（MSI）、合成孔径雷达（SAR）和数字高程模型（DEM）。MSI 捕捉来自电磁波谱不同波长的反射或发射的辐射能量，从可见光到热红外辐射 [9] 。SAR 图像由主动传感器捕获，该传感器向地球发射微波能量，并测量

其反弹回传感器的散射程度。SAR 图像的优点是不受天气或云层覆盖影响。DEM 包含像素级表面高程数据。

在卫星影像领域，关于多模态自监督预训练的研究此前已经进行过研究 [1, 11, 12, 20] 。Nedungadi 等人 [1] 在掩码自动编码器 (MAE) [13] 的基础上，提出了在单模态输入下进行多模态图像重建的方案。他们通过多种特定模态的 MAE 重建解码器实现了这一目标。Han 等人 [12] 和 Astruc 等人 [11] 也在 MAE 的基础上进行研究，但使用多模态输入进行多模态重建。他们通过多种特定模态的嵌入器、跨模态编码器和多种特定模态的重建解码器实现了这一目标。最近，Tseng 等人 [20] 提出了一种创新的 “全局和局部” 跨模态潜在表示重建任务用于自监督学习。先前所有关于卫星影像多模态自监督预训练的工作都使用了一种形式的掩码图像重建。使用位置预测任务进行卫星影像上的多模态自监督预训练仍未被探索。

2.4 用于卫星影像的掩码自编码器

掩码自编码器 [13] 是基于 ViT 的自监督学习者。类似于自然语言处理中的掩码语言模型（例如 BERT [7] ），MAE 通过根据可见的图像块重建被掩码的图像块来学习图像表示。MAE 在卫星影像领域已经被广泛研究 [1, 6, 11, 12, 15, 17, 19] 。

与位置预测一样，MAE 鼓励空间推理，因此学习适合语义分割迁移的图像表示。然而，MAE 使用空间推理进行重建，而位置预测任务则将其用于定位。由于语义分割在某种程度上本质上是一个定位任务，我们认为位置预测任务将学习出更适合语义分割迁移的表示。我们将 MAE 与我们的工作进行迁移性能比较。

3 方法论

我们的工作基于 LOCA [5] ，将其用于多模态卫星图像。我们详细说明了我们的改编部分以及从 LOCA 借鉴的内容。

采样查询和参考视图. 多模态图像对在通道维度上连接以形成单个输入图像 x 。遵循 LOCA 方法，我们从 x 中采样一个查询视图 x_q 和一个参考视图 x_{ref} ，然后对每个视图应用独立的随机增强（即，翻转、裁剪、重新缩放）。为了最大化对应的查询视图和参考视图之间的重叠，同时确保查询能够表示局部图像区域，参考视图被采样为覆盖原始图像的较大区域，而查询视图则覆盖原始图像的小部分。遵循 LOCA 方法，我们每个参考视图采样 10 个查询视图。

查询和参考补丁位置对应关系. 查询视图和参考视图被划分为不重叠的 $P \times P$ 块。每个查询视图因此产生块 x_q^i 对于 $i \in \{1, ..., N_q\}$ ，其中 $N_q = \lfloor H_q/P \rfloor \times \lfloor W_q/P \rfloor$ 和 $H_q \times W_q$ 是查询分辨率。我们使用 $H_q = W_q = 96$ 和 $P = 16$ 产生每个查询的 $N_q = 36$ 块。类似地，参考视图产生块 x_{ref}^j 对于 $j \in \{1, ..., N_{ref}\}$ 。我们使用 $H_{ref} = W_{ref} = 224$ 和 $N_{ref} = 196$ 。为了在增强过程中保持空间位置的一致性，我们追踪每个块的原始位置。这使我们能够定义一个映射函数 $h(i) = j$ ，识别与查询块 x_q^i 重叠最大的参考块 x_{ref}^j 。

