Report

汇报组：

汇报日期：xxxx.xx.xx

1 组员及职责（Team member responsibilities）

1.1 小组成员

Xxx…

1.2 成员职责

Xxx…

2 背景（Background）

2.1 图像检索/位置识别（VPR）任务场景

图像检索/位置识别任务通常被称为**视觉地点识别**（Visual Place Recognition, VPR）：

给定一张查询图像，系统需要在一大规模、带地理坐标标注的图像库中，找到与其对应或相近的位置。

其基本要素可以概括为：

输入：一张由车辆、行人或无人机采集的环境图像；

数据库：覆盖城市或区域的大量参考图像，每张图带有位置信息；

输出：若干相似参考图像（检索结果），或对应的地理位置（分类或回归结果）；

任务目标：在给定的空间误差阈值内（如 25m）最大化正确定位的比例。

近年的工作例如最优传输 Optimal Transport 聚合的 VPR 描述子[1]、预训练大模型到 VPR 的适配方法[2]，以及基于 Game4Loc 的无人机地理定位基准[3]等，都在这个统一模板下，围绕更优的视觉特征表示和更鲁棒、高效的匹配/决策策略展开。

2.2 本课设采用的场景

现有视觉地点识别（VPR）方法大多被建模为图像检索问题：首先用深度网络提取全局或局部特征，再在预先编码好的数据库上进行 k 近邻搜索，典型代表包括基于 VLAD 聚合的 NetVLAD[4] 以及面向大规模城市场景的 CosPlace[5]。这类方法在中小规模数据集上表现优异，但在如 SF-XL 这类覆盖整座城市、包含数千万张图像的城市级场景中，检索的时间与存储成本随数据库规模线性增长，即便采用近似最近邻（ANN）结构也不可避免地在速度与精度之间做折中。

另一条研究路线是将 VPR 视为地理分类问题：如 PlaNet[6]、HGE[7]、CPlaNet[8]等工作将地球划分为一系列 geo-classes，并用卷积网络直接预测图像所属的地理单元，其中心坐标即为最终定位。这类方法在全球尺度、稀疏采样下具有良好的时间和空间复杂度，但其划分方案通常采用百公里级别的粗粒度单元，并假设图像分布高度不均匀，因此在需要米级精度、且图像在城市范围内密集采样的场景下，会面临类间视觉别名严重、分类精度不足的问题。

Divide&Classify（D&C）[9]正是在上述背景下提出：作者首次系统地将误差≤25m、覆盖面积>100km² 的城市级精细 VPR 明确建模为分类任务，指出直接套用全球尺度分类方法在城市级场景中存在明显局限，并针对城市路网的致密采样与强视觉混叠，设计了适用于城市级划分和推理流程的新方案。同时，D&C 还展示了如何将分类预测与检索方法结合，用分类结果限制检索的搜索空间，从而在保持或提升精度的同时显著降低推理时间。

本课设基于 Divide&Classify（D&C）框架，关注的是城市级、精细尺度的 VPR 场景，具体而言：

任务场景：

地图被划分为边长约 20m 的网格单元，每个网格视为一个“地点类别”；

查询图像被分类到某一网格，再用该网格中心坐标作为位置预测；

评价采用“预测坐标与真值距离在 25m 以内”的定位成功率。

3 本工作贡献（Contribution）

本课设在 Divide&Classify（D&C）[9]提出的城市级视觉地点识别框架基础上，引入了 DINOv2 预训练视觉模型[10]作为特征提取骨干网络，并结合长尾数据增强策略进行改进。整体创新点主要体现在以下几个方面：

1. 面向城市级精细 VPR 的预训练特征引入与对比分析

原始 D&C 使用卷积网络（如 Efficient Net）作为 backbone，需要在大规模城市数据集上从头训练，模型性能和训练稳定性在很大程度上依赖于充足的数据与训练资源。而 DINOv2 作为自监督预训练的视觉基础模型，在多种下游视觉任务中展现出更强的全局表征能力和跨场景鲁棒性[10]。本组在城市级 VPR 场景下，将 DINOv2 作为特征提取器嵌入 D&C 的分类框架，系统考察预训练视觉特征与城市级地点分类任务的适配性和优势，为后续将基础特征提取器模型引入精细定位任务提供一个具体基线。

