计算机科学
COMPUTER SCIENCE

基于BEV感知的视觉平面图定位

陈集伟, 陈泽彬, 谭光

引用本文

陈集伟, 陈泽彬, 谭光. 基于BEV感知的视觉平面图定位[J]. 计算机科学, 2026, 53(1): 216-223.

CHEN Jiwei, CHEN Zebin, TAN Guang. Visual Floorplan Localization Based on BEV Perception[J]. Computer Science, 2026, 53(1): 216-223.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250300045

基于 BEV 感知的视觉平面图定位

陈集伟　陈泽彬　谭　光
中山大学智能工程学院　广东 深圳 510275
(chenjw269@mail2.sysu.edu.cn)

摘　要　视觉平面图定位任务通过视觉观测数据与场景平面图表示的匹配实现位姿估计。实际应用中,有效融合视觉观测与平面图之间的几何和语义关联对提升定位精度至关重要。然而,现有方法存在两个主要局限:一是未能充分挖掘相机视野内的语义信息;二是缺乏几何与语义线索的联合匹配机制。针对上述问题,提出基于鸟瞰图(Bird Eye View,BEV)感知的视觉平面图定位框架,其包含 3 个核心模块:首先,BEV 语义建图模块通过多模态图像投影变换构建局部场景的 BEV 语义表征,实现观测数据的结构化表示;其次,预期观测生成模块在平面图空间内生成预期观测数据库,通过可微分渲染方法实现观测数据的快速生成;最后,多层次匹配定位模块提出几何-语义联合匹配机制,通过层次化匹配策略融合 BEV 观测中的几何布局和语义类别信息,实现与平面图的精确匹配。实验结果表明,该框架在公开数据集 Structured3D 和自建仿真环境数据集 IndoorEnv 上的定位召回率分别从 0.32%和 4.82%提升到了 3.12%和 58.77%,显著优于现有基线方法 Laser 和 F3Loc,从而验证了所提方法在复杂室内场景中的有效性和鲁棒性。

关键词:BEV 感知;平面图定位;视觉定位;几何-语义联合匹配
中图分类号　TP391.4;U495

Visual Floorplan Localization Based on BEV Perception

CHEN Jiwei,CHEN Zebin and TAN Guang
School of Intelligent Systems Engineering,Sun Yat-sen University,Shenzhen,Guangdong 510275,China

Abstract　Visual floorplan localization task achieves pose estimation by matching visual observation with scene floorplan representation. In practical applications, how to effectively integrate geometric and semantic correlations between observation and floorplan in matching is particularly important for improving localization accuracy. However, existing methods have two main limitations. Firstly, they fail to fully utilize the semantic information within the camera's field of view. Secondly, they lack a joint matching mechanism for geometric and semantic clues. To address these issues, this study proposes a visual floorplan localization framework based on BEV perception, which includes three core components. Firstly, the BEV semantic mapping module constructs the BEV semantic representation of local scenes through multimodal image projection transformation, achieving structured representation of observation data. Secondly, the expected observation generation module generates an expected observation database within the floorplan space, and achieves rapid generation of observation data through differentiable rendering method. Finally, the multi-level matching and localizing module proposes a geometric-semantic joint matching mechanism, which integrates the geometric layout and semantic category information from BEV observations through a hierarchical matching strategy to achieve accurate matching with the floorplan. The experimental results show that the framework achieves an improvement in localization recall from 0.32% and 4.82% to 3.12% and 58.77% on the publicly available dataset Structured3D and the self built simulation environment dataset IndoorEnv, respectively, which is significantly better than the existing baseline methods Laser and F3Loc. This validates the effectiveness and robustness of the proposed method in complex indoor scenes.

Keywords　BEV perception, Floorplan localization, Visual localization, Geometric-semantic joint matching

1　引言

定位算法在现代智能系统中至关重要,被广泛应用于自动驾驶汽车、虚拟现实(Virtual Reality,VR)和自主机器人等领域。在定位算法中,视觉平面图定位因为能够利用轻量级平面图估计查询图像的相机位姿而备受关注。传统视觉定位方法[1-7]从图像离线构建场景的 3D 点云模型或其他表示,然后在观测图像与场景模型之间进行匹配定位;或者在场景中

到稿日期:2025-03-11　返修日期:2025-05-09
基金项目:国家自然科学基金(62372488)
This work was supported by the National Natural Science Foundation of China(62372488).
通信作者:谭光(tanguang@mail.sysu.edu.cn)

预先采集标注精确位姿[8-10]的图像数据库，通过匹配与观测图像最相似的数据库图像来预测当前位姿。这些方法在计算和存储方面成本较高，因此在无法提前采集数据的陌生场景中，其应用受到限制。

为了在陌生场景定位，学者提出了平面图定位方法[11-18]，其有两点优势：1）大多场所有完整的建筑平面图，获取方便；2）人类可以在陌生场景中使用平面图定位，神经网络同样能够学习这种能力。目前的平面图定位方法中，基于布局预测的方法[11,19]从图像预测房间布局并与平面图进行匹配定位，匹配的依据是不同位置的布局结构点云或者点云在像素平面的投影的相似程度。这类方法依赖于已知[20]或估计的[21]天花板高度，难以处理未知高度的场景；同时，这类方法不使用平面图中的语义信息作为定位参考，缺乏对平面图语义信息的挖掘。

