

动态场景下的SLAM相关技术总结报告

作者 Author *

August 26, 2019

1 序言

视觉SLAM技术在过去的十年取得了飞速的发展，并在计算机视觉和机器人领域有着广泛的应用。但是，绝大多数的视觉SLAM算法都建立在静态环境的假设下。当环境因为物体运动而发生改变时，系统的定位精度将会受到很大影响。在这篇技术报告中，我们首先整理归纳了总结了从静态环境到多刚体和非刚体运动下的SLAM/SfM的基本建模方法和求解思路。随后，我们对现有针对运动场景的SLAM和SfM技术进行了调研。目前常见的算法可以分为两大类：一部分算法将运动区域的输入数据作为外点（outlier）剔除，来维护静态世界（static world）的基本假设。在第3章，我们总结归纳了常见的运动分割方法。这类方法主要通过找到违背SLAM问题基本几何约束，比如极线约束（epipolar constraint）、重投影约束（reprojection）以及图像映射（warping）来进行运动状态的判断。另一类算法则同时解决静态环境和动态物体的结构和姿态估计，这类方法不但需要确定观测数据的运动状态类别，还需要将属于不同运动模型的区域拆分开来。在第4章，我们总结归纳了常见的动态物体分割的方法。这类方法主要通过基于统计方法的模型选择或者子空间聚类方法将观测数据拆解为各个部分。在第5章，我们归纳总结了深度学习利用高阶特征提取上的优势

*作者介绍 Brief introduction

来端到端地获取运动区域或动态物体掩模的主要方法。最后，我们针对长时变化环境下的地图更新进行了相关的文献综述。

2 非刚体和多刚体运动下的SfM技术

为了处理场景中的动态物体，如之前所述将场景中不同运动物体进行分割，并对这些不同运动的物体分别进行三维重建是一个比较直接的方案。但考虑到所有物体的运动和三维信息同时都反应到了视频序列中，理论上这些物体的运动和三维信息可以同时进行求解[1]。在给定特征对应或者像素对应关系的基础上，基于矩阵分解的方式可以从表示了图像序列的特征矩阵中同时求解出动态物体的分割、恢复出各自物体的运动信息以及场景和物体的三维信息。如图[2]所示，这些方法根据场景三维结构在相机运动的模型下生成图像序列的过程，推导出最终的特征矩阵的特性。根据不同物体运动不同在矩阵中反应出的不同性质，对矩阵进行重新组织，并可以根据图像序列的生成模型将每部分的矩阵分解乘包含了相机运动的矩阵乘以三维信息的形式，利用矩阵中提供的约束同时完成运动物体分割、运动求解以及三维信息恢复。

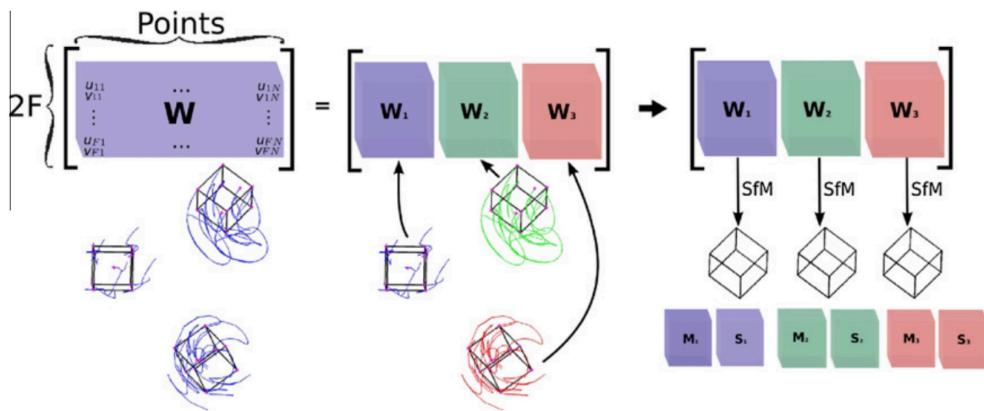


Figure 1: 多刚体下SfM矩阵可以拆解成各个子矩阵的组合来逐个单独求解 [3]。

2.1 视频序列帧中特征变化的子空间约束

与一般对特征的处理相同，假如我们可以跟踪到视频序列中的一系列特征，比如使用光流等方法[4]，如图2所示。为了便于推导，我们在这里假设这些特征在1至f帧之间均连续观测到，噪声和缺失的情况会另外探讨。我们将观测到的特征点记做 $x_{ij} = (u_{ij}, v_{ij})$ ，其中下标中*i*代表第*i*帧，*j*代表第*j*个特征点。由于这些特征点是由相机在三维空间中运动生成的，所以这一系列特征点应当连续变化并与三维场景和相机运动对应。我们将这些特征点的坐标根据编号和时间序列排布到一个矩阵中，如式(1)所示，纵坐标方向上按照帧的时间顺序排列，横坐标按照特征点的编号排列。

$$W = \begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \vdots & \vdots \\ u_{f1} & \cdots & u_{fp} \\ v_{11} & \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ v_{f1} & \cdots & v_{fp} \end{pmatrix} \quad (1)$$

矩阵分解方法认为这个矩阵是由相机的运动和三维结果矩阵合成而成，可以通过矩阵分解恢复出原始的信息。

通过矩阵分解的方式联合求解摄像机运动和三维信息，是SfM中一个重要的方法。这种方法具有优雅的数学描述，充分的考虑到了特征之间在空间和时间上的关联约束。这种方法最早由Tomasi和Kanade[1]根据秩理论在1992年提出。他们的理论指出，在一个面向静态场景的较短的序列中，包含了在整个帧序列上所有跟踪到的特征点的观测矩阵(measurement matrix)，它的秩最多是4。特别的对欧式坐标系下的垂直投影来说秩最多是3[5, 1, 6, 7]。这种秩下的约束表明了这些特征点在时间变化是有明显关联的。

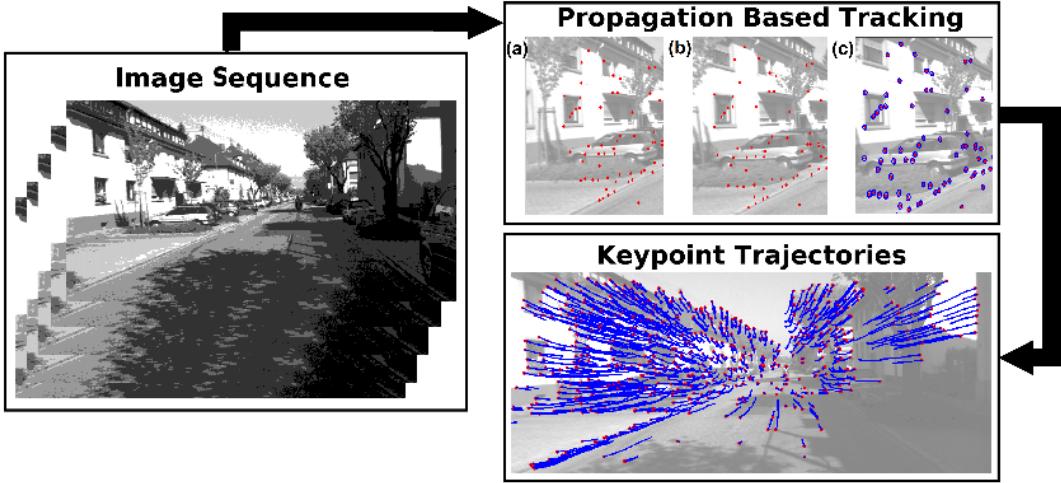


Figure 2: 跟踪得到的特征点序列轨迹[4]，观测矩阵是将图中连续出现的特征点坐标放到矩阵中，能够表示特征点在空间和时域上的关系。

2.2 观测矩阵与相机运动及三维结构的关系

观测矩阵是由相机的运动和三维点共通生成的，本节主要讲观测矩阵与这两个信息之间有什么样的关系。由于垂直投影垂直投影形式较为简单，故我们先从垂直投影的情形来做说明[1]。垂直投影的形式如式(2)所示，是三维点的 x, y 分量的直接映射。

$$x_j = \begin{pmatrix} R_{11} & R_{12} & R_{13} & t_1 \\ R_{21} & R_{22} & R_{23} & t_2 \end{pmatrix} \begin{pmatrix} X_j \\ Y_j \\ Z_j \\ 1 \end{pmatrix} \quad (2)$$

其中 R_{ij} 是旋转矩阵 R 的第 (i, j) 个元素， t_i 是平移分量的元素，上述矩阵中仅包含了旋转矩阵的前两行，第三行可以用这两行的叉乘求得。 $[X_j, Y_j, Z_j, 1]$ 是第 j 个三维点齐次坐标，我们把第 j 个三维点的齐次坐标写作 X_j ，则对第 n 帧中的 p 个点来说，三维点的齐次坐标集合可以写作 $[X_1, \dots, X_p] \in \mathbb{R}^{4 \times p}$ 。设 R_x^j 和 R_y^j 分别代表了第 j 帧姿态的旋转矩阵的第一行和第二行， t_x^j 和 t_y^j 代表了第 j 帧姿态的平移分量。我们对每个相机对整个三维点集用式(2)进行投影，将得到的二维点堆叠起

来就可以得到式(1)中的观测矩阵，如式(3)所示。

$$\begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \vdots & \vdots \\ u_{f1} & \cdots & u_{fp} \\ v_{11} & \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ v_{f1} & \cdots & v_{fp} \end{pmatrix} = \begin{pmatrix} R_x^{1T} & t_x^1 \\ \vdots & \vdots \\ R_x^{fT} & t_x^f \\ R_y^{1T} & t_y^1 \\ \vdots & \vdots \\ R_y^{fT} & t_y^f \end{pmatrix} \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix} \quad (3)$$

一个更复杂的情况是在仿射相机 (affine camera) 的模型下进行推导[8]。在这个假设下，相机的投影模型可以简化成如式(4)所示的形式，旋转矩阵中最后一行为0。

$$x_j = \pi \left(\begin{pmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} & R_{13} & t_1 \\ R_{21} & R_{22} & R_{23} & t_2 \\ 0 & 0 & 0 & d_0 \end{pmatrix} \begin{pmatrix} X_j \\ Y_j \\ Z_j \\ 1 \end{pmatrix} \right) \quad (4)$$

其中 f 是相机的焦距， (c_1, c_2) 是图像中心， $d_0 \in \mathbb{R}$ 是一个常量， R_{ij} 是旋转矩阵 R 的第 (i, j) 个元素， t_i 是平移分量的元素， $\pi(\cdot)$ 是讲齐次量变化为非齐次量的过程，即 $\pi([X, Y, Z] = [X/Z, Y/Z])$ 。我们定义如下的观测矩阵，其中每个位置为当前帧的二维位置减去第一帧的位置 $x_{ij} - x_{1j}$ 。则对每一个三维点 X_j 来说这组观测矩阵元素可以写成式(5)。

$$\begin{pmatrix} u_{ij} \\ v_{ij} \end{pmatrix} = x_{ij} - x_{1j} \quad (5)$$

$$= \frac{f}{d_0} \begin{pmatrix} (R_{11}^i - 1)X_j + R_{12}^i Y_j + R_{13}^i Z_j + t_x^i \\ R_{21}^i X_j + (R_{22}^i - 1)Y_j + R_{23}^i Z_j + t_y^i \end{pmatrix} \quad (6)$$

则显然的在仿射投影关系下，观测矩阵也是由包含姿态的矩阵乘以包含了三维点信息的矩阵得到。

对投影相机来说，上述的简单关系更复杂一些，因为齐次化过程依赖于每个像素的深度信息[9]。考虑到齐次坐标到非齐次坐标的变换，为了能够通过矩阵分解的方式进行求解，我们把齐次化中的尺度因子作为一个需要先求解的参数进行处理。我们把投影关系下的尺度因子记做 $\lambda \in \mathbb{R}$ 。我们记 $P_i \in \mathbb{R}^{3 \times 4f}_{i=1}$ 是一系列相机的投影矩阵，包含了旋转和投影及内参的关系，把齐次坐标表示的三维点 $X_j \in \mathbb{R}^{4p}_{j=1}$ 映射到第*i*个相机里，即满足投影关系 $\lambda_{ij}x_{ij} = P_iX_j$ 。在整个视频序列以及整个三维点集上完整的投影投影过程如式(7)所示，

$$W = \begin{pmatrix} \lambda_{11}x_{11} & \cdots & \lambda_{1p}x_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{f1}x_{f1} & \cdots & \lambda_{fp}x_{fp} \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_f \end{pmatrix} \begin{pmatrix} X_1, & \dots, & X_p \end{pmatrix}. \quad (7)$$

