陨石图像识别：基于 DINOv2 与置信度分诊的多模态融合方法

组 15：侯哲妍、赵萱芝、凌莹、江语欢　Public F1 = 0.84571（#7）

摘要

本项目研究真实用户开放世界图片中的陨石识别问题。任务形式为二分类：给定一张石头图像，判断其是否为陨石。相比标准闭集分类，该任务更接近数据科学中的非平稳分布问题：测试图像来自真实用户，存在光照、尺度、背景、分辨率和拍摄习惯的显著变化；同时训练数据规模有限，陨石、铁矿石、玄武岩、熔渣和河石在外观上高度相似，单纯依赖端到端微调容易过拟合。

最终方案采用“强表征主干 + 参数高效微调 + 结构正则化 + 置信度分诊”的路线。首先以 DINOv2 ViT-B/14 作为自监督视觉表征底座，保留原图上下文而非强行裁剪背景；随后在末 4 个 Transformer block 注入 LoRA，训练少量 adapter 与 MLP 分类头；最后对 LoRA 置信度不稳定的边界样本调用 GPT-4o-mini 进行多模态复核，并用逐样本固定规则融合。该方案在 public leaderboard 上达到 F1=0.84571。

1. 引言

陨石识别具有明确的实际意义：陨石样本稀缺、外观差异大，普通用户上传的图片又往往缺少专业采集条件。若模型只记住训练集中的颜色、背景或尺度模式，在真实测试分布上会快速失效。因此，本项目的研究问题不是“如何在固定数据集上拟合得更好”，而是如何在小样本、强域偏移、类别边界模糊的条件下构建可迁移的判别系统。

本报告的核心贡献包括三点。第一，选择 DINOv2 作为通用视觉特征抽取器，用大规模自监督预训练得到的鲁棒表征抵御开放世界图像分布变化。第二，采用 LoRA 进行任务自适应微调，在冻结主干的同时只训练低秩适配参数，降低小样本过拟合风险。第三，将模型置信度转化为“分诊信号”，只把不确定样本交给多模态大模型复核，避免全量 API 依赖和 Top-N 刷榜式先验。

研究问题	设计原则	对应实现
测试分布未知	不依赖测试集正类比例	逐样本阈值融合，而非 Top-N 选择
训练样本有限	减少可训练参数	冻结 DINOv2，仅训练 LoRA adapter 与分类头
类别边界模糊	把不确定性显式建模	中间概率区间调用 VLM 复核并记录 reason

表 1 研究问题、设计原则与实现映射

2. 数据集介绍

课程数据由有标签训练图像、无标签测试图像与提交模板组成。根据测试阶段材料，Stage 2 测试集共有 194 张真实用户图片；提交文件必须包含 id 与 label 两列，其中 label=1 表示 meteorite，label=0 表示 non-meteorite。数据 README 明确要求以 sample_submission.csv 为模板生成结果，从而保持官方测试图片顺序，并在保存前检查是否存在未预测样本。

字段/文件	含义	在本项目中的作用
train images + labels	课程提供的有标签二分类训练集	用于 LoRA 微调、5 折交叉验证与阈值规则设计
test_images	194 张开放世界无标签图片	最终推理对象，存在强域偏移与背景噪声
sample_submission.csv	官方 id 顺序与提交格式	融合脚本 merge final_label 后生成 submission_fusion.csv

表 2 数据文件与输出格式说明

2.1 数据分析与主要挑战

从图像内容看，数据包含三个层面的困难。第一是域偏移：测试图片在自然光、桌面光、阴影、背景材质、拍摄距离和压缩质量上高度异质，训练分布难以完全覆盖。第二是类间相似：陨石的熔壳、气印和金属颗粒在低分辨率图片中可能并不明显，而铁矿石、熔渣、玄武岩和河石又常常呈现相似的深色粗糙表面。第三是小样本学习：如果直接全参数微调大模型，模型很容易把背景、尺度或图像采集习惯当作捷径特征。

