Abstract

摘要

动态场景下的SLAM相关技术总结报告

作者 Author *

August 24, 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

1.2 基于运动物体跟踪的SLAM技术

1.3 基于运动物体跟踪的SLAM技术

1.3.1 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) [4]。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 [1]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 [2]，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。

*作者介绍 Brief introduction

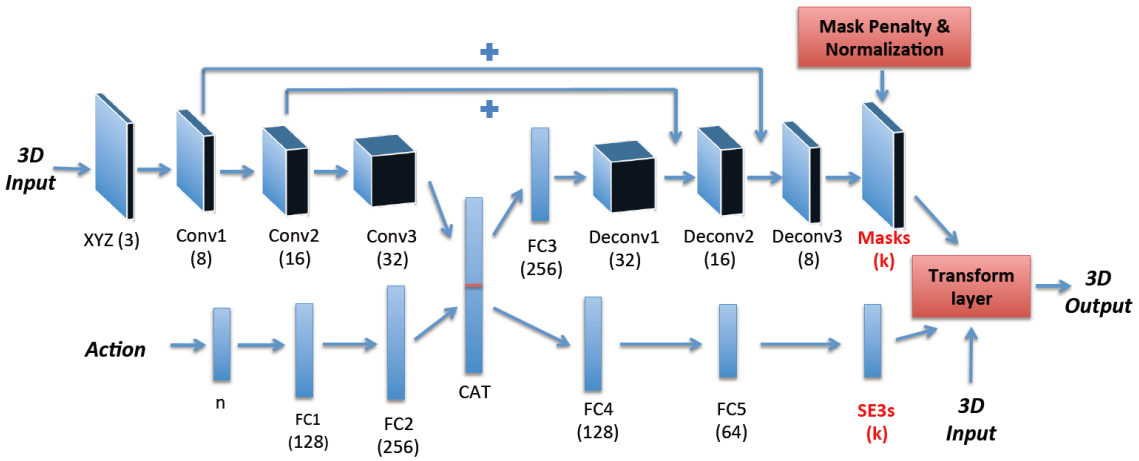

Figure 1: SE3-Net的网络结构图

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 [12]。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。

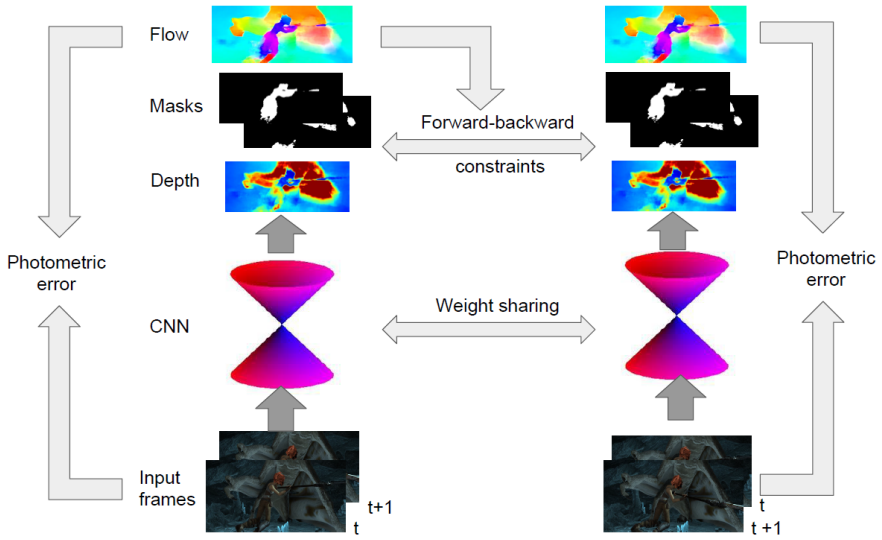

Figure 2: SfM-Net的流程图

与SfM-Net类似，Yin等人提出的GeoNet [16]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 1.3.1，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

