文章编号：1671-7449(2025)01-0054-09

金字塔局部聚合描述符的视觉位置识别研究

张婉怡[1,2]，王　佳[1,2]，宋明星[1,2]

（1. 吉林师范大学 信息技术学院，吉林 四平 136000；
2. 吉林师范大学 吉林省光电子材料与器件工程研究中心，吉林 四平 136000）

摘　要：视觉位置识别是计算机视觉和机器人领域中重要的研究内容。自然场景中由于视点改变所带来的图像内容变化会对位置识别造成一定的难度。为了解决这一问题，提出一种基于位置聚类的特征重组方法。首先，提出一种通用的金字塔扩展方法PyramidVLAD用于直方图特征提取。此外，为了进一步提升效率，将距离在一定阈值内的图像聚类至同一位置，然后再进行相似性计算。通过多组实验验证所提方法的有效性，使用PyramidVLAD与先进方法APANet进行比较，在Recall@1方面，所提方法在两个数据集中分别取得了1.02和2.54百分点的提升，实验结果表明所提方法能够在两个位置识别的基准数据集中获得比现有方法更好的效果。

关键词：视觉位置识别；金字塔主成分；位置聚类；图像处理
中图分类号：TP911.73　　**文献标识码**：A　　**doi:** 10.62756/csjs.1671-7449.2025009
引用格式：张婉怡，王佳，宋明星．金字塔局部聚合描述符的视觉位置识别研究[J]．测试技术学报，2025，39(1)：54-62.
ZHANG Wanyi，WANG Jia，SONG Mingxing. Learning PyramidVLAD for visual place recognition[J]. Journal of Test and Measurement Technology，2025，39(1)：54-62.

Learning PyramidVLAD for Visual Place Recognition

ZHANG Wanyi[1,2]，WANG Jia[1,2]，SONG Mingxing[1,2]

（1. College of Information Technology，Jilin Normal University，Siping 136000，China；
2. Jilin Engineering Research Center of Optoelectronic Materials and Devices，Jilin Normal University，Siping 136000，China）

Abstract：Visual place recognition is an important issue in both computer vision and robotics. Changes in image content caused by viewpoint changes in natural scenes still pose a challenge to location recognition. To solve this problem，a novel feature reorganization method based on location clustering is proposed. Firstly，a general pyramid expansion scheme is extracted based on histogram features，called PyramidVLAD. To maximize the effect of the new function，the similarity is evaluated by clustering images with a certain threshold into same location. Extensive experiments have been conducted to verify the effectiveness of the proposed method using Pyramid VLAD to compare with the best method，APANet. These two datasets achieve improvements of 1.02 and 2.54 percent points in Recall@1，respectively. The results show that this method can consistently obtain better performance than the state-of-the-art methods on the two standard place recognition benchmarks.

Key words：visual place recognition；pyramid principal phases；place clustering；image processing

收稿日期：2024-04-12
基金项目：吉林省教育厅科研项目(JJKH20230506KJ，JJKH20230510KJ)
作者简介：张婉怡(1985—)，女，副教授，博士，主要从事光电图像处理及光电测试研究。E-mail：zhangwanyi@jlnu.edu.cn。

0 引　言

视觉位置识别是计算机视觉[1]和机器人领域[2]的一个重要研究方向，近年来受到广泛关注。在复杂的室外环境中，光照变化、视角变化和部分遮挡等因素对视觉位置识别提出了挑战，传统方法在这些情况下效果较差。

视觉位置识别本质上是一个实例检索任务[3-4]：给定一个查询图像，需要在一个包含地理标记的数据库中找到最匹配的图像。这个过程通常分为两步：1）通过提取局部或全局特征对地理标记数据库进行训练；2）使用相同的特征提取方法对查询图像进行处理，并估计最佳匹配。

视觉位置识别的研究方法主要分为两大类：传统手工标注特征方法和基于神经网络特征的自动学习方法[5]。传统的手工标注特征方法有词袋模型（BoW）[6-7]、费舍尔向量[8]和局部聚合描述符（VLAD）[9]。BoW简单易实施，适用于大规模图像检索；费舍尔向量提供了一种强大的图像表示方式，能够捕捉更细微的图像差异；VLAD相比BoW更精确，能更好地处理图像的局部特征，然而，这些方法在处理视点和遮挡变化时效果不佳，容易因场景复杂性增加而失效。虽然传统VLAD比BoW表现更好，但在特征冗余和匹配误差方面仍存在限制，尤其在视角和光照变化大的环境中表现不佳。大多数传统方法无法动态调整特征以适应不同环境和场景，因此在快速变化且复杂的环境中，往往无法提供稳定且准确的识别结果。这些方法在不同地点和环境中难以保证准确性，限制了其应用的普遍性。基于神经网络的特征自动学习方法近年来得到广泛关注。例如，NetVLAD网络[10-11]结合了VLAD聚合方法和卷积神经网络（CNN）[12-13]特征提取，显著提高了图像检索效果。然而，这类方法计算资源需求高，对光照、视角、遮挡等变化敏感，导致泛化能力不足。此外，深度学习模型在对抗样本下鲁棒性较差。图神经网络（GNN）[14]和几何变换方法在视觉位置识别中利用图结构和空间几何信息，可以进行更复杂的特征融合和匹配，但其计算复杂度高，数据依赖性强，模型训练难度大。本文尝试通过分析卷积神经网络的表征来提高视觉位置识别的性能。为此，我们需要考虑以下问题：能否从一个位置中选择最重要的局部特征，并通过对这些特征进行重组构建新的特征来表示这一位置？为了解决这一问题，本文提出了3个改进策略。

