

# Dense Monocular SLAM using Learned Outlier Mask

## 论文情况

---

- 标题: A Front-End for Dense Monocular SLAM using a Learned Outlier Mask Prior
- 作者: Yihao Zhang, John J. Leonard
- 会议: ICRA, 2021
- 源码: 未开源

## 1 Introduction

---

单眼 SLAM 需要解决的内容:

- 需要仅根据拍摄的照片序列协同解决图像中点的深度的计算和相机位姿的计算。

其中存在的问题:

- 由于深度和位姿带来的第一帧中像素点与第二帧中像素点的数据关联问题。

一种可行的解决方法:

- 使用 CNN 为点的深度提供一个合理的推测, 目前已有很多的方法[1, 2, 3, 4, 5, 6]使用基于学习的深度图作为点的深度的初始值。

本文解决方法:

- 提出一种稠密的 CNN 协助的 SLAM 前端。其关注于无后端干预的位姿估计。
- 当 CNN 训练目标与在线位姿估计目标对齐时, 从训练过程中学习到的离群点掩码

可以与经典概率模型一起使用，通过处理静态遮挡和动态对象引起的离群点，来提高姿态估计精度。

- 随着语义重建对视觉系统变得越来越重要，本文的 SLAM 前端也使用前向传播处理语义分割图像，以提高质量。

## 2 Preliminaries

---

### 2.1 Photometric Consistency Loss

直接法和深度的无监督学习法的核心是光度一致性，即如果一帧中的像素和另一帧中的像素对应于相同的 3D 点，它们在图像中应该具有相同的强度。因此，本文中网络的光度一致性损失定义为：

$$L_{pho}(I', I) = \frac{1}{|V|} \sum_{p \in \Omega(I)} W(p) \|I'(\Pi(p)) - I(p)\|_m \quad (1)$$

$$\Pi(p) = \Pi(p, D(p), R, t) = \begin{bmatrix} u' \\ v' \end{bmatrix} \quad (2)$$

$$d' \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = K[R \quad t]homo \left( K^{-1}D(p) \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \right) \quad (3)$$

其中：

- $I$  和  $I'$  为相邻的两帧， $p = [u, v]^T \in \Omega(I)$  为在图像空间  $\Omega(I)$  的像素， $W(p)$  为在像素  $p$  的损失权重， $V$  为投影到图像内的有效像素集合， $\|\cdot\|_m$  为 m 范式， $\Pi(p)$  为投影函数（对应到 (3) 式中，有  $D(p)$  为像素  $p$  的深度， $R, t$  为  $I$  到  $I'$  的旋转矩阵和平移向量， $K$  为  $3 \times 3$  的相机内参矩阵， $homo$  表示转换为齐次坐标， $d'$  表示在  $I'$  中的投影深度）。
- 1 范式用于 CNN 训练寻找 outliers，2 范式用于直接的 VO 方法。
- $\Pi(p)$  返回非整数值，不能用于像素索引，因此使用到双线性插值。双线性插值还保留了光度损失的可微性，用于基于梯度的优化方法。
- 投影像素  $\Pi(p)$  可能落在图像边界之外，需要排除这些无效像素，并按照有效像素

数量  $|V|$  对光度损失进行归一化。(1) 中的光度损失可以很容易地扩展到 CNN 训练中经常使用的 rgb 通道。

## 2.2 Learning Depth and Outlier Mask from Video

---

[7, 8] 中讨论了如何从视频中学习深度和自我运动，本文着重讨论离群掩码。为图像中的每个像素预测  $[0, 1]$  之间的离群掩码值，这个值在训练过程中作为(1)式中的权重  $W(p)$ 。直观地说，如果一个像素对应于动态物体或遮挡表面上的 3D 点，作为异常值，其光度一致性损失不应被计算， $W(p)$  将趋于零。为了避免在训练过程中学习到的所有像素的离群掩码为 0 的平凡情况，使用具有 all-ones 掩码的交叉熵损失作为正则化。

## 3 Method

---

算法 1 给出了系统概述。该系统是一个 SLAM 前端，最终输出是里程计、深度图和关键帧的语义分割图像。

---

## Algorithm 1 System Overview

---

Input: image sequence  $\{I_i\}$ ,  $i = 1, \dots, n$

Output: camera poses  $\{T_i\}$  for all frames, depth maps  $\{D_j\}_{kf}$  and semantically segmented images  $\{S_j\}_{kf}$  for keyframes,  $\{j\} \subseteq \{1, \dots, n\}$

**function** INSERT A KEYFRAME ( $I$ )

Predict the depth map, outlier mask, and semantic segmentation for  $I$ .

