



# 动态场景下的SLAM相关技术总结报告

作者 Author \*

August 25, 2019

注释测试:

ZK:[\*这句话用来test comment command\*]

WX:[\*这句话用来test comment command\*]

RK:[\*这句话用来test comment command\*]

PJ:[\*这句话用来test comment command\*]

ZS:[\*这句话用来test comment command\*]

SK:[\*这句话用来test comment command\*]

## 1 动态环境下的SLAM系统

### 1.1 基于运动分割的SLAM技术

---

\*作者介绍 Brief introduction

动态分割（又称为移动物体检测/分割[?, ?, ?]）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）[?]和特定的距离度量方法（如桑普森距离[?])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份，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[?, ?]。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（现存方法的总结参见表1）。

### 1.1.1 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[?, ?]。Wangsiripitak等人[?]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[?]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[?]假设移动中的物体上的SURF特征描述子集是已知的并存在数据库中。通过比较在特征检测步骤得到的描述子，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[?]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[?, ?]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，背景初始化类似于background subtraction技术，设定一个背景模型[?, ?]。Zhang等人[?]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[?]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。

### 1.1.2 几何约束

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来分割静态和动态特征。这些约束可以从极线、三角化、基本矩阵估计或重投影误差的等式中得出。

Kundu等人[?]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[?]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[?]，将测量的不确定性加入他们检查不同视线关系的过程。最后通过统计假设测试PJ: [\*statistical hypothesis test\*]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[?]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[?]将先前帧的特征投影到当前帧上，测量这些跟踪到特征的距离。通过这个重投影的距离分类静态和动态的特征。Tan等人[?]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。

### 1.1.3 光流

光流的定义了两个连续的图像间，图案亮度的表面运动。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[?]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[?]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[?, ?]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。

### 1.1.4 相机运动约束

一般的SfM和视觉SLAM通过8点的均值[?]或5点法[?]计算相机的运动。这种普通的相机运动估计并没有考虑任何相机移动方式的假设。另外一种方式是假设相机根据特定参数提供的外在信息(如车轮里程计的信息)估计相机位姿。加入这种相机运动约束，通过匹配特征点是否符合相机运动的约束，分类静态特征。

Scaramuzza[?]提出利用轮式车辆的非完整性约束计算相机运动的方法。他假设相机运动是平面或圆形的，以此建模相机运动。由此约束，相机运动可以被参数化为1自由度并且可由1点法计算[?]。同样地，Sabzevari等人[?]也利用阿克曼转向几何提供的轮式车辆约束来计算相机运动。满足估计的相机运动的特征点会被认为是静态特征，其他的特征点则被认为是动态特征。

### 1.1.5 运动分割的深度学习

在基于特征的运动分割中，我们知道可以利用光流分割移动中的物体。Dosovitskiy等人[?]展示了光流的估计是可以有监督学习获得的。他们提出了两个不同结构的卷积神经网络(CNN)来预测光流。其中一个网络(FlowNetS)将两个连续帧的图像作为输入，而另一个网络(FlowNetC)将两个不同的CNN合并来比较两个特征地图。Ilg等人[?]进一步提出了FlowNet 2.0，将FlowNetS和FlowNetC放入一个更深的网络，同时加入并行的网络来处理小的位移。实验表明FlowNet 2.0的结果可以与最新的方法并肩。Mayer等人[?]延伸出了用双目图像计算场景流的方法。这个光流可以被放大更深的网络，检测运动特征[?]。尽管因为网络不是显式地为了解决运动分割问题而设计，很难说同样的一个网络能够用来分割移动的物体并且提供运动的边界范围，但是这些运动特征对于动作识别非常有用[?, ?]。

Lin和Wang[?]构建了一个可以显式地在图像空间分割移动物体的网络，他们使用Reconstruction Independent Component Analysis(RICA)自编码器[?, ?]来学习时空特征，但是由于时空特征无法学习运动的三维几何信息，因此仍然使用几何特征帮助分割任务。几何特征和时空特征都放进循环神经网络(RNNs)来计算最终的运动分割。Fragkiadaki等人[?]将RGB图像和光流的置信度PJ:[\*objectness score?]\*。类似于AlexNet[?]的两个并行网络用来处理RGB图像和光流，接着放入回归网络生成运动估计PJ:[\*motion proposal\*]。Valipour等人[?]提出了Recurrent Fully Convolutional Network(R-FCN)，使用时序数据在线地在图像序列分割前景物体。Fully Convolutional

