doi:10.6041/j.issn.1000-1298.2026.04.015

面向温室移动机器人视觉定位场景识别方法

周云成　于美玲　吴帛航　张福宁
（沈阳农业大学信息与电气工程学院，沈阳 110161）

摘要：针对温室移动机器人定位需求及高动态变化且高度相似的温室场景识别难题，本文提出一种基于局部特征选择与聚合的视觉定位场景识别方法模型。模型以预训练视觉 Transformer（DINOv2）为骨干网络，提取场景图像局部特征，并设计一种基于可学习查询词局部特征选择与聚合的全局描述符生成策略，通过交叉注意力机制筛选判别性强的局部特征，生成紧凑的全局描述符。融合对比学习与三元组学习优势，应用多重相似性损失函数对识别模型进行优化。构建了包含 2 100 个场景、25 000 幅图像的温室场景数据集，数据涵盖了光照、视角、距离和时序变化等多维干扰因素。试验结果表明，在自建的温室场景数据集上，模型 top－1 召回率（*R*@1）、*R*@5 和 *R*@10 分别达到 88.79%、96.49% 和 97.96%，较现有场景识别基准方法 NetVLAD、GeM、CosPlace、EigenPlaces、MixVPR 和 SALAD 等，*R*@1 分别提高 23.70、19.24、10.64、3.30、3.90、0.44 个百分点。强光和弱光条件下 *R*@1 波动小于 5 个百分点，视角变化 15°内精度保持为 93.12%，距离变化 2 倍时 *R*@1 为 63.94%，5 d 内作物生长变化下 *R*@1 为 61.14%。模型对光照、视角及采样距离变化具有一定的鲁棒性，但大幅度的视角和距离变化会降低模型性能，作物长期生长变化对模型性能影响较大。在温室移动机器人上进行场景识别验证，平均识别率为 85.88%，具有一定的实用性。构建的基于可学习查询词的特征选择与聚合机制及所选取的特征提取骨干网络，有效提升了温室场景识别精度，提出的技术框架可为温室移动机器人视觉系统设计提供参考。

关键词：移动机器人；温室；视觉定位；场景识别
中图分类号：S24；TP183　　**文献标识码**：A　　**文章编号**：1000-1298(2026)04-0151-11　　**OSID**：

Visual Place Recognition for Localization of Mobile Robots in Greenhouse

ZHOU Yuncheng　YU Meiling　WU Bohang　ZHANG Funing
(*College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110161, *China*)

Abstract: Aiming to address the challenges of greenhouse mobile robot localization and scene recognition in highly dynamic environments with visually similar scenes, a novel scene recognition model was proposed based on local feature selection and aggregation. The model employed a pre-trained vision Transformer (DINOv2) as its backbone network to extract local image features and introduced a learnable query-based feature selection and aggregation strategy to generate discriminative global descriptors. By leveraging cross-attention mechanisms, the model selectively aggregated the most informative local features into compact global representations. Furthermore, a hybrid loss function combining contrastive learning and triplet learning was applied to optimize the recognition model. A comprehensive greenhouse scene dataset containing 2 100 scenes and 25 000 images was constructed, covering multiple challenging factors such as illumination variations, viewpoint changes, distance scaling, and temporal crop growth. Experimental results demonstrated that the proposed model achieved Top－1 recall rates (*R*@1) of 88.79%, 96.49% (*R*@5), and 97.96% (*R*@10) on the collected dataset, outperforming state-of-the-art scene recognition benchmarks, including NetVLAD, GeM, CosPlace, EigenPlaces, MixVPR, and SALAD by 23.70, 19.24, 10.64, 3.30, 3.90, and 0.44 percentage points in *R*@1, respectively. The model exhibited strong robustness under varying illumination conditions (*R*@1 fluctuation < 5 percentage points), moderate viewpoint changes (93.12% accuracy within 15° deviation), and scaling variations (63.94% *R*@1 at 2 × distance). However, performance declined under extreme viewpoint/distance changes and long-term crop growth variations (61.14% *R*@1 after 5 days). Real-world

收稿日期：2025－08－06　修回日期：2025－09－15
基金项目：国家重点研发计划项目(2023YFD1501303、2021YFD1500204)
作者简介：周云成(1979—)，男，教授，博士，主要从事农业人工智能技术研究，E-mail：zhouyc2002@syau.edu.cn

validation on a greenhouse mobile robot confirmed the model's practicality, achieving an average recognition rate of 85.88%. The proposed learnable query-based feature aggregation mechanism, combined with the carefully selected feature extraction backbone, significantly improved recognition accuracy in greenhouse environments. This framework can provide a viable technical solution for vision systems in agricultural mobile robotics.

Key words: mobile robot; greenhouse; visual localization; visual place recognition

0 引言

日光温室是我国北方设施农业的核心形态[1],然而,由于温室狭小空间的物理限制(通常垄宽小于 1 m),大型农业机械难以应用,发展空间适应性强的小型智能装备成为破局的关键[2]。基于视觉的场景识别(Visual place recognition, VPR)通过提取查询图像与参考图像的深度特征关联,可解决两类核心问题:当参考图像位置已知时,VPR 可为智能农机提供有效的空间定位,可解决卫星导航在钢架结构温室内难以有效工作的问题;当参考图像位置信息缺失时,也可应用 VPR 进行回环检测,可有效防止农业机械冗余作业行为。因此,开展温室环境下的 VPR 研究,对推动小型化温室智能农机发展具有重要意义。

VPR 作为计算机视觉领域的重要研究方向,其方法论经历了从传统图像处理到深度学习的演进过程。早期研究主要依赖手工设计特征,LOWE[3] 提出的尺度不变特征变换与 BAY 等[4] 改进的加速稳健特征,通过关键点匹配实现场景辨识。视觉词袋(Bag-of-Words, BoW)模型[5] 将局部特征编码为全局表征,但受限于特征鉴别力不足,在相似场景中易失效。SÜNDERHAUF 等[6] 指出鲁棒的 VPR 系统需同时具备 3 个关键特性:光照不变性、视角不变性及场景泛化性。卷积神经网络(Convolution neural network, CNN)凭借其强大的局部特征提取能力与层次化表征机制,为复杂场景下视觉识别提供了有效范式。JEGOU 等[7] 以深度神经网络(Deep neural network, DNN)特征为基础,提出向量局部聚合描述符(Vector of locally aggregated descriptors, VLAD)编码方法,由此构建有效的视觉位置识别模型;NetVLAD 方法[8] 引入概率分配机制,将 CNN 局部特征动态分配至聚类中心,构建了端到端的 VPR 模型,显著提升了识别精度,但其推理时需依赖训练过程中得到的聚类中心先验,泛化能力受限。为突破此局限,KHALIQ 等[9] 提出 MultiRes - NetVLAD,通过编码低分辨率图像金字塔并聚合多尺度特征,增强表征学习能力。张冬灿等[10] 则融合方向梯度直方图的几何约束与 CNN 特征,强化了视角变化下的描述符鲁棒性。AMAR 等[11] 提出了 MixVPR 方法,通过全连接层全局感受野特性,实现了局部特征空间关联性显式建模,实现了高效特征混合。视觉 Transformer(Vision Transformer, ViT)凭借其独特的自注意力机制,为视觉场景识别提供了新的技术思路。研究者通常采用 MixVPR 或 NetVLAD 等先进特征聚合模块与 Transformer 主干网络相结合的方式构建识别系统。KEETHA 等[12] 使用 Transformer 架构预训练模型进行局部特征提取,随后采用无监督聚合策略生成具有判别性的全局描述符。该范式不仅充分发挥了 Transformer 在特征提取方面的优势,还通过特征聚合模块有效提升了场景描述的紧凑性和鲁棒性。