Sturm 和 Triggs[9] 根据基本矩阵和极点的约束对式(7)中的 λ 进行了求解。由于单目视频本来就无法恢复尺度，我们可以任意选择一个深度尺度，比如设 $\lambda_{1p} = 1$ 作为初始值。根据不同帧之间的极线约束，这些射影相机观测矩阵中的深度尺度可以按照式(8)进行更新求解：

$$\lambda_{mp} = \frac{(e_{mn} \times x_{mp}) \cdot (F_{mn}x_{np})}{\|e_{mn} \times x_{mp}\|} \lambda_{np}. \quad (8)$$

其中 $m, n \in 1, 2, \dots, f$, F_{mn} 和 e_{mn} 分别为*m*对*n*帧之间定义的基本矩阵和极点。在求解得所有的深度尺度 λ_{mp} 之后我们就可以得到射影相机模型下的观测矩阵。

针对射影相机的另一个处理方式是Liu等人[8]所采用的小运动近似的方式。这种方式中假设在视频序列中相机的运动相对与帧率来说非常缓慢，每帧之间的旋转运动较小，可以使用李代数到旋转矩阵的一阶泰勒展开做为近似，在这个近似下线性关系更加明确，能够得到简单的观测矩阵。

2.3 针对多刚体系统的观测矩阵分解

我们先从静态世界或者整体就是一个刚体的系统进行考虑，通过2.1节的方式，我们可以在一个视频序列中得到它的观测矩阵 $W \in \mathbb{R}^{20 \times 1}$ ，这里*f*是帧数*p*是三

维点数。根据2.2节中的推导，我们可以看出这个观测矩阵可以分解成运动矩阵 $M \in \mathbb{R}^{2f \times 4}$ 和形状矩阵 $S \in \mathbb{R}^{4 \times p}$ 即如式(9)所示：

$$W = MS. \quad (9)$$

在求解过程中，我们在得到观测矩阵 W 之后，基于秩约束 (Rank constraint) [10]，矩阵 W 可以用奇异值分解 (SVD) 进行分解，得到

$$W = U\Sigma V, \quad (10)$$

的形式。其中 $\Sigma \in \mathbb{R}^{4 \times 4}$ 是一个对角阵，包含了最大的四个特征值， $U \in \mathbb{R}^{2f \times 4}$ 和 $V \in \mathbb{R}^{p \times 4}$ 是对应到最大的四个特征值的特征向量。之后我们可以用 $\hat{M} = U\Sigma^{1/2}$ 和 $\hat{S} = \Sigma^{1/2}V^T$ 来表示运动矩阵和形状矩阵。但是式(10)中的分解并不是唯一的，真实的运动矩阵 M 和形状矩阵 S 还需要再找到一个映射矩阵 A 使得整个分解过程如下式所示：

$$W = MS = (\hat{M}A)(A^{-1}\hat{S}). \quad (11)$$

其中矩阵 A 可以通过旋转矩阵和平移所带有的先验约束进行求解，并可以转化成一个最小二乘形式线性求解过程 [10, 1]。

上述就是通过分解形式联合求解静态场景问题的基本框架，这个框架可以比较容易的推广到场景中存在独立运动的多个刚体的情形 [10]。我们假设场景中包含了 n 个独立运动的刚体，则我们可以通过列交换的形式把观测矩阵 W 中的特征序列根据刚体分组成 $[W_1, \dots, W_n]$ ，这个过程可以用一个排列矩阵 Γ 来表示，如式(12)所示，其中 $\Gamma \in \mathbb{R}^{p \times p}$ 是一个未知的排列矩阵。

$$\bar{W} = W\Gamma = (W_1, \dots, W_n) \quad (12)$$

在没有噪声的情况下，每一个独立的 W_i 即和前述静态场景中的观测矩阵等价应当在一个秩不超过4的子空间中。这样每个 W_i 可以进行单独的分解，得到该刚

体的运动矩阵 M_i 以及形状矩阵 S_i , 如下式所示:

$$\bar{W} = \bar{M}\bar{S} = (M_1, \dots, M_n) \begin{pmatrix} S_1 & & \\ & \ddots & \\ & & S_n \end{pmatrix}. \quad (13)$$

这样对多刚体运动的问题来说, 最关键的就是求解排列矩阵 Γ 让分解得到的矩阵 \bar{S} 具有块对角的性质。

多刚体运动恢复结构 (Multibody Structure from Motion) 对标准SfM下刚体相机的运动进行了拓展, 变成了n个刚体的刚性运动模型。为了解决多刚体运动恢复结构问题, 在仿射相机模型的假设下Costeira和Kanade[10] 引入了一个形状交互矩阵的概念。这个理论里对物体形状构造了一个数学上的可证明的、对刚体运动具有不变性、不依赖于坐标系的描述。这里的结构交互矩阵被证明可以保持在原始子空间里的结构。我们设 $\bar{W} = U\Sigma V^T$ 是一个秩r的观测矩阵SVD 分解结果, 其中 $U \in \mathbb{R}^{2f \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{p \times r}$ 。则形状交互矩阵 Q 就可以定义为:

$$Q = VV^T \in \mathbb{R}^{p \times p}. \quad (14)$$

式(14)具有一个特殊的形状, 当两个特征序列分别属于是两个刚体的时候, Q 的元素会是0。这个特性可以在数学上得到证明[11]。在这个理论基础上, 矩阵分解的求解过程可以基于对 Q 的排序和元素大小的限制来或得不同刚体的分割以及三维信息的重建。在[10]中, 这种运动刚体的分割和聚类是通过最大化式(13)中 S 的对角元素, 并且利用 Q 来约束对角线上的每个块应当属于不同的刚体的约束这样的方式进行求解。Ichimura[12] 使用在[13]中最大化不同子空间的差异性的判别准则将 Q 中的不同刚体到不同的刚体。

2.4 针对非刚性运动的观测矩阵分解

假如场景中的物体是非刚性运动的时候, 情况会非常复杂。Bregler[14]等针对非刚性运动恢复结构上针对垂直投影相机提出了第一个观测矩阵分解的方式。

他们的核心想法是将一个非刚体的物体表示成一组基的组合，认为这个刚体在这些帧中的运动过程可以在这个状态空间中进行近似表示。比如我们选择 k 个关键帧中的结构作为一组基 $B_{i=1}^k$ ，其中每个 B_i 代表了一个 $3 \times p$ 的矩阵，表示了 p 个特征点。这组基的线性组合 $B = \sum_{i=1}^k l_i B_i$ 可以确定一个刚体的描述，其中 $l_i \in \mathbb{R}$ 是一组系数。通过[1]中的方法进行中心化并消去平移向量之后，可以将观测矩阵表示成式(15)所示：

$$\tilde{W} = NB = \begin{pmatrix} l_{11}R'_1 & \cdots & l_{1k}R'_1 \\ \vdots & \ddots & \vdots \\ l_{f1}R'_f & \cdots & l_{fk}R'_f \end{pmatrix} \begin{pmatrix} B_1 \\ \vdots \\ B_k \end{pmatrix}. \quad (15)$$

其中 R' 是旋转矩阵的前两行，由于垂直投影的假设这里 R 就是旋转矩阵不包含内参的过程，所以旋转矩阵的第三行可以通过前两行的叉乘得到。针对观测矩阵 \tilde{W} 进行SVD分解，根据秩3的约束，我们选择3个最大的奇异值及对应的特征向量。则旋转矩阵元素 R'_i 和形状基的系数 l_{ij} 可以从 N 中，通过重新排列 N 的顺序并且对它进行SVD分解恢复出来。与刚性问题中面临同样的问题即SVD可以得到的结果是不唯一的，最后可以通过正交约束求解得到一个映射矩阵 G 得到 R'_f 和 B_k 的唯一解。另一中约束是[15]中引入了一个新的基约束，使得非刚性的分解问题能够使用闭解的形式进行求解。除了直接使用度量约束 (metric constraints)，Paladini等[16]将运动矩阵投影到一个矩阵流形的约束上，让整个分解过程可以通过迭代的方式进行处理。在这些工作的基础上，Dai等[17]尝试去掉针对非刚性重建的额外假设，比如之前使用的非刚性基、针对非刚性场景的先验等，提出了一个没有额外先验的仅使用低秩约束的方法。Kumar等[18]提出了融合多刚体和非刚体的方法，将问题建模成多个非刚体变换的系统。他们讲整个特征轨迹建模乘联合的多个线性或者仿射的空间。可以允许同时优化非刚性的重建和刚性的重建。

3 基于运动分割的SLAM技术

动态分割（又称为移动物体检测/分割[19, 20, 21]）透过将特征区分为两类特征，静态和动态特征，以检测图像中移动的区块。明确地说，给定在图像空间中的特征点集合，动态分割将特征点聚类为静态集合和动态集合。传统的视觉SLAM使用鲁棒的静态方法，计算几何模型（如基本矩阵、单应矩阵）来实现动态分割。比如利用随机抽样一致算法（RANSAC）[22]和特定的距离度量方法（如桑普森距离[23]）将不服从模型的点剔除。当静态的特征点占了主体时，这种方法能够有很好的效果。当动态物体占了相机的绝大部分，或是捕捉到的场景被巨大的移动物体遮挡时，这种方法可能就会失败。其他方法结合其他传感器来解决这个问题，如使用惯性测量单元（IMU）估计相机自身的运动[24, 25]。通过IMU得到的位姿估计可以用来初始化相机位姿并且鲁棒地分割静态和动态特征。在这一章节，我们讨论传统视觉SLAM和视觉惯性SLAM之外，分割静态和动态物体的其他方法。

3.1 背景 - 前景初始化

背景-前景初始化技术假设系统有对于环境的先验知识，利用这个信息分割静态和动态特征。这个先验知识可以归于背景（静态特征）或是前景（动态特征）。如果系统的先验知识是关于前景目标，表示系统知道相机前方运动物体的类型或形状。

大部分前景初始化的实现方法使用tracking-by-detection结构[26, 27]，如图4所示。Wangsiripitak等人[28]假设对于一个三维物体，它的动态特征的位置是已知的。他们用边上的控制点集建模一个三维多面体模型，然后使用Harris' s RaPid tracker[29]对其进行跟踪。如果之前跟踪的特征位于跟踪中的三维多面体上，当检测到这个物体处于移动状态时，便将特征剔除。被这个物体遮挡的静态特征也会被剔除。相似地，Wang等人[30]假设移动中的物体上的SURF特征描述子[31]是已知的并存在数据库中。通过比较在特征检测步



Figure 3: Tracking-by-detection [26]示意图。蓝色框内为检测结果，红色框内为预测的匹配结果。

骤得到的描述子，就可以识别移动中的物体，也可以估算它的位移和旋转。Chhaya等人[32]使用deformable wireframe object class model对建模相机前方的车辆。这个模型使用Principal Component Analysis (PCA) 在3D CAD data上训练而得。这个模型用在位姿估计得过程将车子识别并分割出来。另一方面，Lee等人[27, 33]基于tracking-by-detection结构，他们使用Constrained Multiple-Kernel (CMK) 法，利用深度信息处理跟踪过程中的遮挡问题，同时使用预训练的人类检测器来跟踪行人。

有别于初始化前景的物体，背景初始化采用背景提取 (background subtraction) 技术 [34, 35]，如图??所示。Zhang等人[36]初始化属于背景的特征点集合，并令这个集合为背景的模型。他们假设当首次进行视觉的初始化时，是没有前景物体的。当经过新的一帧后，使用GPCA[37]进行三维动态分割。分割出来的动态部分，对于先前的背景模型具有最高响应值部分，将会用来更新背景。根据新的背景模型，运用标准的对极几何法估计位姿。

3.1.1 基于几何约束的运动分割

由于动态的特征会违反静态场景中的多视角几何约束，依赖此约束的技术是利用极限几何的性质[23]来分割静态和动态特征。这些约束可以从极线、三角



Figure 4: 基于 [34]的背景提取结果。

化、基本矩阵估计或重投影误差的等式中得出。

Kundu等人[21]根据机器人里程计构建一个基本矩阵来定义两个几何约束。第一个约束是极线几何约束，在随后的视角下，成功匹配的点应该要在对应的极线上。如果跟踪到的特征离极线过远，就会被认为是动态特征。第二个约束是Flow Vector Bound(FVB)，目的是分割当三维点沿着极线移动时产生的退化运动。通过设定跟踪到的特征流的上界和下界，超过这个范围的特征就会被作为运动的特征检测出来。最后通过循环的贝叶斯滤波器决定将特征分类为静态特征或动态特征。不同于使用极线约束，Migliore等人[38]利用三角化的原理分割静态和动态特征。他们在概率滤波器的框架下，持续地检查三个不同视角投影出来的视线的交点。如果特征是动态的，则这个交点在运动的过程中不会一样，甚至不会产生交点。但是由于传感器存在的噪声，他们使用Uncertain Projective Geometry[39]，将测量的不确定性加入他们检查不同视线关系的过程。最后通过统计方式分类静态和动态的特征。

