档　号	件　号

编　　号_____

密　　级　　公开

阶段标记_____

页　　数_____

CG-EarthEye 大模型技术报告

会　签

编　写：_____

校　对：_____

审　核：_____

标　审：_____

批　准：_____

长光卫星技术股份有限公司

目录

一、引言

“吉林一号”卫星星座是目前全球最大的亚米级商业遥感卫星星座，具备全球一年覆盖 6 次、全国半月覆盖 1 次、全球任意地点每日 38~40 次重访的能力，在国土安全、地理测绘、土地规划、农林生产、生态环保、智慧城市等各领域发挥了重要作用。面对高频次、超大数据量的吉林一号卫星遥感数据，传统基于机器学习、人工判读的遥感解译方式，难以满足现阶应用需求，如何利用海量吉林一号卫星遥感影像构建基础大模型，支撑各种解译任务，是一个长期的挑战。

深度学习的兴起，使得遥感影像解译技术快速发展。大量计算机视觉领域深度学习模型，例如 ResNet[1]、DeepLabV3[2]、HRNet[3]、ConvNeXt[4]等被用于遥感影像解译[5][6]，并在一些特定的遥感解译场景中表现出优越性。上述方法均采用迁移学习的方式，将计算机视觉领域的预训练模型权重应用到遥感领域，但是由于计算机视觉领域图像与遥感影像之间存在巨大的领域差异，使得模型仍然极度依赖高质量的遥感标注数据，并且泛化性能受限[7]。

近年来，随着大模型技术的发展，利用自监督算法从海量遥感影像中学习领域知识，构建的遥感大模型在模型规模、精度以及泛化性能上表现出优越性。张良培团队[9][10]利用生成式模型 MAE 在 MillionAID 数据集上预训练了遥感大模型，并进一步利用多任务学习的方式进行了模型优化；李彦胜团队[11]利用对比学习的方式，在 2150 万样本规模的多源遥感影像数据上预训练了包含 20 亿参数的 SkySense；Liu 等[11]利用对比学习的方式，在大规模遥感影像和文本数据集上预训练了遥感多模态大模型。总之，生成式和对比式框架已经成为构建遥感领域大模型的主要方法[13]。

尽管现阶段涌现出大量遥感大模型，但仍然存在以下方面的不足：在数据方面，相较于计算机视觉，遥感领域自监督样本的规模有限，高分辨率遥感影像约在百万量级，并且融合了多个遥感公开数据集，可能数据冗余。同时，受限于高分辨遥感影像的来源，样本采集区域分布不均，无法有效表征地表信息；在模型方面，基于生成式的遥感大模型偏向于高频信息的建模，缺乏对影像高级语义信息的理解；基于对比式的遥感大模型偏向于高级语义信息建模，对于相对低级的密集预测任务缺乏优势。在前人的基础上，针对上述不足，本文构建了吉林一号遥感大模型，主要贡献如下：①在全球范围内，基于 2023 年吉林一号全球一张图影像以及 4 张中国季度一张图影像，构建了 1500 万亚米级时序高分辨率遥感自监督数据集；②结合生成式和对比式框架，提出了基于掩码重建与对比学习的多粒度遥感影像自监督学习算法，从高级语义

理解与低级像素预测两方面充分建模高分辨率遥感影像；③基于分布式架构，构建了一个吉林一号大模型应用推理集群，能实现全球目标的自动化稳定提取。实验结果表明，吉林一号大模型在多项任务中达到目前先进水平，能高效稳定建模吉林一号卫星遥感影像特征，并能通过微调的方式快速应用于各种遥感解译任务，对于提升吉林一号卫星遥感影像的智能化水平有着重要意义。

二、吉林一号自监督数据集

吉林一号自监督数据集——JLSSD，是一个专为吉林一号大模型预训练构建的大规模高分辨率卫星遥感影像数据集。该数据集是通过全球分布的 10000K 个采样点位，从 0.75 米空间分辨率吉林一号全球一张图影像以及 4 张中国季度一张图影像裁剪而来。包含 2015K 个中国点位的 8060K 张 4 季度影像样本，以及 7985K 个点位的 7985K 张 2023 年年度合成一张图影像样本，如图 1 所示，展示了 JLSSD 覆盖范围。同时，它具有以下 3 个关键属性：

