DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion

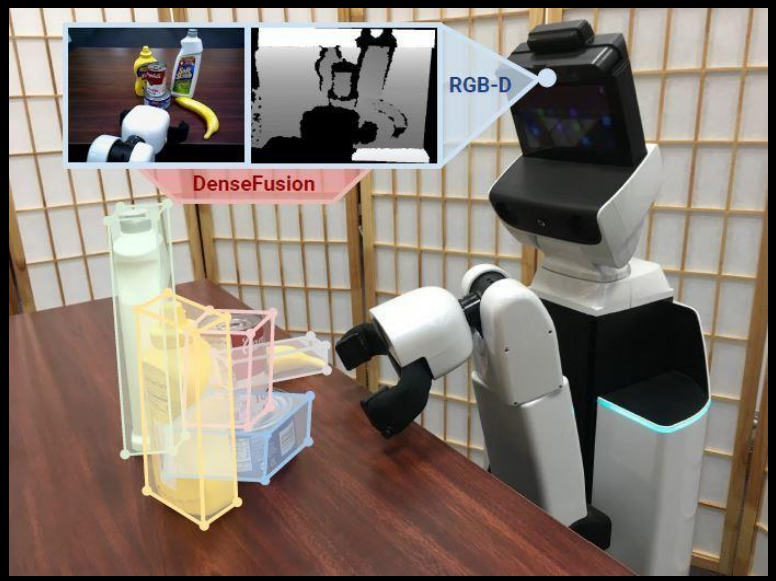
Chen Wang[2], Danfei Xu[1], Yuke Zhu[1], Roberto Martín-Martín[1], Cewu Lu[2], Li Fei-Fei[1], Silvio Savarese[1]

[1]Department of Computer Science, Stanford University

[2]Department of Computer Science, Shanghai Jiao Tong University

引言

- 6D目标姿态估计（旋转矩阵$\mathbf{R}$与平移向量$\mathbf{t}$）
  - 机器人抓取与操控
  - 自动驾驶
  - 增强现实
  - ……
- 主要问题
  - 各种各样的形状和纹理
  - 重度遮挡
  - 传感器噪声
  - 光照环境变化
  - 实时性的要求



相关工作

- Pose from RGB images
  - 传统方法
    - 依赖于检测并匹配已知物体上的关键点
  - 新兴方法
    - 使用学习来预测关键点并用PnP计算pose
  - 在缺少纹理或低分辨率的情况下不可靠
  - Pose estimation in 3D remains a challenge for the lack of depth information
- Pose from depth / point cloud
- Pose from RGB-D data

相关工作

- Pose from RGB images
- Pose from depth / point cloud
  - 应用于自动驾驶（KITTI数据集）
  - 在离散化的3D体素空间中处理——20s / 帧
  - 直接处理3D点云数据
    - 基于PointNet结构
    - 与仅需要点云数据的自动驾驶应用不同，普通物体的姿态估计需要结合几何与外观信息（2D-3D融合）
- Pose from RGB-D data

相关工作

- Pose from RGB images
- Pose from depth / point cloud
- Pose from RGB-D data
  - 传统方法
    - 从RGB-D数据中提取特征，进行特征分组和假设检验
    - 依赖于手动特征和固定的匹配过程
    - 在重度遮挡、灯光变化环境下表现不佳
  - PoseCNN
    - 需要高代价的后处理步骤
  - PointFusion

相关工作

- Pose from RGB images
- Pose from depth / point cloud
- Pose from RGB-D data
  - 传统方法
  - PoseCNN
  - PointFusion
    - 使用异构网络融合几何与外观信息
    - 有很好的实时性
    - 在重度遮挡环境下效果不好

主要贡献1

- 现有方法
  - 单独处理RGB图像与深度信息，没有合适的方法结合两种数据
- DenseFusion
  - 提出了一种稠密的融合网络结构，对颜色和深度信息在像素级进行了融合，得到更有利于姿态估计的color-depth feature embedding