本项目没有把数据问题简单理解为“清洗后训练分类器”，而是把它视为泛化问题。对于确定性较强的样本，LoRA 分类器应独立给出稳定判断；对于落在类别边界附近的样本，单一视觉特征可能不足，需要引入外部视觉常识进行复核。因此，数据分析直接推动了最终的置信度分诊设计。

数据现象	对性能的影响	对应策略
背景与光照差异大	模型可能学习拍摄环境而非石头本体	保留原图上下文，同时做轻量几何/颜色增强
陨石与普通岩石外观相似	边界样本容易被单模型过度自信地误判	用 LoRA probability 识别不确定区间
训练样本少且可能不平衡	全参数微调和准确率指标都不稳定	LoRA、pos_weight、OOF F1 与多 seed 集成

表 3 数据问题、风险与应对策略

2.2 数据预处理与增强

训练阶段保持完整原图上下文。早期尝试过 SAM/CLIP 风格的背景裁剪，但结果从保留原图的 frozen baseline F1=0.740 下降到约 0.700。原因在于陨石识别并不只依赖局部纹理：尺度、边缘形态、阴影、接触面和自然摆放方式都可能为模型提供弱证据。因此最终 LoRA 主干输入采用完整图像，并依赖 DINOv2 的 patch–level 表征自动选择有用区域。

训练增强采用温和扰动而非强破坏：RandomResizedCrop(scale 0.72–1.0, ratio 0.90–1.10)、水平/垂直翻转、12 度以内随机旋转、轻量 ColorJitter，并使用 ImageNet mean/std 标准化。增强目标不是制造过多人工分布，而是模拟真实用户拍摄中的尺度、方向和光照变化，提高模型对非稳定输入的鲁棒性。

VLM 复核分支中，融合脚本保留了 Otsu 阈值、形态学闭运算与最大连通域前景裁剪模块，可在前景比例足够时裁到 512×512；同时最终 GPT 调用以 high–detail 图像输入为核心，避免过度裁剪破坏上下文。这一处理与 LoRA 主干“保留原图”的原则并不冲突：LoRA 负责稳定主判别，VLM 只服务边界样本复核。

2.3 提交格式与可复现检查

README 中的推理流程强调三步：用 predict(model, test_loader, device) 得到以图片文件名为 key 的预测字典；读取 sample_submission.csv 保持官方 id 顺序；用 map 将预测写入 label，并在保存前检查缺失值。最终融合脚本沿用同一原则：先读取 LoRA 详细概率表，再与复核结果按 id 合并，输出 submission_fusion.csv 与 submission_fusion_detail.csv。这样可以同时满足 leaderboard 提交与事后误差分析。

环节	具体做法	目的
训练输入	保留原图；DINOv2 resize 到 518	保留尺度、背景和物理上下文
增强	随机裁剪、翻转、旋转、轻量颜色扰动	模拟真实拍摄变化，防止记忆背景
归一化	ImageNet mean/std	匹配预训练视觉模型输入分布
提交检查	基于 sample_submission.csv 合并并检查 NA	避免顺序错误和漏预测

表 4 数据预处理、增强与提交检查

3. 模型设计与实现

3.1 DINOv2–LoRA 主干

最终模型以 DINOv2 ViT–B/14 为视觉骨干。DINOv2 的优势在于大规模自监督预训练得到的通用视觉特征：在训练样本有限时，它提供了更高的泛化下界。ViT 结构将图像划分为 patch 序列，既能保留全局语义，也能通过 patch mean 表征局部纹理；因此我们将 CLS token 与 patch mean 拼接为 1536 维表示，再送入轻量 MLP 分类头。系统整体结构（主干、LoRA、分类头到分诊路由的端到端数据流）详见附录图 A1。

微调阶段拒绝全参数训练，而是冻结 DINOv2 主干，在最后 4 个 Transformer blocks 的 qkv、proj、fc1、fc2 注入 LoRA adapter。LoRA 的秩 r=8、缩放系数 alpha=16、dropout=0.05。这样做的直觉是：主干中已经包含广泛视觉先验，任务只需要在高层注意力和前馈层中调整少量方向；同时低秩约束降低了模型在小训练集上记忆噪声的能力。