Network (FCN) [?] 用于学习空间特征和生成像素的稠密预测，但是 Gated Recurrent Unit (GRU) 用于在反卷积之前对时间特征进行建模。

## 1.2 基于运动物体跟踪的SLAM技术

### 1.2.1 深度学习方法用于动态物体分割

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) [1]。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 [2]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 [3]，能够从三维点云中预先定义好的 $n$ 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。

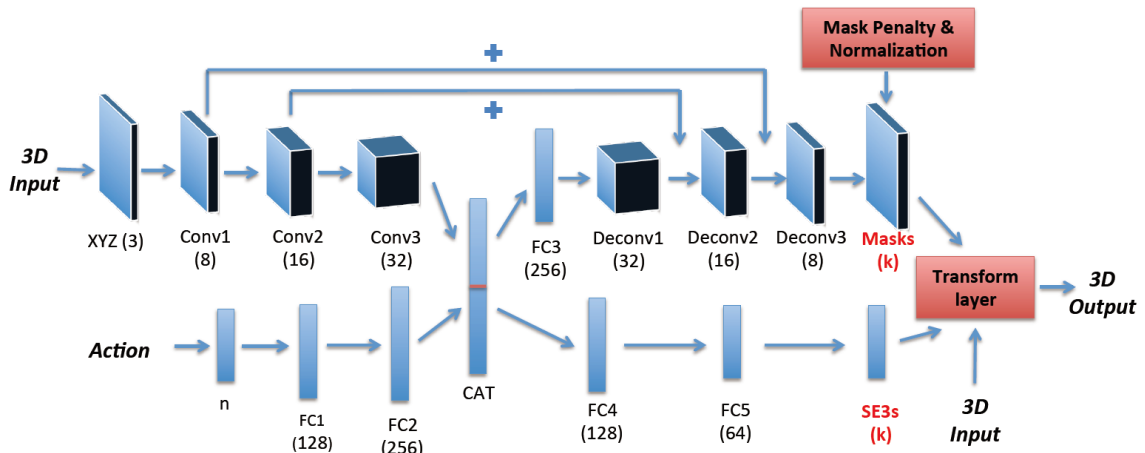


Figure 1: SE3-Net的网络结构图

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 [4]。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。

与SfM-Net类似，Yin等人提出的GeoNet [5]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 1.2.1，GeoNet利用前向和反