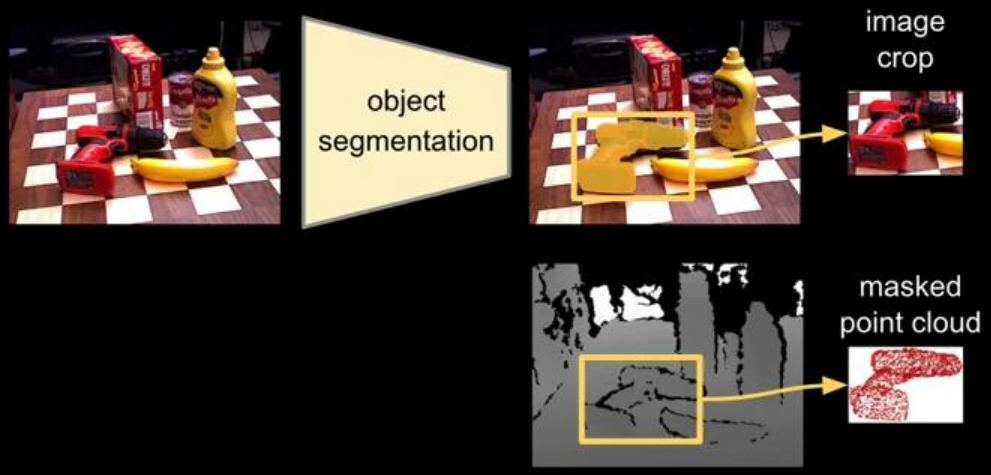
6D姿态估计模型

- 使用齐次变换矩阵$p \in SE(3)$表示目标的6D姿态

$$SO(3) = \{\boldsymbol{R} \in \mathbb{R}^{3\times3} | \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}, det(\boldsymbol{R}) = 1\}.$$

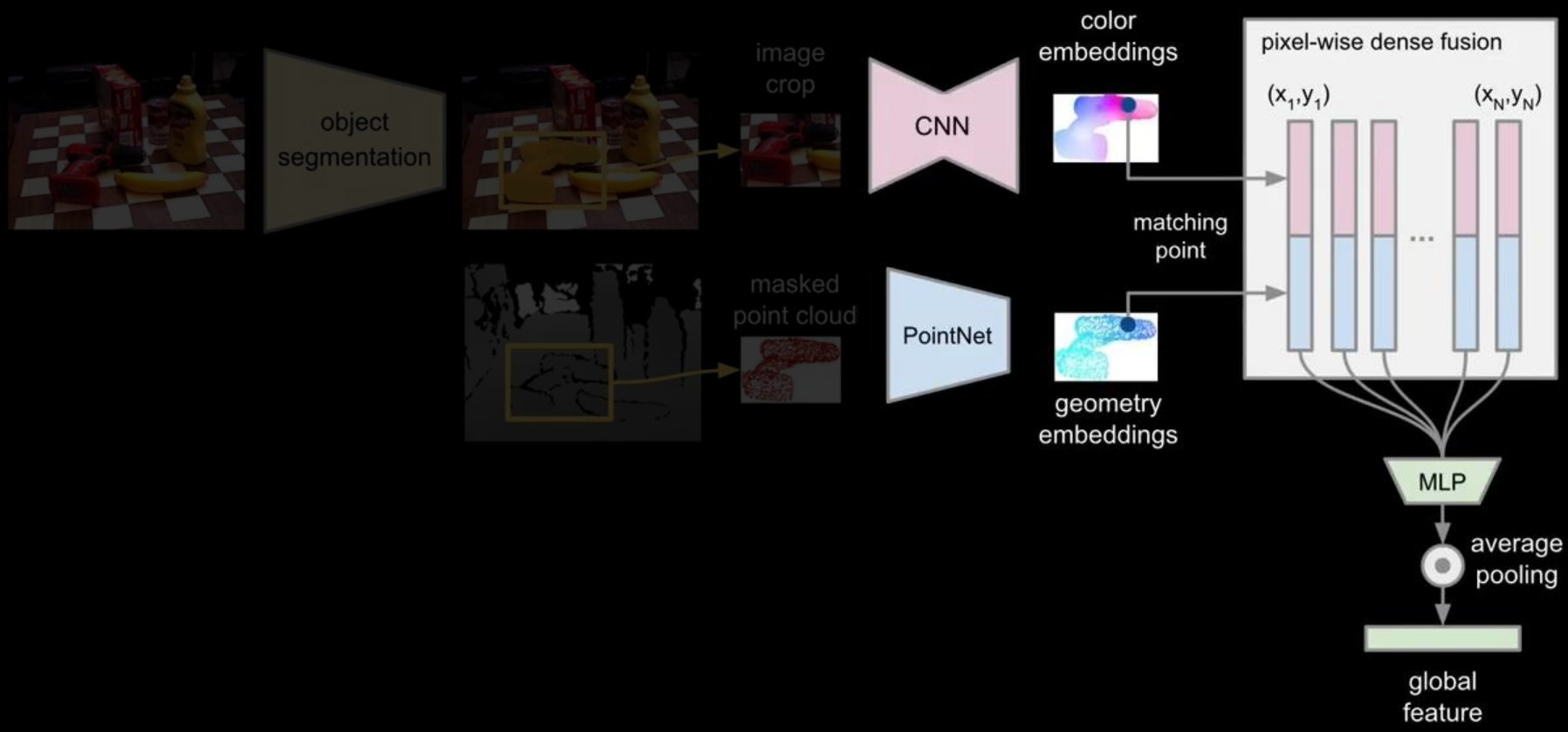
$$SE(3) = \left\{ \boldsymbol{T} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} | \boldsymbol{R} \in SO(3), \boldsymbol{t} \in \mathbb{R}^3 \right\}$$

