CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction

CNN-SLAM:实时稠密单目 SLAM 及深度估计

摘要

鉴于深度卷积神经网络(CNN)在深度估计方面的最新进展，本文研究了如何用 CNN 和单目相机重建出准确且稠密的三维场景。我们提出了一个方法，它可让 CNN 预测的稠密深度图与单目 SLAM 估计的深度图自然融合。该融合方案在图像定位中由于传统的单目 SLAM 方法，在单目 SLAM 失效的时候（如低纹理区域）仍然能够很好地工作。我们展示了用深度预测来估计重建场景的绝对尺度，克服了单目 SLAM 无法得到绝对尺度这一主要缺陷。最后，我们提出了一个框架，它可将单幅图像中提取的语义标签和稠密 SLAM 高效地融合在一起，从单视角产生语义一致性(coherence，也可译作“语义相干”)的场景重构。两个基准数据集的评估结果显示了我们方法的鲁棒性和准确性。

1 引言

传统 SLAM 有几点缺陷：

1.工作范围小。

2.基于主动感知的 SLAM 方法在日光场景下表现很差。

3.单目 SLAM 无法获得绝对尺度，存在尺度漂移现象。若无法获得绝对尺度，则在 VR/AR 和机器人应用中都会遇到很大问题。

4.单目 SLAM 无法在纯旋转下工作，因为此时的 stereo-baseline 丢失，无法进行立体估计。

近来，通过机器学习方法从单幅图像估计场景深度的研究取得了新进展。其中，深度卷积神经网络(CNN)表现出色，其优点是可以得到高分辨、含绝对尺度的深度估计图，甚至在特征点稀疏或重复性纹理(repetitive patterns)的情况下也表现出色。由于场景的绝对深度可用机器学习通过样本训练得到，因此该方法不需要针对场景做出模型假设，也不需要几何上的约束。然而，此方法的缺陷是：虽然深度预测在全局上是准确的，但深度图中物体的边界处是局部模糊的，这样会丧失物体深度、形状的细节信息。

CNN-SLAM 的主要思想是：将深度估计和单目 SLAM 结合。为了解决深度图中边缘模糊的问题，我们将 CNN 预测的深度图作为稠密重建的初始猜测，并连续地通过直接法 SLAM 来优化（依靠 small-baseline stereo matching，与文献【4】的 LSD-SLAM 相似。LSD-SLAM 是典型的直接法)。非常重要的一点是，small-baseline stereo matching 拥有改善边缘区域深度估计的潜力。与此同时，CNN 预测的深度图含有绝对尺度，可为位姿估计提供更多约束条件，显著提高了位姿估计、路径和重建场景的精度。由于该方法可在纯旋转等恶劣情况下工作，因此 tracking 的稳定性大大增强。在实践方面，该框架可在 PC 上实时运行，CPU 和 GPU 同时运算，用 GPU 进行 CNN 深度估计，用 CPU 进行深度优化。

除可用于深度估计之外，CNN 还可成功地用于其他高维回归(high-dimension regression)的任务，其中一个典型的例子就是语义分割(semantic segmentation)。我们基于语义分割提出了 CNN-SLAM 的扩展版本：用像素级的(pixel-wise) labels 将语义 labels 和稠密 SLAM 准确、高效率地融合。

