激光雷达与相机外参的语义初始化与校准

激光雷达和相机的在线语义初始化与校准

论文名称: SOIC: Semantic Online Initialization and Calibration for LiDAR and Camera
作者: Weimin Wang, Shohei Nobuhara, Ryosuke Nakamura and Ken Sakurada

论文下载：https://arxiv.org/pdf/2003.04260v1.pdf

论文代码：https://github.com/−−/SOIC

本文的核心思想

文章提出了一种基于语义信息的雷达和相机的标定方法. 以往的在线标定方法需要先验信息作为算法的初始值,我们通过引入语义质心(SCs)将初始化问题转化为PnP问题，从而消除了这种局限性.尽管现有的工作已经给出了PnP问题的解析解, 但是由于点云的质心通常与其对应图像的质心不匹配, 即使经过非线性化处理, 参数的精度也无法得到提高.我们基于点云和图像数据的语义元素之前的对应关系, 构造了一个代价函数, 通过最小化代价函数来估计最优的外部参数. 并在KITTI数据集上对算法进行了评估.

研究背景

雷达传感器能够稳定地获取物体的空间数据,但是分辨率低, 无法获得物体的颜色信息; 而相机传感器能够获得高分辨率的RGB图像,但是其对光照敏感同时也无法得到距离信息. 当代移动机器人和无人驾驶汽车通过激光雷达与相机传感器的信息融合,来弥补彼此的不足. MV3D[1], AVOD[2],F-PointNet[3]等神经网络利用融合过后的信息来提高物体检测和语义分割任务的性能. 准确的外参标定是进行融合的前提条件,通常是第一步,也是最重要的一步.
在过去的几年中, 研究人员提出了许多激光雷达与相机的标定方法, 传统方法使用棋盘格标定法, 需要人工从点云和图像中选择特征之间的对应关系. 有些学者提出了更为便利的方法, 可以自动检测特征之间的匹配关系.为了提高灵活性, 在线无目标检测方法得到广泛使用,一种方法是基于观测数据, 利用观测到的点云和图像数据之间的强度或边缘特征的相关性来寻找外部参数. 同时也有基于神经网络的方法. 这类方法通常需要良好的初值. 另外一种方法是通过匹配两个传感器的运动来获得标定参数, 为了达到较高的精度, 需要充分和准确的自我运动估计。
在本文中, 我们提出了在线语义标定方法, 通过利用点云和相机的语义信息来解决初值的问题. 如图1所示.SOIC利用语义分割的结果估计初值和最终的外参.因为只要有足够多的语义对象之间的变化而不需要整个场景SOIC就可以工作. 它甚至可以用于离线场景(如室内机器人的校准). 本文的主要贡献如下:

- (1).提出了一种基于语义分割的雷达和相机的外参标定方法
- (2).引入语义质心(SC)来估计优化的初始值.
- (3).在KITTI数据集上对方法进行了评估, 以验证可行性.
- (4).源代码已开源
  图1 SOIC通过点云和图像的语义标签匹配计算外参