基于跨域匹配的方法[12,18,20,22]将平面图和查询图像映射到特征空间进行匹配。Laser[12]将 RGB 图像分块进行编码，使用平面图中嵌入布局结构的门窗语义信息作为布局结构特征向量的补充成分，训练模型学习图像像素与平面图 2D 布局的关联；F3Loc[18]借鉴布局预测方法，从 RGB 图像预测 2D 平面射线深度表示，并与平面图在 2D 射线深度特征空间进行匹配，未使用平面图中的语义信息。目前的平面图定位方法要么不考虑平面图中的语义信息[18,20,22]，仅比较视觉观测和平面图几何结构的相似程度；要么仅考虑和布局结构嵌套的门窗信息[12]来作为描述不同位置布局特征的补充信息，并未独立设计几何语义联合匹配机制。这些缺点导致两方面不足：1）平面图包含重复布局结构，如角落和平滑墙段，定位中会因相似布局而出现歧义；2）场景中家具比较稠密而遮挡布局结构时，会错误匹配布局特征。融合多帧估计[23-24]虽然可以减少单帧误差，但需要运动轨迹来收敛且需要高效数据吞吐，因此提高单帧估计精度仍然重要。针对这些问题，本文扩展已有的平面图定位方法，结合语义几何信息实现定位。

本文提出了一种基于 BEV 感知的平面图定位方法，主要贡献包括 3 个部分。1）该方法将语义 BEV 网格作为视觉观测和平面图匹配定位的中间模态，相应设计了一种渲染平面图不同位姿预期观测的方法，对观测数据和平面图中的布局语义信息进行建模；2）设计了自适应平衡不同类别视觉线索重要性的匹配模型，通过多尺度度量真实观测与预期观测的相似度，增强平面图标注存在错误时的稳定性；3）该方法在公开数据集和仿真环境数据集上的定位召回率超过了已有方法，并且针对布局结构存在歧义的定位结果明显优于已有方法。

2 相关工作

2.1 传统视觉定位方法

视觉定位任务的目标是关联视觉观测和场景表示，估计相机位姿。传统方法通过多视图几何构建 3D 模型[25]，匹配视觉观测和场景模型进行定位。这类基于 3D 场景模型的方法[1-2,5-7,26]，首先使用三角测量（Structure from Motion，SfM）[27]或神经辐射场[28-29]等方法从不同视角重建 3D 场景结构，之后建立查询图像和场景模型之间的 2D-3D 关联，最后使用 ICP[30]或者 PNP 方法计算相机位姿。这类方法构建场景模型、匹配查询图像和存储稠密点云地图需要消耗大量计算和存储资源，无法应用于不允许提前探索的场景以及计算资源受限的情况。

2.2 基于深度学习的视觉定位

随着深度学习的发展，出现了多种数据驱动的视觉定位方法，包括基于回归的方法和基于检索的方法。其中，基于回归的方法主要分为场景坐标回归和绝对位姿回归。场景坐标回归方法[4,29,31-37]分为两个步骤：首先，预测查询图像和场景模型之间的 2D-3D 关联；其次，使用 PNP 方法从 2D-3D 关联计算相机位姿。绝对位姿回归方法[38-42]通过网络参数学习隐式的场景表示，网络直接预测输入的查询图像的位姿。这类基于回归的方法要么需要重建场景三维模型，要么网络模型只编码训练集中的场景信息，在新场景中需要重新采集数据进行重建或者重新训练，仍有较大的局限性。

基于检索的方法[8-10,43,46]预先在场景中采集图像数据库并标注位姿，之后使用模型度量查询图像和数据库图像之间的相似度，以相似度最高的数据库图像预测查询图像的位姿。虽然模型匹配图像的能力可以泛化到新场景，但部署到新场景仍需重新采集图像数据库，其应用受到较大限制。

2.3 平面图视觉定位

以上方法或者依赖于场景模型，或者需要提前采集图像，不能处理无法提前采集数据的陌生场景。为解决该问题，学者提出了使用平面图定位的方法。平面图获取和处理的成本较低，因此该方法是在陌生场景中进行定位的可行方法。机器人领域已有使用雷达[13]以及其他低成本传感器，如深度相机重建点云[47]感知布局进行平面图定位的方法。基于布局预测和匹配的方法[11]从 RGB 图像预测房间布局，与平面图中的布局结构进行匹配定位。该类方法需要假设固定的天花板高度，并且难以处理相似布局结构造成定位歧义的情况。

除了显式匹配布局的方法之外，基于跨模态匹配的方法[12,20,22,24]通过度量学习[48]创建一个特征隐空间，将不同模态的查询图像和平面图映射到同一特征空间进行匹配。已有基于布局几何结构的方法[11]和大多跨模态匹配[18,20,22]的方法仅使用了平面图中的几何信息，少部分参考语义信息的方法[12]局限于与布局紧密结合的门窗，仍未能充分利用平面图中的家具物体等语义信息。室内平面图通常包含丰富的语义物体，如桌椅和橱柜，一方面可以补充被遮挡的布局结构，另一方面可以减小重复布局的歧义，提高准确率。

