第 58 卷 第 16 期/2021 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于余弦相似的视觉语言导航算法

金杰[1]，刘凯燕[1]，查顺考[2*]

[1]天津大学电气自动化与信息工程学院，天津 300072；

[2]中国科学技术大学软件学院，江苏 苏州 215123

摘要 为了解决视觉语言导航任务中存在的导航准确率低与泛化能力弱的问题，在 Regretful 模型的基础上，提出了一种基于余弦相似的视觉语言导航算法。通过增加余弦相似损失函数来指导神经网络，学习预测导航方向，减小了特征空间中类内特征的差异，增大了类间特征的分布范围，提升了无搜索策略模型的导航准确率。同时提出了一种全景视图特征平滑方法来进行数据增强，提升了模型的泛化性能。实验结果表明，该算法改善了模型在 R2R(Room-to-room)数据集上的导航准确率等多项指标，效果优于 Regretful 模型，验证了所提方法的优越性与鲁棒性。

关键词 机器视觉；视觉语言导航；余弦相似度；数据增强

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.1615001

Vision-Language Navigation Algorithm Based on Cosine Similarity

Jin Jie[1], Liu Kaiyan[1], Zha Shunkao[2*]

[1] *School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072, *China*;

[2] *School of Software Engineering, University of Science and Technology of China, Suzhou, Jiangsu* 215123, *China*

Abstract This paper proposes a vision-language navigation algorithm based on cosine similarity using the Regretful model to solve the problems of low navigation accuracy and weak generalization ability in vision-language navigation tasks. By increasing the cosine similarity loss function to guide neural network learning and predict navigation direction, the difference in intraclass features in feature space is reduced. The distribution range of interclass features increases, and the navigation accuracy of the model without search strategy improves. Simultaneously, a feature-smoothing method of panoramic view is proposed to enhance data and improve the generalization performance of the model. Experimental results show that the algorithm improves the navigation accuracy and other model indicators on the R2R(Room-to-room) dataset. Additionally, its effect is better than that of the Regretful model, confirming the superiority and robustness of the proposed method.

Key words machine vision; vision-language navigation; cosine similarity; feature enhancement

OCIS codes 150.1135; 100.5010; 330.5000

1 引 言

在三维环境中进行熟练的移动操作是当前人工智能的主要研究课题之一。随着机器学习技术的提升与感知推理方法的持续进步，机器视觉导航任务得到了快速发展。在过去的几年中，这一领域的创造性工作激增[1-7]，复合型任务视觉语言导航(VLN)就是其中一项具有代表性的任务。视觉语义导航是引导智能体在真实三维环境中执行自然语言指令的任务。与其他导航不同的是，智能体仅仅根据自然语言指令信息和实时的周围环境图像信息，就能够在陌生的环境中导航到目标区域，并且该

收稿日期：2020-10-12；**修回日期**：2020-11-30；**录用日期**：2020-12-06

基金项目：国家自然科学基金(61571320)

通信作者：*zhashunkao@gmail.com

过程中智能体能够自动避开障碍物和不可到达的区域。VLN 任务的研究能够使机器人更加智能化和自动化，提供了高效的智能人机交互技术，并且对于一些人类由于特殊原因不能到达的场景如火灾、地震等具有很大的实用价值。

最初的 VLN 模型思想来源于 VQA（Visual Question Answering），Anderson 等[8]提出了一种基于视觉的 Seq-to-Seq 模型，对语言和图像进行编码，然后解码输出动作序列。在基线模型建立后，最近提出了很多 VLN 的方法与模型。Wang 等[9]在 2018 年 3 月提出了 RPA（Reinforced Planning Ahead）模型，结合 model-based 和 model-free 强化学习来学习环境模型，直接优化导航。Fried 等[10]在 2018 年 6 月提出了 Speaker-Follower 模型，生成合成指令，以作为训练过程的数据增强，并在测试时进行语用推理。Ma 等[11]在 2019 年 1 月提出 SMNA（Self-Monitoring Navigation Agent）模型，引入视觉和文本协同注意机制和导航进度估计器，解决了多模态监督的问题。Speaker-Follower 和 SMNA 模型都采用 Beam search 方法[12]进行导航探索，Beam search 的使用极大地提高了准确率，但是基于这种暴力搜索策略的模型并不高效。针对该问题，Ma 等[13]于 2019 年 3 月提出了 Regretful 模型，如果导航错误可以回退到之前的位置继续探索，该方式减少了导航累积误差，进而提高了准确率。