2. D&C 框架与特征提取器的可泛化结构

本课设在保留 D&C 核心思想（城市网格划分、多组分类器、AAMC 判别）的前提下，完成了 DINOv2 特征到 AAMC 分类器之间的特征投影与归一化设计，形成一套适配良好的特征提取器+D&C 框架实现流程，有助于后续工作在此基础上进一步尝试更多预训练模型或多模态扩展。

3. 结合长尾数据增强缓解类别不平衡与提升泛化能力

针对城市级 VPR 数据集中存在的长尾类别分布，导致在训练过程中出现头

部类过拟合、尾部类表达不足的现象的问题，本课设设计了基于类别频次的随机强增强策略，验证了这一思路在不增加模型复杂度的前提下，可提升尾部类泛化能力，为后续引入更系统的长尾学习与不平衡学习方法提供了参考。

4 算法原理（Methodology）

本节介绍本课设采用的整体算法流程。整体上，我们遵循 Divide&Classify[9] 的城市级地点分类框架，并将原始卷积 backbone 替换为 DINOv2[10]。本节依次分为以下内容：问题建模与城市划分、网格分组与训练集构建、DINOv2 特征提取、AAMC 分类器设计与训练、推理与定位。

4.1 问题建模与城市网格划分

假设训练集记为：$T=\{(I_i,e_i,n_i)\}_{i=1}^{N}$，其中，$I_i$ 为第 i 张城市街景图像，(e_i,n_i) 分别为其在 UTM 坐标系下的东向（East）与北向（North）坐标。

为了将连续的地理位置转化为可分类的离散标签，本研究将整座城市的平面区域划分为边长为 M 的正方形网格单元（cell），例如：每个网格的东向与北向索引分别记为 (p,q)，对应的物理范围为

$$[pM,(p+1)M),\ \ [qM,(q+1)M) \tag{4.1}$$

对于每个训练样本 (I_i,e_i,n_i)，通过简单的取整操作将其分配到某个网格：

$$p_i=\left[\frac{e_i}{M}\right],q_i=\left[\frac{n_i}{M}\right] \tag{4.2}$$

对应的“地点类别”记为 C_{p_i,q_i}。这样，每个网格单元 $C_{p,q}$ 可以看作一个类别，网格中心坐标：

$$\text{Class2UTM}\left(C_{p,q}\right)=\left((p+0.5)M,(q+0.5)M\right) \tag{4.3}$$

在推理阶段将用于从类别预测恢复到地理坐标。

4.2 网格分组与多分类器训练集

在城市级、细粒度划分下，相邻网格的图像往往外观非常相似，若直接在所有网格上训练一个单一的大分类器，会导致严重的视觉混叠和标签冲突。为缓解

这一问题，我们采用 D&C 中的“网格分组（Groups）”策略。

设定一个整数 N，将所有网格按照坐标对 N 取模，划分为 N^2 个组（Group）：

$$G_{u,v}=\left\{C_{p,q}\mid p\xrightarrow{\bmod}N=u,q\xrightarrow{\bmod}N=v\right\} \tag{4.4}$$

其中 $u,v\in\{0,1\ldots,N-1\}$。

这种分组方式保证了同一 Group 内的网格在地理上互不相邻：相邻网格会被分配到不同的 Group 中。对每个 Group $G_{u,v}$，我们单独训练一个分类器，仅区分其内部包含的网格类别。这样可以显著降低单个分类器面对的“相似但标签不同”样本的比例。