3 基于 BEV 感知的多层次定位框架

3.1 任务定义

视觉平面图定位的目标是，使用平面图作为参考信息，通过视觉观测预测自身在场景中的位置。由于平面图缩放了空间距离，并且缺失物体的详细信息，本文认为平面图符合该定义：平面图是真实世界经过缩放的场景表示，物体形状没有过多细节，只能反映区域方位和类别。

图 1 为基于 BEV 感知的平面图定位流程图。BEV 语义建图模块输入观测数据，输出语义 BEV 网格作为周围环境的表示。预期观测生成模块构建了相机在平面图不同位姿的预

期观测数据库。多层次匹配定位模块匹配真实观测构建的语义 BEV 网格和预期观测数据库，来预测当前位姿。

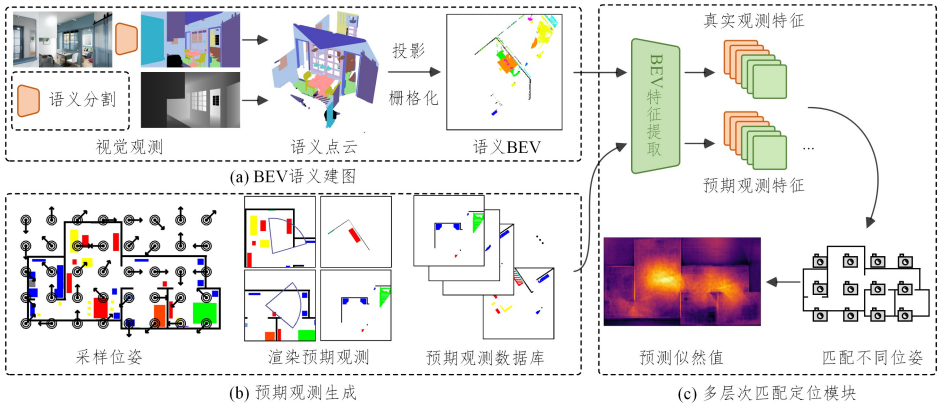

图 1 基于 BEV 感知的平面图多层次定位框架

Fig. 1 BEV perception based floorplan localization framework

3.2 BEV 局部地图构建

与 SLAM 方法从第一人称视角的 RGB 观测输入提取特征点，并计算对应三维空间点云不同，本文设计的 BEV 语义建图模块将视觉观测投影到俯视视角来构建 2D 占用网格。t 时刻从相机观测获取的 BEV 占用网格 m_t 是一个形状为 $M\times M$ 的矩阵，占用网格每个单元对应真实空间 25 cm^2 (5 cm×5 cm)的面积，网格中单元的取值{0,…,C}分别表示不同语义类别，每个单元表示对应的空间区域是空白/占用及占用的语义类别。BEV 网格中的信息反映三维空间相机视野范围内的内容，相机自身的位置对应 BEV 底部中间坐标为 $(M/2,0)$的网格单元。

为了从第一人称相机观测构建带有语义信息的 BEV 网格，如图 1(a)所示，使用深度图计算点云。深度图中每个像素对应空间点云的计算式如式(1)所示。

$$x=\frac{(u-c_x)d_{uv}}{f_x},y=\frac{(v-c_y)d_{uv}}{f_y},z=d_{uv} \tag{1}$$

在 RGB 图像的语义分割结果中，将每个像素与对应位置的点云进行关联，得到点云的语义类别。接着，语义点云从相机视角投影到俯视视角的 2D 平面，并栅格化，得到语义 BEV 网格，实现对视觉观测的几何和语义建模。

3.3 预期观测生成

从相机观测构建语义 BEV 网格后，需要匹配平面图预测相机位姿。匹配过程中，需要一种反映平面图上不同位姿观测数据的中间表示。当已知相机位姿和平面图数据时，相机观测范围可以用位姿前方的视锥表示，投影到 2D 平面为位姿前方的扇形区域。本文定义平面图上的预期观测作为定位的中间表示：平面图上相机位姿前方扇形范围的平面图信息为相机在该位姿下的预期观测，其含义是相机处于对应位姿时的观测信息。考虑到实际场景存在墙体遮挡，以位姿为起点，在水平视野范围内进行射线碰撞检测，排除被墙壁遮挡的内容。从平面图已知位姿渲染预期观测的算法如算法 1 所示。

算法 1 2D 平面图预期观测渲染算法

输入：平面图 $M_{L\times L}$，相机位姿 $P=\{x,y,\theta\}$

输入：预期观测 $m_{L'\times L'}$

1. 根据相机位姿对平面图 $M_{L\times L}$ 进行仿射变换，得到平面图 $M^*_{L\times L}$，使得相机位于坐标原点且朝向对齐图像 y 轴
2. 获取位姿前方扇形范围平面图 $m^*_{L\times L}$
3. 以原点为起点，在水平范围内发射 2D 射线，计算射线离散落点坐标$\{\mathbf{p}_i\}=\{[d_i,\theta_i]\mid\theta_i\in(-40°,40°),d_i\in(0,D]\}$
4. 考虑布局结构信息，对不同朝向射线上的落点进行离散积分，判断射线与布局结构的碰撞次数是否大于阈值，大于阈值的部分被认为不可见
5. 根据射线落点可见性构建可见性矩阵 $V^*[\mathbf{p}_i]$
6. 预期观测 $m_{L'\times L'}=V^*_{L'\times L'}\times m^*_{L'\times L'}$