将移动的物体误分类为静态物体并将其加入位姿估计，将会严重地使SLAM系统的性能降低，Lin等人[40]通过观察这一性质来检测移动中的物体。他们计算两种不同条件下的位姿估计之间的差异，其中一个不加入新检测到的特征，另外一个假设新检测到的特征是静态的并加入位姿估计。借由计算两个结果的距离，设定一个门槛，通过二分贝叶斯滤波器整合，就能够精确地分割静态和动

态的特征。

另外一个几何的方法是利用重投影误差。Zou和Tan[41]将先前帧的特征投影到当前帧上，测量这些跟踪到到特征的距离。通过这个重投影的距离分类静态和动态的特征。Tan等人[42]使用同样的投影原理检测动态特征。他们同时将遮挡问题作为考量提出了一个鲁棒的视觉SLAM。在一个特征投影到当前帧上时，他们检测图像上的外观差异，也就是图像中的某一部分是否改变了。如果外观差异巨大，有极大的可能这个区域被动态物体遮挡，或是因为视点改变被静态物体所遮挡。因为上述原因被遮挡的三维点，会被保留并用来估算相机位姿。

3.2 基于光流的运动分割

光流的定义了两个连续的图像间，图案亮度的表面运动[43]。通常它对应了图像的运动场，因此可以用于分割移动的物体。Klappstein[20]根据光流定义了运动度量描述移动中的物体的似然。测量当场景中有运动物体时，光流错误的范围。Graph-cut算法根据运动度量分割移动的物体。

Alcantarilla等人[44]通过运动似然残差得到的场景流(三维的光流)中的三维运动向量系数，分割移动物体。马氏距离用于考量基于稠密光流和双目重建计算场景流的测量不确定性。如果残差小，特征点很可能属于静态物体。在运动似然残差设立门槛，属于运动物体的特征点可以从SLAM过程中删除，使得视觉里程计的估计更鲁棒。Derome等人[19, 45]计算预测的预想与双目相机观测得到的图像的残差，以此计算光流。预测的图像是根据估计的相机自身运动，将当前帧的双目图像转换到过去的帧上。接着从残差场上的斑点检测移动中的物体。

4 基于动态物体分割的SLAM技术

基于动态物体分割的SLAM方法将场景中的特征对应聚类成不同类别，并跟踪它们的三维轨迹，将复杂动态场景拆解为一个一个刚体运动的子空间解耦求解。

动态物体分割(也称为multi-body segmentation或eorumotion segmentation)将图像中的所有特征对应关系聚类成 n 个不同的运动类别。由于该问题是一个鸡生蛋、蛋生鸡的问题(chicken-egg problem)，因此求解较为困难。为估计动态物体的运动，应首先对特征进行聚类；另一方面，聚类特征需要知道所有运动物体的运动模型。由于遮挡、运动模糊和特征跟踪丢失而带来的噪声、异常值和特征对应缺失使问题变得更加复杂。另一个挑战是如何处理退化运动(如当物体与相机在相同平面上以相同方向和速度移动)或相关运动(例如两个人一起移动，关节运动)的场景。

4.1 静态模型选择

在静态场景中，可通过一个运动模型来描述相邻帧特征点的对应关系。相反，动态场景中的特征点可能来源于具有不同运动模型的物体。运动模型通常基于以下几种模型：基础矩阵(F)，仿射基本矩阵(F_A)，本质矩阵(E)，单应性/投射(H)或仿射(A)矩阵。选择模型时，通常会尝试将所有可能的运动模型与数据拟合，并选择最适合数据的模型。如果数据可以由动态场景中的多个模型描述，则基于这些运动模型，需要建立许多假设才能进行动态物体的分割。

基于统计的3D动态物体分割方法会对数据的子集进行采样，并通过RANSAC [?]或Monte-Carlo采样迭代 [46]用模型进行数据拟合。该类方法使用运动模型将数据划分为局内点(inlier，即符合该运动模型的数据)和局外点(outlier，不符合该运动模型的数据)。之后，对剩余的数据(即当前模型对应的局外点)再次采样，用新采样的数据拟合另一个运动模型，实现动静物体分割。该过程迭代重复，直到所有数据都可以由 n 个运动模型描述，或者剩余的异常值不足以产生更多的运动模型。迭代收敛后，该运动分割过程可从头开始，以生成许多候选的运动分割假设供后续选择。

目前，学者们已提出许多基于统计模型选择的方法来实现动态物体分割。Torr [47]对邻近的特征对应进行采样，并在RANSAC迭代下计算不同的运动模型(F, F_A, H, A)。该方法用GRIC选择最能符合聚类局内点的运动模型。

然而，当所选模型的局内点数量低于阈值时，则使用EM算法(expectation maximization)。为避免耗时费力的暴力计算，Schindler和Suter [48] [46]提出了局部蒙特卡罗采样法，从图像上定义的子区域中进行采样。他们提出了一种从数据中估计噪声大小(noise scale)的方法，从而可以恢复每个运动的残差分布及其标准差。此外，他们推导出了一个新的似然函数，允许运动模型(F, H)重叠，由GRIC选择最佳模型。

虽然之前的方法是对两个图像序列进行操作，但Schindler等人 [49]将 [48]中的技术扩展到一般运动模型(本质矩阵 E)下的多视角图像。为了从多于两张图像的序列中将若干被选的本质矩阵关联起来，他们通过仅连接那些具有相似内点集的基本矩阵，来加强时间一致性。最后，使用类似MDL的方法来选择描述运动的最佳模型。这种方法已被Schindler等人泛化为适用于任何相机模型(不仅是投影相机)和运动模型(不仅是基本矩阵 E)的动态物体分割方法 [50]。另外，Ozden等人也考虑了许多实际因素 [51]，他们研究了如何将之前识别出的动态物体与静态背景融合，或如何将聚类分成两类不同运动模式的问题。

Thakoor等人 [52]将模型选择描述为组合优化问题。采用branch-and-bound技术，通过将优化问题分解为较小的子问题，使用AIC作为代价函数来优化运动分割。对对应关系的局部采样也用于生成运动模式，而空假设(null hypothesis)用于处理局外点。最近，Sabzevari和Scaramuzza [53]利用了基于投影轨迹矩阵分解的统计模型选择技术，通过对极几何生成运动模型，用重投影误差剔除不合理的假设。通过一次次迭代时不断优化结构估计和运动分割估计后，评估当前假设的合理性。该方法已在 [54] 中得到拓展，通过施加相机位姿约束，相机和动态物体的位姿可以分别用单点法(one-point algorithm) [55] [56]和两点法(two-point algorithm) [57]求得。

4.2 子空间聚类

子空间聚类是基于以下观察而被提出的：许多高维数据可以由低维子空间的并集来表示。数据点的子空间可以表示成基向量和数据的低维表示。子

空间聚类框架下的三维运动分割问题基本可描述为找到与每个物体运动相关的子空间，并将数据拟合到子空间中。然而，由于在实际中子空间和数据分割是未知的，因此只能同时进行子空间参数估计和数据的子空间聚类。Costeira-Kanade [58] 和 Gear [59] 通过观察发现独立的刚体运动位于线性子空间中，从而首次提出了用子空间聚类进行运动分割的方法。通过施加秩(rank)的约束，可以恢复每一个线性子空间。

Kanatani [?] 提出“子空间分离”的概念，将其作为聚类低维子空间的一般方法(不仅限于运动分割)，借用统计模型选择实现子空间分离。该方法通过平衡在拟合数据点到子空间时残差的增加及子空间合并时自由度的减小，来选择最佳的子空间划分，并使用最小平方中值(least median of squares)法拟合带有局外点的数据。不同的是，Vidal等人 [60] 扩展了主成分分析(principal component analysis)的概念，提出广义主成分分析(generalized principal component analysis, GPCA)法。PCA仅适用于数据都位于同一线性子空间内的情况，而GPCA可适用于数据位于多个线性子空间的情况。GPCA通过多项式嵌入(或Veronese映射)用 n 阶齐次多项式拟合数据，并通过计算特定点处多项式的导数找到每个子空间的法线，以解决寻找子空间的问题。然后，通过从法向量之间的角度计算相似性矩阵，使用谱聚类进行分割。为了在分割中考虑实际情况，在聚类前，该方法将数据投影到较低维空间中 [60]，然后通过找到多项式嵌入的秩来计算运动模式的数量 n 。

由于之前的算法假设运动是刚性的，与实际情况还有一定差距，因此Yan 和 Pollefeys [61] 提出了一个称为局部子空间仿射(local subspace affinity, LSA)的通用框架，可用于独立、铰接、刚性、非刚性、退化以及非退化的运动模式。LSA通过对点及其最近邻的区域进行采样，用局部子空间拟合采样数据，来估计子空间。最近邻的点可通过计算矢量之间的角度或距离得到。然后，通过计算两个局部子空间的夹角求解仿射矩阵，并对仿射矩阵进行谱聚类得到聚类结果。在估计子空间之前，还要将数据向低维子空间投影。与LSA类似，Goh 和 Vidal [62] 也用一个点及其最近邻点拟合局部子空间。基于局部线性嵌入法

(locally linear embedding, LLE) [63]，学者们提出了局部线性流形聚类法 (locally linear manifold clustering, LLMC)。该方法使用LLE将数据转换为低维表示，并计算由LLE生成矩阵的零空间(null space)来将与每个运动相关联的分离的流形进行聚类，而其中数据的分离可由零空间中的向量表示。

Elhamifar和Vidal [64] [65]给出了另一种观点，它利用稀疏表示将运动模式进行聚类。他们发现线性或仿射子空间的并集中的点可以表示为子空间中所有数据点的线性或仿射组合，并提出了稀疏子空间聚类法(sparse subspace clustering, SSC)。然而，只有当点可被表示为同一子空间中数据的线性或仿射组合时，才能获得稀疏表示。在无噪声数据下，可以通过求解 L_1 最小化问题来估计稀疏系数。给定稀疏系数，可以构建仿射矩阵，通过谱聚类来完成聚类。Rao等人开发了SSC的扩展方法 [66]，它们融合了稀疏表示和数据压缩以处理实际问题，例如数据丢失，不完整或包含局外点。最近，Yang等人 [67]提出了矩阵补全技术来改进SSC算法。与稀疏表示相反，Liu [68]和Chen等人 [69]采用低秩表示(low rank representation, LRR)，使用谱聚类来定义仿射矩阵，实现子空间分割。

值得注意的是，大多数子空间聚类技术以批(batch)处理的形式进行。Vidal [70]设计了一种迭代聚类技术，适用于位于多个移动超平面中的数据。该方法用一组时变多项式模拟了一组移动超平面，利用梯度下降，通过估计超平面的归一化法向量来递归地完成分割。Zhang等人提出了在线子空间聚类的另一种实现方式 [71]。他们修改了K-flats算法，使其能够增量地获取输入数据。他们将 L_1 用作目标函数(而非 L_2)，以便在数据包含噪声和局外点的情况下增强算法的鲁棒性。

在过去的几十年中，子空间聚类已经成为被广泛研究的课题，并且出现了许多方法。有关子空间聚类更详细的介绍，可阅读文献 [72]。

4.3 动态物体的3D跟踪

在实际应用中，跟踪三维空间中的动态物体并估计其三维坐标和深度具有非常重要的意义。该问题的挑战在于，一般的视觉SLAM系统假设场景是刚性、完全静止的，并用三角化方法 [73] 估计场景的三维结构，但该方法并不适用于动态物体，因为特征点重投影之后并不重合。假设 x_1 和 x_2 分别为第一帧和第二张帧图像中对应的特征点，则它们对应的3D点 X 的坐标应该能通过经相机投影矩阵 P_1 和 P_2 ，计算 x_1 和 x_2 重投影光线的交点得到。但是，由于动态物体的运动和相机不同，因此从第一帧到第二帧重投影的光线也在移动，因此二者不相交。为解决动态物体的3D跟踪问题，需要其他的技术作为辅助。本节主要讨论估计动态物体的3D轨迹的相关工作。

由前面的论述可知，标准的三角化测量 [73] 不能用于重建动态物体的3D结构，因为重投影的光线不相交。Avidan和Shashua [74] [75] 提出了轨迹三角化测量方法，可在目标轨迹已知或满足某种参数形式的前提下，重建动态物体的三维点云。该方法假设3D点沿着三维空间中一条未知的直线运动，于是该问题就变成了寻找一条直线，与 t 个视角的投影光线都相交的问题。在 t 至少为5时，该问题有唯一解。因为3个视角的投影光线交叉将形成二次曲面，第4个视角的投影光线与该二次曲面交于2点，此时而第5个视角的投影光线便可使交点唯一确定。