这种设计也给后续分诊提供了概率基础。若主干完全过拟合，输出概率会在错误样本上也极端自信，分诊机制将失效；而 LoRA 与正则化后的模型在多数简单样本上给出接近 0 或 1 的概率，在外观含混样本上保持中间概率，从而使“是否交给 VLM”成为模型自身不确定性的自然函数。

3.2 分类头、损失与训练设置

分类头从单层 Linear 改为 MLP：1536 维输入经 Linear(1536,256)、ReLU、BatchNorm1d、Dropout(0.3) 后输出单个 logit。损失函数为带 label smoothing=0.05 的 BCE，并加入 pos_weight 缓解训练端类别不均衡。优化器使用 AdamW，学习率 1e−4，weight decay 1e−3，Cosine 调度，梯度裁剪为 1.0。训练最多 15 epochs，以验证 F1 早停，patience=3。

验证设计采用分层 5 折交叉验证与全局 out−of−fold 评估，并重复 3 个随机种子。最终 5 folds × 3 seeds 共 15 个模型进行概率平均，测试时使用原图、水平翻转和垂直翻转的 flip TTA。该设计的重点是降低单一划分、单一随机种子和单一视角对结果的偶然影响。

模块	配置
Backbone	DINOv2 ViT−B/14, image size 518, frozen main weights
Pooling	CLS token + patch mean, 1536−dimensional representation
LoRA	last 4 blocks, targets: qkv/proj/fc1/fc2, r=8, alpha=16, dropout=0.05
Head	Linear−>256, ReLU, BatchNorm1d, Dropout(0.3), Linear−>1
Loss/Optimizer	Label smoothing BCE(0.05) + pos_weight; AdamW lr=1e−4, wd=1e−3, cosine schedule
Validation/Ensemble	Stratified 5−fold, 3 seeds, 15−model average, flip TTA

表 5 最终训练配置与实现细节

3.3 置信度分诊与多模态复核

LoRA 输出每张测试图属于陨石的概率 lora_prob。融合脚本首先读取 teammate_code/dinov2_lora_v2_detail.csv；若 lora_prob < 0.2，则直接判为非陨石；若 lora_prob > 0.85，则直接判为陨石。对概率处于 [0.20, 0.85] 的样本，调用 GPT−4o−mini 进行高细节图像复核，要求模型输出 JSON 格式的 confidence 和 reason。各概率区间的分诊规则与样本分流示意详见附录图 A2。

输入	DINOv2 特征	任务适配	输出
原始 RGB 图像	ViT−B/14 patch tokens + CLS	LoRA adapter + MLP head	lora_prob
边界样本图像	high−detail VLM visual input	结构化陨石/非陨石提示词	llm_confidence + reason

表 6 主干模型与 VLM 复核分支的信息流

3.4 融合推理脚本实现

新增训练/推理代码将 VLM 复核实现为可复现的流水线，而不是人工挑选样本。脚本配置包括 MODEL_NAME=gpt−4o−mini、MAX_WORKERS=8、temperature=0.1、max_tokens=500，以及 high−detail image_url 输入。Prompt 明确列出正向陨石证据（熔壳、气印、金属颗粒、球粒）和强负向证据（光滑水磨面、几何形、均匀颜色、贝壳状断口、晶洞、铜绿等），并要求只返回 JSON，降低解析不确定性。

融合规则保持逐样本固定，不使用 Top−N 后处理，也不假设测试集中正类比例。对被复核样本，若 llm_confidence >= 0.5，则判为正类；否则判为负类。对未进入 VLM 查询范围的样本，脚本保留 LoRA 置信度作为默认 confidence，并通过相同 make_decision 逻辑生成 final_label。最终输出包括 submission_fusion.csv 与 submission_fusion_detail.csv；后者保留 lora_prob、llm_confidence、llm_reason 与 final_label，便于追踪每个样本由谁决定。