首先，将直方图特征中的每一组数据定义为一个阶段，并通过统计概率匹配这些阶段特征，这是直方图型特征框架的显著特点。本文之所以将直方图中的每一组数据称为一个阶段，是因为阶段信息在匹配结果中起到了重要作用。具体来说，本文首先通过训练提取出图像中最重要的阶段特征，同时抑制其他不重要的阶段特征，然后将这些重训练的阶段特征扩展为金字塔形式，再将每一层金字塔成分独立进行匹配，并将金字塔中最优的部分融合在一起重组为一个虚拟的完整金字塔特征。通过金字塔形式的特征重组，使得模型能够更好地适应图像在不同尺度上的变化；通过重组和聚类，提高了模型对视点变化和部分遮挡的鲁棒性。

其次，通过将位置定义为一个更大的区域，以便最大化金字塔聚合描述符的作用。在传统的位置识别过程中，最优匹配往往是通过查询图像与数据库中的图像一对一进行粗略匹配实现的，这一过程被称为图像到图像（I2I）的匹配[15]。为了实现特征重组，本文将位置定义为一个更大的区域，即将具有相同位置和相邻位置的所有图像聚类至同一个新的位置。通过将图像基于位置的相似性进行聚类，降低了单个图像误差对整体识别效果的影响。动态调整特征聚类能够根据环境变化调整聚类阈值，提高位置识别的精度和适应性。

最后，本文提出一种弱监督三重损失函数的变体，其思路来源于传统的三重损失函数，能够更好地适应本文的特征框架和聚类位置，此外还提出一种端到端的学习过程用于解决位置识别任务，提高了学习效率和特征表达能力。

1 相关研究综述

1.1 视觉位置识别的传统方法

在深度学习方法普及之前，传统视觉位置识别方法通常采用两步框架：人工特征提取和分类器训练。提取可区分的图像特征是计算机视觉领域的核心工作之一，这些特征可以用于训练支持向量机（SVM）[16]和Boosting[17]等分类器进行训练，此外，BoW被证明不如局部特征敏感。随后，基于直方图的特征被应用到VLAD和费舍尔向量模型中，从而建立了更高阶的局部图像特征统计模型。

由于 AlexNet 在深度学习中的成功，深度学习在图像检索和位置识别领域引起了广泛关注。Torii 等[18-19]提出了一种表示方法，可以解决城市场景中具有重复结构的视觉位置识别问题，并提出了通过结合视点生成和密集 VLAD 的变体[15]来增强鲁棒性识别。在图像检索和视觉位置识别任务中，研究发现三重排序损失在微调预训练 CNN 模型时具有显著效果，并且将 VLAD 应用于学习策略中，结果优于未学习的图像表征。然而，相同场景类别在不同位置的出现可能会降低 NetVLAD 的表示性能。

1.2 位置定义与分类

无论是传统方法还是基于深度学习的方法，对于位置的定义始终是一个基本问题。Lowry 等[20]总结了常见的位置定义：位置的定义因场景不同而异，既可以是一个精确的地点——“位置将部分环境区域描述为一个零维点”，也可以是一个大的区域——“位置可以定义为一个区域的抽象”。本文采用后一种定义，但稍作改动：表示同一位置的不同视点图像以及距离在一定阈值内的图像将被归为一个新的位置。

2 基于 PyramidVLAD 的位置识别方法

本文将具有相邻几何位置的图像聚类为一个新位置，而不是将每张图像独立识别。这种方法的主要优势在于能够利用周围信息进行位置识别。通常认为，查询图像只会与字典特征中的少量词汇匹配，意味着特征中存在主要阶段特征，并且这些特征贡献最大。因此，本文仅保留主要阶段特征，并将其扩展为金字塔形式。

本文提出的视觉位置识别框架如图 1 所示，将一种通用的金字塔扩展方法 PyramidVLAD 用于直方图特征提取，PyramidVLAD 模块为每张图像生成 S 个组成成分。当输入查询图像时，框架首先提取卷积神经网络的局部特征。此阶段可采用任意预训练的卷积神经网络模块，本文所使用的是 VGG[21]网络。

