基于最优传输特征聚合的温室视觉位置识别方法

侯玉涵，周云成※，刘泽钰，张润池，周金桥
（沈阳农业大学信息与电气工程学院，沈阳 110866）

摘 要：为实现温室场景中基于视觉的位置识别，解决现有视觉位置识别模型局部特征聚合范式对训练样本归纳偏置的强依赖，以及聚合过程中存在的冗余信息问题，构建了一种基于最优传输局部特征聚合的温室视觉位置识别方法。将温室场景图像局部特征聚合过程视为最优传输问题，根据局部特征集动态生成分配矩阵，解耦模型对归纳偏置的强依赖，同时在分配中引入“垃圾”簇来解决特征冗余。结合卷积神经网络（convolution neural network，CNN）和 Transformer 的优势，优化设计温室场景图像局部特征提取网络。试验结果表明，在种植作物为番茄的温室场景中，所提方法的位置识别 top-1 召回率（$R_{@1}$）为 88.96%，与 NetVLAD、MixVPR 和 EigenPlaces 3 种方法相比，$R_{@1}$ 分别提高 29.67、2.97 和 2.89 个百分点。与 NetVLAD 和 MixVPR 的聚合器相比，基于最优传输局部特征聚合的 $R_{@1}$ 分别提高 21.65 和 1.09 个百分点。相较于 CNN 网络，研究构建的温室场景图像局部特征提取网络在 $R_{@1}$ 指标上提升了 5.45 个百分点。所提方法的实际温室场景位置识别率不低于 81.94%，具有一定的实际应用能力。基于最优传输局部特征聚合及全局描述符生成方法对位置识别是有效的，场景图像局部特征提取网络能够提高位置识别性能，研究结果可为温室智能农机装备视觉系统设计提供技术参考。

关键词：温室；视觉位置识别；最优传输；特征聚合；深度神经网络；Transformer
doi：10.11975/j.issn.1002-6819.202407155
中图分类号：S24；TP183 **文献标志码**：A **文章编号**：1002-6819(2024)-22-0161-12

侯玉涵，周云成，刘泽钰，等. 基于最优传输特征聚合的温室视觉位置识别方法[J]. **农业工程学报**，2024，40(22)：161-172. doi: 10.11975/j.issn.1002-6819.202407155 http://www.tcsae.org
HOU Yuhan, ZHOU Yuncheng, LIU Zeyu, et al. Recognizing visual position in the greenhouse using optimal transport feature aggregation[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(22): 161-172. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202407155 http://www.tcsae.org

0 引 言

近年来，日光温室等设施农业在中国北方地区发展迅猛，对保障农产品供应和增加农民收入发挥了重要作用。设施生产属劳动密集型行业，随着人口老龄化和城镇化进程的加快，农业劳动力短缺现象日益凸显，大力发展具有自主工作能力的智能农机装备势在必行[1]。同步定位与地图构建（simultaneous localization and mapping，SLAM）[2]被认为是移动机器人在未知环境下实现自主作业的关键，视觉位置识别（visual place recognition，VPR）[3]则是视觉 SLAM 所需的重要功能之一，也是后者实现回环检测和重定位的基础[4]。VPR 能基于视觉传感器（相机）回传的当前场景图像（查询图像），使用特征描述子或深度神经网络提取其特征，并与已具有位置信息的参考图像数据库进行匹配，如匹配成功，则参考图像的位置范围就是查询图像所在的位置。温室的金属棚架结构阻隔电磁信号，卫星定位技术难以在该相对封闭的环境中应用。在平面结构已知的温室中，如提前将特定位置的场景设定为参考（路标）[5]，应用 VPR 技术，智能农机可实现粗定位，即 VPR 可作为一种卫星定位的替代方案。另外，对于采用自主漫游工作的温室移动机器人[6]，如将其所经过位置记为参考，通过 VPR，也可避免对同一位置的多次重复作业。因此，VPR 在温室智能农机装备的视觉系统上具有广泛的应用潜力。

VPR 技术的关键是生成具有良好表征能力的图像全局描述符（向量）。早期的 VPR 技术主要应用人工设计的局部特征描述子，如 SIFT[7]、SURF[8]或 ORB[9]等，通过提取图像局部关键点特征，并进一步用词袋（bag-of-words，BoW）[10]模型或局部聚合描述子向量（vector of locally aggregated descriptors，VLAD）[11]将图像的多个局部特征聚合成全局描述符。温室环境及作物颜色、纹理单一，传统人工特征描述能力不足，难以提取高可分的局部特征。另外，BoW 模型需通过离线方式学习出一个视觉词典，例如通过对大量图像局部特征描述子进行 k 均值聚类，将聚类中心作为词典，词典中的每个单词表示一种特定的视觉特征，图像的全局描述符则被表示为视觉单词的频率直方图。VLAD 则需通过聚合图像的局部特征与相应的聚类中心的残差来获得全局描述符。上述方法[10-12]生成的全局描述符均忽略了局部特征在图像上的空间位置关系，而温室场景中作物、设施、地面等目标物的空间关系也是位置识别的关键依据之一。

近些年，具有更好语义描述能力的深度神经网络特

收稿日期：2024-07-17 修订日期：2024-10-10
基金项目：国家重点研发计划资助（2021YFD1500204，2023YFD1501303）
作者简介：侯玉涵，研究方向为深度学习与计算机视觉。
Email：hyh530@stu.syau.edu.cn
※通信作者：周云成，博士，教授，研究方向为机器学习在农业信息处理中应用。Email：zhouyc2002@syau.edu.cn

征被广泛应用于各类计算机视觉任务中，如农业领域中的果实目标检测[13]、图像语义分割[14]等，其中也包括VPR任务。BABENKO等[15]直接将CNN特定层的激活特征图连接成图像全局描述符，用于位置识别。ARANDJELOVIĆ等[16]在VLAD基础上提出一种支持端到端学习的NetVLAD特征聚合方法，该方法通过软分配将CNN提取的局部特征分配到各聚类中心，取得了超越传统方法的识别精度，是当前广泛采用的特征聚合范式之一。NetVLAD的聚类中心是该模型的可学习参数，是在大量训练样本上获得的归纳偏置，在推理过程中会依赖这些先验知识，这将影响其在VPR应用中的泛化性[17]。通用均值（generalized mean，GeM）[18]通过可学习全局池化操作在通道维度上将CNN的卷积特征图直接转化为全局描述符，通过该方法聚合的全局描述符，局部特征间的空间关系将全部丢失。KEETHA等[19]应用基于Transformer的预训练大模型DINOv2[20]提取图像局部特征，进而用无监督聚合方法生成全局描述符。ALI-BEY等[21]提出一种称为MixVPR的方法，其特征混合器应用全连接层的全局感受野特性，建立局部特征的空间相关性。在温室场景图像中，存在着区分度低的建筑墙体、地面等区域，对应的特征可视为冗余或无用特征，现有方法在将图像局部特征聚合为全局描述符时，对冗余或无用特征考虑较少。温室环境狭窄，目标物类型单一，植株随生长及生产管理高动态变化，关于温室场景下的VPR研究还不多见，VPR在该场景下的适用性值得进一步探讨。

针对温室智能农机装备视觉系统对VPR技术的实际需求，以及现有局部特征聚合方法对训练样本归纳偏置的强依赖，和聚合过程中存在的冗余或无用局部特征问题，本研究构建一种基于最优传输局部特征聚合的温室VPR方法。在场景图像局部特征聚类中引入“垃圾”簇来解决冗余或无用特征问题，将局部特征聚合视作最优传输问题，由局部特征集动态生成分配矩阵，解耦对归纳偏置的强依赖。结合CNN和Transformer的优势，优化设计温室场景图像局部特征提取网络，以期实现局部特征的高效提取和特征间空间关系的有效融合，为智能农机装备视觉系统设计提供技术参考。

1 温室视觉位置识别方法构建

1.1 视觉位置识别的基本思路

本研究将温室视觉位置识别建模为图像实例检索任务，整体技术框架如图1。该技术框架的基本思路为，预先采集同一温室内多个已知位置场景的图像，记为参考图像集Ω，同时为每幅参考图像标注位置信息，进一步用包括特征提取和特征聚合过程的映射f，将每幅参考图像$I_j \in \Omega$转换为对应的全局描述符，如式（1）所示：

$$\boldsymbol{F}_j = f(I_j) \tag{1}$$

式中$\boldsymbol{F}_j$为I_j的全局描述符，该描述符是一个紧凑的图像全局特征描述向量。所有参考图像的全局描述符构成参考集Λ。对于温室移动机器人或其他智能农机装备，将其在移动过程中通过相机获取的当前场景图像记为查询图像I_q，同样用f将I_q转换为全局描述符$\boldsymbol{F}_q$。用$\boldsymbol{F}_q$在Λ中检索最相似参考场景图像，首先将$\boldsymbol{F}_q$与Λ中的$\boldsymbol{F}_j$逐一进行相似度比较，计算相似度得分S_{qj}，如式（2）所示：

$$S_{qj} = \mathrm{sim}(\boldsymbol{F}_q, \boldsymbol{F}_j) \tag{2}$$

式中sim表示相似度度量函数，本研究用向量夹角的余弦值，即余弦相似度作为sim的实现方式。进一步对全部相似度得分进行统一的全局排序，当最高得分超过设定阈值时，可认为I_q与具有最高得分的I_j属于同一场景，机器人位于I_j对应的位置范围内，从而实现当前场景的位置识别。