由观测构建的语义 BEV 网格和通过相机位姿获取的预期观测都处于俯视视角，具有形式相近的几何语义信息，能够有效匹配平面图的不同位姿。

如图 1(b)所示，参考粒子滤波器的初始化过程，在平面图上按照一定距离和朝向间隔均匀地采样位姿，对不同的位姿渲染其预期观测构成数据库。对当前观测和不同位姿的预期观测进行匹配，选择相似度最高的对应位姿作为预测结果。

3.4 多层次匹配定位模块

多层次匹配定位模块的结构如图 2 所示。该模块的目标是将观测构建的语义 BEV 网格和平面图预期观测映射到同一特征空间进行匹配。由于现实环境在细节上往往和平面图不同，尤其是随着时间推移，物体位置可能与平面图不匹配，此时主要参考不易随时间变化的布局结构定位，因此需要动态调整布局结构和语义物体作为定位依据的参考比例。

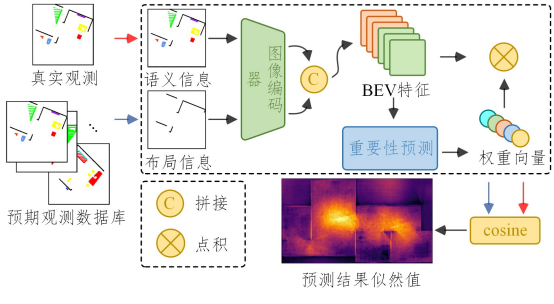

图 2 多层次匹配定位模块

Fig. 2 Multi-level matching and localizing module

为了建模语义 BEV 网格中不同信息对定位的不同参考作用，本文在使用基于 AlexNet[49] 结构的图像编码器提取 BEV 特征的基础上，设计了重要性预测模块来对不同类别的视觉特征进行参考比例的调整。该模块借鉴了 Squeeze-and-

Excitation(SE) 网络[50]的通道注意力思想，具体做法如图 2 所示。将语义 BEV 网格和平面图预期观测按照不同语义层次(布局结构和语义物体)划分不同通道，由图像编码器进行初步特征提取，得到 BEV 特征 $F\in\mathbb{R}^{(C,H,W)}$，其中不同通道包含特定的语义和布局特征而具有不同的重要性。

首先进行压缩操作，将输入特征图按照式(2)压缩为通道描述符 $z\in\mathbb{R}^C$ 输入重要性预测模块。

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F_c(i,j) \tag{2}$$

然后进行激活操作，通过多层感知机对描述符进行压缩再解压得到重要性权重 $W\in\mathbb{R}^C$，并将其作为输出。最后，使用该重要性权重，按照式(3)对原始特征图进行加权，得到最终的语义 BEV 网格特征，根据重要性重新调整权重后的特征用于度量真实观测和不同预期观测之间的相似度。

$$F'=W\odot F \tag{3}$$

重要性预测模块能够动态调整布局结构和语义物体在匹配定位过程的参考比例，实现对平面图标注和真实观测不符的特殊情况的稳定处理。

3.5 度量学习训练

本文使用度量学习[48]的方法训练模型学习特定的特征空间，在该空间进行观测数据与不同位姿预期观测的匹配。涉及的各个变量定义如下：相机的真实观测 m_i 经过模型编码后的特征为 F_i，平面图上任意位姿渲染得到的预期观测 m_j 经过模型特征编码器得到的特征为 F_j，这两组特征之间的相似度为 $S(F_i,F_j)$；定位时，匹配相机真实观测和平面图上不同位姿的预期观测，选择相似度最大的预期观测 $\hat{m}=\arg\max_{m_j}S(F_i,F_j)$ 对应的位姿作为位姿预测值；在具体代码实现中，使用 cosine 相似度作为特征相似度的度量方式。

本文使用对比损失作为损失函数监督特征空间的学习。度量学习中的三元组，即锚点、正样本和负样本，分别由真实观测构建的语义 BEV 网格、平面图上位姿真值处的预期观测和随机位姿处的预期观测构成。训练目标是：BEV 网格编码后的特征空间中，正样本和锚点的特征具有较高相似度，而负样本和锚点的特征之间具有较低相似度。基于该训练目标，设计如式(4)所示的对比损失。

$$\begin{aligned}Loss=&\min(\alpha-S(F_I,F^+),0.0)+\\&\min(S(F_I,F^-)-\beta,0.0)\end{aligned} \tag{4}$$

由于平面图信息存在一定丢失，因此锚点和正样本相似但不完全相同，相似度接近于 1，式(4)中使用阈值 α 来约束这一范围；同理，锚点和负样本特征的相似度接近于 0，使用一个接近 0 的阈值 β 来约束。