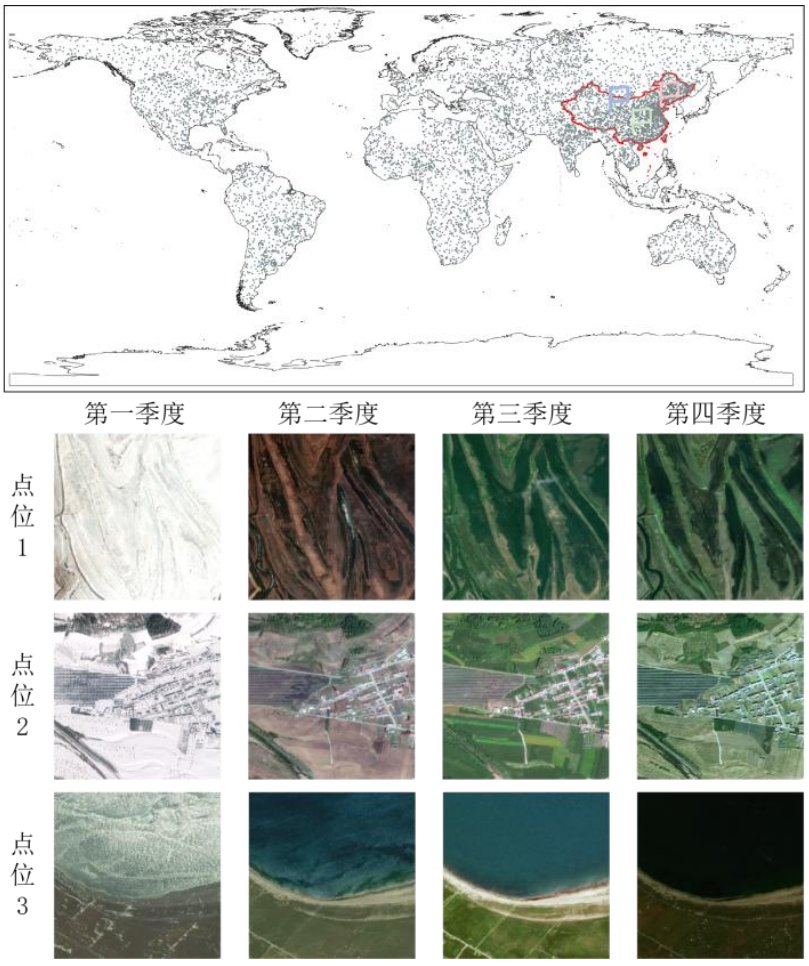

图 1 JLSSD 全球分布图

（一）多样性

JLSSD 具备地表场景多样性特点，首先利用 1 公里×1 公里格网将全球进行划分，然后利用 WorldCover 全球地表覆盖分类数据，全球 DEM 数据、全球行政区划数据在内的多源地学数据筛选出样本格网。所筛选样本全球分布，涵盖各种地形地貌。

筛选过程采用基于分层抽样思想，对不同行政区、不同地物类别、不同高程区间的网格进行划分，并基于具体的类别统计信息计算抽样频次。具体的，预设高程区间宽度为 h，总样本目标网格数为 M，各类别抽样比例为$A_1, A_2, ..., A_7,$，统计全球网格各类别网格数分别为$N_1, N_2, ..., N_7,$，其中公式中角标 1-7 对应林地，草地，耕地，水体，湿地，建筑区，其他。如果行政区 K 内类别 p 对应网格高程区间为$[\sigma_1, \sigma_n)$，则有高程区间集合 $D:\{D_i = [\sigma_1 + (i-1)h, \sigma_1 + ih) | i \in 1,2, ..., N\}$，其中 $N = \left\lfloor \frac{\sigma_n - \sigma_1}{h} \right\rfloor$。设各高程区间网格数分$k_1, k_2, , ..., k_7$，则有行政区 K 中类别为 p 且高程区间为$D_i$的层次采样频次计算方法如下：

$$\alpha_{Kpi} = \max(1, M \times A_p \times \frac{k_i}{N_p}) \tag{1}$$

（二）时序性

得益于“吉林一号”卫星星座的强大数据获取与处理能力，目前已生产了 2023 年 4 个季度的全国亚米一张图影像。因此，JLSSD 包含了中国区域的 x 组对齐的季度样本。据我们所知，这是目前最大的，季度性的亚米级遥感时序自监督数据集。

（三）异质性

JLSSD 采用基于聚类的数据过滤策略来减小冗余类型（如沙漠、水体）场景影像和低质量图像，平衡了影像之间的多样性和图像内的异质性。异质性会增加自监督影像建模的难度，从而增强模型的特征表示能力。