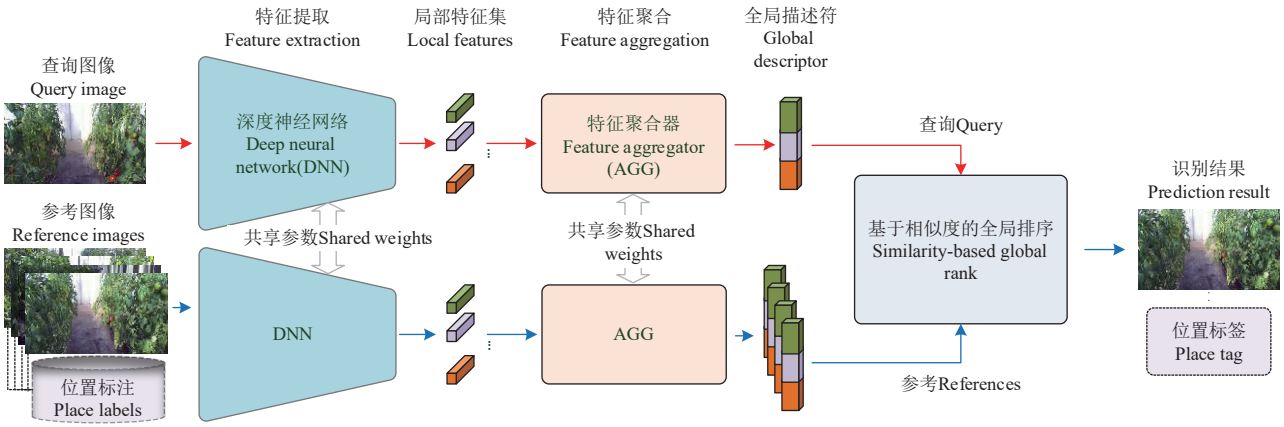


图1 温室视觉位置识别技术框架

Fig.1 Greenhouse visual place recognition technical framework

实现温室VPR的关键是构建有效的映射f，使采集自同一场景的图像全局描述符高度相似，不同位置场景的图像描述符差异显著，且f应具有视角、距离、光照、场景动态变化的鲁棒性。图像I的全局描述符$\boldsymbol{F}$可视为对应温室场景的语义信息表示。如同文本语义信息由构成该文本的各分词决定，图像的语义信息也由图像各局部区域的特征聚合而来。鉴于深度神经网络（deep neural network，DNN）良好的图像局部特征提取能力，本研究用其提取I的各局部区域特征，得到I的局部特征集，并进一步用本研究设计的特征聚合器（aggregator，

AGG）将局部特征集聚合为 $\boldsymbol{F}$，该过程可用式（3）、（4）表示：

$$\left\{\boldsymbol{t}_i | i \in [1,N], \boldsymbol{t}_i \in \mathbb{R}^D\right\} = \mathrm{DNN}(I) \tag{3}$$

$$\boldsymbol{F} = \mathrm{AGG}(\{\boldsymbol{t}_i\}) \tag{4}$$

式中$\{\boldsymbol{t}_i | i \in [1,N], \boldsymbol{t}_i \in \mathbb{R}^D\}$为 I 的局部特征集，以下简记为$\{\boldsymbol{t}_i\}$，$\boldsymbol{t}_i$表示 I 的局部区域特征向量，D 为该向量维度，N 为$\{\boldsymbol{t}_i\}$的元素数。结合式（1）、（3）和（4），则有式（5）：

$$\boldsymbol{F} = f(I) = \mathrm{AGG}(\mathrm{DNN}(I)) \tag{5}$$

映射 f 可由 DNN 和 AGG 复合得到。本研究通过优化设计 AGG 和 DNN，来构建适于温室 VPR 的映射 f。

1.2　基于最优传输的特征聚合器设计

DNN 提取的图像局部特征集$\{\boldsymbol{t}_i\}$，反映了温室场景图像 I 各局部区域信息，需进一步通过特征聚合器 AGG，将多个局部特征$\boldsymbol{t}_i$聚合为反映该场景语义信息的全局描述符 $\boldsymbol{F}$。NetVLAD 等模型[11, 16]是当前广泛采用的特征聚合范式，其首先将$\{\boldsymbol{t}_i\}$的多个元素软分配给 X 个簇$\{\boldsymbol{C}_x | x \in [1,X], \boldsymbol{C}_x \in \mathbb{R}^D\}$，$\boldsymbol{C}_x$表示第 x 个簇的中心特征向量（聚类中心），通过累加每个簇的局部特征残差，并级联各簇的特征残差和来获取全局描述符。特征聚合过程使图像全局描述符向量具有统一的维度，且维度与局部特征的数量无关。然而聚类中心$\boldsymbol{C}_x$则需要通过训练样本学习得到，并在推理时固定下来，且不再随图像特征改变。

温室类型及内部场景纷繁复杂，种植模式与作物类型多样，推理应用时的场景可能与训练样本的分布不同，固定的聚类中心$\boldsymbol{C}_x$使模型完全依赖训练中获得的归纳偏置，难以适应不同的温室环境以及多变的场景类型，进而限制 VPR 模型的鲁棒性。针对这一问题，本研究提出根据温室场景图像局部特征集分布，动态调整软分配策略，即聚类中心是非固定的，以适应温室场景的变化，即 AGG 不再学习固定的聚类中心，而是学习出根据局部特征集分布而动态调整分配策略的能力。

1.2.1　基于最优传输的局部特征分配

为生成固定维度的图像全局描述符，本研究仍将温室场景图像局部特征集$\{\boldsymbol{t}_i\}$分配给 X 个簇（用ω_x表示第 x 个簇）。NetVLAD 等模型在训练结束时，由固定的聚类中心$\boldsymbol{C}_x$来决定$\boldsymbol{t}_i$分配到各个簇的比例，与此不同，本研究将这一分配过程视为最优传输（optimal transport，OT）[22]问题，并改由局部特征集$\{\boldsymbol{t}_i\}$来决定各元素分配给各个簇的代价。为此本研究基于$\{\boldsymbol{t}_i\}$，用具有 2 个全连接（fully connected，FC）层的多层感知机（multilayer perceptron，MLP），来预测分配代价，如式（6）所示：

$$\boldsymbol{S}^{\mathrm{T}} = \mathrm{MLP}\left(T^{\mathrm{T}}\right) = \boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1 T^{\mathrm{T}} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2 \tag{6}$$

式中$\boldsymbol{S} \in \mathbb{R}^{N\times X}$为预测的代价矩阵，元素值$\boldsymbol{S}(i,x)$表示将$\boldsymbol{t}_i$分配给$\omega_x$的代价（得分），$\boldsymbol{T} \in \mathbb{R}^{N\times D}$为由$\{\boldsymbol{t}_i\}$各元素拼接构成的局部特征矩阵，$\boldsymbol{W}_1 \in \mathbb{R}^{M\times D}$、$\boldsymbol{b}_1 \in \mathbb{R}^M$、$\boldsymbol{W}_2 \in \mathbb{R}^{X\times M}$、$\boldsymbol{b}_2 \in \mathbb{R}^X$分别为 2 个 FC 层的权重矩阵和偏置向量，$M$ 为第一层神经元的数量，σ表示非线性激活函数。$\boldsymbol{S}$ 将决定$\{\boldsymbol{t}_i\}$的分配策略，直接用$\{\boldsymbol{t}_i\}$来预测 $\boldsymbol{S}$，使 AGG 根据场景图像动态调整聚类中心成为了可能，这为提高 VPR 模型的鲁棒性提供了基础。

在温室场景图像中，不可避免会出现地面、建筑墙体等区域，这些区域的局部特征难以区分，对 VPR 任务来说，属于冗余或无用的信息，若将这些特征全部聚合到全局描述符中，可能会干扰有用的关键特征。受 SuperGlue 方法[23]的启发，在 X 个簇的基础上，本研究进一步引入 1 个“垃圾”簇，用于将无用特征分配到该簇，并用$\bar{\boldsymbol{s}} \in \mathbb{R}^N$表示将$\{\boldsymbol{t}_i\}$中各局部特征分配给“垃圾”簇的代价。一种类型的特征是否对 VPR 有效，需结合对大量场景图像的分析学习才能做出判断，即$\bar{\boldsymbol{s}}$应由 AGG 根据对大量样本的学习来确定。因此，本研究在 AGG 中引入一个可学习参数$w_z \in \mathbb{R}$，并通过式（7）生成$\bar{\boldsymbol{s}}$：

$$\bar{\boldsymbol{s}} = w_z 1_N,\quad 1_N = [1,\ldots,1]^{\mathrm{T}} \in \mathbb{R}^N \tag{7}$$

此时，本研究 AGG 的可学习参数包括 MLP 的权重参数和w_z，分别负责有用簇和“垃圾”簇的分配代价预测，在训练过程中，二者将同步更新调整，MLP 逐渐学习出对$\{\boldsymbol{t}_i\}$的动态分配策略，w_z将归纳出冗余特征的类型。最终的代价矩阵$\bar{\boldsymbol{S}}$由$\boldsymbol{S}$和$\bar{\boldsymbol{s}}$级联得到，即有式（8）：

$$\bar{\boldsymbol{S}} = [\boldsymbol{S}, \bar{\boldsymbol{s}}] \in \mathbb{R}^{N\times(X+1)} \tag{8}$$

在代价矩阵$\bar{\boldsymbol{S}}$的基础上，进一步计算将$\{\boldsymbol{t}_i\}$中各元素分配到各个簇的份额，即获取分配矩阵$\overline{\boldsymbol{P}}$。$\overline{\boldsymbol{P}}$能够将$\{\boldsymbol{t}_i\}$代表的资源分布$\mu = 1_N$转换为 $(X+1)$ 个簇代表的分布$\tau = [1_X^{\mathrm{T}}, N-X]^{\mathrm{T}}$上，此时$\overline{\boldsymbol{P}}$需满足式（9）所示约束：

$$\overline{\boldsymbol{P}}^{\mathrm{T}} 1_N = \tau,\quad \overline{\boldsymbol{P}}\, 1_{X+1} = \mu \tag{9}$$

