doi:10.6041/j.issn.1000-1298.2025.02.045

基于局部图像特征聚合的温室场景识别方法

于美玲　周云成　侯玉涵　刘峻淳
(沈阳农业大学信息与电气工程学院, 沈阳 110866)

摘要: 场景识别可作为温室环境空间定位的替代方案,也是智能农机装备视觉系统的重要功能之一。针对以特征聚类为基础的场景识别范式无法适应高动态变化且高度相似的温室场景识别的问题,提出一种基于深度特征聚合的温室场景识别方法,以预训练的视觉 Transformer 网络为基础,提取场景图像局部特征,应用多层感知机全局感受野特性,考虑局部特征空间关系,融合图像局部特征,生成场景图像全局描述子,以多重相似性损失最小化为优化目标,构建温室场景识别模型。试验结果表明,模型场景识别 R@1(top-1 召回率)、R@5 和 R@10 分别达到 78.43%、89.21% 和 92.47%,具有较高的场景识别精度。所提出的基于多层感知机的特征混合方法是有效的,与采用池化操作进行特征聚合相比,R@1 提高 8.01 个百分点。模型对光照条件变化具有一定的鲁棒性,与正常的中等光照条件相比,强光及弱光条件下,R@1 下降未超过 4.00 个百分点。相机视角及采样距离的变化也会影响模型识别性能,20°以内的视角变化,R@1 下降 6.61 个百分点,2 倍以内的距离变化,R@1 下降 17.87 个百分点。与现有场景识别基准方法 NetVLAD、GeM、Patch-NetVLAD、MultiRes-NetVLAD 和 MixVPR 相比,R@1 分别提高 7.82、6.59、3.56、4.14、1.88 个百分点,在温室场景识别任务上模型性能有较大提升。该研究构建的基于多层感知机的图像全局特征聚合方法,能够生成可靠的全局描述子,用于温室场景识别,且具有一定的光照、视角、距离及时间变化的鲁棒性,研究结果可为智能农机视觉系统设计提供技术参考。

关键词: 日光温室; 场景识别; 特征聚合; 视觉 Transformer; 图像处理
中图分类号: S126; TP183　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)02-0485-10　　**OSID**:

Greenhouse Scene Recognition Method Based on Local Image Feature Aggregation

YU Meiling　ZHOU Yuncheng　HOU Yuhan　LIU Junting
(*College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110866, *China*)

Abstract: Scene recognition could be used as an alternative for spatial positioning in greenhouse environments, and it was also one of the important functions of the visual system of intelligent agricultural machinery equipment. Addressing the issue that scene recognition paradigms based on feature clustering could not adapt to the recognition of greenhouse scenes with high dynamic changes and high similarity, a greenhouse scene recognition method based on deep feature aggregation was proposed. This method, grounded on a pre-trained visual transformer network, extracted local features from scene images. It applied the global receptive field characteristics of multi-layer perceptron, took into account the spatial relationships of local features, fused the local features of the images, and generated global descriptors for the scene images. With the goal of minimizing multi-similarity loss as the optimization objective, a greenhouse scene recognition model was constructed. The test results indicated that the R@1 (top-1 recall rate), R@5, and R@10 of the model's scene recognition reached 78.43%, 89.21%, and 92.47%, respectively, and it possessed high scene recognition accuracy. The proposed feature mixing method based on multi-layer perceptron was proven effective, with an improvement of 8.01 percentages in R@1 compared with that of feature aggregation using pooling operations. The model demonstrated a certain robustness to changes in lighting conditions, with the R@1 metric decreased by no more than

收稿日期: 2024-09-18　修回日期: 2024-10-28
基金项目: 国家重点研发计划项目(2021YFD1500204、2023YFD1501303)
作者简介: 于美玲(2002—),女,硕士生,主要从事智慧农业研究,E-mail: 1048615322@qq.com
通信作者: 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

4.00 percentages under strong and weak lighting conditions compared with that under normal medium lighting conditions. Changes in camera angle and sampling distance also impacted the model's recognition performance, with a decline of 6.61 percentages for angle changes within 20 degrees, and a drop of 17.87 percentages for distance changes within twice the original distance. Compared with the existing scene recognition benchmark methods, including NetVLAD, GeM, Patch - NetVLAD, MultiRes - NetVLAD, and MixVPR, the R@1 of proposed model was improved by 7.82, 6.59, 3.56, 4.14, and 1.88 percentages, respectively, demonstrating a significant performance enhancement on the greenhouse scene recognition task. The image global feature aggregation method based on multi-layer perceptron constructed was able to generate reliable global descriptors for greenhouse scene recognition, and exhibited robustness to changes in lighting, viewpoint, distance, and time. The research findings would provide technical references for the design of visual systems for intelligent agricultural machinery.

Key words: solar greenhouse; scene recognition; feature aggregation; vision Transformer; image processing

0 引言

日光温室等设施农业在我国北方地区取得了长足发展,为农民增收和解决居民菜篮子问题做出了巨大贡献。随着城镇化和人口老龄化,农业劳动力短缺问题日益凸显。受工作环境约束,大型农机设备难以应用于设施生产,发展小型化温室智能装备[1]对解决劳动力短缺具有重要意义。场景识别是指应用传感器采集的当前场景图像(查询图像),通过计算机视觉技术,提取其特征,并与先前已获取场景图像(参考图像)的特征进行比较,实现场景辨识的过程。场景识别是温室智能农机装备视觉系统的重要功能之一,当参考图像位置已知时,场景识别可用于工作环境中的空间定位,如果位置未知,场景识别也可用于回环检测[2],避免在同一场景反复作业。开展温室场景识别方法研究,对促进智能农机装备发展具有一定意义。

