arXiv:2602.01673v1 [cs.CV] 2 Feb 2026

基于NetVLAD与Faiss的视觉SLAM实时回环检测

Enguang Fan伊利诺伊大学厄巴纳-香槟分校计算机科学系美国伊利诺伊州厄巴纳 enguang2@illinois.edu

摘要——**回环检测（LCD）是同步定位与建图（SLAM）的核心组成部分：它识别再次访问的地点，并提供能够校正累积漂移的位姿图约束。DBoW等经典词袋方法效率较高，但在外观变化和感知混淆条件下往往性能下降。与此同时，基于深度学习的视觉地点识别（VPR）描述子（例如NetVLAD和基于Transformer的模型）具有更强的鲁棒性，但其计算成本常被视为实时同步定位与建图的障碍。本文实证评估了将NetVLAD用作回环检测模块的效果，并在KITTI数据集上将其与DBoW进行比较。我们提出了细粒度Top-K精确率–召回率曲线，以更好地反映查询可能没有匹配或存在多个有效匹配的回环检测设定。借助Faiss加速的最近邻搜索，NetVLAD在达到实时查询速度的同时，相较DBoW提升了准确率与鲁棒性，从而成为同步定位与建图中回环检测切实可行的直接替代方案。**

索引词—**回环检测、同步定位与建图、视觉地点识别、NetVLAD、Faiss、KITTI、精确率–召回率。**

I. 引言

回环检测（LCD） [1], [2] 是同步定位与建图（SLAM）的基本组成部分，使机器人能够识别先前访问过的地点，并通过消除位姿估计中的累积漂移来校正其估计轨迹。通过识别重访地点，回环检测可显著提高定位精度，并确保SLAM长期保持一致，尤其是在单目SLAM中，单个相机无法直接恢复深度，因此会导致尺度漂移。图1展示了未使用回环检测模块时ORB-SLAM2 [3] 在KITTI [4] 序列 00上的估计轨迹，其中累积漂移导致轨迹大幅偏离真实路径。相比之下，图2展示了应用回环检测后的校正轨迹，说明回环闭合如何约束误差并维持长期定位精度。

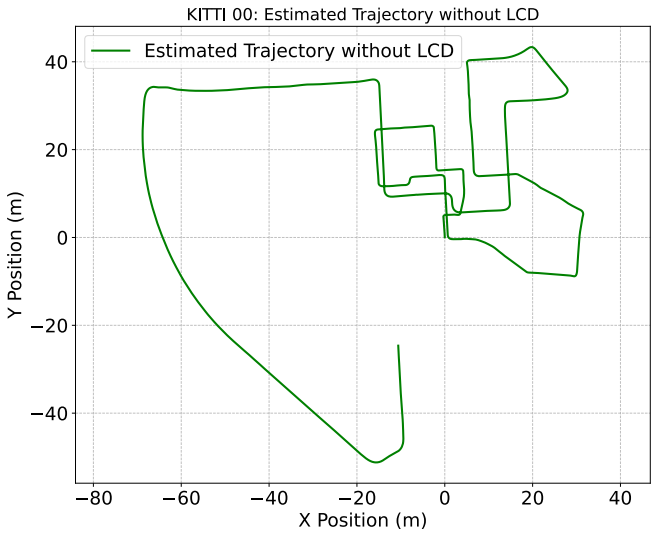

图1. 未进行回环检测时，KITTI序列00的单目SLAM估计轨迹。由于缺乏回环闭合约束，估计轨迹会产生累积漂移，进而导致定位误差。由于这是一个单目同步定位与建图系统，轨迹仅具有未知尺度，坐标轴上的绝对米制数值并不准确。

传统回环检测方法，例如动态词袋（DBoW） [5], 依赖手工设计的局部特征（例如 ORB [6],SIFT [7]），并通过快速特征提取和倒排索引搜索结构实现高效率。一个实际优势是，SLAM 流水线通常可以复用已为视觉里程计（帧间匹配）计算好的特征，因此回环检测几乎不增加额外的特征计算开销。剩余开销是倒排索引查找，它通常仅将查询与一部分帧进行比较这些帧共享相同的视觉词。然而，这些方法在动态或具有挑战性的环境中，以及在极端外观变化（如昼夜转换或季节变化）下，鲁棒性有限。对缺乏空间关系的局部特征描述子的依赖，也常常因感知混淆而导致错误回环，尤其是在重复性的结构化环境（如城市街道）中。图6和图7展示了一个假阳性示例，其中 DBoW 将属于不同位置的两帧检索为回环。

相比之下，基于深度学习的视觉地点识别（VPR）方法，例如 NetVLAD [8] 和基于Transformer的模型（如 DINO v2 [9], ViT-base [10]），利用从大规模图像数据集中学习到的全局特征嵌入，展现出更高的准确率。这些模型提取的地点描述子更具判别性，能够隐式捕捉几何关系，因此在具有挑战性的场景中表现出色。然而，基于深度学习的方法的主要缺点是计算成本较高，