三、CGEarthEye 架构

本文提出的基于掩码重建与对比学习的多粒度遥感影像自监督学习算法框架如图 2 所示。该模型包含数据增强模块、特征计算模块以及损失计算模块。其过程为：输入一个影像样本，经过数据增强模块，得到 10 种样本的数据增强，然后经过特征计算模块，编码影像特征，最后利用对比学习损失对特征计算模块进行优化。

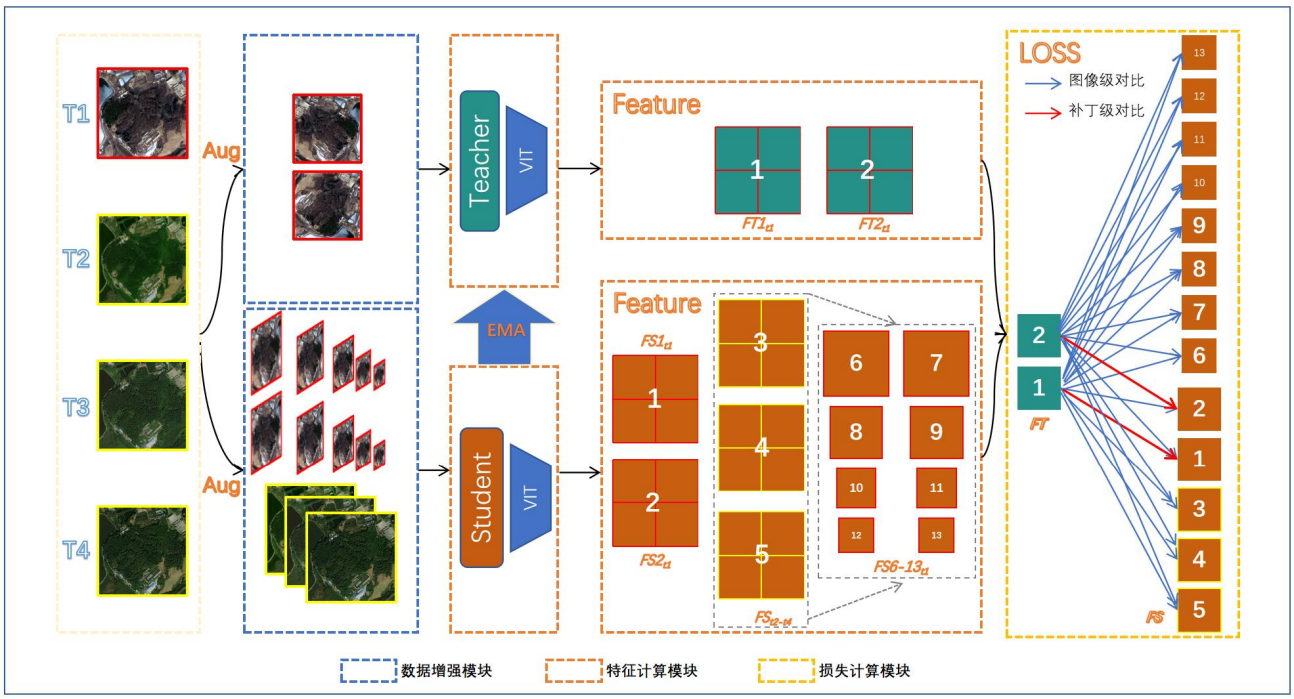

图 2 吉林一号大模型预训练框架图

（一）数据增强模块

对比学习的核心在于，同一影像的不同数据变换经过模型编码后具备相同的语义信息。因此，对于输入影像 T1，首先进行多尺度影像裁剪，通过在不同缩放比例下对图像进行全局裁剪和局部裁剪，生成多尺度的影像输入，增强模型对局部细节和全局语义的理解能力。然后对裁剪后的影像进行随机的颜色抖动，包括调整亮度、对比度和饱和度，增强模型对光照变化的鲁棒性。接着，对影像进行几何变换，包括随机的水平翻转或垂直翻转。最后，对于全局裁剪的影像，随机进行 10%~50%的像素块掩码，用于掩码重建任务。经过一系列数据变换后，T1 样本影像得到 2 个全局裁剪数据增强影像、对应掩码影像以及 8 个多尺度局部裁剪数据增强影像。同时，对于具备季度时序影像的样本，获取对应 3 个季度影像样本，经过多尺度裁剪以及几何变换后，随机替换原始 8 个多尺度局部裁剪数据增强影像中的 3 个。因此，一个影像样本在特征计算之前，将变换得到 12 个数据增强后的影像，用于自监督训练。