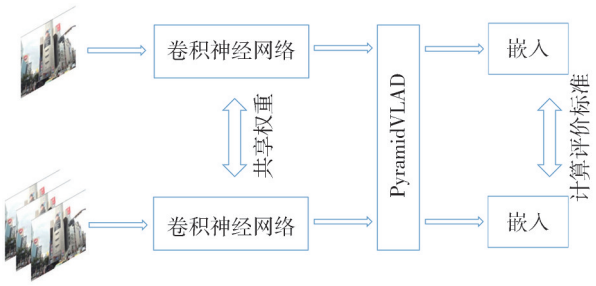


图 1 视觉位置识别的整体框架

Fig. 1 Evaluation framework of visual place recognition

2.1 PyramidVLAD 特征提取方法

2.1.1 直方图特征描述符

本文所提 PyramidVLAD 网络构造如图 2 所示，包括一个预训练的卷积神经网络。卷积神经网络最后一层的输出经过激活函数后输入到 NetVLAD 池化层，最终生成一组扩展特征，本文将局部聚合描述符 NetVLAD 作为基准，当图像 $I^{W'\times H'}$ 经过网络后，最后一层卷积层和激活函数的输出被视为张量 $F\in\mathbf{R}^{(W\times H\times C)}$，其中 F 通常表示数量为 N 的 C 维局部特征，其中 $N=W\times H$。$X=\{x_i\in R^C|i\in\{1,\cdots,N\}\}$。作为字典特征族中的一员，局部聚合描述模块将局部特征编码为 K 个簇，每个簇都由一定数量的特征组成。

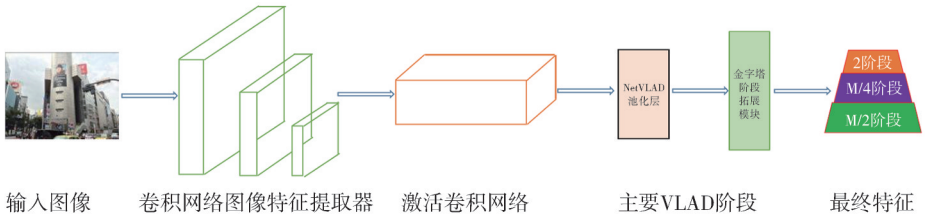


图 2 PyramidVLAD 的整体框架

Fig. 2 Evaluation framework of PyramidVLAD

不同于 BoW 和 NetVLAD 等方法存储的是每个词的残差和，聚合描述符的计算公式如下

$$v_k=\sum_{i=1}^{N}a_k(x_i)(x_i-b_k),\tag{1}$$

式中：b_k 为第 k 个词汇；$a_k(x_i)$ 表示描述符 x_i 到 b_k 的成员关系，即如果 b_k 是最接近描述符 x_i 的簇，则 $a_k(x_i)=1$，否则为 0。整个聚合描述符被表示为

$$\boldsymbol{v}=[\boldsymbol{v}_0^{\mathrm{T}},\boldsymbol{v}_1^{\mathrm{T}},\cdots,\boldsymbol{v}_{(K-1)}^{\mathrm{T}}]^{\mathrm{T}}。\tag{2}$$

在标准 VLAD 形式中，$a_k(x_i)$ 是二值化的；但在深度学习的条件下，为了使 $a_k(x_i)$ 可微，可将

$a_k(x_i)$改写为

$$a_k(x_i)=\frac{e^{-\alpha\|x_i-b_k\|_2}}{\sum_{k'}e^{-\alpha\|x_i-b_{k'}\|_2}}。\tag{3}$$

每张图像对应的K个主要阶段直方图特征通过NetVLAD进行提取。

2.1.2　PyramidVLAD特征生成过程

为了确定特征中哪些阶段是重要的，哪些阶段是不重要的，需要设计一种评分规则对每一个阶段进行评价。根据NetVLAD，采用以下规则对每个阶段进行评分：

1）阶段的特征数量越多，得分越高；

2）阶段的累计残差越小，得分越高。

为了平衡以上两个判断标准，分数的计算标准为

$$r_k=\alpha e^{\frac{c_k}{\sum_{k'}c_{k'}}}+(1-\alpha)e^{-\frac{err_k}{\sum_{k'}err_{k'}}}。\tag{4}$$