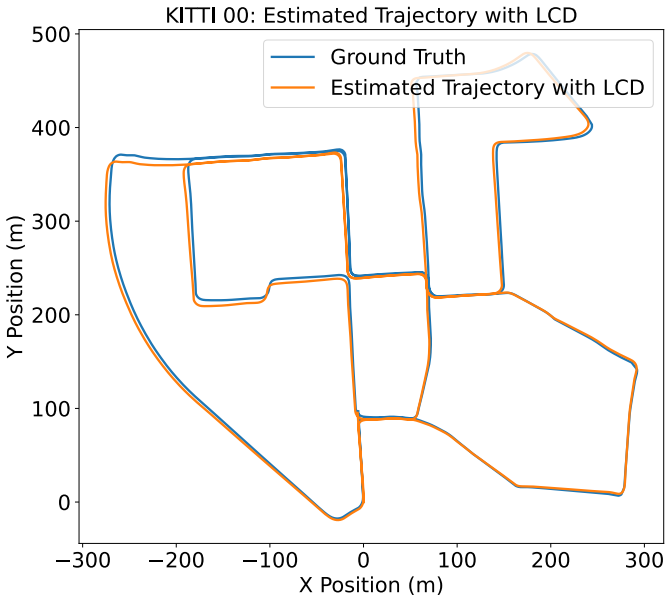

图 2. KITTI序列00中采用回环检测的单目同步定位与建图估计轨迹和真实轨迹。由于这是一个单目同步定位与建图系统，估计轨迹仅确定到一个未知尺度，并经过缩放以尽可能匹配真实轨迹。

因为它们需要大型神经网络，并且必须针对高维嵌入向量进行穷尽式最近邻搜索，因此不适用于实时 SLAM 应用。

此外，尽管视觉地点识别（VPR）与回环检测密切相关，但两者在匹配的定义和处理方式上存在根本差异。VPR 通常假设查询与数据库图像之间是一对一映射 [11]，，因此必须保证高精确率，以避免错误的地点识别。相比之下，回环检测在同步定位与建图框架中运行，在该框架下：

- 如果查询帧属于未曾重访的轨迹，则可能没有匹配项。
- 如果查询帧有多个候选项满足相对位姿差约束，则可能存在多个有效匹配项。
- 回环检测系统可以通过时间和几何一致性检查来剔除错误匹配，因此能够容忍假阳性。

尽管存在这些差异，大多数基于深度学习的VPR方法并非专为回环检测设计，而是假设数据库是静态的，并优先考虑高精确率的单图像检索。本文对将 NetVLAD（一种具有代表性的深度VPR描述子）应用于回环检测进行了系统的实证研究，评估其在同步定位与建图场景下的准确性和查询效率。本文通过实证评估 NetVLAD 作为 DBoW 回环检测替代方案的有效性。具体而言，我们旨在：

- **比较准确率：** 比较 NetVLAD 和 DBoW 在回环检测中的表现，重点关注 NetVLAD 在外观变化和动态场景中的鲁棒性。
- **评估效率：** 使用 Faiss 加速的最近邻搜索来测量查询效率，以评估实时可行性。
- **使用回环检测专用指标进行评估：** 使用针对回环检测的指标对这两种方法进行基准测试，包括我们提出的细粒度Top-K精确率–召回率曲线。

II. 相关工作

回环检测（LCD）在同步定位与建图（SLAM）中发挥着关键作用，它使机器人能够识别先前访问过的地点并校正位姿估计漂移。传统的回环检测方法主要依赖手工设计的视觉特征描述子，而近年来，基于深度学习的视觉地点识别（VPR）方法取得了进展，带来了更鲁棒、精度更高的替代方案。然而，这些深度学习模型在实时回环检测中的适用性尚未得到广泛研究。本节将回顾传统回环检测方法和现代基于深度学习的视觉地点识别技术。

A. 传统回环检测方法

经典的回环检测方法主要依赖手工设计的视觉特征，例如 ORB [6]，SIFT [7]，和 SURF [12]，，以检测重访地点。这些特征提取局部关键点及其描述子，并将其用于图像匹配。

一种广泛采用的方法是动态词袋（DBoW） [5]，，它通过构建层次词汇树来实现高效图像检索，从而扩展了传统词袋（BoW）范式。DBoW 通过以下方式实现回环检测：从图像中提取局部描述符，并将其与预训练的视觉词关联；将每幅图像表示为视觉词直方图，从而支持基于倒排索引的快速搜索；复用视觉里程计中的特征描述符，以降低计算开销。DBoW 凭借紧凑的表示形式和快速的搜索结构，具有很高的效率。然而，它也存在若干局限：在动态环境中（例如昼夜变化、季节变化）泛化能力有限；容易出现感知混淆，即视觉上相似但实际不同的地点产生错误回环，尤其是在城市场景等结构化环境中；依赖词汇表，需要使用离线训练的特征词汇表，而该词汇表可能无法泛化到未见过的环境中。其他传统回环检测方法，例如 FAB-MAP[13]，，引入概率模型以提高地点识别的可靠性，但仍依赖人工设计的描述符，因此容易受到外观变化和场景动态变化的影响。

B. 基于深度学习的视觉地点识别方法

深度学习的最新进展催生了视觉地点识别（VPR）方法，这些方法在具有挑战性的条件下优于传统的手工设计方法。与依赖人工设计特征的经典方法不同，基于深度学习的 VPR 模型能够从大规模图像数据集中学习具有判别性的特征嵌入。

最具代表性的基于深度学习的 VPR 方法之一是 NetVLAD [8], , 它通过基于 VLAD 的 [14], [15] 池化层增强 CNN 骨干网络，从而生成用于地点识别的全局描述符。与基于 BoW 的方法不同，NetVLAD：