（二）特征计算模块

特征提取模块主要用于对数据增强模块输出进行特征编码，包含教师分支与学生分支，采用相同的 VIT 模型[15],模型结构如图 3 所示。吉林一号大模型参数量共计 21 亿，包含了 5 种不同参数量大小的 ViT 模型，参数量从 22M~1100M，以适应不同的业务场景，如表 1 所示。

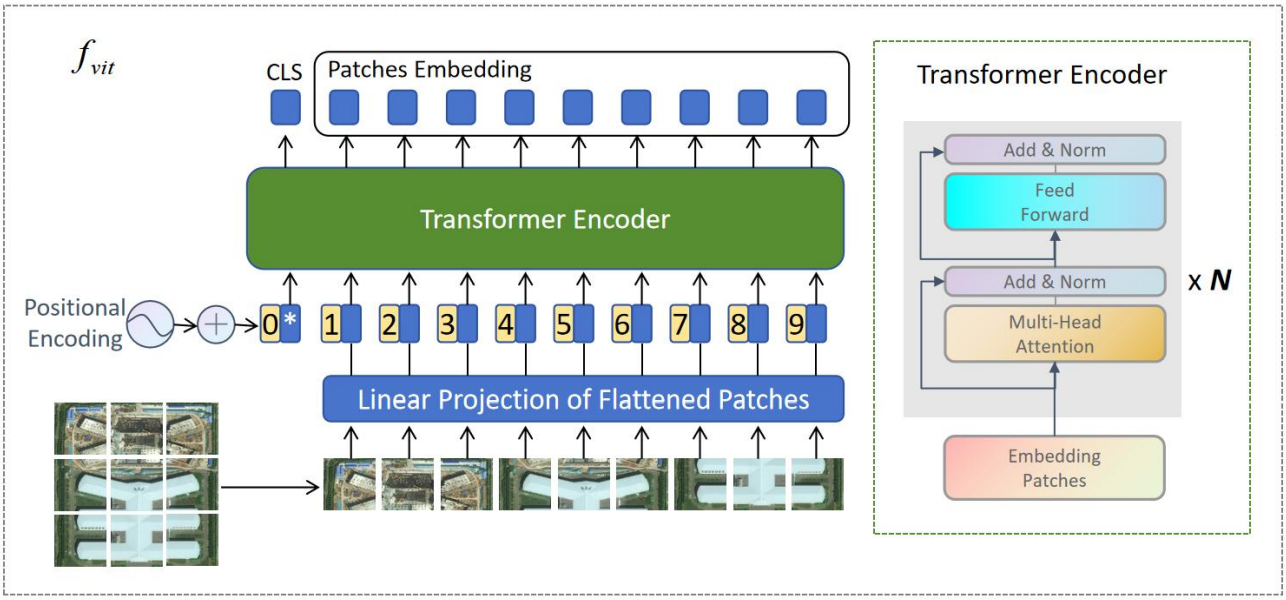

图 3 ViT 模型架构图

表 1 CGEarthEye 模型参数表

模型	层数	编码维度	隐藏层维度	注意力头	参数量/M
CGEarthEye-Small	12	384	1536	6	22
CGEarthEye-Base	12	768	3072	12	86
CGEarthEye-Large	24	1024	4096	16	307
CGEarthEye-Huge	32	1280	5120	16	632
CGEarthEye-Giant	40	1536	6144	24	1100

经过全局裁剪的 2 个影像首先经过教师分支编码后得到特征 $FT1_{t1}$、$FT2_{t1}$，对应经过掩码后的影像输入到学生分支进行编码后得到特征 $ST1_{t1}$、$ST2_{t1}$，然后 8 种局部裁剪以及 3 种季度性的影像利用学生分支编码后得到特征 $FSi_{t1}, i \in [6,13]$、$FSi_{t1}, i \in [3,5]$。

（三）损失计算模块

为了建模模型对于遥感影像的全局理解能力，本文引入了对比学习损失[16]，将 2 个分支输出的分类 token 编码特征经过 3 层全连接神经网络变换后，进行损失的计算，具体计算方式如下：

$$L1 = -\sum_{T}\sum_{S} p_t \log p_s \tag{2}$$

其中，p_t表示教师分支 2 个全局增强影像的经过全连接层后的分类编码，p_s表示学生分支 8 个局部增强影像的经过全连接层后的分类编码，对于中国区域的样本，随机取 3 个局部裁剪增强影像，利用季度性影像进行替换。

