引用格式：毕盛, 林华伟, 董敏. 多特征融合的目标物体导航方法 [J]. 电子科技大学学报, 2025, 54(3): 401-410.
BI S, LIN H W, DONG M. Target object navigation method based on multi-feature fusion[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(3): 401-410.

多特征融合的目标物体导航方法

毕 盛*，林华伟，董 敏

（华南理工大学 计算机科学与工程学院，广州 510006）

摘要：目标物体导航是在未知的环境中根据视觉观察到达预期的目标物体。其中，如何从视觉观察中找到目标物体的方向是至关重要的。针对这一问题，提出一种基于多特征融合的目标物体导航方法。该方法通过特征融合模块融合包含导航环境整体信息、局部信息的视觉特征和指代目标物体语义的文本特征，得到表征导航方向的方向特征和导航环境的环境特征，将视觉表示与导航方向相关联，从而指导导航动作的生成，约束代理朝目标物体方向导航，提高模型的导航成功率和效率。AI2-Thor 数据集上的实验表明，和基准模型对比，导航成功率 SR 提升 11.7%、导航成功路径长度加权比率 SPL 提升 0.093；和目前先进的方法对比，SR 提升 2.1%、SPL 提升 0.008。实验结果证明了该方法的准确性和高效性。

关键词：目标物体导航；多特征融合；多头注意力机制；深度强化学习
中图分类号：TP183 **文献标志码：**A DOI: 10.12178/1001-0548.2024122

Target object navigation method based on multi-feature fusion

BI Sheng*, LIN Huawei, and DONG Min

(School of Computer Science & Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: Target object navigation is the process of reaching the expected target object based on visual observation in an unknown environment. In this process, it is crucial to find the direction of the target object from visual observation. A target object navigation method based on multi-feature fusion is proposed to address this issue. Using a feature fusion module to fuse visual features that contain the overall and local information of the navigation environment and the text features that refer to the semantics of the target object, the method obtains the directional features that represent the navigation direction and environmental features of the navigation environment. And then this method associates the visual representation with the navigation direction, thereby guiding the generation of navigation actions, constraining the agent to navigate towards the direction of the target object, and improving the success rate and efficiency of the model's navigation. Experiments on the AI2 Thor dataset show that compared to the benchmark model, the navigation success rate SR has increased by 11.7 percentage points, and the navigation success path length weighted ratio SPL has increased by 0.093; Compared with current advanced methods, SR has increased by 2.1 percentage points and SPL has increased by 0.008. The experimental results have demonstrated the accuracy and efficiency of this method.

Key words: target object navigation; multi feature fusion; multi head attention mechanism; deep reinforcement learning

目标物体导航[1]是视觉导航[2]中的一个重要子任务和研究方向，导航环境是在室内家居场景中[3]，其主要目的是利用代理对环境的视觉观察到达给定目标物体，该代理一般是部署导航模型的机器人。所以在目标物体导航任务中，需要构建一个模型能够建立视觉观察和目标物体的强关联，从而指引代理在视觉观察中找到目标物体所处的位置方向。现有的目标物体导航方法强调从视觉观察中提取物体语义、位置、物体关系等丰富的视觉表示，然后和通过独热编码[4]或者预训练[5]的嵌入层网络获得的目标物体特征简单拼接，其主要关注点在于通过模型构建丰富的视觉表示，以此来告诉模型周围的环境是什么，但是缺乏建立视觉表示和目标物体语义特征的强关联，进而影响导航的成功率

收稿日期：2024−05−22
基金项目：广东省科技计划（2020A0505100015）；广东省高校教师特色创新研究项目（2022DZXX03）
作者简介：毕盛，博士，副教授，主要从事智能机器人方面的研究。
*通信作者 E-mail: picy@scut.edu.cn

和效率。因此，如何识别出当前视觉观察中和目标物体关联的区域的位置方向信息，从而指导代理向目标物体方向导航是提高导航成功率和导航效率的关键。

本文采用多特征融合[6]的方法实现从视觉表示中找到目标物体方向的目的，多特征融合的目标物体导航方法根据目标物体名称的指导在视觉观察中找到目标物体或者和目标物体同时出现可能性最大的物体，并根据该物体的方向产生导航动作，执行动作朝物体前进。该方法先将视觉观察特征和目标物体语义特征嵌入同一维度的语义空间，并通过设计的独特 Transformer[7]架构利用多头自注意力机制[8]构建两者的关联，从而不仅告诉代理周围的环境是什么，而且指导了代理找到目标物体所在的位置方向。具体来说，该方法由特征提取模块、特征融合模块和策略学习模块构成。首先是特征提取模块，包括利用预训练的 ResNet18 网络[9]提取表征视觉观察整体信息的全局特征、预训练的 DETR 网络[10]提取表征视觉观察局部语义和位置信息的局部特征和可学习的嵌入层网络提取指代目标物体的目标特征；然后是特征融合模块融合全局特征、局部特征、目标特征得到方向特征和环境特征；最后，通过策略学习模块利用历史上下文信息调整方向特征和环境特征，并根据两个特征生成导航动作。在正式训练之前，本文还通过模仿学习[11]的方法对模型预训练，从而使模型获得一个较好的初始化权重，缓解目标物体导航的稀疏奖励[12]问题，为特征融合模块快速学习并收敛提供良好的前提条件。