根据式（9）所示约束求解$\overline{\boldsymbol{P}}$的复杂度极高，CUTURI 等[24]提出，在代价矩阵基础上，应用 Sinkhorn 算法求解 OT 所需的分配矩阵，本研究也基于该算法求解$\overline{\boldsymbol{P}}$。Sinkhorn 算法通过在多次迭代过程中，交替地对$\exp\left(\bar{\boldsymbol{S}}\right)$的行和列进行归一化，来得到分配矩阵，该过程可表示为式（10）：

$$\overline{\boldsymbol{P}} = \mathrm{Sinkhorn}\left(\exp\left(\bar{\boldsymbol{S}}\right); \kappa\right) \tag{10}$$

式中κ表示 Sinkhorn 算法的迭代次数，本研究采用 $\kappa = 12$[23]。由于 Sinkhorn 算法过程是可微分的，因此其可直接应用于神经网络训练，实现端到端学习。

1.2.2　全局描述符生成方法

在分配矩阵$\overline{\boldsymbol{P}}$的基础上，将场景图像 I 的局部特征集聚合为全局描述符 $\boldsymbol{F}$。为降低 $\boldsymbol{F}$ 的维度，同时增强 AGG 的表达能力，首先用具有 2 个 FC 层的 MLP 对 I 的局部特征矩阵$\boldsymbol{T}$进行非线性降维变换，如式（11）所示：

$$\hat{\boldsymbol{T}}^{\mathrm{T}} = \mathrm{MLP}\left(\boldsymbol{T}^{\mathrm{T}}\right) = \boldsymbol{W}_{f2}\sigma\left(\boldsymbol{W}_{f1}\boldsymbol{T}^{\mathrm{T}} + \boldsymbol{b}_{f1}\right) + \boldsymbol{b}_{f2} \tag{11}$$

式中$\hat{\boldsymbol{T}} \in \mathbb{R}^{N\times D'}$为 I 的降维后的局部特征矩阵，D'为降维后的局部特征向量维度，且 $D' < D$，$\boldsymbol{W}_{f1} \in \mathbb{R}^{M\times D}$、$\boldsymbol{b}_{f1} \in \mathbb{R}^M$、$\boldsymbol{W}_{f2} \in \mathbb{R}^{D'\times M}$、$\boldsymbol{b}_{f2} \in \mathbb{R}^{D'}$分别为 2 个 FC 层的权重矩阵和偏置向量。

基于分配矩阵$\overline{\boldsymbol{P}}$，通过式（12）所示运算，将矩阵$\hat{\boldsymbol{T}}$分配（聚类）到 ($X$+1) 个簇上，则有：

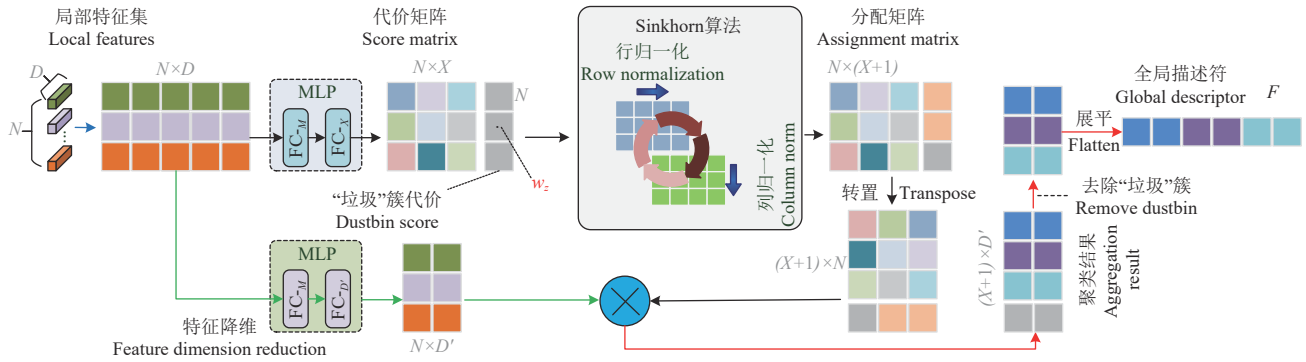
$$\overline{\boldsymbol{O}} = \overline{\boldsymbol{P}}^{\mathrm{T}}\hat{\boldsymbol{T}} \tag{12}$$

式中$\overline{\boldsymbol{O}} \in \mathbb{R}^{(X+1)\times D'}$为分配结果矩阵。$\overline{\boldsymbol{O}}$的最后一行为局部特征集在“垃圾”簇上的分配，将其从$\overline{\boldsymbol{O}}$中去除，得到有效局部特征分配结果矩阵$\boldsymbol{O} \in \mathbb{R}^{X\times D'}$，其中每一行分别表示局部特征集在一个有用簇上的聚类结果。通过对 $\boldsymbol{O}$ 的行向量进行级联，获取温室场景图像的全局描述符$\boldsymbol{F}$，如式（13）所示：

$$\boldsymbol{F} = [\boldsymbol{o}_{1,*}, \boldsymbol{o}_{2,*}, \ldots, \boldsymbol{o}_{X,*}]^{\mathrm{T}},\ \boldsymbol{F} \in \mathbb{R}^{XD'} \tag{13}$$

式中$\boldsymbol{o}_{1,*}$等为 $\boldsymbol{O}$ 的第 1 行中元素构成的向量。

基于最优传输的局部特征聚合器 AGG 的整体结构设计可表示为图 2，其中 D、M、D'和 X 将作为 AGG 的超参数，决定 2 个 MLP 的权重参数数量。AGG 可将 DNN 提取的图像局部特征集聚合为全局描述符。



注：N 为特征向量个数；D 和 D'分别为原始和降维后的特征向量维度；X 表示聚类的簇数；MLP 表示多层感知机；FC$_M$ 等表示神经元数量为 M 的全连接层；w_z 为可学习参数。下同。

Note: N is the number of feature vectors; D and D' respectively are the dimensions of the original and reduced feature vectors; X represents the number of clusters; MLP stands for multilayer perceptron; FC$_M$ represents a fully connected layer with a neuron count of M; w_z is a learnable parameter. The same below.

图 2 基于最优传输的局部特征聚合器

Fig.2 Optimal-transport-based local feature aggregator

1.3 温室场景图像局部特征提取网络优化设计

在机器人等智能农机装备获取的温室场景图像中，绿色植株、建筑墙体、温室地面等是图像的主体。以植株部分为例，其各局部区域高度相似，难以区分。人类视觉系统在解决这类问题时，主要依赖于超越局部区域的更大尺度的环境视觉信息。AGG 通过聚合局部特征集，生成图像的全局描述符，局部特征的可分性对全局描述符的可分性至关重要，即温室场景图像局部特征也应通过融入更大尺度的环境信息来提高可分性。另外，局部特征间在聚类前建立互信息，也有利于高质量聚类结果的生成。因此本研究在构建图像局部特征提取网络 DNN 时，将设计能够在局部特征中融入全局环境信息并建立特征间互信息的网络结构，同时考虑网络的图像局部特征提取能力。

1.3.1 特征提取网络设计思路

当前可用于图像特征提取的网络主要有 CNN[25] 和 Transformer[26] 两种类型。CNN 在捕获图像局部细节信息方面具有优势，但对全局信息的捕捉能力较弱；Transformer 的多头注意力（multi-head self-attention，MHSA）机制具有全局感受野特性，能够快速建立局部特征间的互信息，但缺少对图像局部细节信息的提取能力。现有研究[18-21] 通常只采用一种网络结构来提取图像特征，本研究则结合 CNN 和 Transformer 的优势来设计 DNN，使其既具有良好的图像局部特征提取能力，又能通过 MHSA 在局部特征间建立互信息并在特征中融入环境信息，使 DNN 的设计能够更加针对温室场景图像特点。本研究 DNN 的总体设计思路是，在视觉 Transformer（vision transformer，ViT）[27] 架构基础上，通过用良好设计的 CNN 图像嵌入模块取代 ViT 的线性嵌入，同时改进 ViT 编码器的 Transformer 构造块，使其在满足建立互信息并提取全局信息的前提下，进一步降低计算复杂度。