Initialize the probability model parameters  $\alpha$ ,  $\beta$ ,  $\mu$ , and  $\sigma$  for each pixel in  $I$  using the predicted depth map and outlier mask.

**end function**

INSERT A KEYFRAME ( $I_1$ )

**for**  $i = 2$  to  $n$  **do**

Estimate  $T_i$  for  $I_i$  against the keyframe.

Update the depth map of the keyframe given  $T_i$ .

Update  $\alpha$ ,  $\beta$ ,  $\mu$ , and  $\sigma$  for each pixel in the keyframe.

**if** the keyframe criteria are satisfied **then**

    INSERT A KEYFRAME ( $I_i$ )

    Fuse the semantic class probabilities between the last keyframe and this keyframe.

**end if**

**end for**

---

### 3.1 Keyframe Insertion

第一帧以及满足一定条件的帧作为关键帧，关键帧包含一系列属性（位姿、彩色图、灰度图、语义分割图、深度图、离群掩码、概率模型参数）。在插入关键帧后，彩色图被输入到 CNN 模型中预测语义分割图、深度图和离群掩码。离群掩码和深度图用于计算之后概率模型的先验分布。

假设时间  $k$  获得一个像素的深度测量  $d_k$ ，则可以建模此深度的分布为：

$$p(d_k|\hat{d}, \rho) = \rho\mathcal{N}(d_k|\hat{d}, \tau_k^2) + (1 - \rho)\mathcal{U}(d_k|d_{min}, d_{max}) \quad (4)$$

其中  $\rho \in [0, 1]$  为深度测量为 inlier 的概率， $\hat{d}$  为真实深度， $\tau_k^2$  为深度测量为 inlier 时的方差， $[d_{min}, d_{max}]$  为深度测量是 outlier 时所服从的均匀分布的区间。注意，此处假设当深度测量为 inlier 时其服从高斯分布，否则服从均匀分布。 $\rho$  为服从参数为  $\alpha$  和  $\beta$  的 Beta 分布。

为使用此概率计算后验概率 ( $p(\hat{d}, \rho|d_1, \dots, d_k)$ ) 的更新，需要一个在  $d$  上的先验高斯分布以及  $\rho$  上的 Beta 分布：

- 对于参数为  $\alpha_0$  和  $\beta_0$  的 Beta 分布，设  $\alpha_0/(\alpha_0 + \beta_0)$  即 Beta 均值  $\mathbb{E}[\rho]$  为离群掩码的预测值。进一步，将  $\alpha_0 + \beta_0$  设为固定的调优参数，其控制着 Beta 分布的方差。
- 先验高斯分布的均值  $\mu_0$  设为 CNN 预测的深度值，标准差  $\sigma_0$  设为设为预测深度的百分比。这个百分比设置为一个调优参数，也可以使用 CNN 进行预测，但这会使得 SLAM 和 CNN 之间引入额外的复杂性。

先验参数  $\alpha_0, \beta_0, \mu_0, \sigma_0$  将在后验计算时被更新，并设当前时间这些被更新的参数为  $\alpha_k, \beta_k, \mu_k, \sigma_k$ 。