1 相关工作

视觉导航作为一项机器人和人工智能交叉领域的重要任务，近年来引起了越来越多的关注，并且得到了快速发展，在自动驾驶汽车[13]、智能家居[14]等领域都有广泛的应用，可以提高效率并丰富用户体验。传统的视觉导航，依赖于利用同步定位与建图（simultaneous cocalization and mapping, SLAM）[15]的方法事先构建地图，在导航过程中利用视觉特征定位自身位置和目标点位置进行点对点导航。文献 [16] 提出了一种单目视觉 SLAM，该模型不需要明确提取和匹配特征点，而是直接利用图像的灰度信息来估计相机的运动和场景的结构。文献 [17] 提出了 ORB-SLAM2，是一种基于特征点的视觉 SLAM 系统。它结合了特征提取、匹配、姿态估计和地图优化等关键步骤，实现了高效的实时定位和地图构建。但是，面对各种各样未知的家居环境，传统的方法往往束手无策，近年来研究人员使用深度强化学习的方法在家居环境下进行视觉导航。

基于视觉编码器的目标物体导航方法主要关注如何从环境中提取有效的视觉特征，以帮助导航任务中的目标识别与方向定位。文献 [18] 使用 ResNet 网络提取视觉图像的全局特征，通过深度强化学习模型进行导航。然而，仅依赖全局特征，缺少对局部物体的语义与位置信息的细化处理，在复杂环境中的细节处理能力不足。文献 [19] 提出的物体关系图将全局特征与局部特征相结合，使用 faster-RCNN[20]提取局部物体特征，较为全面地捕捉了环境中的重要信息。这种结合全局与局部特征的方法，使得导航在复杂环境中表现更为出色。文献 [21] 基于空间注意力机制，计算视觉特征与语义、动作和历史信息特征的相似性，并根据这些相似性分配注意力权重，从而突出环境中重要区域的视觉特征。但该方法缺乏局部物体特征，影响模型感知环境并确定目标物体导航方向。文献 [22] 通过构建层次化的对象到区域图来提取导航过程中与目标区域相关的特征，但其对局部物体的语义信息的提取较弱，限制了在复杂环境中的细节处理。文献 [23] 使用视觉 Transformer[24]来表征环境特征，提升了特征的提取能力，特别是在全局信息的建模上有较大优势。然而，Transformer 模型的应用仅限于视觉特征，对语义和其他类型特征的融合不足。文献 [25] 提出对象记忆 Transformer，使用对象场景存储器存储长时间的场景与物体语义，通过 Transformer 关注历史观察中的对象。虽然该方法在静态环境中表现良好，但其设计依赖历史数据，在动态环境中存在局限。

基于导航策略学习的目标物体导航方法主要致力于如何通过策略学习模型，优化机器人在环境中的行动决策。文献 [26] 提出了基于元强化学习的自适应导航方法，能够在测试过程中自适应调整策略，提升了在新环境中的泛化能力。然而，该方法依赖于 ResNet18 提取的整体视觉特征，缺乏对局部信息的充分利用。文献 [27] 结合元学习与分层语义信息，将上下文向量、目标语义特征与知识图的层次关系特征输入 A3C 网络[28]，提升了策略学习的泛化能力，尤其在不熟悉的环境中表现出色。文献 [29] 通过图 Transformer[30]知识推理网络 GTV，利用对象关系图构建环境中的最佳导航策略，分解

导航动作以提高策略的学习效率。尽管 GTV 能够在复杂场景中生成有效的策略，但其在大规模动态环境中的处理能力仍有局限。

本文方法的创新在于通过视觉和目标文本特征的联合提取实现多模态特征的深度融合，从而更全面地捕捉环境信息，增强了对导航方向的理解。

2　方法

目标物体导航模型如图 1 所示，其中，特征提取模块从第一视角视觉观察中提取表征视觉观察中各个区域物体语义和位置信息的 100×256 局部特征和表征导航环境中整体构造信息的 49×256 全局特征，从目标物体名称中提取表征目标物体语义的 1×256 目标特征；特征融合模块利用设计的独特 Transformer 架构强化并融合局部特征、全局特征和目标特征得到 1×256 的方向特征和 1×256 的环境特征。策略学习模块接收方向特征、环境特征进行导航序列建模并学习导航策略，从而生成得到 1×6 的动作特征和 1×1 的评分特征，动作特征表征离散动作空间的动作概率分布函数，代理根据概率选择一个动作执行；评分特征评估当前状态的价值。

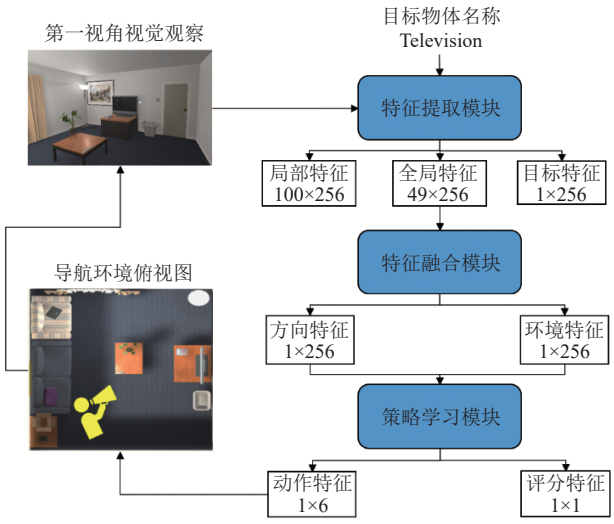


图 1　目标物体导航框架

2.1　特征提取模块

特征提取模块如图 2 所示，本文使用在 COCO 数据集上预训练的 DETR 目标检测网络来提取第一视角 RGB 图像，得到 100×256 的局部特征，再将局部特征输入线性层和 ReLU 激活函数得到 100×249 的局部特征，然后将每一行特征对应的 100×7 关联特征和局部特征连结起来得到 100×256 的局部特征。其中，关联特征通过 DETR 网络的推理结果得到，分别是位置信息、标签信息、置信度和关联因子，关联因子表征当前特征和目标物体的关联程度，如果是目标物体即为 1，否则为 0。接着，使用在 ImageNet 数据集上预训练过的 ResNet18 网络提取第一视角 RGB 图像得到 7×7×512 的全局特征，再使用 1×1 卷积层将全局特征的通道数降维至 256，然后通过正余弦位置编码得到的位置特征和全局特征相加并展平，得到 49×256 的全局特征。正余弦位置编码公式表示为：