近年来,国内外学者在视觉场景识别(Visual place recognition, VPR)方法上开展了一定的研究工作。SÜNDERHAUF 等[3]指出,稳健的 VPR 系统必须满足 3 项核心要求,即光照条件不变性、视角不变性以及通用性。ARANDJELOVIC 等[4]基于深度神经网络(Deep neural network, DNN)特征,提出特征向量局部聚合描述符(Vector of locally aggregated descriptors, VLAD)编码方法,构建了一种支持端到端学习的场景识别网络模型。CHEN 等[5]将场景识别任务视作图像分类问题,在标准场景环境数据集上对网络进行了深入训练,实现了场景识别性能的提升。CHENG 等[6]提出一种基于语义聚合描述子向量(Vectors of semantically aggregated descriptors, VSAD)的方法,该方法在标准场景识别任务中表现出了良好的性能。ZHAO 等[7]设计了一种网络判别模型,通过提取多尺度特征,生成输入图像的映射判别特征。张学典等[8]构建一种基于 VGG16[9]的快速场景识别算法,以解决深度卷积神经网络(Deep convolution neural network, DCNN)在场景识别中计算时间随数据空间尺度增长的问题。张冬灿等[10]提出了一种融合方向梯度直方图(Histogram of oriented gradients, HOG)和 CNN 的场景识别算法,通过结合 HOG 几何信息和卷积特征,强化图像描述子对于视角变化的鲁棒性,提高了算法在外观变化场景下的稳定性。目前,场景识别研究大多集中在静态室内或单一室外环境上[11-13]。随作物生长及采摘、采收等多种生产管理活动,温室场景处在高度的变化之中。温室建筑墙体及植株形态、颜色、纹理单一,不同场景高度相似。这些因素给温室场景识别带来极大困难,基于深度神经特征的视觉场景识别方法在温室环境下的适用性值得进一步探讨。

针对智能农机视觉系统对场景识别的实际需求,及温室场景高动态变化、高度相似的特点,提出一种基于深度特征融合的温室场景识别方法,以预训练的深度特征提取网络为基础,提取场景图像局部特征,应用多层感知机全局感受野特性,考虑局部特征空间关系,融合图像局部特征,生成场景图像全局描述子,构建温室场景识别模型,以多重相似性损失最小化为优化目标,训练识别模型。在种植作物为番茄的温室场景中开展测试试验,验证方法的有效性,以期为温室智能农机装备视觉系统设计提供参考。

1 材料与方法

1.1 温室场景图像数据集构建

于 2023 年 11—12 月,在沈阳市辽中区某温室内,应用 Huawei Mate 40 Pro 型智能手机采集温室场景图像。温室种植作物为番茄,处于坐果期,株高 1.0~1.8 m,行距约 0.7 m,株距约 0.3 m。分别于晴天、多云或雪天 08:00—10:00、11:00—13:00、16:00—18:00 不同时间段采集图像。在温室内,采

集多个场景的图像。农业机器人等智能农机移动过程中，会在不同视角、距离、光照条件下观察到同一场景，为模拟该特性，针对不同场景，分别按不同视角、不同距离，在不同光照条件下成像，间隔一定天数对同一场景成像，构建视角、距离、光照、时间变化的场景数据集。针对视角变化，用三脚架固定相机拍摄位置，0 ~ 20°范围向左右旋转载物台，同时成像，采集具有视角变化的场景图像。针对距离变换，将相机置于滑轨的滑块上，移动滑块，调节相机与前方场景的距离，同时成像，采集具有距离变化的图像。在不同天气的不同时间段对同一场景成像，采集具有光照变化的图像。同时，对同一场景，分别间隔不同天数采集图像，使同一场景的图像体现作物生长变化。先后共采集 1 180 个场景的 36 900 幅图像，每个场景约 30 幅图像，标识场景编号和图像编号，构建温室场景数据集。进一步按 7∶3将数据集划分为训练集和测试集，分别用于后续场景识别模型的训练和测试。

1.2　温室场景识别技术架构

在温室环境中，对于智能农机通过视觉传感器采集的当前场景图像 I（查询图像），首先用由深度神经网络构建的图像特征提取器 h，提取图像局部特征 H，即有 $H=h(I)$，进一步通过特征聚合器 g，将局部特征 H 聚合为图像的全局描述子 F，即有 $F=g(H)$。对于已完成采集的历史参考图像，同样提取各图像的全局描述子，并存储于参考数据库。以查询图像的特征描述子作为检索条件，通过相似度比较，从参考数据库中检索最相似的参考图像，当相似度超过设定阈值时，则认为参考图像与查询图像来自同一场景。该过程即为温室场景识别的基本技术流程，如图 1 所示。当预先采集的参考图像已与温室环境具体位置绑定时，场景识别即可用于智能农机装备的空间定位。如果参考图像是机器行进过程中采集的，预先未与具体位置绑定，该技术也可用于判断机器是否回到了曾经到过的场景。

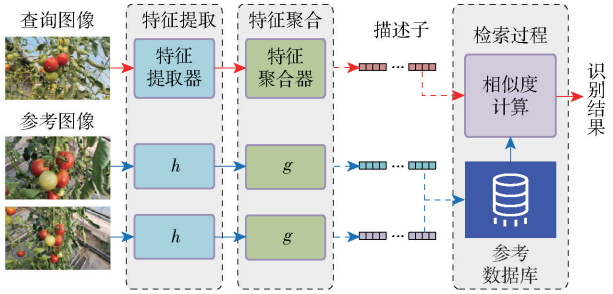


图 1　温室场景识别基本技术流程

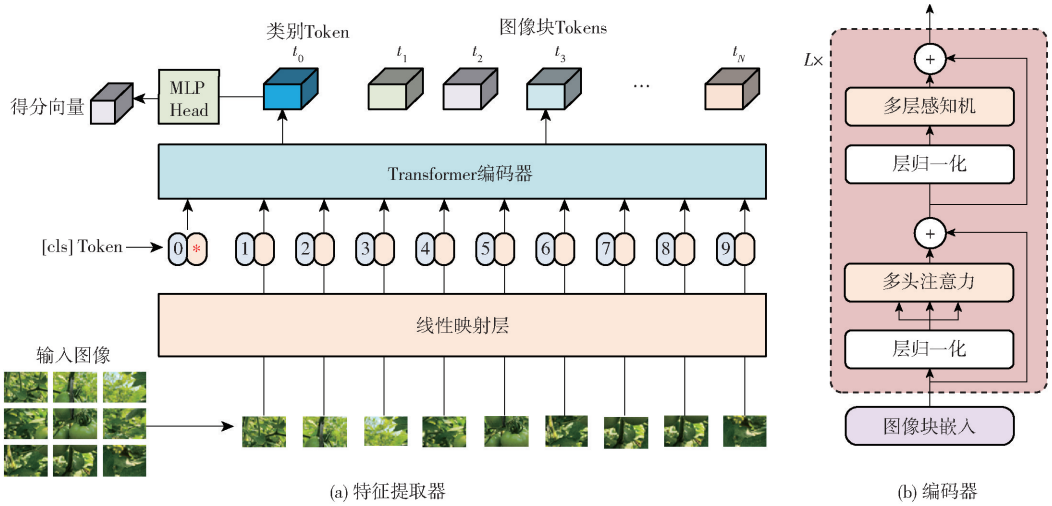
Fig. 1　Basic technical process of greenhouse scene recognition