1.3.2 CNN 图像嵌入模块设计

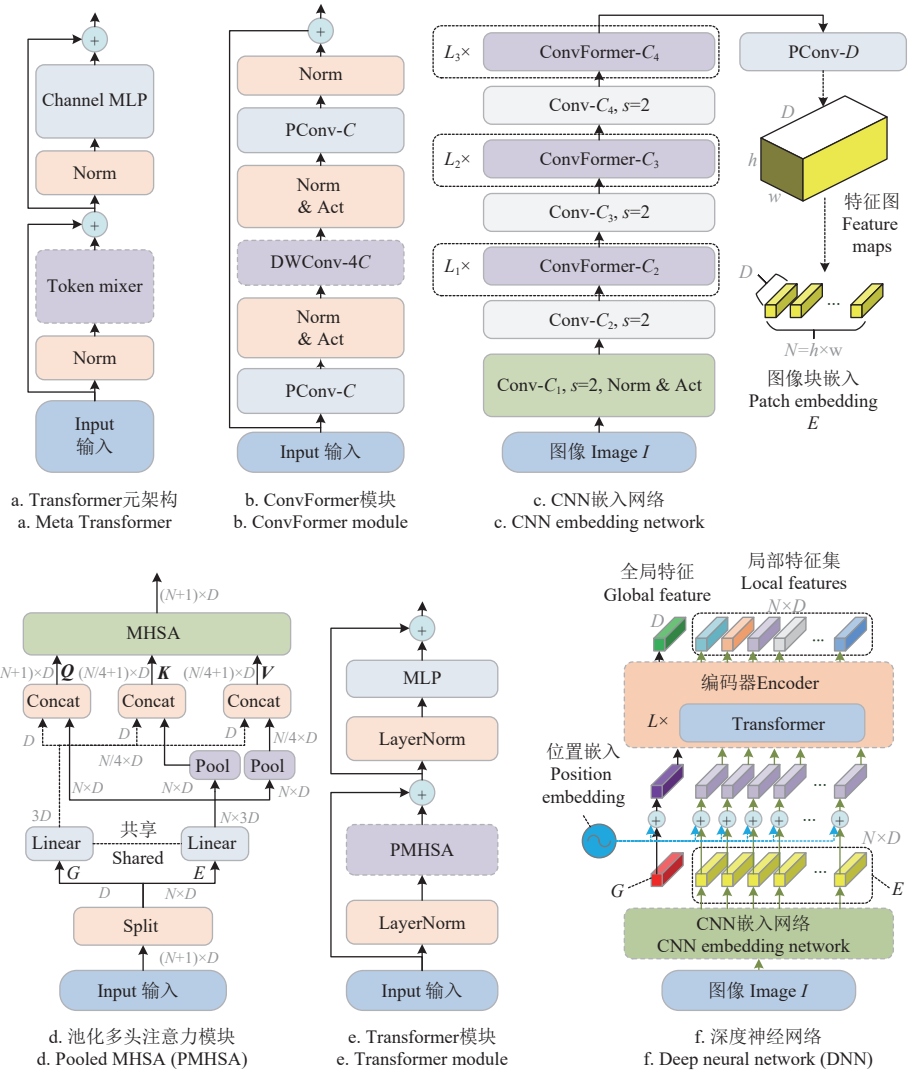
在参考深度神经网络相关研究[28-29] 基础上，设计 CNN 嵌入模块。YU 等[28] 通过将 Transformer 的 MHSA 抽象为词混合器，提出了图 3a 所示的 Transformer 元架构。LI 等[29] 进一步证明用轻量的深度化卷积（depth-wise convolution，DWConv）作为元架构的词混合器也是有效的，本研究也采用该处理，同时用等效的点卷积（point convolution，PConv）取代通道 MLP 中的线性层，且进一步将 DWConv 移入两个 PConv 之间，以优先执行局部词混合，此时 Transformer 元架构变化为全卷积模块（ConvFormer），结构如图 3b，其中超参数C表示模块的通道数，且用批归一化（batch normalization，BN）作为 Norm 层。由于本研究的 ConvFormer 从 Transformer 元架构变化而来，因此其具有卷积和 Transformer 的综合特性。

通过用 ConvFormer 堆叠及跨步卷积降维，构建 CNN 图像嵌入模块，结构如图 3c，其中的超参数 C_1 等决定网络宽度，本研究分别取 C_1=64、C_2=128、C_3=256 和 C_4=512，L_1 等决定网络深度，分别取 L_1=2、L_2=4 和 L_3=4 [29]。输入图像 I 通过嵌入模块实现嵌入的过程可表

示为式（14）：

$$E=\mathrm{CNNEmb}(I)，\ E\in\mathbb{R}^{h\times w\times D} \tag{14}$$

式中 CNNEmb 表示 CNN 图像嵌入模块，$\boldsymbol{E}$ 为输出特征图张量，$h\times w$表示 E 的空间维度。将 E 在通道维度上的每个特征向量$\boldsymbol{E}_{x,y}\in\mathbb{R}^D, x\in[0,w), y\in[0,h)$，作为局部图像块的嵌入。由于 CNNEmb 的卷积特性，$\boldsymbol{E}_{x,y}$包含了局部图像块的细节特征信息。



a. Transformer元架构 a. Meta Transformer　b. ConvFormer模块 b. ConvFormer module　c. CNN嵌入网络 c. CNN embedding network

d. 池化多头注意力模块 d. Pooled MHSA (PMHSA)　e. Transformer模块 e. Transformer module　f. 深度神经网络 f. Deep neural network (DNN)

注：Norm 为归一化层；PConv-C 表示通道数为 C 的点卷积；Act 表示激活层；DWConv-4C 表示通道数为 4C 的深度化卷积；Conv-C_1, s=2 等表示通道数为 C_1，步长为 2 的跨步卷积；L_1×等表示模块重复堆叠 L_1 次；h, w 为特征图空间维度；MHSA 表示多头注意力；Concat 表示特征级联操作；Pool 为池化操作；Linear 表示线性变换；Split 为特征分割操作；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询、关键字和值矩阵；G 表示全局嵌入；LayerNorm 表示层归一化。下同。

Note: Norm denotes normalization layer; PConv-C represents point convolution with C channels; Act means activation layer; DWConv-4C represents a depth convolution with 4C channels; Conv-C_1, s=2, etc. represent stride convolutions with a channel number of C_1 and a stride of 2; L_1× represents module stacking repeated L_1 times; h, w is the spatial dimension of the feature maps; MHSA stands for multi-head self-attention; Concat is used for feature concatenation; Pool is a pooling operator; Linear represents a linear transformation; Split is a feature splitting operator; $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ represent the query, key, and value matrices respectively; G represents the global embedding; LayerNorm stands for layer normalization. The same below.

图 3　图像特征提取网络

Fig.3　Image feature extraction network

1.3.3　位置嵌入及全局嵌入引入

目标物之间的空间关系也是场景位置识别的重要依据，因此图像块的位置信息也应通过编码融入到图像块嵌入 $\boldsymbol{E}$ 中。参考 DETR 模型[30]，本研究用固定的 sine 和 cosine 位置编码方法将位置信息编码为位置嵌入向量，如式（15）所示：

$$\boldsymbol{P}_{x,y}=\mathrm{PosEmb}(\vartheta(\boldsymbol{E}_{x,y}))，\ \vartheta(\boldsymbol{E}_{x,y})=x+y\cdot w+1 \tag{15}$$

式中$\boldsymbol{P}_{x,y}$为$\boldsymbol{E}_{x,y}$的位置嵌入向量，PosEmb 表示 sine 和 cosine 位置编码操作，ϑ表示获取$\boldsymbol{E}_{x,y}$在 $\boldsymbol{E}$ 上的位置信息。将位置嵌入 $\boldsymbol{P}$ 通过加操作融入到 $\boldsymbol{E}$ 中后，作为最终的图像块嵌入，即有$\boldsymbol{E}\leftarrow\boldsymbol{P}+\boldsymbol{E}$。

为使温室场景图像局部特征能有效融合全局环境信息，本研究进一步引入 1 个可学习的全局嵌入（命名为$\boldsymbol{G}\in\mathbb{R}^D$），并在其中融入位置嵌入，即有$\boldsymbol{G}\leftarrow\boldsymbol{G}+\mathrm{PosEmb}(0)$。全局嵌入 $\boldsymbol{G}$ 与图像块嵌入 $\boldsymbol{E}$ 一起，经后续 ViT 编码器的处理，使 $\boldsymbol{G}$ 通过 MHSA 逐渐融合多个$\boldsymbol{E}_{x,y}$的信息，以获取输入图像的全局特征，$\boldsymbol{E}_{x,y}$也在该过程中

融合 $\boldsymbol{G}$，获得全局环境信息。

1.3.4 Transformer 构造块的改进

原 Transformer 构造块[27]主要由带残差连接的 MHSA 和 MLP 两部分构成，其中 MHSA 在计算全局注意力时，需要对查询矩阵 $\boldsymbol{Q}$、关键字矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$ 做大量的矩阵乘法运算，产生很高的计算延时。农业机器人等设备计算能力有限，在满足识别精度的前提下，应降低网络计算量。本研究在执行 MHSA 计算时，先用池化（Pool）在 $\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵的序列维度上执行降维操作，此后再进行注意力计算，以降低矩阵计算量，将改进后的 MHSA 称为池化多头注意力模块（pooled MHSA，PMHSA），其结构如图 3 d。由于 PMHSA 将 $\boldsymbol{K}$、$\boldsymbol{V}$ 的序列维度降为原来的 1/4，其计算量下降为原来的 25%。

用 PMHSA 置换原 Transformer 构造块中的 MHSA，构建改进的 Transformer 块（图 3e），以降低运算量，将其重复堆叠多次，作为 DNN 的 ViT 编码器，结构如图 3f，其中超参数 L 决定编码器的深度，本研究设 L=8[27]。用编码器在全局嵌入 $\boldsymbol{G}$ 和图像局部块嵌入 $\boldsymbol{E}$ 上融合环境特征并建立互信息，编码器的最后一个 Transformer 块的输出 $\boldsymbol{G}^{(L)}$、$\boldsymbol{E}^{(L)}$ 分别为全局嵌入特征向量和图像块嵌入特征张量，其中 $\boldsymbol{G}^{(L)}$ 融合了图像全局特征，$\boldsymbol{E}^{(L)}$ 则提取了图像局部特征。对 $\boldsymbol{E}^{(L)}$ 进行归一化操作，并将归一化结果作为输入图像 I 的局部特征矩阵 $\boldsymbol{T}$，如式（16）所示：

