

---

# 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<sup>[1]</sup>、预训练大模型到 VPR 的适配方法<sup>[2]</sup>, 以及基于 Game4Loc 的无人机地理定位基准<sup>[3]</sup>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 $\leq 25\text{m}$ 、覆盖面积 $>100\text{km}^2$ 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，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 EfficientNet）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. 解决训练成本较高、泛化能力有限的问题

相比于完全提出新算法，本课设更侧重于在经典框架基础上的问题驱动型优

化：针对城市级 VPR 中“从头训练卷积 backbone 成本较高、泛化有限”的问题，引入 DINOv2 以提升特征质量和训练效率；同时保持 D&C 在城市划分和多分类器设计上的优点。

## 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

$$[pM, (p+1)M), [qM, (q+1)M) \quad (4.1)$$

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：

$$p_i = \left\lfloor \frac{e_i}{M} \right\rfloor, q_i = \left\lfloor \frac{n_i}{M} \right\rfloor \quad (4.2)$$

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：

$$Class2UTM(C_{p, q}) = ((p+0.5)M, (q+0.5)M) \quad (4.3)$$

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，相邻网络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)：

$$G_{u,v} = \{C_{p,q} \mid p \xrightarrow{\text{mod}} N = u, q \xrightarrow{\text{mod}} N = v\} \quad (4.4)$$

其中  $u, v \in \{0, 1, \dots, N-1\}$ 。

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：

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

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  $\{I^{(s)}\}$
- 若干低分辨率局部视图  $\{I^{(l)}\}$ （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

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

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：

$$q_{teacher}(y|I^{(t)}) \quad (4.7)$$

学生输出对应的分布

$$p_{student}(y|I^{(s)}) \quad (4.8)$$

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：

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

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：

$$f_{DINO}: I \mapsto h \quad (4.10)$$

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：

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

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：

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

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：

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

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  $y_i$ ：

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

- 对其他类别  $k \neq y_i$ ：

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

然后通过 Softmax 获得类别概率：

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

交叉熵损失为：

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

通过最小化  $L_{\text{AAMC}}$ ，我们同时优化投影层参数  $W_{\text{proj}}$  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 训练策略

### 4.5.1 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，计算 DINOv2 特征特征  $h_i$ ；
3. 通过对应的 AAMC 分类头计算 logits 与损失  $L_{\text{AAMC}}$ ；
4. 仅更新 DINOv2 中最后若干层、投影层  $W_{\text{proj}}$  以及该 Group 的原型  $W^{(u,v)}$ ；
5. 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。

### 4.5.2 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：

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

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：

$$P^{(u,v)}(k | h_q), k = 1, \dots, K_{u,v} \quad (4.19)$$

3. 在所有 Group、所有类别中寻找全局概率最大的类别：

$$(u^*, v^*, k^*) = \arg\max_{u,v,k} P^{(u,v)}(k | h_q) \quad (4.20)$$

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：

$$x = \text{Class2UTM} C_{p,q}^{*,*} \quad (4.21)$$



---

---

## 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.
- [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.