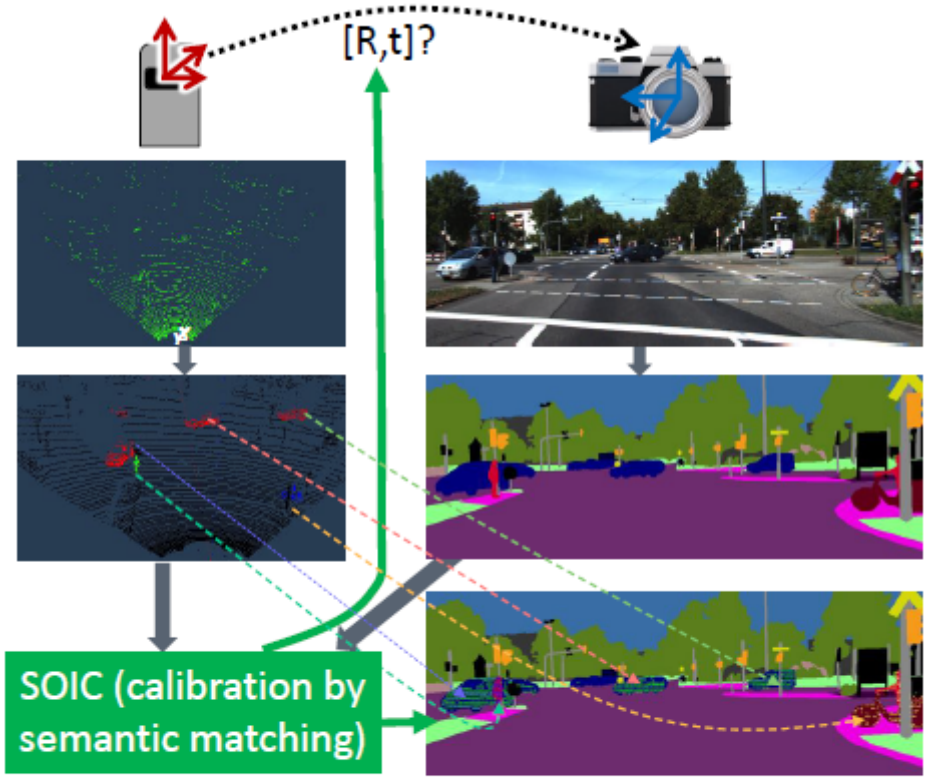


Fig. 1: SOIC estimates the extrinsic calibration parameters between LiDAR and camera sensors based on the semantic matching of point cloud and image data.

方法

我们的方法包括三个步骤:

- (1).利用现有的方法对图像和点云进行了预训练, 得到语义分割结果, 基于这些语义分割结果,通过语义质心(SCs)来得到一个初始的位姿估计值.
- (2).在语义对应信息约束下, 定义了代价函数.
- (3).最后以初始值为基准, 对代价函数进一步的优化,得到更加精确的参数.

A.公式形式

对于点云和图像的匹配点, 我们描述$P^L = \{\mathbf{p}_1^L, \mathbf{p}_2^L, \ldots, \mathbf{p}_n^L\}$, 其中$\mathbf{p}_i^L = (x_i, y_i, z_i) \in \mathbb{R}^3$, n是点云的数量, L表示为激光雷达坐标系.
另外, 我们标记每个点p_i的语义标签,$\ell_{[l,m]}^{img} \in S$, $S = \{0, 1, 2 \ldots N\}$表示语义集合.
类似地, 我们对WxH的图像每个像素也进行了标注, $\ell_{[l,m]}^{img} \in S$表示像素$I[l, m]$的类别. 其中
$l \in [0, W]$ and $m \in [0, H]$, 由于分辨率的差异像素的数量远大于点云的数量.
我们定义旋转角$\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$和平移向量$\mathbf{t} = (t_x, t_y, t_z)$表示从点云$P^L$到相机$P^C$的坐标变换. 因此许多点被投影到相同类别的图像像素上.
激光雷达坐标系下的点可以通过旋转角和平移向量通过下式变换到相机坐标系下:

$$\mathbf{p}_i^C = \mathcal{R}(\theta) \cdot \mathbf{p}_i^L + \mathbf{t}$$

如果我们知道相机内参$\mathbf{K}$和投影函数$\mathcal{P}$, 可以通过如下方法将位于相机坐标系下的空间点投影到像素坐标$[u^i, v^i]$上.