$$\mathrm{PE}(i,2j)=\sin\left(\frac{i}{1\,000^{\frac{2j}{d}}}\right) \tag{1}$$

$$\mathrm{PE}(i,2j+1)=\cos\left(\frac{i}{1\,000^{\frac{2j}{d}}}\right) \tag{2}$$

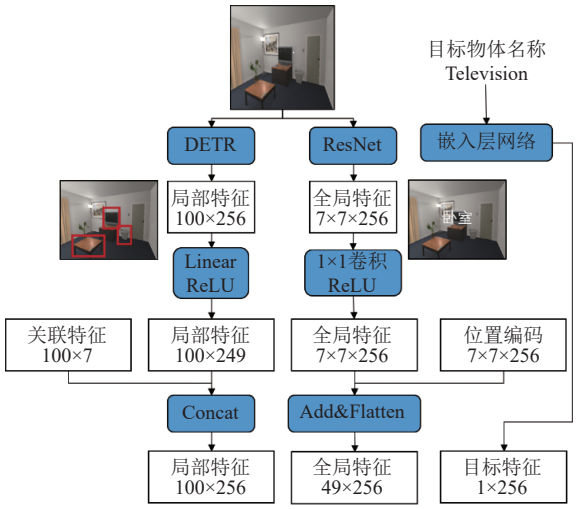


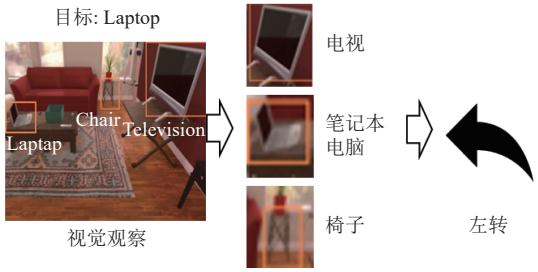
图 2　特征提取模块

给定词汇表的大小 V=22 和嵌入向量的维度 D=256 来初始化嵌入矩阵 $\boldsymbol{E}$，将目标物体名称在词汇表中单词的索引 I 输入嵌入矩阵 $\boldsymbol{E}$ 得到目标特征，目标特征在训练中会被学习，使得相似的信息在嵌入空间中更接近，有利于特征融合模块构建文本语义信息和视觉语义信息的关联。经过特征提取模块，表征局部物体语义和位置信息的局部特征、表征环境整体信息的全局特征和表征目标物体语义的目标特征维度都变成 256 维，处在同一维度的嵌入空间下，帮助后续特征融合模块融合多特征构建视觉观察和目标物体的强关联。

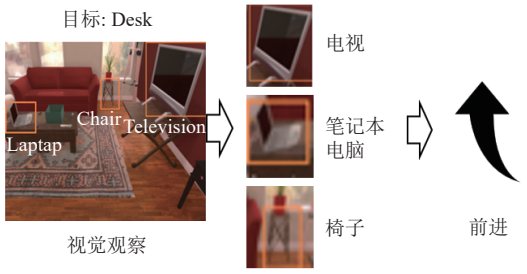
2.2　特征融合模块

在进行室内目标物体导航时，肉眼首先识别当前视角内物体的语义和位置信息，如果发现目标物体，直接导航到目标物体前方；如果没有发现目标物体，就在当前视角的物体中选择和目标

物体共同出现可能性最大的物体方向走去，继续寻找目标物体。如图 3a 所示，当导航目标是笔记本电脑并且在代理第一视角的 RGB 图像中识别出笔记本电脑、椅子和电视，此时，代理执行左转动作朝目标物体方向靠近；但当如图 3b 所示，此时导航目标是桌子而在代理第一视角的 RGB 图像中并没有识别到目标物体时，代理应该选择和桌子同时出现可能性最大的椅子方向前进，因此代理执行前进动作。所以，在目标物体导航过程中，代理应该从第一视角的 RGB 图像中提取全局特征，定位自身目前所处的位置；提取局部特征，感知周围环境中物体语义和位置信息；接着构建目标特征和局部特征、全局特征的强关联，指导生成方向特征；同时，还要融合全局特征和局部特征得到环境表征，构建方向特征和环境特征一一对应的序列，帮助后续模块根据这些序列历史经验信息学习导航策略。



a. 第一视角发现目标物体



电视

b. 第一视角未发现目标物体

图 3 导航示例图

因此，在特征融合模块中引入了本文设计的独特 Transformer 架构，如图 4 所示，该模块分为 encoder 和 decoder 两个部分。

首先是 encoder 部分，本文将局部特征先输入 encoder 网络，输入 encoder 的局部特征是由 100 个维度为 256 的序列构成的，每一个序列都表示环境中局部物体的语义和位置信息，在 encoder 中，每一个序列通过多头自注意力和其他序列交互，学习了环境中物体共同出现的关联信息，有助于在当前视角未发现目标物体时指引机器人朝和目标物体关联的物体方向前进，进而发现目标物体。随后，把强化的局部特征作为键值对输入 decoder 部分。