图 1 技术框架中的特征提取器 h 和特征聚合器 g 共同构成温室场景识别模型 f，即有 $f=g(h(I))$，场景识别模型的性能将由特征提取器和特征聚合器共同决定。本研究用预训练的深度神经网络作为场景图像特征提取器 h，通过优化设计特征聚合器 g，将 h 提取的图像局部特征 H 融合为反映场景综合特征的全局描述子 F，以期提高温室场景识别模型性能。

1.3　场景图像局部特征提取器设计

视觉 Transformer（Vision Transformer，ViT）[14] 和 CNN[15-19] 都能够有效提取图像局部特征，二者在图像识别、目标检测、语义分割等多种计算机视觉任务中都得到了广泛应用。近几年，Transformer 在大模型中表现优异，本文采用 Transformer 构建温室场景图像特征提取器 h，其结构如图 2a 所示。

采用 ViT 构建特征提取器 h，首先将输入图像 $I\in\mathbf{R}^{c\times h_r\times w}$ 划分为大小为 $p\times p$ 的不重叠图像块，图像块数量共有 $N=h_r w/p^2$ 个，其中每个块都承载着图像的局部信息，将这些图像块展平为向量序列，随



(a) 特征提取器　　(b) 编码器

图 2　图像特征提取器结构图

Fig. 2　Image feature extractor

后将序列数据送入线性映射层,对每个展平后的图像块向量做线性变换,生成表示该图像块局部信息的向量(用 Token 表示)。在局部 Tokens 基础上,引入一个类别 Token(在图 2a 中表示为 *),用于聚合图像的全局特征表示。为了让模型学习到输入序列之间的空间位置信息和相关性,在 Tokens 上引入位置嵌入。所有 $N+1$ 个 Tokens 共同作为 Transformer 编码器的输入,编码器则由 L 个 Transformer 块(图 2b)堆叠而成,Transformer 块由 1 个多头注意力(Multi-head self-attention,MHSA)模块和 1 个多层感知机(Multi-layer perceptron, MLP)构成。参考 ViT[14],本文设定 $p=14$、$L=12$。Transformer 编码器输出 $N+1$ 个 Tokens,其中类别 Token 通过多层 Transformer 块中的 MHSA 处理,融合了图像的全局信息。进一步将由多个全连接层构成的 MLP 头作为分类器,并以类别 Token 作为输入,最终输出类别得分。

用 ImageNet 作为训练集,预训练大模型 DINOv2[20]作为教师网络,采用知识蒸馏方法将 DINOv2 的图像特征提取能力蒸馏到 h 上。蒸馏过程中,用 DINOv2 在 ImageNet 上输出的类别得分来监督 h 输出的类别得分。在特征提取器 h 完成预训练后,去除分类器部分,将 Transformer 编码器输出的 $N+1$ 个 Token 特征作为图像特征集,表示为

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

式中 t_i——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,通过后续特征聚合器聚合生成输入图像 I 的描述子 F。

1.4 基于特征混合的全局描述子生成

对于由特征提取器提取的图像特征集 $\boldsymbol{H}=\{t_i\}\in\mathbf{R}^{(N+1)\times D}$,如果 t_i 为图像块 Token,则其携带了图像局部块的深度特征,同时也通过 Transformer 的多头注意力机制,融合了图像的全局特征,如果 t_i 为类别 Token,则该 t_i 携带了图像全局深度特征。

直接将温室场景图像特征矩阵 $\boldsymbol{H}$ 展平(Flatten)成图像的全局描述子 $\boldsymbol{F}$,将使全局描述子的维度过高,描述子之间的比较将消耗过长的时间。同时,由于温室场景图像中存在着地面、墙体等难以区分的区域,因此 $\boldsymbol{H}$ 中也会存在大量的冗余特征。本研究应用多层感知机 MLP 的全局注意力特性,通过对其进行训练,使其学习出从 $\boldsymbol{H}$ 中选择有效特征的能力,并最终将有效特征融合为场景图像的全局描述子 $\boldsymbol{F}$。

应用由全连接(Fully connected, FC)层构建的 MLP 对 $\boldsymbol{H}$ 进行第 1 次融合,该过程表示为

$$\boldsymbol{Y}=\mathrm{MLP}(\boldsymbol{H})=\boldsymbol{W}_2(\sigma(\boldsymbol{W}_1\boldsymbol{H}))\quad(\boldsymbol{Y}\in\mathbf{R}^{(N+1)\times D}) \tag{2}$$

式中 $\boldsymbol{W}_1$——第 1 层全连接的权重参数

$\boldsymbol{W}_2$——第 2 层全连接的权重参数

σ——非线性激活函数 ReLU

$\boldsymbol{Y}$——MLP 的输出特征

MLP 的权重参数维度分别为 $\boldsymbol{W}_1\in\mathbf{R}^{rM\times M}$、$\boldsymbol{W}_2\in\mathbf{R}^{M\times rM}$,$M=N+1$,$r>0$ 为参数调节因子,用于调整 FC 层权重参数数量。MLP 对 $\boldsymbol{H}$ 的第 1 次融合是在 $\boldsymbol{H}$ 的行方向进行的,相当于对所有 Token 特征的逐个维度做特征混合。由于 FC 层的神经元具有全局感受野,每个神经元均可选择并融合 $\boldsymbol{H}$ 每一列的全局信息并建立局部特征之间的相关性,因此 $\boldsymbol{Y}$ 中融合了输入特征矩阵 $\boldsymbol{H}$ 的全局信息及局部特征间的相关性。将 $\boldsymbol{Y}$ 进一步混入 $\boldsymbol{H}$ 中,表示为

$$\boldsymbol{H}\leftarrow\boldsymbol{Y}+\boldsymbol{H}=\mathrm{MLP}(\boldsymbol{H})+\boldsymbol{H} \tag{3}$$