基于 Transformer 的两阶段检索策略在场景识别领域取得了显著进展[13-14]。该策略采用分层次的特征匹配机制:第 1 阶段采用高效的全局描述符进行快速粗检索,快速缩小候选范围至相似图像;第 2 阶段则基于局部特征进行精细匹配和几何验证,从而提高识别精度。然而,现有的场景识别研究主要聚焦于静态室内或单一室外环境[15-17],而温室场景具有显著不同的特性。由于作物生长周期中形态变化(如植株高度、叶片密度、果实成熟度)以及频繁的农事操作(如修剪、采摘),温室场景呈现高度动态性。此外,温室建筑结构重复性(如统一的墙体裁料)以及作物单一性(如品种、颜色、纹理)导致不同场景之间具有极强的相似性,使得基于深度神经特征的场景识别方法在温室环境中面临巨大挑战,其适用性及针对温室特殊环境的设计改进方法亟需探明。

鉴于温室移动机器人等智能农业机械视觉系统在复杂环境中的定位需求,结合温室场景存在动态变化和结构高度相似性的问题,在基于多层感知机特征混合方法研究[18] 基础上,本文提出基于可学习查询词特征选择与聚合的温室场景识别方法。以预训练深度特征提取框架为基础,获得当前图像局部特征,借助可学习查询词,使用交叉注意力机制选择并聚合场景图像局部特征,通过捕捉特征间依赖关系,将多层次特征聚合为全局描述向量,采用多重相似性度量损失函数,为模型提供明确的优化目标,建立温室场景识别模型。在番茄栽培温室场景中开展试验,通过系统性能评估,验证所提方法的有效性,

以期为温室智能农业机械视觉系统开发提供理论支持。

1　材料与方法

1.1　温室场景数据采集

2024 年 10—12 月在沈阳农业大学现代农业科研基地的多座日光温室内，进行了图像数据采集工作。温室内采用基质栽培番茄，吊蔓生长，株高为 1.0 ~ 1.8 m，行距约 0.9 m，株距约 0.3 m。为构建具有充分多样性的数据集，采用多设备、多条件采集策略。使用华为 Mate 60（4 000 像素 ×3 000 像素）、小米 Note 14 Pro（4 000 像素 ×3 000 像素）和海康威视 E12 摄像头（1 920 像素 ×1 080 像素）3 种成像设备，分别在晴天、多云和雪天条件下，于 08:00—10:00、11:00—13:00、16:00—18:00 不同时段进行图像采集。

为模拟机器人在实际工作环境中的视角、距离和光照条件的变化，设计了不同的数据采集方案。对于视角变化，采用三脚架固定相机，以 0°为基准，旋转 45°范围进行多角度拍摄。对于距离变化，通过滑轨调节相机与拍摄场景间的距离，进行不同距离成像。对于光照变化，同一场景下，分别在不同天气及不同时段多次成像。同时为模拟场景自身的变化，间隔不同时间对同一场景进行重复拍摄，使得数据包含作物生长带来的场景差异。最终构建了包含 2 100 个独立场景的数据集，共计 25 000 幅图像，按比例 4∶1 划分为训练集和测试集，分别用于温室场景识别模型训练与测试。该数据集充分考虑视角、距离、光照和时间等多维度的变化因素，可为模型训练提供丰富的样本。

1.2　温室场景识别技术

机器人在温室作业时，通过搭载视觉感知系统，实时采集当前环境数据（即查询图像 I）。系统首个处理环节采用由深度神经网络构建的特征提取器 h' 对查询图像进行初步编码，提取具有判别性的局部特征（表示为 H）。通过专门设计的特征聚合器 g 将这些局部特征融合为紧凑的全局描述符（表示为 F），该过程可表示为：$F \leftarrow g(H)$，$H \leftarrow h'(I)$。

对于预先建立的参考图像数据库，系统对库中所有参考图像执行统一的处理流程，提取其全局描述符并存储。在识别阶段，系统以查询图像的全局描述符为检索依据，通过计算余弦相似度等度量方式，在参考数据库中筛选匹配度最高的候选图像。当相似度高于预设阈值时，即可判定查询图像与匹配到的参考图像源自同一场景。当参考图像与温室空间位置信息绑定时，该系统可应用于空间定位；当参考图像为智能装备作业过程动态采集时，则可用于闭环检测，判断农业机械是否重返历史作业区域。温室场景识别基本技术框架如图 1 所示。

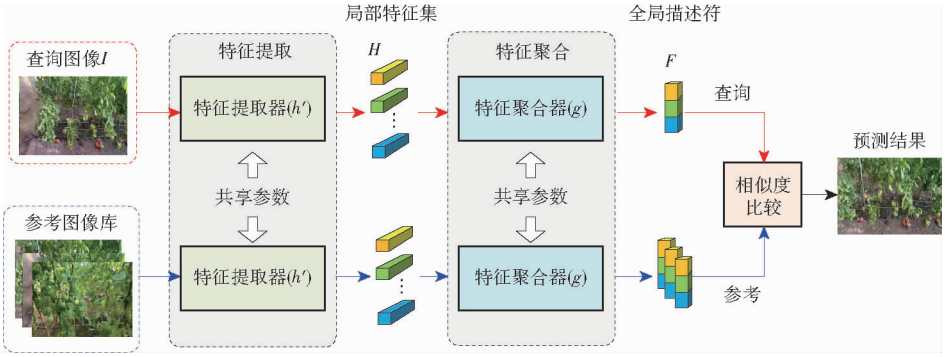


图 1　温室场景识别技术框架

Fig. 1　Technical framework for greenhouse visual place identification

由图 1 可知，温室场景识别模型 f 是由特征提取器 $h'(\cdot)$ 与特征聚合器 $g(\cdot)$ 共同组成的，即有 $F \leftarrow f(I) = g(h'(I))$。该模型识别准确率取决于 2 个关键组件协同作用，即特征提取器负责捕获图像的底层视觉特征，而特征聚合器则实现特征的语义级融合。

