文献阅读

20/06/18

Deep Hough Voting for 3D Object Detection in Point Clouds

Charles R. Qi[1] Or Litany[1] Kaiming He[1] Leonidas J. Guibas[1,2]

[1]Facebook AI Research [2]Stanford University

Introduction

- 3D 物体检测：estimate oriented 3D bounding boxes + semantic classes of objects from point clouds.
- 3D 点云：对光照更鲁棒； 几何信息更准确；
- 传统工作:
  - 将不规则的点云网格划分为规则点云； 利用3D CNN detector；
    3D卷积计算消耗大；无法处理稀疏点云
  - 将3D点云投应为2D俯视图；用3D detector定位；
    牺牲了几何细节
- 本篇工作：从原始数据3D点云直接进行3D检测

Method

- 由于3D点云的稀疏性，3D物体的中心可能和任何点都不相邻;基于点的网络很难在物体中心附近聚集场景上下文信息;
- 提出一种基于投票的机制，生成一个靠近物体中心的新的点集;利用这些点，可以汇总聚集获得box proposal

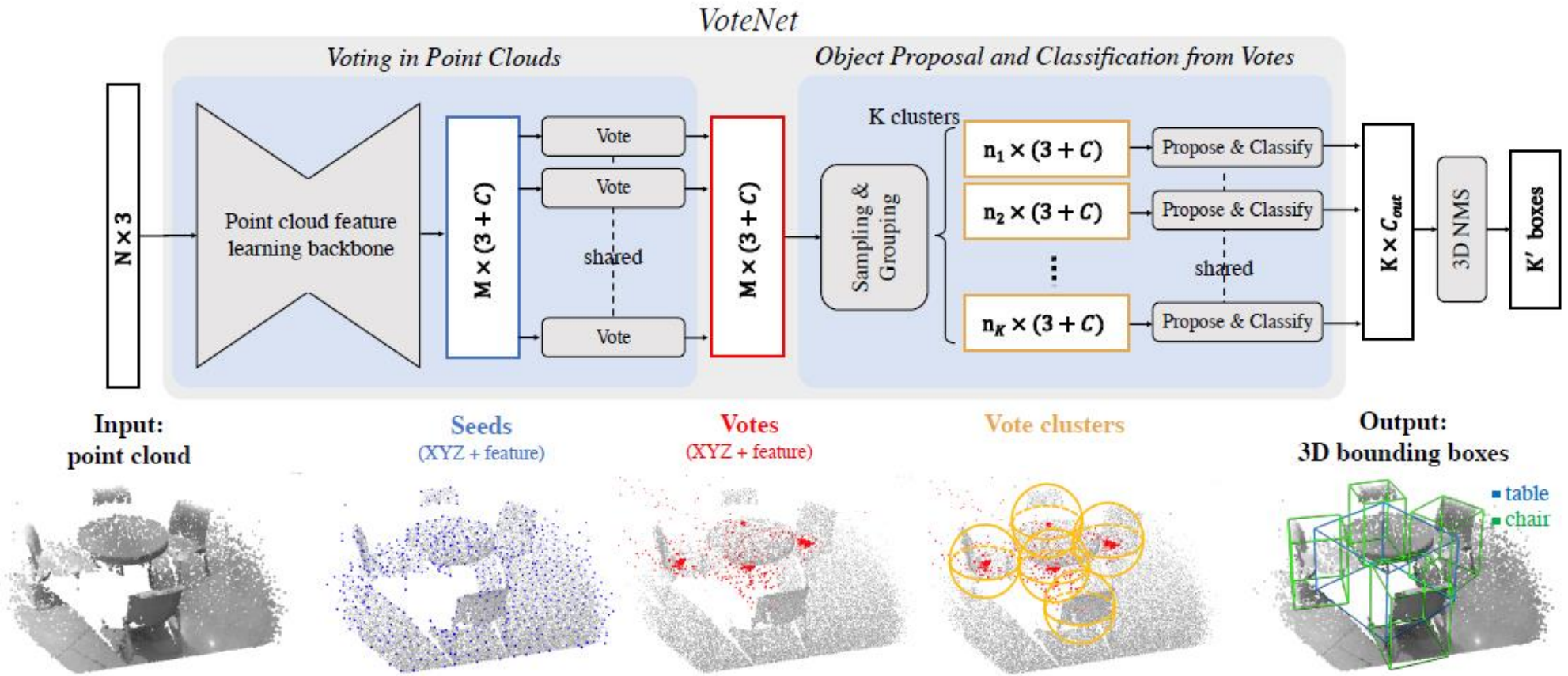
Method

- Deep Hough voting：
- 离线：建立codebook：image patches – offsets to object center