2 相关研究工作

SLAM 按传感器分，可分为基于深度相机的和基于单目相机的，按方法分可分为直接法和特征点法。

对于单目 SLAM（或者说纯几何的 SLAM），ORB 可以说是 state-of-the-art 的。

由于深度学习的发展，从单视角进行深度估计的研究得到了越来越多的重视。

3 单目语义 SLAM

我们在该部分详细地介绍 CNN-SLAM，图 2 显示了该算法的框架和流程。

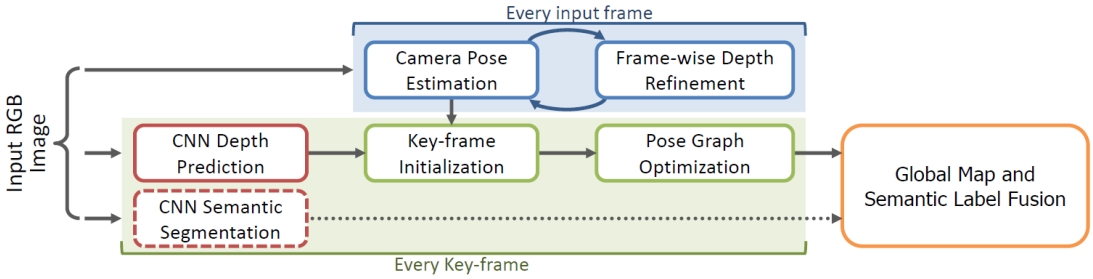

Figure 2. CNN-SLAM Overview.

我们使用基于关键帧的 SLAM 框架，并用文献【4】的方法（直接法、半稠密）作为 baseline，从视觉上显著的(visually distinct)帧集合中选出关键帧。这些关键帧的位姿将参与全局优化。与此同时，每输入一帧图像，程序就通过该帧与其最近关键帧的变换(transformation)来估计该帧的相机位姿。

为保持算法能以高帧率运行，我们只预测关键帧的深度图（用 CNN 生成）。如果当前的位姿估计和已有的关键帧位姿相差很多，则算法基于当前帧生成一个新的关键帧，并用 CNN 生成其深度图。另外，我们还通过对每个深度预测图进行像素级的置信衡量(confidence measure)，生成一个不确定性图(uncertainty map)。由于一般情况下，SLAM 用的相机和拍摄 CNN 训练图像的相机并不同，因此我们提出了将深度图规范化（normalized）的方法。这样，对不同内参的相机拍摄的视频流都有较好的鲁棒性。当进行语义 label fusion 时，我们用另一个卷积网络对输入帧进行语义分割的预测。最后，生成关键帧的位姿图，来全局优化它们的相对位姿。

该框架一个很重要的环节（也是我们的主要贡献），就是通过 small-baseline stereo matching 来改善 CNN 预测的关键帧深度图。我们通过在关键帧和输入帧之间最小化 color consistency 来实现。深度图中的边缘区域将会主要通过该环节得到优化，epipolar matching 可提供更高的准确性（这部分内容将在 3.3 和 3.4 中介绍）。优化后的深度的传播方式受各深度值的不确定性的影响，这种不确定性由我们专门提出的置信衡量而得出（在 3.3 中定义）。该框架的每一个步骤都将在以下小节中详细阐述。

3.1 相机位姿估计

CNN-SLAM 的位姿估计受 LSD-SLAM 的关键帧法的启发而得。系统保存了一系列关键帧 $k_1, k_2, \ldots k_n$ 并将它们作为结构元素参与 SLAM 重建。每个关键帧 k_i 都与相应的关键帧位姿T_{k_i}、深度图D_{k_i}和深度不确定图U_{k_i}对应。与【4】不同的是，我们的深度图D_{k_i}是稠密的，它由 CNN 进行深度预测而得（详见 3.2 节）。D_{k_i}估计了每个深度值的 confidence。【4】将D_{k_i}初始化为一个大的、恒定值。与之不同的是，我们根据深度预测的 measured confidence 初始化D_{k_i}。我们将深度图的元素标记为 $u=(x, y)$，并用$\dot{u}$作为 u 的齐次坐标表示(homogeneous representation)。

我们对每一帧 t 估计它的相机位姿$T_t^{k_i} = [R_t, t_t]$，即离 t 最近的关键帧k_i到 t 的变换。它由 3*3 的旋转矩阵R_t和 3D 平移向量t_t组成。该变换由最小化当前帧图像I_t最近关键帧图像I_{k_i}的光度残差(photometric residual)，并通过高斯-牛顿优化目标函数得到：

$$E\left(T_t^{k_i}\right) = \sum_{\tilde{u}\in\Omega} \rho\left(\frac{r\left(\tilde{u}, T_t^{k_i}\right)}{\sigma\left(r\left(\tilde{u}, T_t^{k_i}\right)\right)}\right)$$