1.3　场景图像局部特征提取器

在计算机视觉领域，ViT[19] 与 CNN[20-24] 作为 2 种主流特征提取架构，在各类视觉任务中均展现出卓越性能。近年来，随着 Transformer 架构在自然语言处理领域取得突破性进展，在计算机视觉任务中的应用价值也日益凸显。特别是在大规模预训练模型方面，Transformer 架构展现出了显著优于传统 CNN 的性能优势。基于 ViT 架构的预训练大模型 DINOv2[25]，通过堆叠多个 Transformer 块，实现了从局部到全局的层次化特征提取，能够逐步建立对图像的理解，从识别局部细节到把握整体场景，有助于提升场景识别精度。选择 DINOv2 作为温室场景图像特征提取器，其结构如图 2a 所示。

DINOv2 模型首先对输入图像 $I \in \mathbf{R}^{c \times h \times w}$ 进行预

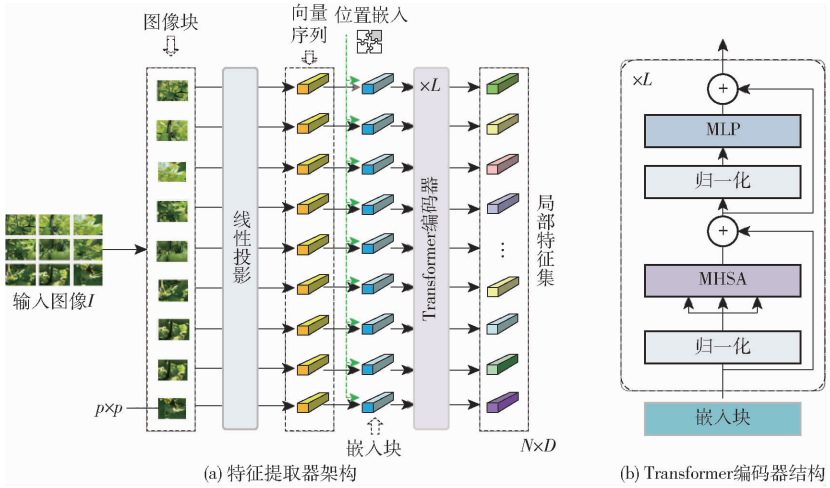


图 2 基于 DINOv2 的图像特征提取器

Fig. 2 DINOv2-based image feature extractor

处理,其中 c、h、w 表示图像通道数、高度和宽度。模型将输入图像划分成 $N=(h/p)\times(w/p)$ 个尺寸为 $p\times p$ 的非重叠图像块,每个图像块都包含原始图像重要局部信息。这些图像块经过展平处理后形成向量序列,通过线性映射层进行特征变换,将每个向量序列投影到 D 维特征空间,生成对应的 Token 表示。为了保留图像空间结构信息,模型在 Token 序列中加入可学习的位置嵌入编码,使得 Transformer 能够感知各图像块之间的相对位置关系。经过位置编码增强的 N 个 Token 将作为输入,送入由 L 个 Transformer 块堆叠而成的编码器。每个 Transformer 块采用标准的架构设计,包含 1 个多头自注意力模块(MHSA)和 1 个多层感知机(MLP)模块,其中 MHSA 模块负责捕捉图像块之间的长距离依赖关系,MLP 则对特征进行非线性变换和增强。这种层次化特征提取方式使得模型能够逐步建立从局部到全局的图像理解能力。

采用 DINOv2 的基础版作为温室场景图像特征提取器,其 $p=14$、$L=12$。DINOv2 的训练采用自监督学习方法,通过改进的 iBOT 目标函数进行训练,其核心是通过掩码图像建模和在线目标蒸馏,结合跨视图一致性约束,使模型学习具有语义一致性的视觉表征。训练过程使用超大规模未标注数据集(数亿图像),DINOv2 学习到丰富多样的通用视觉特征表示,提升模型泛化能力和特征表达能力,在多种下游任务中表现出极佳的性能。DINOv2 通过自监督学习得到的特征具有语义一致性,即不同视角或光照条件下同一场景图像能够提取出相似特征表示,这种特性对于温室场景识别任务尤为重要。因此采用该模型骨干网络作为温室场景图像局部特征提取器。去除 DINOv2 分类器部分,并冻结网络中的权重,将最后一层 Transformer 编码器输出的 N 个 Token 特征作为温室场景图像的局部特征集,表示为

$$\boldsymbol{H}=\{t_i\mid i\in[1,N],t_i\in\mathbf{R}^D\}\quad(\boldsymbol{H}\in\mathbf{R}^{N\times D})\tag{1}$$

式中 t_i——局部 Token 特征

D——特征向量维度

$\boldsymbol{H}$——由 $\{t_i\}$ 组成的特征矩阵

在 $\boldsymbol{H}$ 基础上,借助后续设计的特征聚合器进行聚合处理,最终生成温室场景图像 I 的全局描述符 F。

1.4 特征聚合器设计

1.4.1 特征聚合器结构

针对特征提取器 DINOv2 输出的图像块 Token 序列 $\boldsymbol{H}\in\mathbf{R}^{N\times D}$(其中每个 Token 编码了图像局部区域的深度特征),展平处理会导致全局描述符 F 的维度高达 ND,不仅增加计算开销,还会引入大量冗余信息(如温室场景中低区分度的地面或墙体特征)。MixVPR 等方法通过在局部特征集上应用自注意力和线性变换聚合并降维生成全局描述符,没有考虑特征冗余性,并将各特征重要性等同对待,可能会出现判别力不足的背景特征干扰识别结果的问题。为此,设计了一种基于可学习查询词与交叉注意力的局部特征动态聚合机制(图 3),其核心思想是通过由 M 个可学习的向量构成的全局查询词袋 $\boldsymbol{Q}\in\mathbf{R}^{M\times D}$ 与局部特征序列 $\boldsymbol{H}$ 进行交叉注意力计算,自适应地筛选具有场景识别判别性的局部特征,并生成维度固定且紧凑的全局描述符 F。

设计的特征聚合器通过 l 个聚合模块堆叠并附加聚合头而成,每个聚合模块又由两部分构成,一部分为 Transformer 编码器模块,一部分为基于查询词的局部特征选择模块。为了保持 DINOv2 的通用视觉表示,在温室场景识别模型的训练过程中冻结 DINOv2 的权重。为使 DINOv2 输出的局部特征序