- 在线：从图像采样关键点并获得image patch；与codebook进行比较获得offset并计算投票

- Interest points, Vote, Vote aggregation：均可利用网络进行训练学习
- Object proposals：location, dimension, orientation. Semantic class

Method-VoteNet

- existing points生成投票 ---> virtual points (votes) 进行物体检测



Method-Learning to Vote in Point Clouds

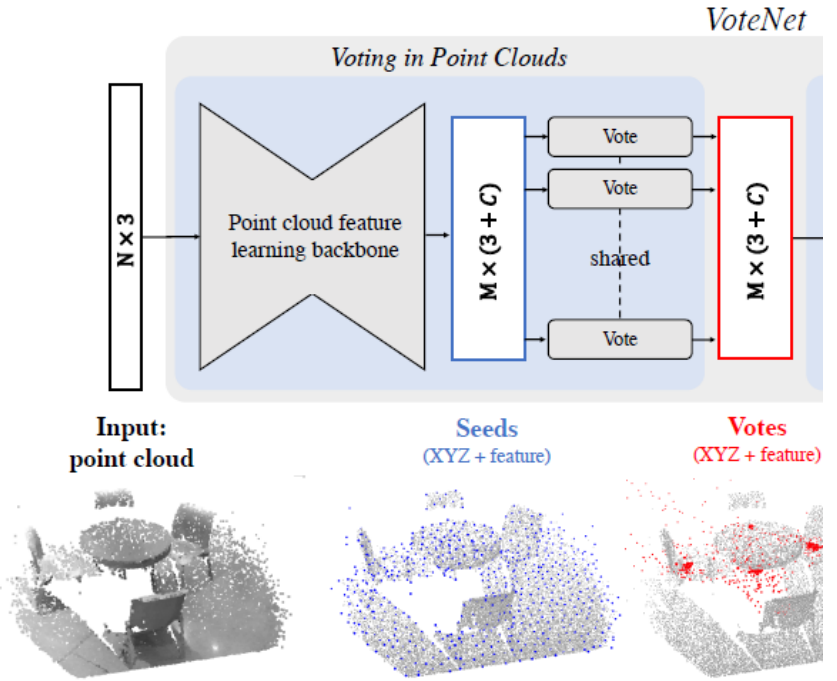
Input: N*3 point Output: M* (3+C) votes

Point cloud feature learning: PointNet++

Hough voting with deep networks: MLP+ FC

$$f_i \rightarrow \Delta x_i, \Delta f_i \qquad y_i = x_i + \Delta x_i \quad g_i = f_i + \Delta f$$

$$L_{\text{vote-reg}} = \frac{1}{M_{\text{pos}}} \sum_i \|\Delta x_i - \Delta x_i^*\| \mathbb{1}[s_i \text{ on object}],$$