VLN 任务的关键在于有序地感知视觉与文本。传统的方法利用模态接地跨视觉和文本特征来解决这一问题，但是却忽略了环境中包含丰富的语义信息，比如导航图中的信息和已走路径的信息。因此，Zhu 等[14]提出了 AuxRN 模型，该模型使用 4 种自监督辅助任务来帮助智能体更好地获取环境的语义信息。Majumdar 等[15]利用大量“无实体”的 web 视觉和语言语料库来学习视觉基础，从而在这一相对缺乏数据的具体化感知任务上提高了性能。

综上所述，目前的 VLN 模型根据有无搜索策略大致分为两类。有搜索策略的模型导航准确率相对较高，但是其是以牺牲路径长度为代价的，综合指标路径长度加权成功率（SPL）不高，在实际的机器人导航中并不实用，如 Speaker-Follower[10]、SMNA[11]、AuxRN[14]模型；无搜索策略是通过采用自监督模仿学习或者回退策略等方法来平衡导航准确率和路径长度，如 RCM（Reinforced Cross-Modal Matching）[16]、Regretful 模型[13]。

因此，本文主要研究无搜索策略的视觉语言导航任务。在 Regretful 模型的基础上进行算法改进，主要内容包括：

1）引入余弦相似度作为损失函数的约束条件，可以描述实际导航方向与预测方向的误差，提升了无搜索策略模型的导航准确率。

2）通过全景视图特征平滑方法来进行数据增强，提升了模型的泛化性能。

2　原理与方法

2.1　Regretful 模型

本文算法模型是基于 Regretful 模型[13]进行改进的，其算法的组成框架如图 1 所示。主要分为两个部分，视觉语言匹配模块和导航决策模块。视觉语言匹配模块，主要将摄像头获取的全景图像信息与文本指令信息进行多模态信息融合，采用 Attention 机制获取视觉与文本特征，从而根据视觉特征及

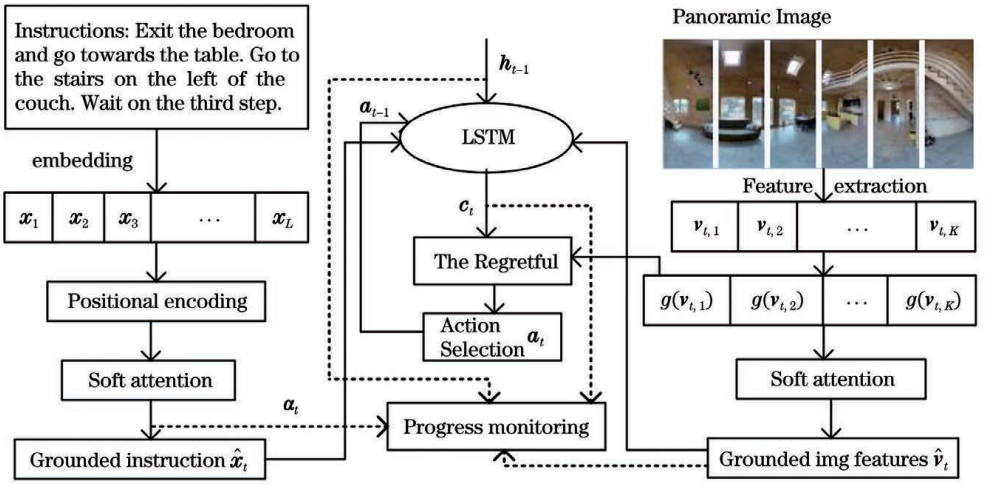

图 1　Regretful 代理框架[13]

Fig. 1　Regretful agent framework[13]