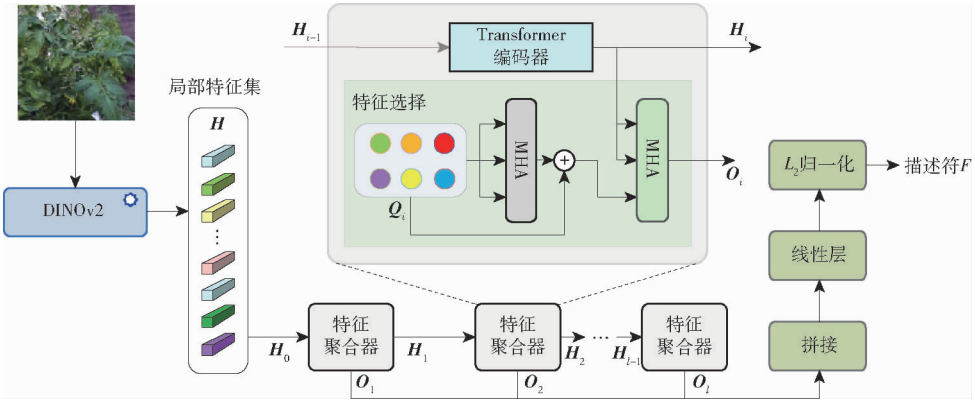


图 3　图像局部特征聚合器

Fig. 3　Image local feature aggregator

列适应场景识别需求，在聚合模块中引入 Transformer 编码器（Encoder）模块，进一步对 DINOv2 输出特征集 $\boldsymbol{H}$ 进行处理，以适配场景识别任务需要。聚合模块所采用的 Transformer 编码器结构与 DINOv2 中的编码器结构（图 2b）完全相同，不再赘述。Transformer 编码器将输入的图像局部特征序列映射为输出，即

$$\boldsymbol{H}_i = \mathrm{Encoder}_i(\boldsymbol{H}_{i-1}) \tag{2}$$

式中　$\boldsymbol{H}_{i-1}$——第 i 个聚合模块中 Encoder 的输入特征序列

$\boldsymbol{H}_i$——经 Encoder 转换后的输出特征序列

其中 $\boldsymbol{H}_0$ 为 DINOv2 特征提取器输出特征序列。

1.4.2　基于查询词的特征选择

基于查询词的特征选择模块，应用多头注意力（Multi-Head Attention，MHA）机制[26]。MHA 包括 3 种输入：查询（表示为 $\boldsymbol{q}$）、关键字（表示为 $\boldsymbol{k}$）和值（表示为 $\boldsymbol{v}$），经过线性映射将 3 种输入转换为多个并行的头（head），每个 head 输出为

$$\mathrm{MHA}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}) = \mathrm{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{D}}\right)\boldsymbol{v} \tag{3}$$

在 MHA 中，$\boldsymbol{q}$ 发挥选择器作用，能够决定哪些输入信息（由 $\boldsymbol{k}$ 和 $\boldsymbol{v}$ 构成）是任务相关的。由 Softmax 生成的注意力图用来表征模型对相应部分输入的注意程度。当 $\boldsymbol{q}$ 和 $\boldsymbol{k}$、$\boldsymbol{v}$ 来自于相同输入时，称为自注意力，当来自于不同输入时，则称为交叉注意力。

在特征选择模块中，定义固定数量的可学习查询词 $\boldsymbol{Q}_i = [q_{i,1}, q_{i,2}, \cdots, q_{i,M}] \in \mathbf{R}^{M\times D}$，$M$ 为查询词数量。通过大量样本训练，期望 $\boldsymbol{Q}_i$ 学习出对 $\boldsymbol{H}_i$ 的选择性。在使用 $\boldsymbol{Q}_i$ 选择 $\boldsymbol{H}_i$ 前，先在 $\boldsymbol{Q}_i$ 之间使用自注意力机制，即

$$\boldsymbol{Q}_i \leftarrow \mathrm{MHA}(\boldsymbol{Q}_i,\boldsymbol{Q}_i,\boldsymbol{Q}_i) + \boldsymbol{Q}_i \tag{4}$$

通过式（4）自注意力过程，允许查询词在模型训练过程中，相互聚合各查询词的共享信息。应用 $\boldsymbol{Q}_i$ 通过交叉注意力机制从 $\boldsymbol{H}_i$ 中选择对视觉位置识别任务具有重要作用的局部特征，即

$$\boldsymbol{O}_i = \mathrm{MHA}(\boldsymbol{Q}_i,\boldsymbol{H}_i,\boldsymbol{H}_i) \tag{5}$$

式中　$\boldsymbol{O}_i$——基于查询词的特征选择模块的输出，$\in \mathbf{R}^{M\times D}$

该过程通过计算查询词组与场景图像局部特征序列 $\boldsymbol{H}_i$ 之间的注意力权重，动态评估各局部特征重要性，并自适应地选择对场景识别任务重要的局部特征，并将其聚合到输出 $\boldsymbol{O}_i$ 中。

1.4.3　全局描述符生成

通过聚合各个特征选择模块输出，生成场景图像全局描述符。通过级联操作 Concat，将各查询模块输出特征序列 $\boldsymbol{O}_i$ 连接为单一输出，即

$$\boldsymbol{O} \leftarrow \mathrm{Concat}(\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots,\boldsymbol{O}_l) \tag{6}$$

式中　$\boldsymbol{O}$——级联后的单一输出，$\in \mathbf{R}^{1\times lMD}$

级联各查询模块输出，能够使 $\boldsymbol{O}$ 包含网络中不同层的特征，具有更丰富多样的局部特征类型。为与所选取的特征提取器 DINOv2 结构相匹配，同时考虑计算复杂度，设置聚合器的超参数 l 为 2、聚合模块查询词数量 M 为 32。

$\boldsymbol{O}$ 直接级联了多个查询模块的输出，维度仍然过高。因此通过线性变换，对其进行降维，同时通过 L_2 归一化（L_2norm）对降维结果进行归一化处理，使最终生成的全局描述符 F 位于单位超球面内，即

$$F = L_2norm(\boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{O})) \tag{7}$$

式中　$\boldsymbol{W}_1$——第 1 层线性变化权重矩阵，$\in \mathbf{R}^{D_1\times lMD}$

$\boldsymbol{W}_2$——第 2 层线性变化权重矩阵，$\in \mathbf{R}^{D_2\times D_1}$

以 F 作为场景图像的全局描述符，通过不同图像 F 间的比对，实现基于场景图像视觉位置识别。

1.5　模型损失函数定义

提出的温室场景识别模型通过预训练的特征提取网络获取图像局部特征序列，并利用所设计的特征聚合器生成全局描述符。模型需经过训练才能用于描述符生成，训练的核心目标是使同类场景描述

符在特征空间中相互靠近,异类场景描述符彼此远离。为此,将多重相似性损失函数(Multiple-similarity loss, MS Loss)[27-28]的最小化作为模型训练目标,该函数综合对比损失[29]和三元组损失[30]的优势,使用动态加权机制,同时考虑样本对的绝对距离和批次内整体分布特征,让度量学习更全面地覆盖样本特性。MS Loss 定义为

$$L_{MS}=\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{\alpha}\lg\left(1+\sum_{k\in P_i}e^{-\alpha(S_{ik}-\lambda)}\right)+\frac{1}{\beta}\lg\left(1+\sum_{k\in N_i}e^{\beta(S_{ik}-\lambda)}\right)\right] \tag{8}$$