Method-Object Proposal and Classification from Votes

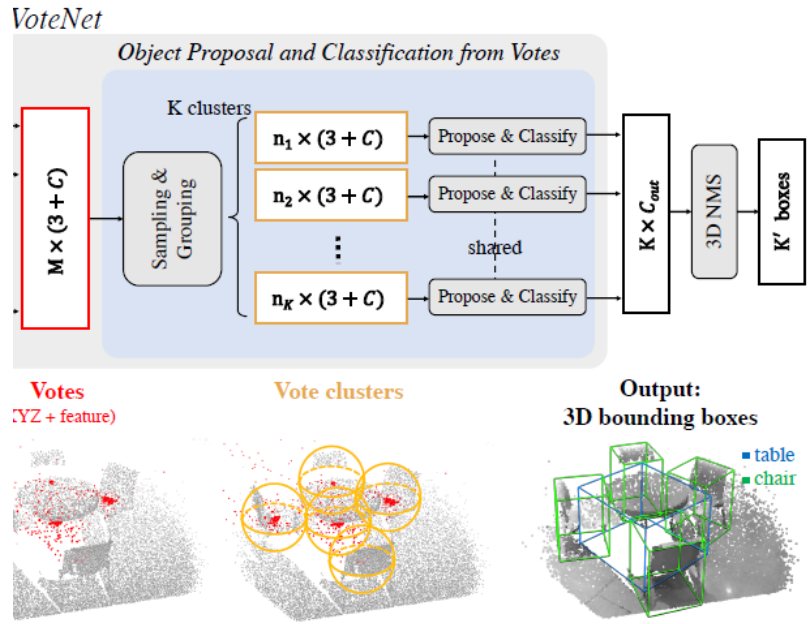
Vote clustering through sampling and grouping

- Farthest point sampling -> K cluster

Proposal and classification from vote clusters： shared PointNet

- For each vote cluster：

$$p(\mathcal{C}) = \mathrm{MLP}_2 \left\{ \max_{i=1,\ldots,n} \{\mathrm{MLP}_1([z_i'; h_i])\} \right\}$$



Method-Loss function

- voting loss + objectness loss + a 3D bounding box estimation loss + semantic classification loss

$$L_{\text{VoteNet}} = L_{\text{vote-reg}} + \lambda_1 L_{\text{obj-cls}} + \lambda_2 L_{\text{box}} + \lambda_3 L_{\text{sem-cls}}$$

- Objectness loss: cross-entropy loss of 2 classes
- Semantic classification loss: cross-entropy loss of NC classes
- Box loss: L1-smooth loss (center, size, heading)

Experiments

- Dataset： SUB RGB-D / ScanNetV2 3D reconstructed meshes

	Input	mAP@0.25	mAP@0.5
DSS [42, 12]	Geo + RGB	15.2	6.8
MRCNN 2D-3D [11, 12]	Geo + RGB	17.3	10.5
F-PointNet [34, 12]	Geo + RGB	19.8	10.8
GSPN [54]	Geo + RGB	30.6	17.7
3D-SIS [12]	Geo + 1 view	35.1	18.7
3D-SIS [12]	Geo + 3 views	36.6	19.0
3D-SIS [12]	Geo + 5 views	40.2	22.5
3D-SIS [12]	Geo only	25.4	14.6
VoteNet (ours)	Geo only	**58.6**	**33.5**

	Input	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS [42]	Geo + RGB	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [38]	Geo + RGB	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	**51.3**	70.1	47.6
2D-driven [20]	Geo + RGB	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet [34]	Geo + RGB	43.3	81.1	**33.3**	64.2	24.7	**32.0**	58.1	61.1	51.1	**90.9**	54.0
VoteNet (ours)	**Geo only**	**74.4**	**83.0**	28.8	**75.3**	22.0	29.8	**62.2**	**64.0**	47.3	90.1	**57.7**

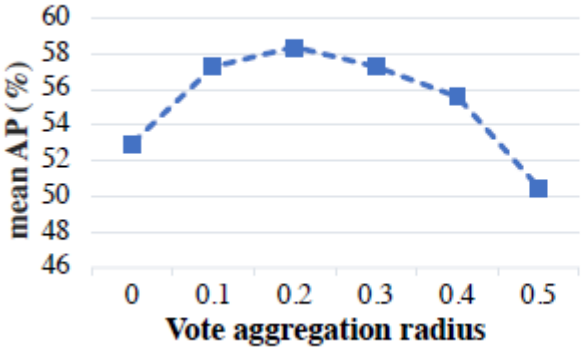
Table 1. **3D object detection results on SUN RGB-D val set.** Evaluation metric is average precision with 3D IoU threshold 0.25 as

Experiments

- To Vote or Not To Vote
  - BoxNet: directly proposes boxes from sampled scene points

Method	mAP@0.25	
	SUN RGB-D	ScanNet
BoxNet (ours)	53.0	45.4
VoteNet (ours)	**57.7**	**58.6**