文本的时序信息来判断已完成的指令与下一条指令，为动作选择决策做准备。

具体地，在 t 时刻，获取当前时刻的全景图像 $\boldsymbol{o}_t$，然后进行特征提取，该模型直接获取预训练好的 ImageNet 的 ResNet-152 卷积特征，定义在不同方向所得的视觉特征为 $\boldsymbol{v}_t=(\boldsymbol{v}_{t,1},\boldsymbol{v}_{t,2},\cdots,\boldsymbol{v}_{t,K})$，$K$ 为导航方向的最大角度。同时语言信息通过词嵌入工具 GloVe 进行语言编码，把每一个单词表达成一个由实数组成的向量，L 个单词经过映射后定义为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_L)$，这些向量捕捉到了单词之间的语义特征。在这些基础特征和历史语境的制约下，LSTM 编码生成当前状态的隐含语义信息 $\boldsymbol{h}_t$，具体表达式为

$$(\boldsymbol{h}_t,\boldsymbol{c}_t)=\mathrm{LSTM}([\hat{\boldsymbol{x}}_t,\hat{\boldsymbol{v}}_t,a_{t-1}],\boldsymbol{h}_{t-1},\boldsymbol{c}_{t-1}),\tag{1}$$

其中，$\boldsymbol{c}_t$ 为 LSTM 的单元格状态，$\hat{\boldsymbol{x}}_t$ 与 $\hat{\boldsymbol{v}}_t$ 分别为协同文本和图像特征，a_{t-1} 表示上一状态的导航动作。然后将卷积特征和语义特征进行共注意力机制编码，获得特征的权重概率分布。

导航决策模块，主要是根据视觉语言匹配模块处理之后的多模态信息，来判断当前的导航进度与下一步的动作选择。具体来讲，由于注意力机制不会保留序列位置信息，因此结合上一模块输出，首先进行位置编码[17]获取位置信息，然后进行进度监控，估计当前导航的进度。另一方面，多模态信息通过 LSTM 网络[18]进行解码，得到动作序列，从而决定下一步的方向，最后输出导航轨迹。动作选择公式为

$$o_{t,k}=(\boldsymbol{W}_{\mathrm{a}}[\boldsymbol{h}_t,\hat{\boldsymbol{x}}_t])^{\mathrm{T}}g(\boldsymbol{v}_{t,k}),\tag{2}$$

$$\boldsymbol{p}_t=\mathrm{Softmax}(\boldsymbol{o}_t),\tag{3}$$

其中，$\boldsymbol{W}_{\mathrm{a}}$ 为网络学习参数，$g(\cdot)$ 表示多层感知器（MLP）。

进度监控是通过执行匹配模块的注意力权重分布，来估计代理完成指令的进度，预测与导航终点的距离，同时进一步加强视觉图像与语言指令之间的对齐与匹配。进度监控的输出表示为 $p_{\mathrm{pm},t}$，具体公式为

$$\boldsymbol{h}_{\mathrm{pm},t}=\sigma(\boldsymbol{W}_{\mathrm{h}}[\boldsymbol{h}_{t-1},\hat{\boldsymbol{v}}_t]\otimes\tanh\boldsymbol{c}_t),\tag{4}$$

$$\boldsymbol{p}_{\mathrm{pm},t}=\tanh(\boldsymbol{W}_{\mathrm{pm}}[\boldsymbol{\alpha}_t,\boldsymbol{h}_{\mathrm{pm},t}]),\tag{5}$$

其中，$\boldsymbol{W}_{\mathrm{h}}$ 和 $\boldsymbol{W}_{\mathrm{pm}}$ 表示网络学习参数，$\boldsymbol{c}_t$ 为 LSTM 的单元格状态，⊗表示点积，σ 是 Sigmoid 函数，$\boldsymbol{\alpha}_t$ 为文本特征的注意力权重。