- 直接从数据中学习特征表示，无需预定义词汇表。
- 捕捉特征之间的全局空间关系，减少感知混淆。
- 比基于手工设计描述子的 方法准确率更高。

其他值得关注的基于深度学习的视觉地点识别方法包括：

- Patch-NetVLAD [16], 采用局部描述子，以提升视角鲁棒性。
- 自监督学习方法，例如 CALC [2],, 通过训练自编码器来重建基于 HoG 的描述子。
- 视觉Transformer（ViT） [10]及其变体，例如 DINO [17] 和 DINOv2 [9], , 利用自注意力机制进行特征学习，并取得了最先进的成果。

尽管这些深度学习方法在视觉地点识别任务中表现出色，但它们并非专门为回环检测而设计。仍有若干挑战：

- 计算成本：ViT 和 NetVLAD 等方法会生成高维描述子，需要进行大量最近邻搜索，因此若不进行加速，实时回环检测将难以实现。
- 检索标准不匹配：视觉地点识别通常优先考虑高精确率（每个查询对应一个正确匹配），而回环检测允许根据轨迹约束存在多个正确匹配，因此需要进一步的后处理步骤。

III. 方法

本节介绍我们如何将视觉地点识别（VPR）技术应用于同步定位与建图（SLAM）中的回环检测问题。我们采用基于深度学习的 NetVLAD 方案，取代传统的基于动态词袋（DBoW）的回环检测模块。与在光照变化、动态物体等环境变化下表现不佳的手工特征方法不同，我们的方法利用学习得到的全局描述符，在保持实时可行性的同时提高鲁棒性和准确性。图 3 展示了 NetVLAD 在 KITTI [4] 序列 00 上生成的帧间相似度分数。该热力图表示各帧描述子之间的欧几里得距离，颜色较深的区域表示距离较小（相似度较高）。此外，图 4 展示了回环的真实聚类结果，说明所定义的位姿约束能够有效识别重访地点。我们首先介绍 SLAM 流水线，重点说明 NetVLAD 如何集成为回环检测模块。接下来，我们将正式定义用于评估的数学框架

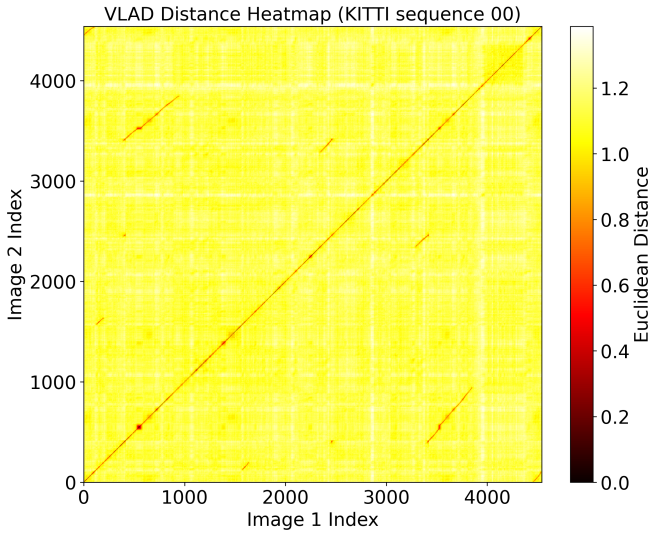

图 3. 基于 NetVLAD 欧几里得距离的 KITTI序列00相似度热力图。颜色较深的区域表示帧之间的距离较小（相似度较高）。对角线表示自身与自身及相邻帧之间的匹配，这些匹配的相似度通常较高，在回环检测中应予以忽略。

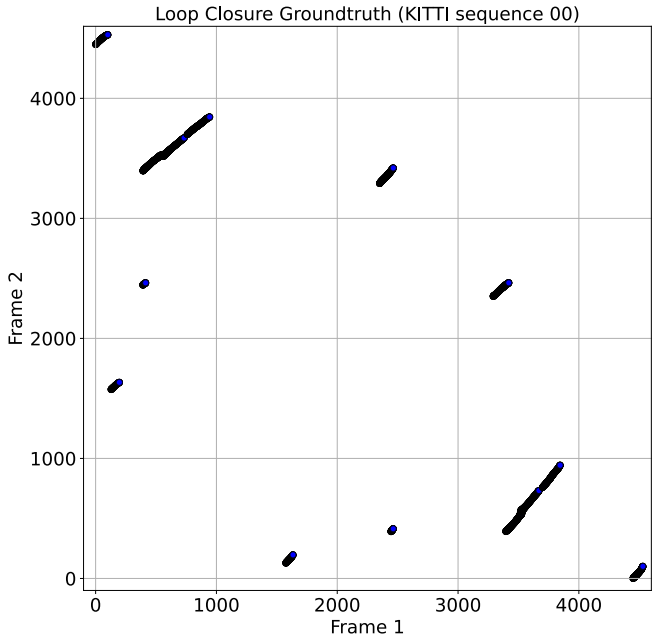

图 4. KITTI序列00的位姿真值聚类结果。相对位姿差异较小的帧自然形成簇。

回环检测性能。最后，我们提出一种细粒度 Top-K 精确率–召回率曲线，以反映 VPR 与回环检测评估之间的关键差异。我们的方法旨在弥合准确但计算开销大的深度学习方法与高效但鲁棒性较差的手工设计方法之间的差距，使回环检测在真实世界同步定位与建图应用中既可扩展又可靠。