进一步用 MLP 对 $\boldsymbol{H}$ 进行第 2 次融合,表示为

$$\boldsymbol{Y}^{\mathrm{T}}=\mathrm{MLP}(\boldsymbol{H})=\boldsymbol{W}_4(\sigma(\boldsymbol{W}_3\boldsymbol{H}^{\mathrm{T}})) \tag{4}$$

式中 $\boldsymbol{W}_3$——第 3 层全连接权重参数

$\boldsymbol{W}_4$——第 4 层全连接权重参数

权重参数的维度分别为 $\boldsymbol{W}_3\in\mathbf{R}^{rD\times D}$、$\boldsymbol{W}_4\in\mathbf{R}^{D\times rD}$。同时将 $\boldsymbol{Y}$ 再次融入到 $\boldsymbol{H}$ 中,即有 $\boldsymbol{H}\leftarrow\boldsymbol{Y}+\boldsymbol{H}$。MLP 对 $\boldsymbol{H}$ 的第 2 次融合是在 $\boldsymbol{H}$ 的列方向进行的,相当于对 Token 特征进行通道维度的特征提取与混合。

将上述计算过程视为特征混合器(Feature mixer, FM),将特征混合器 FM 重复堆叠多次,通过 FC 的全局感受野,不断提取融合每个 Token 特征的全局信息,并建立局部特征间的相关性,最终输出融合全局信息的同维度张量,具体表示为

$$\boldsymbol{Z}=\mathrm{FM}_l(\mathrm{FM}_{l-1}(\cdots\mathrm{FM}_1(\boldsymbol{H}))) \tag{5}$$

式中 $\boldsymbol{Z}$——第 l 层 FM 输出特征矩阵

特征混合器的最终输出 $\boldsymbol{Z}\in\mathbf{R}^{(N+1)\times D}$ 维度与输入特征矩阵 $\boldsymbol{H}$ 维度相同,仍属于高维特征,需进一步对其进行降维。在通道维度上利用线性映射层对 $\boldsymbol{Z}$ 进行降维,表示为

$$\boldsymbol{Z}'=\boldsymbol{Z}\boldsymbol{W}_d \tag{6}$$

式中 $\boldsymbol{Z}'$——通道维度上降维后特征矩阵

$\boldsymbol{W}_d$——线性映射层权重参数

线性映射层权重参数维度为 $\boldsymbol{W}_d\in\mathbf{R}^{D\times C}$,其中 C 为线性映射层的神经元数量,降维后的特征矩阵维度为 $\boldsymbol{Z}'\in\mathbf{R}^{(N+1)\times C}$,且 $C<D$。应用线性映射层在通道维度进行降维,能够进一步在通道维度上融合不

同类型特征，同时起到参数化池化降维的作用。进一步在空间维度及 $\boldsymbol{Z}'$ 的行方向上用全连接层进行线性投影变化，即

$$\boldsymbol{O} = \boldsymbol{W}_r \boldsymbol{Z}' \tag{7}$$

式中　$\boldsymbol{O}$——降维后的最终特征矩阵

$\boldsymbol{W}_r$——全连接层的权重参数

权重参数 $\boldsymbol{W}_r \in \mathbf{R}^{K\times(N+1)}$，$K < N+1$，其中 K 为全连接层的神经元数量，最终特征矩阵维度为 $\boldsymbol{O} \in \mathbf{R}^{K\times C}$。式(6)和式(7)表示的 2 个全连接层分别在通道维度和空间维度上进行了特征降维和信息融合，进一步在其输出特征 $\boldsymbol{O}$ 的基础上，生成温室场景图像全局描述子。首先将 $\boldsymbol{O}$ 展平为向量，进而对向量进行 L_2 归一化，可表示为

$$\boldsymbol{F} = l_2 - \mathrm{norm}(\mathrm{Flatten}(\boldsymbol{O})) \quad (\boldsymbol{F} \in \mathbf{R}^{K\times C}) \tag{8}$$

式中　l_2 – norm——L_2 归一化操作

归一化操作计算方法为 $l_2 - \mathrm{norm}(x) = x/\|x\|_2$。将基于特征混合的全局描述子生成过程用图 3 所示的特征聚合器表示。

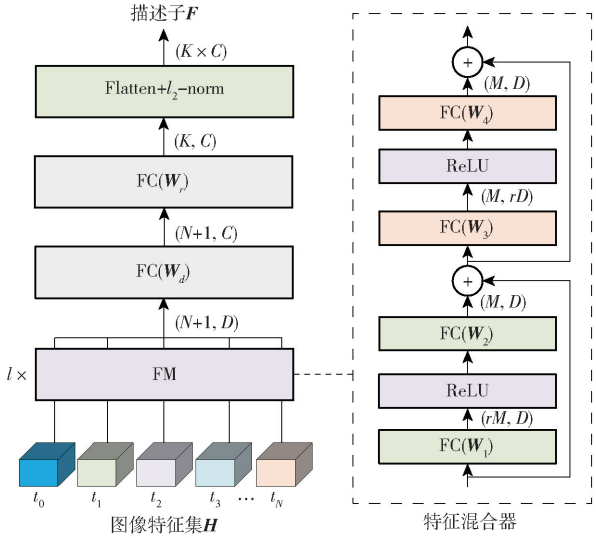


图 3　温室场景图像特征聚合器

Fig. 3　Feature aggregator for greenhouse scene image

1.5　模型优化目标定义

输入图像经预训练的特征提取网络，提取出局部特征矩阵 $\boldsymbol{H}$，进一步经特征聚合器聚合后生成图像的全局描述子 $\boldsymbol{F}$，二者构成了温室场景识别模型。特征提取器和特征聚合器需经过训练才能生成有效的全局描述子。温室场景识别模型的基本优化目标为：拉近同一场景图像的描述符距离，而推远不同场景图像的描述符距离。现有研究[21]表明，多重相似性损失函数(Multiple-similarity loss，MS Loss)[22-23]能够实现该目标，该函数兼顾了对比损失函数[24]和三元组损失函数[25]的特性，结合样本对间的绝对距离与批次内其他样本对的整体距离分布，采用加权策略调整损失，综合考虑了样本对间的多种关系，使模型能更全面地学习数据分布，能够实现更精确的度量学习。本文采用 MS Loss 作为模型优化目标，其定义可表示为