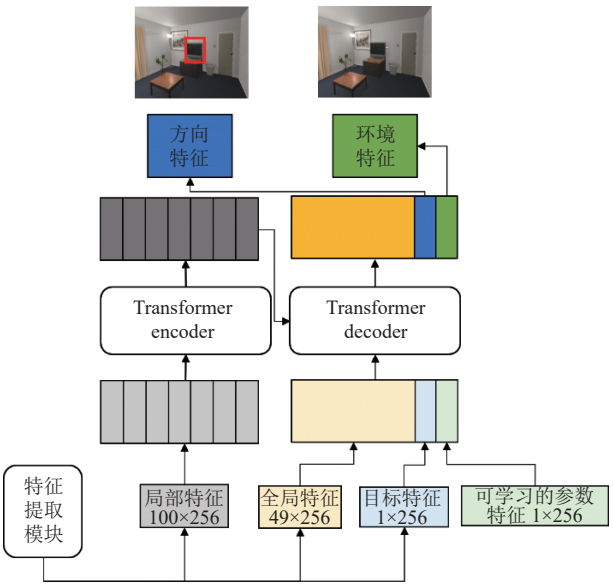


图 4 特征融合模块

在 decoder 部分，本文先将 1×256 的目标特征、1×256 的可学习的参数特征和 49×256 的全局特征连结，得到 51×256 的特征，然后输入 Transformer decoder 网络。再将 51×256 维特征作为查询和 encoder 输出的 100×256 维键值对特征得到 51×256 维的特征，将 1×256 的目标特征、1×256 的可学习的参数特征对应的两个特征输出得到方向特征和环境特征。其中，目标特征通过多头自注意力从全局特征和局部特征中寻找和目标语义特征关联的信息，确定导航方向，从而生成方向特征。可学习的参数特征则通过多头自注意力根据全局特征和局部特征自适应调整表征当前环境，从而生成环境特征。

2.3 策略学习模块

策略学习模块如图 5 所示，本文将特征融合模块输出的方向特征和环境特征展平为 1×512 维的特征，然后将上一时刻 1×6 维的动作特征输入线性层得到 1×64 维的动作特征，最后将两个特征连结得到 1×576 维的特征输入双层 LSTM 网络[31]，得到 1×512 维的状态特征。然后是 A3C 强化学习网络部分，动作生成网络由单层线性层构成，动作评分网络由双层线性层构成；向动作生成网络输入状态特征，网络经过线性变换得到 1×6 维的动作特征；向动作评分网络输入状态特征，网络经过两次线性变换，将特征依次变成 1×64 维的中间特征和 1×1 维评分特征。

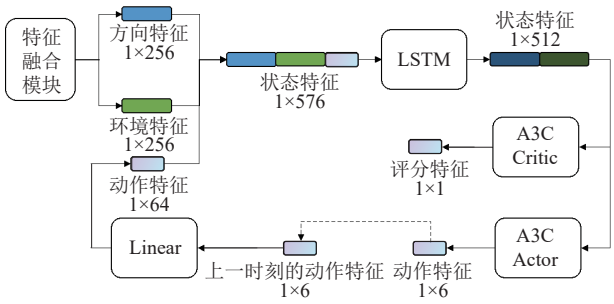


图 5　策略学习模块

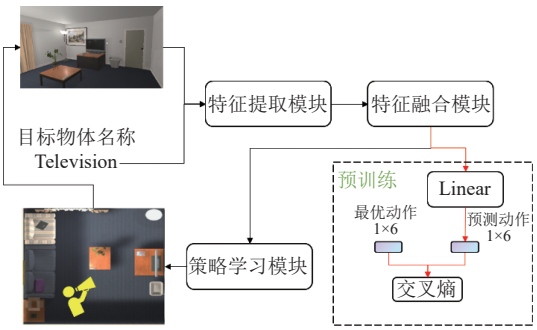


图 6　预训练流程

2.4　训练流程

目标物体导航属于弱强化学习问题，其奖励函数是：

$$R=\begin{cases}5 & \text{机器人到达目标}\\ -0.001 & \text{机器人未到达目标}\end{cases} \tag{3}$$

在执行目标物体导航任务过程中，当监督信号是由强化学习提供的弱奖励时，特征融合模块输出的方向特征和环境特征在导航前期就不能很好地提供正确的方向和环境信息，导致后续的策略学习模块无法学到有效正确的导航策略。因此，本文在正式训练模型前设置了一个预训练阶段，如图 6 所示，采用一种模仿学习的方法来指导代理融合环境视觉特征和文本语义特征，得到方向特征和环境特征，从而给予模型一个较好的初始化权重，便于正式训练时模型快速收敛并学习到有效正确的导航策略。具体来说，在正式训练之前，本文使用预训练数据集 $D_{\text{pre}}=\{(s_0,t_0,a_0),(s_1,t_1,a_1),\cdots,(s_n,t_n,a_n)\}$，其中，$(s_n,t_n,a_n)$表示第 n 条专家导航轨迹；s_n、t_n、a_n分别表示第 n 条专家导航轨迹的视觉观察、目标物体名称和最佳动作。将第 n 条专家导航轨迹的视觉观察、目标物体名称输入特征提取模块、特征融合模块得到方向特征和环境特征，再输入线性层得到 1×6 维预测动作，并通过最小化预测动作和最优动作之间的交叉熵来预训练模型。其公式如下：