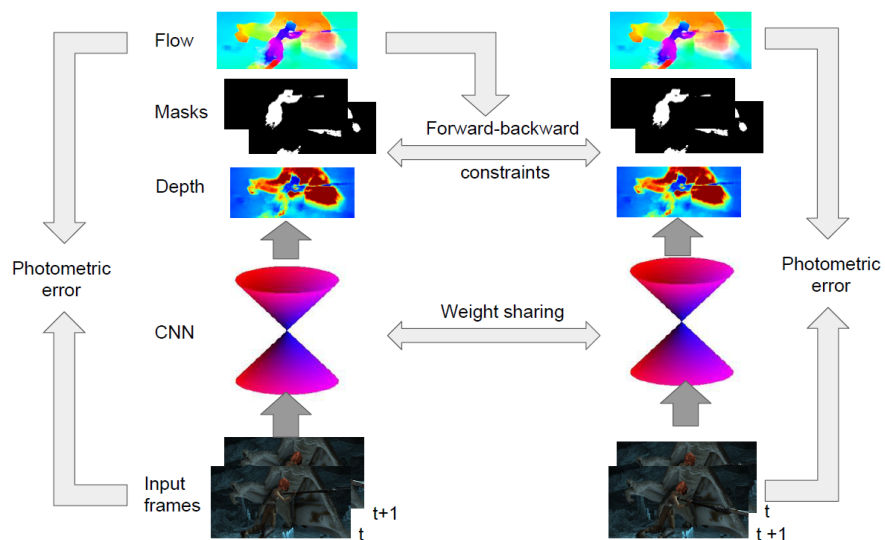


Figure 2: SfM-Net的流程图

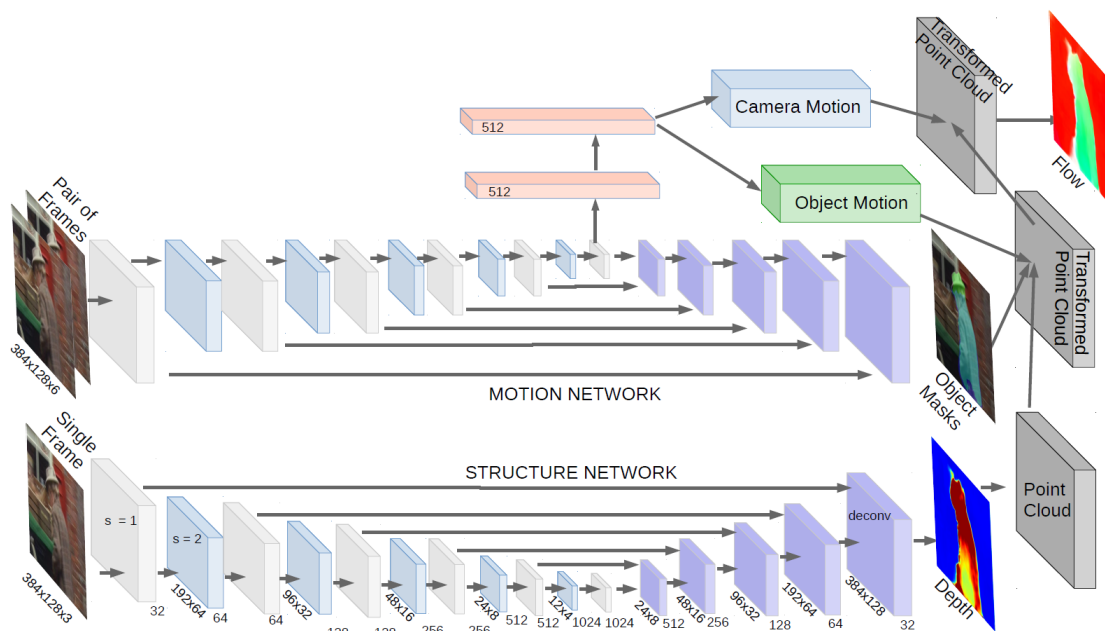


Figure 3: SfM-Net的网络结构图

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。

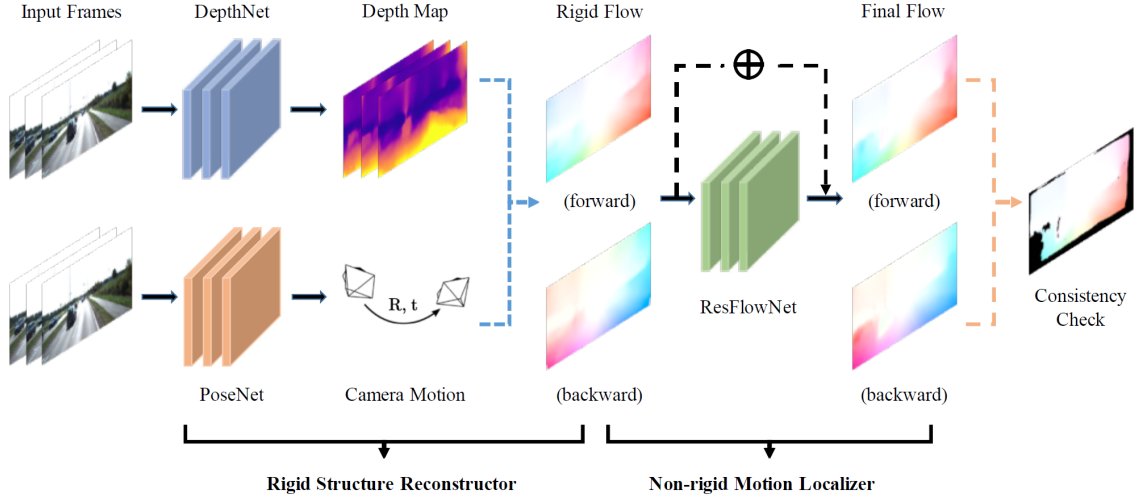


Figure 4: GeoNet的流程图

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 [6]，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 1.2.1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。

