CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

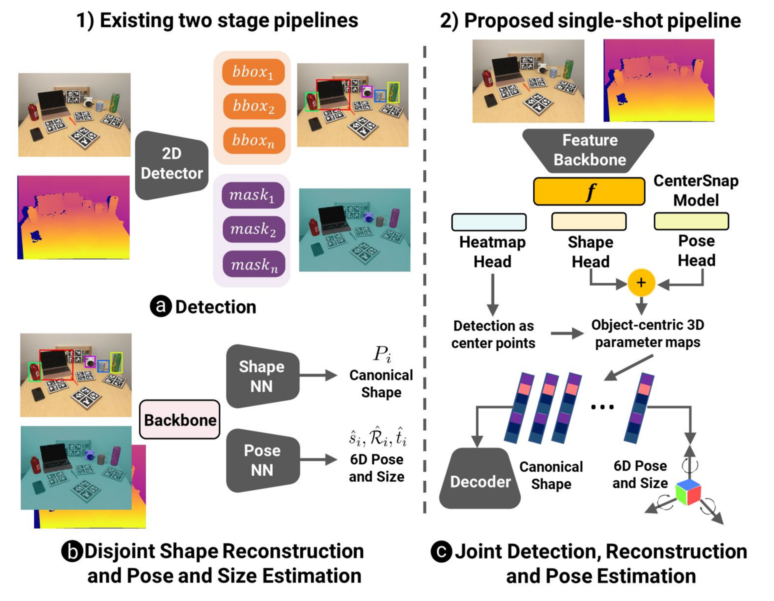
Georgia Tech

Muhammad Zubair Irshad Thomas Kollar Michael Lasky Kevin Stone Zsolt Kira

TOYOTA RESEARCH INSTITUTE

Motivation

- Joint Detection, Reconstruction and Pose Estimation
- Category-level 3D object understanding
- Applications: Robotics Grasping, Manipulation

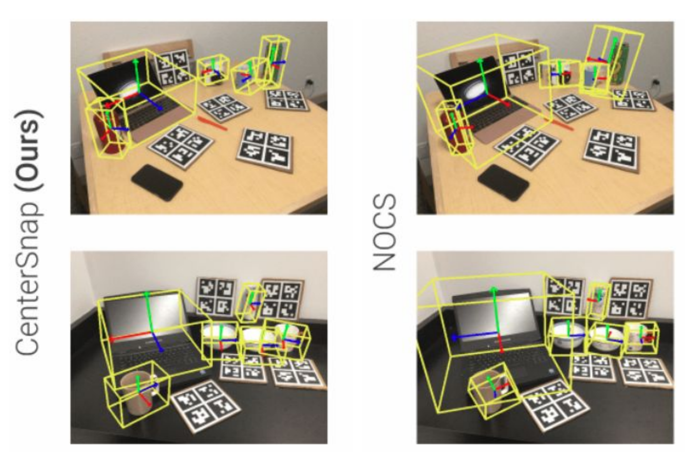


Given Input	$I \in \mathbb{R}^{h_o \times w_o \times 3}, D \in \mathbb{R}^{h_o \times w_o}$
Predict	$P \in \mathbb{R}^{K \times N \times 3}, \tilde{\mathcal{P}} \in SE(3), \hat{s} \in \mathbb{R}^3$