$$L_{MS} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{\alpha}\lg\left(1+\sum_{k\in P_i}\mathrm{e}^{-\alpha(S_{ik}-\lambda)}\right)+\frac{1}{\beta}\lg\left(1+\sum_{k\in N_i}\mathrm{e}^{\beta(S_{ik}-\lambda)}\right)\right] \tag{9}$$

式中　L_{MS}——多重相似性损失

S_{ik}——样本对的相似度

α、β、λ——损失定义的超参数

m——训练过程中一小批样本的图像数

P_i——小批次样本中相对于样本 i 的正样本(同一场景)索引集合

N_i——小批次样本中相对于样本 i 的负样本(不同场景)索引集合

对于每一个样本对，MS Loss 通过挖掘与加权两步迭代，综合考虑样本对自相似性、正负样本相对相似性，以全面利用局部样本分布，更有效地采集和加权重要信息样本对。

1.6　场景特征相似性度量方法

在 MS Loss 定义中，以及在以查询图像的全局特征描述子作为关键字检索参考图像特征库时，本研究均采用向量夹角的余弦值，即余弦相似度作为图像全局描述符 $\boldsymbol{A}$、$\boldsymbol{B}$ 的近似性度量，相似度越接近 1，表示样本对属于同一场景的概率越高。

2　结果与分析

2.1　场景识别模型训练与测试

基于 PyTorch 深度学习计算框架，采用 Python 编程语言，建立本研究温室场景识别模型。在一台配置有 48 GB 显存 Nvidia Tesla A40 图形处理器(Graphics process unit，GPU)、Intel Xeon Silver 4314 CPU、128 GB 内存、Windows Server 2022 操作系统、CUDA 12.1 加速库的计算机上开展模型的训练和测试试验。

采用温室场景训练集训练场景识别模型，将输入图像尺寸设置为 320 像素 × 320 像素，使用 AdamW 优化器迭代更新模型权重参数，初始学习率为 0.002，迭代次数(epoch)为 60，模型损失可收敛到稳定值。训练过程中，使用旋转、颜色抖动等数据增广技术[26]对图像进行变换，避免模型在训练过程中可能出现的过拟合问题，提高模型鲁棒性。

训练结束后，用测试集测试模型的温室场景识别效果，测试过程中，输入图像不再采用数据增广。采用 top-K 召回率(Recall@K，R@K)[27]评估模型场景识别性能。R@K 的含义为，用查询图像检索

参考场景图像,在返回的相似度最高的 K 个样本中,包含真正与查询图像为同一场景参考图像的概率,如包含查询场景,则识别成功,该指标的定义为

$$R@K = \frac{M_K}{X} \times 100\% \tag{10}$$

式中 X——查询图像个数

M_K——保留前 K 个查询结果前提下,成功进行了场景识别的样本数

2.2 不同方法场景识别性能比较

将本文方法应用于温室场景图像测试集,测试其在该数据集上的场景识别性能,同时和一些现有的基准场景识别方法进行比较,以验证方法的有效性。NetVLAD[28]、GeM[29]、Patch - NetVLAD[30]、MultiRes - NetVLAD[31]和 MixVPR[32]是几种常用的场景识别基准方法,采用与本研究相同的温室场景图像数据集训练并测试这几种方法,训练方法及超参数设置与文献中相同,方法的测试结果如表 1 所示。

表 1 不同方法场景识别性能比较

Tab. 1 Comparison of scene recognition performance among different methods %

方法	R@1	R@5	R@10
NetVLAD	70.61	76.45	80.76
GeM	71.84	80.83	83.57
Patch - NetVLAD	74.87	85.22	90.44
MultiRes - NetVLAD	74.29	85.37	89.59
MixVPR	76.55	87.64	90.63
本文方法	78.43	89.21	92.47

表 1 的测试结果表明,在温室场景识别任务上,本文方法具有一定优势,在几种方法中,具有最高的温室场景识别精度,其 R@1、R@5 和 R@10 分别达到 78.43%、89.21% 和 92.47%。上述 5 种对比方法均采用了相同的特征提取网络,因此几种方法在场景识别性能上的差异,主要是由全局特征描述子生成质量决定的,这也说明本研究构建的局部特征融合方法,及应用全连接层的全局感受野特性,在描述子中聚合图像的全局信息,并建立局部信息的空间相关性,对提高温室场景识别性能是有效的。与 NetVLAD 相比,本文方法 3 种指标分别提高 7.82、12.76、11.71 个百分点,NetVLAD 采用聚类的方法,将图像局部特征聚合为多个簇,通过连接多个簇的局部特征残差来构建全局描述子,没有很好地利用局部特征间的空间关系,且聚类中心是在训练过程中确定的,场景识别时推理过程中,聚类中心不再随场景的不同而变化,难以适应复杂多变的温室场景,因此其在温室场景识别任务上没有表现出良好性能。与 GeM 相比,本文方法 3 种指标分别提高 6.59、8.38、8.9 个百分点,GeM 应用参数化全局池化来生成图像的全局特征描述子,但池化过程中,丢失了过多的图像局部特征。相较于 Patch - NetVLAD,本文方法 3 种指标分别提高 3.56、3.99、2.03 个百分点,Patch - NetVLAD 由于缺乏更细粒度的局部描述子,导致可能无法达到传统局部特征在特征匹配方面的同等精度。相较于 MultiRes - NetVLAD,本文方法 3 种指标分别提高 4.14、3.84、2.88 个百分点,MultiRes - NetVLAD 方法在面对具有相似场景结构的温室数据集中,受到的影响更大。相较于使用整体特征聚合技术的 MixVPR,本文方法 3 种指标分别提高 1.88、1.57、1.84 个百分点。本文方法在保留局部信息的基础上,充分建立了各局部特征的空间相关性,融入了多样化全局特征,使描述子具有了较好的全局和局部可区分性。总体上,本文方法在温室场景识别任务上具有较好的性能。