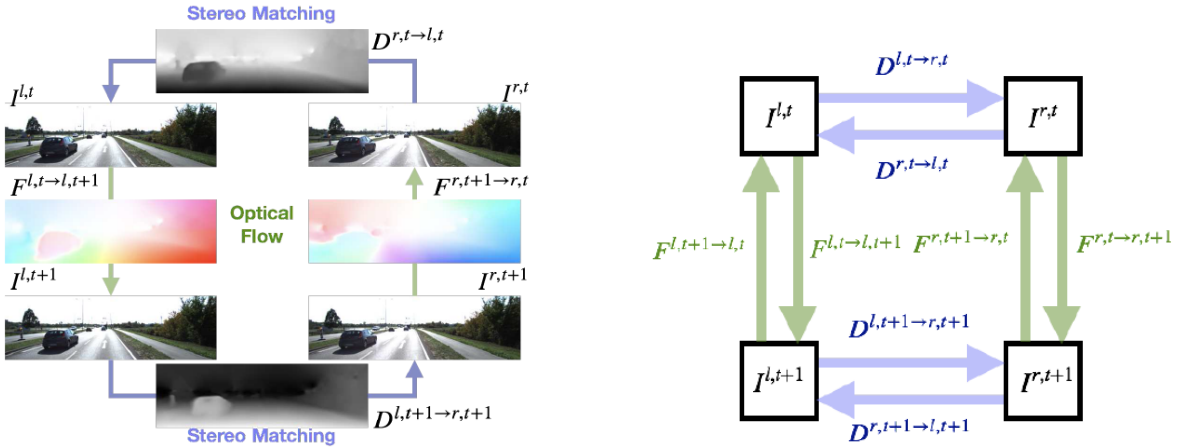


Figure 5: 左边为 [6] 的流程图，右边为该方法利用的各种帧间几何约束

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 [7]。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。



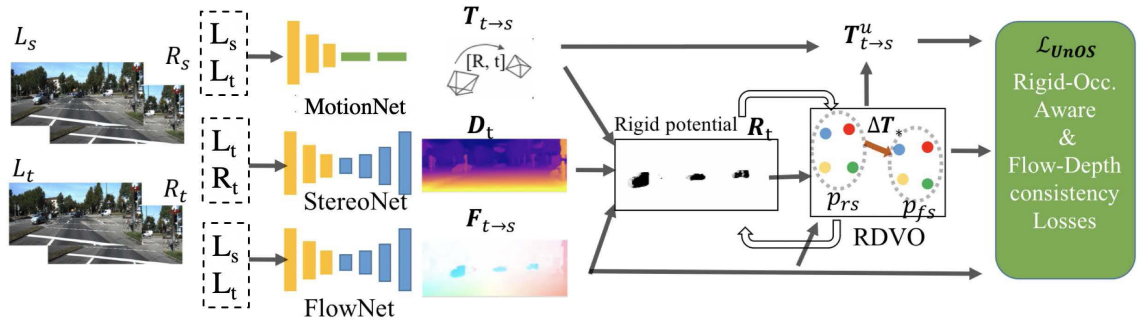


Figure 6: UnOS的流程图

### 1.3 非刚体和多刚体运动下的SfM技术

## 2 长时变化环境下的地图更新

### 2.1 动态环境下静态部分地图的构建

motion removal

对于存在运动物体的场景，动态的观测数据违反了配准能量函数的基本假设，需要被视为离群点从地图中剔除。在视觉SLAM应用于稠密建图的这一方向，一个主要的策略就是只维护静态部分的地图。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，以及如何补全地图中因为运动物体遮挡而未被获取的数据。通过维护一个高质量、可服用的静态地图，相机位姿估计可以在静态世界的假设中鲁棒地进行估计，而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。

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 [12] 发现，马尔科夫定位 (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，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。

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 [8] 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 [9]、回环调整 [10] 以及计算效率 [11] 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。

ElasticFusion [13] 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，[14] 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。

BaMV0 [15] 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，BaMV0不可避免地引入了累计误差。

BaMV0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 [16] 认为有效地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。通过三维数据融合，这种长时的时序信息不会带来额外的计算代价。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 [17] 提出了一种在线的算法，可以同时单目、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 [18] 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，DynaSLAM则结合了多目立体视觉和深度学习的算法进行运动物体检测。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误地发生运动的情形，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，建立更稳定可靠的静态地图来帮助定位。

然而，在动态环境中构建静态地图依赖静态世界这一假设。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。

## 2.2 静态背景和动态物体的同时重建

尽管动态物体对于相机位姿的求解会造成干扰，但在某些应用情形而言，动态物体的三维信息也是我们感兴趣的部分。在这种情形下，算法就需要完成静态背景与动态物体的同时重建。一般而言该类方法会更为复杂与困难。算法除了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配准和融合。

对于静态背景和动态物体的同时重建问题，其核心过程可分为两大部分：静态与动态的分割，以及每一部分的数据融合。良好的分割结果可提升相机位姿的求解精度，从而使得重建结果的准确度提升；而对于不同动态物体使用合适的融合方式则影响了所关注物体重建的结果。