2.2　基于余弦相似度的损失函数

VLN 任务是指智能体遵循自然语言指令，在 Matterport3D 模拟器[19]中从起始姿势导航到目标位置。在离散空间环境中确定一张图，其中每个节点都是智能体可以到达的位置，两个节点之间的每个边表示它们之间存在的导航路线。形式上，智能体在每一轮的开始，输入一个自然语言指令 $\bar{\boldsymbol{x}}=\langle\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_L\rangle$，其中 L 是指令的长度，$\boldsymbol{x}_i$ 是单个单词标记。智能体能够在每个节点处获取周围全景图像信息 $\boldsymbol{o}_t$，其观察的初始 RGB 图像 $\boldsymbol{o}_0$ 由初始姿势确定，包括 3D 位置 $v\in V$、方向 $\psi\in[0,2\pi)$ 和摄像机仰角 $\theta\in\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ 的元组 $s_0=\langle v_0,\psi_0,\theta_0\rangle$，其中 V 是场景中与全景视点相关联的 3D 点集。智能体通过执行一系列操作 $\langle s_0,a_0,s_1,a_1,\cdots,s_T,a_T\rangle$，每一个动作 a_t 对应新的姿态 $s_{t+1}=\langle v_{t+1},\psi_{t+1},\theta_{t+1}\rangle$，生成新的图像观测 $\boldsymbol{o}_{t+1}$，其中 T 为导航到终点的时刻。当智能体选择停止动作时，事件结束。如果动作序列将智能体交付到接近预设的目标位置 v^*，则任务成功完成。

模拟器动作空间是状态相关的，允许智能体在不同的前向轨迹之间根据细粒度做出选择。为了简化任务，基线模型将动作空间简化为 6 个动作，分别对应于左、右、上、下、前进和停止。前向动作被定义为始终移动到距离智能体视野中心最近的可到达视点。左、右、上和下动作被定义为将相机移动 30°。要使智能体成功地导航到目标位置，需要其作出正确的动作预测，因此本文增加余弦相似损失函数来指导深度神经网络的学习，减小特征空间中类内特征的差异，增大类间特征的分布范围，从而提升动作预测效果。

Regretful 模型损失函数由三部分组成，训练交叉熵损失代理进行行动选择，训练均方误差损失进行进度监控，训练熵损失以鼓励代理探索其他行动。在此基础上，为了使实际动作预测方向与参考方向向量更加相似，增加了余弦相似损失函数 L_{cosine} 进行约束学习，以提高模型导航方向预测的准确率。Regretful 模型损失函数表达式为

$$L_{\mathrm{loss}}=\underbrace{\lambda\sum_{t=1}^{T}\boldsymbol{y}_{\mathrm{nv},t}\log(\boldsymbol{p}_{t,k})}_{\text{action selction}}+\underbrace{(1-\lambda)\sum_{t=1}^{T}(\boldsymbol{y}_{\mathrm{pm},t}-\boldsymbol{p}_{\mathrm{pm},t})^2}_{\text{progress monitor}}-\underbrace{\beta\sum_{t=1}^{T}\sum_{k=1}^{K}-\boldsymbol{p}_{t,k}\log(\boldsymbol{p}_{t,k})}_{\text{entropy loss}}+L_{\mathrm{cosine}},\tag{6}$$

式中：$\boldsymbol{p}_{t,k}$ 为时间为 t 时每个导航方向动作概率组成的向量；$\boldsymbol{y}_{\mathrm{nv},t}$ 为时间为 t 时参考导航方向组成的向量；$\lambda=0.5$ 为平衡交叉熵损失的权值；$\beta=0.01$ 为熵损失的权值；$\boldsymbol{y}_{\mathrm{pm},t}$ 为从当前视点到目标的长度单位的归一化距离组成的向量；$\boldsymbol{p}_{\mathrm{pm},t}$ 为进度监控实际输出，表示文本指令的完成度。

余弦相似度通过计算向量空间中两个向量的夹角余弦值来评估它们的相似度[20]。余弦值越接近 1，就表明夹角越接近 0°，也就是两个向量越相似。模型中网络输出预测的是下一步动作方向的概率，因而可以增加余弦相似度来描述方向的误差，作为损失函数的约束条件。