A. 带有基于 *NetVLAD* 的回环闭合检测的 *SLAM* 流水线

通过将 NetVLAD 集成为回环检测模块，我们的 SLAM 流水线对环境变化实现了更强的鲁棒性，同时借助基于 Faiss 的近似最近邻搜索进行优化检索，保持了实时可行性 [18]。

图 5 展示了我们采用基于 NetVLAD 的回环检测的 SLAM 系统。该流程由以下关键模块组成：

- **视觉里程计（VO）**：提取连续帧之间的局部特征对应关系，以估计增量运动。
- **关键帧选择**：动态选取一部分帧，以避免冗余计算。
- **NetVLAD 编码器**：每个关键帧均由 NetVLAD 处理，生成紧凑的高维嵌入作为其全局描述子。
- **嵌入数据库**：提取的描述子存储在数据库中，以便在回环检测过程中进行高效检索。
- **基于 Faiss 的候选检索**：使用基于 NetVLAD 的全局描述符将当前关键帧与过去的关键帧进行匹配，并通过 Faiss 检索排名靠前的候选帧。
- **时间一致性检查**：我们不将回环闭合视为单帧匹配，而是采用序列到序列的匹配策略。有效的回环闭合应由多个连续关键帧构成的一致组提供支持，而非依赖孤立的帧间匹配。这有助于减少由感知混淆导致的假阳性，并确保系统能够稳健应对视点变化。
- **位姿图优化**：通过应用位姿图优化，利用有效的回环闭合校正估计轨迹中的累积漂移。

与 DBoW 相比，我们基于 NetVLAD 的回环检测具有以下优势：

- **更高的鲁棒性**：NetVLAD 学习到的特征嵌入对视点、光照和季节变化具有不变性。
- **更高的准确性**：与依赖基于直方图的视觉词匹配的 DBoW 不同，NetVLAD 在其学习到的表示中保留了空间关系。
- **高效检索**：通过基于 Faiss 的近邻搜索，我们实现了与 DBoW 相当的查询速度，同时显著提高了准确率。

通过利用基于深度学习的描述符，我们的方法弥合了传统手工设计方法与计算开销巨大的基于Transformer的模型之间的差距，使其成为同步定位与建图中回环检测的可行方案。所提出的流程兼顾速度与准确率，适用于实际机器人应用。

B. 数学定义

我们定义以下术语，以便对回环检测问题进行定量分析：

- **查询帧** q**：** 由相机捕获、用于搜索匹配的先前访问帧的当前帧。
- **查询检索集** $R(q)$**：** 返回的查询结果是一个集合，其中每个元素的格式为 [帧 ID，分数]:

$$R(q) = \{(r_1, s_1), (r_2, s_2), \ldots, (r_k, s_k)\} \tag{1}$$

  – 对于 DBoW，分数基于加权视觉词相似度度量。

  – 对于基于深度学习的方法（例如 NetVLAD、ViT），我们将匹配分数定义为相似度值：余弦相似度，或查询帧与匹配帧的全局描述符之间的负欧几里得距离（因此，分数越大表示匹配越接近）。

  – 为确保时间一致性，我们排除时间上相邻的帧，不将其视为有效的回环闭合，因为它们无法提供用于校正漂移的额外信息。这也是许多 SLAM 系统中的常见做法 [1], [3], [19]。
- **真实匹配集** $G(q)$**：** 定义为一组帧，这些帧是给定查询的有效回环闭合匹配项。

$$G(q) = \{g_1, g_2, \ldots, g_m\} \tag{2}$$

其中：

– $G(q) = \emptyset$若查询帧不存在回环闭合。– $|G(q)| = 1$ 若存在唯一的重访位置。– $|G(q)| > 1$ 若多个帧满足相对位姿差异阈值。