与之相反，Shashua等人 [76] 假设物体在圆锥截面上移动。在这种假设下，需要至少9个视角才能获得唯一解。如果已知圆锥曲线的类型（圆、椭圆等），则只需要至少7个视角即可。他们通过将圆锥方程拟合二维空间中的运动点，或通过最小化三维空间中估计圆锥半径的误差来解决该非线性优化问题。基于之前的相关工作，Kaminski和Teicher [77] [78] 将轨迹三角化问题泛化为在投影空间中找到一个超曲面簇。这种多项式的表示形式将非线性轨迹估计问题转化为未知参数空间中的线性问题。另一方面，为了处理缺失的数据，Park等人 [79] 将3D轨迹表示为轨迹基矢量的线性组合，这样就可以用最小二乘法鲁棒地估计三维点云的位置。他们还通过分析相机运动、图像点运动和轨迹基

矢量之间的关系，提出了可重构性标准(reconstructability criteria)。由于可重构性与3D重建误差成反比，因此该标准可以准确地考察精确重建的可能性 [80]。

由于可观测性问题（观察者与目标之间的距离无法观测得到），单目相机进行动态物体的3D跟踪可被视为仅可跟踪(bearing-only-tracking, BOT)问题。单目相机可以被视为一种BOT传感器，它只能提供动态物体上被跟踪特征点的方向信息（例如上一帧和当前帧中对应的特征相对相机中心的角度）。对于BOT问题，往往使用基于滤波的方法，因为它可建模相机和目标的位置和速度的不确定性，并且已得到广泛的研究 [81] [82]。

Kundu等人 [83]采用粒子滤波器来估计动态物体的位置和速度，他们用瞬时匀速运动模型(instantaneous constant velocity motion model)建模未知运动，并用李代数(Lie algebra)参数化动态物体的刚体变换。在初始化时，用几何约束和流向量绑定(flow vector bound)对动态物体进行分割 [?] [84]，并且使粒子沿投影光线均匀分布。利用由静态场景的三维点云估计得到的地平面和允许的最大深度值来约束粒子的空间。之后，将每个粒子投影到当前帧，计算重投影误差(投影点和实际特征点位置的误差)来更新每个粒子的权重。由于较低误差或较高权重的粒子具有较高的重采样概率，因此它们集中于能够产生最小重投影误差的深度值周围。

5 基于深度学习方法的动态分割

随着深度学习在越来越多的计算机视觉任务中表现出优异的性能，近年来也有不少研究将深度学习用于解决动态物体的分割问题。在动态物体分割的研究中，许多学者使用了空间变换网络(Spatial Transformer Networks) [85]。这是因为动态物体分割的过程中涉及动态物体的识别，而同一类物体的大小、位置、姿态往往在不同图片中不尽相同，因此需要网络能抵抗这些因素的干扰，准确识别出物体，即网络的识别需要有空间不变性(spatially invariant)。而

空间变换网络能以自监督学习的方式在网络内部对空间数据进行变换处理，使其具有空间不变性。该网络结构可作为一个模块嵌入到任何物体识别、检测、分割网络中，提高网络的性能。

由于动态物体分割本质上是将视频流中的动态物体识别、分离的过程，因此它可用深度学习中的注意力机制(attention)解决。近年来，出现了许多将注意力机制用于动态物体分割的研究工作，例如 [?]用强化学习的方式训练循环神经网络(recurrent neural network, RNN)，引入注意力机制使其输出图片中的多个物体。

目前，深度学习用于动态物体分割的研究工作往往需要预定义刚体的类型、运动模式或数量。将三维点云或光流作为输入，深度网络预测出动态物体的掩膜。Byranvan和Fox等人提出了SE3-Net [86]，能够从三维点云中将预先定义好的 n 个动态物体的6自由度位姿以 $SE(3)$ 的形式预测出来。SE3-Net设计了一个编码-解码网络，使用卷积和反卷积预测每个动态物体的掩膜和6自由度位姿。其中，编码网络由两个并行的卷积和全连接网络构成，将输入的三维点云分别变换成隐变量和控制向量(control vector)。随后，将隐变量和控制向量拼接起来，由解码器(同样由两个并行的反卷积和全连接网络构成)输出稠密的物体掩膜和 $SE(3)$ 变换参数。最后，用一个非线性变换层将三维点云、物体掩膜和 $SE(3)$ 变换融合，生成动态物体的三维点云。

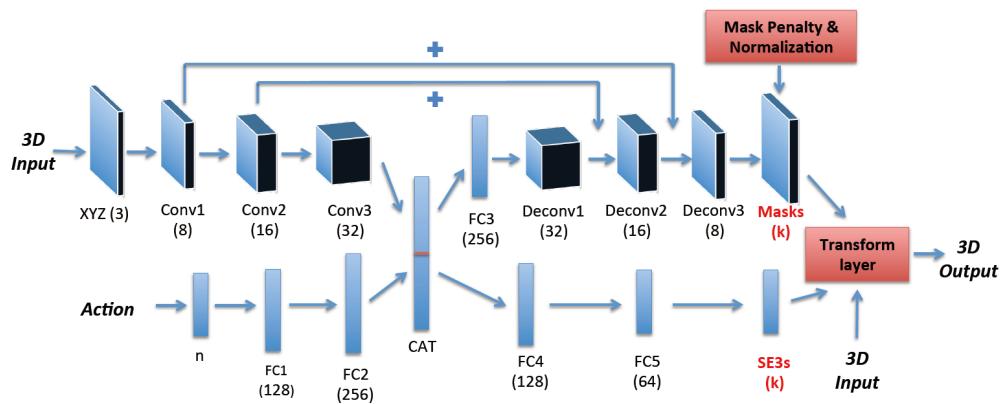


Figure 5: SE3-Net的网络结构图

Vijayanarasimhan等人通过实验证明，可以借助光流用深度学习分割场景中的动态物体 [87]。他们设计了SfM-Net，通过显式的几何约束训练网络，使其可预测场景深度、相机运动和动态物体分割。SfM-Net由两个主流的卷积、反卷积网络构成，它们分别作为结构网络(structure network)和运动网络(motion network)。其中，结构网络通过学习预测场景深度，运动网络估计相机和物体位姿。在经过卷积网络的嵌入层(embedding layer)之后，通过两个全连接层输出动态物体的位姿估计。同时，嵌入层经过反卷积输出运动物体的掩膜估计。之后，通过估计的深度图，利用估计的相机和物体位姿将一帧RGB图像中的像素变换到另一帧的视角下，合成新视角下的图片，从而计算场景的光流。利用显示的几何约束关系，便可以自监督学习的方式通过最小化光度误差(photometric error)进行训练。

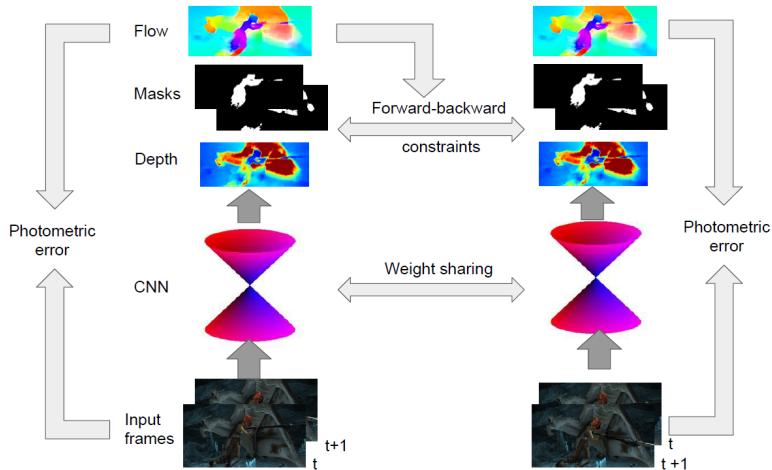


Figure 6: SfM-Net的流程图

与SfM-Net类似，Yin等人提出的GeoNet [88]自监督学习的方式，利用三维几何约束，将单目深度估计、光流估计和相机运动估计联合学习求解。为了能够恢复出完整场景的光流信息，GeoNet同样将场景显式地分为静态和动态部分，将各种估计通过视角合成生成新视角下的图片，并将损失函数建立在生成图片和拍摄图片的误差上，进行联合的自监督训练。如图 7，GeoNet利用前向和反向两个操作，判断区域内运动是由静态背景还是动态物体造成的，然后分

别求解静态和动态部分的光流，合成整个场景的完整光流。另外，为增加对局外点(outlier)、光照变化、遮挡、无纹理和重复纹理区域的鲁棒性，GeoNet还使用了自适应的几何一致性损失函数。目前，GeoNet在室外车辆驾驶环境下，在深度、光流估计上均取得了非常不错的效果。

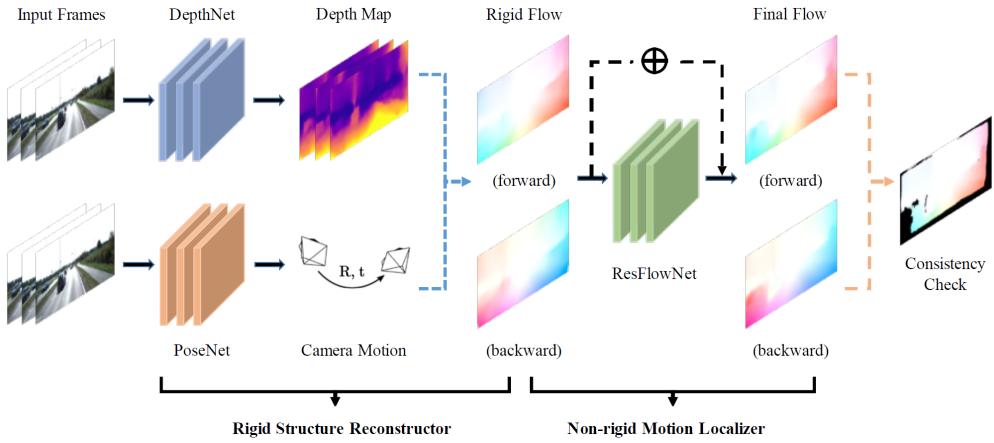


Figure 7: GeoNet的流程图

Lai等人认为，双目的深度估计和光流估计有其相同的地方，即寻找对应点的匹配和计算移动距离(视差)，而此前的许多工作将二者分别用不同的网络估计，只在损失函数中将其耦合。因此，Lai等人将场景深度估计和光流估计用同一个网络求解 [89]，共享高维的特征表示，并在SfM-Net和GeoNet这类工作的思路下，充分利用了两个时刻双目图像之间的各种几何约束(如图 8所示)，使光流、深度估计的精度更高，从而有助于动态物体的分割。

最近，Wang等人提出了UnOS网络，同样用双目自监督学习方法联合估计光流和场景深度 [90]。UnOS使用3个网络同时求解深度、相机位姿和刚性场景下的光流(rigid optical flow)，并将刚性光流与FlowNet估计的光流作比较，找出符合刚性场景假设(rigid-scene assumption)即静态的部分。然后，促使两个光流估计在静态部分尽可能一致，那么余下的部分即是动态物体，由此得到动态物体的初步掩膜。之后，使用视觉里程计优化初步估计的掩膜，得到更精准的动态物体分割。在整个自监督训练过程中，除了图片合成作为损失函数之