- Effect of Vote Aggregation



Aggregation method	mAP
Feature avg.	47.2
Feature max	47.8
Feature RBF avg.	49.0
Pointnet (avg.)	56.5
Pointnet (max)	57.7

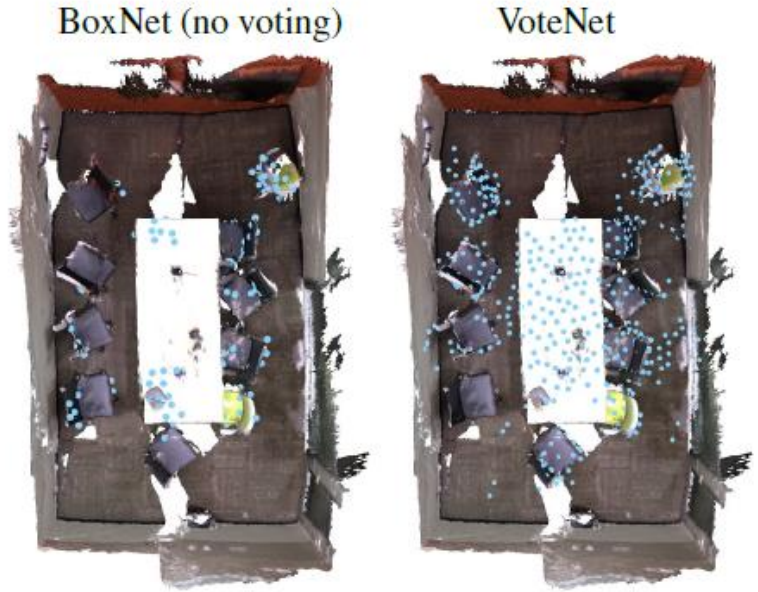


Figure 3. **Voting helps increase detection contexts.** Seed

PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation

Yisheng He[1] Wei Sun[2] Haibin Huang[3] Jianran Liu[2] Haoqiang Fan[2] Jian Sun[2]
[1]Hong Kong University of Science and Technology
[2]Megvii Inc. [3]Kuaishou Technology

Introduction

- Task：估计物体从世界坐标系到相机坐标系的刚性变换
- Challenge: 光照变化、场景遮挡
- 传统工作：
  - 直接回归旋转平移

  对非线性的旋转空间泛化能力较弱
  - 检测2D关键点，用PnP算法估计6D姿态

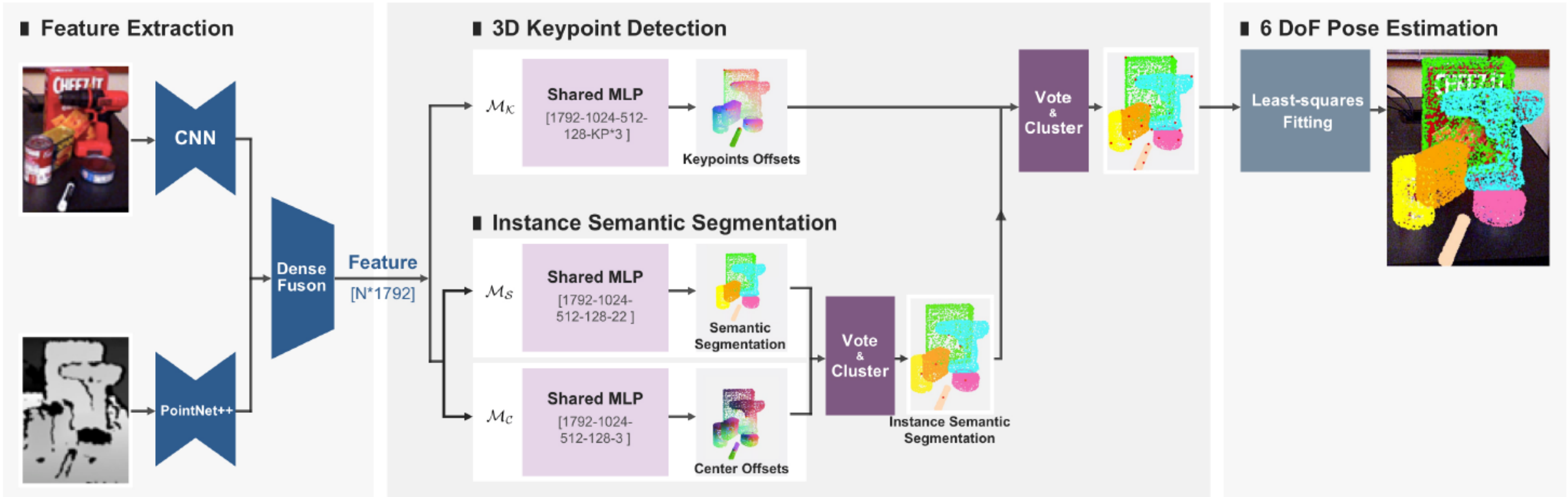
  投影上的小误差可能在3D空间误差较大
- 本篇工作：将基于2D关键点的方法延申到3D关键点，从而可利用物体的空间几何限制