对于 L_{cosine} 的设计，本研究将预测轨迹方向 $\boldsymbol{x}$ 和真实轨迹方向 $\boldsymbol{y}$ 当作两组序列向量，计算两组序列向量的余弦相似度。其中：

$$\boldsymbol{x}_t=\mathrm{Softmax}(\boldsymbol{p}_{t,k}),\tag{7}$$

$$\boldsymbol{y}_t=\boldsymbol{y}_{\mathrm{nv},t}。\tag{8}$$

在每个轨迹点，设置一组参数，用这些参数记录真实的轨迹点方向和预测的轨迹点方向，在每次预测一个轨迹方向时，计算一次损失函数值，并累计前面轨迹点的损失函数值，即

$$\cos\theta=\frac{\sum_{t=1}^{T}(\boldsymbol{x}_t*\boldsymbol{y}_t)}{\sqrt{\sum_{t=1}^{T}\boldsymbol{x}_t^2}*\sqrt{\sum_{t=1}^{T}\boldsymbol{y}_t^2}},\tag{9}$$

$$L_{\mathrm{cosine}}=1-\cos\theta,\tag{10}$$

式中“$*$”表示卷积。

3　实验与结果分析

3.1　实验设置

3.1.1　模拟器与数据集

实验选取国际公开数据集 R2R（Room-to-room）[8]进行训练检测。整个数据集包含 90 个不同的房间环境，分为训练集、已知验证集、未知验证集和测试集。其中训练集包括 61 个环境和 14025 条指令；已知验证集，是指在训练集中出现过的导航环境图片，包括 1020 条指令；未知验证集则是指未在训练集出现过的导航环境图片，由其他 11 个环境中的 2349 条指令组成；测试集由未知的 18 个环境中的 4173 条指令组成。

R2R 数据集建立在 Matterport3D 数据集[19]上，包含了 90 栋建筑 194400 张 RGB 图像中的 10800 张全景图，并从导航图中采样了 7189 条路径。每条路径都有三个由人类编写的地面真实导航指令，平均长度为 29 个单词。

3.1.2　数据增强

在公开数据集 R2R 中有 61 个场景，打印其中的两个场景的节点分布，如图 2 所示，可以看到图中大部分场景中相连节点的距离都小于 3 m。考虑到类似于图像均值滤波的方式，对于某一个节点的全景图像特征，在小于 3 m 的范围内应该是相似的，因此，本文随机搜索当前轨迹视点小于 3 m 的相邻节点，使用它们的平均值替代原来的特征，进行数据扰动，起到了数据增强的作用。

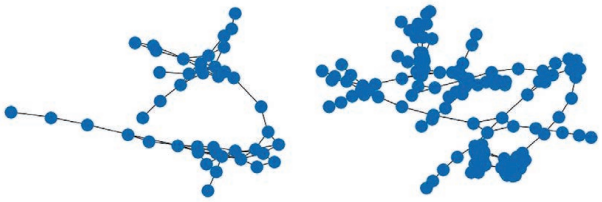

图 2　数据集中场景节点分布图

Fig. 2　Distribution of scene nodes in dataset

3.1.3　网络框架及实验设置

本文在 PyTorch 深度学习框架上实现基于余弦相似度损失函数的视觉语言导航算法，实验参数设置主要参考 Regretful 模型的训练设置。模型使用 ImageNet 上预先训练好的 ResNet-152 来提取 2048 维图像特征。用于映射原始图像特征的 MLP 由 BN→FC→BN→Dropout→ReLU 组成。BN 指 Batch Normalization 层，FC（Fully Connected）层将 2176 维输入向量映射到 1024 维向量，并将 dropout 设置为 0.5。设置语言指令编码器的嵌入维度为 256，之后加一个比率为 0.5 的 Dropout 层。LSTM 隐藏状态设为 512 维。本实验使用 1 块 GPU（显卡型号为 NVIDIA GeForce GTX 1050 Ti）进行训练，迭代 1 万次。采用自适应矩估计（Adam）的训练策略，其中，批处理大小为 64，初始学习率设为 10^{-4}。在动作选择的训练中执行分类抽样。

3.2　实验结果

VLN 挑战使用 4 个评估指标：导航错误（NE）、成功率（SR）、Oracle 成功率（OSR）和（标准化逆向）路径长度加权成功率（SPL）[21]。导航错误（NE）为测量预测路径中的最后一个节点与参考路径最后一个节点之间的距离。成功率（SR）为测量预测路径中的最后一个节点在参考路径最后一个节点的阈值距离 d_{th}（本文中阈值距离为 3 m）内的频率。Oracle 成功率（OSR）表示 agent 成功停止在最接近目标点的比率。路径长度加权成功率

