【开放探索】遥感影像数据自动标注与清洗 Pipeline 搭建

1. 项目背景

目前遥感影像语义分割任务主要依赖人工标注。虽然人工标注质量较高，但面临“三慢”痛点（标注慢、审核慢、返修慢），导致数据生产周期长，难以满足模型快速迭代的需求。

2. 整体目标

利用现有的先进技术（如语义分割模型、SAM 系列、多模态视觉大模型、Agent 等），构建一套**自动化的数据标注与清洗流水线**，实现低成本、高效率的数据生产。

3. 任务详情

3.1 技术路线参考

- **核心参考：** 借鉴 SAMRS (GitHub) 的思路，利用由粗到细（Coarse-to-Fine）或提示工程（Prompt Engineering）的方法进行自动标注。
- **创新点：** 鼓励引入 Agent 或多模态大模型辅助清洗错误标签，提升自动化标注的准确率。

3.2 数据来源

- **数据集：** 本次任务均基于 **SIOR 数据集**（因数据量适中，适合初期验证）。
- **数据路径：** OneDrive 下载链接（注：请优先处理目录下的 `SIOR` 文件夹）。

4. 具体执行要求

4.1 标注产出要求

- **全图标注：** SAMRS 原有方法主要针对部分目标（基于检测框生成 Mask），本任务要求实现**全图（Full-Image）**的语义分割标注，需尽可能覆盖图中的背景与前

景要素。

- **数据量拆分：** 需完成总计 **1400 张** 影像的自动标注，并按以下比例划分数据集：
  - **训练集 (Train):** 1000 张
  - **验证集 (Val):** 200 张
  - **测试集 (Test):** 200 张
  - *注意：验证集和测试集建议进行少量人工抽检或精修，以确保作为评估基准的可靠性。*

4.2 效果验证与交付（考虑到可能无 GPU 服务器，放置低优先级）

- **验证闭环：** 使用生成的“训练集”训练一个标准的语义分割模型（如 DeepLabV3+, U-Net 等），并在“测试集”上进行评估。
- **考核指标：** 测试集上的平均交并比（**mIoU**）需达到 **0.4 及以上**。
- **交付物：**
  a. 自动标注后的数据集文件（Mask 图像或 JSON/XML）。
  b. 数据清洗/自动标注的代码 Pipeline。
  c. 验证实验的训练日志及 mIoU 测试报告。