C. 真正例、假阳性、假负例和真负例

令 θ 为匹配分数阈值。若检索帧 r_i满足 $s_i \geq \theta$，则接受该帧；否则拒绝。根据上述定义，我们将检索结果分类如下：

- **真正例 (TP)**：检索到的帧中属于真实值集合且通过阈值的帧：

$$TP(q) = R(q) \cap G(q) \cap \{r \mid s_r \geq \theta\} \tag{3}$$

- **假阳性 (FP)**：检索到的帧中不在真实值集合内但分数超过阈值的帧：

$$FP(q) = R(q) \setminus G(q) \cap \{r \mid s_r \geq \theta\} \tag{4}$$

- **假负例 (FN)**：未被检索到或分数低于阈值的真实值帧：

$$FN(q) = G(q) \setminus R(q) \cup G(q) \cap \{g \mid s_g < \theta\} \tag{5}$$

- **真负例 (TN)**：检索到的帧不属于真实值集合，且被正确拒绝：

$$TN(q) = R(q) \setminus G(q) \cap \{r \mid s_r < \theta\} \tag{6}$$

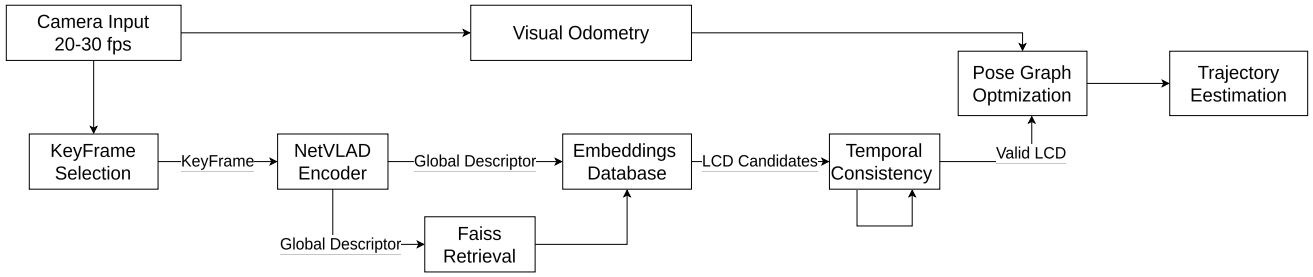

图 5. 基于 NetVLAD 的回环检测 SLAM 流水线。基于 NetVLAD 的回环检测模块取代了传统的基于 DBoW 的方法，提供了更稳健、更准确的替代方案，同时通过基于 Faiss 的检索保持实时可行性。

D. 细粒度 *Top-K* 精确率-召回率曲线

现有 VPR 基准使用基于**Top-1 检索**的精确率-召回率 (PR) 曲线，并假设每个查询始终恰好有一个正确匹配 [11]。然而，对于回环检测而言，这一假设并不成立，因为：

1) 一个查询可能有多个有效的真实匹配，因为在特定相对位姿阈值范围内的任何历史帧都可以视为正确匹配。2) 如果查询不属于重访轨迹，则可能完全没有真实匹配。3) 与 VPR 不同，假阳性可通过时间一致性和几何一致性进行过滤；在 VPR 中，假阳性会直接影响准确率。

为了更好地评估回环检测性能，我们提出细粒度 Top-K 精确率–召回率曲线，它通过以下方式扩展了标准 PR 曲线：

- 逐个考察检索到的候选项，而不是像 Recall@N 那样，只要检索集包含至少一个正确匹配，就将整个检索集视为正例。
- 评估**Top-K 检索**结果，而非仅看 Top-1 匹配，以充分利用时间和几何一致性筛选。

Top-K 检索的定义：

给定查询帧 q，检索集 $R(q)$由按匹配分数排序的前 K 个候选项组成：

$$R_K(q) = \{(r_1, s_1), (r_2, s_2), \ldots, (r_K, s_K)\} \tag{7}$$

我们根据 Top-K 检索结果定义精确率和召回率：

$$\text{Precision@K} = \frac{|TP_K(Q)|}{|TP_K(Q)| + |FP_K(Q)|} \tag{8}$$

$$\text{Recall@K} = \frac{|TP_K(Q)|}{|TP_K(Q)| + |FN_K(Q)|} \tag{9}$$

$TP_K(Q)$、 $FP_K(Q)$和 $FN_K(Q)$分别表示前 K 个检索候选项中的真正例、假阳性和假阴性集合，而 Q 是所有查询帧的集合。细粒度Top-K精确率–召回率曲线通过遍历阈值 θ 的所有可能取值并绘制精确率与召回率的关系得出，与标准PR曲线相同。

IV. 评估

像 DBoW 这样的传统回环检测方法面临的一大挑战是感知混淆，即视觉上相似但实际不同的地点被错误地匹配为回环闭合。图6和图7展示了一个假阳性示例：尽管两帧属于完全不同的地点，却被检索为回环闭合。这凸显了手工设计的视觉特征的一项关键局限：它们难以应对城市街道等结构化环境，因为重复特征可能导致假阳性。为系统地评估我们的方法，我们在一台配备 AMD Ryzen 7700X 中央处理器和 Nvidia RTX 4070 Ti GPU 的系统上进行实验。



图 6. KITTI 序列 00 中的第 4537 帧



图 7. KITTI 序列 00 中的第 667 帧

A. 实验设置与指标

1) 数据集：*KITTI*：KITTI [4] 数据集是评估视觉里程计和 SLAM 系统的常用基准。它提供城市驾驶环境中的相机流和真值位姿（通过高精度 GPS/IMU 获取）。在我们的实验中，我们使用频率为 20 Hz 的左侧单目灰度相机流，并预先计算闭环真值。如果两个帧的相对位姿满足以下条件，则认为它们之间的回环闭合有效：

- 平移差小于 1.5 米。
- 旋转差异小于 0.3 弧度。

为确保回环闭合具有意义，我们不将连续帧视为有效匹配，因为它们自然会呈现相似性，但不会为位姿图优化提供额外约束。为此，我们采用简单的时间一致性检查，排除最近的 100 帧。

B. 精度基准测试结果

我们分析了 DBoW 和 NetVLAD 在不同 **Top-K** 值 $\{1, 5, 10, 25\}$ 下的细粒度 Top-K 精确率-召回率（PR）曲线。图 8–11 展示了当我们调整回环闭合的前 K 个候选项数量时，每种方法的精确率和召回率如何变化。