$$[u^i, v^i] = \mathcal{P}\left(\mathbf{K}, \mathbf{p}_i^C\right)$$

注意: 由于错误的外参估计可能会使一些空间点投影到图像后, 超出像素范围.
我们定义代价函数来使得点云的标签$\ell_i^{pcd} \in S$和像素标签$\ell_{[u^i,v^i]}^{img} \in S$的一致性最大化, 因此我们定义代价函数为:

$$\mathcal{C} = 1 - e^{-\epsilon^{-1}|\left(\ell_i^{pcd} - \ell_{[u^i,v^i]}^{img}\right)|}$$

其中ϵ是一个很小量这也导致了$e^{-\epsilon^{-1}}$接近于0. 如果点p_i和像素$\left[u^i, v^i\right]$标签是一致的, 那么代价函数$\mathcal{C}$近似为0, 如果类别不相同代价函数$\mathcal{C}$将会趋近于1.
对于转换到相机坐标系的点p_i^C, 如果超出图像或者与像素标签不一致, 则通过定义一个距离函数$\mathcal{D}$来计算原始激光雷达坐标系下的点p_i^L的损失.

$$\mathcal{D}\left(\mathbf{p}_i^L\right) = \min_{\ell_{[l,m]}^{img} = \ell_i^{pcd}} \left(\mathcal{M}\left([u,v]^i, [l,m]\right)\right) \left|p_i^L\right|^2$$

因此这个函数计算图像中具有相同标签的曼哈顿距离. 因此上式的$\mathcal{M}$定义如下:

$$\mathcal{M}([u,v],[l,m]) = |u-l| + |v-m|$$

由于语义分割会有多个类别, 我们对不同类别进行了加权, 这样可以根据不同类别对损失函数进行计算, 我们定义的类别加权函数$\mathbf{1}_A$如下:

$$1_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

结合上述的公式, 我们可以得到点云和图像的最终的代价函数, 分母表示有效的语义标签的个数

$$\mathcal{L} = \frac{\sum_{s \in S} \sum_i^n 1_{\{s\}} \left(\ell_i^{pcd}\right) \mathcal{C}\left(\mathbf{p}_i^L\right) \mathcal{D}\left(\mathbf{p}_i^L\right)}{\sum_{s \in S} \sum_i^n 1_{\{s\}} \left(\ell_i^{pcd}\right)}$$

最终我们通过最小化代价函数来求解外参变量$\hat{\boldsymbol{\theta}}$和$\hat{\mathrm{t}}$

$$\hat{\boldsymbol{\theta}}, \hat{\mathbf{t}} = \arg\min_{\boldsymbol{\theta}, \mathbf{t}} \mathcal{L}$$

B. 代价函数的初始化

受到解决PnP问题的控制点决策的启发, 我们提出了语义质心来构造一个可以解析求解的含噪声的PnP问题. 如图2所示, 我们通过如下式子类别s的语义信息.

$$P_s^L = \left\{\mathbf{p}_{s,1}^L, \mathbf{p}_{s,2}^L, \ldots \mid \mathbf{p}_{s,i}^L \in P^L, \ell_i^{pcd} = s\right\}$$

语义质心如下

$$\mathbf{SC}_\mathbf{s}^L = \frac{\sum_{\mathbf{p}_i \in \{P_d^L\}} \mathbf{p}_i}{\left|\left\{P_s^L\right\}\right|}$$

在PnP问题中, 我们将点云的语义质心$3D-SC$与图像的语义质心$2D-SC$视为3D-2D的匹配点对.需要注意的是, 点云的三维语义质心通常与图像的几何中心不一致. 这意味着点云的语义质心和图像的语义质心可能没有很好的对应. 因此, 这里计算的结果仅作为代价函数的初始值.
图2 展示了点云和图像的语义分割结果以及语义质心