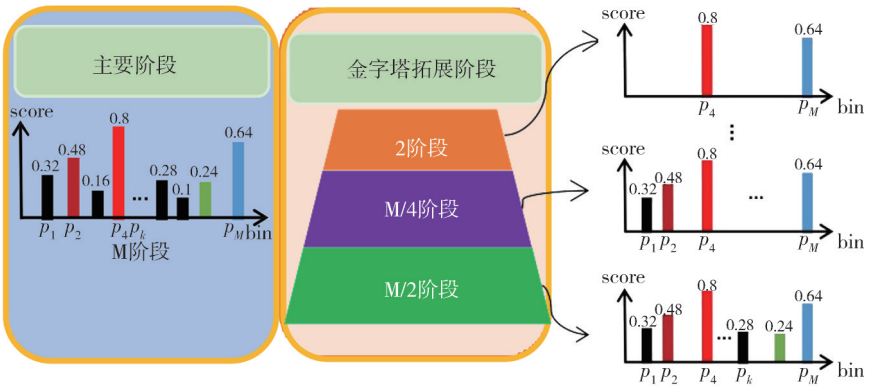
在NetVLAD中，图像的局部特征仅与词汇的一部分子集匹配。现将评分排名前M的阶段特征保留并作为主要阶段特征，其余特征置为零。

$$\boldsymbol{v}'=[0,\cdots,\boldsymbol{v}_{p1}^{\mathrm{T}},\cdots,\boldsymbol{v}_{pk}^{\mathrm{T}},\cdots,0]^{\mathrm{T}},\tag{5}$$

式中：$\boldsymbol{v}_{pk}^{\mathrm{T}}$为第$k$个主要阶段；$\boldsymbol{v}'$为主要阶段特征，维度为$k$。

进一步分析主要阶段特征发现诸如建筑物的主要特征在灯光和广告牌频繁变化的情况下依然保持稳定。使用这些主要阶段特征进行匹配，可以有效抑制因变化特征带来的影响。

基于主要阶段特征，进一步将特征扩展至金字塔形式以提高表达能力。金字塔成分扩展示意图如图3所示。为了简化表述，只考虑特征中的M个主要阶段特征。首先，保留分数排名前$M/2$个的阶段特征，将其他特征置为零；然后对上述过程进行迭代，直到只剩两个主要阶段特征停止。



(a) 直方图特征　　(b) 主要阶段特征　　(c) 金字塔拓展

图3　PyramidVLAD网络

Fig. 3　PyramidVLAD network

最后，每一层金字塔被视为一个组成成分，每个组成成分具有相似的结构和相同的维度，如式(5)所示，在特征扩展之后共产生S个组成成分。

其中，图3(a) NetVLAD产生K个阶段的直方图特征，通过公式(4)对每一阶段特征进行计算，最终保留M个得分更高的主要阶段特征。图3(b)对图3(a)中得到的特征进行扩展。首先保留$M/2$个高分阶段，将其余置为零，然后开始迭代，直到剩余两个主要阶段为止。最终将会得到S个金字塔层特征。图3(c)示例说明金字塔扩展如何实现。

2.2　聚类位置定义

PyramidVLAD对直方图特征进行重组，使得不同图像能够被聚类到不同的位置。在训练数据集中，假设有P张图像，首先将每张图像视为一个独立的聚类中心，并将其定义为一个位置。如果其他图像与该位置的距离在设定的阈值T之内，则这些图像被归为一组。由于这种聚类方法会将包含多个地点的图像分入不同的组，因此需要对重复图像进行剔除，最终只保留Q个位置。图4展示了新位置的定义方法，最终识别出的位置根据式(6)进行计算。

$$C^*=\mathrm{argmin}_c I2P-Dist(F_q,F_c),\tag{6}$$

式中：F_q和F_c分别为查询q的特征和位置c的特征；$I2P-Dist(F_q,F_c)$为对q和c之间的I2P距离进行计算。

根据传统的位置定义，每个点独立作为一个

位置，因此在本图中有23个位置。由于图像可能包含不同位置的相似场景信息，可能会出现误匹配（图中绿色十字星是真实结果但最终匹配到了红色圆）。按照本文提出的聚类位置定义，最终只定义了3个位置，这样不仅不同视角的图像归为一个位置，还将地理位置相邻的图像分为一类；此外，不同位置可能包含同一张图像（图中位置2和位置3包含了重复的图像，意味着该图像中同时包含了2个位置）

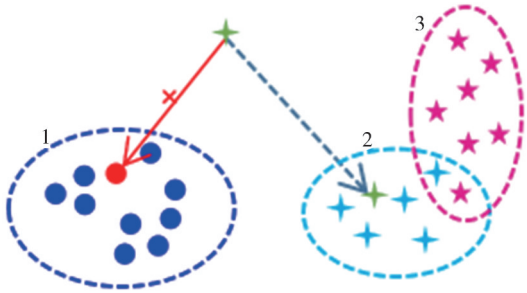


图 4 聚类位置定义

Fig. 4 The clustered place definition

2.3 距离测量和训练损失计算

2.3.1 I2P距离

PyramidVLAD产生s个组成成分。在识别过程中，首先通过下式对查询图像和该位置的第s个成分的最小距离进行计算，

$$d_s(f_s,f_{c,s})=\min_j d_\theta^2(f_s,f_{c,j,s}), \tag{7}$$

式中：$j\in\{1,2,\cdots,N_c\}$；f_s和$f_{c,j,s}$分别为查询q的第s个分量和位置c的第j张图像的s个分量。

将所有分量的最小化距离求和即可以得到查询图像f到位置c的距离，

$$I2P-Dist(F_q,F_c)=\operatorname{argmin}_c\left\{\sum_{s=1}^{S}\omega_s^c d_s(f_s,f_{c,s})\right\}, \tag{8}$$