- **Top-1：** 在这一严格设置下，两种方法都能达到相对较高的精确率，但召回率非常低。尽管得分最高的匹配通常是正确的，但许多真实回环闭合会被遗漏。由于具有更具判别性的全局描述符，NetVLAD 往往比 DBoW 保持更高的精确率。
- **Top-5 和 Top-10：** 随着**K** 增大，召回率会提高，因为更多候选项会被视为有效匹配。不过，精确率会略有下降，因为一些错误匹配不可避免地会进入排名前列。NetVLAD 在总体精确率方面仍优于 DBoW，体现了其对真正例进行排序时的稳健性。
- **Top-25：** 在**K**取值最高时，召回率的提升会更加显著。不过，精确率的下降也更为明显，表明系统更容易出现假阳性。即便如此，在大多数召回率水平下，NetVLAD 的精确率仍然更高，这说明它更擅长从更大的候选集中检索出相关回环。

总体而言，这些 PR 曲线凸显了精确率与召回率之间固有的权衡：放宽标准（增大**Top-K**）通常会提高召回率，但会牺牲精确率。**NetVLAD** 展现出更优的平衡，始终能够在保持有竞争力的召回率的同时取得更高的精确率。最佳**Top-K**选择取决于同步定位与建图应用的具体需求：有些系统可能优先考虑精确的回环检测，而另一些系统则可能倾向于扩大检索范围，以最大限度地提高召回率。

C. 时间开销分析

查询速度是实时 SLAM 应用中的另一个关键因素。在实际部署场景中，视觉 SLAM 系统运行于搭载相机的机器人上，相机帧率为 20–30 帧/秒 [4]。由于通常每隔 3 到 4 帧选取一个关键帧，回环检测检索系统必须满足严格的计算约束：为保持实时性能，每个关键帧的回环检测总耗时应显著低于 100 毫秒；换言之，系统每秒至少需要处理 10 个关键帧。

总查询时间还可以进一步分解为两个部分：编码时间和检索时间，表示为：

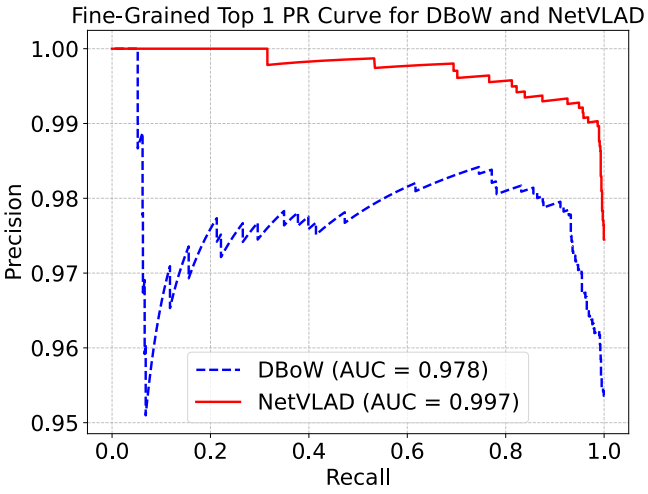

图 8. KITTI序列00的细粒度 Top-1 精确率-召回率曲线

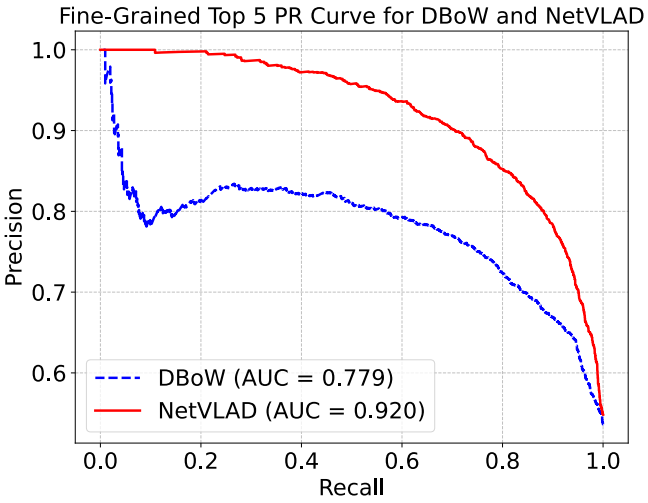

图 9. KITTI序列00的细粒度 Top-5 精确率-召回率曲线

$$T_{\text{query}} = T_{\text{encoding}} + T_{\text{retrieval}} \tag{10}$$

其中：

- 特征编码时间（$T_{\text{编码}}$）——计算图像描述符（例如，提取 NetVLAD 描述符、生成词袋直方图）所需的时间。
- 检索时间（$T_{\text{检索}}$）——在参考数据库中搜索匹配项所需的时间。