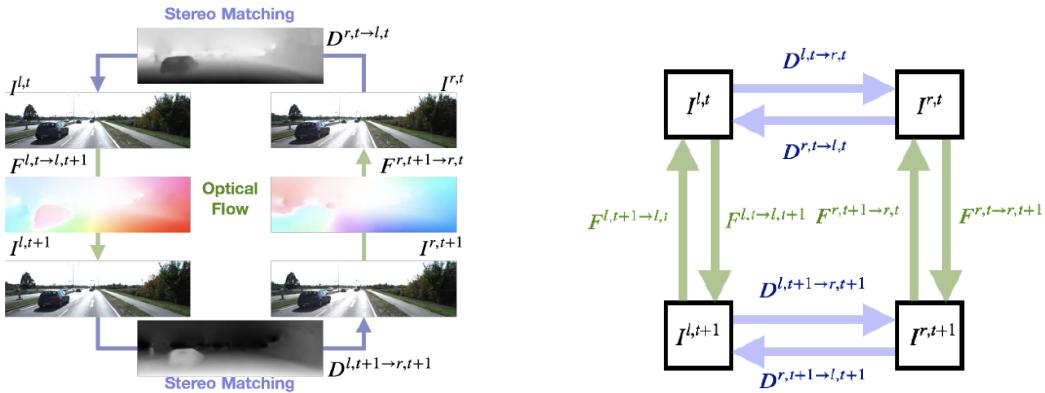


Figure 8: 左边为 [89]的流程图，右边为该方法利用的各种帧间几何约束

外，UnOS还使用了光流-深度一致性损失函数，使该方法在双目光流、深度估计、动态物体分割任务上均取得了很好的效果。

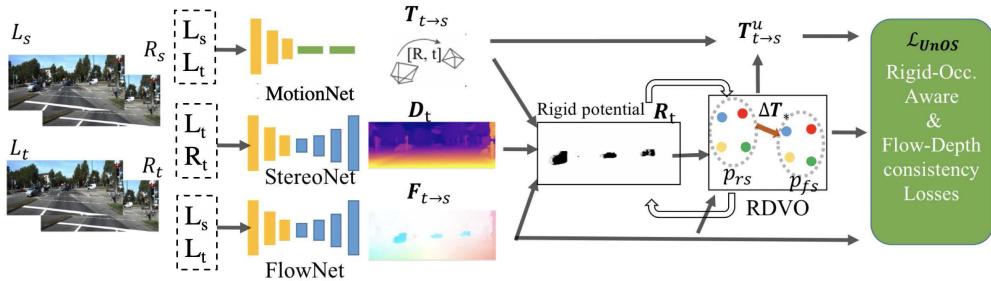


Figure 9: UnOS的流程图

6 长时变化环境下的地图更新

相比于基于稀疏特征的视觉SLAM算法，稠密视觉SLAM技术（dense visual SLAM）通过更注重于维护高质量、可复用的三维地图来帮助传感器定位。由于便携的消费级深度传感器的出现，室内场景的稠密视觉SLAM在近些年取得了不错的进展。KinectFusion [91]首次利用RGBD数据实现了实时的稠密定位和数据融合，并在场景尺度 [92]、回环调整 [93] 以及计算效率 [94] 上有着一系列的拓展。这类方法建立在场景完全静态的严格假设下，当运动物体区域的点云

数据被融合到三维地图中，将会带来系统不可逆的崩塌。现有的针对动态场景和环境变化下的定位方法可以分为三类：一类方法只将观测数据中的静态区域融合到三维地图中，确保基于地图的定位方法仍然建立在静态世界的假设下；一类方法分别构建静态地图和动态地图，利用动态地图的历史时序信息作为先验，以提升系统的精度和鲁棒性；还有一类方法维护整个地图在时序上的变化情况，通过引入时间维度的信息将环境描述成一个随时间而转移的状态量，通过反映出的环境变化情况提供更好地预测信息。

6.1 动态环境下静态部分地图的构建

如前文所述，对于存在运动物体的场景，动态的观测数据违反了基本的几何约束，需要被视为离群点从地图中剔除，这一思想与基于运动分割的SLAM技术十分相似。对于稠密视觉SLAM任务来说，维护静态的三维地图能充分地进行数据融合，也为观测提供了更加完整的运动状态先验。相应地，应对运动物体的挑战主要包括如何避免将运动部分的数据融入地图，如果运动部分数据没有被很好地消除，用于定位的地图信息就会使问题变得复杂起来。



Figure 10: 博恩德意志博物馆中的交互式解说机器人Rhino可以在人群中进行准确的自定位 [3]。

事实上，通过维护静态地图和采用鲁棒的定位策略在很早的时候就被广泛研究。Fox等人 [3]发现，Markov localization通过维护整个状态空间的概率密

度，可以在环境偶尔变化的情况下能够保持稳定，比如门的开关或人的走动。然而，当大量物体没有包含在静态地图中，比如摄像头被室内的人群包围时(如图 10所示)，相机定位将会失败，其主要原因在于马尔科夫假设在高动态环境下并不成立。Fox等人利用entropy filter和distance filter两种滤波方法选出输入数据中没在地图中的部分，将状态空间离散化，从而高效准确地更新置信状态，保证传感器在复杂动态场景下的定位鲁棒性。

ElasticFusion [95]可以应对画面中存在少量运动物体的场景。算法并未显式地检测运动物体，而是将动态环境下的稠密重建作为一个鲁棒估计问题，通过统计的方式自主地将动态区域作为外点剔除。在这个工作的基础上，[96] 从重建的角度出发，认为每个面元只有在多个连续帧被反复观测到才可以融合到三维模型中。当输入的点云数据与匹配上的地图点位置距离过远时，这部分点云会被作为种子点，通过区域生长将当前帧分割成静态和动态区域。相应地，地图上与动态区域有着匹配关系的部分将从地图上剔除掉。通过这种不断更新地图的方式，当之前静态的物体发生运动时，系统可以有效地检测出运动状态的变化，以消除这部分数据对系统鲁棒性的影响。

BaMVO [97] 利用背景提取领域(background subtraction)广泛使用的非参数化背景模型进行稠密视觉里程计估计。通过存储连续的4帧深度图并对齐到同一个视角，背景区域可以根据多帧对齐后的深度值差异来进行判别。这样的多帧判别方法建立了时域上的连续性，但是由于采用帧到帧(frame-to-frame)的定位策略，BaMVO不可避免地引入了累计误差。

BaMVO说明时序多帧的反馈对动态环境下有效的运动物体检测与分割至关重要，而StaticFusion [98]认为有效的时序信息传播可以通过维护一个只包含场景中静态部分的三维地图来实现。三维数据可以有效地进行长时的三维时序信息融合，而数据融合有效地压缩了冗余信息，降低了整个系统的计算代价和内存开销。如图 11所示，通过同时检测运动物体并重建静态环境，staticFusion实现了动态环境下的鲁棒稠密的RGBD SLAM。点云数据被聚类到一个个聚类簇中，每个聚类簇再进行运动状态估计和刚体运动估计的联合求解，以获得每个

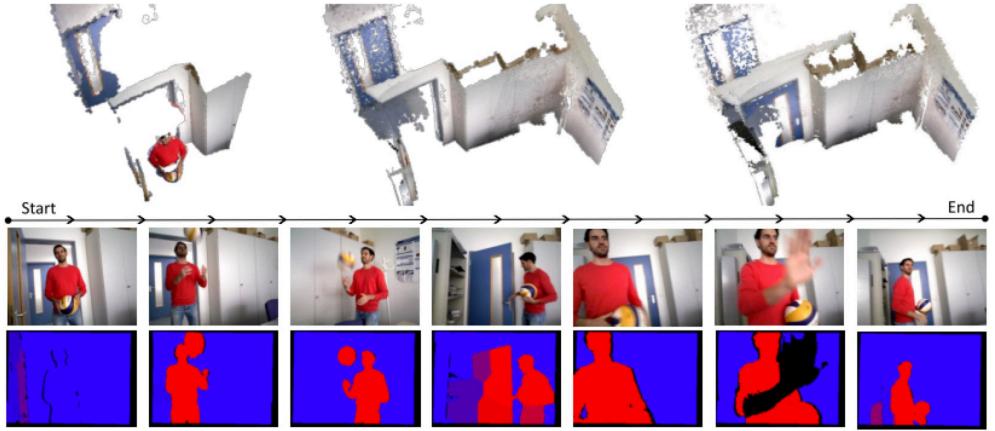


Figure 11: 三维静态地图提供了更加完整的先验，有助于提升运动分割和相机位姿估计这两个子问题的联合求解 [98]。

聚类簇属于静态或动态的概率。被判定为静态的聚类簇内数据会被融合到静态地图中，而被判定动态的聚类簇会进行场景流估计，以实现运动物体时序上的信息传递。由于采用了帧到模型（frame-to-model）的定位策略，相机位姿估计可以有效地消除由于累计误差带来的漂移。

DynaSLAM [99]提出了一种在线的算法，可以同时在单目、双目和RGBD相机设定下应对环境中的运动物体。整个系统建立在ORB-SLAM [100]的前端基础上，而核心出发点是通过建立可复用的三维地图进行更加精确的相机位姿估计。对于单目相机和双目相机，DynaSLAM采用卷积神经网络（CNN）进行像素级的物体分割，作为运动状态估计的先验。在RGBD相机的设定下，由于有着更加可靠的深度信息，DynaSLAM结合了基于稠密地图的几何约束和深度学习的算法进行运动物体检测。如图 12，通过语义信息与几何约束相结合的方式，DynaSLAM可以应对一些复杂的情形：一类是可能运动的物体在数据采集过程中处于静止状态的情形，比如停着的汽车或者坐着不动的人；另一类是没有运动先验的物体被错地发生运动的情形，比如人推着椅子行进。这种深度学习与几何相结合的方法可以更好地应对长时复杂多变的环境，在运动状态易变的情况下尽可能地剔除可能发生运动状态变化的数据，建立更稳定可靠的静态地图来

帮助定位。



Figure 12: 将基于稠密地图的几何约束（左）和基于深度学习的语义信息（中）相结合，可以更好地在复杂的动态环境下进行运动分割（右） [99]。

当然，在动态环境中构建静态地图依赖静态世界（static world）这一基本假设。在仓库、停车场和住宅这种环境的组成容易发生变化的场景下，环境变化将持续很长的时间，而这种变化可能有利于相机的定位。在极端情形下，可见范围内的静态地图占比很少或者信息量很小的时候，对动态物体运动的推断就对相机位姿估计起到了至关重要的作用。

6.2 静态背景和动态物体的同时建图

尽管动态物体对于相机位姿的求解会造成干扰，对动态物体的运动估计对于整个系统而言仍然至关重要。一些方法将静态背景与动态物体拆分开来，分别进行三维地图的构建。相比于只维护静态地图的方法，动态物体的运动和几何结构的推断可以带来更好的静态地图和更加可靠的运动分割结果。当然，该类方法的复杂性和计算代价也明显变高，因为算法不仅要通过识别静态背景以获得良好的位姿信息，还需对于每一个运动物体都维护独立的坐标系和地图以进行相应的位姿估计和数据融合。

Meta-room[101]提供了一种杂乱的办公环境的静态建模思路。Meta-room通过迭代地处理点云数据来拆分静态环境，识别环境中的变换元素，并更新地图。发生改变的动态元素将从地图中移除，而之前被遮挡区域会由新的观测数据填充。遮挡问题和传感器噪声在这一框架中需要额外注意。如图13所示，维

护一个好的meta-room可以帮助我们找到环境中椅子一类容易频繁移动的物体，构建出meta-object模型，更好地记录下环境随时间的动态变化情况。

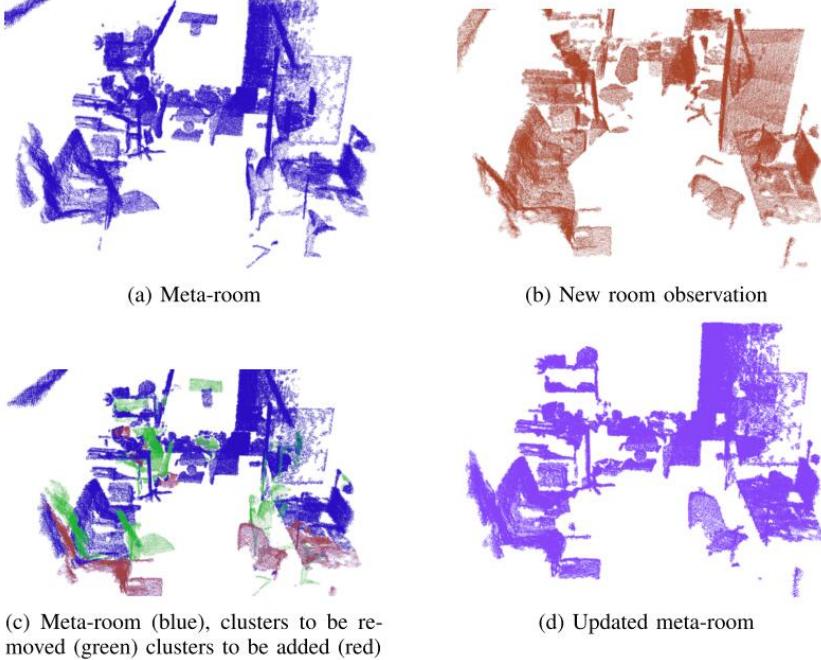


Figure 13: meta-room更新过程示意。

Caccamo等人 [102] 使用了自底向上 (bottom-up) 的特征分类的方式进行物体的识别与分割，如图 14a所示。算法首先利用第一帧数据初始化静态背景地图，然后将每一帧与静态地图进行配准。运动检测模块将特征分类并将输入数据分到维护的静态地图或物体模型。整个系统建立在基于关键帧的SLAM框架上维护了一个静态的地图，并对输入的每一帧进行特征计算与配准。根据配准之后的误差，将误差较高的部分聚合分离出来，从而判断出与相机运动不一致的动态物体，并维护该动态物体的地图，完成融合。