目前，对于动态物体与静态场景的分割问题，研究者往往使用RGBD传感器作为算法的输入，以获得更为优良准确的单帧三维信息。Zhang和Xu [19]维护了每一个时刻的场景模型(Scene Model)，以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。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 [20]，使用了一种基于图节点的运动表示结构(Graph Node-based Motion Representation)来进行动态物体的融合，以完成对非刚性运动物体（如人体、窗帘等）的模型融合。不过，由于算法本身对于动态物体采用的是DynamicFusion的融合框架，该方法只用到了输入的深度信息而丢弃了RGB图像信息，并且难以处理动态物体发生拓扑变化的情况。Caccamo等人 [21]使用了自底向上的特征分类的方式进行物体的识别与分割。算法维护了一个静态的地图，并对于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。该算法假设场景中只有一个刚性运动的物体，使用场景较为受限。

对于多个物体的跟踪与重建，Rünz和Agapito [22]提出了Co-Fusion，用于处理多个不同物体的运动。该方法通过运动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 [23]，算法将Mask-RCNN [24]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 2.2所示。该类方法将语义信息与物体形状相结合，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。

相较于使用语义信息进行自顶向下的分割，Xu等人 [25]使用几何和运动信息进行场景物体的分割。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、建图以及融合。该方法对于每个物体都维护了一个基于体素的子图(Object-level dynamic volumetric map)，从而相应的定位和融合算法可以在物体层面上增量的进行。

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将静态与动态部分分割开，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。

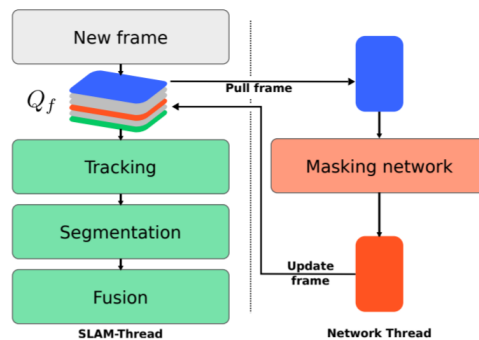


Figure 7: MaskFusion算法结构框图。

### 2.3 四维地图构建与长时定位

对于动态场景建模，先前的一部分方法完全将动态区域作为离群点予以剔除，另一部分方法则同时维护静态和动态地图，以提供一个更好的环境静态地图和更可靠的动态区域检测。但无论哪种途径，都依赖于静态世界的假设，这使得这些方法在部署到不断变化的环境或是动态性较低的环境中时效果不佳。

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当中建模环境的动态性，并最终达到进行lifelong建图的目的。Chen等人[26]以及后来的Brechtel等人[27]提出并拓展了传统的占据网格的框架，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。在这个视角下，针对动态性他们建议了以物体为中心的表示，他们认为占据网格中格子的占据概率由环境中的物体决定，当物体发生运动，其对应的占据网格也会发生相应的运动。因此，在该框架中，他们需要自始至终追踪每一个网格的运动。与该思想相反的，采用以地图为中心的方法也可以对环境中的动态进行建模。

Schindler和Dellaert等人[28]利用自下而上的启发式方法，将从SfM管道中的点观测分组为建筑假设和概率时间模型来推断建筑物存在的时间间隔，建立了一个“4D城市”模型。

Yang和Wang[29]建议用一个“可能性”网格来同时表示静态区域和动态区域。一对对偶传感器模型被用来在移动机器人定位中判别静态及动态物体，然而，他们的工作假定机器人的位置是已知的，具有一定的精度来进行计算以及更新地图，所以该方法并不适合于全局定位问题。

之后，Saarinen[30]等人提出用一系列独立的马尔科夫链去建模整个环境，将状态之间的转换参数建模为两个泊松过程并在线学习这些参数，采用基于近因加权的方法处理非平稳单元的动力学问题。而同样的，该方法也无法普适地应对真实环境下的不同的动态场景及物体。