### 3.2 Pose Estimation

在关键帧建立之后，根据关键帧跟踪后续的帧。帧相对于关键帧的位姿估计直接通过图像对齐活的：

$$\min_{T,a,b}\{L_{pho}(I_f, aI_{kf} + b) + w[(a - 1)^2 + b^2]\} \quad (5)$$

$T = [R, t]$  为位姿矩阵,  $a$  和  $b$  为仿射变换的参数,  $I_f$  和  $I_{kf}$  为当前帧和关键帧的灰度图。(5) 式的第二项表示仿射变换参数的正则化, 此正则损失通过  $w$  进行加权。计算  $L_{pho}$  时的深度图如(2)式, 其由每个像素的高斯均值构成。(1) 中  $W(p)$  则为像素  $p$  对应的  $\mathbb{E}[\rho] = \alpha_{k-1}/(\alpha_{k-1} + \beta_{k-1})$ 。也就是说, 如果一个像素根据概率模型得到可能是离群值, 则对该像素的光度损失进行加权。

(5) 的优化求解通过牛顿法得到 (考虑 TensorFlow 对计算 Hessian 矩阵的优势)。为进一步减少离群值的影响, 使用 Huber 损失函数, 以迭代重新加权最小二乘的方式实现, 作用于光度损失  $L_{pho}$ 。在迭代优化过程中, 将增量累加到 SE(3) 位姿矩阵上。迭代开始时的初始位姿估计是由给定过去位姿估计的恒定运动模型计算的。

### 3.3 Discrete Depth Search

在估计到位姿后, 通过给定的位姿估计重新估计关键帧的深度图  $D_{kf}$  以及放射变换参数  $a, b$ 。对于  $I_{kf}$  中的每个像素, 最大化像素在  $I_{kf}$  和  $I_f$  上对应的周围  $3 \times 3$  区域 ( $\omega$ ) 的归一化互相关 (Normalized Cross-Correlation, NCC) :

$$\max_{D_{kf}(\omega)} \frac{\sum_{p \in \omega} I'_{kf}(p) I_f(\Pi(p))}{\sqrt{\sum_{p \in \omega} I'_{kf}(p)^2} \sqrt{\sum_{p \in \omega} I_f(\Pi(p))^2}} \quad (6)$$

其中  $I'_{kf} = aI_{kf} + b$ 。通过简单网格化方法解决此优化问题。对于  $\omega$  中的每个像素, 将其在高斯均值  $\mu_{k-1}$  上下两个标准差 ( $2\sigma_{k-1}$ ) 的深度范围离散化为数个点。将整个区域 ( $D_{kf}(\omega)$ ) 离散化在一起, 与 TensorFlow 中的批处理兼容。中心像素被赋值为最大化 NCC 分数的离散深度点。这个重新估计的深度在后验更新中被视为深度测量 ( $d_k$ )。

### 3.4 Posterior Update

(4) 式给出的测量模型是不平凡的, 但是后验可以近似为 Beta 分布和高斯分布的乘积:

$$q(\hat{d}, \rho) = \text{Beta}(\rho | \alpha_k, \beta_k) \mathcal{N}(\hat{d} | \mu_k, \sigma_k^2) \quad (7)$$

因此, 计算后验仅包含了从时间  $k - 1$  的值更新  $\alpha_k, \beta_k, \mu_k, \sigma_k$ 。详细的更新步骤见文献 [9]。

在更新过程中，(4) 式中深度测量的标准差  $\tau_k$  是需要的。一些方法中，这种测量不确定性是通过在深度搜索步骤中假设图像中一个像素的标准偏差，并使用几何方法将这个像素的不确定性反向投影到深度不确定性来获得的。相反，本文使用简单近似来反向传播像素的不确定性：

$$\tau_k^2 = \left(\frac{\delta_d}{\delta_\lambda}\right)^2 \tau_\lambda^2 \quad (8)$$

其中  $\tau_\lambda$  为假设的一个像素的标准差， $\delta_d$  为 3.3 中的深度搜索范围， $\delta_\lambda$  为在图像  $I_f$  上对应像素的极线搜索范围。通过这种方式，使用投影函数的数值近似雅可比矩阵反向传播不确定性。

### 3.5 Keyframe Criteria and Keyframe Propagation

3.2 - 3.4 的方法在每一个进来的帧中重复，直到当前帧离关键帧足够远。使用两个标准来判断当前帧是否为新的关键帧：

- 为关键帧之后允许通过的最大帧数设置一个阈值。
- 在姿态估计步骤 3.2 的最后一次迭代中，计算投影在图像边界内的有效像素数量，即 (1) 中的  $|V|$ 。如果有效像素的百分比低于阈值，就插入一个新的关键帧。