6D姿态估计模型



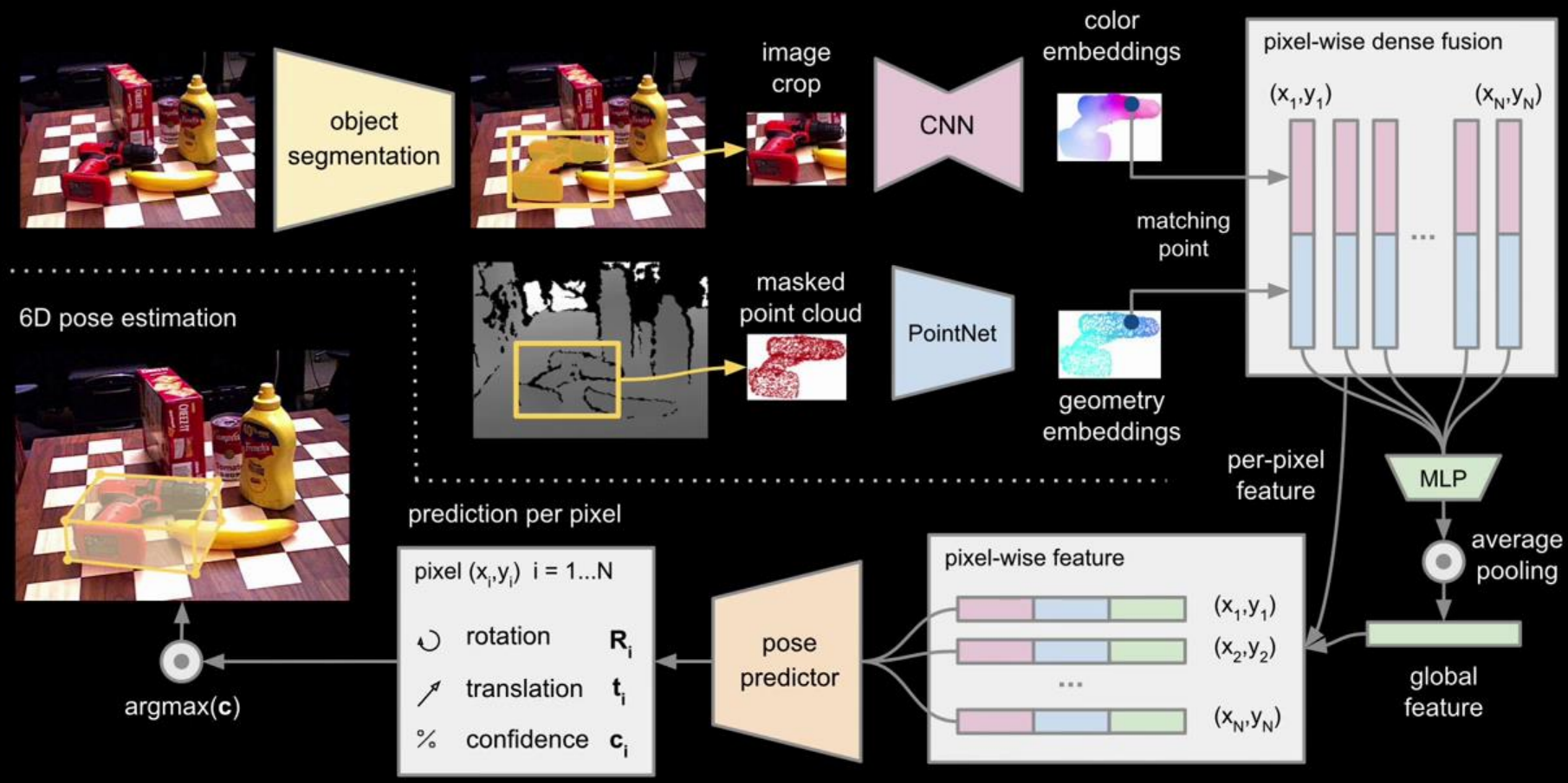
STEP1: Generate object segmentation masks and bounding boxes from RGB images. Process the masked depth into point cloud.