2.3 特征聚合有效性分析

采用由 MLP 构建的多层 FM,在骨干网络提取的图像局部特征中融合全局信息,并建立局部特征的空间相关性,FM 在该过程中发挥了重要作用。CNN 中常用的池化操作也能够实现特征聚合,将特征聚合器中的 MLP 替换为平均池化层,并设置池化窗口为 3×3,滑动步长为 1,使骨干网络提取的局部特征图经多层池化后,维度保持不变,即与 FM 相同,后经 2 个线性层分别在通道维度和空间维度上进行降维,生成全局特征描述子,进一步验证基于特征混合的全局描述子生成方法的有效性。两种处理的训练和测试方法相同,结果如表 2 所示。

表 2 基于特征混合的全局描述子有效性分析

Tab. 2 Effectiveness analysis of global descriptors based on feature fusion %

聚合方法	R@1	R@5	R@10
特征混合	78.43	89.21	92.47
特征池化	70.42	79.86	84.90

由表 2 可以看出,采用特征混合的全局描述子生成方法,其在温室场景识别任务上,识别率更好,R@1、R@5 和 R@10 比采用特征池化的方法分别高 8.01、9.35、7.57 个百分点,基于 MLP 的特征混合器能够更好地根据场景识别任务聚合描述子所需的图像特征。池化操作虽然可以聚合池化窗口范围内的局部特征,但其感受野有限,难以建立局部特征间的全局空间相关性。由 MLP 构成的特征混合器,应用全连接层,具有全局感受野,能够融合全局特

征，同时还可根据场景识别任务需求，通过权重参数加强或抑制图像局部特征，因此能够生成更具鲁棒性的全局特征描述子，本文的描述子生成方法是有效的。

2.4　场景识别模型有效性检测

温室场景光照条件复杂，环境狭窄，遮挡严重，农业机器人等智能农机离目标场景远近各异，作物随时间生长变化。进一步开展模型有效性检测试验，验证其对光照、角度、距离和场景随时间变化的有效性。检测试验在一个与数据集构建不同的日光温室内进行，模型已经使用温室场景图像训练集训练。

2.4.1　光照对场景识别性能的影响

根据采样天气和时间段，将光照强度划分为 3 个等级，将晴朗天气的 08:00—10:00 时间段视为正常的中等亮度光照条件，将晴朗天气的 11:00—13:00 时间段视为强光照条件，将多云和雪天的 16:00—18:00 时间段视为弱光照条件。在不同光照条件下分别采集了 100 组场景的样本图像，每个场景约 20 幅图像，不同光照条件下图像的成像效果如图 4 所示。



(a) 正常光照　(b) 强光照　(c) 弱光照

图 4　光照变化场景示例

Fig. 4　Illumination change scene example

为探究光照条件对本研究模型场景识别性能的影响，用不同光照条件下采集的场景样本，测试模型场景识别性能，结果如表 3 所示。

表 3　光照对场景识别性能的影响

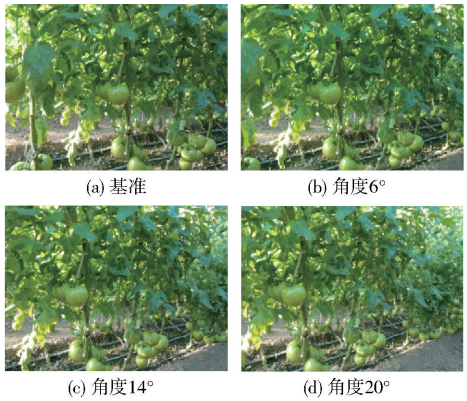
Tab. 3　Influence of illumination variation on scene recognition performance　%

光照条件	R@1	R@5	R@10
正常光照	80.29	90.71	92.70
强光照	76.35	83.40	88.87
弱光照	77.70	86.97	90.25

由表 3 可知，在不同光照条件下，本研究温室场景识别模型都表现出了较高的识别性能，模型对光照变化具有较好的适应性。同时光照条件对模型的识别精度也具有一定影响，在正常光照条件下，模型的场景识别性能最高，R@1、R@5、R@10 分别为 80.29%、90.71% 和 92.70%，在强光照条件下，R@1 下降 3.94 个百分点，在弱光照条件下，则下降 2.59 个百分点。下降在 4 个百分点以内，说明模型具有一定的鲁棒性。在正常光照条件下，相机捕捉的图像颜色、对比度等具有良好的稳定性，纹理清晰，更利于深度神经网络对图像局部特征的提取。强光照条件下图像会出现过曝现象，细节丢失，使得特征提取和匹配变得困难。弱光照条件下，图像普遍偏暗，缺乏足够的细节和纹理信息，不利于特征提取。总体上，模型在不同光照条件下都具有良好的识别性能，模型具有一定的鲁棒性，光照条件也会在一定程度上影响场景识别精度。

2.4.2　角度对场景识别性能的影响

将相机置于带旋转刻度载物台的三脚架上，在 0 ~ 20° 范围内左右旋转载物台，使相机对场景产生视角变化，过程中采集场景，共采集 56 组场景图像，每组约 25 幅图像。同一场景，不同角度下采集的图像示例如图 5 所示。



(a) 基准　(b) 角度6°　(c) 角度14°　(d) 角度20°

图 5　不同角度采集的场景示例

Fig. 5　Angle change scene examples

为探究角度变化对模型场景识别性能的影响，以起始位置拍摄的图像作为参考图像，将视角变化时采集的图像作为查询图像，测试本研究模型的场景识别性能，以验证视角变化对性能的影响，结果如表 4 所示。