Murphy[31]等人建议应用Rao-Blackwellized粒子滤波器来解决SLAM问题并理论上展示了其在动态场景下的可行性。但他们的方法假设了状态转换的概率与环境的当前状态独立并且给定了一个先验，且只能在一个小尺度的环境下工作。之后，Avots等人[32]，Petrovskaya[33]等人分别提出对它的改进，前者用Rao-Blackwellized粒子滤波器来估计机器人的姿态和环境中的门的状态，他们使用一个参考占用网格来表示环境，而非他们的状态（其中门的位置是已知的）；后者与前者相似，但将门的开关状态这一二元模型改为一个参数化模型（门的打开角度）。而Stachniss和Burgard[34]也使用Rao-Blackwellized粒子滤波器对聚类后的局部网格图确定的一组可能的环境配置来对机器人进行定位，并从该集合中估计环境的配置。Meyer和Delius[35]跟踪那些由环境中使用临时局部地图的离群对象引起的观测结果，然后用上述粒子滤波器来估计机器人的姿势，该滤波器不仅依赖于这些临时地图，也依赖于环境的参考地图，然而，这项工作仍然依赖于全局定位的静态映射，只有在位置跟踪失败时才会创建临时映射。

另外，对于lifelong的动态环境建图，Konolige[36]提出了一个有趣的方法，该方法主要侧重于可视化地图，并提供了一个框架，在该框架中，可以随着时间的推移更新本地地图（视图），并在环境配置更改时添加/删除新的本地地图。Kretzschmar[37]等人也给出了类似的想法，他们利用一种有效的信息论图形修剪策略进行图形压缩。该方法可用于偏倚最近的观察结果，以获得与前者工作的类似的表现。然而，这两种方法主要集中在长期操作中出现的可伸缩性问题上，而不是环境随时间变化的动态方面。从这个想法出发，Walcott-Bryant[38]等人提出了一个名为Dynamic Pose Graph（DPG）的局部表示来建模长时下低动态环境的SLAM问题。

Churchill和Newman[39]提出了关于lifelong建图的另一个视角。他们认为导航不需要一个全局参考框架，并介绍了“经验”的概念，即具有相对测量信息的机器人路径。“经验”可以通过基于外观的数据关联方法连接在一起，随着时间的推移而变化的地方由一组不同的“经验”表示。Tipaldi等人[40]改进并综合了上述基于粒子滤波的方法，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，可以移动并改变配置的物体，例如箱子、架子或门，以及静止不移动的物体，例如墙壁。该方法在二维网格上用一个隐马尔科夫模型描述空间的占据和它的动态性，并通过EM算法学习其参数，联合估计机器人姿态以及全局定位中的环境状态，然后应用一个Rao-Blackwellized粒子滤波器（其中机器人姿态为被采样部分滤波器，网格占据状态为分解的解析部分），同时通过考虑相关马尔科夫链的混合时间来建立一种基于局部地图表示的地图管理方法以能够最小化内存需求，并以合理的概率方式来忘记变化。

之后，Krajník等人[41]提出在光谱域中表示环境动力学，并将其用去图像特征以改进定位，之



后也陆续有研究者将该方法应用于占用网格以减少内存需求、应用于拓扑图以改进路径规划。

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，因此存在一个主要缺陷，即需要对环境进行定期和定期的观测。这意味着机器人的活动必须分为一个学习阶段，当它经常访问各个位置建立其动态环境模型时，以及当它使用其模型执行有用任务时的部署阶段。这一划分意味着，虽然机器人可以创建更适合长期操作的动态模型，但它不能维护这些模型。因此，机器人不适应那些不存在学习阶段的动力学问题，这会导致其效率随着时间的推移而降低。Krajník等人[42]又提出了一种lifelong移动机器人时空动态环境探测的新思路，该方法假设世界处于不断变化的状态，这将为探索空间增加一个额外的时间维度，使探索任务成为一个永无止境的数据收集过程。为了创建和维护一个动态环境的时空模型，机器人不仅要确定在哪里，还要确定何时进行观察。我们将信息论探索应用于世界表征，将环境状态的不确定性建模为时间的概率函数，从而解决这一问题。

另外，Ambrus等人[43]提出了一种新的方法来重新创建杂乱的办公环境的静态结构，他们将其定义为“meta-room”，它基于一个配备了rgb-d 深度摄像头的自主机器人在长时间内收集到的多个观测结果进行实验。该方法通过识别从一个观测点到下一个观测点的变化，移除动态元素，同时添加先前被遮挡的对象，以尽可能准确地重建底层静态结构，直接与点簇一起工作。构建meta-room的过程是迭代的，它被设计为在可用时合并新数据，并对环境变化具有鲁棒性。meta-room的最新估计用于区分和提取动态物体群与观测结果。该方法之后也被应用在一些导航机器人平台来得到更好，更细节的物体模型。