ρ 是 Huber 范数(Huber norm)，σ 是衡量残差不确定度的函数（详见【4】）。r 是光度残差，它的定义如下：

$$r\left(\tilde{u}, T_t^{k_i}\right) = I_{k_i}(\tilde{u}) - I_t\left(\pi\left(KT_t^{k_i}\tilde{V}_{k_i}(\tilde{u})\right)\right)$$

由于深度图是稠密的，因此从效率上考虑，我们仅对颜色梯度高的区域中的像素子集计算光度残差，由$\tilde{u}$表示，$\tilde{u} \subset u \in \Omega$。$\pi$是将 3D 点投影到 2D 图像上的投影函数：

$$\pi([xyz]^T) = \left(\frac{x}{z}, \frac{y}{z}\right)^T$$

$V_{k_i}(u)$表示顶点贴图(vertex map)中的 3D 元素，该顶点贴图由关键帧深度图计算生成：

$$V_{k_i}(u) = K^{-1}\dot{u}D_{k_i}(u)$$

K 是相机的内参矩阵。

一旦得到了$T_t^{k_i}$，当前帧在世界坐标系下的的相机位姿即可用$T_t = T_t^{k_i}T_{k_i}$计算。

3.2 基于 CNN 的深度估计和语义分割

每当生成一个新的关键帧时，CNN 便产生一幅对应的深度预测图。我们使用的深度预测方法属于目前的最高水平（文献【16】的方法），它基于 Residual Network (ResNet)延伸出一个全卷积网络。该架构的第一部分基于 ResNet-50（文献【9】），由 ImageNet 训练生成的权重对该部分进行初始化。框架的第二部分将 last pooling 和 fully connected layers (由 ResNet-50 提出)替换为一系列上采样的 blocks，这些 blocks 由 unpooling 和卷积层组成。在上采样之后，drop-out is applied before a final convolutional layer which outputs a 1-channel output map representing the predicted depth map. The loss function is based on the reverse Huber function（文献【16】）.

我们参考其他利用相同的架构进行深度预测、语义分割的成功范例（文献【3，29】），也训练了该网络，用于从 RGB 图像中预测像素级的语义 label。我们修改了网络，使它的输出通道数和类别数相同，并用了 soft-max 层和 cross-entropy loss function，通过反向传播和 Stochastic Gradient Descent(SGD)最小化此 loss function。需要指明的是，虽然原则上可以使用任何语义分割算法，但我们此工作的目的是展现帧级的分割图是如何成功地和单目 SLAM 框架融合在一起的（详见 3.5 节）。

3.3 生成关键帧和位姿图优化

使用预先训练的 CNN 来预测深度的局限之一是：如果 SLAM 所用传感器的内参和拍摄训练集图像的相机的内参不同，那么在 3D 重建时绝对尺度的估计必然不准确。为解决该问题，我们用当前相机的焦距f_{cur}和训练集所用相机的焦距f_{tr}的比值，通过 CNN 调整深度：