式中 L_{MS}——多重相似性损失

S_{ik}——样本对全局描述符间的相似度

α、β、λ——损失超参数

m——训练批次中的样本图像数量

P_i——小批次样本中相对于样本 i 正样本(同类场景)索引集合

N_i——小批次样本中相对于样本 i 负样本(异类场景)索引集合

MS Loss 通过两阶段优化策略实现对样本的精细化学习。首先进行困难样本挖掘,识别具有判别价值的样本对;随后执行自适应加权,综合考虑样本自身的相似性度量及其在批次中的相对分布关系。该双重机制不仅充分挖掘了局部样本分布的特征信息,还能动态调整不同样本对的贡献权重,使模型更加聚焦于具有关键判别力的样本对,提升特征学习有效性。

1.6 模型训练与评估方法

基于 PyTorch 2.7.1 深度学习框架实现温室视觉位置识别模型。训练和测试用计算机配置为,Nvidia L20 GPU(48 GB 显存)、Intel Xeon Silver 4314 CPU(2.4 GHz 主频)、128 GB 内存,运行 Windows Server 2022 操作系统,并借助 CUDA 12.8 加速计算。

训练过程中,将输入图像分辨率统一调整为 322 像素 × 322 像素,应用 AdamW 优化器,初始学习率设置为 2×10^{-4},权重衰减系数为 10^{-3}。训练中动态调整学习率,每 5 轮迭代后,将学习率衰减至原值的 30%,并设置 3 轮线性预热阶段。总训练轮次为 50 轮,确保模型充分收敛。应用数据增广技术[31],包括随机旋转、颜色抖动等,对场景图像进行变换处理,提升模型泛化能力。

用测试集评估训练后模型性能,评估指标为 top-K 召回率($R@K$)[32],其定义为查询图像检索回的前 K 个最相似参考图像中包含真实匹配场景的概率。测试过程中,采用余弦相似度作为距离度量,且不采用任何数据增广技术。$R@K$ 计算式为

$$R@K=\frac{M_K}{X}\times100\% \tag{9}$$

式中 X——查询图像个数

M_K——保留前 K 个查询结果前提下,场景识别成功的样本数

2 结果与分析

2.1 模型对比

为系统评估本文方法在温室场景识别任务中的性能,开展对比试验。将本文方法与当前主流的视觉场景方法,包括 NetVLAD[8]、GeM[33]、CosPlace[34]、EigenPlaces[35]、MixVPR[11] 和 SALAD[36],在相同的温室场景数据集上进行对比测试,并采用相同的评估标准。所有对比方法均采用其原始文献推荐的网络结构和超参数设置,并使用与本试验相同的数据划分方式进行训练和测试,结果如表 1 所示。

表 1 不同方法场景识别性能对比

Tab.1 Performance benchmarking for scene identification across methods %

方法	$R@1$	$R@5$	$R@10$
NetVLAD	65.09	78.22	83.59
GeM	69.55	79.38	83.26
CosPlace	78.15	89.28	92.62
EigenPlaces	85.49	94.83	96.45
MixVPR	84.89	94.89	96.99
SALAD	88.35	95.79	97.59
本文方法	88.79	96.49	97.96

注:$R@1$ 等表示 top-1 召回率。下同。

由表 1 可知,本文方法 $R@1$、$R@5$ 和 $R@10$ 均优于对比方法,达到 88.79%、96.49% 和 97.96%,表明选取的特征提取骨干网络及所提出的特征聚合机制是有效的。特征聚合机制引入基于查询词的场景图像局部特征选择策略,通过交叉注意力机制,特征选择策略经大量样本训练后,能够根据任务选择最相关的局部特征,而 NetVLAD 的固定聚类中心无法适应多样化的动态场景。与 GeM 方法相比,本文方法 3 个指标分别提高 19.24、17.11、14.70 个百分点。GeM 方法通过参数化全局池化实现特征聚合,在池化过程丢失了大量具有判别性的局部信息,本文方法在利用特征选择机制可选择最具判别性的多个局部特征,并通过不同层级特征的进一步聚合生成描述符,使描述符更具判别力。与 CosPlace 和 EigenPlaces 相比,本文方法展现出更优异的特征表达能力。本文方法在 3 个指标上分别比 CosPlace 提高 10.64、7.21、5.34 个百分点,比 EigenPlaces 提高

3.30、1.66、1.51 个百分点。结果表明，本文方法所得的全局描述符能够更全面地保留场景语义信息，具备更强的泛化能力。与 MixVPR 和 SALAD 方法相比，本文方法仍具有性能优势，3 个指标分别比 MixVPR 提高 3.90、1.60、0.97 个百分点，比 SALAD 提高 0.44、0.7、0.37 个百分点。本文方法通过所设计的特征聚合机制，在保留关键局部信息的同时，有效建立了特征间空间相关性，实现了更优的全局-局部特征平衡。总体上，提出的基于查询词选择的局部特征聚合机制可生成具有高度判别性的场景描述符，可为温室场景识别提供新的解决方案。

2.2　特征提取器对比

为验证选取的 DINOv2 特征提取器在温室场景识别任务中的有效性，选取 ResNet50[37] 和 EfficientNet－B0[38] 作为基准骨干网络，与 DINOv2 骨干网络进行对比试验。3 种骨干网络均与本文特征聚合器相结合，构建完整的场景识别模型。3 种模型训练策略和评估方法完全一致，结果如表 2 所示。

表 2　特征提取器性能对比

Tab.2　Feature extractor performance comparison

%

骨干网络	*R*@1	*R*@5	*R*@10
ResNet50	81.30	92.36	95.46
EfficientNet－B0	78.93	89.73	92.62
DINOv2	88.79	96.49	97.96

由表 2 可知，与 ResNet50 相比，本文模型 *R*@1、*R*@5 和 *R*@10 分别提升 7.49、4.13、2.5 个百分点；相较于 EfficientNet－B0，提升幅度更为明显，分别达到 9.86、6.76、5.34 个百分点。温室场景中作物空间分布和环境光照条件等全局上下文信息对识别至关重要。ResNet50 的 CNN 架构，受限于局部感受野的逐步扩展机制，虽然通过下采样操作可以扩大感受野范围，但在捕捉长距离依赖关系方面仍存在不足，导致部分全局信息丢失。相比之下，DINOv2 基于 Transformer 的自注意力机制能够直接建立图像块之间的全局关系，更有效地捕获温室场景中关键上下文特征。EfficientNet－B0 虽然通过神经网络架构搜索实现了模型轻量化，但其特征提取能力在复杂的温室场景中表现出明显的局限性。温室环境具有场景相似度高、作物特征变化细微等特点，这对网络特征判别能力提出了更高要求。DINOv2 通过预训练学习了通用视觉特征表达能力，在温室视觉位置识别任务上表现出更优的识别性能。总体上，DINOv2 可用于温室场景识别任务的图像局部特征提取。