式中：ω_s^c用于评估位置c中第s个成分的重要性。

2.3.2 弱监督三重损失

从训练数据集中可以得到三元数$\{q,p,n\}$，对于每一张查询图像q，将会存在相应的正样本p和负样本n。因此，PyramidVLAD的弱监督三重损失可以定义为

$$loss=\sum_{s=1}^{S}\omega_s\sum_j l(\min_k)d_\theta^2(q_s,p_{k,s})-d_\theta^2(q_s,n_{j,s})+m), \tag{9}$$

式中：$l(x)=\max(x,0)$；m为常量；$n_{j,s}$为PyramidVLAD中第j个负样本的第s个分量；q_s为查询q的第s个分量；$p_{k,s}$为第k个正样本的第s个分量；ω_s表征每种分量的重要性。对于每一层扩展特征都进行三重损失的计算，并将这些损失求和作为最终的损失。

3 实验设计与结果分析

3.1 数据集与实现细节

本节介绍了实验中使用的基准数据集和PyramidVLAD的实现细节，并讨论了位置聚类的优点。最后，对PyramidVLAD的定性和定量结果进行分析，并与NetVLAD及其他先进方法进行比较。

3.1.1 数据集

实验使用两个公开数据集Pitts250k-test和Tokyo 24/7[22]。Pitts250k是一个来自匹兹堡谷歌街景的视觉位置识别数据集，包含83 000张数据库图像和8 000张查询图像。Tokyo 24/7是一个在白天、日落和夜晚使用不同移动相机拍摄的挑战性数据集，包含75 000张数据库图像和315张查询图像。针对不同的测试集，使用Pitts30k-train或TokyoTM-train对预训练模型进行微调。

3.1.2 实验过程

1）常规参数设置：基础结构由去掉最后一层卷积层的VGG-16和NetVLAD池化层组成，重用了开源网络NetVLAD的参数，并将数据库图像聚类至64个中心，margin设置为0.1，batchsize设置为8。训练过程中的优化算法为SGD，Pitts30k训练集的学习率为$1e^{-4}$，TokyoTM训练集的学习率为$5e^{-4}$。训练20轮后性能不再有显著提高时结束训练。

2）位置聚类：在进行位置聚类时，需要考虑两个关键参数：聚类中心和聚类阈值。

将数据库中的每一张图像视为一个聚类中心，低于阈值的其他图像聚类至同一个中心。若一张图像具有多个中心，只选择一个中心作为新的位置，舍弃其他中心。在实验中对不同的聚类阈值进行了分析。

3）PyramidVLAD生成：设置0.95作为平衡因子，对NetVLAD中每一个阶段特征进行评分，保留得分最高的8个阶段（$M=8$），逐渐减半高分阶段数量，直到剩余两个阶段特征。ω_s^c设置为1.0，表示每层扩展特征具有同等重要性。实验中对主要阶段特征和金字塔扩展的作用进行了分析。

3.2　PyramidVLAD有效性验证

3.2.1　PyramidVLAD参数选取

为了选取PyramidVLAD的最优参数，本文在Tokyo24/7数据集上测试了不同主要阶段特征数量M和金字塔层数PL的性能，实验结果如表1所示。

表1　Tokyo24/7数据集性能对比

Tab. 1　Tokyo24/7 dataset performance comparison

M	PL	Recall@1/%	运行时间/ms
8	1	67.94	22.3
8	2	69.21	24.2
8	3	69.52	25.6
16	1	68.57	24.7
16	2	69.84	26.2
16	3	70.20	27.9
16	4	70.48	28.3
32	1	67.57	26.6
32	2	67.30	28.3
32	3	69.89	30.6
32	4	70.18	32.9
32	5	70.32	33.5

从表1中可知，随着金字塔层数的增加，尽管模型运行时间略有增加，但召回率逐步提高，这表明金字塔式特征扩展通过逐步提取高分特征有效提取了重要特征，并抑制了不重要特征的干扰，提高了模型表达能力。此外，随着主要阶段特征数量M的增加，模型召回率先增加后减少，当M=16时，取得了最优结果。这表明主要阶段特征M数量应选取合适的范围，过多的主要阶段特征会引入更多干扰，同时降低性能并增加运行时间。最终，选择主要阶段特征M为16，金字塔特征扩展层数PL为4，此时的模型Recall@1的性能取得了最优结果为70.48%，后续实验均选用此配置。