当一帧被选择为关键帧后，该帧的估计位姿被存储为关键帧位姿。对于上一关键帧中具有高内点概率的像素，以类别概率形式的语义标签使用透视投影函数 (2) 与最新的深度图估计和相对位姿估计传播到新关键帧。将传播后的类概率与新预测的 softmax 概率进行融合：

$$P(c|I_{0,\dots,j}) = \frac{1}{Z} P(c|I_{0,\dots,j-1}) P(O = c|I_j) \quad (9)$$

$j$  为关键帧索引， $P(O = c|I_j)$  为预测的像素为类别  $c$  的 softmax 概率， $P(c|I_{0,\dots,j-1})$  为从前一关键帧传播来的类别概率， $P(c|I_{0,\dots,j})$  为融合的类别概率。

## 4 Results

---

- 数据集: KITTI, ScanNet
- 硬件: NVIDIA GTX-1070

## 4.1 Training on KITTI

对于 KITTI 数据集，之前关于深度无监督学习的工作提供了高质量的训练权重。由于本文的网络架构与 [10] 中的网络架构兼容，本文使用其公开的权重来初始化网络中的权重。语义分割网络在 Cityscapes 数据集上进行训练。使用 [11] 在 [12] 上进一步训练深度网络，以实现离群点掩码预测。