设 t_e 表示特征编码时间，t_m 表示匹配两幅图像的特征描述符所需的时间，如 [11] 中所述。总查询时间可表示为 t_R，其中：

$$t_R = t_e + O(N) \times t_m \tag{11}$$

这里，$O(N)$表示图像匹配搜索算法的时间复杂度，其中 N 是图像总数。

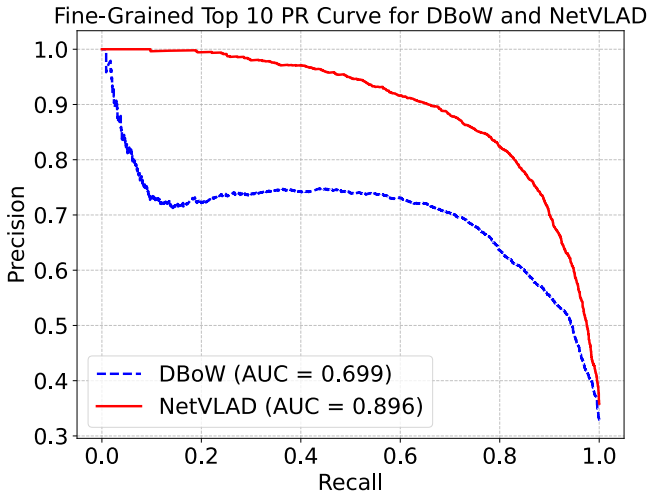

图10. KITTI序列00的细粒度Top-10精确率-召回率曲线

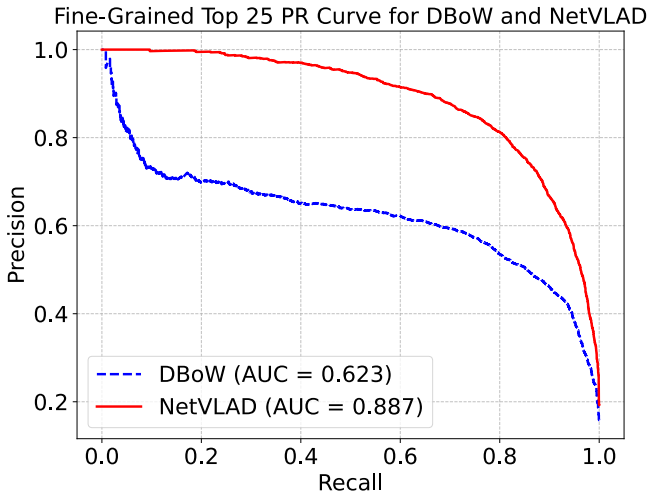

图11. KITTI序列00的细粒度Top-25精确率-召回率曲线

- 对于 DBoW，由于采用倒排索引搜索，它的查询速度相比朴素的线性搜索显著提升；该方法通常只将查询与共享相同视觉词的候选子集进行比较。评估所有查询时，整体搜索复杂度为：

$$O(M \cdot N) \tag{12}$$

其中，M 是每个查询对应的缩减候选集，$M \ll N$。（对于每个查询，开销为 $O(M)$。）

- 对于NetVLAD，在大小为 N 的数据库上进行朴素线性扫描，其复杂度为：

$$O(N) \tag{13}$$

每次查询（如果对所有查询进行评估，则为 $O(N^2)$）。若无加速，在大规模场景下，这一过程的成本会变得很高。为此，我们集成了基于 Faiss 的近似最近邻搜索 [18]，，它利用底层优化（例如单指令多数据）大幅缩短检索时间。

D. 编码与检索时间比较

我们比较了 KITTI [4] 序列 00 数据集中基于 DBoW 的回环检测与我们基于 NetVLAD 并使用 Faiss 的回环检测的特征编码时间和检索时间。该数据集包含 4541 张图像。

1) 编码与检索时间比较： 表I列出了两种方法进行编码与检索所需的总时间，该时间成本针对全部4541幅图像。

表I 编码与检索总时间比较

方法	编码时间（秒）	检索时间（秒）
ORB +DBoW	19.028	9.389
NetVLAD + Faiss（本文方法）	42.600	2.870

2）结果分析：

- **DBoW 计算图像描述子更快**，耗时 19.028 秒，而 NetVLAD 则需要 42.6 秒。这是意料之中的，因为 DBoW 依赖手工设计的特征提取方法（ORB [6],SIFT [7]），计算开销较低，可通过多线程在中央处理器上运行。相比之下，NetVLAD 会先通过深度卷积神经网络（VGG16 [20]）处理图像，再应用 VLAD [14] 池化，因此计算开销更大。

- **NetVLAD + Faiss 在检索方面显著快于**DBoW，仅耗时 2.87 秒，而 DBoW 需要 9.389 秒。DBoW 使用倒排索引搜索，虽然效率较高，但在大规模数据集上的速度仍然较慢。NetVLAD 与 Faiss 结合后，检索速度比 DBoW 快得多 3×，更适合实时 SLAM 应用。

尽管 NetVLAD 的嵌入时间比 DBoW 长，但其检索速度显著更快，因此足以满足实时同步定位与建图的需求。DBoW 和 NetVLAD 的单张图像平均总处理时间分别为 6.257 ms 和 10.013 ms，均远低于回环检测 100ms 的实时限制。鉴于 NetVLAD 在准确性和应对环境变化的鲁棒性方面表现更优，嵌入时间上的权衡是合理的，因为这能带来更可靠的回环检测和更优的同步定位与建图性能，同时仍适用于实时应用。

五、结论

本文研究了基于深度学习的视觉地点识别（VPR）方法——尤其是 NetVLAD——应用于同步定位与建图中的回环检测的可行性。我们在 KITTI数据集上将其与动态词袋（DBoW）进行了对比实验，得出了三项关键发现：

- **准确性与鲁棒性：**NetVLAD在多个Top-K设置下始终优于DBoW，在动态且视觉重复度高的城市场景中实现了更高的精确率和召回率。其学习得到的全局描述符缓解了感知混淆问题，而这类问题常常困扰手工特征方法。

- **细粒度评估：**通过引入专为回环检测设计的细粒度Top-K精确率-召回率曲线，我们凸显了VPR（单匹配假设）与回环检测（多个有效匹配）之间的根本差异。结果表明，NetVLAD在Top-K值较低和较高时均保持了出色的性能。
- **实时可行性：**尽管深度学习产生了更高维的描述子，我们将基于 Faiss 的最近邻搜索集成后，总查询速度仍可与 DBoW 相媲美。这表明基于深度学习的方法可以经过优化，在实时同步定位与建图所要求的严格时间约束下运行。