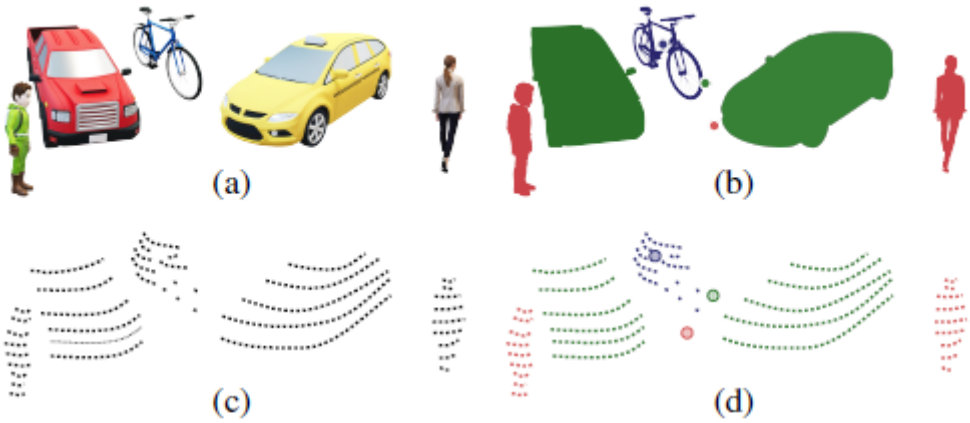
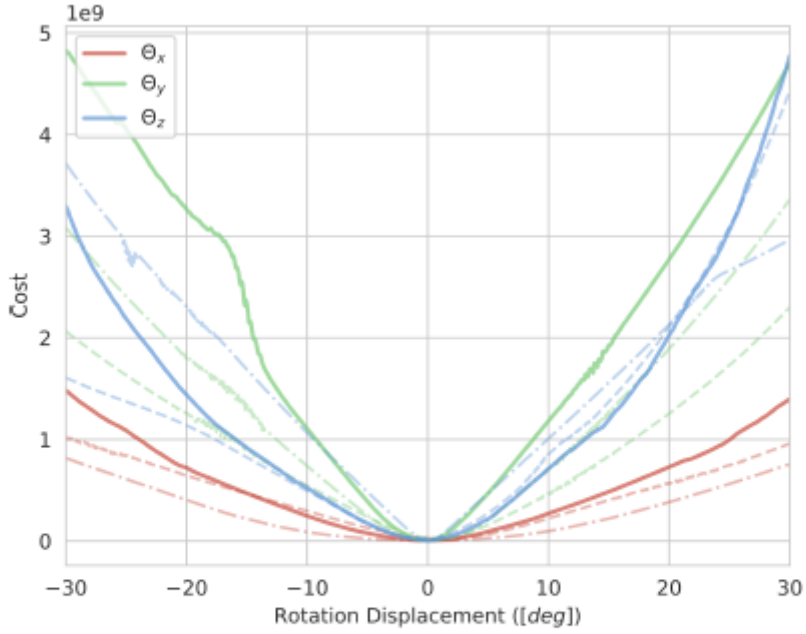


Fig. 2: (a) and (c) are RGB image and the corresponding 3D point cloud acquired by camera and LiDAR sensor. (b) and (d) are semantic segmentation results of (a) and (c). Red, green, blue represents pedestrians, vehicles, bicycles class respectively. Three filled circles in (d) and (d) indicates semantic centroids (SC) of each class.