(SPL)为综合考虑成功率和标准化路径长度的指标。

将本文方法与近几年视觉语言导航算法的模型进行比较,对比结果如表 1 所示。比较的模型主要包括:1)随机方法:agent 在环境中随机导航;2)Seq-to-Seq[8]:序列到序列模型;3)RPA[9]:一种结合基于模型和无模型强化学习的导航方法;4)Speaker-Follower[10]:一种引入了数据增强和全景动作空间的导航方法;5)RCM[16]:一种具有跨模态匹配,并且将模仿学习与强化学习相结合的方法;6)Self-Monitoring[11]:具有自我监控导航进度的方法;7)Regretful[13]:在模型(6)式的基础上增加回退策略减小导航累积误差的方法。Regretful * 指的是带有数据增强的模型。本文评估了当前无搜索策略的方法,发现这些算法的 SPL 指标都很高,由此说明本文方法具有实际的应用意义。

表 1　不同模型对比结果

Table 1　Comparison of different models

Model	Validation-Seen				Validation-Unseen				Test (unseen)			
	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑
Random	9.45	0.16	0.21	–	9.23	0.16	0.22	–	9.77	0.13	0.18	0.12
Seq-to-Seq	6.01	0.39	0.53	–	7.81	0.22	0.28	–	7.85	0.20	0.27	0.18
RPA	5.56	0.43	0.53	–	7.65	0.25	0.32	–	7.53	0.25	0.33	0.23
Speaker-Follower	3.36	0.66	0.74	–	6.62	0.36	0.45	–	6.62	0.35	0.44	0.28
RCM	3.37	0.67	0.77	–	5.88	0.43	0.52	–	6.01	0.43	0.51	0.35
Self-Monitoring	3.22	0.67	0.78	0.58	5.52	0.45	0.56	0.32	5.99	0.43	0.55	0.32
Regretful	3.69	0.65	0.72	0.59	5.36	0.48	0.61	0.37	–	–	–	–
Regretful *	3.23	0.69	0.77	0.63	5.32	0.50	0.59	0.41	5.69	0.48	0.56	0.40
Ours	3.23	**0.70**	0.76	**0.65**	5.33	**0.52**	**0.61**	**0.42**	5.69	**0.50**	**0.57**	**0.41**

如表 1 所示,本文所提出的方法在 3 个验证数据集上相比以前的模型,指标都有所改善。改进模型在不可见测试环境下 SR 达到了 0.50,比之前的 Regretful * 模型高了 0.02。同时,改进模型也极大地提高了在已知和未知验证集中的 SR 和 SPL。可以看出,本文所提出的改进模型提高了导航成功率,提升了模型的泛化性能,并且验证了余弦相似损失函数和全景视图特征平滑数据增强策略的有效性。

为了更直观地展示本文算法的特点,本文选取验证集中不可见环境的复杂样例进行展示,如图 3 所示。从图中可以看出,语言指令要求智能体对每一步的实际导航方向作出正确的预测,前期的方向判断错误会导致误差积累,从而影响最终的导航正确率。本文方法能够精确地对实际导航方向进行预

图 3　复杂导航场景示例

Fig. 3　Examples of complex navigation scenarios

测，与其他模型相比，本文模型能够使智能体更加准确地判断不同方向之间的细微差别，并选择正确的路径。由此可知，本文模型学习了一种更可靠的策略来关注智能体在自主导航过程中的方向信息和跨模态匹配信息，能够提升模型的导航准确率与泛化性能。

4 结　论

提出一种基于余弦相似的视觉语言导航算法，该算法采用无搜索策略的 Regretful 模型进行训练，通过引入余弦相似损失函数对网络模型进行优化训练，学习了一种更可靠的策略来关注智能体在自主导航过程中的方向信息和跨模态匹配信息。同时，采取全景视图特征平滑的方法进行数据增强，提高了模型的导航准确率与鲁棒性。对比实验结果证明了本文算法的有效性与泛化性。后续研究将主要致力于获得更为明确的跨模态匹配信息，使得视觉语言导航算法性能得到进一步提升。

参 考 文 献

[1] Brahmbhatt S, Hays J. DeepNav: learning to navigate large cities[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE Press, 2017: 3087-3096.