查询和参考补丁都有 C 个通道：$x_q^i, x_{ref}^j \in \mathbb{R}^{P \times P \times C}$ 。根据 SatMAE [6] ，我们将这些通道分为 G 个通道组，每组有 g 个通道。每个组经过一个独立的补丁嵌入处理以生成 $S_q^g \in \mathbb{R}^{N_q \times d}$ 和 $S_{ref}^g \in \mathbb{R}^{N_{ref} \times d}$ 的令牌序列。这些序列沿序列维度连接，产生 $S_q \in \mathbb{R}^{GN_q \times d}$ 和 $S_{ref} \in \mathbb{R}^{GN_{ref} \times d}$ 。通道分组使我们能够灵活地形成令牌序列，例如，可以从多种模态的混合中形成，或为每种模态单独形成。我们对不同的通道组设置进行了消融实验。

根据 SatMAE [6] 的方法，我们应用组和位置编码以保留空间和通道组信息。每个标记接收一个组编码 $GE_g \in \mathbb{R}_{GE}^d$ 和位置编码 $PE_i \in \mathbb{R}^{d_{PE}}$ ，其中 $d_{GE} + d_{PE} = d$ 。这些编码被连接后添加到上述 S_q 和 S_{ref} 中的相应标记上。

组采样. 通道分组将查询的序列长度从 N_q 增加到 GN_q 个标记（以及参考的从 N_{ref} 增加到 GN_{ref} ）。为了维持计算效率，我们从每个空间位置采样一个标记（每个位置的每组有 G

个标记），保持原始序列长度 N_q 和 N_{ref} 。我们在通道组中均匀采样以确保平衡的表示，生成 $S'_q \in \mathbb{R}^{N_q \times d}$ 和 $S'_{ref} \in \mathbb{R}^{N_{ref} \times d}$ 。

采样序列 S'_q 和 S'_{ref} 通过 transformer 编码器块独立处理，生成查询和参考表示 $Z_q \in \mathbb{R}^{N_q \times d}$ 和 $Z_{ref} \in \mathbb{R}^{N_{ref} \times d}$ 。

查询-参考交互. Caron 等人 [5] 声称，为了解决相对补丁位置预测任务，查询补丁表示必须关注到相应的参考补丁表示。根据 LOCA [5] 的方法，我们使用一个单交叉注意块来实现这点，其查询是由 Z_q 计算得到，键/值则来自 Z_{ref} ，产生输出 $U \in \mathbb{R}^{N_q \times d}$ 。

为了在没有注释的情况下学习空间关系，我们遵循 LOCA [5] ，并解决一个相对补丁位置预测任务。这被表述为一个 N_{ref} 类分类任务，其中每个查询补丁从 N_{ref} 个位置中预测其对应的参考补丁位置。特别是，一个分类层处理查询补丁表示 U 以输出每个查询补丁的位置预测 $O \in \mathbb{R}^{N_{ref} \times N_q}$ 。我们最小化损失

$$\frac{1}{|\Omega|} \sum_{j \in \Omega} \ell(O_j, h(j)) \tag{1}$$

。

其中 Ω 是查询补丁在参考视图中具有对应补丁位置的集合，ℓ 是 softmax 交叉熵损失。

为了促进跨组和跨模态的交互，我们在自注意力和交叉注意力块中阻止同组内的 patch 相互参与。这促使模型基于来自不同组和模态的信息形成表示，而不是过度依赖同组内的 patch。具体来说，我们定义一个二元掩码 M ，其中当 patch i 和 j 属于同一组时 $M_{i,j} = 0$ ，否则为 $M_{i,j} = 1$ 。该掩码应用于：

- 自注意力：防止查询块或参考块之间的组内注意力。
- 交叉注意：防止查询补丁关注同一组中的参考补丁

蒙版注意力计算为：

$$E = \text{softmax}(\frac{QK^\top}{\sqrt{d}} \odot M)V$$

这里 K 、Q 和 V 是标准的键、查询和值注意力矩阵，$\odot$ 表示元素级乘法。

为了改变位置预测任务的复杂性，我们按照 LOCA [5] 的方法，遮掩住查询补丁表示可见的参考补丁表示 Z_{ref} 的比例 η 。

为了学习对像素级分类（语义分割的一个基本部分）有效的表示而不依赖标签，我们遵循 LOCA 方法，通过聚类生成伪标签。伪标签（软聚类分配）是基于（可学习的）聚类原型 $Q \in \mathbb{R}^{K \times \tilde{d}}$ 与参考视图 $\tilde{Z} \in \mathbb{R}^{N_{ref} \times \tilde{d}}$ 的投影块表示之间的相似度获得的。查询中的块 i 因此将拥有一个伪标签

$$y^j = \text{Sinkhorn-Knopp}\left(\text{softmax}\left(\tilde{Z}^j_{ref} \cdot Q/\tau\right)\right)$$