4 实验及分析

4.1 实验设置

4.1.1 数据集

本文实验中使用了公开数据集 Structured3D(S3D)[51]和在搭建的 Unity 仿真环境中采集的数据集 IndoorEnv。S3D 数据集是一个包含了 3296 个场景的合成数据集，每个场景有不同的布局结构，并采集若干视角的相机观测，共有 78453 组水平方向视野为 80°的视觉观测。该数据集提供了不同观测样本的 RGB 图像、深度图像和语义分割，以及布局结构的平面图和每组观测数据在平面图上的位姿。另外，由于已有的室内数据集要么不提供连续运动过程的观测序列，要么不提供平面图标注，因此本文用 Unity 搭建仿真环境，采集了一个水平方向视野为 80°的前向视角相机图像数据集 IndoorEnv，并标注了室内环境平面图。IndoorEnv 中的场景布局结构参考现实生活中真实住宅的布局户型图，每个场景都包括完整的客厅、走廊、餐厅、厨房、厕所、卧室等不同类型的室内场景功能区域，不同区域空间分布上的位置关系同样参考真实世界中的住宅室内环境。同时，每个区域都放置了 5 件以上与区域类型有一定相关性的语义类别的家具物体，例如客厅中的沙发、茶几和电视，餐厅的餐桌和餐椅，卧室的床和床头柜；所使用的数据采集设备为水平方向视野为 80°、垂直方向视野为 64°的 RGBD 摄像头；场景平面图的标注流程为获取布局结构和家具物体的 x 和 y 轴位置和尺寸，并将占用区域按照 0.05 m/pixel 的比例缩放到平面图上。数据集包含了 20 个场景，其中 16 个场景的数据用于训练，4 个场景的数据用于测试。相机在仿真环境中按照一定的轨迹向前移动，并采样前向视野的视觉观测。IndoorEnv 数据集的部分场景和视觉观测数据如图 3 所示。

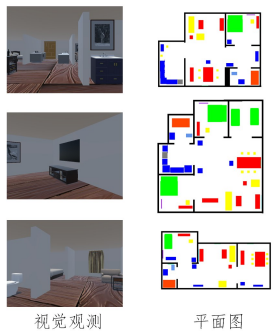

图 3 IndoorEnv 数据示例

Fig. 3 Example data in IndoorEnv dataset

4.1.2 基线方法

将本文方法与以下代表性方法进行比较。

1) Laser[12]：使用神经网络从 RGB 观测图像提取隐式特征，并从平面图布局采样离散 2D 点云构建密码本来渲染不同位姿的隐式特征。通过匹配由密码本渲染的不同位姿的隐式特征和相机观测数据的隐式特征，来预测相机位姿。

2) F3Loc[18]：使用 2D 平面上从相机发射的射线深度作为周围环境布局的表示，以平面图上不同位姿的射线深度作为不同位姿的特征。通过从 RGB 观测图像预测的机器人视线所处水平面的 2D 射线表示，匹配平面图上不同位姿的 2D 射线表示预测相机位姿。

为确保公平比较，实验在 IndoorEnv 数据集上，按 Laser 和 F3Loc 需要的平面图输入进行了标注，为 Laser 提供平面图布局边界的离散点云，为 F3Loc 提供平面图不同位姿的射线深度。

4.1.3 性能指标

使用召回率作为评估定位准确度的指标，其定义如式(5)所示。

$$Recall=\frac{Nums(Err<Threshold)}{Total} \tag{5}$$

根据定义，召回率为定位结果的误差在一定阈值范围内的比例。例如，1m 召回率指的是在所有定位结果中，定位误差小于 1m 的比例。

4.1.4 实现细节

本文方法使用 Python 3.7.11 和 PyTorch 1.10.0 在 CUDA 11.1 上实现，主要的实验平台是配备 NVIDIA Quadro RTX 8000 的服务器，操作系统为 CentOS 8.2。

4.2 整体精度对比

表 1 和表 2 比较了 2 个数据集上不同方法的各项召回率。如表 1 所列，本文方法取得了最佳性能。本文提出的定位方法在不使用平面图的语义标注的情况下，在 S3D 数据集上的召回率超过了两种对比方法；在 IndoorEnv 数据集上，不使用语义标注的情况下超过了对比方法，参考语义标注时进一步提高了准确度。当视觉观测存在歧义时，该方法可以利用语义线索解决歧义；当视野内出现复杂布局时，该方法又能够利用深度相机提供的三维空间距离信息，从而比仅使用 RGB 图像观测的这两种方法取得更好的性能。

表 1 S3D 数据集上定位召回率的对比

Table 1 Comparison of recall rate on S3D dataset

(%)

method	0.1m	0.5m	1m	1m30°
Laser	0.7	6.4	10.4	8.7
F3Loc	1.5	14.6	22.4	21.3
Ours w/ Semantic	**3.00**	**17.72**	**23.55**	**21.62**

表 2 IndoorEnv 数据集上定位召回率的对比

Table 2 Comparison of recall rate on IndoorEnv dataset

(%)

method	0.1m	0.5m	1m	1m30°
Laser	0.83	1.32	5.87	3.72
F3Loc	1.42	4.28	11.43	8.57
Ours w/o Semantic	**2.42**	**34.68**	**55.65**	**55.40**
Ours w/ Semantic	**5.65**	**43.55**	**67.74**	**67.34**