类似的，针对多个物体的同时跟踪与场景模型重建，Rünz和Agapito [104] 提出了Co-Fusion，可以处理多个不同物体的运动。该方法通过几何约束和语义信息将物体从场景中分割出来，然后对这些物体分别进行跟踪和重建。算法分割出物体后，可对每一部分的三维数据分别进行基于面元的数据融合，以处理

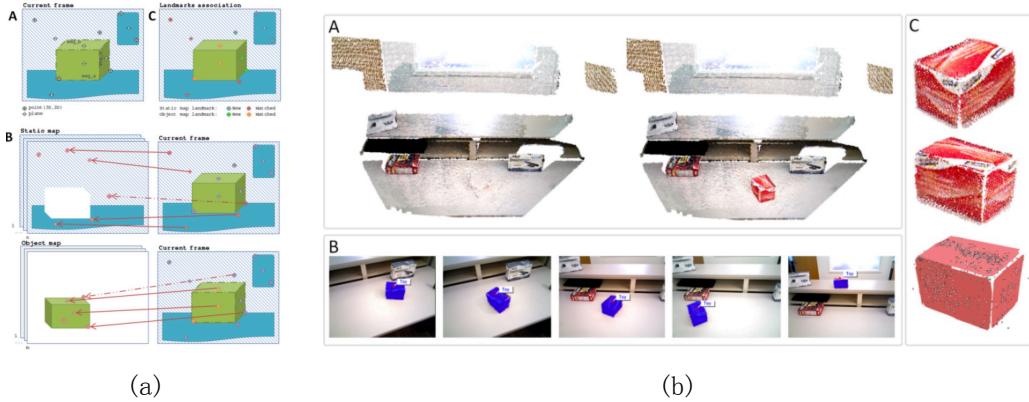


Figure 14: (a) 观测数据中的平面被组合起来，形成不同的分割区域。基于静态背景的配准可以去除由于运动造成的错误匹配。 (b) 真实场景下重建得到的完整场景模型。

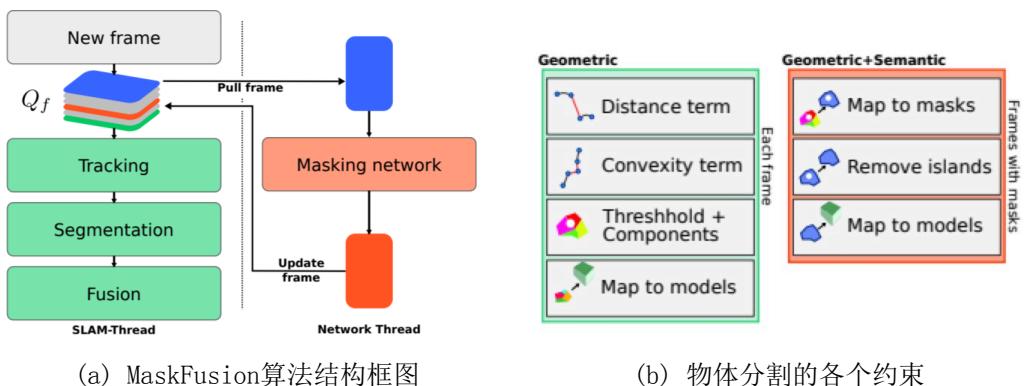


Figure 15: MaskFusion [103]利用二维图像的语义推断维护了场景中每个物体以及整个静态背景独立的三维模型。

不同物体的刚体运动，获得它们的三维模型。这种基于物体分割的动态物体重建会更适用于机器人相关的应用。算法可以对运动的物体获得较为准确的三维信息，从而使得机器人可以与环境进行更为丰富的交互。Rünz等人之后基于深度学习的方法提出了MaskFusion [105]，算法将Mask-RCNN [103]的分割结果与形状信息相结合，替代了原有的分割模块，从而在物体的分割边缘上能得到更好的表现，如图 15a所示。该类方法将语义信息与几何边缘信息相结合，从而获得更加完善的室内场景的物体分割结果。但从另一个角度来说，物体的语义信息依赖于模型的训练集。实验过程中的运动物体需要在训练集中出现过才能得到合理的分割结果，这也是使用语义作为分割标准的一个无法避免的弊端。

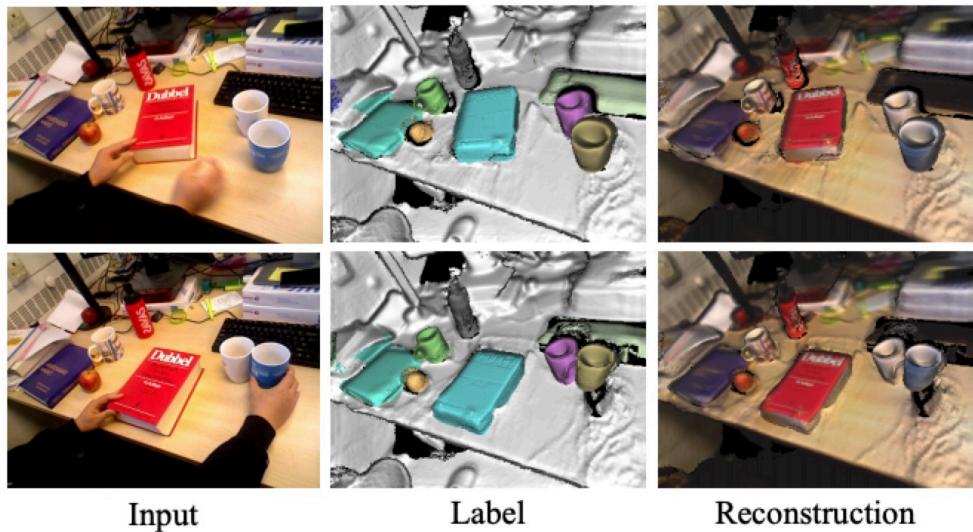


Figure 16: MidFusion算法 [106]构建了物体级 (object-level) 的稠密体素地图，可以应对移动的物体，并忽略场景中人的运动。

相较于使用语义信息进行自顶向下的分割，Xu等人 [106] 使用实例分割 (instance segmentation)，并通过几何和运动信息进行分割结果的优化，获得更好的分割边缘。如图16所示，三维地图中不仅仅只维护了几何和颜色信息，也保留了语义类别和运动状态的先验，以便为系统提供更加鲁棒的预测。对于分割后的物体，算法分别对这些物体进行物体姿态的估计、建图以及数据

融合。由于维护了基于体素结构的物体级的三维地图，算法对环境变化和未占用空间有着更强的感知能力，在室内场景的移动机器人领域有着更广泛的应用前景。MidFusion采用基于体素空间的截断符号距离函数（TSDF）作为隐式的空间表示方式。通过计算与最近表面的距离作为空间几何的参量，TSDF在未占用区域（free space）和未探索过的区域相比于点云或面元的表示形式有着更加完整的信息。Fehr等人[107]也利用TSDF表示三维空间，对静态地图和运动物体模型进行不断优化。这样的特性是之前基于点云、面元等空间表示形式所不具备的。



Figure 17: 基于TSDF的变化检测框架 [107]。

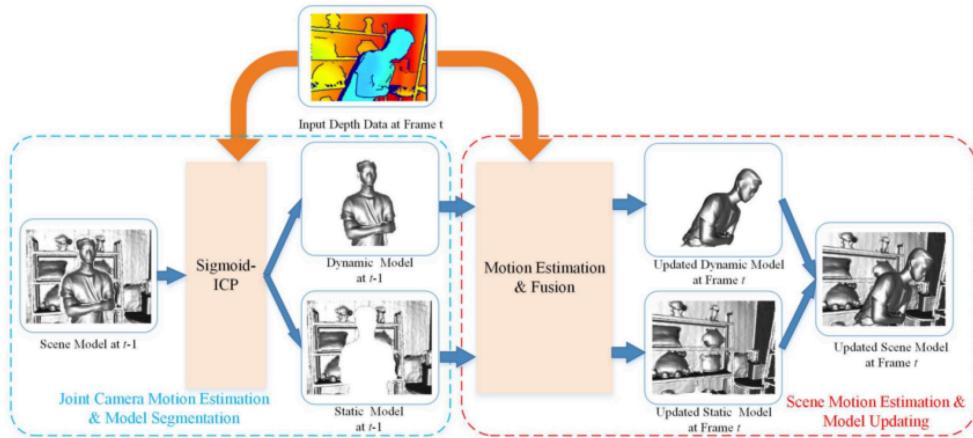


Figure 18: MixedFusion算法 [108]的流程图。蓝色虚线框内是运动分割和相机位姿估计的联合优化，而红色虚线框内是动态物体运动估计和三维数据融合。

相比于前面几种多刚体的环境，MixedFusion [108]能够重建出完全非刚体的人体模型。MixedFusion分别维护了每一时刻的运动物体模型和静态背景地图作为整个场景模型，并对输入的每一帧深度信息进行初步配准，区分出静态部分和动态物体。如图 18所示，场景的静态部分用于相机位姿的估计，而对于动态的部分作者则参考了Newcombe提出的DynamicFusion [109]，使用了图节点(graph Node based Motion Representation)将非刚体运动转化为以节点为控制点的多段刚体运动估计。通过运动物体的非刚体运动估计建立当前帧与模板模型(canonical model)的映射关系，来进行动态物体的数据融合。由于mixedFusion只用到了输入的深度信息而丢弃了RGB图像信息，在数据配准时容易受到深度弱纹理的影响，并且难以处理动态物体发生拓扑变化的情况。

总体而言，动态物体与静态场景的同时重建问题是一个较为困难的问题，即便输入为信息最为丰富的RGBD数据，目前也很难给出一个普适性的解决方案，均需要根据情况增加约束以使得问题可解。研究大多着眼于如何区分静态与动态部分，并使用适当的模型来描述动态物体的运动。尽管目前对于单一物体的简单运动可以恢复出较好的模型，但对于多物体复杂运动，考虑到相应的运算开销，常常难以获得较为鲁棒、准确的结果。另一方面，虽然运动部分的输入数据也被保留在地图中，这类方法仍然遵循着静态世界的基本假设，在动态变化较弱的环境下，同时维护静态地图和动态物体模型仍然会和只保留静态地图的方法遇到相似的挑战。

6.3 四维地图构建与长时定位

对于动态场景下的地图构建，无论将动态区域作为离群点予以剔除，或是分别维护静态背景和动态物体模型，都依赖于静态世界的假设。为了克服静态世界假设的局限性，一些研究人员致力于在一个统一的表示下建模环境的动态性，并最终达到长时甚至lifelong的建图的目的。早期，Murphy等人 [110]应用Rao-Blackwellized粒子滤波器(RBPF)来解决SLAM问题，并展示了其在动态场景下的理论可行性。如图19所示，他们的方法假设状态转移的概率与环境的当

前占据状态独立，并且只能在一个小尺度的局部范围内工作。之后，Avots等人 [111] 和 Petrovskaya 等人 [112] 针对门开合导致的环境变化这一问题对粒子滤波的方案进行了改进：前者用不断维护栅格地图，以得到确切的门的位置；后者则将门的开关状态从是或非的0/1状态改为基于门的打开角度的参数化模型。Stachniss 和 Burgard [113] 也使用 Rao-Blackwellized 粒子滤波器对局部的栅格地图进行聚类，来判断环境中可能存在的物体间的相互关系。Meyer 和 Delius [114] 维护了一些临时的局部地图，用来跟踪环境中未知的观测数据。传感器的位姿估计则是通过粒子滤波器中维护的全局 reference 地图和局部地图状态进行估计。然而，这项工作仍然依赖于静态地图进行位姿估计，只有在定位失败时才会利用局部地图。