，其中 $j = h(i)$ 和 τ 是控制 softmax 分布锐度的温度参数。我们使用 $\tau = 0.05$ 。$\tilde{Z}$ 是通过一个两层 MLP 投影的 Z 。使用 Sinkhorn-Knopp 算法防止模型崩溃为一个平凡解 [5] 。我们最小化目标

$$\frac{1}{|\Omega|} \sum_{j \in \Omega} \ell((Q^\top \tilde{Z}_q)_j, y_j) \tag{2}$$

。

与 LOCA [5] 中的做法类似，我们通过最大化平均熵来正则化此损失，以鼓励网络使用所有聚类原型。

组合目标包括方程 1 和 2 ，并且权重相同。

我们在 MMEarth 多模态卫星影像数据集上对我们的模型和基线方法进行预训练。我们使用来自 MMEarth 的 300,000 个样本来减少预训练时间，并且仅使用 Sentinel 2、Sentinel 1 和 Aster DEM 模态。我们使用 AdamW 优化器进行预训练，学习率为 6.25×10^{-5} ，采用余弦

Table 1. 哨兵 2 的通道分组。在有和没有通道分组的哨兵 2 图像上的洪水类别的 IoU 和 Sen1Floods11 的 mIoU 比较。

Channel grouping. Pretraining	Finetuning	IoU (flood)	mIoU
		69.12	82.51
	☒	73.06	84.75
☒	☒	73.90	85.24

Table 2. 群体采样。群体采样对 Sen1Food11 的计算成本和洪水分割性能的影响。

Group setting	Group sampling	Speedup	IoU (flood)	mIoU
S2 Similar		—	73.90	85.24
	☒	$\times$ 4.2	73.08	84.76
Best		—	72.86	84.64
	☒	$\times$ 12.2	72.82	84.61

调度，批量大小为 64，权重衰减为 0.1。我们和基线方法的模型都预训练了 100 个轮次。我们使用各自的公开实现源码来训练基线方法。通过在 Sen1Floods11 洪水制图语义分割数据集上进行端到端微调来完成评估。我们使用一个轻量解码器，它包含四个转置卷积层和一个最终输出分割概率的卷积层，以防止预训练权重被重解码器消耗。报告的评估结果经过三次运行取平均值。

我们比较了在有和没有通道分组的情况下对 Sentinel 2 图像进行预训练的表现。根据 SatMAE，我们按空间分辨率和波长的相似性对 Sentinel 2 波段进行分组，如下所示。（关于波段的详细信息，请参见附录。）我们称这个分组为“S2 相似性”。表 1 中的结果表明，在处理多光谱图像时，通道分组非常重要，在微调和预训练阶段应用时可以提高性能。在预训练阶段，采用通道分组，一个特定通道组中的查询补丁，例如 SWIR 波段，会预测其在包括所有组的参考视图中的位置。我们假设这种跨组交互挑战了模型提取和关联每个组中特殊信息的能力，从而获得更丰富的聚合信息。

为了管理预训练的计算成本，我们在每个补丁位置随机抽取一组，从而保持恒定的序列长度。表 2 显示，对于 S2 相似组设置，我们在千兆浮点运算上减少了 $\times 4.2$ ，但代价是 mIoU 降低了 -0.48 。最佳组设置是最终表现最好的组设置（参见段落“将 DEM 模式添加为通道组”）。它包含 6 组，因此组抽样导致千兆浮点运算减少了 $\times 12.2$ 。有趣的是，性能仅下降了 -0.03 的 mIoU。