总体而言，我们的研究结果证实，NetVLAD 是同步定位与建图中回环检测的一种可行且有效的 DBoW 替代方案，能够在环境条件变化时保持稳健性，同时仍满足实际的运行时间要求。未来工作将进一步优化深度描述子（例如模型压缩、GPU 加速），并将我们的方法扩展到其他数据集和传感器模态。我们还计划研究更先进的时间与几何一致性检验，以进一步减少假阳性。

参考文献

[1] R. Mur-Artal、J. M. M. Montiel 和 J. D. Tard´os，《ORB-SLAM：一种通用且精确的单目同步定位与建图系统》，*IEEE*机器人学汇刊，第 31 卷，第 5 期，第 1147–1163 页，2015 年 10 月。 [在线]。网址：https://ieeexplore.ieee.org/document/7219438/ [2] N. Merrill 和 G. Huang，《轻量级无监督深度闭环检测》，发表于机器人学：科学与系统（*RSS*）会议论文集，美国宾夕法尼亚州匹兹堡，2018 年 6 月 26–30 日。[3] R. Mur-Artal 和 J. D. Tard´os，《ORB-SLAM2：面向单目、立体和 RGB-D 相机的开源同步定位与建图系统》，*IEEE*机器人学汇刊，第 33 卷，第 5 期，第 1255–1262 页，2017 年。[4] A. Geiger、P. Lenz 和 R. Urtasun，《我们是否已准备好自动驾驶？KITTI视觉基准套件》，发表于计算机视觉与模式识别会议（*CVPR*），2012 年。[5] D. G´alvez-L´opez 和 J. D. Tard´os，《用于图像序列快速地点识别的二进制词袋》，*IEEE*机器人学汇刊，第 28 卷，第 5 期，第 1188–1197 页，2012 年 10 月。[6] E. Rublee、V. Rabaud、K. Konolige 和 G. Bradski，《ORB：SIFT 或 SURF 的高效替代方案》，发表于*2011*年国际计算机视觉会议。IEEE，2011 年，第 2564–2571 页。[7] D. G. Lowe，《基于局部尺度不变特征的目标识别》，发表于第七届 *IEEE* 国际计算机视觉会议论文集，第 2 卷。IEEE，1999 年，第 1150–1157 页。[8] R. Arandjelovic、P. Gronat、A. Torii、T. Pajdla 和 J. Sivic，《NetVLAD：用于弱监督地点识别的卷积神经网络架构》，发表于*IEEE*计算机视觉与模式识别会议论文集，2016 年，第 5297–5307 页。[9] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等，《DINOv2：无需监督学习稳健视觉特征》，*arXiv* 预印本 *arXiv:2304.07193*，2023 年。[10] A. Dosovitskiy，《一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer》，*arXiv* 预印本 *arXiv:2010.11929*，2020 年。[11] M. Zaffar、S. Garg、M. Milford、J. Kooij、D. Flynn、K. McDonald-Maier 和 S. Ehsan，《VPR-Bench：一种用于视觉地点识别的开源评估框架，可量化视点和外观变化》，国际计算机视觉杂志，第 129 卷，第 7 期，第 2136–2174 页，2021 年。[12] H. Bay，《SURF：加速稳健特征》，计算机视觉—*ECCV*，2006年。[13] M. Cummins 和 P. Newman，《FAB-MAP：外观空间中的概率定位与建图》，国际机器人研究杂志，第 27 卷，第 6 期，第 647–665 页，2008年。[14] H. J´egou、M. Douze、C. Schmid 和 P. P´erez，《将局部描述子聚合为紧凑的图像表示》，发表于*2010*年*IEEE*计算机学会计算机视觉与模式识别会议。IEEE，2010 年，第 3304–3311 页。[15] R. Arandjelovic 和 A. Zisserman，《关于 VLAD 的一切》，发表于*IEEE*计算机视觉与模式识别会议论文集，2013 年，第 1578–1585 页。[16] S. Hausler、S. Garg、M. Xu、M. Milford 和 T. Fischer，《Patch-NetVLAD：用于地点识别的局部—全局描述符多尺度融合》，发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021 年，第 14 141–14152 页。[17] H. Zhang、F. Li、S. Liu、L. Zhang、H. Su、J. Zhu、L. M. Ni 和 H.-Y. Shum，《DINO：采用改进去噪锚框的 DETR 端到端目标检测》，*arXiv* 预印本 *arXiv:2203.03605*，2022 年。[18] M. Douze、A. Guzhva、C. Deng、J. Johnson、G. Szilvasy、P.-E. Mazar´e、M. Lomeli、L. Hosseini 和 H. J´egou，《Faiss库》，2024 年。[19] M. J. Milford 和 G. F. Wyeth，《SeqSLAM：适用于晴朗夏日和暴风雪冬夜的基于视觉路线的导航》，发表于*2012*年*IEEE*机器人与自动化国际会议，2012 年，第 1643–1649 页。[20] K. Simonyan 和 A. Zisserman，《用于大规模图像识别的超深卷积网络》，*arXiv* 预印本 *arXiv:1409.1556*，2014 年。