# 周报（2023. 5.16-2023.5.23）姓名：孙瑞阳

## 每日小结

| | 周一 | 周二 | 周三 | 周四 | 周五 |
|---|---|---|---|---|---|
| 早 | 修改网络代码，上课 | 遥感预训练 VIT 代码 | 遥感预训练 VIT 代码 | 多 loss 代码 | 阅读文献 |
| 中 | 多 loss 平衡，对比学习代码 | 论文阅读 | 多 loss 平衡 | 遥感预训练 VIT 代码，上课 | 多 loss 平衡 |
| 晚 | | 上课 | 整理结果 | 遥感预训练 VIT 代码 | 遥感预训练 VIT 代码 |

注：简单表述当前时间段工作，如看文献 1，整理数据等

## 科研详情

文献阅读

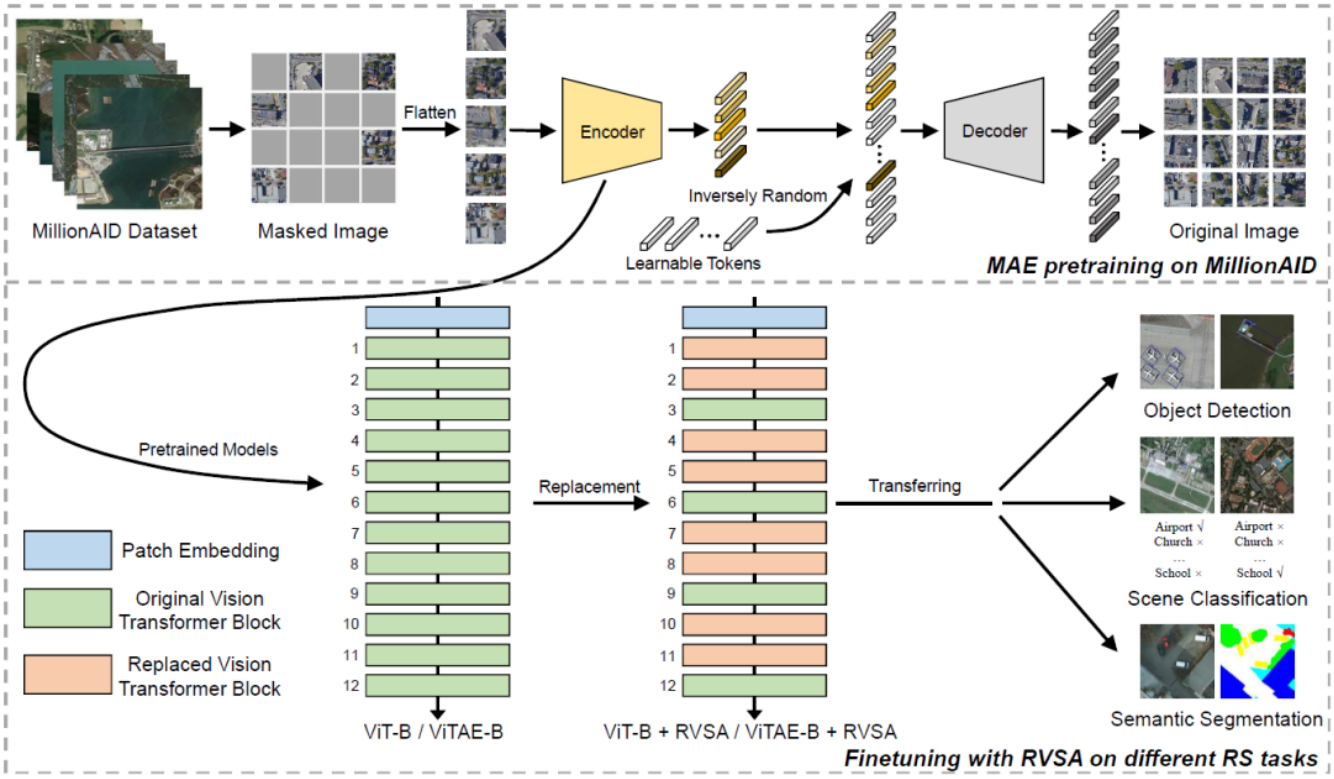
**文献1**

**题目：Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model**

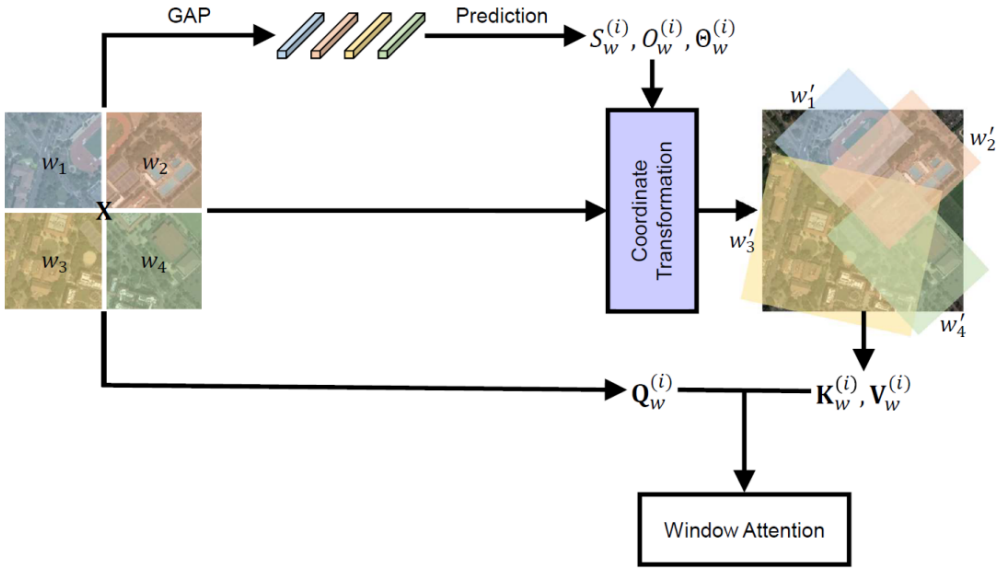