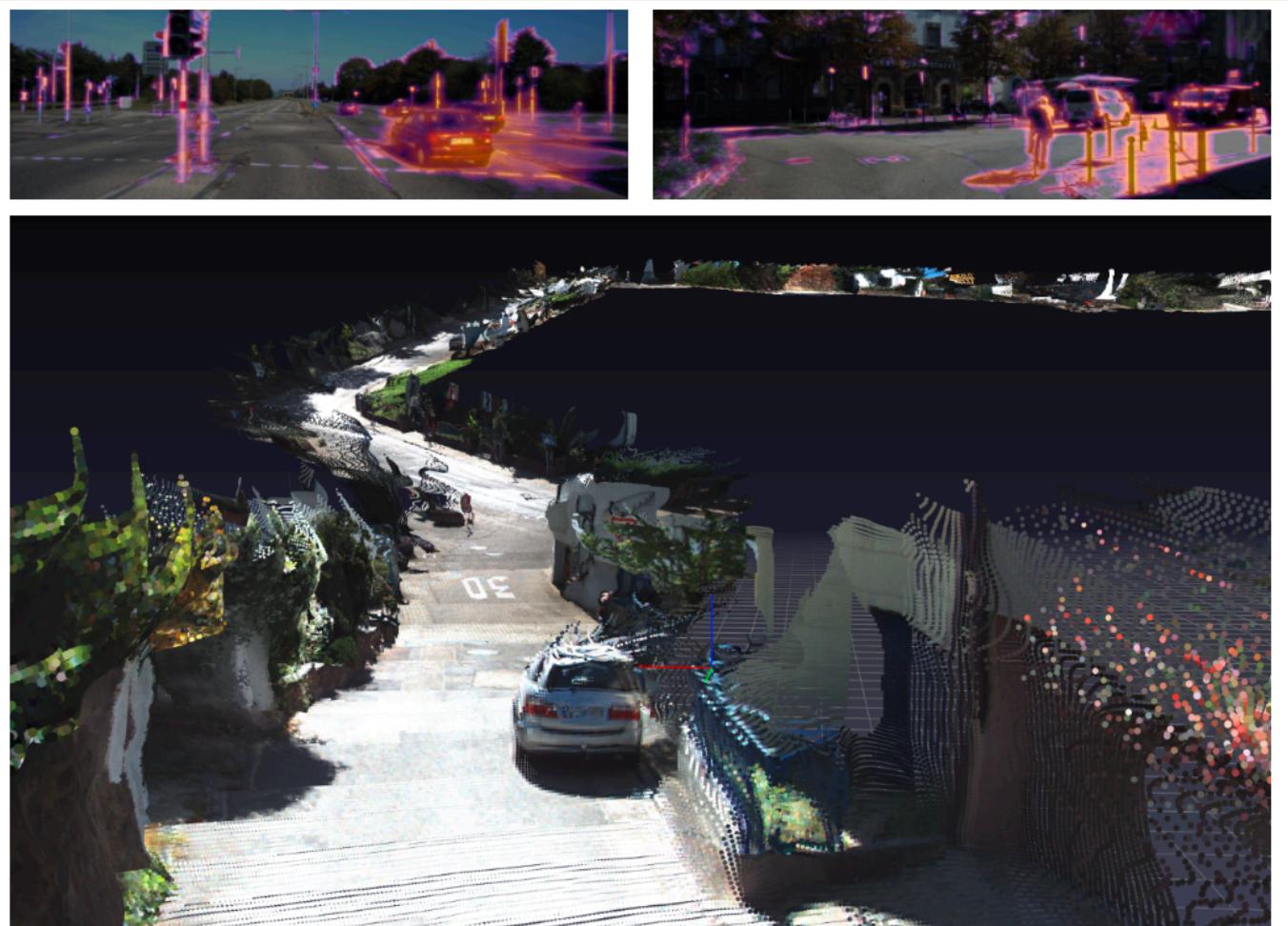


Fig. 1: Top row: The learned outlier mask. The brighter and yellower regions indicate the predicted outliers in the photometric consistency due to dynamic objects and static occlusions. Bottom row: A dense point cloud of KITTI [15] odometry sequence 10 generated by our SLAM front-end. (Far points are excluded for cleaner visualization.)



Fig. 2: Visualization of the predicted outlier mask. The darker regions indicate inlier pixels, whereas the brighter and yellower regions represent the predicted outlier pixels. Dynamic objects such as cars and people, and static occlusions such as poles and object boundaries are clearly identified.

## 4.2 Trajectory Evaluation on KITTI

TABLE I: Absolute Trajectory Error (ATE RMSE) on the KITTI odometry sequences 09 and 10 computed on 5-frame snippets with the evaluation method in [4] (lower is better). O.M., D.W., and P.U. stand for outlier mask, down-weighting in the pose estimation, and posterior update respectively.

Method	Seq. 09	Seq. 10
ORB-SLAM (full)	$0.014 \pm 0.008$	$0.012 \pm 0.011$
ORB-SLAM (short)	$0.064 \pm 0.141$	$0.064 \pm 0.130$
O.M.+P.U.+D.W.	<b><math>0.060 \pm 0.140</math></b>	<b><math>0.035 \pm 0.064</math></b>
O.M.+D.W.	$0.063 \pm 0.152$	$0.036 \pm 0.075$
P.U.+D.W.	$0.077 \pm 0.175$	$0.045 \pm 0.090$
O.M.+P.U.	$0.075 \pm 0.167$	$0.045 \pm 0.090$
None	$0.075 \pm 0.168$	$0.044 \pm 0.090$

TABLE II: Absolute Trajectory Error (ATE RMSE) on four KITTI raw sequences (2011\_09\_26) for the full system and the system without the posterior update and down-weighting.

Method	0009	0046	0059	0084
Full	<b>4.09</b>	<b>0.59</b>	<b>2.78</b>	<b>3.35</b>
None	4.88	0.64	2.98	3.90