$$\boldsymbol{T}=\psi\left(\boldsymbol{E}^{(L)}\right) \tag{16}$$

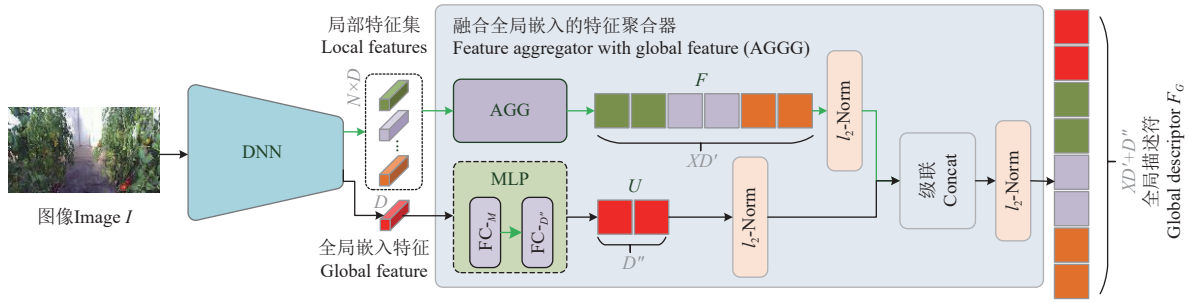
式中 ψ 表示归一化操作。

1.3.5 描述符与全局特征的融合

用 AGG 聚合图像的局部特征矩阵 $\boldsymbol{T}$，可生成描述符 $\boldsymbol{F}=\mathrm{AGG}(T)$。编码器输出的全局嵌入特征向量 $\boldsymbol{G}^{(L)}$ 也融合了场景图像的全局特征信息，本研究将其插入到 $\boldsymbol{F}$ 中，以期进一步提高温室场景 VPR 能力。为降低全局描述符的维度，首先用 MLP 对 $\boldsymbol{G}^{(L)}$ 进行非线性降维变换，得到降维后的全局特征向量 $\boldsymbol{U}$。级联 $\boldsymbol{U}$ 和 $\boldsymbol{F}$ 并归一化，获取最终的图像全局描述符 $\boldsymbol{F}_G$，如式（17）所示：

$$F_G=\psi([\psi(U),\psi(F)]),\ U=\mathrm{MLP}\left(G^{(L)}\right) \tag{17}$$

式中 $\boldsymbol{F}_G$ 表示最终的全局描述符。由图像 I 生成 $\boldsymbol{F}_G$ 的过程如图 4 表示，其中将融合 AGG 聚合结果与全局嵌入特征的操作部分视为另一特征聚合器（AGGG），其对应的超参数分别设置为 D = 768、M = 512、D' = 128、K = 64、D'' = 256[17]。



注：D''为降维后的全局特征向量维度；l_2-Norm 表示 L2 归一化；U 表示降维后的全局特征。
Note: D'' is the dimension of the reduced global feature vector; l_2-Norm represents L2 normalization; U denotes reduced global features.

图 4 图像全局描述符生成过程

Fig.4 The process of generating image global descriptor

1.4 温室视觉位置识别模型优化目标定义

用本研究设计的 DNN 和 AGGG 构成最终映射 f（温室 VPR 模型），该模型需经过采集自日光温室的大量场景图像的训练，才能用于 VPR 任务。f 的优化目标是通过不断训练，拉近同一位置场景图像全局描述符的距离，同时推远不同位置场景图像全局描述符的距离。多重相似性（multiple-similarity，MS）损失[31]在这类目标的度量学习中具有良好的效果，本研究以该损失的最小化作为 f 的优化目标，其定义如式（18）所示：

$$\ell_{MS}=\frac{1}{m}\sum_{q=1}^{m}\left\{\frac{1}{\alpha}\log\left[1+\sum_{r\in\mathcal{P}_q}e^{-\alpha\left(S_{qr}-\lambda\right)}\right]+\frac{1}{\beta}\log\left[1+\sum_{r\in\mathcal{N}_q}e^{\beta\left(S_{qr}-\lambda\right)}\right]\right\} \tag{18}$$

式中 ℓ_{MS} 表示 MS 损失，α、β、λ 为 ℓ_{MS} 的超参数，m 表示训练中一个小批次训练样本的图像数，S_{qr} 表示图像 I_q、I_r 对应全局描述符间的余弦相似度，$\mathcal{P}_q$、$\mathcal{N}_q$ 为小批次样本中相对于 I_q 的正样本（同一场景）和负样本（非同一场景）索引集合，而正、负样本则通过文献 [31] 的方法挖掘产生。

2 温室视觉位置识别试验

2.1 温室场景数据集构建

于 2023 年 10—12 月，在沈阳农业大学教学科研基地多栋日光温室内开展场景图像采集，种植作物为番茄，处于膨果转色期，株高约 1.1～1.9 m，株距约 0.3 m，行距约 1.2 m。在 09:30—17:00 时段采集图像。温室场景图像训练样本的多样性对提高 VPR 模型泛化性能至关重要，应用相机物理特性及参数差异产生的成像变化来增加样本的多样性，即在采样阶段就通过使用不同相机来实现图像样本的物理增广。分别使用 HUAWEI Mate 50、VIVO X16、XIAOMI Note 10 手机，及 Intel RealSense D435 摄像头采集图像，其中手机分辨率均设置为 4 000×

3 000 像素，摄像头分辨率设置为 1 920×1 080 像素。采用不同传感器采集的样本训练 VPR 模型，在实际部署时也可增加模型对不同视觉传感器的适应性。将温室内的不同位置，如不同植株行，定义为不同场景。图像采集时，记录采集参数、场景位置等相关数据。

在数据采集过程中，变化相机与场景间的距离、视角，并在不同光照条件下采样，同时对作物生长过程中在同一场景采样，以构建不同距离、视角、光照和作物生长变化的场景样本。由于行间距限制，相机视线垂直作物行时，相机距作物行近，距离的微小变化即可在图像上产生显著变化，应用该特性，将相机置于滑轨上，调节距作物行远近并成像，获取同一场景下不同成像距离的图像。将相机固定于三脚架顶具旋转刻度载物台上，旋转载物台并成像，获取同一场景不同视角的图像。在一天内不同时段对同一场景成像，获取同一场景光照变化的图像。考虑作物生长动态变化，对同一场景，分别在第 1、5、9、13 天成像，获取同一场景作物生长动态变化图像。考虑不同距离时，主要针对作物行场景进行短距离、小场景成像，其他类型场景均考虑大场景，即图像中包含更多的温室场景目标。共在 5 个温室，采集了 24 000 幅图像，包含 2 000 个不同场景，每个场景 12 幅图像，构成温室场景数据集。

2.2　模型训练及测试方法

基于 PyTorch 2.1 计算框架[32]，用 Python 编程实现本研究温室视觉位置识别模型。在配置有 Intel Xeon Silver 4 314 处理器，128 GB 内存，NVIDIA Tesla A40 计算卡，Windows Server 2019 操作系统，CUDA 12.1 加速库的计算机上开展模型的训练及测试试验。

将温室场景数据集按 4:1 比例分为训练集和测试集。应用训练集，通过小批量随机梯度下降法训练 VPR 模型，每一小批样本包括 16 个场景的共 64 幅图像，每个场景随机选择 4 幅图像，训练过程中，将输入图像调整为 320×320 像素，并用随机数据增强方法[33]来进一步增加训练样本的多样性。初始学习率设为 10^{-4}，在第一阶段的 10^5 次迭代中，学习率线性下降为 10^{-6}，固定学习率后，再进行 10^5 次迭代训练[17]，模型损失值可收敛到稳定状态。

应用测试集，测试 VPR 模型识别能力，用场景识别相关研究[16,21,34]常用的 top-K 召回率（$R_{@K}$）评价其性能。测试集包含 400 个场景，从每个场景随机选择 1 幅图像，连同场景编号构成参考图像集 Ω，剩余图像构成查询集 H。用本研究方法计算 Ω 样本的全局描述符，构成参考集 Λ。对于 H 的每幅查询图像I_q，同样计算其全局描述符，以余弦相似度作为距离度量，在 Λ 中搜索离该描述符最近的 K 个样本，如结果中含有与I_q为同一场景的图像，则记此次查询的分数$s_q=1$，否则$s_q=0$，$R_{@K}$即为 H 中所有查询图像的s_q平均值，如式（19）所示：

$$R_{@K}=\frac{1}{n}\sum_{q}^{n}s_q \tag{19}$$

式中n表示 H 的样本数。

3　结果与分析

3.1　方法比较

为验证本研究方法的有效性，和几种优秀 VPR 方法 NetVLAD[16]、GeM[18]、CosPlace[35]、EigenPlaces[36]、MixVPR[21] 进行比较。各方法均在本研究构建的温室场景数据集上进行相同的训练和测试，结果如表 1。

表 1　不同视觉位置识别方法性能比较

Table 1　Comparison of performance between different visual place recognition methods　　%

方法 Methods	$R_{@1}$	$R_{@5}$	$R_{@10}$
NetVLAD	59.29	76.43	82.44
GeM	65.73	77.33	82.06
CosPlace	76.84	88.37	91.82
EigenPlaces	86.07	95.52	97.33
MixVPR	85.99	93.45	95.41
本文方法 Our method	88.96	96.06	97.65

注：$R_{@1}$、$R_{@5}$、$R_{@10}$ 分别表示通过式（19）计算的 top-1、top-5 和 top-10 召回率。下同。

Note: $R_{@1}$, $R_{@5}$, $R_{@10}$ respectively represent the top-1, top-5, and top-10 recalls calculated by equation (19). The same below.