为了建模模型对于遥感影像的像素预测能力，本文进一步将对比损失的计算应用到掩膜块的编码上，利用掩膜块对应位置教师分支的输出进行监督，具体计算方式如下：

$$L2 = -\sum_{i} p_{ti} \log p_{si} \tag{3}$$

其中 i 是掩码标记的补丁索引，利用教师网络对应输出对学生网络进行监督。

最终模型的损失计算方式如下：

$$L2 = L1 + L2 \tag{4}$$

模型通过正向传播计算完损失后，进行反向传播求导，在反向传播的过程中只激活学生分支的网络参数，并利用随机梯度下降算法进行参数的更行。对于教师分支的网络参数，利用动量的方式进行参数的更新，来避免训练过程中模型的坍塌，具体计算方式如下：

$$\theta_t = m\theta_{t-1} + (1-m)\theta_s \tag{5}$$

式中θ_s、θ_t表示当前时刻更新后的学生分支网络参数与教师分支网络参数、θ_{t-1}表示当前时刻还未更新的教师分支网络参数，m表示动量权重，本文取 0.992。

四、实验与分析

（一）模型训练

表 2 实验软硬件环境配置

实验环境	配置说明
硬件环境	CPU:2*Intel8358P 2.6GHz/32 核/48MB/240W
	GPU:2*8*A800GPU/1280GB
	内存：2*32*32GB DDR4
	存储：4*7.68TB NVMe SSD
	网络：4*200G IB+ 1*100G IB
软件环境	操作系统：Ubuntu 20.04.5 LTS
开发环境	Visual Studio Code 集成开发环境
算法框架	Pytorch2.0.0 深度学习框架

为了加速模型的训练，本文引入的多种加速技术进行全方面的训练加速。在数据层面，利用 LMDB 进行训练数据的存储与管理，能极大提升海量数据的读取效率。在模型层面，采用了 FlashAttention[17]算法，获得了约 2 倍注意力计算的效率提升。在训练策略方面，采用了 FSDP[18]技术，将模型、优化器、梯度参数进行切片，并采取混合精度，有效扩大的训练的批量大小。同时，受 DINOv2[19]模型训练启发，本文先利用 224x224 大小全局裁剪与 98x98 局部大小裁剪进行模型训练，然后进一步将全局裁剪设置为 518x518。最终实现超 2 倍训练效率的

提升，约 3 倍内存的节省。实验利用 16 张 80GB A800 GPU 训练了 150 天，训练环境如表 2 所示。

（二）场景分类实验

本文进一步利用 RESISC-45、AID 在目前主流的算法中进行了比较，本节实验在 MMPretrain 框架下完成，结果如表 3 所示。

表 3 CGEarthEye 场景分类实验结果

方法	骨干	RESISC-45	AID
SeCo[26]	ResNet50	0.9291	0.9347
GASSL[27]	ResNet50	0.9306	0.9355
CACo[28]	ResNet50	0.9194	0.9088
SatLas*[29]	Swin-B	-	0.6598
SatLas[29]	Swin-B	0.9470	0.9496
CMID*[30]	Swin-B	-	0.8780
CMID[30]	Swin-B	0.9553	0.9611
RingMo[31]	Swin-B	0.9567	0.9690
GFM*[32]	Swin-B	-	0.7942
GFM[32]	Swin-B	0.9464	0.9547
SatMAE[33]	ViT-L	0.9410	0.9502
Scale-MAE*[34]	ViT-L	-	0.7643
Scale-MAE[34]	ViT-L	0.9504	0.9644
SSL4EO[35]	ViT-B	0.9127	0.9106
RVSA[9]	ViT-B	0.9569	0.9703
SkySense*[11]	Swin-H	-	0.9407
SkySense[11]	Swin-H	0.9632	0.9768
MTP[10]	InternImage-XL	0.9627	-
CGEarthEye*	ViT-G	0.9584	0.9760
CGEarthEye	ViT-G	**0.9675**	**0.9769**

从实验结果可以看出，CGEarthEye 模型实现了性能 SOTA，尤其 CGEarthEye 冻结骨干的精度超过了大多数算法全量微调的精度。这有助于缓解大模型应用过程中微调困难的问题，通过冻结大模型参数，只添加可训练分类头的方式，CGEarthEye-Giant 1.1B 参数量可在 12GB 显存上进行应用微调。

（三）语义分割实验

为了测试 CGEarthEye 在更精细的像素级任务上的微调性能，例如语义分割，它是遥感目标和土地覆盖物识别与提取中最重要的应用之一。本文在 LoveDA[36]、iSAID[37]、Potsdom，3 个数据集上，利用 UperNet 头进行微调与对比实验。本节实验在 MMSegmentation 框架下完成，