2.3　模型对多因素变化的适应性

针对温室环境特有的光照多变、空间受限、作物动态生长等复杂条件，设计了鲁棒性验证试验，分析模型对光照变化、视角变化、距离变化、作物生长变化的适应性。

2.3.1　光照强度对场景识别的影响

采用不同光照条件采集的场景样本，对模型场景识别性能进行测试。将光照条件划分为 3 个典型等级，晴朗天气 08:00—10:00 时段为正常光照，晴朗天气 11:00—13:00 正午时段为强光照，多云或雪天 16:00—18:00 时段为弱光照。每种光照条件分别包含 100 组场景的测试样本，结果如表 3 所示。

表 3　光照强度变化对场景识别的影响

Tab.3　Effect of lighting changes on place identification accuracy

%

光照条件	*R*@1	*R*@5	*R*@10
正常光照	88.81	97.96	99.71
强光照	83.86	90.08	95.78
弱光照	84.92	93.92	97.47

由表 3 可知，在正常光照条件下，模型识别性能最好，*R*@1、*R*@5 和 *R*@10 分别达到 88.81%、97.96% 和 99.71%。在极端光照条件下，模型仍具有稳定的识别能力。相比正常光照条件，强光照下，*R*@1 下降 4.95 个百分点，弱光照下，*R*@1 下降 3.89 个百分点，波动幅度均控制在 5 个百分点以内。性能差异产生机制可归因于不同光照条件下图像质量变化。正常光照条件下，图像色彩还原准确，对比度适中，纹理特征丰富，为深度特征提取提供了理想条件。强光照导致图像高光区域细节丢失，特征难以提取。弱光照造成信噪比降低，纹理细节模糊，增加了特征匹配难度。尽管在强光和弱光下，特征提取存在挑战，得益于预训练 DINOv2 的良好视觉特征表达，即动态特征聚合机制，模型对不同光照条件下的场景识别具有一定的鲁棒性。

2.3.2　观测角度变化对模型性能的影响

采用带刻度的旋转载物台（刻度精度 ±0.5°）固定相机，以正视场景 0°视角为基准，按 1°间隔旋转至 45°采集多角度图像，构建包含 56 组场景的评估样本。以 0°图像为查询图像，检验其对不同视角参考图像的匹配表现，定量分析视角变化对识别效果的影响，结果如表 4 所示。

由表 4 可知，模型小视角变化范围 0°～15°内，保持优异识别性能，*R*@1 达 93.12%。当视角增至 15°～30°时，*R*@1 下降至 71.35%，降幅达 21.77 个百分点。性能下降主要原因为：视角增大导致场景

表 4 观测角度变化对场景识别的影响

Tab. 4 Effect of viewpoint variation on place identification accuracy %

视角/(°)	R@1	R@5	R@10
0 ~ 15	93.12	97.65	98.48
15 ~ 30	71.35	88.61	92.17
30 ~ 45	58.93	70.79	78.76

重叠区域减少,物体间遮挡效应增强,具有判别性的共享特征减少。当视角变化超过 30°后,性能衰减加剧,R@1 下降到 58.93%。主要是因为原场景可视部分减少,场景表观特征发生本质变化,背景结构产生明显位移。总体上,在视角变化小于 15°时,模型可保持良好性能,在视角 15° ~ 30°范围内模型仍具有实用价值,超过 30°时,性能急剧下降。本文方法在适度视角变化下具有鲁棒性。

2.3.3 距离对模型识别性能的影响

采用带刻度的滑轨(刻度精度 ±0.1 mm)控制相机位移,以距离株行 50 cm 处为基准位置,固定视角,按 5 cm 间隔逐步增大至 100 cm,构建包含 100 组场景的评估样本集。以 50 cm 处图像为查询图像,测试不同距离参考图像的匹配性能,分析距离对识别精度的影响,结果如表 5 所示。

表 5 距离变化对场景识别性能的影响

Tab. 5 Effect of distance variation on place identification accuracy %

距离/cm	R@1	R@5	R@10
50 ~ 65	97.16	99.24	100
65 ~ 80	80.21	88.14	92.91
80 ~ 100	63.94	80.82	83.91

由表 5 可知,模型性能随距离增加呈下降趋势。当距离为 50 ~ 65 cm 时,模型具有最优性能 R@1 达 97.16%。当距离增至 65 ~ 80 cm 时,R@1 下降至 80.21%,降幅达 16.95 个百分点;距离继续扩大至 80 ~ 100 cm 时,R@1 降至 63.94%。根据针孔相机工作原理,距离倍增导致视场面积扩大 4 倍,使得场景目标表观尺寸缩小 50%,有效特征密度降低,新增干扰物进入视野,新增特征干扰了全局描述符的比对。另一方面,即使是在 2 倍基准距离条件下,模型 R@1 仍保持 63.94%。本文方法在应对距离变化方面具有较强的鲁棒性,能够在一定程度上适应智能农机作业过程的距离波动。

2.3.4 模型对场景动态变化的适应性

试验共采集 58 组番茄坐果期的温室场景图像,时间跨度为 13 d,分别在第 1 天、第 5 天、第 9 天、第 13 天采集各场景图像。通过固定第 1 天图像作为参考集,分别测试第 5 天、第 9 天和第 13 天查询图像匹配性能,分析模型在作物生长干扰下的场景识别稳定性,结果如表 6 所示。

表 6 作物动态生长变化对场景识别的影响

Tab. 6 Effect of dynamic crop growth variations on place identification accuracy %

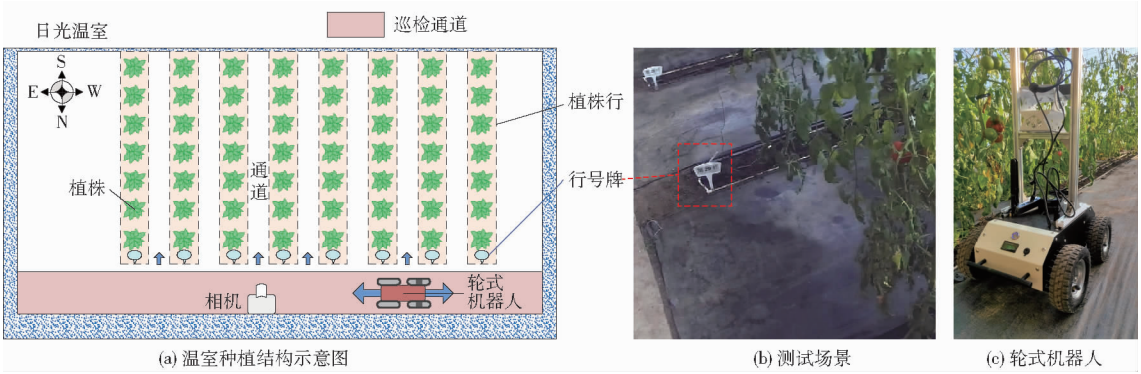
生长时间/d	R@1	R@5	R@10
5	61.14	68.82	76.10
9	53.78	59.52	64.28
13	34.42	42.62	52.45