对于 Group $G_{u,v}$，其训练集可写为：

$$T_{u,v}=\{(I_i,y_i)|\ C_{p_i,q_i}\in G_{u,v},y_i=\text{index}(C_{p_i,q_i}\ \text{in}\ G_{u,v})\} \tag{4.5}$$

其中 y_i 为该样本在 Group 内的类别索引。

4.3 DINOv2 特征提取骨干网络

4.3.1 DINOv2 的核心思想（自监督教师–学生自蒸馏）

DINOv2 延续了 DINO 系列方法中“teacher–student 自监督蒸馏”的设计思想：在无人工标签的数据上，通过教师网络为不同视图提供软目标，由学生网络进行拟合。在此基础上，DINOv2 在数据规模、训练稳定性和模型家族等方面进行了系统扩展，使其在多种下游任务上具有更强的迁移能力。总体流程为：

1. 对同一张原始图像 I 生成多种增强视图：

- 若干高分辨率视图 $\left\{I^{(g)}\right\}$
- 若干低分辨率局部视图 $\left\{I^{(l)}\right\}$（multi-crop 策略）。

2. 将这些视图分别送入教师网络和学生网络，两者结构相同（如 Vision Transformer），但参数更新方式不同：

- 学生网络参数通过标准反向传播更新；
- 教师网络参数为学生网络参数的指数滑动平均（EMA），即

$$\theta_{teacher}\leftarrow\tau\theta_{teacher}+(1-\tau)\theta_{student} \tag{4.6}$$

3. 教师对每个视图输出一个经过温度缩放、中心化与“sharpening”的概率分布：

$$q_{teacher}\left(y|\ I^{(t)}\right) \tag{4.7}$$

学生输出对应的分布

$$p_{student}\left(y|\ I^{(s)}\right) \tag{4.8}$$

训练目标是最小化两者之间的交叉熵/KL 散度，使学生在不同视图之间学习到视角和裁剪不变的语义表征。整体损失形式可写为：

$$L_{\text{DINO}} = \sum_{I^{(t)},I^{(s)}} CE\left(q_{teacher}\left(\cdot|\ I^{(t)}\right), p_{student}\left(\cdot|\ I^{(s)}\right)\right) \tag{4.9}$$

4. 通过合适的温度参数、输出中心化与多视图设计，DINOv2 避免了自监督训练中“塌缩到常数向量”的问题，使网络自发学到具有**语义聚类性质**的特征：同类物体/场景在特征空间中自然聚在一起，不同类相互分离。

与传统卷积 backbone 相比，DINOv2 具有以下特点：

- 采用 VisionTransformer 结构，更善于捕获全局上下文信息；
- 通过在大规模、干净的图像集合上自监督预训练，获得**跨数据集、跨任务**的迁移能力；
- 在分类、检测、分割及检索等多种下游任务上都有较强表现。

这些性质与 VPR 的需求高度契合：VPR 需要对视角变化、光照变化等保持鲁棒，同时又要保留区分不同地点的细粒度结构信息。

4.3.2 在 D&C 框架中的使用方式

在原始 D&C 中，作者采用卷积网络（如 EfficientNet）作为 backbone，从零开始在城市数据集上训练特征。本课设将 backbone 替换为 DINOv2[10]，以利用其在大规模无监督预训练中学到的、更鲁棒的全局视觉特征。

设 DINOv2 backbone 记为函数：

$$f_{\text{DINO}}: I \mapsto h \tag{4.10}$$

对于输入图像 I_i，我们首先将其按 DINOv2 标准进行预处理（缩放、中心裁剪、归一化），使用预训练的 DINOv2 模型进行特征提取，得到高维特征向量：

$$h_i = f_{\text{DINO}}(I_i) \in \mathbb{R}^D \tag{4.11}$$

在实现上，我们冻结 DINOv2 主干，仅训练投影层和分类头，以平衡训练成

本与性能。

4.4 加性角度间隔分类器（AAMC）

对于每个 Group $G_{u,v}$，我们构建一个独立的加性角度间隔分类器（AAMC）。设该 Group 内共有 $K_{u,v}$ 个类别，对应的类别原型向量为：

$$W^{(u,v)} = [w_1^{(u,v)}, w_2^{(u,v)}, \ldots, w_{K_{u,v}}^{(u,v)}] \in \mathbb{R}^{d \times K_{u,v}} \tag{4.12}$$