结果如表 4 所示。

表 4 CGEarthEye 语义分割实验结果

方法	骨干	LoveDA	iSAID	Potsdom
		mIoU		mF1
SeCo[26]	ResNet50	0.4363	0.5720	0.8903
GASSL[27]	ResNet50	0.4876	0.6595	0.9127
CACo[28]	ResNet50	0.4889	0.6432	0.9135
SatLas*[29]	Swin-B	-	0.5603	-
SatLas[29]	Swin-B	-	0.6871	0.9128
CMID*[30]	Swin-B	-	0.5940	-
CMID[30]	Swin-B	-	0.6621	0.9186
RingMo[31]	Swin-B	-	0.6720	0.9127
GFM*[32]	Swin-B	-	0.6086	-
GFM[32]	Swin-B	-	0.6662	0.9185
Scale-MAE*[34]	ViT-L	-	0.6577	-
Scale-MAE[34]	ViT-L	-	0.4653	0.9154
SSL4EO[35]	ViT-B	-	0.6401	0.9154
RVSA[9]	ViT-B	0.5244	0.6449	-
SkySense*[11]	Swin-H	-	0.6540	-
SkySense[11]	Swin-H	-	**0.7091**	**0.9399**
MTP[10]	InternImage-XL	0.5417	-	
DINOv2[19]	ViT-G	0.5514	0.6833	-
CGEarthEye*	ViT-G	<u>0.5667</u>	<u>0.6951</u>	<u>0.9353</u>

受制于算力开销，CGEarthEye 在语义分割上等后续的下游微调均采用冻结骨干方式进行。实验结果表明，CGEarthEye 在 LoveDA 数据集上呈现出 SOTA 水平，mIoU 达到 56.67%，优于全量微调 MTP 的 54.17%。在 iSAID 和 Postom 数据集上分别低于全量微调的 SkySense 模型 1.4%、0.46%，但高于其冻结精度 4.11%，并高于除全量微调的 SkySense 模型外所有方法，不管冻结还是开放。但是 SkySense 使用超过 1250 万对多模态数据训练，采用地理位置感知、多模态与多时相对比，并利用超 80 台 8 卡 A100 服务器进行训练，整体花销显著高于 CGEarthEye。总之，CGEarthEye 在像素级任务中表现出显著优势，具备优异的图像表征能力，可快速应用于下游语义分割任务。

（四）变化检测实验

变化检测从不同时期的遥感影像中自动识别地表变化，支持资源监测、环境评估和城市管理等时空动态分析。为了评估 CGEarthEye 在变化检测下游任务上的微调性能，本文在 SYSU-CD[38]、CDD[39]、LEVIR-CD[40]3 个数据集上，将 CGEarthEye 冻结骨干提取的特征图进

行逐层融合至下游的 ChangeFormer 变化检测分支网络中进行模型微调，实验在 OpenCD 框架下完成，具体实验结果如表 5 所示。

表 5 CGEarthEye 变化检测实验结果

模型	骨干	LEVIR-CD	SYSU-CD	CDD
ChangeFormer[41]	MiT-B2	0.9111	0.8311	-
BiT-18[42]	ResNet-18	0.8931	-	-
STANet[43]	-	-	0.7736	-
HANet[44]	ResNet-101	0.9028	0.7741	0.8923
CGNet[45]	VGG-16	0.9201	0.7992	0.9473
SGSLN[46]	-	0.9233	0.8307	0.9624
C2FNet[47]	VGG-16	0.9183	0.7797	0.9593
MutSimNet[48]	-	0.9200	0.8234	
CACG-Net[49]	-	0.9229	0.8335	0.9473
MTP[10]	InternImage-XL	0.9267	-	**0.9837**
SkySense[11]	Swin-H	0.9258	-	-
ChangeClip[50]	ViT-B	0.9201	0.8332	0.9789
CGEarthEye*	ViT-G	0.9246	0.8347	0.9804

CGEarthEye 冻结骨干的精度在三个变化检测数据集上表现为最优或次优，仅在 CDD 数据集上低于全量微调的 MTP 模型 0.33%，体现出良好的泛化能力和鲁棒性。此外，与 ChangeFormer 的对比实验也表明了在融入了大模型特征后，能有效提升下游算法性能。

（五）目标检测实验

目标检测是遥感解译中的一项基本任务，旨在自动识别和定位遥感图像中的特定目标，如建筑物、车辆等。根据检测框的朝向不同，遥感目标检测任务可以分为水平检测和定向检测。为了测试 CGEarthEye 在目标检测任务上的微调性能，本文利用 Faster-RCNN[51]、Oriented-RCNN[52]分别在水平框目标检测数据集 DIOR 以及旋转框检测数据集 DIOR-R[53]上进行了对比实验。对于水平框检测任务，本文的实验在 MMDetection 框架下完成，对于旋转框检测任务，本文的实验在 MMRotate 框架下完成。实验结果表 6 所示。