由表 6 可知,模型性能随生长时间呈衰减趋势。在第 5 天时,模型还能保持较好的识别能力,R@1、R@5 和 R@10 还能保持在 61.14%、68.82% 和 76.10%,表明模型对作物初期形态变化具有一定适应性。至第 9 天时,R@1 降至 53.78%,至第 13 天时,R@1 下降至 34.42%。这一性能衰减主要源于作物生长及农事管理导致的场景变化。第 5 天时番茄果实膨大,并出现叶片黄化,第 9 天时果实转色并对老叶进行了梳理;第 13 天时部分果实成熟。整个试验期内,场景因作物生长发生了较大的变化,而温室场景图像又主要以植株部分为主,进而使场景局部特征发生较大的改变,全局描述符变化导致召回率下降。值得注意的是,模型在短期(不大于 5 d)生长变化中表现出较强的鲁棒性(R@1 大于 60%)。这得益于模型稳定特征提取能力,以及基于查询词进行特征选择的机制对稳定特征的筛选。然而,长期生长导致的表观特征本质变化仍是场景识别面临的挑战。综合而言,温室视觉场景识别模型对于短期作物生长变化具备一定的鲁棒性。

2.4 温室移动机器人视觉定位试验

选择一个种植相同作物的另一独立温室(图 4)开展基于场景识别的视觉定位实地验证试验。在温室巡检通道(位于植株行首)和植株行间开展定位试验。首先使用三脚架固定相机(高度 1.5 m,俯角 15°),沿东西向巡检通道在作物行首拍摄图像,包含植株行首行号牌,以其作为位置已知的参考图像。在随机选定 4 个植株行间分别采集参考图像,并以番茄植株先后顺序作为图像位置参考。

采用轮式移动机器人平台(移动速度 0.2 ~ 1.0 m/s)搭载 RGB 相机(安装高度 1.5 m ±5 cm,俯角 15° ±2°)在不同光照条件下沿巡检通道进行 9 组重复试验。每组试验控制机器人沿同一巡检路径匀速行进,同时以帧率 15 f/s 采集连续视频(分辨率 1 920 像素 ×1 080 像素)。考虑实际作业中设备安装误差,每组试验随机调整相机位姿参数。用机器人在随机选定的 4 个株行间开展定位试验。通过视频帧提取构建各试验组别的独立查询集(每组包含



(a) 温室种植结构示意图　　(b) 测试场景　　(c) 轮式机器人

图 4　温室移动机器人视觉定位试验

Fig. 4　Vision-based localization test for greenhouse mobile robots

312 ~ 558 帧)。应用本文模型进行视觉位置识别,在巡检通道进行的各组试验,以行号牌(图 4b)匹配(由人工判读)作为定位成功的判定标准,即通过查询图像召回的参考图像中出现了与查询图像相同的行号牌;在植株行间进行定位试验,则依据作物形态由人工判定查询图像与参考图像是否匹配。最终以识别准确率(成功匹配帧数/总帧数)作为评估指标,结果如表 7 所示。

表 7　视觉定位试验结果

Tab. 7　Results of visual localization tests

视频号	试验场景	光照条件	总帧数/帧	识别帧数/帧	识别准确率/%
1	行首	正常光	349	301	86.25
2	行首	强光照	382	324	84.81
3	行首	弱光照	377	312	82.75
4	行首	正常光	446	405	90.80
5	行首	强光照	473	418	88.37
6	行首	弱光照	454	369	81.27
7	行首	正常光	509	437	85.85
8	行首	强光照	558	459	82.26
9	行首	弱光照	535	446	83.36
10	行间	正常光	312	285	91.34
11	行间	正常光	320	261	81.56
12	行间	正常光	480	402	83.75
13	行间	正常光	520	451	86.73

由表 7 可知,本文方法在实际生产环境中展现出良好的实用性能。9 次行首重复试验显示,模型识别准确率为 81.27% ~ 90.80%,平均识别率 85.88%。不同光照条件下识别率表现稳定,正常光、强光照和弱光照的平均值分别为 87.63%、85.15% 和 82.46%,表明光照变化对模型性能影响不显著,系统具有良好的光照鲁棒性。在 4 个不同植株行间场景中,模型平均识别准确率为 85.85%,尽管存在更复杂的背景干扰和植株遮挡,仍保持较高识别精度,验证了对行间环境的适应能力。结果表明本文方法具有较强的泛化能力和实用性。此外,在机器人不同行进速度下,识别率保持稳定(标准差小于 3.5%),速度对识别性能的影响不明显,相机位姿的随机扰动也未对模型性能产生显著影响。图 5 为行间定位试验中当前帧匹配的参考图像,表明本文模型能够较好地召回具有位置信息的参考图像。总体而言,本文提出的视觉定位方案能够满足智能农机在实际作业环境中的可靠性要求,为解决温室自主导航中的定位问题提供了可选技术途径。

3　结论

(1)所提出的基于特征选择与聚合的温室场景识别方法模型 $R@1$、$R@5$ 和 $R@10$ 分别达到 88.79%、96.49% 和 97.96%,性能明显优于 NetVLAD、GeM、CosPlace 等基准方法,与模型



实时采集的查询帧　　匹配的参考图像

图 5　查询帧匹配与参考图像的匹配示例

Fig. 5　Query frame and reference image matching examples

MixVPR 和 SALAD 相比也具有一定的优势。

(2)模型在强光和弱光条件下 $R@1$ 下降均未超过 5 个百分点,表明 DINOv2 预训练特征和特征选择聚合的描述符生成机制对光照变化具有较强的适应性。视角变化在 15°以内时,$R@1$ 保持在 93.12%;距离变化在 2 倍范围内,$R@1$ 为 63.94%。

(3)基于特征选择与聚合的温室场景识别方法,通过 DINOv2 骨干网络和可学习查询词特征选择聚合机制,有效提升了温室场景识别精度,为智能农机视觉系统设计提供了可靠的技术方案选择。模型在光照、视角和距离变化下表现出较强的鲁棒性,但在长期作物生长场景中性能有待优化。

参 考 文 献

[1] 冀荣华, 王文轩, 安冬, 等. MPMS-SGH:面向日光温室的多参数多步预测模型[J]. 农业机械学报, 2025, 56(7): 265-278.
JI Ronghua, WANG Wenxuan, AN Dong, et al. MPMS-SGH: multi-parameter multi-step prediction model for solar greenhouse[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(7): 265-278. (in Chinese)

[2] 李天华, 董广胜, 姚玉康, 等. 连栋温室分段变距喷雾机器人设计与试验[J]. 农业机械学报, 2024, 55(2): 170-179.
LI Tianhua, DONG Guangsheng, YAO Yukang, et al. Design and experiment of segmented and variable distance spraying robot for multi-span greenhouse[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(2): 170-179. (in Chinese)

[3] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[4] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[5] GALVEZ-LOPEZ D, TARDOS J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.