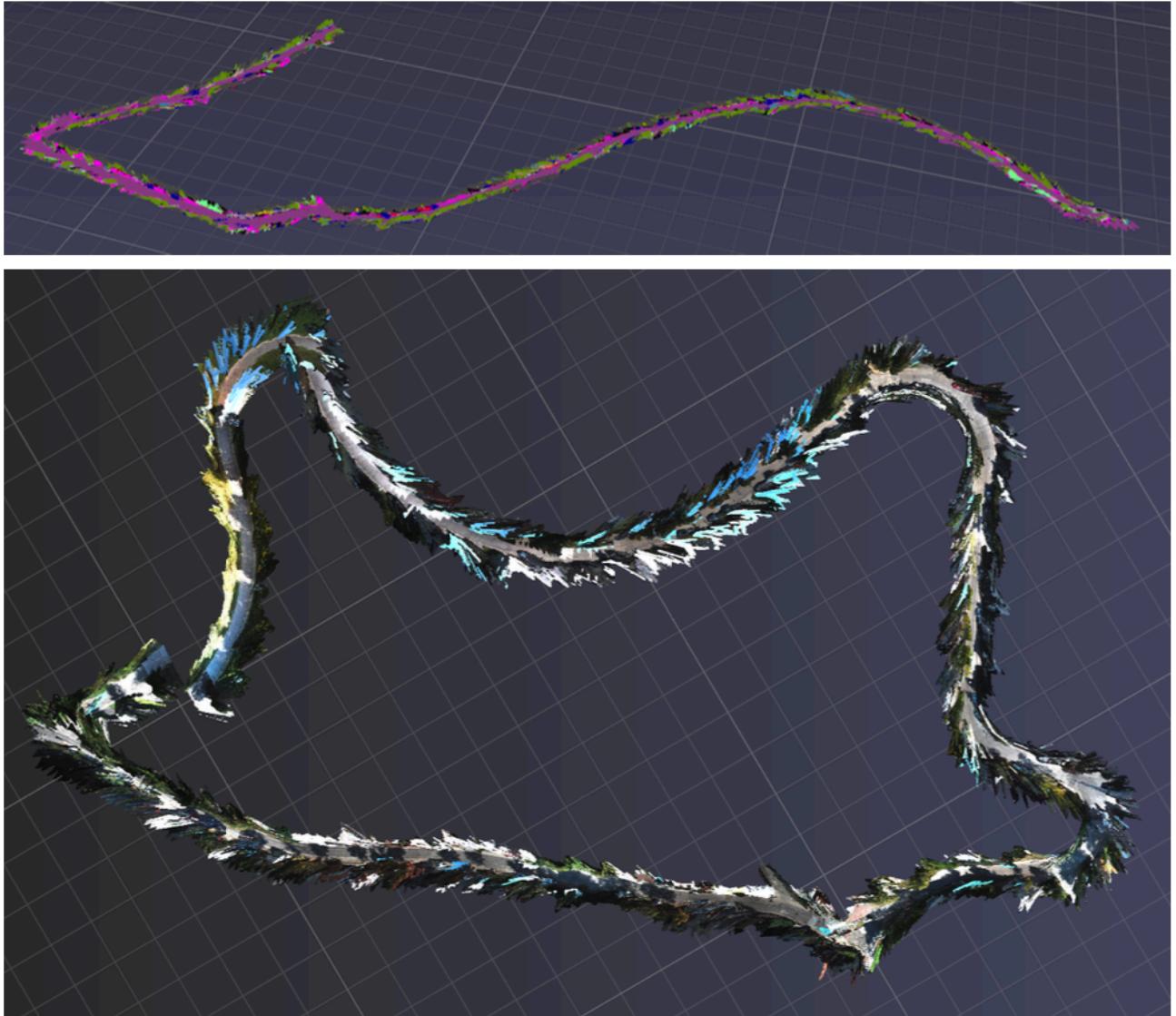


Fig. 3: Top row: A semantically labeled dense point cloud of KITTI odometry sequence 10 generated by our front-end. Bottom row: A dense point cloud of sequence 09. (Far points are excluded for cleaner visualization.)

- O.M.: outlier mask
- D.W.: down-weight
- P.U.: posterior update

### 4.3 Evaluation on ScanNet

TABLE III: Mean Absolute Trajectory Error (ATE RMSE) on 30-frame snippets made from ScanNet [29] sequences. O.M., D.W., and P.U. stand for outlier mask, down-weighting in the pose estimation, and posterior update respectively.

Method	0144_00	0559_01	0565_00	0606_02
Full	<b>0.017</b>	<b>0.021</b>	<b>0.018</b>	<b>0.012</b>
O.M.+D.W.	0.025	0.042	0.035	0.018
P.U.+D.W.	0.075	0.054	0.020	0.020
O.M.+P.U.	0.081	0.024	0.019	0.015
None	0.018	0.027	0.021	0.026

TABLE IV: mIoU of the raw CNN semantic segmentation prediction and the fused semantic labeling through keyframe propagation (IV-E).