从实验结果可以看出，无论是在水平框目标检测任务还是旋转框目标检测任务上，CGEarthEye 都展现出了优越的性能，即使在冻结骨干的情况下，依然超越其它模型，在 DIOR 和 DIOR-R 数据集上实现了 SOTA 的水平。这说明通过 CGEarthEye 预训练的骨干具备良好的目标级特征提取能力。

表 6 CGEarthEye 目标识别实验结果

方法	骨干	DIOR	DIOR-R
GASSL[27]	ResNet50	0.6740	0.6565
CACo[28]	ResNet50	0.6691	0.6410
SatLas[29]	Swin-B	0.7410	0.6759
CMID[30]	Swin-B	0.7511	0.6637
RingMo[31]	Swin-B	0.7590	--
GFM[32]	Swin-B	0.7284	0.6767
SatMAE[33]	ViT-L	--	0.6566
Scale-MAE[34]	ViT-L	0.7381	0.6647
SSL4EO[35]	ViT-B	0.6482	0.6123
RVSA[9]	ViT-B	0.7322	0.7105
SkySense[11]	Swin-H	0.7873	0.7427
MTP[10]	ViT-L+RVSA	0.8110	0.7454
CGEarthEye*	ViT-G	0.8262	0.7520

五、结论

本文结合地学知识在全球区域构建了一个包含 1500 万吉林一号亚米级卫星遥感影像样本的高质量时序自监督数据集，并提出了基于掩码重建与对比学习的多粒度遥感影像自监督学习算法，训练了包含 21 亿参数的 CGEarthEye。算法结合了生成式框架与对比式框架的优势，并将季节性信息与多尺度信息融入到算法当中，使得模型具备对高分辨率卫星遥感影像全局与局部理解能力。对比视觉领域大模型，CGEarthEye 优势显著，仅仅微调解码器的情况下各项任务优于全量微调的视觉领域大模型。对比遥感领域大模型，CGEarthEye 具备大多数遥感领域大模型不具备的冻结微调能力，在极大缩短应用微调时间与显存的情况下，精度上保持优势，在 4 项任务 10 个数据集上实现性能 SOTA。同时，为服务于吉林一号行业应用，本文微调了 20 种应用模型，并通过分布式推理架构，实现 CGEarthEye-Base 23 小时提取全国地物的效率。后续有望进一步加入吉林一号多光谱卫星遥感影像数据、SAR 卫星遥感影像数据,扩充 CGEarthEye 在多模态卫星遥感影像上的特征提取能力。

参考文献

[1]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[2]Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[3]Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[4]Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[5]徐丹青,吴一全.光学遥感图像目标检测的深度学习算法研究进展[J].遥感学报,2024,28(12):3045-3073.

Xu D Q and Wu Y Q. 2024. Progress of research on deep learning algorithms for object detection in optical remote sensing images. National Remote Sensing Bulliten， 28（12）：3045-3073.

[6]张永军, 李彦胜, 党博, 武康, 郭昕, 王剑, 陈景东, 杨铭. 多模态遥感基础大模型：研究现状与未来展望[J]. 测绘学报, 2024, 53(10): 1942-1954.

Yongjun ZHANG, Yansheng LI, Bo DANG, Kang WU, Xin GUO, Jian WANG, Jingdong CHEN, Ming YANG. Multi-modal remote sensing large foundation models: current research status and future prospect[J]. Acta Geodaetica et Cartographica Sinica, 2024, 53(10): 1942-1954.

[7]张良培, 张乐飞, 袁强强. 遥感大模型：进展与前瞻[J]. 武汉大学学报（ 信息科学版), 2023, 48(10): 1574-1581.

ZHANG Liangpei, ZHANG Lefei, YUAN Qiangqiang. Large Remote Sensing Model: Progress and Prospects[J]. Geomatics and Information Science of Wuhan University, 2023, 48(10): 1574-1581.

[8]Long Y, Xia G S, Li S, et al. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid[J]. IEEE Journal of selected topics in applied earth observations and remote sensing, 2021, 14: 4205-4230.

[9]Wang D, Zhang Q, Xu Y, et al. Advancing plain vision transformer toward remote sensing foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 61: 1-15.