[2] Gupta S, Davidson J, Levine S, et al. Cognitive mapping and planning for visual navigation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7272-7281.

[3] Chen J H, Jiang H H. Multi-scale segmentation for ridge row in vision navigation[J]. Laser & Optoelectronics Progress, 2020, 57(8): 081017.
陈颉颢, 蒋红海. 视觉导航中垄行多尺度分割算法[J]. 激光与光电子学进展, 2020, 57(8): 081017.

[4] Parisotto E, Salakhutdinov R. Neural map: structured memory for deep reinforcement learning[C]//Proceedings of 2018 International Conference on Learning Representations (ICLR), April 30-May 3, 2018, Vancouver, BC, Canada. La Jolla: ICLR, 2018.

[5] Wu B, Wang X R. Inertial navigation aided image feature matching method[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101509.
吴斌, 王旭日. 惯性导航辅助图像特征匹配方法研究[J]. 激光与光电子学进展, 2020, 57(10): 101509.

[6] Savinov N, Dosovitskiy A, Koltun V. Semi-parametric topological memory for navigation[C]//Proceedings of 2018 International Conference on Learning Representations (ICLR), April 30-May 3, 2018, Vancouver, BC, Canada. La Jolla: ICLR, 2018.

[7] Zhu Y K, 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), May 29-June 3, 2017, Singapore. New York: IEEE Press, 2017: 3357-3364.

[8] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3674-3683.

[9] Wang X, Xiong W H, Wang H M, et al. Look before you leap: bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11220: 38-55.

[10] Fried D, Hu R, Cirik V, et al. Speaker-follower models for vision-and-language navigation[C]//Proceedings of the 32th Conference on Neural Information Processing Systems (NeurIPS), December 3-8, 2018, Montreal, Canada. New York: ACM, 2018: 3314-3325.

[11] Ma C Y, Lu J, Wu Z, et al. Self-monitoring navigation agent via auxiliary progress estimation[C]//Proceedings of 2019 International Conference on Learning Representations (ICLR), May 6-9, 2019, New Orleans, Louisiana, United States. La Jolla: ICLR, 2019.

[12] Xu Y, Fern A, Yoon S. Discriminative learning of beam-search heuristics for planning[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), January 6-12, 2007, Hyderabad, India. New York: ACM, 2007: 2041-2046.

[13] Ma C Y, Wu Z X, AlRegib G, et al. The regretful agent: heuristic-aided navigation through progress estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 6725-6733.

[14] Zhu F D, Zhu Y, Chang X J, et al. Vision-language navigation with self-supervised auxiliary reasoning tasks[C]//2020 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10009-10019.

[15] Majumdar A, Shrivastava A, Lee S, et al. Improving vision-and-language navigation with image-text pairs from the web[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12351: 259-274.

[16] Wang X, Huang Q Y, Celikyilmaz A, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 6622-6631.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proceedings of the 31th Conference on Neural Information Processing Systems (NIPS 2017), December 4-9, 2017, Long Beach, Canada. New York: ACM, 2017: 5998-6008.

[18] Luong T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, September 17-21, 2015, Lisbon, Portugal. Stroudsburg: Association for Computational Linguistics, 2015: 1412-1421.

[19] Chang A, Dai A, Funkhouser T, et al. Matterport 3D: learning from RGB-D data in indoor environments[C]//2017 International Conference on 3D Vision (3DV), October 10-12, 2017, Qingdao, China. New York: IEEE Press, 2017: 667-676.

[20] Wu H H, Su H S, Liu G H, et al. Facial expression recognition algorithm based on cosine distance loss function [J]. Laser & Optoelectronics Progress, 2019, 56(24): 241502.
吴慧华, 苏寒松, 刘高华, 等. 基于余弦距离损失函数的人脸表情识别算法[J]. 激光与光电子学进展, 2019, 56(24): 241502.

[21] Anderson P, Chang A, Chaplot D S, et al. On evaluation of embodied navigation agents[EB/OL]. (2018-07-18)[2020-10-19]. https://arxiv.org/abs/1807.06757.