所有接下来的实验都是通过组采样进行的。

我们使用通道组架构添加了 Sentinel 1 模式。我们定义了新的通道组设置，其中包括 Sentinel 1 波段，如下所示。

- S2+S1 分离：S2 相似 + { (A-VV, A-VH, D-VV, D-VH), (A-HH, A-HV, D-HH, D-HV) }
- RGBN+S1 Separate : { (B2), (B3), (B4), (B8), (A-VV, A-VH, D-VV, D-VH), (A-HH, A-HV, D-HH, D-HV) }
- S2+S1 混合：S2 相似 + { (B1, A-VV, A-VH, D-VV, D-VH), (B1, A-HH, A-HV, D-HH, D-HV) }

表中的结果显示，只要两种模式分开分组（如 S2+S1 相似和 RGBN+S1 组设置），添加 Sentinel 1 波段就可以提高性能。这些设置鼓励跨模态交互，因为来自一种模态的查询块表示必须通过关注所有模态来预测其位置。我们假设这种跨模态交互教会模型更有效地从多模态数据中提取和结合信息。我们还观察到，将 Sentinel 1 模式作为独立通道组添加使预训练任务更

Table 3. 关于使用通道组架构添加 SAR 模态的消融研究

Group setting	IoU (flood)	mIoU	Pretraining objective (acc@1)
S2 相似	73.08	84.76	60.5
S2+S1 Separate	73.68	84.87	35.33
RGBN+S1 Separate	73.38	85.10	30.93
S2+S1 Mixed	72.40	84.35	54.92

Table 4. 添加 DEM。添加第三种模态对 Sen1Flood11 性能的不同策略的影响

Group setting	η	IoU (flood)	mIoU	Pretraining objective (acc@1)
S2+S1 分离	80 %	73.68	84.87	35.33
S2 + S1 + DEM Separate	80 %	72.11	84.38	13.8 %
	100 %	73.88	85.21	1.54 %
Best	80 %	72.44	84.35	35.20 %
	100 %	74.52	85.52	1.57 %

具挑战性，导致预训练目标的准确性降低 -25% 。混合不同模态的波段（如在 S2+S1 混合组设置中）并没有提高性能。混合波段减少了跨模态交互的需要，因为查询块表示已经拥有来自所有模态的信息，并且可以依赖方便的模态来解决预设任务。我们还发现，在混合设置中预设任务并不比单模态设置中的任务更难（54.92% 对比 60.50% ）。

将 DEM 模态添加为通道组. 为了通过通道组添加 DEM，我们引入了两个新的通道组设置：

- S2 + S1 + DEM 分别：S2 + S1 分别 + { (DEM) }
- 最佳：{ (B1, B2), (B3, B7), (B4, B8A), (B11), (DEM, A-VV, A-VH, D-VH), (A-HH, A-HV, D-VV, D-HH) }

“最佳” 组设置是在 Sen1Floods11 上提供最佳性能的设置。它将 MSI 和 SAR 模态分开，但将 DEM 混合到 SAR 中。

表格 4 显示，将 DEM 作为一个单独的通道组添加使得预训练任务更具挑战性，导致位置预测精度降低（-16.53 %）。将 DEM 混合进现有模式，使预训练任务相对简单，导致位置预测精度小幅下降（-0.13 %）。这表明，整合模态的策略是一个可以调整的超参数，用以控制预训练任务的难度并提高迁移性能。

有趣的是，一个参考掩膜率为 $\eta = 100\%$ 时达到最佳性能，显示查询补丁没有必要 “查看” 参考视图表示。因此，通过不包括交叉注意力块，我们方案的复杂性和计算成本可以降低。

同组注意力屏蔽. 我们尝试使用同组注意力掩码作为一种技术，通过促进跨模态交互来改善多模态学习。表 5 显示，当参考掩码比例较低（$\eta = 60\%$ ）时，同组注意力掩码显著提高了迁移性能（$+1.89$ IoU）。然而，将参考掩码比例增加到 $\eta = 100\%$ 则导致性能轻微下降（-0.06 mIoU）。同组注意力掩码使预训练任务变得更具挑战性，从而帮助模型学习更好的表示。然而，将其与激进的参考掩码结合使用会降低其效果，因为可供关注的参考表示很少，并且可能会使预训练任务对模型来说过于具有挑战性，不利于学习良好的表示。