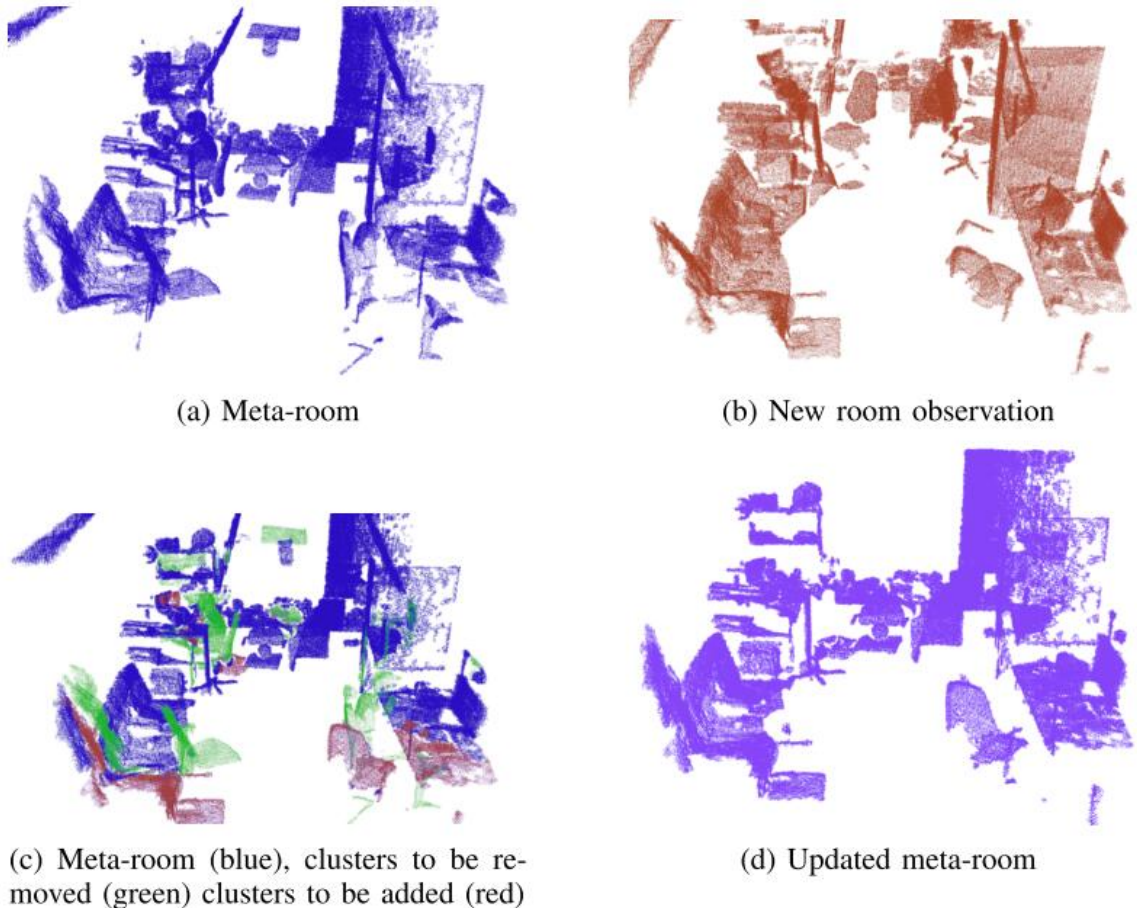


Figure 8: meta-room更新过程示意。

上面提到的Krajník和Ambrus的工作重点都是使用变化检测算法的结果来分析变化的时空行为，而有的研究人员只关注观测结果之间的变化。Fehr等人[44]提出了一种新的基于扩展截断有符号距离函数（TSDF）的动态场景下的三维重建算法，该算法能够在场景中同时获得动态对象的三维重建

的同时，对静态地图进行连续的细化。这是一个具有挑战性的问题，因为地图更新是递增的，并且常常是不完整的。以前的工作通常在点云、曲面或地图上执行变化检测，这些点云、曲面或地图无法区分未探测空间和空白空间。相比之下，该方法基于TSDF的表示自然包含了这些信息，从而使其能够更有力地解决场景差异问题。