**作者：**Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao and Liangpei Zhang

**出处：**TGRS 2023

**方法：**




1. 首先使用具有约一亿参数的 Plain ViT 模型和研究院最近提出的更先进的 ViTAE 模型，并采用掩码图像建模算法 MAE 在大规模遥感数据集 MillionAID 上对其进行预训练，从而得到很好的初始化参数。**（本周已完成该预训练权重和代码调用）**
2. **采用 RVSA 进行微调：**与自然图像相比，遥感图像通常尺寸更大。由于完全自注意力具有和图片分辨率呈平方关系的计算复杂度，直接将采用完全自注意力的预训练模型应用于下游任务时会显著增加训练成本。为此，在微调阶段采用窗口自注意力替换原始的完全自注意力，这将计算代价降低到与图像大小线性相关的复杂度。RVSA 是引入了一系列变换参数来学习可变方向、大小和位置的窗口，包括相对参考窗口的偏移量、尺度缩放因子以及旋转角度。结构如图：




3. 基于 ViT 和 ViTAE 模型，文章将上述自注意力方法应用于三种遥感感知任务（场景分类、语义分割和目标检测），并提供了预训练模型和权重**（该部分用于场景分类的预训练模型正在改 bug，很快能完成在自己的工作中）**
4. 实验比较（只选取了场景分类）：