补丁簇预测. 表显示，包含补丁集群预测任务显著正面影响了迁移性能（$+1.77$ mIoU）。

与其他卫星图像 SSL 预训练方案的比较. 我们将我们的预训练方案与其他流行方案进行比较。我们使用他们公开可访问的实现源代码，在 MMEarth 数据集上对 ViT-Small 编码器

Table 5. **同组注意力屏蔽。同组注意力屏蔽和参考屏蔽对 Sen1Floods11 迁移性能的影响**

η	Same-group atten. masking	IoU (flood)	mIoU	Pretraining objective (acc@1)
60 %		71.99	84.07	53.15
	☒	73.88	85.21	44.11
100 %		74.62	85.52	1.57
	☒	74.56	85.49	1.57

Table 6. **补丁簇预测。包括补丁簇预测任务的效果。**

Cluster loss	IoU (flood)	mIoU
☒	73.88	85.21
	72.11	84.06

Table 7. **与其他 SSL 预训练方案在 Sen1Floods11 上的比较**

Scheme	Encoder	IoU (flood)	mIoU
Satellite LOCA (ours)	ViT-Small	74.62	85.49
MMEarth [1]	ConvNext-T	68.92	82.34
ScaleMAE [17]	ViT-Small	68.85	82.29
SatMAE++ [15]	ViT-Small	67.37	81.47
SatMAE [6]	ViT-Small	65.28	80.56

（或对 MMEarth 方案的 ConvNext-T 编码器 [1] ）进行 100 次迭代的预训练。对于 MMEarth 方案，我们仅使用 Sentinel 1 和 Sentinel 2 模式进行预训练。通过端到端微调轻量解码器（包括 4 个转置卷积层加上一个最终像素级分类卷积层）进行评估。我们报告方案的单次微调运行结果。

表 7 显示，我们采用的 LOCA 方法在 Sen1Floods11 数据集的卫星图像语义分割中表现明显优于其他方法。

4 结论

我们调整了 LOCA，一种位置预测的自监督学习方法，用于多模态卫星图像的语义分割。我们的主要贡献包括扩展通道分组以处理多模态数据，引入同组注意力掩码以促进跨模态交互，并使用组采样在预训练期间保持计算效率。在 Sen1Floods11 上的实验结果表明，我们的方法显著优于现有基于重建的卫星图像自监督方法。未来的工作可以探索在预训练中引入尺度不变机制，如 ScaleMAE [17] ，利用卫星数据的时间维度，扩展到其他模态，并在更多样化的下游任务中评估迁移学习。

References

[1] Ankit, Oehmcke Stefan, Belongie Serge, Igel Christian, Lang Nico Nedungadi Vishal, and Kariryaa. 2025. MMEarth: Exploring Multi-modal Pretext Tasks for Geospatial Representation Learning. In *Computer Vision–ECCV 2024* (Cham), Elisa, Roth Stefan, Russakovsky Olga, Sattler Torsten, Varol Gül Leonardis Aleš, and Ricci (Eds.). Springer Nature Switzerland, 164–182.

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Geography-Aware Self-Supervised Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 10161–10170. doi:10.1109/ICCV48922.2021.01002

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (Feb. 2019), 423–443. doi:10.1109/TPAMI.2018.2798607

[4] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. 2020. Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2020-June (6 2020), 835–845. doi:10.1109/CVPRW50498.2020.00113

[5] Mathilde Caron, Neil Houlsby, and Cordelia Schmid. 2024. Location-Aware Self-Supervised Transformers for Semantic Segmentation. *Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision, WACV 2024* (1 2024), 116–126. doi:10.1109/WACV57701.2024.00019

[6] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B Lobell, and Stefano Ermon. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. *Advances in Neural Information Processing Systems* 35 (12 2022), 197–211. https://sustainlab-group.github.io/SatMAE/

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North* (2019), 4171–4186. doi:10.18653/V1/N19-1423

[8] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1422–1430. doi:10.1109/ICCV.2015.167