其中每个原型向量 $w_k^{(u,v)}$，在训练过程中同样约束为 L_2 归一化。

给定样本的嵌入特征 h_i，其与第 k 个类别原型的余弦相似度为：

$$\cos\theta_k = (w_k^{(u,v)})^\top h_i \tag{4.13}$$

在 AAMC 中，我们对真实类别 y_i 引入角度间隔 m 和缩放因子 s，定义 logits 为：

- 对真实类别 y_i：

$$l_{y_i} = s \cdot \cos(\theta_{y_i} + m) \tag{4.14}$$

- 对其他类别 $k \neq y_i$：

$$l_{y_i} = s \cdot \cos(\theta_k) \tag{4.15}$$

然后通过 Softmax 获得类别概率：

$$P(k|\ h_i) = \frac{\exp(l_k)}{\sum_{j=1}^{K_{u,v}} \exp(l_j)} \tag{4.16}$$

交叉熵损失为：

$$L_{\text{AAMC}} = -\frac{1}{|T_{u,v}|} \sum_{(I_i, y_i) \in T_{u,v}} \log P(y_i|\ h_i) \tag{4.17}$$

通过最小化 L_{AAMC}，我们同时优化投影层参数 W_{proj} 和 Group 内的类别原型 $W^{(u,v)}$，获得具有较大类间角度间隔的判别性特征。

4.5 长尾数据增强策略

在城市级 VPR 的网格划分下，训练集中各类别（网格）之间的样本量高度

不均衡：

部分区域由于采样密集或路网复杂，属于头部类（head classes），样本数量较多，模型容易充分学习；

大量区域属于尾部类（tail classes），样本数量有限，在标准训练过程中容易出现过拟合单一视角/表达不足的问题，导致测试时对这些区域的泛化能力较弱。

为缓解这一“长尾分布”问题，本课设在 DINOv2+D&C 框架上设计了一种基于类别频次的随机数据增强策略，核心思路是：

1. 统计每个类别c的训练样本数n_c，将其划分为“头部类”和“尾部类”；
2. 为每个类别设定一个增强概率$p_{\text{aug}}(c)$，使其与样本量呈反比：

样本量大的头部类n_c越大，$p_{\text{aug}}(c)$越小；

样本量少的尾部类n_c越小，$p_{\text{aug}}(c)$越大；

3. 在每个训练 step 中，对于来自类别c的图像I，以概率$p_{\text{aug}}(c)$对其施加强数据增强（strong augmentation）；
4. 如果该图像被增强，则原图和增强后的图像同时加入本次训练 batch。

该策略不改变 DINOv2 + AAMC 的整体网络结构，只是在数据采样与增强环节对“头部类 / 尾部类”区别对待，使得尾部类在特征空间中得到更多“视角”与“风格”的覆盖，从而提高模型对尾部类的泛化能力。

在此基础上，我们还额外观察到：大部分 VPR 数据集都存在训练集和测试集差距过大，导致测试集出现未见类（out-of-distribution, OOD）的情况，后续或可根据 OOD 学习方法，进一步提高 VPR 效果。

4.6 训练策略

4.6.1 训练阶段

实际训练时，我们对所有 Group 交替进行更新，以控制显存和训练时间。一个典型的训练轮次可以描述为：

1. 选择某个 Group $G_{u,v}$；
2. 从中 $T_{u,v}$ 采样一个批次图像；
3. 对批次中的每个样本，以概率$p_{\text{aug}}(c)$对其施加强数据增强；
4. 扩展后的 batch 中所有图像，计算 DINOv2 特征特征 h_i；
5. 通过对应的 AAMC 分类头计算 logits 与损失 L_{AAMC}；
6. 仅更新 DINOv2 中最后若干层、投影层 W_{proj} 以及该 Group 的原型 $W^{(u,v)}$；
7. 轮换到下一个 Group，重复上述过程。