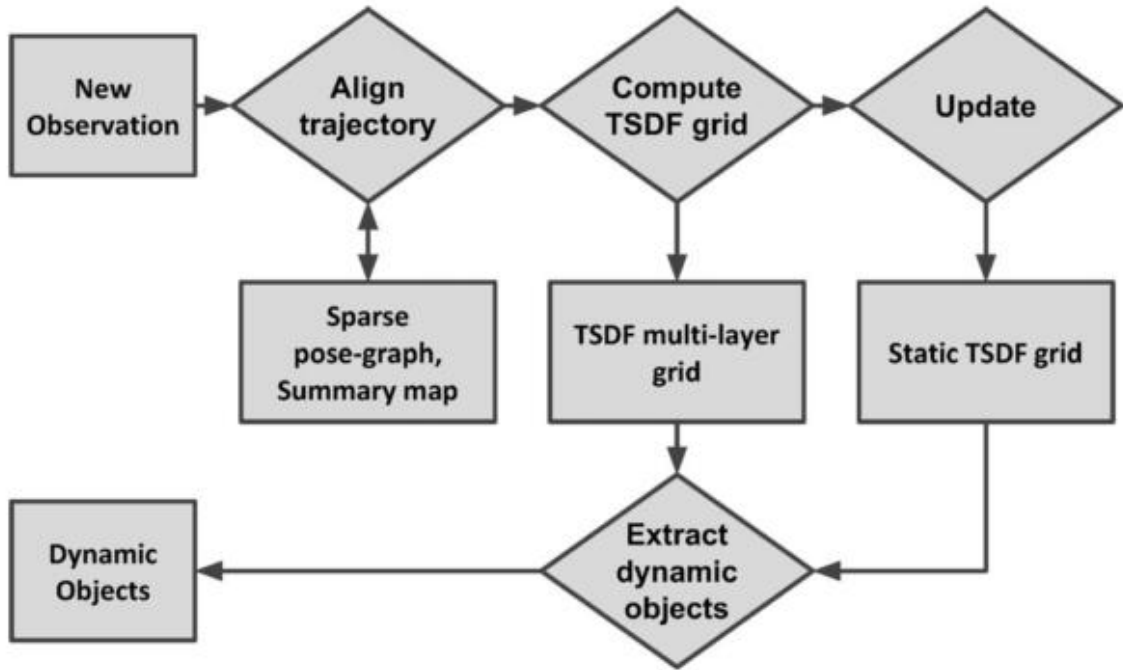


Figure 9: 基于TSDF的变化检测框架。

总而言之，关于如何将静态和动态场景置于一个统一优美的空间表示形式下，早期研究人员在基于滤波的框架下做了很多的探索，而在面对实际问题时，大部分在真实场景下拥有鲁棒效果的方法却仍然需要沿着前面几个章节所述的技术路线进行。

## References

- [1] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems (NIPS), pages 2017 – 2025, 2015.
- [2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. 2015.
- [3] 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.
- [4] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. arXiv, 2017.
- [5] 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.

- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] T Whelan, M Kaess, MF Fallon, H Johannsson, JJ Leonard, and JBM Kintinuous. Kintinuous: Spatially extended kinectfusion. In RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, 2012.
- [11] 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.
- [12] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Markov localization for mobile robots in dynamic environments. Journal of artificial intelligence research, 11:391 – 427, 1999.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.

- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Intl. Conf. on Computer Vision (ICCV), pages 2961 – 2969, 2017.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] Grant Schindler and Frank Dellaert. Probabilistic temporal inference on reconstructed 3d scenes. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.
- [29] Shao Wen Yang and Chieh Chih Wang. Feasibility grids for localization and mapping in crowded urban scenes. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2011.
- [30] Jari Saarinen, Henrik Andreasson, and Achim J. Lilienthal. Independent markov chain occupancy grid maps for representation of dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2012.
- [31] K. Murphy. Bayesian map learning in dynamic environments. In Advances in Neural Information Processing Systems (NIPS), 1999.
- [32] 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.
- [33] 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.
- [34] Cyrill Stachniss and Wolfram Burgard. Mobile robot mapping and localization in non-static environments. In National Conference on Artificial Intelligence, 2005.

- [35] 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.
- [36] Kurt Konolige and James Bowman. Towards lifelong visual maps. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2009.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] Gian Diego Tipaldi, Daniel Meyer-Delius, and Wolfram Burgard. Lifelong Localization in Changing Environments. 2013.
- [41] Tomáš Krajník, 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.
- [42] Tomas Krajník, Joao M. Santos, and Tom Duckett. Life-long spatio-temporal exploration of dynamic environments. In European Conference on Mobile Robots, 2015.
- [43] 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.
- [44] Ivan Dryanovski, Jürgen 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.