| Pretrain | Backbone | Method | UCM-55 | AID-28 | AID-55 | NWPU-19 | NWPU-28 |
|---|---|---|---|---|---|---|---|
| IMP | VGG-16 | LSENet [42] | 98.53 | 94.41 | 96.36 | 92.23 | 93.34 |
| IMP | ResNet-50 | $F^2$BRBM [5] | 98.64 | 96.05 | 96.97 | 92.74 | 94.87 |
| IMP | ResNet-50 | GRMANet [73] | 99.29 | 95.43 | 97.39 | 93.19 | 94.72 |
| IMP | ResNet-101 | EAM [74] | 98.81 | 94.26 | 97.06 | 91.91 | 94.29 |
| IMP | ResNet-101 | MSANet [75] | 97.80 | 93.53 | 96.01 | 90.38 | 93.52 |
| ASP [76] | ResNet-101 | — | — | 95.40 | — | — | 94.20 |
| IMP | DenseNet-121 | MGML-FENet [77] | — | 96.45 | **98.60** | 92.91 | 95.39 |
| IMP | MobileNet-V2 [78] | RBFF [79] | 95.83 | 91.02 | 93.64 | 84.59 | 88.05 |
| IMP | ViT-B | — | 99.15 | 93.81 | 96.08 | 90.96 | 93.96 |
| IMP | Swin-T | — | 99.43 | 96.55 | 98.10 | 92.73 | 94.70 |
| CSPT [80] | ViT-B | — | — | 96.75 | — | — | 95.11 |
| CSPT | ViT-L | — | — | 96.30 | — | — | 95.62 |
| RingMo [32] | ViT-B | — | — | 96.54 | 98.38 | 93.46 | 95.35 |
| RingMo [32] | Swin-B | — | — | 96.90 | 98.34 | 94.25 | 95.67 |
| IMP | ViTAEv2-S | — | 99.43 | 96.61 | 98.08 | 93.90 | 95.29 |
| RSP | ViTAEv2-S | — | **99.62** | 96.91 | 98.22 | **94.41** | 95.60 |
| MAE | ViT-B | — | **99.81** | **97.47** | **98.56** | **94.56** | **95.78** |
| MAE | ViTAE-B | — | 99.60 | **97.20** | 98.42 | **94.43** | **95.82** |
| MAE | ViT-B + VSA | — | 99.41 | 96.85 | 98.30 | 93.74 | 95.29 |
| MAE | ViTAE-B + VSA | — | 99.43 | 96.90 | 98.34 | 93.98 | 95.53 |
| MAE | ViT-B + RVSA | — | **99.70** | 96.92 | 98.33 | 93.79 | 95.49 |
| MAE | ViT-B + RVSA$^\diamond$ | — | 99.58 | 96.86 | 98.44 | 93.74 | 95.45 |
| MAE | ViTAE-B + RVSA | — | 99.56 | **97.03** | 98.48 | 93.93 | **95.69** |
| MAE | ViTAE-B + RVSA$^\diamond$ | — | 99.50 | 97.01 | **98.50** | 93.92 | 95.66 |

**启发：**

1. 已跑通 MillionAID 的 VIT-B-16 预训练模型和权重代码，还没跑完
2. 用于场景分类的 ViT-B + RVSA-**UCM-55** 预训练模型代码正在改 bug，很快能完成在自己的工作中。

**文献2**

**题目：** Multimodal Learning with Transformers: A Survey

**作者：** Peng Xu, Xiatian Zhu, and David A. Clifton

**出处：** TPAMI 2023

**方法：**

这篇综述论文梳理了面向多模态任务的 Transformer 发展。

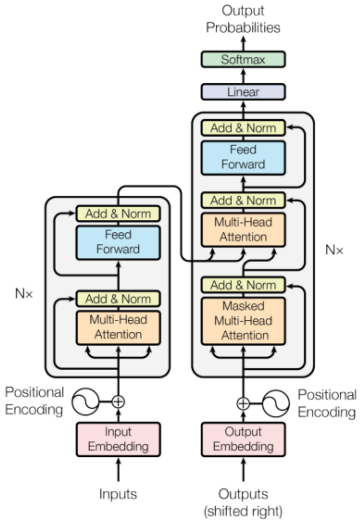



Fig. 1. Overview of Transformer [2].

全文的**主要内容包括**：

（1）对多模态学习、Transformer 生态体系、多模态大数据时代的背景介绍；

（2）以几何拓扑的思想角度对 Transformer、视觉 Transformer、多模态 Transformer 进行了系统性回顾和总结；

（3）从多模态预训练和面向特定多模态任务的两个维度对多模态 Transformer 的应用和研究进行了总结；

（4）对多模态 Transformer 模型及应用中的一些共通的技术挑战和设计思想进行了对比与总结；

（5）并且讨论了该研究社区内的一些开放问题和潜在的研究方向。

**全文的主要观点和特色包括**：

（1）全文以尽可能公式化的方式在多模态上下文中讨论 Transformer 的关键组件。

（2）强调了，在基于 Transformer 的多模态模型中，跨模态的相互交互（例如，融合，对齐）实质上是由自注意力机制及其变体所感知并处理的。所以，从自注意力设计与演变的角度，归纳总结了基于 Transformer 的多模态学习实践中的公式化表达，将常见的基于 Transformer 的多模态交互过程归纳为了 6 种自注意力操作。