$$L(\theta)=\min-\frac{1}{N}\sum_{n=1}^{N}a_n\log\hat{a}_n \tag{4}$$

式中，θ是模型参数；$L(\theta)$表示目标函数；N是预训练数据集数量；a_n是最优动作分布；$\hat{a}_n$是预测动作分布。

预训练完成后，本文将经过模仿学习，初始化权重的特征提取模块和特征融合模块，接入后续的策略学习模块来正式训练整个模型。

3　实验

3.1　数据集

和现有的目标物体导航方法[21-23, 25, 27, 29]一样，本文采用 AI2-Thor 的离线数据集来进行实验，测试本文方法的精度。AI2-Thor 是一个仿真的人工 3D 环境，基于 Unity 3D，可以实现对于物体和场景的模拟。它包含厨房、客厅、卧室、卫生间 4 种场景。在每种类型的场景中，都有 30 个不同的房间，拥有 2 000 个家具物品。本文选择 22 个物体作为导航目标物体，其中包括椅子、笔记本电脑、台灯、桌子和沙发等家居生活常见家具。从厨房、客厅、卧室、卫生间 4 种场景的房间中选择 120 个房间，每个房间内至少存在 4 个物品，选择其中 80 个作为训练集，每个场景包含 20 个房间，将剩余的 40 个房间平均分为验证集和测试集。本文使用验证集上成功率最高的模型在测试集中测试模型精度。

3.2　评价指标

本文通过成功率（success rate, SR）和成功路径长度加权比率 (success weighted by path length, SPL) 来评估模型性能。SR 衡量导航的准确性，SPL 衡量导航效率，其公式如下：

$$\text{SR}=\frac{1}{N}\sum_{n=1}^{N}S_n \tag{5}$$

式中，N 表示导航次数；S_n表示第 n 次的导航指标，导航成功为 1，失败为 0。SPL 表示为：

$$\text{SPL}=\frac{1}{N}\sum_{n=0}^{N}S_n\frac{l_n}{\max\left(l_n,l_{\text{opt}}\right)} \tag{6}$$

式中，l_n 表示第 n 次导航的路径长度；l_{opt} 表示第 n 次导航的最优路径长度。SPL 表示模型导航路径长度和最优导航路径长度的关系，取值范围 0～1，该指标越高，说明导航路径长度越短, 导航效率越高。

3.3 实验参数

本文在预训练阶段，使用预训练数据集训练 10 个 epoch，每个 epoch 迭代 377 次，每次的 batch size 为 1 024。选取测试精度最高的模型作为后续正式训练的模型。正式训练时，使用预训练的模型权重初始化特征提取模块和特征融合模块，使用 Adam 优化器以学习率10^{-5}更新经过预训练的模块，以学习率10^{-4}更新策略学习模块。本文使用 18 个异步代理训练 3 百万次 episode，每 10 万次保存一次模型参数，然后将得到的 30 个模型在验证集测试精度并将精度最高的模型在测试集上测试精度。

3.4 对比实验

3.4.1 对比方法

Random Policy：随机策略模型的代理根据平均的概率分布选择动作空间中的动作，因此，代理将在房间里随机行走或停止。

Baseline：基准模型将特征提取模块得到的视觉特征和文本特征送入策略学习模块学习导航策略，用于对比验证本方法中特征融合模块的作用。

EOTP[21]：利用余弦相似度计算视觉特征分别和语义特征、动作特征、历史信息特征的关联程度作为注意力评分权重分布来表示不同区域视觉特征的重要程度。

HOZ[22]：HOZ 图由场景节点、区域节点和对象节点构成。通过训练构建的 HOZ 图可以在导航过程中定位自身所处的区域和目标所处的区域并规划导航路径。

Vtnet[23]：Vtnet 使用视觉 Transformer 表征环境特征并输入强化学习网络学习导航策略。

OMT[25]：OMT 提出对象场景存储器（OSM）存储长期场景和对象语义，通过注意力机制关注 OSM 中先前观察到的场景和对象序列，从而在室内环境高效导航。

CVVMN[27]：CVVMN 将当前环境的上下文向量表示、目标物体名称的语义特征和知识图表示的层次关系特征连结起来，输入元强化学习网络学习导航策略。

GTV[29]：GTV 通过构建图像中的对象关系图来学习最佳的导航动作，在此过程中模型利用强化学习分解动作，并根据先验知识自适应调整状态表示，从而优化导航策略。

3.4.2 对比结果

和以上流行方法进行对比实验，结果如表 1 所示。本文方法的导航指标在全部路径长度上（ALL）的 SR 为 74.3%、SPL 为 0.457，在路径长度大于等于 5 上（$L\geqslant 5$）的 SR 为 66.3%、SPL 为 0.442，其性能都最好。其中，和基准模型对比，在全部路径长度上 SR 提升 11.7%、SPL 提升 0.093，在路径长度大于等于 5 上 SR 提升 14.8%、SPL 提升 0.097，都超过了基准模型。和当前表现最好的 Vtnet 模型相比，本文方法在全部路径长度上 SR 提升 2.1%、SPL 提升 0.008，在路径长度大于等于 5 上 SR 提升 2.9%、SPL 提升 0.002。这验证了本文提出的多特征融合的目标物体导航方法的有效性和高效性。这是因为本文方法将视觉表征和目标物体特征放入同一嵌入空间，并利用多头自注意力机制融合多个特征，从而构建了视觉表示和目标物体的关联。这在路径长度大于等于 5 的较远距离的导航中提升更加明显，因为此时机器人一般无法在第一视角观察中找到目标物体，需要探索环境定位目标物体，而本文方法在多特征融合过程中学习到了物体关系特征，且建立了视觉表示和目标物体的关联，可以找到当前视角中和目标物体关联的物体，从而提升导航性能。