6D姿态估计模型



STEP2: The RGB image crop and point cloud are then encoded into **color** and **geometry** embeddings and fused at each **corresponding** pixel.

6D姿态估计模型



STEP3: The pose predictor produces a pose estimate for **each pixel** and the predictions are voted to generate the final 6D pose prediction of the object.

6D姿态估计模型

- 对于每个融合feature预测一个pose，损失函数：

$$L_i^p = \frac{1}{M} \sum_j ||(Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i)||$$

- 对于形状对称的物体：

$$L_i^p = \frac{1}{M} \sum_j \min_{0<k<M} ||(Rx_j + t) - (\hat{R}_i x_k + \hat{t}_i)||$$

6D姿态估计模型

- Naïve loss function:

$$L = \frac{1}{N} \sum_i L_i^p$$

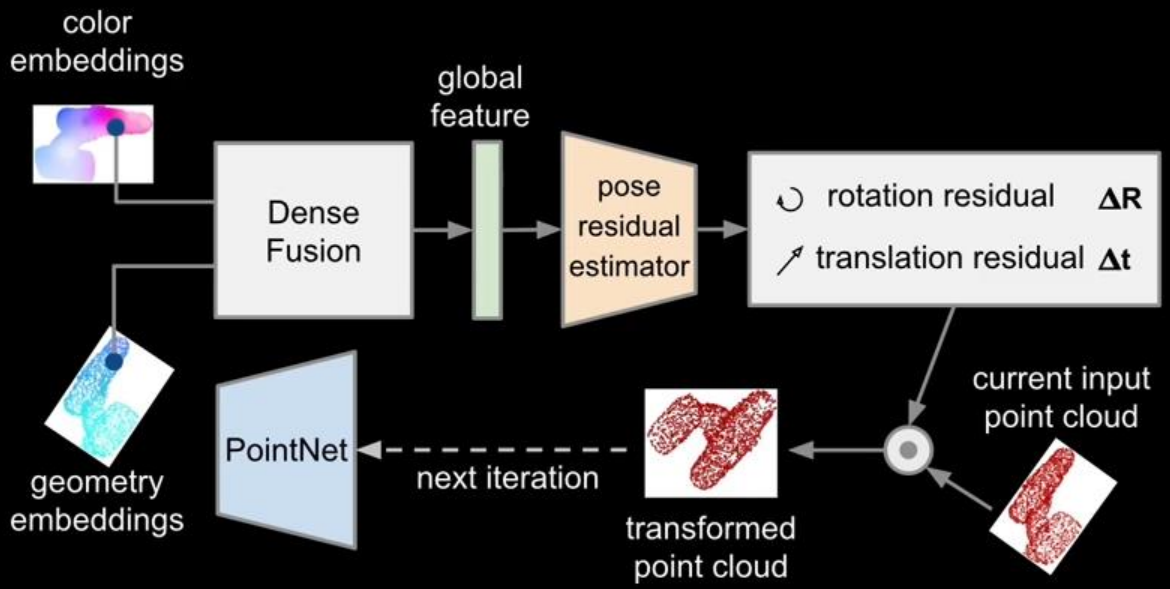
- With confidence regularization term:

$$L = \frac{1}{N} \sum_i (L_i^p c_i - w \log(c_i))$$

主要贡献2

- 现有方法
  - 需要耗时的pose refinement过程，限制了在实时应用中的表现
- DenseFusion
  - 提出了一种基于神经网络的迭代refinement过程，能够在保证实时性的条件下优化预测结果

Iterative Pose Refinement



STEP1: Consider the previously predicted pose as an **estimate of canonical frame** of the target object and transform the input point cloud into this frame.

STEP2: Feed the transformed point cloud back into the network and predict a **residual pose** based on the previously estimated pose.

Apply iteratively and generate potentially finer pose estimation each iteration.

实验结果

总结

- 设计了一种稠密(dense)的像素级融合方式，将RGB数据的特征和深度特征以一种更合适的方式进行了整合
- 设计了一种可与整体网络联合训练优化的refinement模块，可以代替传统方式中的离线后处理方法（如ICP）
- 提供了端到端的workflow
- 提高模型运行的实时性