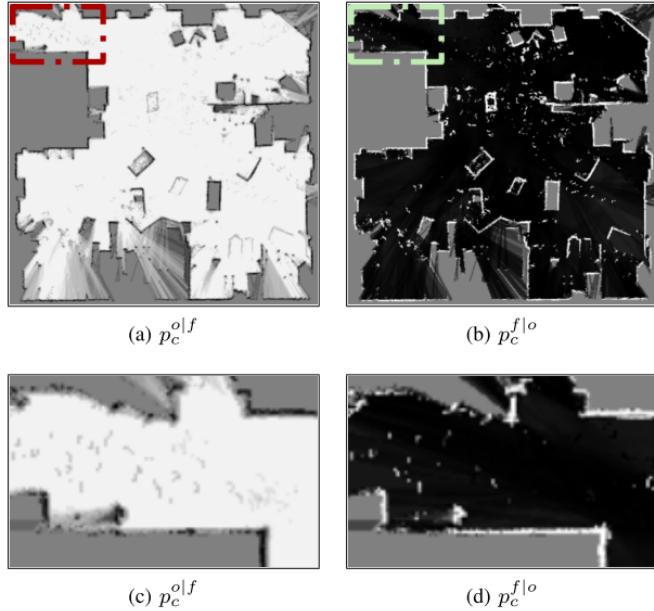


Figure 19: 基于粒子滤波的状态转移概率示意图。（颜色越深，概率越大）。

Churchill 和 Newman [115] 提出了关于 lifelong 建图的一种新视角。他们认为导航不需要一个全局的地图，而更多地利用拓扑信息。他们利用传感器的已有路径和对应的观测数据构建了环境的拓扑地图，并称之为“experience”。这种数据驱动的方式可以采用图像匹配进行定位，并通过不断累加的历史

路径信息记录环境中experience的变化情况。类似的，Tipaldi等人[116]利用这种拓扑地图的思路改进了已有的粒子滤波的方法，提出了一种新的适应环境变化的lifelong定位方法。它明确地考虑了环境的动态变化，且能够区分：表现出高动态行为的物体，例如汽车和人；可以移动并改变摆放位置的物体，例如箱子、架子或门；以及静止不移动的物体，例如墙壁。该方法在二维占据栅格上用一个隐马尔科夫模型（HMM）描述环境的动态变化，并通过Rao-Blackwellized粒子滤波器高效地迭代更新位姿和地图的状态。

除了基于RBPF的滤波方法，占据栅格也是一种常见的空间表示形式。Chen等人[117]以及后来的Brechtel等人[118]提出并拓展了传统的占据栅格（occupancy grid）框架，使之包含了对动态物体的建模，并用贝叶斯滤波的方式对其进行更新，如图20所示。在这个视角下，他们认为占据栅格的占据概率由环境中的物体决定。当物体发生运动，其对应的占据栅格也会发生相应的运动。因此，在该框架中，他们需要自始自终追踪每一个栅格的运动。这种从物体出发的建图思路也被成为以物体为中心的环境建模（object-centered reconstruction）。

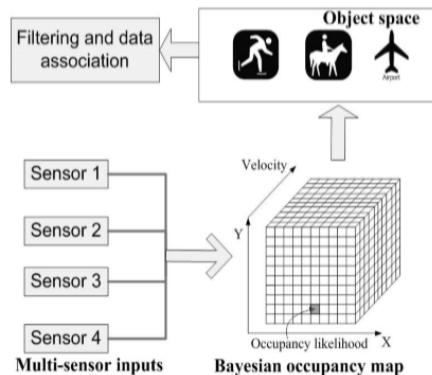


Figure 20: 基于贝叶斯滤波器的占据栅格状态更新过程。

为了在lifelong情形下建模变化的环境，Konolige等人 [119]提出了一套侧重于地图可视化的框架。在该框架中，局部栅格地图可以随环境的变化进行更新、添加和删减。Kretzschmar等人[120]也给出了类似的想法，他们利用一种

基于信息论的图修剪策略 (graph pruning) 进行图压缩操作。这两种方法更侧重于解决长时建图带来的尺度问题。Walcott-Bryant等人[121]则提出了一个名为Dynamic Pose Graph (DPG) 的图模型，针对长时低动态环境进行定位和建图。

与object-centered reconstruction相对应的是从地图角度出发的环境建模 (map-centered reconstruction)。这种地图驱动的环境建模方式不过多考虑引起环境变换的原因，而只记录场景在时序上的变化情况。Schindler等人 [122]自底向上地 (bottom-up) 地定义了一种时变的场景几何表示方式。在传统的SfM基础上，他们给每个相机加上了对应的时间戳，并为地图中的每一点记录了存在的时间范围。利用概率时序模型，可以从如图21建立的“4D城市”几何模型中推断出观测数据对应的确切年份。

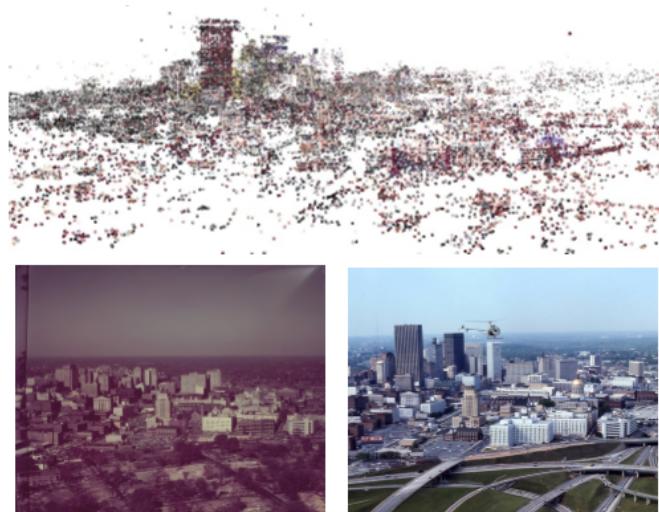
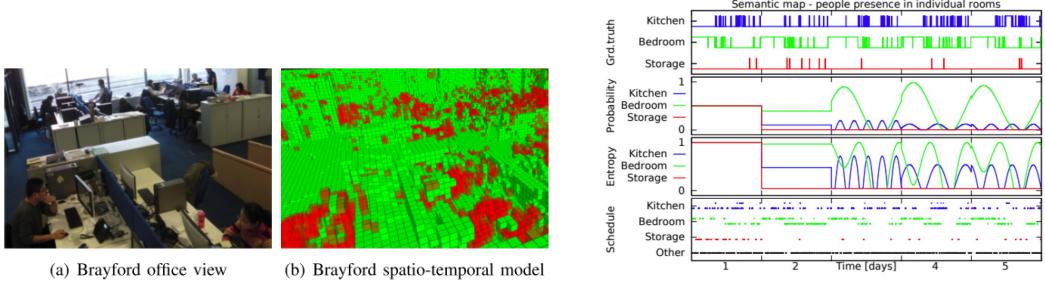


Figure 21: “4D城市”示意图（1956年–1971年）。

之后，Krajiník等人[124]利用频谱建模环境的时空动态信息。这种频谱的空间表示形式可以高效地识别、分析和记忆环境的周期性变化，对于长时的机器人相关应用来说，这种周期性的信息使得模型有着很强的预测能力。文章通过傅里叶逆变换 (inverse Fourier transform) 来识别长时变化环境下的频谱参数，并用来预测环境的局部状态。傅里叶频谱中最显著的频率组成部分对应



(a) 时变的占据栅格示意图（绿色为静态栅格，红色为周期性变化栅格）。
(b) 时空信息熵随时间和场景动态发生改变。

Figure 22: Lifelong 设定下移动机器人在变化环境的探索示意 [123]。

着环境中最明显的周期性变化。虽然上述方法适用于移动机器人中使用的大多数环境模型，但由于其依赖于传统的快速傅立叶变换（FFT）方法，因此需要对环境进行定期频繁的观测。这意味着机器人需要首先频繁地访问环境来构建动态模型，然后再利用学习到的动态模型开展导航、路径规划的任务。也就是说，虽然机器人可以创建适合长期操作的动态模型，但它不能在线地维护这些模型。Krajiník等人[123]在此基础上提出了一种新的思路来解决lifelong设定下移动机器人建模环境的时空变化问题。空间的时空变化模型仍然通过频谱进行表示，而路径规划通过基于信息增益的蒙特卡洛方法实现。机器人利用概率模型和时空信息熵（spatio-temporal entropy）来规划观测，并学习场景的动态变化。该方法增加了一个额外的时间维度，使得机器人不仅可以构建动态模型并不断更新，还可以观察和理解环境自身的变化情况。这种自主智能的环境探索能力能在机器人运动过程中不断优化对环境动态变化的理解。

总而言之，关于如何将静态和动态场景置于一个统一优雅的参数空间下这一问题，研究人员针对时空特性进行了广泛的研究。而在面对实际应用时，基于静态世界假设的三维地图表示仍然是目前的主流技术路线。

References

- [1] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *Intl. J. of Computer Vision*, 9(2):137 – 154, 1992.
- [2] Luca Zappella, Alessio Del Bue, Xavier Lladó, and Joaquim Salvi. Joint estimation of segmentation and structure from motion. *Computer Vision and Image Understanding*, 117(2):113 – 129, 2013.
- [3] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Markov localization for mobile robots in dynamic environments. *Journal of artificial intelligence research*, 11:391 – 427, 1999.
- [4] Nolang Fanani, Matthias Ochs, Henry Bradler, and Rudolf Mester. Keypoint trajectory estimation using propagation based tracking. In *IEEE Intelligent Vehicles Symposium (IV)*, 2016.
- [5] João Costeira and Takeo Kanade. A multi-body factorization method for motion analysis. In *Intl. Conf. on Computer Vision (ICCV)*, 1995.
- [6] Michal Irani. Multi-frame correspondence estimation using subspace constraints. *Intl. J. of Computer Vision*, 48(3):173 – 194, 2002.
- [7] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. Subspace video stabilization. *ACM Trans. Graphics*, 30(1):1 – 10, 2011.

- [8] Liu Feng, Yuzhen Niu, and Hailin Jin. Joint subspace stabilization for stereoscopic video. In Intl. Conf. on Computer Vision (ICCV), 2013.
- [9] Peter Sturm and Bill Triggs. A factorization based algorithm for multi-image projective structure and motion. In European Conf. on Computer Vision (ECCV), pages 709 - 720, 1996.
- [10] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. Intl. J. of Computer Vision, 29(3):159 - 179, 1998.
- [11] Ken-ichi Kanatani. Motion segmentation by subspace separation and model selection. In Intl. Conf. on Computer Vision (ICCV), volume 2, pages 586 - 591. IEEE, 2001.
- [12] Naoyuki Ichimura. Motion segmentation based on factorization method and discriminant criterion. In Intl. Conf. on Computer Vision (ICCV), 1999.
- [13] N. Ostu. A threshold selection method from gray-histogram. 9(1): 62 - 66, 2007.
- [14] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.
- [15] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. Intl. J. of Computer Vision, 67(2):233 - 246, 2006.

- [16] Marco Paladini, Alessio Del Bue, Marko Stosic, Marija Dodig, João M. F. Xavier, and Lourdes De Agapito. Factorization for non-rigid and articulated structure using metric projections. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.
- [17] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.
- [18] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Multi-body non-rigid structure-from-motion. 2016.
- [19] Maxime Derome, Aurelien Plyer, Martial Sanfourche, and Guy Le Besnerais. Moving object detection in real-time using stereo from a mobile platform. Unmanned Systems, 3(04):253 - 266, 2015.
- [20] Jens Klappstein, Tobi Vaudrey, Clemens Rabe, Andreas Wedel, and Reinhard Klette. Moving object segmentation using optical flow and depth information. Lecture Notes in Computer Science, 5414:611 - 623, 2008.
- [21] Abhijit Kundu, K. Madhava Krishna, and Jayanthi Sivaswamy. Moving object detection by multi-view geometric techniques from a single camera mounted robot. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009.
- [22] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6): 381 - 395, 1981.

- [23] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. *Kybernetes*, 30(9/10):1865 – 1872, 2008.
- [24] Eagle S. Jones and Stefano Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *International Journal of Robotics Research*, 30(4):407 – 430, 2011.
- [25] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research*, 34(3):314 – 334, 2014.
- [26] Michael D. Breitenstein, Student Member, Fabian Reichlin, Bastian Leibe, Esther Kollermeier, and Luc Van Gool. Online multi-person trackingby-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1820 – 1833, 2010.
- [27] K. H. Lee, J. N. Hwang, Greg Okpal, and James Pitton. Driving recorder based on-road pedestrian tracking using visual slam and constrained multiple-kernel. In *IEEE International Conference on Intelligent Transportation Systems*, 2014.
- [28] S. Wangsiripitak and D. W. Murray. Avoiding moving outliers in visual slam by tracking moving objects. In *IEEE International Conference on Robotics and Automation*, 2009.
- [29] Chris Harris and Carl Stennett. Rapid – a video rate object tracker. In tracker. In *Br. Mach. Vis. Conf.*, 1990.