3.2.2　PyramidVLAD与NetVLAD性能比较

为进一步验证PyramidVLAD的有效性，本文对比了PyramidVLAD和NetVLAD的性能，除了局部聚合描述符提取方式不同，其余配置均相同，实验结果如表2所示。在运行时间相近的情况下，PyramidVLAD的Recall@1性能提高了3.5百分点，充分验证了其有效性。

综上所述，在视觉位置识别中，一些特征如建筑物、商场和地标等难以改变，而广告牌和LED灯光等频繁变化。这些保持不变的特征是最容易区分的，而变化的特征会带来噪声。基于直方图特征的方法能通过统计学手段抑制变化内容的影响，PyramidVLAD进一步提升了这一效果。通过提取主要阶段特征，去除了非主要且变化的内容；通过金字塔扩展，有效消除了变化或误导匹配的内容。

表2　不同局部聚合描述符提取方式对比

Tab. 2　Comparison of different local aggregation descriptor extraction methods

局部聚合描述符提取方式	Recall@1/%	运行时间/ms
NetVLAD	66.98	27.4
PyramidVLAD	70.48	28.3

3.3　I2P有效性验证

3.3.1　I2P参数选择

为了选取I2P最优参数，本文在Tokyo24/7数据集上设置不同阈值T进行性能对比，实验结果如表3所示。从表3中可得，当阈值T设为25时，模型Recall@1和Recall@5分别达到了最佳性能70.48%和82.22%，这表明选取合适的阈值对算法性能至关重要。

表3　不同阈值T性能对比

Tab. 3　Performance comparison with different thresholds T

T	Recall@1/%	Recall@5/%
0	66.35	76.23
9	67.30	77.45
25	70.48	82.22
36	68.32	80.10
49	66.75	79.25

3.3.2　与I2I对比

为了进一步验证I2P有效性，本文将其与I2I方法进行了对比，实验结果如表4所示。与I2I方法相比，I2P方法的Recall@1和Recall@5分别提高了5.72和6.96百分点，充分证明了I2P方法的有效性。

表4　不同检索方式性能对比

Tab. 4　Performance comparison of different retrieval methods

不同检索方式	Recall@1/%	Recall@5/%
I2I	64.76	75.26
I2P	70.48	82.22

综上所述，本文将一个区域抽象为一个位置，即将一个地理位置周围及附近的图像聚类为同一个位置。这与传统方法不同，传统方法将每张图像视为一个独立的位置，即使有些图像的地理位置是相同的。因此，本文将传统的图像对图像（I2I）检索转变为图像对位置（I2P）检索。实验总结出位置聚类定义的两个主要优势：

1）减少错误匹配：将相同地理位置的图像归为一类，有效避免了错误匹配，并能够在特定视角下得到最准确的匹配查询；

2）提高召回率：新位置的定义能够帮助机器人识别获得更高的召回率，从而提高后续模块的准确率，如全局路径规划、自动驾驶和机器人导航等。

3.4 实验结果与分析

3.4.1 与基准方法的比较

与基准方法对比结果如表 5 所示，结果表明本文方法在 Recall@1 指标上优于 NetVLAD。在 Pitts250k-test 和 Tokyo 24/7 数据集中，Recall@1 分别提升了 4.01 和 10.48 百分点。

表 5 对比 NetVLAD 的召回率比较

Tab. 5 Comparison of recalls with NetVLAD

方法	Tokyo24/7		Pitts250k-test	
	Recall@1/%	Recall@5/%	Recall@1/%	Recall@5/%
NetVLAD	60.00	73.65	80.66	90.88
本文	**70.48**	**82.22**	**84.67**	**92.08**

图 5 展示了本文方法与基准算法检索结果的可视化。每 1 列代表一个查询，第 1 行是查询图像，第 2 行是本文方法的结果，第 3 行是 NetVLAD 的结果。绿色矩形表示正确结果，红色矩形表示错误结果。

图 5 对比 NetVLAD 的可视化结果比较

Fig. 5 Comparison recognized result with NetVLAD

综述所示，相较于基准模型，本文方法通过引入金字塔式扩展特征，抑制了不重要特征并减少了冗余特征的干扰。同时，采用 I2P 检索方法，将相同地理位置归为一类，有效避免了错误匹配，提升了检索性能。

本文所使用的微调神经网络为 VGG-16。图 6 展示了 Tokyo 24/7 数据集中一些成功的案例。每 1 行代表一个查询，第 2 列是查询图像，第 2 列是 NetVLAD 的结果，第 3 列是本文方法的结果。绿色矩形表示正确结果，红色矩形表示错误结果。