Method

- 基于如下观察：3D空间重刚性物体的两个点的空间关系是固定的
  物体表面的一个点到关键点的偏移是固定且可学习
- 引入语义分割模块，处理多个物体同时存在的情况

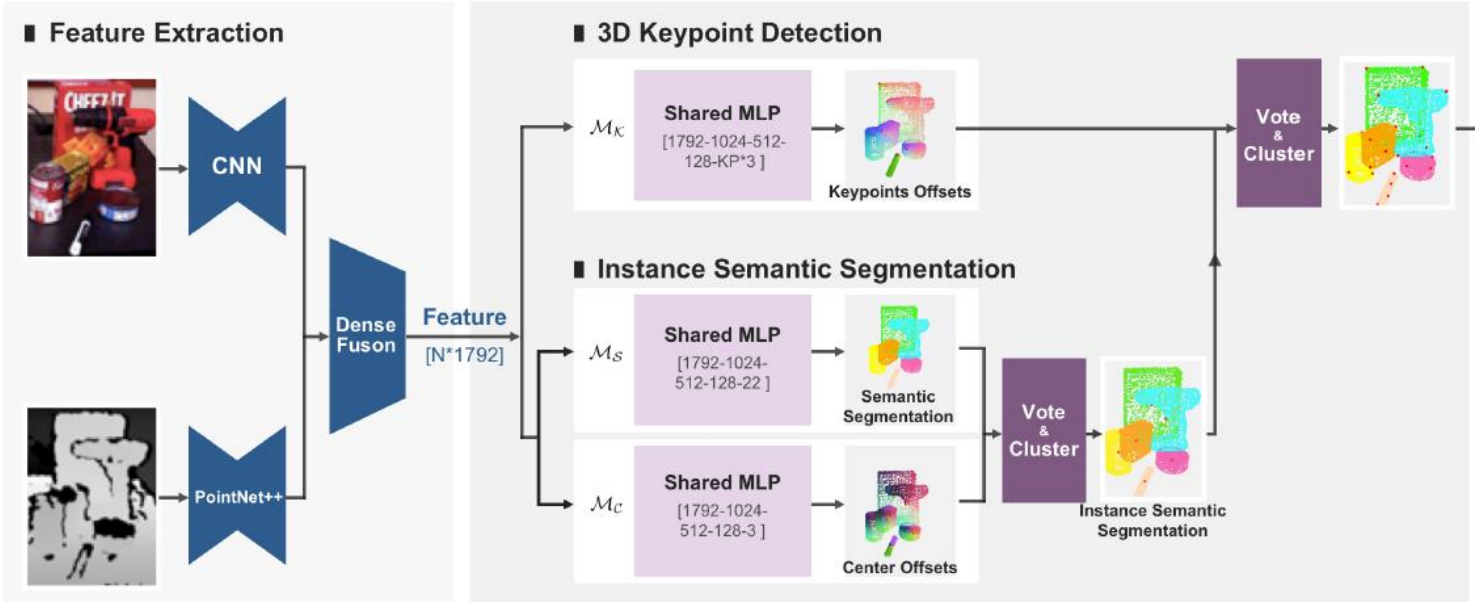
Method

- 3D Keypoint detection + pose parameter estimation



Method

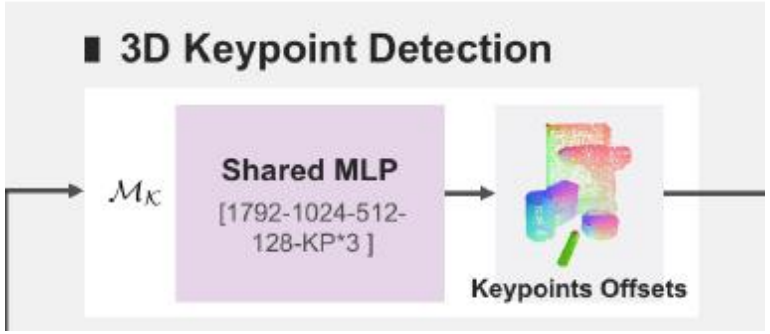
- Input: RGBD image
- 3D keypoint detection： 预测每个点到关键点的偏移
- Instance semantic segmentation： 预测每个点的语义标签
- Center voting module： 预测每个点到物体中心的偏移



Method- keypoint detection

- $f_i \rightarrow of_i^j \quad vkp_i^j = x_i + of_i^j$