表 3 进一步统计了在定位误差小于 1m，即定位成功的样本上的位置误差和朝向误差的中位数，这两个指标可以说明系统在细微尺度定位上的准确性。本文方法在位置误差和朝向误差指标上都超过了对比方法，说明本文方法在细粒度精确定位方面能取得良好性能。

表 3 IndoorEnv 数据集上定位距离和朝向误差的对比

Table 3 Comparison of transition and rotation error on IndoorEnv dataset

Metric	Laser	F3Loc	Ours w/o Semantic	Ours w/ Semantic
Success rate @ 1m(%)	3.72	8.57	**55.40**	**67.34**
<1m med terr(cm)	72.4	46.11	**41.23**	**35.46**
<1m med rerr(deg)	9.99	7.50	**2.76**	**2.60**

4.3 样本定位结果分析

以 IndoorEnv 数据集的几组定位结果为例，图 4 给出了本文方法和 F3Loc 方法在单帧观测下的定位结果。本文方法有效使用视觉观测中的语义几何特征准确定位，其似然概率收敛到了正确的相机位姿；而 F3Loc 方法只使用简单的几何信息进行定位，其似然概率发散到平面图各处。

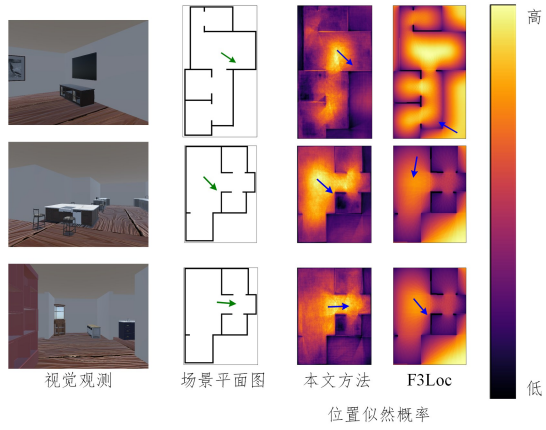

图 4 定位结果示例

Fig. 4 Example of localization result

图 5 给出了定位失败结果示例，其中绿色箭头为位姿真值，蓝色箭头为似然概率最大的 3 组位姿。在第一和第二组数据中，由于缺少语义物体参考，语义 BEV 网格只包含了几何结构信息，在平面图中不同位置的定位歧义导致了定位失败。在第三组结果中，视野内语义物体处于视野边缘，不能完整观察到充足的语义物体方位信息作为自身定位的相对位置参考。

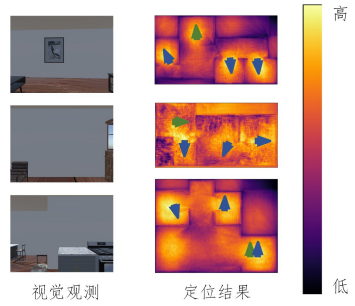

图 5 定位失败结果示例(电子版为彩图)

Fig. 5 Example of localization failure result

4.4 动态权重效果验证

为了验证不同级别线索定位作用动态权重模块的作用，设计了 5 组不同条件的实验：实验 A 去掉了平面图和真实观测 BEV 中的语义线索；实验 B 保留了真实观测中的语义线索，去掉了平面图上的语义线索，并且不使用重要性预测模块；实验 C 保留了真实观测中的语义线索，去掉了平面图上的语义线索，但使用重要性预测模块；实验 D 保留了真实观测中的布局和语义线索，去掉了平面图中的布局结构；实验 E 保留平面图和真实观测中的语义线索。表 4 列出了在 IndoorEnv 数据集上的消融实验结果。实验 A，D 和 E 的对照表明，当平面图上有完整的语义物体标注时，网络能够有效地结合布局和语义线索达到最高准确度，此时各项定位的召回率最高；同时，当布局结构和语义物体单独作为定位参考时，布局结构比家具物体更重要。实验 B 和 C 表明，当平面图语义物体标注缺失时，真实观测和平面图之间存在不匹配，重要性预测模块能够动态地调整 BEV 网格中几何信息和语义信息的参考比例，适当减轻语义物体作为定位线索的权重，避免

定位失败；在实验 C 中，由于使用了重要性预测模块，虽然语义物体缺失的平面图与包含语义物体的真实观测之间存在不匹配，但模型能够适当调整定位参考线索的比例，提高定位准确率。

表 4 动态权重召回率效果

Table 4 Effectiveness of adaptive weighting Recall

(%)

实验条件	0.1 m	0.5 m	1 m	1 m30°
实验 A	2.42	34.68	55.65	55.4
实验 B	0.81	5.65	18.55	18.55
实验 C	2.82	26.61	52.82	52.42
实验 D	2.02	30.65	46.37	46.37
实验 E	5.65	43.55	67.74	67.34