$$D_{k_i}(u) = \frac{f_{cur}}{f_{tr}}\widetilde{D}_{k_i}(u)$$

$\tilde{D}_{k_i}$是通过关键帧图像I_{k_i}直接由 CNN 退化得到的深度图。

图 3 显示了用上式校正的有效性，我们使用的是 *ICL-NUIM* 数据集。如图所示，经过校正后，深度图和轨迹的准确性都有明显的提升。

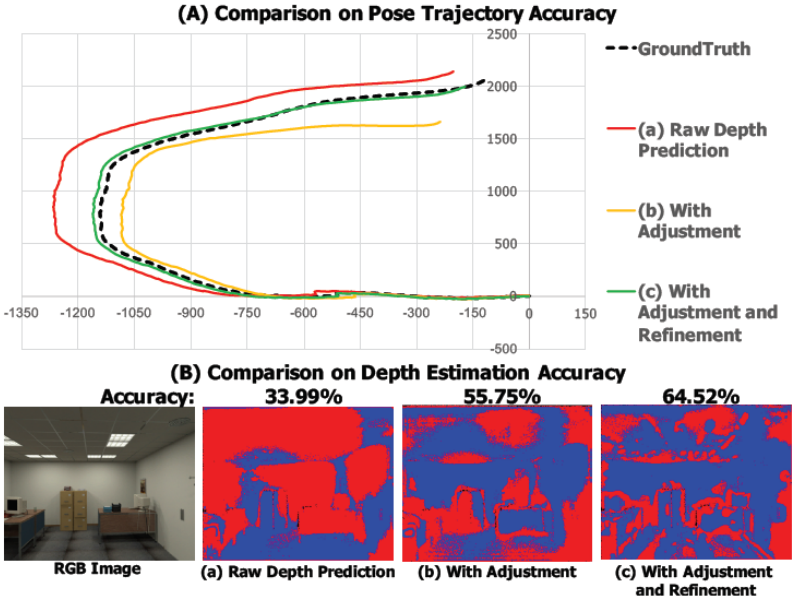

Figure 3. Comparison among (a) direct CNN-depth prediction, (b) after depth adjustment and (c) after depth adjustment and refinement, in terms of (A) pose trajectory accuracy and (B) depth estimation accuracy. Blue pixels depict correctly estimated depths, i.e. within 10 % of ground-truth. The comparison is done on one sequence of the *ICL-NUIM* dataset [8].

另外，我们将每个深度图D_{k_i}和不确定图U_{k_i}联系在一起。文献【4】将U_{k_i}的每个元素初始化为一个大的、恒定值。由于 CNN 可在不依赖 temporal regulation 的情况下产生深度图，因此我们根据当前深度图D_{k_i}和最近关键帧中对应的场景点(respective scene point)的差别，计算出置信值，从而将U_{k_i}初始化。这样以来，置信值衡量了在不同帧中，每个深度估计值的一致性程度(this confidence measures now coherent each predicted depth value is across different frames)。高置信值的元素在连续优化中将比文献【4】中的更快、更有效。

具体而言，我们将U_{k_i}定义为当前关键帧k_i的深度图D_{k_i}和最近关键帧k_j的深度图对应元素差的平方，k_j的深度图由估计的$T_{k_j}^{k_i}$(k_i到k_j的位姿变换矩阵)得到：

$$U_{k_i}(u) = \left(D_{k_i}(u) - D_{k_j}\left(\pi\left(KT_{k_j}^{k_i}V_{k_i}(u)\right)\right)\right)^2$$

为进一步提高每个新初始化关键帧的准确性，在经过它们和输入帧的优化后，我们将其D_{k_i}、U_{k_i}和由最近关键帧产生的 D、U 融合(we propose to fuse its depth map and uncertainty map with those propagated from the nearest key-frame after they have been refined with new input frames)（显然，我们不会对第一个关键帧这样操作），深度优化的详细流程详见 3.4 节。为此，我们首先定义了从最近关键帧k_j传播的不确定地图(propagated uncertainty map)：

$$\tilde{U}_{k_j}(v) = \frac{D_{k_j}(v)}{D_{k_i}(u)} U_{k_j}(v) + \sigma_p^2$$

$$v = \pi\left(KT_{k_j}^{k_i}V_{k_i}(u)\right)$$

根据文献【4】，σ_p^2是白噪声方差，用于增加传播不确定度(propagated uncertainty)。之后，根

据权重规则(weighted scheme)，将两个深度图和不确定图融合：

$$D_{k_i}(u) = \frac{\widetilde{U}_{k_j}(v) D_{k_i}(u) + U_{k_i}(u) D_{k_j}(v)}{\widetilde{U}_{k_j}(v) + U_{k_i}(u)}$$

$$U_{k_i}(u) = \frac{\widetilde{U}_{k_j}(v) U_{k_i}(u)}{\widetilde{U}_{k_j}(v) + U_{k_i}(u)}$$

最终，在位姿图中找到与新加入关键帧视场相似（即位姿相近）的一系列关键帧，根据这些关键帧创建新的边缘，并通过这些新创建的边缘更新位姿图(Finally, the pose graph is also updated at each new key-frame, by creating new edges with the key-frames already present in the graph that share a similar field of view(i.e., having a small relative pose) with the newly added key-frame)。关键帧的位姿每次都会通过位姿图优化法（文献【14】）得到全局的优化。