表 4　视角变化时场景识别性能

Tab. 4　Influence of view angle change on scene recognition performance　%

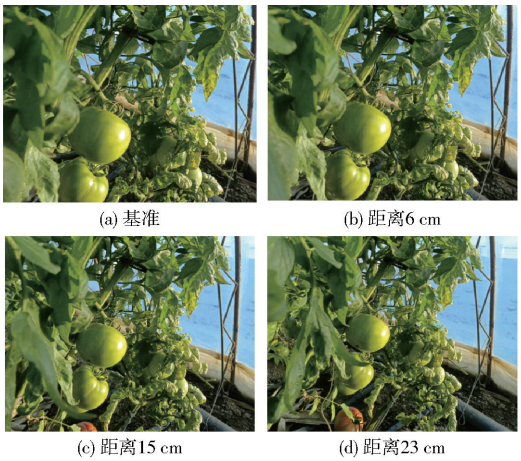
视角/(°)	R@1	R@5	R@10
1 ~ 7	79.22	91.28	94.47
7 ~ 14	77.89	86.79	90.27
14 ~ 20	72.61	74.84	77.52

从表 4 可以看出，随着相机视角增加，场景识别算法的 R@K 呈现逐渐下降的趋势。当角度为 1° ~ 7° 时，算法 R@1、R@5 和 R@10 分别为 79.22%、91.28% 和 94.47%，表明在小视角变化下，算法能够较为准确地进行场景识别。当视角为 7° ~ 14° 时，R@1 下降 1.33 个百分点，当视角变化角度进一步增加时，算法的性能逐渐下降。相机视角的小幅

变化,会使识别性能小幅下降。当视角为在 20°以内时,识别性能下降 6.61 个百分点。相机视角的较大变化会对场景识别算法的性能产生较大影响。随着视角变化幅度的增加,图像中的物体遮挡和背景变化等因素变得更加显著,查询图像和参考图像包含公共部分场景图像变小,不同图像所含语义信息发生较大变化,全局特征描述子的相似度下降,场景识别性能降低。总体上,场景识别模型在图像采样视角的小幅变化下仍具有较高精度,视角的大幅变化会使场景识别性能快速下降。

2.4.3 距离对场景识别性能的影响

当目标场景距离相机较近时,相机距目标的微小距离变化,即可使成像结果发生较大变化,利用该特性,对近距离场景进行采样,构建距离变化场景样本。将相机置于滑轨上的可移动滑块,相机视角垂直植株行,初始位于株行前 30 cm 处,在 0 ~ 30 cm 范围内逐渐远离并采集场景图像。共采集 100 组场景,每组场景约 25 幅图像。同一场景不同距离图像示例如图 6 所示。



(a) 基准 (b) 距离6 cm

(c) 距离15 cm (d) 距离23 cm

图 6 不同距离场景示例

Fig. 6 Distance change scene examples

以原始场景位置拍摄的图像作为参考图像,逐渐远离场景过程中采集的图像作为查询图像,进一步测试本文温室场景识别模型对距离的鲁棒性,结果如表 5 所示。

表 5 距离变化时场景识别性能

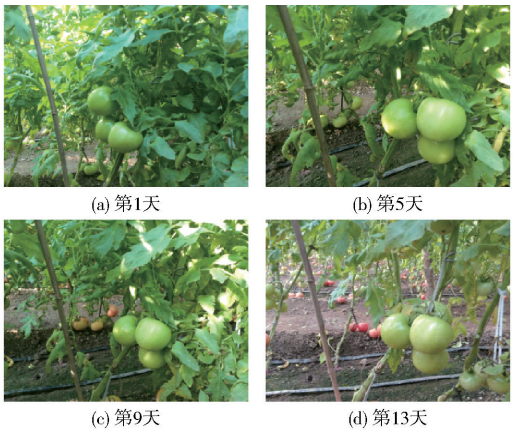
Tab. 5 Influence of distance variation on scene recognition performance %

距离/cm	R@1	R@5	R@10
1 ~ 10	81.81	90.82	93.74
10 ~ 20	74.94	84.11	85.66
20 ~ 30	63.94	76.56	78.96

由表 5 可知,随着采样距离增加,模型场景识别性能逐渐下降。当采样距离超过 10 cm 时,模型 R@1、R@5 和 R@10 分别下降 6.87、6.71、8.08 个百分点,当采样距离超过 20 cm 时,3 种指标分别下降 17.87、14.26、14.78 个百分点。根据针孔相机工作特点,在垂直视线下,当对目标场景采样距离扩大 2 倍时,视场所覆盖场景范围的宽、高也扩大 2 倍,查询图像与参考图像的公共场景大幅缩小,有更多目标进入查询图像,图像语义发生较大变化,场景识别性能下降。由表 5 可以看出,当采样距离为 20 ~ 30 cm 时,此时查询图像与参考图像相比,采样距离接近扩大 2 倍,但模型识别性能 R@1 仍达 63.94%,表明模型具有一定的距离变化鲁棒性。

2.4.4 作物动态生长对模型性能的影响

作物动态生长变化会导致同一地点的图像在外观上产生差异,从而影响场景识别的性能。因此,算法需要具备一定的鲁棒性,以应对作物生长的变化。在测试集上,共采集 58 组场景,每组场景约 40 幅图像,同一场景作物动态生长变化示例如图 7 所示。



(a) 第1天 (b) 第5天

(c) 第9天 (d) 第13天

图 7 作物动态生长变化场景示例

Fig. 7 Examples of dynamic growth and change scenarios for crops

在构建的作物动态生长变化测试集上,以第 1 天的场景数据作为参考集,并将第 5、9、13 天记录的场景数据当作查询数据集。通过评估时间变化反映作物动态生长变化对于场景识别模型性能的影响,了解模型在应对作物动态生长变化时的性能。

由表 6 可知,随着天数的增加,场景识别 R@K 呈下降趋势,随着作物动态生长变化,模型的识别性能逐渐下降。第 5 天采样时,模型 R@1、R@5 和 R@10 分别为69.92%、77.17% 和 79.84%,当采样