由表 1 可知，在温室 VPR 任务中，相较于其他方法，本文方法具有最高的识别性能。与 NetVLAD 相比，本文方法的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别提高 29.67、19.63 和 15.21 个百分点。本研究骨干网络结合了 CNN 和 Transformer 的优势，与 NetVLAD 的 CNN 特征提取网络相比，具有更好的环境特征融合能力，同时本文方法根据图像的局部特征集动态调整全局描述符聚合结果，相较于 NetVLAD 聚合器依赖训练过程中学习到的归纳偏置，具有更好的变化适应性。与 GeM 相比，本文方法的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别提高 23.23、18.73 和 15.59 个百分点。本文的特征聚合器在生成全局描述符时，考虑了局部特征间的空间相关性，相较于 GeM 仅使用参数化空间均值池化获取全局描述符，具有更好的场景语义表达能力。与 CosPlace 相比，本研究方法在 $R_{@1}$ 等 3 种指标上分别高出 12.12、7.69 和 5.83 个百分点，与 EigenPlaces 相比，则分别高出 2.89、0.54 和 0.32 个百分点。相较于 CosPlace 和 EigenPlaces 的图像分类任务，本研究将 VPR 建模成度量学习任务，所生成的图像描述符更能表达场景的完整语义信息，具有更好的泛化性。与 MixVPR 相比，本文方法的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别高出 2.97、2.61 和 2.24 个百分点。本文方法和 MixVPR 在聚合全局描述符向量时均应用了全局注意力机制，且都考虑了各局部特征的重要性，但前者进一步考虑了无用特征对全局描述符的干扰，使其位置识别精度进一步提高。本文方法的位置识别 $R_{@1}$ 指标为 88.96%，且在各召回率指标上均优于其他 5 种方法，表明其在温室 VPR 任务上是有效的。

3.2　特征聚合及骨干网络有效性验证

本研究构建了基于最优传输的全局特征聚合方法，优化设计了温室场景图像局部特征提取网络，为进一步验证这些设计和处理的有效性，开展基于不同设置的验证试验。

3.2.1 特征聚合方法有效性验证

为验证本研究特征聚合方法的有效性，分别用NetVLAD、GeM、MixVPR 等方法的特征聚合器将 DNN 输出的局部特征集聚合为全局描述符，由于 CosPlace 和 EigenPlaces 无独立的特征聚合器，不再参与试验比较。进一步用本研究设计的聚合器 AGG 和 AGGG 生成全局描述符，AGG 仅聚合局部特征集，AGGG 则更进一步地将 DNN 输出的全局嵌入特征融入描述符。同时将 DNN 输出的全局嵌入特征直接作为全局描述符，将其命名为虚拟聚合器 G。用本研究设计的 DNN 作为骨干网络，采用不同的聚合器，构建 6 种模型，在温室场景数据集上训练并测试各个模型，结果如表 2。

表 2 特征聚合方法性能比较

Table 2 Performance comparison of feature aggregation methods

%

模型编号 Model No.	配置 Configures	$R_{@1}$	$R_{@5}$	$R_{@10}$
1	DNN + NetVLAD	67.31	79.86	85.29
2	DNN + GeM	86.04	95.07	96.48
3	DNN + MixVPR	87.87	95.04	96.34
4	DNN + AGG	86.80	95.15	97.01
5	DNN + G	86.15	93.64	95.53
6	DNN + AGGG	88.96	96.06	97.65

由表 2 可知，在骨干网络相同的前提下，本研究基准模型（模型 6）具有最高的位置识别精度。与 MixVPR 的聚合器相比，聚合器 AGGG 的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别提高 1.09、1.02 和 1.31 个百分点，与 NetVLAD 的聚合器相比，$R_{@1}$ 则提高了 21.65 个百分点，表明 AGGG 的设计对提高 VPR 性能是有效的。与 MixVPR 的聚合器只融合局部特征相比，AGGG 在动态聚合局部特征的同时，还融合了 DNN 提取的全局特征，具有更好的场景区分能力。对比模型 4 和模型 3，前者的 $R_{@5}$ 和 $R_{@10}$ 上高于后者。AGG 和 MixVPR 均考虑了不同局部特征的重要性，但前者通过引入“垃圾”簇，排除了无用信息对全局描述符的干扰。对比模型 4 与模型 2、1，前者在 3 项指标上均高于后两者，进一步表明基于最优传输的特征聚合方法的有效性。对比模型 5 和模型 1，即使只用 DNN 输出的全局特征作为全局描述符，前者仍具有高于后者的位置识别性能，表明 DNN 的编码器通过 PMHSA 有效融合了场景图像的全局特征，这也是模型 6 能在模型 4 基础上进一步提高识别性能的原因。总体上，本研究设计的基于最优传输的特征聚合器及全局特征描述符生成方法是有效的，能够提高温室 VPR 性能。

3.2.2 特征提取网络有效性验证

为进一步验证本研究 DNN 网络的有效性，分别用 ResNet50[37]、EfficientNet-B0[38]、Swin Transformer v2-base(Swin2-B)[39]，以及 DNN 作为图像局部特征提取器，考虑到前 3 种网络无全局嵌入特征输出，用本研究 AGG 作为局部特征聚合器，组成 4 种模型，在温室场景数据集上进行相同的训练和测试，结果如表 3。

表 3 局部特征提取网络性能比较

Table 3 Performance comparison of local feature extraction networks

模型编号 Model No.	配置 Configures	$R_{@1}$/%	$R_{@5}$/%	$R_{@10}$/%	参数量 Parameters/M	计算速度 Calculation speed/(帧·s^{-1})
4	DNN + AGG	86.80	95.15	97.01	87.5	55.65
7	ResNet50 +AGG	77.59	91.13	94.55	10.3	109.65
8	EfficientNet-B0 + AGG	81.35	90.91	93.70	7.5	66.36
9	Swin2-B + AGG	76.32	88.85	92.29	89.7	21.35

由表 3 可知，本研究 DNN 网络具有最高的位置识别召回率。与模型 7 相比，模型 4 的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别提高 9.21、4.02 和 2.46 个百分点，与模型 8 相比则分别提高 5.45、4.24、3.31 个百分点。与 ResNet50 和 EfficientNet-B0 的纯 CNN 结构不同，DNN 融合了 CNN 和 Transformer 的优势，在有效提取场景图像局部细节信息的同时，还能在局部特征中融合全局环境特征，进而提高了全局描述符的语义可分性。与纯 CNN 相比，DNN 仍具有较大的参数量和计算延时，但由于优化设计了 MHSA，计算速度仍达 55.65 帧/s，具有一定的实时性。对比模型 4 与模型 9，前者的 $R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 比后者分别高出 10.48、6.3 和 4.72 个百分点，在参数量相近的情况下，前者计算速度是后者的 2.6 倍。相较于 Swin2-B 的纯 Transformer 架构，DNN 在确保建立起局部特征间相关性的同时，还具有 CNN 的图像局部细节特征提取能力。在 4 种模型中，模型 9 的识别性能最低，说明在 VPR 任务中，图像局部细节特征是构造图像全局语义信息的最基本元素。以上分析表明，本研究设计的 DNN 特征提取网络是有效的，能够进一步提高温室 VPR 性能。

3.3 温室视觉位置识别影响因素分析

相机视角、光环境、场景变化等一直是影响 VPR 性能的主要因素。温室环境狭窄，光环境复杂，作物动态生长，种植管理等均会使同一场景图像先后发生显著改变。为探究本研究模型（表 2 模型 6）对多种因素变化的鲁棒性，进一步开展采样距离、成像视角、光照条件、作物生长影响因素变化试验分析。

3.3.1 采样距离变化影响分析

相机离目标场景的远近，影响图像中目标物的大小及所含场景范围。在温室场景测试集中，选取采样时单纯距离变化的数据集。限于有限的行距（约 1.2 m），采集样本时，以离一侧株行 50 cm 处为基准点，相机视线垂直于株行并成像，将滑轨上的相机逐渐远离株行至 100 cm 处，即采样距离变化为 50 cm，过程中不断成像，以基准点处成像为参考集，为便于问题分析，对变化范围进行近似三等分，即分别以＞50～65、＞65～80、＞80～100 cm 范围内成像为查询集，分析采样距离变化对本研究模型的影响，结果如表 4。

表 4　采样距离变化对视觉位置识别性能影响分析

Table 4　Analysis of the impact of sampling distance shifts on visual place recognition performance　%

采样距离 Sampling distance/cm	$R_{@1}$	$R_{@5}$	$R_{@10}$
＞50 ～ 65	93.75	98.75	99.17
＞65 ～ 80	87.08	97.92	99.17
＞80 ～100	63.30	80.85	87.77

表 4 数据表明，随着采样距离增加，位置识别召回率逐渐下降。当采样距离从基准点附近（＞50～65 cm）增加近 1 倍（＞80～100 cm）时，$R_{@1}$、$R_{@5}$ 和 $R_{@10}$ 分别下降超 30、17 和 11 个百分点。当采样距离扩大时，出现在查询图像中的场景范围超过参考图像，查询图像中出现了更多的场景目标物，表征语义的全局描述符会发生大的变化，与参考集的匹配成功率下降，使视觉位置识别召回率降低。在试验采样距离变化范围内，本研究方法的 $R_{@1}$ 均在 63% 以上，模型对采样距离变化表现出了一定的鲁棒性。

3.3.2　相机视角变化影响分析

智能农机等在进行位置识别时，所获查询图像与既有参考图像的视角很难相同。视角的变化会使图像中的场景范围、目标物、遮挡关系发生改变。从测试集中选取单纯视角变化的场景样本，这些样本在采样时，相机置于三脚架顶部具刻度的载物台上，以当前相机视角为基准点（0 点）进行采样，接着将相机向一侧旋转 45°（超过该范围时，图像间已无交叉场景），过程中不断成像，以基准点处成像为参考集，对视角变化范围进行三等分，即分别以＞0°～15°、＞15°～30°、＞30°～45°范围内成像作为查询集，分析视角变化对本研究模型的影响，结果如表 5。

表 5　相机视角变化对视觉位置识别性能影响分析

Table 5　Analysis of the impact of camera viewport shifts on visual place recognition performance　%

视角 Viewpoint/(°)	$R_{@1}$	$R_{@5}$	$R_{@10}$
＞0～15	95.30	98.66	99.33
＞15～30	72.60	87.72	92.70
＞30～45	55.75	78.76	82.30