- [30] Yin-Tien Wang, Ming-Chun Lin, and Rung-Chi Ju. Visual slam and moving-object detection for a small-size humanoid robot. *International Journal of Advanced Robotic Systems*, 7(2):13, 2010.
- [31] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.
- [32] Falak Chhaya, Dinesh Reddy, Sarthak Upadhyay, Vishesh Chari, Zee-shan Zia, and K. Madhava. Monocular reconstruction of vehicles: Combining slam with shape priors. In *IEEE International Conference on Robotics and Automation*, 2016.
- [33] Kuan Hui Lee, Jenq Neng Hwang, Greg Okopal, and James Pitton. Ground-moving-platform-based human tracking using visual slam and constrained multiple kernels. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3602 – 3612, 2016.
- [34] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for background subtraction. *Pattern Recognition*, 2017.
- [35] M. Piccardi. Background subtraction techniques: a review. In *IEEE International Conference on Systems*, 2005.
- [36] D. Zhang and P. Li. Visual odometry in dynamical scenes. *Sensors and Transducers*, 147(12):78 – 86, 2012.
- [37] Vidal René, Ma Yi, and Sastry Shankar. Generalized principal component analysis (gPCA). *IEEE Trans Pattern Anal Mach Intell*, 27(12):1945 – 1959, 2005.

- [38] Davide Migliore, Roberto Rigamonti, Daniele Marzorati, Matteo Matteucci, and Domenico G. Sorrenti. Use a single camera for simultaneous localization and mapping with mobile object tracking in dynamic environments. 2009.
- [39] S. Heuel and W. Forstner. Matching, reconstructing and grouping 3d lines from multiple views using uncertain projective geometry. 2001.
- [40] Chieh Chih Wang, Charles Thorpe, Martial Hebert, Sebastian Thrun, and Hugh Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. International Journal of Robotics Research, 26(9) :889 – 916, 2007.
- [41] Zou Danping and Tan Ping. Coslam: collaborative visual slam in dynamic environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(2) :354 – 366, 2013.
- [42] Tan Wei, Haomin Liu, Zilong Dong, Guofeng Zhang, and Hujun Bao. Robust monocular slam in dynamic environments. In IEEE International Symposium on Mixed and Augmented Reality, 2013.
- [43] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. Artificial Intelligence, 17(1 - 3) :185 – 203, 1980.
- [44] Pablo F. Alcantarilla, Jose J. Yebes, Javier Almazan, and Luis M. Bergasa. On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In IEEE International Conference on Robotics and Automation, 2012.

- [45] Maxime Derome, Aurelien Plyer, Martial Sanfourche, and Guy Le Besnerais. Real-time mobile object detection using stereo. In International Conference on Control Automation Robotics and Vision, 2014.
- [46] Konrad Schindler and David Suter. Two-view multibody structure-and-motion with outliers through model selection. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(6):983 – 995, 2006.
- [47] O Faugeras, PHS Torr, T Kanade, N Hollinghurst, J Lasenby, M Sabin, and A Fitzgibbon. Geometric motion segmentation and model selection-discussion. *PHILOS T ROY SOC A*, 356:1338 – 1340, 1998.
- [48] Konrad Schindler and David Suter. Two-view multibody structure-and-motion with outliers. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume 2, pages 643 – 648. IEEE, 2005.
- [49] Konrad Schindler, U James, and Hanzi Wang. Perspective n-view multibody structure-and-motion through model selection. In European Conf. on Computer Vision (ECCV), pages 606 – 619. Springer, 2006.
- [50] Konrad Schindler, David Suter, and Hanzi Wang. A model-selection framework for multibody structure-and-motion of image sequences. *Intl. J. of Computer Vision*, 79(2):159 – 177, 2008.
- [51] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *IEEE Trans. Pattern Anal. Machine Intell.*, 32(6):1134 – 1141, 2010.

- [52] Ninad Thakoor, Jean Gao, and Venkat Devarajan. Multibody structure-and-motion segmentation by branch-and-bound model selection. *IEEE Trans. Image Processing*, 19(6):1393 – 1402, 2010.
- [53] Reza Sabzevari and Davide Scaramuzza. Monocular simultaneous multi-body motion segmentation and reconstruction from perspective views. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 23 – 30. IEEE, 2014.
- [54] Reza Sabzevari and Davide Scaramuzza. Multi-body motion estimation from monocular vehicle-mounted cameras. *IEEE Trans. Robotics*, 32(3):638 – 651, 2016.
- [55] Davide Scaramuzza. 1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. *International Journal of Computer Vision*, 95(1):74 – 85, 2011.
- [56] Davide Scaramuzza, Friedrich Fraundorfer, and Roland Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point ransac. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2009.
- [57] Diego Ortin and José María Martínez Montiel. Indoor robot motion based on monocular images. *Robotica*, 19(3):331 – 342, 2001.
- [58] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159 – 179, 1998.
- [59] C William Gear. Multibody grouping from motion images. *International Journal of Computer Vision*, 29(2):133 – 150, 1998.

- [60] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gPCA). *IEEE Trans. Pattern Anal. Machine Intell.*, 27(12):1945 – 1959, 2005.
- [61] Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In European Conf. on Computer Vision (ECCV), pages 94 – 106. Springer, 2006.
- [62] Alvina Goh and René Vidal. Segmenting motions of different types by unsupervised manifold clustering. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1 – 6. IEEE, 2007.
- [63] Lawrence K Saul and Sam T Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, 4(Jun):119 – 155, 2003.
- [64] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 2790 – 2797. IEEE, 2009.
- [65] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765 – 2781, 2013.
- [66] Shankar Rao, Roberto Tron, Rene Vidal, and Yi Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Trans. Pattern Anal. Machine Intell.*, 32(10):1832 – 1845, 2009.

- [67] Congyuan Yang, Daniel Robinson, and René Vidal. Sparse subspace clustering with missing entries. In Intl. Conf. on Machine Learning (ICML), pages 2463 – 2472, 2015.
- [68] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 35(1):171 – 184, 2012.
- [69] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In ICML, volume 1, page 8, 2010.
- [70] René Vidal. Online clustering of moving hyperplanes. In Advances in Neural Information Processing Systems (NIPS), pages 1433 – 1440, 2007.
- [71] Teng Zhang, Arthur Szlam, and Gilad Lerman. Median k-flats for hybrid linear modeling with many outliers. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 234 – 241. IEEE, 2009.
- [72] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52 – 68, 2011.
- [73] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146 – 157, 1997.
- [74] Shai Avidan and Amnon Shashua. Trajectory triangulation of lines: Reconstruction of a 3d point moving along a line from a monocular image sequence. In Proceedings. 1999 IEEE Computer Society

Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), volume 2, pages 62 - 66. IEEE, 1999.

- [75] Shai Avidan and Amnon Shashua. Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):348 - 357, 2000.
- [76] Amnon Shashua, Shai Avidan, and Michael Werman. Trajectory triangulation over conic section. In *Intl. Conf. on Computer Vision (ICCV)*, volume 1, pages 330 - 336. IEEE, 1999.
- [77] Jeremy Yirmeyahu Kaminski and Mina Teicher. General trajectory triangulation. In *European Conference on Computer Vision*, pages 823 - 836. Springer, 2002.
- [78] Jeremy Yirmeyahu Kaminski and Mina Teicher. A general framework for trajectory triangulation. *Journal of Mathematical Imaging and Vision*, 21(1-2):27 - 41, 2004.
- [79] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *European Conf. on Computer Vision (ECCV)*, pages 158 - 171. Springer, 2010.
- [80] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d trajectory reconstruction under perspective projection. *Intl. J. of Computer Vision*, 115(2):115 - 135, 2015.
- [81] Vincent Aidala and Sherry Hammel. Utilization of modified polar coordinates for bearings-only tracking. *IEEE Transactions on Automatic Control*, 28(3):283 - 294, 1983.

- [82] J-P Le Cadre and Olivier Trémois. Bearings-only tracking for maneuvering sources. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):179 – 193, 1998.
- [83] Abhijit Kundu, K Madhava Krishna, and CV Jawahar. Realtime multi-body visual slam with a smoothly moving monocular camera. In *Intl. Conf. on Computer Vision (ICCV)*, pages 2080 – 2087. IEEE, 2011.
- [84] Abhijit Kundu, K Madhava Krishna, and CV Jawahar. Realtime motion segmentation based multibody visual slam. In *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, pages 251 – 258. ACM, 2010.
- [85] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2017 – 2025, 2015.
- [86] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 173 – 180, 2017.
- [87] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv*, 2017.
- [88] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1983 – 1992, 2018.

- [89] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 1890 - 1899, 2019.
- [90] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 8071 - 8081, 2019.
- [91] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), volume 11, pages 127 - 136, 2011.
- [92] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. ACM Trans. Graphics, 32(6):169, 2013.
- [93] T Whelan, M Kaess, MF Fallon, H Johannsson, JJ Leonard, and JBM Kintinous. Kintinous: Spatially extended kinectfusion. In RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, 2012.
- [94] Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Volumetric 3d mapping in real-time on a cpu. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 2021 - 2028. IEEE, 2014.

- [95] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. *Robotics: Science and Systems (RSS)*, 2015.
- [96] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *International Conference on 3D Vision (3DV)*, pages 1 - 8. IEEE, 2013.
- [97] Deok-Hwa Kim and Jong-Hwan Kim. Effective background model-based rgb-d dense visual odometry in a dynamic environment. *IEEE Trans. Robotics*, 32(6) :1565 - 1573, 2016.
- [98] Raluca Scona, Mariano Jaimez, Yvan R Petillot, Maurice Fallon, and Daniel Cremers. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1 - 9. IEEE, 2018.
- [99] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4) :4076 - 4083, 2018.
- [100] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robotics*, 33(5) :1255 - 1262, 2017.
- [101] Rares Ambrus, Nils Bore, John Folkesson, and Patric Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2014.

- [102] Sergio Caccamo, Esra Ataer-Cansizoglu, and Yuichi Taguchi. Joint 3d reconstruction of a static scene and moving objects. In International Conference on 3D Vision (3DV), pages 677 - 685, 2017.
- [103] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Intl. Conf. on Computer Vision (ICCV), pages 2961 - 2969, 2017.
- [104] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 4471 - 4478, 2017.
- [105] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), pages 10 - 20, 2018.
- [106] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 5231 - 5237, 2019.
- [107] Ivan Dryanovski Ji § urgen Sturm Igor Gilitschenski Roland Siegwart Marius Fehr, Fadri Furrer and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2017.
- [108] Hao Zhang and Feng Xu. Mixedfusion: Real-time reconstruction of an indoor scene with dynamic objects. IEEE Trans. on visualization and computer graphics, 24(12):3137 - 3146, 2017.

- [109] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 343 - 352, 2015.
- [110] K. Murphy. Bayesian map learning in dynamic environments. In Advances in Neural Information Processing Systems (NIPS), 1999.
- [111] Dzintars Avots, Edward Lim, Romain Thibaux, and Sebastian Thrun. A probabilistic technique for simultaneous localization and door state estimation with mobile robots in dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2002.
- [112] Anna Petrovskaya and Andrew Y. Ng. Probabilistic mobile manipulation in dynamic environments, with application to opening doors. In International Joint Conference on Artificial Intelligence, 2007.
- [113] Cyrill Stachniss and Wolfram Burgard. Mobile robot mapping and localization in non-static environments. In National Conference on Artificial Intelligence, 2005.
- [114] D. Meyer-Delius, J. Hess, G. Grisetti, and W. Burgard. Temporary maps for robust localization in semi. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2010.
- [115] Winston Churchill and Paul Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2012.

- [116] Gian Diego Tipaldi, Daniel Meyer-Delius, and Wolfram Burgard. Lifelong Localization in Changing Environments. 2013.
- [117] C. Chen, C. Tay, C. Laugier, and K. Mekhnacha. Dynamic environment modeling with gridmap: A multiple-object tracking application. In International Conference on Control, 2006.
- [118] Sebastian Brechtel, Tobias Gindele, and Rudiger Dillmann. Recursive importance sampling for efficient grid-based occupancy filtering in dynamic environments. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2010.
- [119] Kurt Konolige and James Bowman. Towards lifelong visual maps. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2009.
- [120] Henrik Kretzschmar and Cyrill Stachniss. Information-theoretic compression of pose graphs for laser-based slam. Intl. J. of Robotics Research, 31(11):1219 – 1230, 2012.
- [121] A. Walcott-Bryant, M. Kaess, H. Johannsson, and J. J. Leonard. Dynamic pose graph slam: Long-term mapping in low dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2012.
- [122] Grant Schindler and Frank Dellaert. Probabilistic temporal inference on reconstructed 3d scenes. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.
- [123] Tomas Krajnik, Joao M. Santos, and Tom Duckett. Life-long spatio-temporal exploration of dynamic environments. In European Conference on Mobile Robots, 2015.

- [124] Tomščák, Krajská, Jaime Pulido Fentanes, Grzegorz Cielniak, Christian Dondrup, and Tom Duckett. Spectral analysis for long-term robotic mapping. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2014.