4.5 重要性预测模块的敏感性分析

重要性预测模块首先通过平均池化获取特征图的通道特征，然后对特征进行压缩和解压缩，以预测每个通道的重要性。压缩幅度的不同会导致结果的不同，如果压缩幅度过大，会丢失信息，不能很好地建模每个通道的语义重要性；如果压缩幅度过小，则模型计算量增加，且可能过拟合，降低精度。为了分析动态权重模块的敏感性，设置不同的特征压缩率，训练同样的轮数，并选择各自正负样本区分度最大的权重测试定位召回率。如图 6 所示，通过不同压缩率下的多组定位召回率可以看出，当压缩率设置为 8 时，能够取得最好的定位精度。

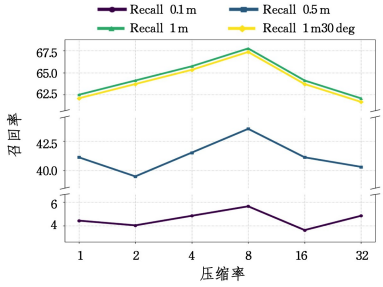

图 6 不同压缩率下的召回率

Fig. 6 Recalls under different reductions

结束语 本文研究了基于 BEV 感知的平面图定位任务，开发了一种从视觉观测构建 BEV 占用网格，匹配平面图实现定位的系统。该系统包含 BEV 语义建图、预期观测采样和渲染以及多层次匹配定位这 3 个关键组件，它们分别实现了从视觉观测构建平面图、渲染平面图上不同位姿的预期观测和匹配不同位姿进行定位。所提方法优于对比方法，在公开数据集和采集的仿真环境数据集上各项召回率分别取得了从 0.32%和 4.82%到 3.12%和 58.77%的提升。其核心优势是参考语义线索进行定位，消除几何结构的歧义性，提高准确率，并且针对平面图中的语义物体可能随着时间推移与真实环境不一致的情况，提出不同尺度定位线索的动态权重机制。

本文主要关注已有建筑平面图，假设语义物体标注在平面图上普遍存在的情况。后续工作计划扩展当前研究，通过设计更高效的多级位姿采样策略，提高定位系统的实时性；以及针对平面图上仅标注模糊区域的情况，通过大语言模型辅助关联区域类型与视野内的语义物体类别，提高定位的稳定性。

参 考 文 献

[1] SARLIN P E, CADENA C, SIEGWART R, et al. From coarse to fine: Robust hierarchical localization at large scale[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12716-12725.

[2] PANEK V, KUKELOVA Z, SATTLER T. Meshloc: Mesh-based visual localization[C]// European Conference on Compu-ter Vision. Cham: Springer, 2022: 589-609.

[3] ZHOU Q, AGOSTINHO S, OŠEP A, et al. Is geometry enough for matching in visual localization? [C]// European Conference on Computer Vision. Cham: Springer, 2022: 407-425.

[4] LIU J, NIE Q, LIU Y, et al. Nerf-loc: Visual localization with conditional neural radiance field[C]// 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 9385-9392.

[5] LIU L, LI H, DAI Y. Efficient global 2d-3d matching for camera localization in a large-scale 3d map[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2372-2381.

[6] SATTLER T, LEIBE B, KOBBELT L. Fast image-based localization using direct 2d-to-3d matching[C]// 2011 International Conference on Computer Vision. IEEE, 2011: 667-674.

[7] SATTLER T, LEIBE B, KOBBELT L. Efficient & effective prioritized matching for large-scale image-based localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(9): 1744-1756.

[8] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[9] BALNTAS V, LI S, PRISACARIU V. Relocnet: Continuous metric learning relocalisation using neural nets[C]// Procee-dings of the European Conference on Computer Vision (ECCV). 2018: 751-767.

[10] SCHINDLER G, BROWN M, SZELISKI R. City-scale location recognition[C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-7.

[11] BONIARDI F, VALADA A, MOHAN R, et al. Robot localization in floor plans using a room layout edge extraction network [C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 5291-5297.

[12] MIN Z, KHOSRAVAN N, BESSINGER Z, et al. Laser: Latent space rendering for 2d visual localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11122-11131.

[13] BONIARDI F, CASELITZ T, KÜMMERLE R, et al. Robust LiDAR-based localization in architectural floor plans[C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 3318-3324.

[14] BONIARDI F, CASELITZ T, KÜMMERLE R, et al. A pose graph-based localization system for long-term navigation in CAD floor plans[J]. Robotics and Autonomous Systems, 2019, 112: 84-97.

[15] LI Z, ANG M H, RUS D. Online localization with imprec ise floor space maps using stochastic gradient descent[C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2020: 8571-8578.

[16] MENDEZ O, HADFIELD S, PUGEAULT N, et al. Sedar: Reading floorplans like a human—using deep learning to enable human-inspired localisation[J]. International Journal of Computer Vision, 2020, 128(5): 1286-1310.

[17] WANG X, MARCOTTE R J, OLSON E. GLFP: Global localization from a floor plan[C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2019: 1627-1632.

[18] CHEN C, WANG R, VOGEL C, et al. F3Loc: fusion and filtering for floorplan localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 18029-18038.

[19] CHU H, KIM D K, CHEN T. You are here: Mimicking the human thinking process in reading floor-plans[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2210-2218.

[20] HOWARD-JENKINS H, RUIZ-SARMIENTO J R, PRISACARIU V A. Lalaloc: Latent layout localisation in dynamic, unvisited environments[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10107-10116.