3.4 帧级深度优化

该步骤的目标是基于每个 new-frame 的深度估计图，不断优化当前活跃的关键帧的深度图。我们用 small baseline stereo matching 的方法（文献【5】中的半稠密框架），基于沿极线的 5-pixel 匹配，遍历当前帧 t 的每个像素生成D_t和U_t。之后，基于相机位姿估计$T_t^{k_i}$，通过关键帧k_i对D_t和U_t进行校准。

估计的D_t和U_t将根据下式直接与最近关键帧k_i的D_{k_i}和U_{k_i}融合：

$$D_{k_i}(u) = \frac{U_t(u) D_{k_i}(u) + U_{k_i}(u) D_t(u)}{U_t(u) + U_{k_i}(u)}$$

$$U_{k_i}(u) = \frac{U_t(u) U_{k_i}(u)}{U_t(u) + U_{k_i}(u)}$$

重要的是，由于关键帧和稠密深度图有关，因此该过程可稠密地执行，即关键帧的每个元素都可得到优化，而不是像文献【5】那样只优化梯度高的区域的深度值。由于低纹理区域的深度值往往有较高的不确定度（即U_t中对应位置的值较高），因此本方法可自动地、有选择性地优化：对于高梯度区域，每一帧都参与优化该区域的深度；对于低纹理区域，其深度值将逐渐稳定在 CNN 估计的深度值附近，而不会受深度值不确定性的影响。图 3-(c)展示了该深度图优化方法的有效性。

3.5 三维模型和语义标签的全局融合

相机获取的一系列关键帧可融合在一起，实现三维场景重建。由于 CNN 经过训练后可提供语义标签和深度图，因此语义信息也可以与全局三维场景的每个元素联系起来，我们将语义信息和三维模型的融合过程称为语义标签融合(semantic label fusion)。

在我们提出的框架中，我们用文献【27】提出的实时算法，该算法旨在渐进地将深度图和 RGB-D 序列中每帧生成的 connected component map 融合在一起。该方法利用 Global Segmentation Model (GSM)，随着时间的推移，将标签平均分配到每个 3D 元素，因此该方法对帧级的分割中的噪声是鲁棒的。在我们的框架中，我们将位姿估计作为算法的输入。这是因为相机位姿是由单目 SLAM 估计的，而输入的深度图只与一系列捕获的关键帧有关。我们使用语义分割图，而不是文献【27】中的 connected component maps。该框架实现的效果是：根据新加入的关键帧渐进地重建三维场景，场景中的每个 3D 元素都关联了一个语义类

别。在此之前，我们在训练 CNN 时使用了这些语义类别。

4 效果评估

在这一小节，我们对 tracking 的精度、三维重建的精度、纯旋转情况下的鲁棒性及语义标签融合进行评估。我们在实现 CNN-SLAM 时，CNN 网络的输入/输出分辨率为 304*228，但输入帧和预测的深度图的分辨率都先转换为 320*240，并将它们作为其余所有阶段的输入。基于 CNN 的深度预测和语义分割在 GPU 上运行，算法中的其余部分都在 CPU 上以两个线程运行，基于这种架构，CNN-SLAM 可实时运行。其中，一个线程用于帧级处理（相机位姿估计和深度优化），另一个线程用于关键帧相关处理（关键帧初始化、位姿图优化和全局语义标签融合）。

我们使用 ICL-NUIM（合成的）和 TUM RGB-D SLAM（由 Kinect 捕获）两个数据集测试。这两个数据集都提供了相机路径和深度图的 ground truth。在所有的实验中，我们都用 NYU Depth v2 数据集的室内场景训练 CNN，得到训练好的 CNN 模型，并用此模型测试该网络在从未见过的环境下的泛化性能(test the generalization capability of the network to unseen environments)。另外，NYU Depth v2 数据集包含了深度的 ground truth 和帧级的语义标签注释，而这恰恰是语义标签融合所必须的。特别需要说明的是，我们在官方的 train split of the labeled subset 上训练语义分割网络，而用原始的 NYU 数据集（里面的图片更多）训练深度网络，如文献【16】所述。语义注释由 4 个超级类(super-class)组成：floor, vertical structure, large structure/furniture, small structure. 值得说明的是，由于相机传感器、视角和场景不同，训练数据集的设置与评价 CNN-SLAM 的数据集设置有所不同。例如，NYU Depth v2 数据集包含了很多起居室、厨房和卧室的图片，而 TUM RGB-D SLAM 数据集中更多的是办公室场景，例如书桌、物体和人物。因此这些起居室、厨房和卧室等场景是 TUM RGB-D SLAM 数据集中没有的。