[6] SÜNDERHAUF N, SHIRAZI S, ADAM J, et al. Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free[C]// Proceedings of Robotics: Science and Systems XI, 2015: 1-10.

[7] JEGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Rewgnition (CVPR), 2010: 3304-3311.

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

[9] KHALIQ A, MILFORD M, GARG S. MultiRes-NetVLAD: augmenting place recognition training with low-resolution imagery [J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3882-3889.

[10] 张冬灿, 张国良, 李俊学, 等. 基于 HOG 特征与卷积神经网络的闭环检测算法[J]. 四川轻化工大学学报(自然科学版), 2022, 35(4): 49-55.
ZHANG Dongcan, ZHANG Guoliang, LI Junxue, et al. Loop closure detection algorithm based on HOG feature and convolutional neural networks[J]. Journal of Sichuan University of Science & Engineering (Natural Science Edition), 2022, 35(4): 49-55. (in Chinese)

[11] AMAR A, BRAHIM C, PHILIPPE G. MixVPR: feature mixing for visual place recognition[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023: 2998-3007.

[12] KEETHA N, MISHRA A, KARHADE J, et al. AnyLoc: towards universal visual place recognition[J]. IEEE Robotics and Automation Letters, 2024, 9(2): 1286-1293.

[13] HAUSLER S, GARG S, XU M, et al. Patch-NetVLAD: multi-scale fusion of locally-global descriptors for place recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14136-14147.

[14] LU F, ZHANG L, LAN X, et al. Towards seamless adaptation of pre-trained models for visual place recognition[C]// Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), 2024.

[15] 张哲, 包德伟, 陶亮, 等. 一种融合时序特征的网络场景识别算法[J]. 重庆理工大学学报(自然科学), 2024, 38(3): 141-149.
ZHANG Zhe, BAO Dewei, TAO Liang, et al. A network scene recognition algorithm that integrates temporal feature[J]. Journal of Chongqing University of Technology (Natural Science), 2024, 38(3): 141-149. (in Chinese)

[16] 李建辛, 司冠南, 田鹏新, 等. 多模态知识图谱的 3D 场景识别与表达方法综述[J]. 计算机工程与应用, 2023, 59(20): 35-50.
LI Jianxin, SI Guannan, TIAN Pengxin, et al. Survey of 3D scene recognition and representation methods of multimodal knowledge[J]. Computer Engineering and Applications, 2023, 59(20): 35-50. (in Chinese)

[17] 刘铁, 段勇. 融合 CNN 和 Transformer 的机器人室内场景识别[J]. 电子测量与仪器学报, 2023, 37(5): 223-229.
LIU Tie, DUAN Yong. Robot indoor scene recognition based on fusion of CNN and Transformer[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(5): 223-229. (in Chinese)

[18] 于美玲，周云成，侯玉涵，等. 基于局部图像特征聚合的温室场景识别方法[J]. 农业机械学报，2025，56(2)：485－494.
YU Meiling，ZHOU Yuncheng，HOU Yuhan，et al. Greenhouse scene recognition method based on local image feature aggregation[J]. Transactions of the Chinese Society for Agricultural Machinery，2025，56(2)：485－494.（in Chinese）

[19] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：transformers for image recognition at scale[C]// Proceedings of the 9th International Conference on Learning Representations（ICLR 2021），2021：1－22.

[20] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition（CVPR），2016：2818－2826.

[21] HUANG G，LIU Z，VAN DER MAATEN L，et al. Densely connected convolutional networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR），2017：2261－2269.

[22] HOWARD A，SANDLER M，CHEN B，et al. Searching for MobileNetV3[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision（ICCV），2019：1314－1324.

[23] MA N N，ZHANG X Y，ZHENG H T，et al. ShuffleNet V2：practical guidelines for efficient CNN architecture design[C]// Proceedings of the 2018 European Conference on Computer Vision（ECCV），2018：122－138.

[24] TAN M X，CHEN B，PANG R M，et al. MnasNet：platform-aware neural architecture search for mobile[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR），2019：2815－2823.

[25] OQUAB M，DARCET T，MOUTAKANNI T，et al. DINOv2：learning robust visual features without supervision[EB/OL].（2023－11－15）[2025－09－14]. https://arxiv.org/abs/2304.07193.

[26] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30（NIPS 2017），2017：5998－6008.

[27] 申秋慧，张宏军，徐有为，等. 知识图谱嵌入模型中的损失函数研究综述[J]. 计算机科学，2023，50(4)：149－158.
SHEN Qiuhui，ZHANG Hongjun，XU Youwei，et al. Comprehensive survey of loss functions in knowledge graph embedding models[J]. Computer Science，2023，50(4)：149－158.（in Chinese）

[28] WANG X，HAN X，HUANG W，et al. Multi-similarity loss with general pair weighting for deep metric learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：5022－5030.

[29] HADSELL R，CHOPRA S，LECUN Y. Dimensionality reduction by learning an invariant mapping[C]// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition（CVPR），2006：1735－1742.

[30] HOFFER E，AILON N. Deep metric learning using triplet network[C]// Proceedings of the 5th International Workshop on Similarity-Based Pattern Recognition，2015：84－92.

[31] 林成创，单纯，赵淦森，等. 机器视觉应用中的图像数据增广综述[J]. 计算机科学与探索，2021，15(4)：583－611.
LIN Chengchuang，SHAN Chun，ZHAO Gansen，et al. Review of image data augmentation in computer vision[J]. Journal of Frontiers of Computer Science and Technology，2021，15(4)：583－611.（in Chinese）

[32] SCHROFF F，KALENICHENKO D，PHILBIN J. FaceNet：a unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2015：815－823.

[33] LOPEZ-PAZ D，RANZATO M. Gradient episodic memory for continual learning[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems（NIPS 2017），2017：6470－6479.

[34] BERTON G，MASONE C，CAPUTO B. Rethinking visual geo-localization for large-scale applications[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR），2022：4868－4878.

[35] BERTON G，TRIVIGNO G，CAPUTO B，et al. EigenPlaces：training viewpoint robust models for visual place recognition[C]// Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision（ICCV），2023：11046－11056.

[36] IZQUIERDO S，CIVERA J. Optimal transport aggregation for visual place recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR 2024），2024：17658－17668.

[37] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(5)：901－912.

[38] TAN M X，LE Q V. EfficientNet：rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning（ICML 2019），2019：10691－10700.