	0144_00	0559_01	0565_00	0606_02
Raw (%)	16.77	11.80	15.64	16.74
Fused (%)	<b>16.82</b>	<b>11.93</b>	<b>15.69</b>	<b>16.80</b>

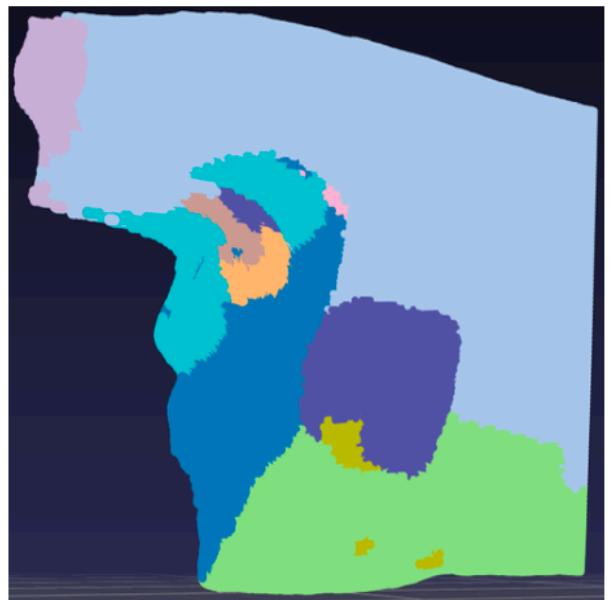
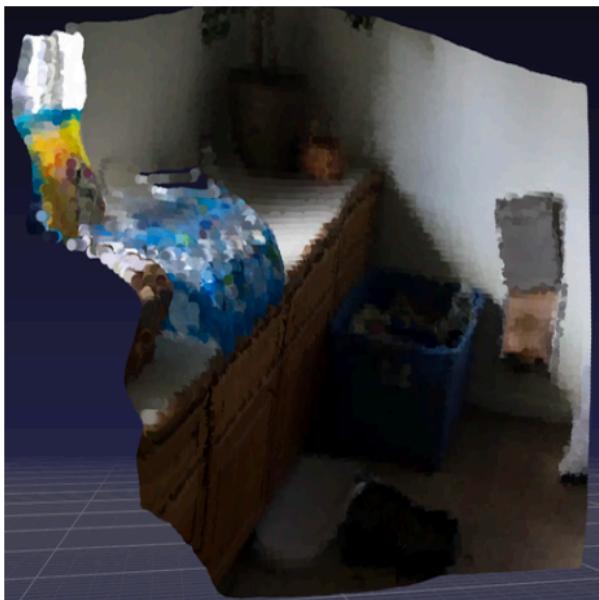


Fig. 4: A dense point cloud and its semantically labeled counterpart of a keyframe generated from the ScanNet [29] data by our front-end.

# 参考

---

- [1] K. Tateno, F. Tombari, I. Laina, and N. Navab, “CNN-SLAM: Real- time dense monocular SLAM with learned depth prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6243–6252.
- [2] N. Yang, R. Wang, J. Stuckler, and D. Cremers, “Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.
- [3] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, “D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1281–1292.
- [4] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, “CodeSLAM – learning a compact, optimisable representation for dense visual SLAM,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2560–2568.
- [5] S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison, “SceneCode: Monocular dense semantic reconstruction using learned encoded scene representations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 776–11 785.
- [6] J.Czarnowski,T.Laidlow,R.Clark, and A.J.Davison, “DeepFactors: Real-time probabilistic dense monocular SLAM,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 721–728, 2020.
- [7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.
- [8] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.
- [9] G. Vogiatzis and C. Hernández, “Video-based, real-time multi-view stereo,” *Image and Vision Computing*, vol. 29, no. 7, pp. 434–441, 2011.
- [10] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.

- [11] Y. Zhang and J. J. Leonard, “Bootstrapped self-supervised training with monocular video for semantic segmentation and depth estimation,” *arXiv preprint arXiv:2103.11031*, 2021.
- [12] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Advances in neural information processing systems*, 2014, pp. 2366–2374.