阶段	关键操作	输出
1. LoRA 概率读取	读取 dinov2_lora_v2_detail.csv 中的 id 与 lora_prob	基础概率表
2. 分诊选择	筛选 0.20 <= lora_prob <= 0.85 的边界样本	VLM query list
3. VLM 复核	并发调用 GPT−4o−mini；解析 JSON	llm_fusion_results.csv

	confidence/reason；失败则回退 0.5	
4. 固定规则融合	lora<0.2–>0, lora>0.85–>1, otherwise confidence>=0.5–>1	final_label
5. 提交生成	按 sample_submission.csv 顺序 merge，并保存 id,label 两列	submission_fusion.csv

表 7 融合推理脚本的可复现流程

4. 结果展示与分析

模型性能随方法演进稳定提升。最初的 frozen DINOv2 原图 baseline 达到 F1=0.740；注入 LoRA 后提升到 0.795，这是最大的单步增益，说明任务自适应微调确实重塑了类别边界；分类头重构、正则化和验证设计进一步将底座推到 0.8148；最终加入置信度分诊与 VLM 复核后达到 0.84571。

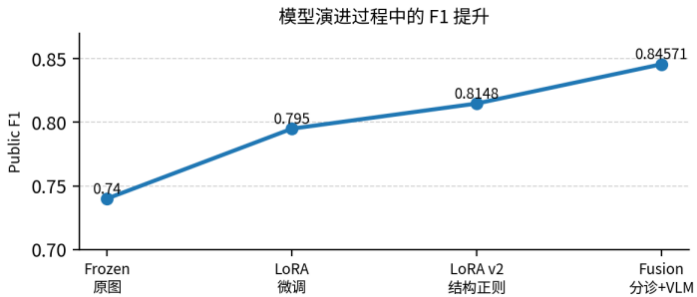

图1 从 frozen baseline 到最终融合系统的 F1 演进

设计决策	对照/现象	F1 或结论
是否裁剪背景	SAM 裁剪 vs 保留原图	约 0.70 –> 0.74，原图更优
是否 LoRA	frozen vs LoRA	0.740 –> 0.795，最大单步提升
分类头与正则化	Linear head vs MLP+BN+Dropout	0.795 –> 0.8148，减少过拟合
VLM 使用范围	全量复核 vs 只复核中间带	约 0.70 < 约 0.81，分诊优于全交 API
最终系统	LoRA 主干 + VLM 边界复核	Public F1 = 0.84571

表 8 关键优化与对照结果

从误差来源看，系统的优势在于将“确定性”和“不确定性”拆开处理。LoRA 负责多数高置信样本，使系统不会退化为外部 API 判别；VLM 只处理 63/194 张边界图片，即 32% 的测试样本，使通用多模态视觉常识被投放到最需要的位置。该分工也解释了为什么全量交给 VLM 的早期方案反而效果较差：通用模型虽然见过更广的视觉世界，但缺少本任务训练集的决策边界，容易在普通深色石头上过度联想。

4.1 其他探索与失败经验

除主线方法外，项目还尝试了两类方向。第一是标签传播：用 DINOv2 特征空间中的近邻关系把训练标签传导到无标签测试样本，但 public F1 仅约 0.69。失败原因是视觉近邻并不等于真实类别近邻，普通岩石和陨石在颜色与粗糙度上高度交织，伪标签传播会造成连锁误判。

第二是 INR/SIREN 纹理描述符。我们为每张图像训练微型隐式神经表示，用重构 loss 曲线、最高 PSNR 和激活权重统计来描述微观纹理复杂度。在 frozen 阶段，该方法曾把 F1 从 0.740 提到约 0.764；但在 LoRA v2 后，融合 INR 反而回落，说明 LoRA 已经吸收了主要纹理信息，额外几何复杂度特征变成冗余甚至噪声。

4.2 局限性与未来改进

当前系统仍有三个局限。第一，VLM 复核带来额外成本与延迟，且依赖外部服务可用性；提交仓库中应避免明文保存 API key，改用环境变量或本地配置文件。第二，分诊阈值 [0.20,0.85] 与判定阈值 0.5 仍带有经验成分，可在未来用校准集、

temperature scaling 或 conformal prediction 估计置信区间。第三，报告中缺少完整训练日志中的训练准确率和 loss 曲线，后续应把每折训练日志、OOF 预测和随机种子固定记录纳入 GitHub README。