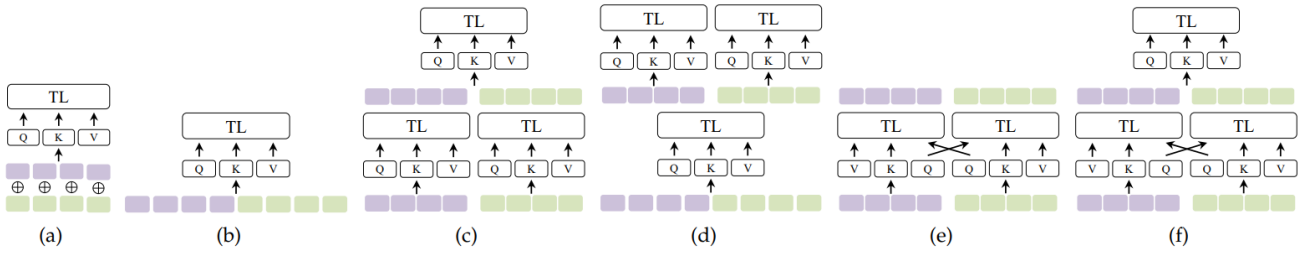



Fig. 2. Transformer-based cross-modal interactions: (a) Early Summation, (b) Early Concatenation, (c) Hierarchical Attention (multi-stream to one-stream), (d) Hierarchical Attention (one-stream to multi-stream), (e) Cross-Attention, and (f) Cross-Attention to Concatenation. “Q”: Query embedding; “K”: Key embedding; “V”: Value embedding. “TL”: Transformer Layer. Best viewed in colour.

**启发**：

该综述的主要观点之一是，强调了 Transformer 的理论优势之一是它能够以模态不可知（modality-agnostic）的方式进行工作，因而可以与各种模态及其组合进行兼容。该文建议将自注意力机制视为一种图式建模，通常在无先验知识的情况下，它将输入序列(单模态和多模态)建模为**全连通图**，自注意力机制将来自任意模态的任意标记令牌的嵌入向量建模为图上的一个节点。这点值得具体看文献学习。