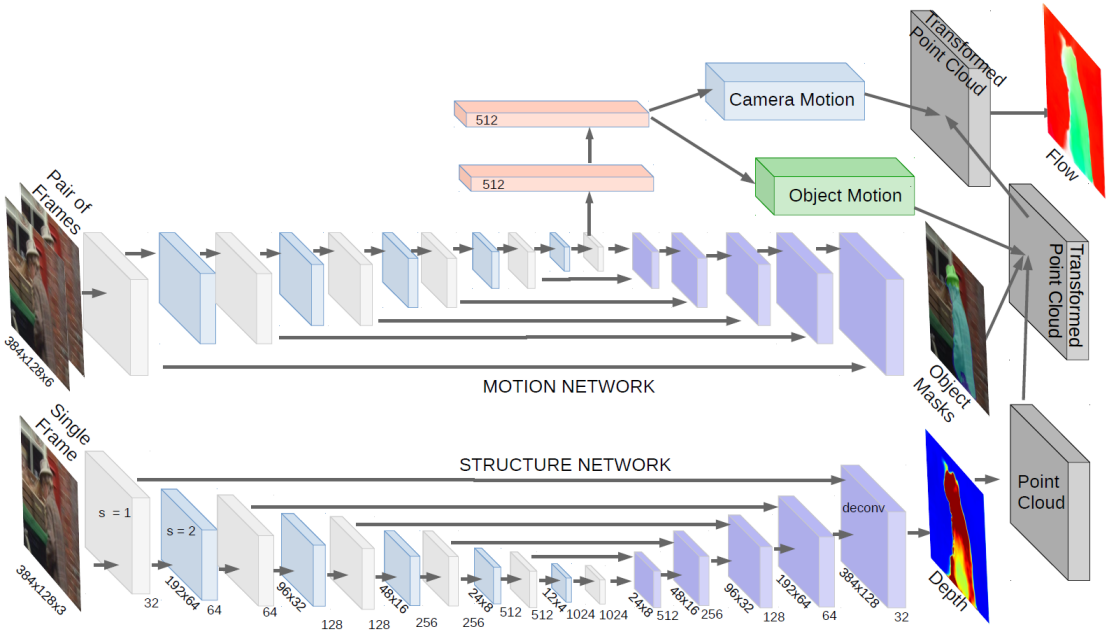

Figure 3: SfM-Net的网络结构图

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。

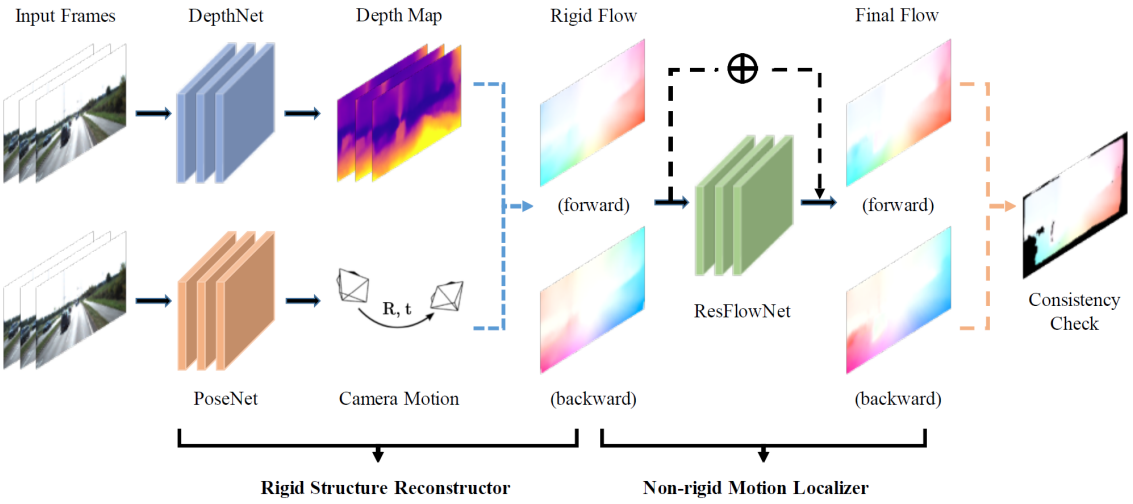

Figure 4: GeoNet的流程图

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 [7]，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 1.3.1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。

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 [13]。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