表 5 数据表明，当相机视角变化逐渐增大时，识别性能逐渐下降。当视角变化范围在＞0°～15°（小幅视角变化）时，位置识别 $R_{@1}$ 不低于 95%，模型对小幅视角变化具有良好的鲁棒性。但当视角变化超过 30°（大幅视角变化）时，$R_{@1}$ 大幅下降超 39 个百分点，相机视角的大幅变化使模型的识别性能迅速下降。随着相机视角变化逐渐增大，查询图像与参考图像的交叉场景逐渐变小，二者的语义相似性逐渐降低，使召回率下降。总体上，相机视角变化会对本研究模型性能产生较大影响，提高模型对视角变化的适应性将是下一步工作重点。

3.3.3　光照变化影响分析

根据采样时的记录，从温室场景测试集中选取不同光照条件的场景样本。对于具有光照变化的场景样本，将场景图像中强光照样本定义为查询集，弱光照样本定义为参考集，然后交换角色。对于仅有强光照的场景样本，随机选择 1 幅图像作为参考，其他样本作为查询集，对于仅有弱光照的场景样本，也采用此种处理。测试本研究模型在各试验组合下的召回率，结果如表 6。

表 6　光照变化对视觉位置识别性能影响分析

Table 6　Analysis of the impact of sunlight shifts on visual place recognition performance　%

查询集光照 Sunlight of query set	参考集光照 Sunlight of reference set	$R_{@1}$	$R_{@5}$	$R_{@10}$
强 Intense	弱	82.02	91.01	93.82
弱 Weak	强	88.89	95.91	97.66
强 Intense	强	80.34	88.20	90.45
弱 Weak	弱	88.30	95.32	96.49

注：强光照指晴朗天气上午和中午的光照条件；弱光照指日落后及阴天的光照条件。

Note: Intense sunlight occurs during morning and noon on sunny days; Weak lighting refers to after sunset and on cloudy days.

表 6 数据表明，在不同的查询集和参考集光照条件下，视觉位置识别 $R_{@1}$ 指标值均超过 80%，本研究模型具有良好的光照变化适应性。光照条件变化主要对成像结果的颜色空间产生影响，由于 DNN 特征提取网络中的多种归一化处理，克服了光照条件变化对图像亮度、饱和度等颜色空间上的影响，同时模型充分提取并聚合了图像的深层语义信息，使光照变化对场景位置识别性能的影响弱于其他因素。对比表 6 的后 2 行数据可知，模型在弱光照条件下的 VPR 性能高于强光照条件。强光照条件下，相机在特定视角下会出现成像过曝光问题，丧失了图像局部区域的纹理细节，不利于模型提取稳定的场景语义特征，使模型识别性能下降。总体上，本研究模型对光照条件变化具有良好的鲁棒性，弱光照条件更易于位置识别。

3.3.4　作物生长变化影响分析

作物生长及生长过程中的生产管理，使不同时间采集自同一场景的图像发生很大变化。为探究在作物生长变化中，做场景位置识别是否仍有可能，从测试集中选取具有作物生长变化的场景样本，这些样本在采集时，以第 1 天成像为基准点，此后分别于第 5 天、第 9 天和第 13 天在相同场景进行采样。以第 1 天的场景图像作为参考集，分别以第 5、9 和 13 天的场景图像作为查询集，测试本研究模型的识别效果，结果如表 7。

表 7　作物生长对视觉位置识别性能影响分析

Table 7　Analysis of the impact of crop growth on visual place recognition performance　%

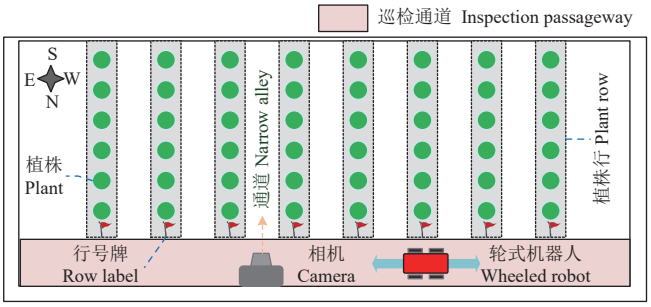
查询集 Query set	$R_{@1}$	$R_{@5}$	$R_{@10}$
第 5 天样本 Day 5 samples	49.75	60.10	67.49
第 9 天样本 Day 9 samples	41.60	52.10	56.72
第 13 天样本 Day13 samples	36.07	45.90	51.64

由表 7 可知，作物生长变化对本研究模型 VPR 性能有较大影响，第 5 天时，$R_{@1}$ 已不足 50%，随着时间增长，模型识别性能持续下降。在作物生长过程中，植株形态会发生变化，如果实不断成熟并被采摘，生产管理不断梳理掉老龄叶子等，使同一场景的语义信息发生较大变化。作物生长变化及生产管理对场景位置识别性能影响大，在动态变化的温室场景下执行 VPR 任务仍具挑战性。

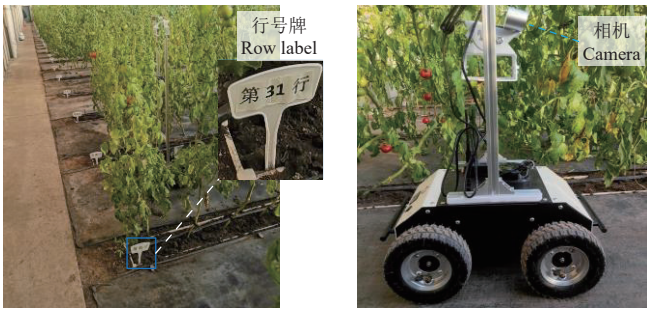
3.4　温室视觉位置识别测试试验

为进一步评估本研究模型是否具有实用性，在一栋

种植作物相同、管理模式类似的日光温室（与场景数据集构建所在温室不同）内，进一步开展位置识别测试试验。温室种植结构示意如图 5a，在植株行首端设置行号牌（图 5b），作为位置标签。首先以人工方式在两株行间通道首端位置应用相机成像，视线平行于株行并微下俯，保证两侧植株行号牌出现在图像中。在所有株行间通道首端成像，构建参考图像集（位置标签已知）。



a. 温室种植结构示意图
a. Schematic diagram of greenhouse planting structure



b. 测试场景
b. Test scenario

c. 轮式机器人
c. Wheeled robot

图 5 温室视觉位置识别测试
Fig.5 Greenhouse visual place recognition test

用轮式机器人小车（图 5c）携带相机，视线同样平行于株行并微下俯，在巡检通道沿东西向在同一路段上遥控行进，过程中拍摄连续视频。共进行 4 次拍摄，每次微调相机高度和俯仰角，得到 4 条视频，编程提取图像帧，构建 4 个独立的查询图像集，用本研究模型分别做 VPR 测试，为每个查询图像保留最相似的参考图像。进一步用人工方式判读查询图像I_q是否与所得参考图像I_r为同一场景，判别标准是两图像中是否出现了共同的行号牌，例如I_q中有$\{a,b,\cdots\}$号，I_r中有$\{b,c\}$号，则判定为识别成功。将每条视频中被成功识别的帧数，与该视频总帧数的比值称为识别率，结果如表 8。

表 8 温室视觉位置识别试验结果
Table 8 Results of greenhouse visual place recognition

视频编号 Video No.	总帧数 Total frames/帧	识别帧数 Recognition frames/帧	识别率 Recognition rate/%
I	360	307	85.28
II	457	376	82.28
III	469	414	88.27
IV	515	422	81.94

由表 8 数据可知，4 次测试的平均识别率在 84% 以上，本研究方法具有较好的视觉位置识别性能。实际测试是在与训练样本采集温室不同的其他温室中进行的，表明本研究方法具有良好的泛化能力。4 次试验均在同一路段采用相同帧率进行采样，由采集视频的总帧数可知，各次试验机器人行进速度有一定差异，对比识别率，未发现速度对识别性能的影响。总体来看，本研究模型在实际温室场景中的视觉位置识别率不低于 81.94%，具有一定的实际应用能力。

4 结 论

构建了基于最优传输特征聚合的温室视觉位置识别方法，开展温室视觉位置识别试验，主要结论如下：

1）本文模型在温室视觉位置识别上是有效的，top-1 召回率（$R_{@1}$）达到 88.96%，与 NetVLAD、MixVPR 和 EigenPlaces 相比，$R_{@1}$ 指标分别提高 29.67、2.97 和 2.89 个百分点。

2）基于最优传输的局部特征聚合及全局特征描述符生成方法能有效提高位置识别精度，在骨干网络相同的前提下，与 NetVLAD 和 MixVPR 的聚合器相比，本研究聚合器的 $R_{@1}$ 分别提高 21.65 和 1.09 个百分点。

3）结合 CNN 和 Transformer 特点设计的局部特征提取网络，能有效提高温室视觉位置识别性能，其 $R_{@1}$ 指标相比优秀的独立 CNN 和 Transformer 网络分别提高 5.45 和 10.48 个百分点。

4）采样距离、相机视角、光照变化、作物动态生长变化等因素影响位置识别性能，当采样距离增加 1 倍时，$R_{@1}$ 下降超 30 个百分点，模型对光照变化具有较好的鲁棒性，作物生长对位置识别性能影响大，5 天的作物生长变化，使模型 $R_{@1}$ 下降到不足 50%，动态变化的温室场景对视觉位置识别任务仍具挑战性。

本文方法在实际温室中的位置识别率不低于 81.94 %，具有一定的实际应用能力，研究结果可为温室智能农机装备视觉系统设计提供技术参考。

［参 考 文 献］

[1] 满忠贤，何杰，刘善琪，等. 智能农机多机协同收获作业控制方法与试验[J]. 农业工程学报，2024，40(1)：17-26.
MAN Zhongxian, HE Jie, LIU Shanqi, et al. Method and test for operating multi-machine cooperative harvesting in intelligent agricultural machinery[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(1): 17-26. (in Chinese with English abstract)

[2] 陈明方，黄良恩，王森，等. 移动机器人视觉里程计技术研究综述[J]. 农业机械学报，2024，55(3)：1-20.
CHEN Mingfang, HUANG Liang'en, WANG Sen, et al. Survey of research on visual odometry technology for mobile robots[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(3): 1-20. (in Chinese with English abstract)

[3] LOWRY S, SUNDERHAUF N, NEWMAN P, et al. Visual place recognition: A survey[J]. IEEE Transactions on Robotics, 2015, 32(1): 1-19.