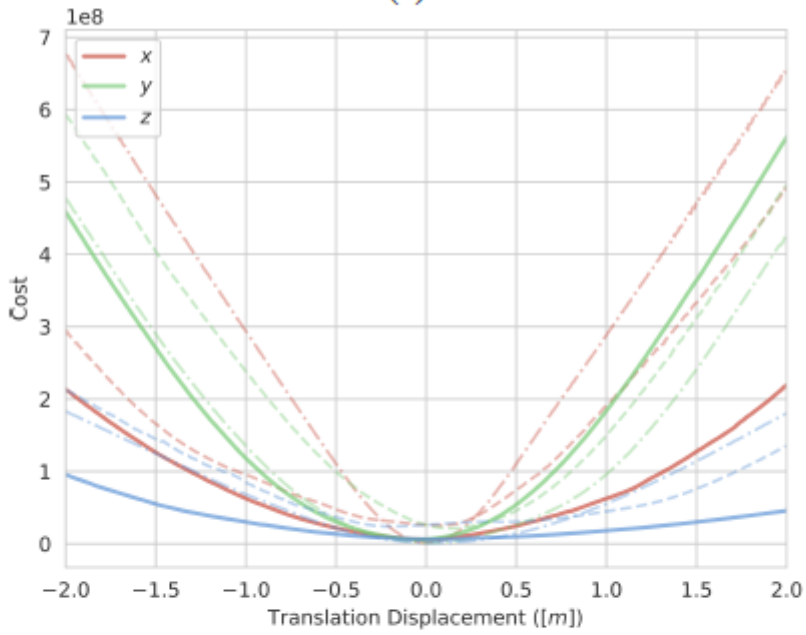
实验结果

我们首先在KITTI有标签的数据集上评估了所提出的方法的性能. 并将其与在线的方法进行了比较. 随后我们证明不同的外参变量对代价函数的影响以及收敛情况

图3 代价函数的收敛情况



(a)



(b)

Fig. 3: Cost change of the designed cost function along with the (a) angular and (b) translation displacement of x−, y−, z−axis respectively. The cost is calculated with 20 pairs. solid line —— : Vehicles; dash-dot line -·-· : Pedestrians; dashed line - - - : Cyclists. The interval for angle displacement is 0.01° and 5[mm] for translation displacement.

图4 展示了100帧数据车辆类别的语义质心分布情况

我们使用RANSAC算法从3D-SCs中估计出绿色平面, 红色箭头表示平面的法线, 可以看出所有的质心都近似地分布在同一个平面上.

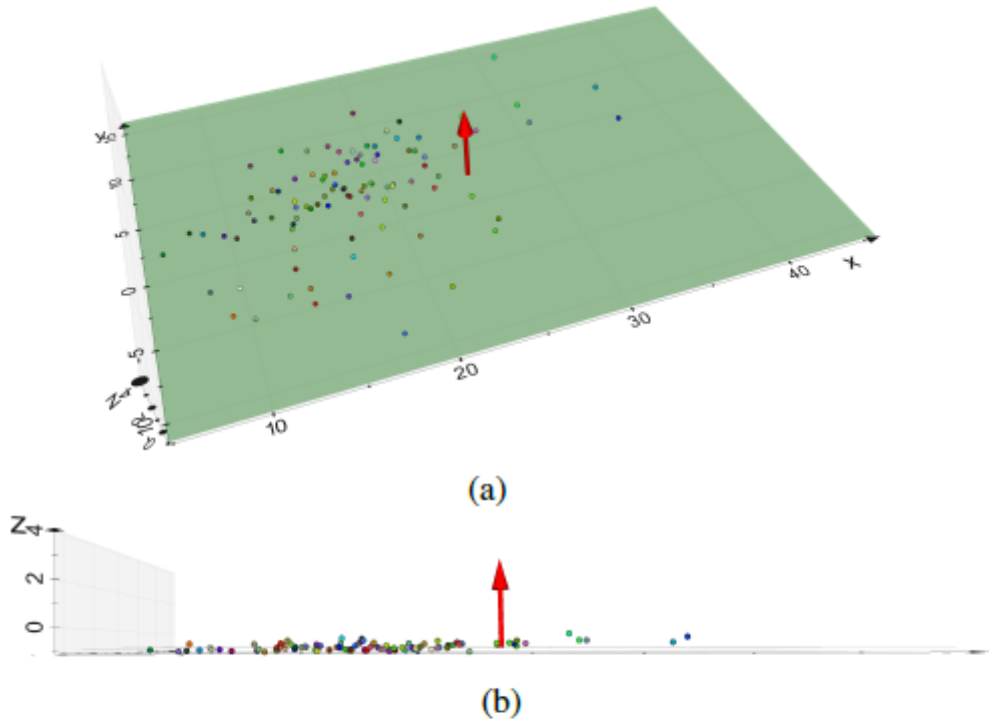


Fig. 4: Distribution of vehicles' semantic centroids of vehicles class from 100 frames. (a) top view, (b) side view. The green plane is estimated from 3D SCs with RANSAC algorithm. The red arrows indicate the normal of the estimated plane. We can see that all centroids are approximately distributed on the same plane.