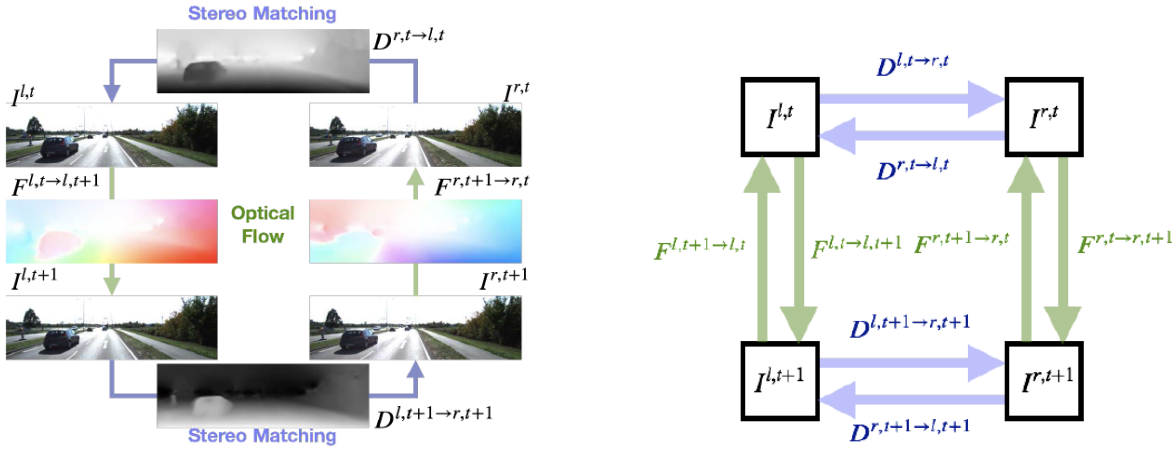

Figure 5: 左边为 [7]的流程图，右边为该方法利用的各种帧间几何约束

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。

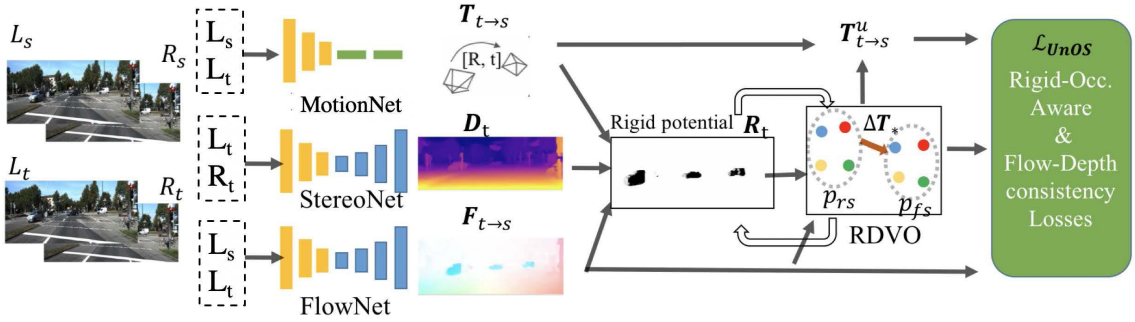

Figure 6: UnOS的流程图

1.4 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

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 [14]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。而对于存在运动物体的场景，动态的观测数据需要被视为离群点从地图中剔除，以避免被定位模块使用。

ElasticFusion [15]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，[5] 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。

BaMVO [6] 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（dense visual odometry）。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。StaticFusion [10]证明维护一个只包含场景中静态部分的三维地图是一种有效地时序信息传播的方式。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。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。

2.2 静态背景和动态物体的同时重建

2.3 四维地图构建与长时定位

[3]

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. 2015.

[2] A. Byravan and D. Fox. Se3-nets: Learning rigid body motion using deep neural networks. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 173 – 180, 2017.

[3] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 2758 – 2766, 2015.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems (NIPS), pages 2017 – 2025, 2015.

[5] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In International Conference on 3D Vision (3DV), pages 1 – 8. IEEE, 2013.

[6] D.-H. Kim and J.-H. Kim. Effective background model-based rgb-d dense visual odometry in a dynamic environment. IEEE Trans. Robotics, 32(6):1565 – 1573, 2016.

[7] H.-Y. Lai, Y.-H. Tsai, and W.-C. Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 1890 – 1899, 2019.

[8] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. 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] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. ACM Trans. Graphics, 32(6):169, 2013.

[10] R. Scona, M. Jaimez, Y. R. Petillot, M. Fallon, and D. 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.

[11] F. Steinbrücker, J. Sturm, and D. 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] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. arXiv, 2017.

[13] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. 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.

[14] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. Kintinuous. Kintinuous: Spatially extended kinectfusion. In RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, 2012.

[15] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison. Elastic-fusion: Dense slam without a pose graph. Robotics: Science and Systems (RSS), 2015.

[16] Z. Yin and J. 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.