[10]Wang D, Zhang J, Xu M, et al. Mtp: Advancing remote sensing foundation model via multi-task pretraining[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.

[11]Guo X, Lao J, Dang B, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 27672-27683.

[12]Liu F, Chen D, Guan Z, et al. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[13]燕琴, 顾海燕, 杨懿, 李海涛, 沈恒通, 刘世琦. 智能遥感大模型研究进展与发展方向[J]. 测绘学报, 2024, 53(10): 1967-1980.

[14]Qin YAN, Haiyan GU, Yi YANG, Haitao LI, Hengtong SHEN, Shiqi LIU. Research progress and trend of intelligent remote sensing large model[J]. Acta Geodaetica et Cartographica Sinica, 2024, 53(10): 1967-1980.

[15]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[16]Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[17]Dao T, Fu D, Ermon S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness[J]. Advances in neural information processing systems, 2022, 35: 16344-16359.

[18]Zhao Y, Gu A, Varma R, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel[J]. arXiv preprint arXiv:2304.11277, 2023.

[19]Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision[J]. arXiv preprint arXiv:2304.07193, 2023.

[20]Cheng G, Han J, Lu X. Remote sensing image scene classification: Benchmark and state of the art[J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.

[21]Xia G S, Hu J, Hu F, et al. AID: A benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.

[22]Christie G, Fendley N, Wilson J, et al. Functional map of the world[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6172-6180.

[23]Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[24]Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[25]Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[26]Manas O, Lacoste A, Giró-i-Nieto X, et al. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9414-9423.

[27]Ayush K, Uzkent B, Meng C, et al. Geography-aware self-supervised learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10181-10190.

[28]Mall U, Hariharan B, Bala K. Change-aware sampling and contrastive learning for satellite images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5261-5270.

[29]Bastani F, Wolters P, Gupta R, et al. Satlaspretrain: A large-scale dataset for remote sensing image understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 16772-16782.

[30]Muhtar D, Zhang X, Xiao P, et al. Cmid: A unified self-supervised learning framework for remote sensing image understanding[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-17.

[31]Sun X, Wang P, Lu W, et al. RingMo: A remote sensing foundation model with masked image modeling[J].

IEEE Transactions on Geoscience and Remote Sensing, 2022, 61: 1-22.

[32]Mendieta M, Han B, Shi X, et al. Towards geospatial foundation models via continual pretraining[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 16806-16816.

[33]Cong Y, Khanna S, Meng C, et al. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery[J]. Advances in Neural Information Processing Systems, 2022, 35: 197-211.

[34]Reed C J, Gupta R, Li S, et al. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4088-4099.

[35]Wang Y, Braham N A A, Xiong Z, et al. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets][J]. IEEE Geoscience and Remote Sensing Magazine, 2023, 11(3): 98-106.

[36]Wang J, Zheng Z, Ma A, et al. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation[J]. arXiv preprint arXiv:2110.08733, 2021.

[37]Waqas Zamir S, Arora A, Gupta A, et al. isaid: A large-scale dataset for instance segmentation in aerial images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 28-37.

[38]Shi Q , Liu M , Li S ,et al.A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection[J].IEEE Transactions on Geoscience and Remote Sensing, 2022, 60.

[39]Lebedev M A , Vizilter Y V , Vygolov O V ,et al.CHANGE DETECTION IN REMOTE SENSING IMAGES USING CONDITIONAL ADVERSARIAL NETWORKS[J]. 2018.

[40]Chen H , Shi Z .A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection[J].Remote Sensing, 2020, 12(10):1662.

[41]Bandara W G C, Patel V M. A transformer-based siamese network for change detection[C]//IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2022: 207-210.

[42]Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[43]Chen H, Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote sensing, 2020, 12(10): 1662.

[44]Han C, Wu C, Guo H, et al. HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 3867-3878.

[45]Han C, Wu C, Guo H, et al. Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote

Sensing, 2023, 16: 8395-8407.

[46]Zhao S, Zhang X, Xiao P, et al. Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-16.

[47]Han C, Wu C, Hu M, et al. C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[48]Liu X, Liu Y, Jiao L, et al. MutSimNet: Mutually reinforcing similarity learning for RS image change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-13.

[49]Liu F, Liu Y, Liu J, et al. Candidate-aware and Change-guided Learning for Remote Sensing Change Detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[50]Dong S, Wang L, Du B, et al. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2024, 208: 53-69.

[51]Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[52]Xie X, Cheng G, Wang J, et al. Oriented R-CNN for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3520-3529.

[53]Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS journal of photogrammetry and remote sensing, 2020, 159: 296-307.