4.1 CNN-SLAM 与当前最高水平 SLAM 的对比

将 CNN-SLAM 与当前单目 SLAM 最高水平的两个框架：LSD-SLAM（直接法）和 ORB-SLAM（特征点法）作对比。为保持完整性，我们还和 REMODE（文献【23】，REMODE 是单目稠密深度图估计的当前最高水平）作对比。REMODE 可根据作者提供的 github 代码来实现。最后，我们也将 CNN-SLAM 和文献【16】的方法作对比。【16】将 CNN 预测的深度图作为基于深度的 SLAM 的输入（基于点的融合，文献【11,27】，当前基于深度 SLAM 的最高水平），其实现基于文献【27】提供的代码。

Given the ambiguity of monocular SLAM approaches to estimate absolute scale, we also evaluate LSD-SLAM by bootstrapping its initial scale using the ground-truth depth map, as done in the evaluation in [4, 20]. As for REMODE, since it requires as input the camera pose estimation at each frame, we use the trajectory and key-frames estimated by LSD-SLAM with bootstrapping.

LSD-SLAM: https://www.github.com/tum-vision/lsd_slam
ORB-SLAM2: https://www.github.com/raulmur/ORB_SLAM2
REMODE: https://www.github.com/uzh-rpg/rpg_open_remode
文献【27】: https://campar.in.tum.de/view/Chair/ProjectInSeg

按照文献【26】中提出的评价方法，表 1 列出了基于绝对轨迹误差(Absolute Trajectory

Error, ATE)的相机位姿准确性评估（计算方均根）。另外，我们通过计算深度误差小于 10%的点的比例，评估了重建准确性和重建密度。从表 1 可看出，CNN-SLAM 优于其他方法，而且三维重建的准确性、稠密性也大大提高。

深度图估计的准确性对比如图 4 所示。从图中可看出，CNN-SLAM 能明显改善 CNN 生成深度估计图中边缘模糊的情况，且对低纹理场景也能正常工作。

Table 1. Comparison in terms of Absolute Trajectory Error [m] and percentage of correctly estimated depth on ICL-NUIM and TUM datasets (TUM/seq1: *fr3/long_office_household*, TUM/seq2: *fr3/nostructure_texture_near_withloop*, TUM/seq3: *fr3/structure_texture_far*.

	Abs. Trajectory Error					**Perc. Correct Depth**					
	Our Method	LSD-BS [4]	LSD [4]	ORB [20]	Laina [16]	Our Method	LSD-BS [4]	LSD [4]	ORB [20]	Laina [16]	Remode [23]
ICL/office0	**0.266**	0.587	0.528	0.430	0.337	**19.410**	0.603	0.335	0.018	17.194	4.479
ICL/office1	**0.157**	0.790	0.768	0.780	0.218	**29.150**	4.759	0.038	0.023	20.838	3.132
ICL/office2	0.213	**0.172**	0.794	0.860	0.509	**37.226**	1.435	0.078	0.040	30.639	16.7081
ICL/living0	**0.196**	0.894	0.516	0.493	0.230	12.840	1.443	0.360	0.027	**15.008**	4.479
ICL/living1	**0.059**	0.540	0.480	0.129	0.060	**13.038**	3.030	0.057	0.021	11.449	2.427
ICL/living2	0.323	**0.211**	0.667	0.663	0.380	26.560	1.807	0.167	0.014	**33.010**	8.681
TUM/seq1	**0.542**	1.717	1.826	1.206	0.809	12.477	3.797	0.086	0.031	**12.982**	9.548
TUM/seq2	0.243	**0.106**	0.436	0.495	1.337	**24.077**	3.966	0.882	0.059	15.412	12.651
TUM/seq3	0.214	**0.037**	0.937	0.733	0.724	**27.396**	6.449	0.035	0.027	9.450	6.739
Avg.	**0.246**	0.562	0.772	0.643	0.512	**22.464**	3.032	0.226	0.029	18.452	7.649