实验结果说明，本文方法相较于基准网络在应对视点和光照变化方面具有更好的鲁棒性。这些优异的结果归功于位置定义的新方法及特征扩展策略的应用。

图 6 成功示例

Fig. 6 Success cases

3.4.2 PyramidVLAD 与 APANet 性能比较

在本文使用的基准数据集中，APANet 表现最佳。因此，本文将 PyramidVLAD 与 APANet 进行了比较。如图 7(a)～图 7(c) 所示，PyramidVLAD 在 Recall@1 方面的两个数据集中分别提升了 1.02 和 2.54 百分点。虽然在 Pitts250k-test 数据集中，PyramidVLAD 在 Recall@5 和 Recall@20 方面略逊于 APANet，但考虑到 APANet 主要通过注意力机制解决位置识别问题，这为未来的工作提供了思路，即在特征选择前的图像预处理阶段引入注意力机制。

图 7(d) 显示了使用 I2P 距离所带来的召回率提升，证明了 I2P 距离会对视觉位置识别效果提升。

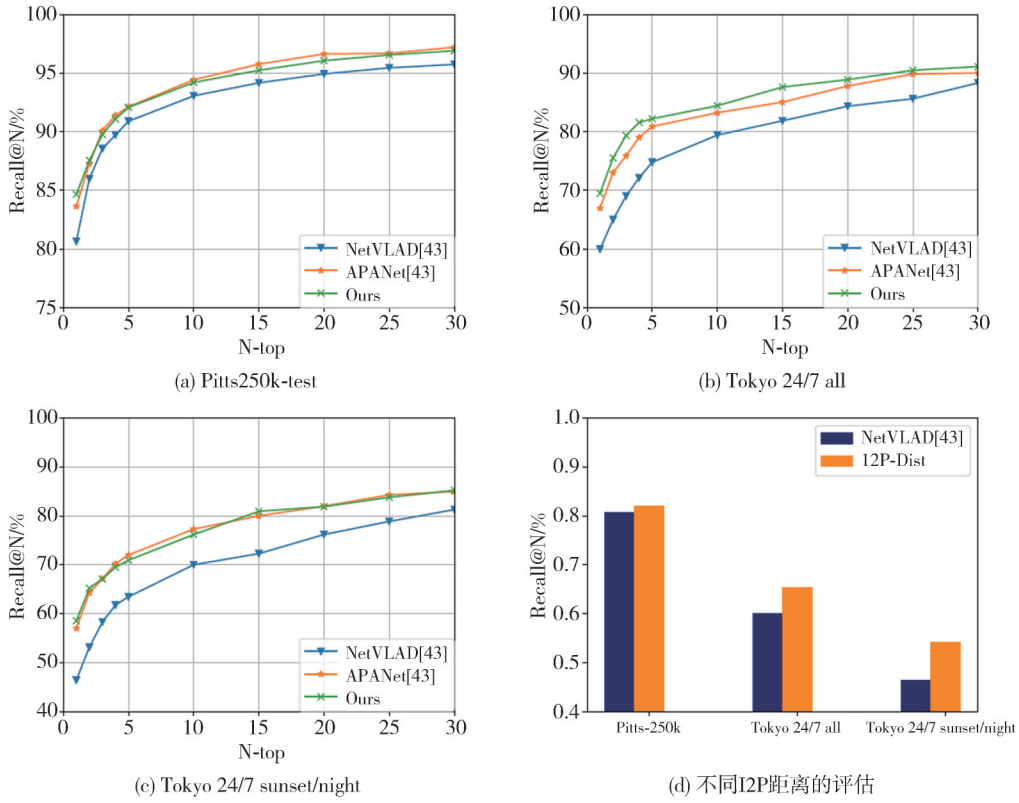


图 7　不同数据方法的比较

Fig. 7　Comparison of different datas

4　结　论

本文提出了一种基于位置聚类的特征机制用于视觉位置识别，能够更有效地捕获和表达位置的关键视觉信息。同时设计了图像到位置距离用于评估新定义位置中扩展特征的相似性，提高了位置识别的准确性和效率。然后，提出一种端到端的网络用于测试和评估提出的方法。实验结果表明，PyramidVLAD 在 Recall@1 方面优于 NetVLAD，在 Pitts250k-test 和 Tokyo 24/7 数据集中分别提升了 4.01 和 10.48 百分比。与最佳方法 APANet 相比，PyramidVLAD 在 Recall@1 方面在两个数据集中分别提升了 1.02 和 2.54 百分比。实验结果证明，本文提出的方法在光照和视点变化方面具有良好的鲁棒性，且具有通用性，可以用于其他基于直方图的位置识别方法。未来研究可以探索将此技术应用于动态环境下的实时处理。

参考文献：

[1] 卢荣胜，史艳琼，胡海兵. 机器人视觉三维成像技术综述[J]. 激光与光电子学进展，2020，57(4)：1-19.
LU Rongsheng，SHI Yanqiong，HU Haibing. Review of three dimensional imaging techniques for robotic vision[J]. Laser & Optoelectronics Progress，2020，57(4)：1-19.（in Chinese）