表 6 作物动态生长变化时场景识别性能

Tab. 6 Impact of dynamic crop growth changes on scene recognition performance %

作物动态生长时间/d	R@1	R@5	R@10
5	69.92	77.17	79.84
9	51.18	55.69	60.97
13	31.83	37.43	38.86

天数变化超过 13 d 时，3 种指标分别下降 38.09、39.74、40.98 个百分点。模型在经过作物动态生长变化后对场景的识别能力在下降。从图 7 可以清晰地观察到温室内部场景随时间推移所发生的变化，与第 1 天初始状态相比，第 5 天番茄果实已显著增大，叶片发黄；第 9 天时部分果实已经发生变色，叶片逐渐减少；在第 13 天时部分果实颜色更为鲜艳，趋于成熟，且叶片掉落明显，生长变化非常显著。从数据可知，在第 5 天时 R@1 为 69.92%，说明了模型在面对 4 d 时间变化内作物生长变化能保持较好的鲁棒性。

3　结论

（1）在温室场景识别任务中，本文模型 R@1、R@5 和 R@10 分别达到 78.43%、89.21% 和 92.47%，具有较高的场景识别精度。所提出的基于多层感知机的特征混合方法是有效的，与采用池化操作进行特征聚合相比，R@1 提高 8.01 个百分点。

（2）光照条件对场景识别性能具有一定影响，与正常的中等光照条件相比，强光及弱光条件下，模型的场景识别性能下降未超过 4 个百分点，即模型对光照条件变化具有一定的鲁棒性。

（3）相机视角及采样距离的变化也会影响模型识别性能，20° 以内的视角变化，使模型性能下降 6.61 个百分点，2 倍以内的距离变化，使性能下降 17.87 个百分点，但 R@1 仍达到 63.94%，模型具有一定稳定性。

（4）作物动态生长变化也会影响模型识别性能，5 d 内时间变化时仍能保持 R@1 为 69.92%，即模型对于作物生长变化条件具有一定鲁棒性。

（5）本研究场景识别方法与现有基准方法进行了比较，与 NetVLAD 相比，R@1、R@5 和 R@10 分别提高 7.82、12.76、11.71 个百分点；与 GeM 相比，分别提高 6.59、8.38、8.9 个百分点；与 Patch - NetVLAD 相比，分别提高 3.56、3.99、2.03 个百分点；与 MultiRes - NetVLAD 相比，分别提高 4.14、3.84、2.88 个百分点；与 MixVPR 相比，则分别提高 1.88、1.57、1.84 个百分点。在温室场景识别任务上模型性能有较大提升。

（6）温室场景复杂多变，类似于 NetVLAD 通过特征聚类实现场景图像全局描述子生成的方法，无法在推理时，根据场景变化改变聚类中心，而类似于 GeM 利用池化聚合局部特征，又难以保留足够的图像信息。在温室场景下，相较于本文所提出的方法，其他几种方法的表现均未能达到同样的数据水平。本研究构建的基于多层感知器的全局特征聚合和空间相关性建立方法，与基准方法生成全局描述子的思路完全不同，没有基准方法存在的相关问题。研究结果可为温室场景识别及智能农机视觉系统设计开发提供技术参考。

参　考　文　献

[1] 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)

[2] 陈劭，郭宇翔，高天啸，等. 移动机器人 RGB - D 视觉 SLAM 算法[J]. 农业机械学报，2018，49(10)：38 - 47.
CHEN Shao, GUO Yuxiang, GAO Tianxiao, et al. RGB - D visual SLAM algorithm for mobile robots[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(10): 38 - 47. (in Chinese)

[3] 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.

[4] ARANDJELOVIC R, GRONAT P, TOFII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 40(6): 1437 - 1451.

[5] CHEN Z, JACOBSON A, SLLNDERHAUF N, et al. Deep learning features at scale for visual place recognition[C]//2017 IEEE International Conference on Robotics and Automation (ICRA), 2017: 3223 - 3230.

[6] CHENG X, LU J, FENG J, et al. Scene recognition with objectness[J]. Pattern Recognition, 2018, 74: 474 - 487.

[7] ZHAO Z Y, LIU Z R, LARSON M, et al. Reproducible experiments on adaptive discriminative region discovery for scene recognition[C]//Proceedings of the 27th ACM International Conference on Multimedia, 2019: 1076 - 1079.

[8] 张学典，顾璋琦，秦晓飞. 基于 VGG16 模型的快速闭环检测算法[J]. 光学仪器，2019，41(3)：20 - 26.
ZHANG Xuedian, GU Zhangqi, QIN Xiaofei. VGG16 model-based fast loop closure detection algorithm [J]. Optical Instruments, 2019, 41(3): 20 - 26. (in Chinse)

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C] // International Conference on Learning Representations. Computational and Biological Learning Society, 2015.

[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 Chinse)

[11] 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)

[12] 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)

[13] 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 Chinse)

[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[C]// The Ninth International Conference on Learning Representations (ICLR 2021). Virtual Conference,2021.

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

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

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

[18] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818 - 2826.

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

[20] OQUAB M, DARCET T, MOUTAKANNI T, et al. Dinov2: learning robust visual features without supervision[EB/OL]. (2023-04-24)[2024-10-22]. https://openreview.net/forum? id = SwNL9a_eYyl.

[21] NIE J W, XUE D Y, PAN F, et al. Efficient saliency encoding for visual place recognition: introducing the lightweight pooling-centric saliency-aware VPR method[J]. IEEE Robotics and Automation Letters, 2024, 9(7): 6035 - 6042.

[22] 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 Chinse)

[23] 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.

[24] HADSELL R, SUMIT C, YANN L. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),2006:1735 - 1742.

[25] HOFFER E, AILON N. Deep metric learning using triplet network[C]// International Workshop on Similarity-based Pattern Recognition. Springer, Cham. , 2015.

[26] 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 Chinse)

[27] 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.

[28] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018,40(6):1437 - 1151.

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

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

[31] 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.

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