图5 展示了语义质心与参数估计值的对应关系

绿色数字表示图像的语义质心, 蓝色数据表示投影点云的语义质心, 数字表示图像点云对的索引值.

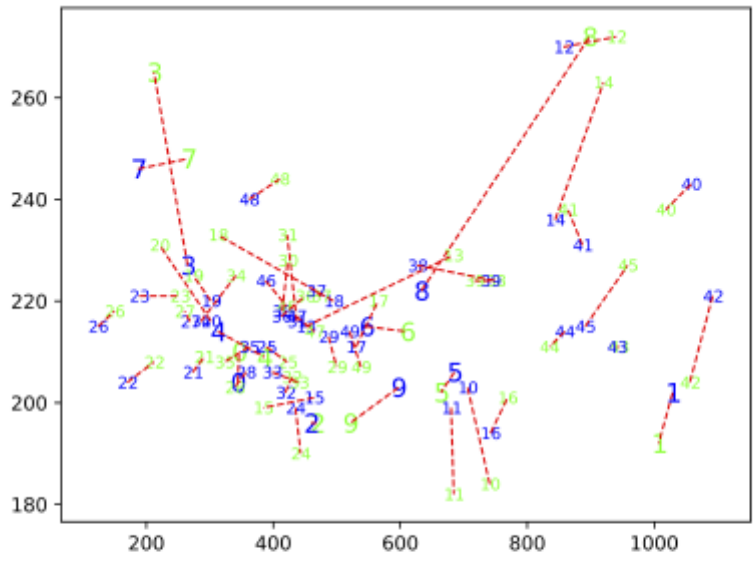


Fig. 5: Correspondence of semantic centroids with the estimated initial parameters from 50 pairs. Green numbers indicate semantic centroids from images and blue numbers show projected point cloud semantic centroids. The number indicate the index of the image-pointcloud pair.

由于我们发现在图4中可视化的语义质心都近似分布在一个平面上, 使用EPnP的方法来求解这个PnP问题将会变得很困难, 因此我们选择使用IPPE方法来进行求解.

显然语义分割的标签的可以帮助提升算法的性能, 因此我们选择训练好的预测KITTI车辆的网络PointRCNN来完成语义标签的预测工作.