$$L_{\text{keypoints}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}||of_i^j - of_i^{j*}||\mathbb{I}(p_i \in I)$$



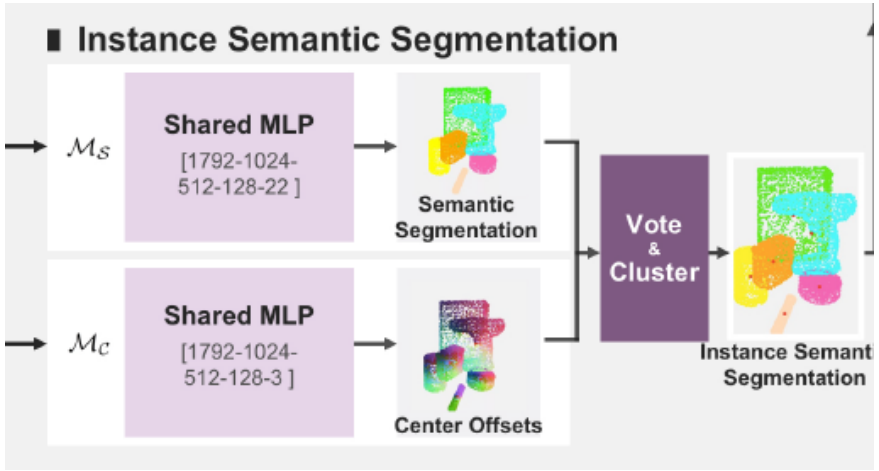
Method-Instance Semantic Segmentation

- Semantic:

$$L_{semantic} = -\alpha(1-q_i)^{\gamma} log(q_i)$$
$$where \quad q_i = c_i \cdot l_i$$

- Center voting: 区分不同的instance

$$L_{\text{center}} = \frac{1}{N}\sum_{i=1}^{N} ||\Delta x_i - \Delta x_i^*|| \mathbb{I}(p_i \in I)$$