[9] William Emery and Adriano Camps. 2017. *Introduction to Satellite Remote Sensing*. Elsevier. https://www.sciencedirect.com/book/9780128092545/introduction-to-satellite-remote-sensing

[10] Pedram Ghamisi, Behnood Rasti, Naoto Yokoya, Qunming Wang, Bernhard Hofle, Lorenzo Bruzzone, Francesca Bovolo, Mingmin Chi, Katharina Anders, Richard Gloaguen, Peter M. Atkinson, and Jon Atli Benediktsson. 2019. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* 7 (3 2019), 6–39. Issue 1. doi:10.1109/MGRS.2018.2890023

[11] Astruc Guillaume, Gonthier, Nicolas, Mallet Clement, and Landrieu Loic. 2025. OmniSat: Self-supervised Modality Fusion for Earth Observation. In *Computer Vision –ECCV 2024* (Cham), Elisa, Roth Stefan, Russakovsky Olga, Sattler Torsten, Varol Gül Leonardis Aleš, and Ricci (Eds.). Springer Nature Switzerland, 409–427.

[12] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. 2024. Bridging Remote Sensors with Multisensor Geospatial Foundation Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27852–27862. doi:10.1109/CVPR52733.2024.02631

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2022-June (2022), 15979–15988. doi:10.1109/CVPR52688.2022.01553

[14] Zhengtao Li, Guokun Chen, and Tianxu Zhang. 2020. A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 847–858. doi:10.1109/JSTARS.2020.2971763

[15] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. 2024. Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (6 2024), 27811–27819. doi:10.1109/CVPR52733.2024.02627

[16] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 69–84.

[17] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. 2023. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. *Proceedings of the IEEE International Conference on Computer Vision* (2023), 4065–4076. doi:10.1109/ICCV51070.2023.00378

[18] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, Hairong Qi, and Min H Kao. 2023. Cross-Scale MAE: A Tale of Multiscale Exploitation in Remote Sensing. *Advances in Neural Information Processing Systems* 36 (12 2023), 20054–20066.

[19] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. 2024. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv:2304.14065 [cs.CV] https://arxiv.org/abs/2304.14065

[20] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. 2025. Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. arXiv:2502.09356 [cs.CV] https://arxiv.org/abs/2502.09356

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[22] Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (10 2023), 12113–12132. doi:10.1109/TPAMI.2023.3275156

[23] Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Y Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua M Susskind. 2022. Position Prediction as an Effective Pretraining Strategy. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 26010–26027. https://proceedings.mlr.press/v162/zhai22a.html

[24] Cheng Zhang, Wanshou Jiang, Yuan Zhang, Wei Wang, Qing Zhao, and Chenjie Wang. 2022. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). doi:10.1109/TGRS.2022.3144894

[25] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5122–5130. doi:10.1109/CVPR.2017.544

[26] Adrian Ziegler and Yuki M. Asano. 2022. Self-Supervised Learning of Object Parts for Semantic Segmentation . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 14482–14491. doi:10.1109/CVPR52688.2022.01410

A 模式和频段

表 8 列出了本工作中使用的模态的波段。

Table 8. 使用的模态和频段，以及用于引用它们的代码。

Modality	Code	Name	Spatial Resolution (metres)
Multspectral Satellite Imagery (Sentinel 2)	B1	Ultra-blue	60
	B2	Blue	10
	B3	Green	10
	B4	Red	10
	B5	Red edge 1	20
	B6	Red edge 2	20
	B7	Red edge 3	20
	B8	Near-infrared	10
	B8A	Red edge 4	20
	B9	Water vapour	60
	B10	Cirrus	60
	B11	Shortwave-infrared 1	20
	B12	Shortwave-infrared 2	20
Synthetic Aperture Radar (Sentinel 1)	A-VV	Ascending orbit VV	10
	A-VH	Ascending orbit VH	10
	A-HH	Ascending orbit HH	10
	A-HV	Ascending orbit HV	10
	D-VV	Descending orbit VV	10
	D-VH	Descending orbit VH	10
	D-HH	Descending orbit HH	10
	D-HV	Descending orbit HV	10
Digital elevation model	DEM	Elevation	30

Received 6 June 2025