图6 展示了使用网络预测出来的语义标签和真实的GT对优化结果的影响

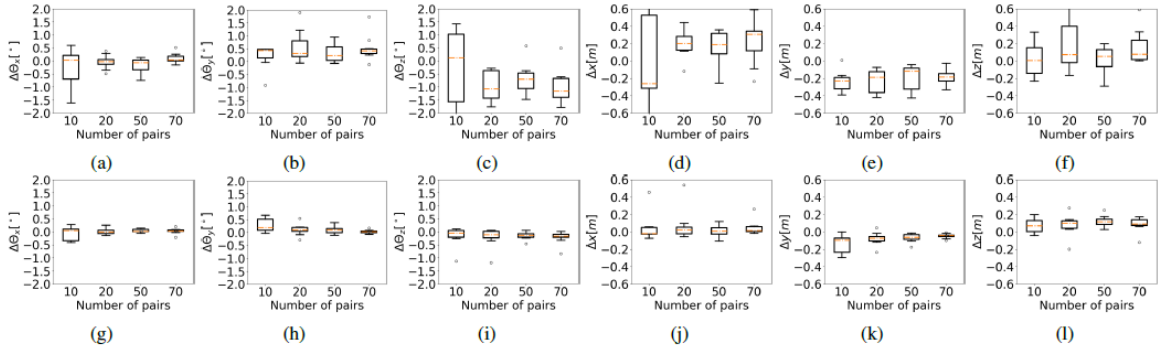


Fig. 6: Calibration results by SOIC. (a)-(f): estimated results for each parameter based on predicted semantics by PointRCNN [30] for point clouds and NVIDIA semantic segmentation model [31] for images; (g)-(l) results with GT semantics. For each number of pairs, we perform SOIC 10 times on randomly selected point cloud and image pairs.

图7 展示了使用MI的结果

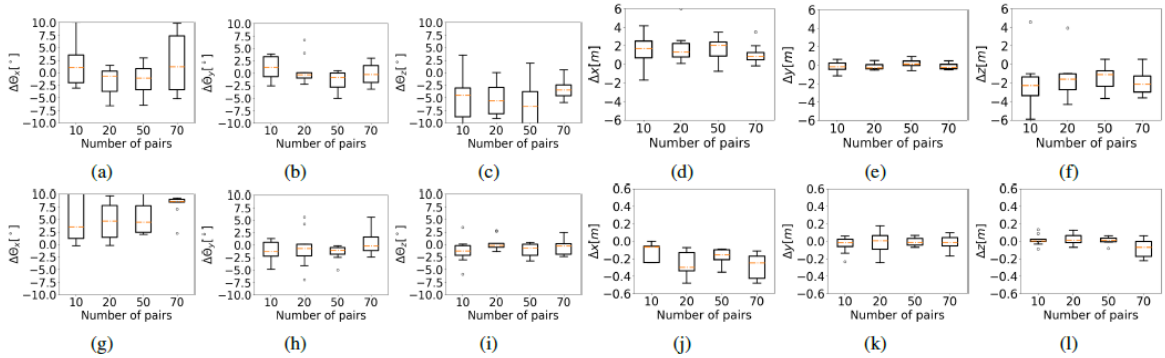


Fig. 7: Calibration results by MI [10]. (a)-(f): Results by MI with the same initial guesses estimated by SOIC; (g)-(l): Results with GT extrinsic values as the initial guess. Note that the range of y axis in (a)-(c) and (g)-(i) is 5 times, in (d)-(f) and (j)-(l) is 10 times greater than that in Fig. 6.

表1 展示了不同条件下的误差值

Parameters estimated with different pairs \ Method		MI [10] (initial guess by SOIC from GT semantics)	MI [10] (initial guess with GT calibration values)	SOIC (w/ Pred. semantics)	SOIC (w/ GT semantics)
$\Delta\Theta_x[^\circ]$	10	-1.395	-0.237	-0.504	-0.399
	20	-1.149	-0.144	-0.352	-0.075
	50	0.775	2.455	0.042	0.070
	70	-1.547	2.233	-0.053	0.016
$\Delta\Theta_y[^\circ]$	10	-0.930	0.530	0.302	0.090
	20	0.006	-0.036	-0.059	-0.073
	50	0.533	-0.563	0.037	0.171
	70	-1.330	-1.390	0.331	-0.001
$\Delta\Theta_z[^\circ]$	10	3.520	0.119	0.613	0.051
	20	0.075	0.023	-0.447	-0.177
	50	-1.179	-0.533	-0.894	-0.233
	70	0.572	-0.570	-0.605	-0.132
Δt_x[m]	10	-1.693	-0.065	-0.291	-0.047
	20	0.274	-0.269	0.115	0.041
	50	0.500	-0.097	0.168	0.061
	70	0.427	-0.110	0.185	0.049
Δt_y[m]	10	0.304	-0.016	-0.214	-0.085
	20	-0.555	0.026	-0.072	-0.015
	50	0.013	0.036	-0.065	-0.086
	70	0.081	-0.015	-0.172	-0.042
Δt_z[m]	10	-1.352	0.010	0.275	0.003
	20	-3.050	-0.070	0.028	0.130
	50	-3.289	-0.010	0.072	0.090
	70	-2.096	0.002	0.055	0.078

TABLE I: Errors of calibration result by MI and SOIC under different conditions. Darker color indicates the greater error. For SOIC, the calibrated parameters with the least cost are selected from 10 trials. For MI, we manually selected parameters with the least errors.