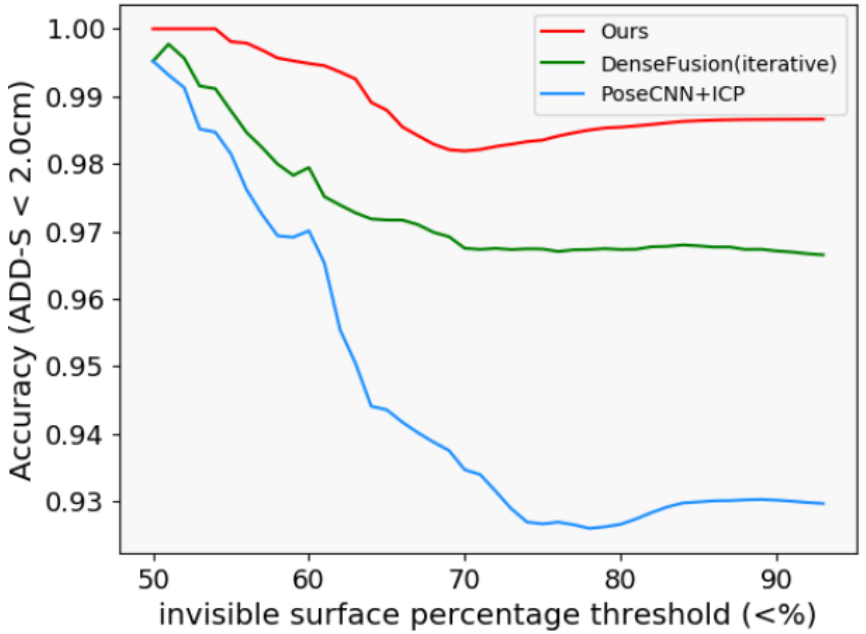
Experiments

- Dataset： YCB-Video Dataset / LineMOD dataset

	RGB			RGBD					
	PoseCNN DeepIM [26, 52]	PVNet [37]	CDPN [27]	Implicit ICP[45]	SSD-6D ICP[22]	Point-Fusion[50]	DF(per-pixel)[50]	DF(ite-rative)[50]	PVN3D
ape	77.0	43.6	64.4	20.6	65.0	70.4	79.5	92.3	**97.3**
benchvise	97.5	**99.9**	97.8	64.3	80.0	80.7	84.2	93.2	99.7
camera	93.5	86.9	91.7	63.2	78.0	60.8	76.5	94.4	**99.6**
can	96.5	95.5	95.9	76.1	86.0	61.1	86.6	93.1	**99.5**
cat	82.1	79.3	83.8	72.0	70.0	79.1	88.8	96.5	**99.8**
driller	95.0	96.4	96.2	41.6	73.0	47.3	77.7	87.0	**99.3**
duck	77.7	52.6	66.8	32.4	66.0	63.0	76.3	92.3	**98.2**
eggbox	97.1	99.2	99.7	98.6	**100.0**	99.9	99.9	99.8	99.8
glue	99.4	95.7	99.6	96.4	**100.0**	99.3	99.4	**100.0**	**100.0**
holepuncher	52.8	82.0	85.8	49.9	49.0	71.8	79.0	92.1	**99.9**
iron	98.3	98.9	97.9	63.1	78.0	83.2	92.1	97.0	**99.7**
lamp	97.5	99.3	97.9	91.7	73.0	62.3	92.3	95.3	**99.8**
phone	87.7	92.4	90.8	71.0	79.0	78.8	88.0	92.8	**99.5**
ALL	88.6	86.3	89.9	64.7	79.0	73.7	86.2	94.3	**99.4**

Experiments

Robust to Occlusion Scenes



Comparisons to Directly Regressing Pose

- 将3D keypoint voting module修改为直接回归每个点的旋转平移参数； 加入置信度估计；选择置信度最高的部分作为最终
- 和非线性的旋转空间相比，3D 关键点偏移的搜索空间更小

	DF(RT)[50]	DF(3D KP)[50]	Ours(RT)	Ours(2D KPC)	Ours(2D KP)	PVNet[37]	Ours(Corr)	Ours(3D KP)
ADD-S	92.2	93.1	92.8	78.2	81.8	-	92.8	**95.5**
ADD(S)	86.9	87.9	87.3	73.8	77.2	73.4	88.1	**91.8**

Table 4. Quantitative evaluation of 6D Poses on the YCB-Video dataset with different formulations. All with our predicted segmentation.

Experiments

Effect of 3D Keypoints Selection

- 8 corners of the 3D bounding box <-> points selected from the FPS algorithm

	VoteNet[38]	BBox 8	FPS 4	FPS 8	FPS 12
ADD-S	89.9	94.0	94.3	**95.5**	94.5
ADD(S)	85.1	90.2	90.5	**91.8**	90.7

Table 5. Effect of different keypoint selection methods of PVN3D. Results of VoteNet[38], another 3D bounding box detection approach are added as a simple baseline to compare with our BBox8.

- bounding box corners are virtual points that are far away from points on the object.
- Therefore, point-based networks are difficult to aggregate scene context in the vicinity of these virtual corner points.