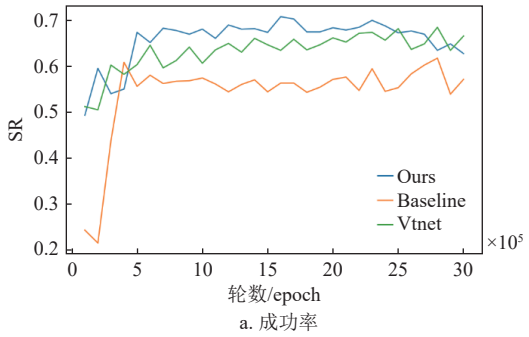
表 1 导航方法性能对比

方法	ALL		$L\geqslant 5$	
	Success/%	SPL	Success/%	SPL
Random	8.1	0.037	0.4	0.002
EOTP	46.2	0.178	32.63	0.159
HOZ	70.6	0.402	62.75	0.392
CVVMN	71.0	0.196	61.92	0.242
OMT	71.1	0.266	\	\
GTV	71.5	0.420	60.10	0.432
Vtnet	72.2	0.449	63.40	0.440
BaseLine	62.6	0.364	51.50	0.345
本文	**74.3**	**0.457**	**66.3**	**0.442**

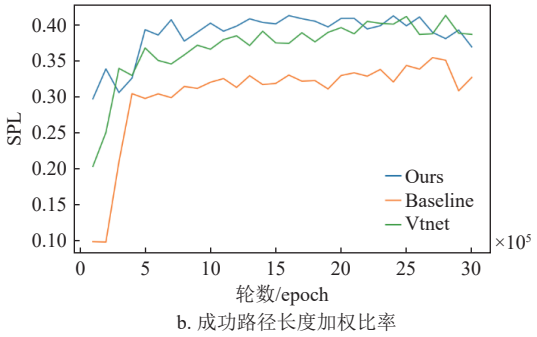
近年来的其他模型，EOTP 只使用全局特征和目标特征，忽略了局部特征对于模型理解环境中局部物体语义和位置的作用，导致代理在导航过程中无法精确定位环境中感兴趣的物体。同时，通过余弦相似度的方法来构建不同模态特征的关联，只考虑了特征向量的方向而忽略了语义内容；余弦相似度假设所有特征具有相同的重要性，但在实际应用中，不同的特征可能具有不同的权重；余弦相似度基于特征向量之间的夹角，因此无法捕捉特征之间的非线性关系。HOZ 将场景中的物体映射到不同区域并进行区域间的导航。但是，一方面忽略了区

域内部的导航问题；另一方面，目标物体的场景是复杂多变的，当进行未知环境下的导航时，模型可能无法适应新场景的布局。CVVMN 用图卷积网络的方式存储环境中对象的关联信息，GTV 使用图 Transformer存储环境中的对象关联信息，也存在未知环境场景变化导致模型无法适应的问题。Vtnet 使用视觉 Transformer 构建丰富的视觉表示，OMT 使用对象记忆 Transformer 记忆环境中场景和对象语义，它们的主要关注点在于通过模型构建丰富的视觉表示，以此来告诉模型周围的环境是什么，但是缺乏建立视觉表示和目标物体语义特征的强关联。

为了消除实验中可能发生的偶然性因素，本文把本文方法和基准（Baseline）以及近 3 年表现最好的 Vtnet 在训练过程中每 10 万 epoch 保存一次的模型在验证集上测试精度。如图 7 所示，将 3 种方法在测试集上的 SR 和 SPL 对比，可以发现本方法的成功率（SR）和成功路径长度加权比率（SPL）普遍表现都优于其他模型。另外，本方法的模型架构学习速度更快，在 10 万次 epoch 就可以达到约 54% 的 SR 和约 0.35 的 SPL，而同时期的基准模型（Baseline）只有约 26% 的 SR 和约 0.12 的 SPL，Vtnet 有约 55% 的成功率和约 0.24 的 SPL。本方法的收敛速度也更快，在 150 万次 epoch 基本收敛到精度最高值，而 Baseline 和 Vtnet 在 270 万次才基本收敛到精度最高值。这一方面得益于本文模型在正式训练前经过了预训练，给予了模型较好的初始化权重，所以模型在前期训练过程就能学到较好的导航策略；另一方面，本文提出的多特征融合的模型架构能够有效地融合归纳视觉表示和目标物体文本特征，并根据不同的场景和不同的目标物体生成不同的导航方向，从而能够有效地提升导航的成功率和效率。



a. 成功率



b. 成功路径长度加权比率

图 7　测试集导航指标变化图

3.5　消融实验

本文通过消融实验分析本文提出的多特征融合的目标物体导航各个组件的作用，如表 2 所示，本文详细分析了特征融合模块中 encoder、decoder、环境特征、方向特征、局部特征和全局特征的作用。

表 2　组件消融实验结果

作用	ALL		$L\geqslant 5$	
	Success/%	SPL	Success/%	SPL
encoder	72.6	0.447	63.7	0.426
decoder	71.2	0.435	61.6	0.413
环境特征	67.9	0.412	56.7	0.389
方向特征	66.3	0.412	54.5	0.375
局部特征	7.9	0.011	2.5	0.006
全局特征	69.8	0.433	62.1	0.420
本文方法	74.3	0.457	66.3	0.442

encoder：为了检验 encoder 网络对于模型精度的影响，本文直接将局部特征输入 decoder，去除了 encoder 网络，并增加了 decoder 的层数，尽量保持模型参数量的一致。实验结果表明在所有路径长度中 SR 降低 1.7%、SPL 降低 0.010，在路径长度大于 5 中 SR 降低 2.6%、SPL 降低 0.016，由于缺失了 encoder 过程中局部特征序列的交互产生的物体关联特征，影响了模型的导航性能。

decoder：为了检验 decoder 网络对于模型精度的影响，将目标特征、参数特征、全局特征和局部

特征联结，去掉了 decoder 网络，并增加了 encoder 的层数，尽量保持模型参数量的一致。实验结果表明在所有路径长度中 SR 降低 3.1%、SPL 降低 0.022，在路径长度大于 5 中 SR 降低 4.7%、SPL 降低 0.029，这表明本文设计的 encoder 和 decoder 架构更有利于多特征之间的融合。