图8 展示了实例场景的实验结果

我们用于语义分割的图像是图8(a), 图8(b)展示了通过PointRCNN网络预测出来的属于车辆类别的点云. 并利用SOIC预测出来的外参将属于车辆类别的点云投影到图像中.图8(c)使用SOIC的预测的外参将GT指定的车辆类别的点云投影到图像上, 图8(d)是使用GT提供的车辆点云的类别, 并且使用GT的外参将属于车辆类别的点云投影到图像上.粉红色框的内容表示车辆类别的点云,在经过外参投影到图像的像素中时会和原始语义图像(图像的语义类别由图像的GT提供)车辆的像素位置发生轻微错位.图8(b)的错位情况较为严重,是由于车辆的类别和外参都是预测得到的.

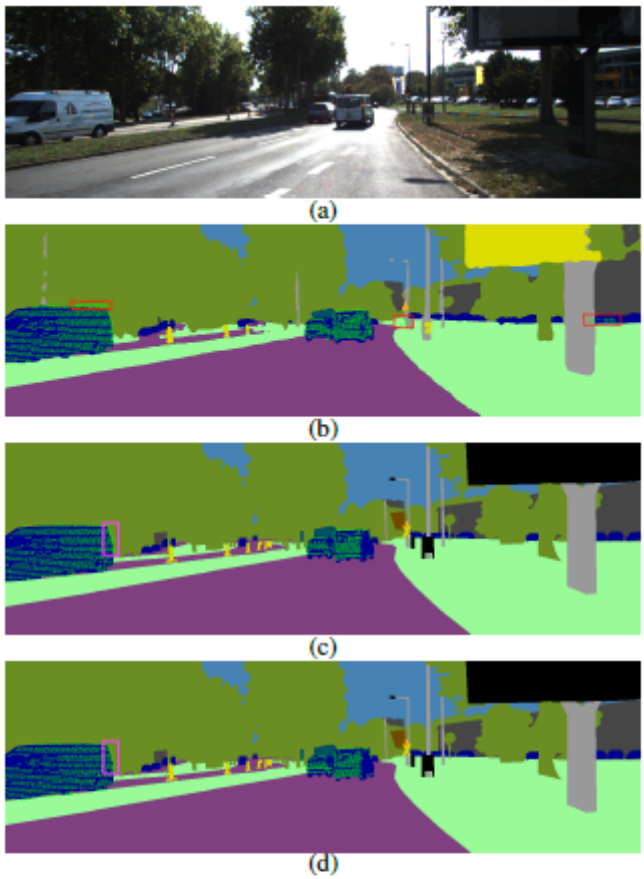


Fig. 8: An example qualitative result. (a): Raw RGB image. (b): Projection of predicted semantic 3D points to the corresponding predicted semantic image with extrinsic parameters estimated by SOIC w/ predicted semantics. (c): Projection of GT semantic 3D points to the corresponding GT semantic image with extrinsic parameters estimated by SOIC w/ GT semantics (d): Projection of GT semantic 3D points to the corresponding GT semantic image with GT extrinsic parameters. For semantic images, the color are conformed with the Cityscapes Dataset [28]. Green points in (b)-(d) are the vehicle's class predicted by PointRCNN. Red boxes in (b) show the points that are segmented as Vehicles class wrongly by PointRCNN. The pink box in (c) shows the slight mis-alignment compared with that in (d).