这样可以在不显著增加模型参数的情况下，分别学习各 Group 内的判别边界，同时共享 DINOv2 和 backbone 的全局表征能力。

4.6.2 推理与定位阶段

在推理阶段，对于一张查询图像 I_q，整体流程如下：

1. 提取特征：

$$h_q = \mathrm{Norm}\left(W_{\mathrm{proj}} f_{\mathrm{DINO}}\left(I_q\right)\right) \tag{4.18}$$

2. 对所有 Group 的 AAMC 分类器分别计算 logits 和概率分布。对于 Group $G_{u,v}$，其输出为：

$$P^{(u,v)}\left(k \mid h_q\right), k = 1, \ldots, K_{u,v} \tag{4.19}$$

3. 在所有 Group、所有类别中寻找全局概率最大的类别：

$$\left(u^*, v^*, k^*\right) = \mathrm{argmax}_{u,v,k} P^{(u,v)}\left(k \mid h_q\right) \tag{4.20}$$

4. 将该类别映射到对应的网格 C_{p^*,q^*}，再通过 Class2UTM 函数得到预测坐标：

$$x = \mathrm{Class2UTM} C_{p^*,q^*} \tag{4.21}$$

5 数据集（Datasets）

5.1 SF-XL 数据集

本课设采用 San Francisco eXtra Large（SF-XL）数据集[5]作为主要实验平台。该数据集专门面向城市级、精细尺度的视觉地点识别（VPR）场景，具有以下特点：

- 覆盖范围大：数据覆盖旧金山城区，地图面积约 170km²；
- 采样密集：训练集中约有 4,120 万张带地理坐标标注的图像，沿道路网络大致以“每米一张图”的密度采集，具备米级定位任务所需的空间分辨率；
- 测试集真实：完整测试集包含两个智能手机采集的测试集，分别记为 testv1 和 testv2，模拟实际用户用手机拍摄的查询图像，具有视角、成像

条件、设备类型多样等特点。

考虑到 SF-XL 测试集规模较大、完整评测开销较高，在本课设的实际实验中，我们从原始测试集中随机采样了约 1/5 的数据作为评测子集，在保持数据分布大致一致的前提下，降低了单次实验的计算成本。

6 实验（Experiments）

6.1 地图划分与标签构建

首先将连续的 GPS/UTM 坐标离散化为网格单元。具体做法如下：

以 M=10m 为边长，将城市区域划分为规则的正方形网格（cell），每个网格对应一个“地点类别”。为缓解网格边界处的视觉混叠问题，进一步按照坐标对 N=2 取模，将所有网格划分为4个 Groups，每个 Group 中的网格在地理上互不相邻；

6.2 实验设置

本节实验基于 SF-XL 数据集展开，在 1/5 测试子集上进行对比。

本研究主要比较三种方法：

1. 原始 D&C：使用 EfficientNet 作为骨干网络，仅采用标准的数据增强和训练策略；
2. DINOv2+D&C：在保持 D&C 整体框架不变的前提下，用 DINOv2 预训练模型替换原始 backbone；
3. DINOv2+D&C+长尾增强：在第 2 种方法基础上，引入基于类别频次的强数据增强策略，对尾部类进行更积极的数据扩展。

为统一评价标准，三种方法均采用 LocalizationRadius@N（LR@N）指标，LR@N 的定义为：若前 N 个预测中至少有一个位置与真值距离不超过 25m，则认为该查询被成功定位。本研究选用 LR@1、LR@5、LR@10、LR@20 作为定量指标。