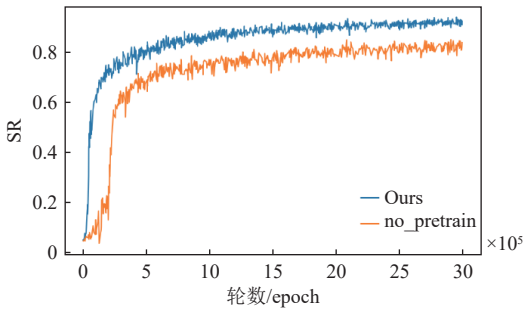
环境特征：为了检验环境特征对于模型精度的影响，decoder 部分不再将参数特征联结到全局特征和目标特征上，特征融合模块只输出目标特征。实验结果表明，在所有路径长度中 SR 降低 6.4%、SPL 降低 0.034，在路径长度大于 5 中 SR 降低 9.6%、SPL 降低 0.053，因为缺失了环境特征对当前环境的归纳，影响了模型对于目标物体方向的感知。

方向特征：为了检验方向特征对于模型精度的影响，decoder 部分本文不再将目标特征联结到全局特征和参数特征上，特征融合模块只输出环境特征。实验结果表明，在所有路径长度中 SR 降低 8.0%、SPL 降低 0.045，在路径长度大于 5 中 SR 降低 11.8%、SPL 降低 0.067，因为缺失了方向特征的指引，模型无法构建视觉表示和正确导航动作的关联。

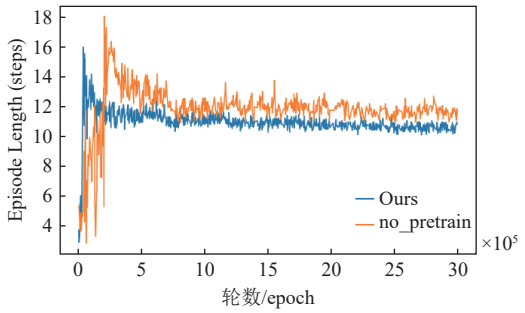
局部特征：为了检验局部特征对于模型精度的影响，去掉了特征提取模块中的提取局部特征部分，将全局特征输入 encoder 并作为键值对输入 decoder 和目标特征、可学习的参数特征融合得到方向特征和环境特征。实验结果表明，模型无法收敛。这说明局部特征所表征的环境中精确的局部物体信息对于导航是至关重要的。

全局特征：为了检验全局特征对于模型精度的影响，去掉了特征提取模块中的提取全局特征部分，decoder 部分本文不再将全局特征联结到可学习的参数特征和目标特征上。实验结果表明，在所有路径长度中 SR 降低 4.5%、SPL 降低 0.024，在路径长度大于 5 中 SR 降低 4.2%、SPL 降低 0.022，由于缺失了全局特征，模型对于整体导航环境和自身在环境中所处位置的感知能力降低了，导致模型精度的下降。

如图 8 所示，分别是经过预训练（Ours）和未经过预训练（no_pretrain）的模型在训练过程中的导航路径长度和导航成功率随 epoch 的变化，未经过预训练的模型学习速度和收敛速度都比较慢，导航成功率也一直较低并且导航路径也相对较长。这是由于目标物体导航任务奖励的稀疏性使得机器人在有限的情况下才能接收到有信息量的反馈，从而较难学到一个正确的导航策略；每次产生导航动作的后果都会被延迟，直到机器人到达目标物体。这种延迟使代理更难将其行动与积极结果联系起来，并学到一个有效的导航策略。此外，模型的特征融合模块是以多层 Transformer 网络为主体的模块，该网络中的自注意力机制允许模型在输入序列的不同位置之间建立长距离的依赖关系，这使得模型更加强大，但也需要更多数据来正确地学习这些依赖关系。因此，当监督信号是由强化学习提供的弱奖励时，特征融合模块输出的方向特征和环境特征不能很好地提供正确的方向和环境信息，导致后续的导航策略学习模块无法学到有效正确的导航策略。



a. 成功率



b. 导航路径长度

图 8 训练集导航指标变化图

本文通过改变特征融合模块中 Head（自注意力头）、Encoder 和 Decoder 层的数量来构建不同的 Transformer 架构。表 3 总结了这些架构的性能。可以观察到，当模型层数太多时，可能无法收敛到最佳策略。另一方面，具有单个编码器和解码器层的变压器不具有足够的网络能力来产生代表性特征。当模型包含 8 个自注意力头、2 层 encoder 和 2 层 decoder 时成功率最高。

表 3　不同参数的 Transformer 的对比

Head	Encoder	Decoder	Accuracy	Success/%	SPL
4	1	1	0.710	70.2	0.423
	2	2	0.712	73.0	0.442
	4	4	0.714	72.0	0.419
8	1	1	0.710	70.1	0.432
	2	**2**	0.720	**74.3**	**0.457**
	4	4	0.715	72.8	0.421
	6	6	**0.722**	73.2	0.430