未来改进可以沿三条路线推进：其一，使用更系统的难样本挖掘，针对熔渣、玄武岩、铁矿石等易混负类构建 hard negative set；其二，对 LoRA 输出做概率校准，使分诊阈值从经验规则变成统计规则；其三，把 VLM reason 转化为可审计特征，例如是否出现熔壳、气印、金属颗粒等结构化标签，再训练轻量二级分类器替代部分 API 调用。

5. 结论

本项目的关键经验是：在小样本与域偏移同时存在的图像任务中，单纯堆叠模型或盲目裁剪并不可靠。更稳健的策略是先建立强通用表征底座，再通过参数高效微调获得任务边界，最后用置信度分诊把高成本外部能力用于少数边界样本。最终系统在 194 张开放世界测试图像上实现 68% 样本由 LoRA 直判、32% 样本由 VLM 复核，public F1 达到 0.84571。

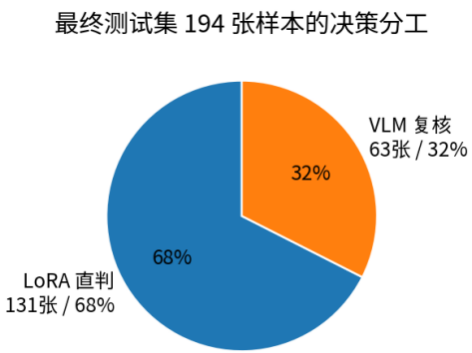

图 2 最终系统中 LoRA 与 VLM 的决策分工

代码、模型与复现链接

代码仓库：https://github.com/K1koli/meteorite-identification-G15

训练模型：链接:https://pan.baidu.com/s/1llby5Bi4owU3WZZVyoZg6A?pwd=g84j; 提取码: g84j。

参考文献

[1] Oquab, M. et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193, 2023.

[2] Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021.

[3] Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.

[4] Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. ICLR, 2019.

[5] OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024.

附录:

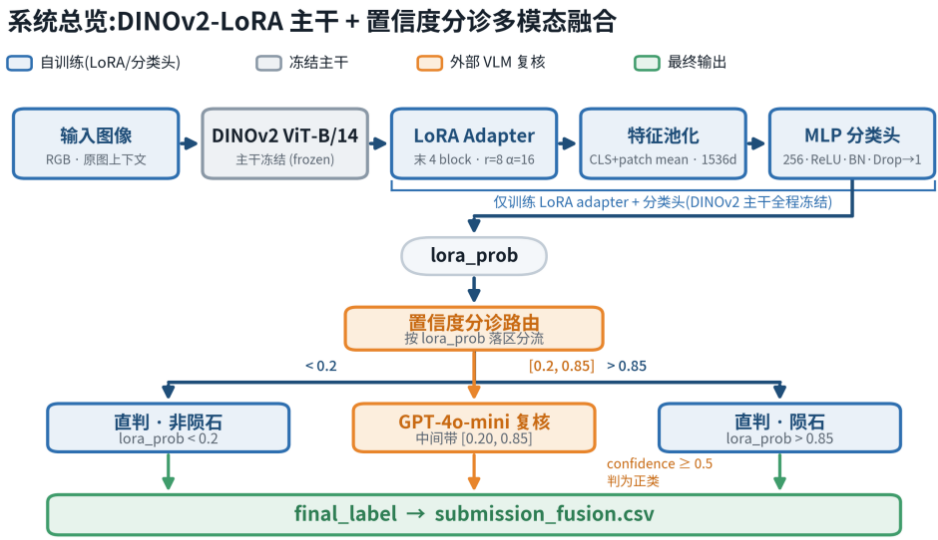

图 A1 系统总览:DINOv2–LoRA 主干 + 置信度分诊多模态融合的端到端结构

图 A2 置信度分诊机制:按 lora_prob 落区将样本分流至直判或 VLM 复核