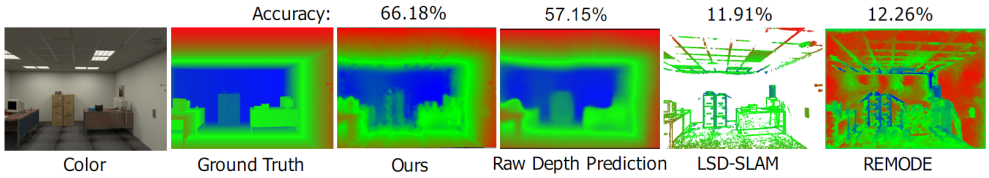

Figure 4. Comparison in terms of depth map accuracy and density among (from the left) the ground-truth, a refined key-frame from our approach, the corresponding raw depth prediction from the CNN, the refined key-frame from LSD-SLAM [4] with bootstrapping and estimated dense depth map from REMODE [23], on the (*office2*) sequence from the *ICL-NUIM* dataset [8]. The accuracy value means correctly estimated depth density on this key-frame.

4.2 纯旋转场景下的准确性评估

效果对比如图 5 所示。我们的方法可在纯旋转场景下正常工作，而 LSD-SLAM 会有显著的噪声；ORB-SLAM 根本无法工作，因为在初始化时若场景是纯旋转的，则无法获得初始化必须的 baseline，因此无法完成初始化。

Figure 5. Comparison on a sequence that includes mostly pure rotational camera motion between the reconstruction obtained by ground truth depth (left), proposed method (middle) and LSD-SLAM [4] (right).

4.3 三维重建场景和语义标签的融合

图 6 给出了 3 个融合的例子，绿色的是估计的相机轨迹。据我们所知，这是该领域第一个用单目相机，将三维重建场景和语义信息融合的实验。其他相关的实验结果和分析在补充材料中。

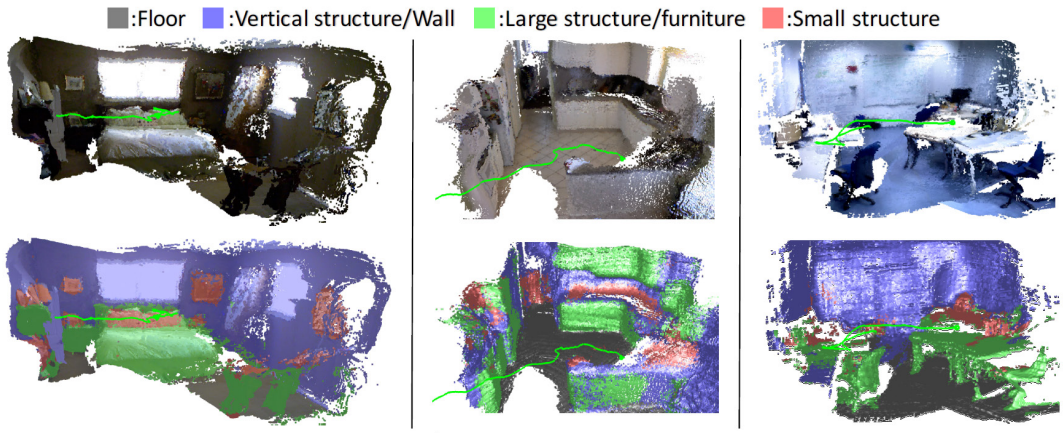

Figure 6. The results of reconstruction and semantic label fusion on the office sequence (top, acquire by our own) and one sequence (*kitchen_0046*) from the *NYU Depth V2* dataset [25] (bottom). Reconstruction is shown with colors (left) and with semantic labels (right).

5 总结

CNN-SLAM 为单目相机进行场景理解提供了新思路。今后的研究方向可以是通过深度预测进行回环检测，即通过几何方法优化深度图，提高深度估计的准确性。