[4] YUAN F M, SCHUBERT S, PROTZEL P, et al. Local positional graphs and attentive local features for a data and runtime-efficient hierarchical place recognition pipeline[J]. IEEE Robotics and Automation Letters, 2024, 9(3): 2686-2693.

[5] 周浩然，刘成菊，安康，等. 基于物体路标的仿人机器人实时里程计[J]. 信息与控制，2021，50(2)：204-215.
ZHOU Haoran, LIU Chengju, AN Kang, et al. Real time odometer for humanoid robot based on object landmarks[J].

Information and Control, 2021, 50(2): 204-215. (in Chinese with English abstract)

[6] 吴雄伟，周云成，刘峻淳，等. 面向温室移动机器人的无监督视觉里程估计方法[J]. 农业工程学报，2023，39(10)：163-174.
WU Xiongwei, ZHOU Yuncheng, LIU Junting, et al. Unsupervised visual odometry method for greenhouse mobile robots[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2023, 39(10): 163-174. (in Chinese with English abstract)

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

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

[9] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]//2011 International Conference on Computer Vision (ICCV), November 6-13, 2011, Barcelona, Spain. Piscataway, NJ: IEEE, 2011: 2564-2571.

[10] 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.

[11] 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 Recognition (CVPR), June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 3304-3311.

[12] PERRONNIN F, LIU Y, SANCHEZ J, et al. Large-scale image retrieval with compressed Fisher vectors[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 3384-3391.

[13] 谭厚森，马文宏，田原，等. 基于改进 YOLOv8n 的香梨目标检测方法[J]. 农业工程学报，2024，40(11)：178-185.
TAN Housen, MA Wenhong, TIAN Yuan, et al. Improved YOLOv8n object detection of fragrant pears[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(11): 178-185. (in Chinese with English abstract)

[14] 张羽丰，杨景，邓寒冰，等. 基于 RGB 和深度双模态的温室番茄图像语义分割模型[J]. 农业工程学报，2024，40(2)：295-306.
ZHANG Yufeng, YANG Jing, DENG Hanbing, et al. Semantic segmentation model for greenhouse tomato images using RGB and depth bimodal[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(2): 295-306. (in Chinese with English abstract)

[15] BABENKO A, SLESAREV A, CHIGORIN A, et al. Neural codes for image retrieval[C]//13th European Conference on Computer Vision (ECCV), September 6-12, 2014, Zurich, Switzerland. Berlin: Springer-Verlag, 2014: 584-599.

[16] 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.

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

[18] RADENOVIC F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655-1668.

[19] 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.

[20] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: Learning robust visual features without supervision[EB/OL]. (2023-11-15) [2023-04-14]. https://arxiv.org/abs/2304.07193.

[21] ALI-BEY A, CHAIB-DRAA B, GIGUERE P. MixVPR: Feature mixing for visual place recognition[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 3-7, 2023, Waikoloa, Hawaii, USA. Los Alamitos: IEEE Computer Society, 2023: 2997-3006.

[22] BONNEEL N, DIGNE J. A survey of optimal transport for computer graphics and computer vision[J]. Computer Graphics Forum, 2023, 42(2): 439-460.

[23] SARLIN P E, DETONE D, MALISIEWICZ T, et al. SuperGlue: Learning feature matching with graph neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 4937-4946.

[24] CUTURI M. Sinkhorn distances: Lightspeed computation of optimal transport[J]. Advances in Neural Information Processing Systems, 2013, 26: 2292-2300.

[25] GU J X, WANG Z H, JASON K, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), December 4-9, 2017, Long Beach, CA, USA. New York: Curran Associates Inc, 2017: 5999–6009.

[27] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C/OL]//9th International Conference on Learning Representations (ICLR), May 3-7, 2021, Virtual Only Conference, Online. (2023-10-27) [2021-06-03] https://arxiv.org/abs/2010.11929.

[28] YU W H, LUO M, ZHOU P, et al. MetaFormer is actually what you need for vision[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-24, 2022, New Orleans, LA, USA. New York: IEEE, 2022: 10809-10819.

[29] LI Y Y, HU J, WEN Y, et al. Rethinking vision transformer for MobileNet size and speed[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV), October 2-6, 2023, Paris, France. Los Alamitos: IEEE Computer Society, 2023: 16843-16854.

[30] GONG C Y, WAND D L, LI M, et al. Nasvit: Neural architecture search for efficient vision transformers with gradient conflict-aware supernet training[C/OL]//10th International Conference on Learning Representations (ICLR), April 25-29, 2022, Virtual Only Conference, Online. (2023-10-15) [2022-03-28]. https://openreview.net/pdf?id=Qaw16njk6L.

[31] WANG X, HAN X T, HUANG W L, et al. Multi-Similarity loss with general pair weighting for deep metric learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 5017–5025.

[32] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An

imperative style, high-performance deep learning library[C]// 33rd Annual Conference on Neural Information Processing Systems, December 8-14, 2019, Vancouver, BC, Canada. Cambridge: MIT Press, 2019: 8026-8037.

[33] CUBUK E D, ZOPH B, SHLENS J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 3008-3017.

[34] 李康宇，王西峰，朱守泰. 基于局部与全局描述符紧耦合联合决策的机器人分层式视觉位置识别方法[J]. 信息与控制，2024，53(3)：400-415.
LI Kangyu, WANG Xifeng, ZHU Shoutai, et al. Hierarchical visual place recognition approach for robots based on joint decision-making of tightly-coupled local and global descriptors[J]. Information and Control, 2024, 53(3): 400-415. (in Chinese with English abstract)

[35] BERTON G, MASONE C, CAPUTO B. Rethinking visual geo-localization for large-scale applications[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-24, 2022, New Orleans, LA, USA. New York: IEEE, 2022: 4868-4878.

[36] BERTON G, TRIVIGNO G, CAPUTO B, et al. EigenPlaces: Training viewpoint robust models for visual place recognition[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV), October 2-6, 2023, Paris, France. Los Alamitos: IEEE Computer Society, 2023: 11046-11056.

[37] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[38] TAN M X, LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//36th International Conference on Machine Learning (ICML), June 9-15, 2019, Long Beach, CA, USA. New York: ACM, 2019: 10691-10700.

[39] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 11-17, 2021, Montreal, QC, Canada. Los Alamitos: IEEE Computer Society, 2021: 9992-10002.

Recognizing visual position in the greenhouse using optimal transport feature aggregation

HOU Yuhan , ZHOU Yuncheng※ , LIU Zeyu , ZHANG Runchi , ZHOU Jinqiao

(*College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110866, *China*)

Abstract: As the foundation for implementing closed-loop detection within the realm of visual SLAM (simultaneous localization and mapping), visual place recognition (VPR) has great potential in various applications of greenhouse robot navigation and other fields. However, the existing VPR cannot fully meet the actual requirements of greenhouse scenes due to the complexity and constant variations in the greenhouse environment. In particular, the local feature aggregation paradigm strongly depends on the induction bias of training samples in VPR models, which leads to the issue of information redundancy during feature aggregation. In this study, a greenhouse VPR was presented, according to the optimal transport of local feature aggregation. The process of aggregating local features into a global descriptor was framed as an optimal transport problem, where the cost matrix was predicted through an MLP (multi-layer perceptron). Thus, a cost matrix was dynamically generated using the local features that was extracted from the greenhouse scene images. Additionally, a 'dustbin' cluster was introduced into the cost matrix to allocate the redundant features. Taking the cost matrix as the input, the Sinkhorn algorithm was employed to determine an optimal solution to the assignment matrix. Furthermore, the soft assignment of local features to various clusters was achieved through the assignment matrix. Ultimately, the assignment was concatenated to form a global descriptor for the scene image, which was used for place recognition. A deep neural network (DNN) was optimized and designed to serve as the backbone for local feature extraction of greenhouse scene images, by combining the advantages of CNN (convolutional neural network) and Transformer. Furthermore, cosine similarity was used as the metric function to calculate the similarity measure between scene image global descriptors, so as to perform descriptor matching. A series of experiments were conducted in a tomato greenhouse. The experimental results showed that the improved model achieved better performance. The top-1 recall rate ($R_{@1}$) for place recognition was achieved at 88.96%, which was 29.67, 2.97, and 2.89 percentage points higher than the those of NetVLAD, MixVPR, and EigenPlaces models, respectively. When compared to the aggregators employed in MixVPR and NetVLAD, our aggregator achieved improvements in $R_{@1}$ by 1.09 and 21.65 percentage points, respectively, showcasing its effectiveness. Compared with the CNN, the improved network achieved an increase of 5.45 percentage points in $R_{@1}$. There was even more pronounced $R_{@1}$ improvement (reaching 10.48 percentage points), compared with a Transformer network. Simultaneously, our network resulted in a 1.6-fold increase in computation speed compared to the previous Transformer. In addition, the experiments further demonstrated that the improved model exhibited excellent performance of place recognition and strong robustness when dealing with factors, such as small sampling distance shifts, small viewpoint shifts, and different sunlight intensities. The greenhouse VPR achieved a place recognition rate of no less than 81.94% in actual greenhouses, indicating its practical application potential. The method based on optimal transport of local feature aggregation and global descriptor generation was effective for place recognition, and the image local feature extraction network can boost the performance of place recognition. These findings can provide technical support to the visual systems of intelligent agricultural machinery in the greenhouse.

Keywords: greenhouse; visual place recognition; optimal transport; feature aggregation; deep neural network; Transformer