工作进展

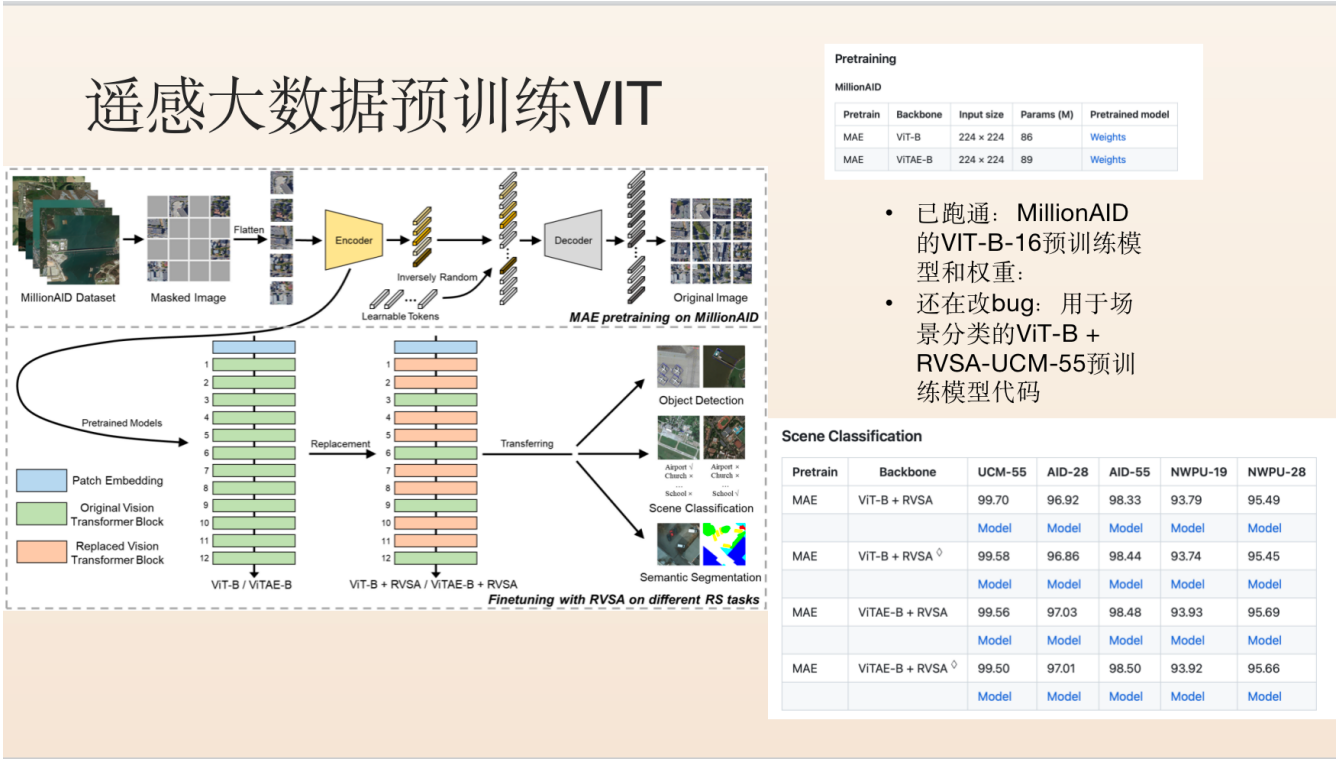
1：阅读文献；

2: 补充了大小数据集上的实验

3: 遥感预训练大数据集VIT：

已跑通MillionAID的VIT-B-16预训练模型和权重代码，还没跑完。跑的有点慢。

用于场景分类的ViT-B + RVSA-**UCM-55**预训练模型代码，快改完了。




Pretraining

MillionAID

| Pretrain | Backbone | Input size | Params (M) | Pretrained model |
|---|---|---|---|---|
| MAE | ViT-B | 224 × 224 | 86 | Weights |
| MAE | ViTAE-B | 224 × 224 | 89 | Weights |

- 已跑通：MillionAID的VIT-B-16预训练模型和权重；
- 还在改bug：用于场景分类的ViT-B + RVSA-UCM-55预训练模型代码

Scene Classification

| Pretrain | Backbone | UCM-55 | AID-28 | AID-55 | NWPU-19 | NWPU-28 |
|---|---|---|---|---|---|---|
| MAE | ViT-B + RVSA | 99.70 | 96.92 | 98.33 | 93.79 | 95.49 |
| | | Model | Model | Model | Model | Model |
| MAE | ViT-B + RVSA ◊ | 99.58 | 96.86 | 98.44 | 93.74 | 95.45 |
| | | Model | Model | Model | Model | Model |
| MAE | ViTAE-B + RVSA | 99.56 | 97.03 | 98.48 | 93.93 | 95.69 |
| | | Model | Model | Model | Model | Model |
| MAE | ViTAE-B + RVSA ◊ | 99.50 | 97.01 | 98.50 | 93.92 | 95.66 |
| | | Model | Model | Model | Model | Model |

4: **多loss平衡正在学习：**

**Balanced Cross Entropy代码已跑通（跑的很慢，loss太大了有点训练），Focal loss已跑通**

5: 和何江师兄交流了 **大数据集去雾** 的问题，准备做大小数据集的泛化性能测试，看效果，才能确定是是因为 缺少数据增强 造成的大数据集效果较差。

下周计划

1. 网络修改：继续看论文想 idea
2. **遥感预训练大数据集VIT：跑完和袁老师交流**
3. 多loss平衡学完：在原来的交叉熵loss上增加了Balanced Cross Entropy、 Focal loss
4. 增加对比学习loss：在找代码，上次找的代码有问题