6.3 定量结果

三种方法的 LR@N 指标如下表所示：

表格 1 不同方法的 LR@N 指标

方法	LR@1	LR@5	LR@10	LR@20
D&C	36.6	47.8	51.7	55.3
DINOv2+D&C	45.4	60.6	64.4	67.0
DINOv2+D&C+长尾增强	45.6	60.8	64.6	67.1

结果表明，与原始 D&C（Efficient Net 骨干）相比，引入 DINOv2 后，模型在四个指标上均有明显提升，这说明在相同的 D&C 城市级分类框架下，预训练的 DINOv2 特征显著优于从头训练的卷积骨干，尤其在更高的 Recall（LR@5/10/20）下，能够为每个查询提供更丰富的“候选位置”，对后续可能的检索或多候选融合都更加有利。

在 DINOv2 基础上引入长尾增强后，指标有小幅但一致的提升。尽管提升幅度不大，但所有指标均有正向变化，说明基于样本量的长尾增强在当前设置下能够在不增加模型复杂度的前提下，进一步改善模型对尾部类及少样本区域的泛化性能。这种改进更可能体现在“减少灾难性错误”和“提升中等难度样本的召回率”上，而不是在极端高频类别上拉高分数。

7 结论（Conclusion）

本课设围绕城市级精细视觉地点识别（VPR）任务，选取 Divide&Classify 作为基础框架，将 DINOv2 自监督预训练视觉模型引入城市网格化地点分类，并在此基础上探索面向长尾类别分布的数据增强策略。整体工作从问题背景出发，对城市级 VPR 的任务场景和现有方法进行了梳理，随后在方法层面对 D&C 的网格划分、Group 分组、AAMC 分类器以及 DINOv2 特征提取流程进行了统一建模与实现。

实验结果表明，在相同的 D&C 框架下，用 DINOv2 替换原始卷积 backbone 后，模型在 LR@1–LR@20 四个指标上均获得了显著提升：相较于原始 D&C，LR@1 从 36.6 提升到 45.4，LR@5/10/20 也分别有较大提升，说明预训练视觉特征在城市级精细定位任务中相较“从头训练”的卷积特征具有明显的优势。在此基础上，我们针对长尾类别设计了基于样本频次的强增强策略，对尾部类进行更积极的数据扩展。尽管提升幅度较小（LR@1–LR@20 均提升约 0.1–0.2 个百分点），但整体趋势稳定向好，表明该策略在不改变模型结构的前提下，能够一定程度上改善尾部类的泛化表现。

未来工作可以在本课设的基础上，引入未见类学习等方法，解决 OOD 问题。

参考文献

[1] S. Izquierdo and J. Civera, “Optimal Transport Aggregation for Visual Place Recognition,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 17658–17668.

[2] F. Lu et al., “Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition,” arXiv preprint arXiv:2402.14505, 2024.

[3] Y. Ji, B. He, Z. Tan, and L. Wu, “Game4Loc: A UAV Geo-Localization Benchmark from Game Data,” in Proc. AAAI Conf. Artificial Intelligence, vol. 39, 2025, pp. 3913–3921.

[4] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1437–1451, 2018.

[5] G. Berton, C. Masone, and B. Caputo, “Rethinking Visual Geo-Localization for Large-Scale Applications,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4878–4888.

[6] T. Weyand, I. Kostrikov, and J. Philbin, “PlaNet: Photo Geolocation with Convolutional Neural Networks,” in Proc. European Conf. Computer Vision (ECCV), 2016.

[7] E. Müller-Budack, K. Pustu-Iren, and R. Ewerth, “Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification,” in Proc. European Conf. Computer Vision (ECCV), 2018.

[8] P. H. Seo, T. Weyand, J. Sim, and B. Han, “CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps,” in Proc. European Conf. Computer Vision (ECCV), 2018.

[9] G. Trivigno, G. Berton, J. Aragon, B. Caputo, and C. Masone, “Divide&Classify: Fine-Grained Classification for City-Wide Visual Place Recognition,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023.

[10] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” arXiv preprint arXiv:2304.07193, 2023..