[21] CRUZ S, HUTCHCROFT W, LI Y, et al. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2133-2143.

[22] HOWARD-JENKINS H, PRISACARIU V A. Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments[C]// European Conference on Computer Vision. Cham: Springer, 2022: 693-709.

[23] SARLIN P E, DUSMANU M, SCHÖNBERGER J L, et al. Lamar: Benchmarking localization and mapping for augmented reality[C]// European Conference on Computer Vision. Cham: Springer, 2022: 686-704.

[24] SARLIN P E, DETONE D, YANG T Y, et al. Orienternet: Visual localization in 2d public maps with neural matching[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21632-21642.

[25] YU T, XIONG S W. Stereo Visual Localization and Mapping for Mobile Robot in Agricultural Environment[J]. Computer Science, 2023, 50(12): 185-191.

[26] PANEK V, KUKELOVA Z, SATTLER T. Visual localization using imperfect 3d models from the internet[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 13175-13186.

[27] AGARWAL S, FURUKAWA Y, SNAVELY N, et al. Building rome in a day[J]. Communications of the ACM, 2011, 54(10): 105-112.

[28] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[29] CHEN L, CHEN W, WANG R, et al. Leveraging neural radiance fields for uncertainty-aware visual localization[C]// 2024 IEEE International Conference on Robotics and Automation(ICRA). IEEE, 2024: 6298-6305.

[30] BESL P J, MCKAY N D. Method for registration of 3-D shapes [C]// Sensor fusion IV: Control Paradigms and Data Structures. Spie, 1992: 586-606.

[31] BRACHMANN E, KRULL A, NOWOZIN S, et al. Dsac-differentiable ransac for camera localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6684-6692.

[32] SHOTTON J, GLOCKER B, ZACH C, et al. Scene coordinate regression forests for camera relocalization in RGB-D images [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2930-2937.

[33] VALENTIN J, NIESSNER M, SHOTTON J, et al. Exploiting uncertainty in regression forests for accurate camera relocalization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4400-4408.

[34] NGUYEN S T, FONTAN A, MILFORD M, et al. Focustune: Tuning visual localization through focus-guided sampling[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 3606-3615.

[35] WANG S, LASKAR Z, MELEKHOV I, et al. Hscnet++: Hierarchical scene coordinateclassification and regression for visual localization with transformer[J]. International Journal of Computer Vision, 2024, 132(7): 2530-2550.

[36] REVAUD J, CABON Y, BRÉGIER R, et al. Sacreg: Scene-agnostic coordinate regression for visual localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 688-698.

[37] LIU T R, YANG H K, LIU J M, et al. Reprojection Errors as Prompts for Efficient Scene Coordinate Regression[C]// European Conference on Computer Vision. Cham: Springer, 2024: 286-302.

[38] KENDALL A, GRIMES M, CIPOLLA R. Posenet: A convolutional network for real-time 6-dof camera relocalization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2938-2946.

[39] WALCH F, HAZIRBAS C, LEAL-TAIXE L, et al. Image-based localization using lstms for structured feature correlation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 627-637.

[40] WU J, MA L, HU X. Delving deeper into convolutional neural networks for camera relocalization[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 5644-5651.

[41] CHEN S, CAVALLARI T, PRISACARIU V A, et al. Map-relative pose regression for visual re-localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 20665-20674.

[42] SONG X，LI H，LIANG L，et al. TransBoNet：Learning camera localization with transformer bottleneck and attention[J]. Pattern Recognition，2024，146：109975.

[43] DING M，WANG Z，SUN J，et al. CamNet：Coarse-to-fine retrieval for camera re-localization[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：2871-2880.

[44] LIU X D，YU P. Cross-view Geo-visual Localization[J]. Computer Science，2023，50(S2)：407-413.

[45] LYU M，GUO X，ZHANG K，et al. A visual indoor localization method based on efficient image retrieval[J]. Journal of Computer and Communications，2024，12(2)：47-66.

[46] ZHANG B J，LIU G H，LI Z，et al. Image retrieval using compact deep semantic correlation descriptors[J]. Information Processing & Management，2024，61(3)：103608.

[47] ITO S，ENDRES F，KUDERER M，et al. W-rgb-d：floor-plan-based indoor global localization using a depth camera and wifi [C]// 2014 IEEE International Conference on Robotics and Automation(ICRA). IEEE，2014：417-422.

[48] HOFFER E，AILON N. Deep metric learning using triplet network[C]// Similarity-based Pattern Recognition：Third International Workshop(SIMBAD 2015). Springer，2015：84-92.

[49] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012.

[50] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7132-7141.

[51] ZHENG J，ZHANG J，LI J，et al. Structured3d：A large photo-realistic dataset for structured 3d modeling[C]// Computer Vision—ECCV 2020：16th European Conference. Springer，2020：519-535.

CHEN Jiwei，born in 2000，postgraduate. His main research interests include visual localization and deep learning.

TAN Guang，born in 1978，Ph.D，professor，is a member of CCF(No. 19464M). His main research interests include mobile computing，distributed computing and networking.

（责任编辑：柯颖）