[2] 叶海峰，赵玉琛. 视觉位置识别中代表地点的标识牌算法[J]. 小型微型计算机系统，2021，42(4)：823-828.
YE Haifeng，ZHAO Yuchen. Algorithm of signboard representing place in visual position recognition[J]. Journal of Chinese Computer Systems，2021，42(4)：823-828.（in Chinese）

[3] 王红君，郝金龙，赵辉，等. 大规模城市环境下视觉位置识别技术的研究[J]. 计算机应用与软件，2021，38(8)：194-198.
WANG Hongjun，HAO Jinlong，ZHAO Hui，et al. Visual position recognition technology in large-scaleurban environment[J]. Computer Applications and Software，2021，38(8)：194-198.（in Chinese）

[4] ZAFFAR M，GARG S，MILFORD M，et al. An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change[J]. International Journal of Computer Vision，2021，129(7)：2136-2174.

[5] 张文炽，陈黎辉，吴炜，等. 基于卷积神经网络特征融合的交通标志识别[J]. 计算机应用，2019，39

(S1): 21-25.
ZHANG Wenchi, CHEN Lihui, WU Wei, et al. Traffic sign recognition based on feature fusion of convolutional neural networks[J]. Journal of Computer Applications, 2019, 39(S1): 21-25. (in Chinese)

[6] LI T, MEI T, KWEON I S, et al. Contextual bag-of-words for visual categorization[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2011, 21(4): 381-392.

[7] 朱小波, 车进. 融合BOW模型的多特征子空间行人重识别算法[J]. 计算机工程与应用, 2019, 55(18): 146-150.
ZHU Xiaobo, CHE Jin. Multi-feature subspace person re-identification based on BOW model[J]. Computer Engineering and Applications, 2019, 55(18): 146-150. (in Chinese)

[8] BHUVANESWARI N R, KUMAR V G S. A multi feature fusion based image classifi cation using multi-class support vector machine[J]. International Journal of Control Theory and Applications, 2016, 9(36): 551-559.

[9] ÖZDEMIR A, SCERRI M, BARRON A B, et al. EchVPR: echo state networks for visual place recognition [J]. IEEE Robotics and Automation Letters, 2022, 7(2): 4520-4527.

[10] HAUSLER S, JACOBSON A, MILFORD M. Multi-process fusion: visual place recognition using multiple image processing methods[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1924-1931.

[11] GONG Q, LIU Y, ZHANG L, et al. Ghost-dilNetVLAD: a lightweight neural network for visual place recognition[J/OL]. [2024-03-12]. arXiv: 2112.11679. https://doi.org/10.48550/arXiv.2112.11679.

[12] CHAITANYA K J, CRISTIAN B, SIMON V M, et al. On the expressive power of geometric graph neural networks[C]//Proceedings of the 40th International Conference on Machine Learning (ICML), 2022, 202: 15330-15355.

[13] CHEN W, LIU Y, WANG W, et al. Deep learning for instance retrieval: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7270-7292.

[14] HAN K, WANG Y H, GUO J Y, et al. Vision GNN: an image is worth graph of nodes: a survey [J]. Neural Information Processing Systems, 2022, 35(4): 1-15.

[15] 贾迪, 朱宁丹, 杨宁华, 等. 图像匹配方法研究综述[J]. 中国图象图形学报, 2019, 24(5): 677-699.
JIA Di, ZHU Ningdan, YANG Ninghua, et al. Image matching methods[J]. Journal of Image and Graphits, 2019, 24(5): 677-699. (in Chinese)

[16] CHEN Y, GAN W, JIAO S, et al. Salient feature selection for CNN-based visual place recognition[J]. IEICE Transactions on Information and Systems, 2018, 101(12): 3102- 3107.

[17] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The Annals of Statistics, 2000, 28(2): 337-374.

[18] TORII A, SIVIC J, PAJDLA T, et al. Visual place recognition with repetitive structures[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(11): 2346-2359.

[19] ARANDJELOVIC R, GRONAT P, TORII A, et al. Netvlad: CNN architecture for weakly supervised place recognition[J]. IEEE Transactions on Ransactions Pattern Analysis and Machine Intelligence, 2018, 40(6): 1437-1451.

[20] LOWRY S, SÜNDERHAUF N, NEWMAN P, et al. Visual place recognition: a survey [J]. IEEE Transactions on Robotics, 2016, 32(1): 1-19.

[21] CAMPOS J, YEE A, VEGA I F. Simplifying VGG-16 for plant species identification [J]. IEEE Latin America Transactions, 2022, 20(11): 2330-2338.

[22] TORII A, ARANDJELOVIC R, SIVIC J, et al. 24/7 place recognition by view synthesis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(2): 257-271.