4　结束语

本文提出了一种基于多特征融合的目标物体导航方法。该方法包括 3 个模块，特征提取模块提取全局特征对整个环境进行全局感知；提取局部特征聚焦于机器人周围的局部区域，用于更精细地感知环境；提取目标特征用于指代目标物体，包含目标物体的语义信息。特征融合模块将不同来源的特征进行融合，使得它们能够在同一维度的语义空间中进行交互和整合。策略学习模块通过强化学习算法，机器人根据每一步的方向特征和环境特征提取上下文信息逐步学习出有效的导航策略，使得在不同环境下能够快速而准确地找到目标物体。其中，特征融合模块通过设计独特的 Transformer 架构，融合目标物体文本特征和视觉观察图像特征得到导航方向。实验结果表明，本文提出的模型能够通过构建视觉观察和目标物体的强关联指导代理朝目标物体方向导航，提高导航的准确率和效率，在目标物体导航方面获得较好的性能。但本文方法目前只在 AI2-Thor 数据集中进行了实验，考虑到该数据集的场景相对较为标准化，不能完全反映现实生活中的复杂环境，未来计划会将本文提出的方法应用到更具多样性和复杂性的仿真环境中进行进一步验证，并最终在真实环境中部署模型，进行导航实验。

参 考 文 献

[1] SUN J, WU J, JI Z, et al. A survey of object goal navigation[J]. IEEE Transactions on Automation Science and Engineering, 2024(3): 1-17.

[2] ZHANG T, HU X, XIAO J, et al. A survey of visual navigation: From geometry to embodied AI[J]. Engineering Applications of Artificial Intelligence, 2022, 114: 105036.

[3] WEN S, LYU X, YU F R, et al. Vision-and-language navigation based on cross-modal feature fusion in indoor environment[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 15(1): 3-15.

[4] TAVARES G M, OYAMADA R S, JUNIOR S B, et al. Trace encoding in process mining: A survey and benchmarking[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107028.

[5] ZOPH B, GHIASI G, LIN T Y, et al. Rethinking pre-training and self-training[J]. Advances in Neural Information Processing Systems, 2020, 33: 3833-3845.

[6] 潘梦竹, 李千目, 邱天. 深度多模态表示学习的研究综述[J]. 计算机工程与应用学报, 2023, 59(2): 48-64.
PAN M Z, LI Q M, QIU T. A review of research on deep multimodal representation learning [J]. Journal of Computer Engineering & Applications, 2023, 59(2): 48-64.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.

[8] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[10] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 213-229.

[11] HUSSEIN A, GABER M M, ELYAN E, et al. Imitation learning: A survey of learning methods[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-35.

[12] OCANA J M C, CAPOBIANCO R, NARDI D. An overview of environmental features that impact deep reinforcement learning in sparse-reward domains[J]. Journal of Artificial Intelligence Research, 2023, 76: 1181-1218.

[13] 许宏鑫, 吴志周, 梁韵逸. 基于强化学习的自动驾驶汽车路径规划方法研究综述[J]. 计算机应用研究学报, 2023, 40(11): 3212-3217.
XU H X, WU Z Z, LIANG Y Y. A survey of autonomous vehicle path planning methods based on reinforcement learning[J]. Application Research of Computers, 2023, 40(11): 3212-3217.

[14] 付心仪, 张鹤, 薛程, 等. 面向未来的智能家居前沿进展[J]. 科技导报, 2023, 41(8): 36-52.
FU X Y, ZHANG H, XUE C, et al. The cutting-edge progress of smart homes for the future[J]. Science & Technology Review, 2023, 41(8): 36-52.

[15] KAZEROUNI I A, FITZGERALD L, DOOLY G, et al. A survey of state-of-the-art on visual SLAM[J]. Expert Systems with Applications, 2022, 205: 117734.

[16] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C]//European Conference on Computer Vision. Zurich: Springer, 2014: 834-849.

[17] MUR-ARTAL R, TARDÓS J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[18] ZHU Y, MOTTAGHI R, KOLVE E, et al. Target-driven visual navigation in indoor scenes using deep

reinforcement learning[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2017: 3357-3364.
[19] DU H, YU X, ZHENG L. Learning object relation graph and tentative policy for visual navigation[C]//Computer Vision–ECCV 2020: 16th European Conference. Glasgow: Springer, 2020: 19-34.
[20] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 1137-1149.
[21] MAYO B, HAZAN T, TAL A. Visual navigation with spatial attention[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 16898-16907.
[22] ZHANG S, SONG X, BAI Y, et al. Hierarchical object-to-zone graph for object navigation [C]//IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 15130-15140.
[23] DU H, YU X, ZHENG L. VTNET: Visual transformer network for object goal navigation[C]//ICLR 2021-9th International Conference on Learning Representations. Minneapolis: [s.n.], 2021: 1-16.
[24] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.
[25] FUKUSHIMA R, OTA K, KANEZAKI A, et al. Object memory transformer for object goal navigation[C]//2022 International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2022: 11288-11294.
[26] WORTSMAN M, EHSANI K, RASTEGARI M, et al. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6750-6759.
[27] LI F, GUO C, ZHANG H, et al. Context vector-based visual mapless navigation in indoor using hierarchical semantic information and meta-learning[J]. Complex & Intelligent Systems, 2023, 9(2): 2031-2041.
[28] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning. New York: ACM, 2016: 1928-1937.
[29] ZHOU K, GUO C, ZHANG H, et al. Optimal graph transformer viterbi knowledge inference network for more successful visual navigation[J]. Advanced Engineering Informatics, 2023, 55: 101889.
[30] SUN C, LI C, LIN X, et al. Attention-based graph neural networks: A survey[J]. Artificial Intelligence Review, 2023, 56(S2): 2263-2310.
[31] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model[J]. Artificial Intelligence Review, 2020, 53(8): 5929-5955.

编 辑 刘飞阳