Overview

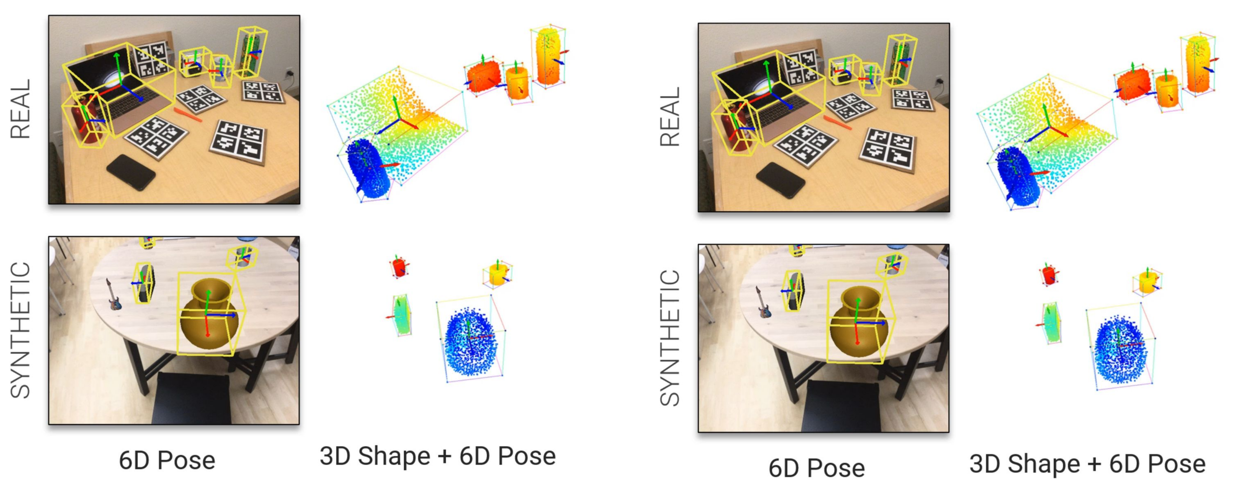
★ Prior Works...

- Computationally expensive
- Multi-stage pipelines [1,2]
- Not Scalable
- Low performance in challenging scenarios



★ Contributions

- Object-centric holistic scene-understanding
- Single-shot ***3D shape reconstruction*** and **6D pose and size estimation** from single-view RGB-D
- **Fast** joint reconstruction and pose estimation system. Our technique runs at **40 FPS**
- Over **12% improvement** in mAP for 6D pose
- Employ a **shape-prior** to learn from a large collection of CAD models

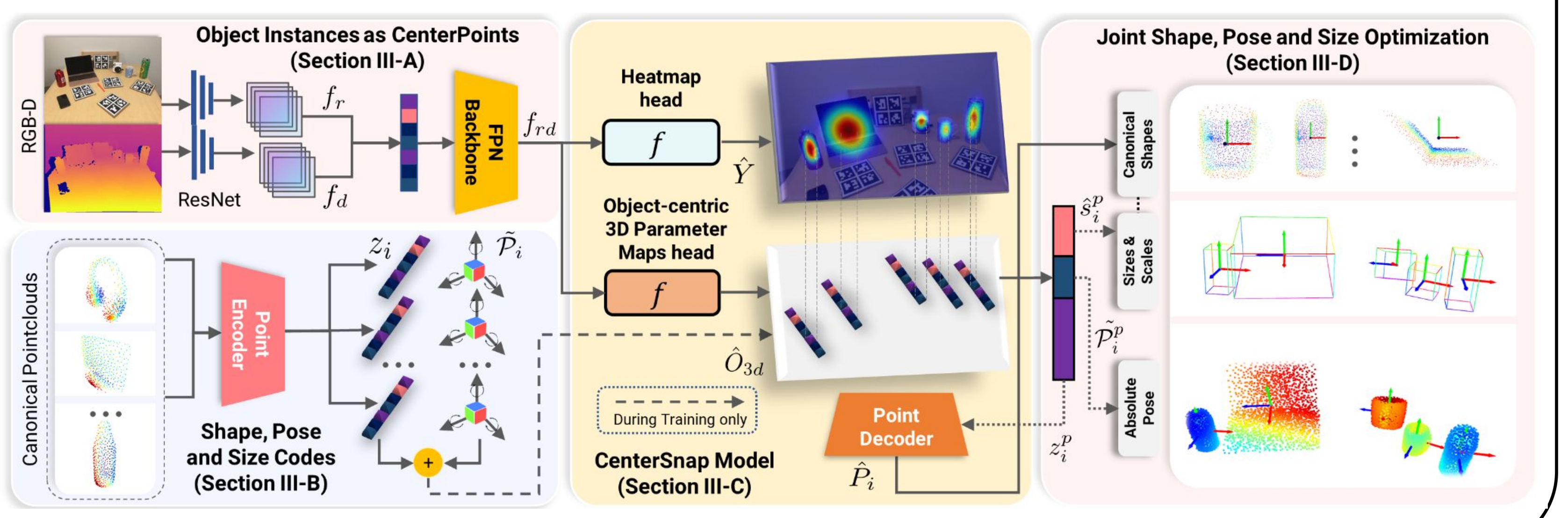


CenterSnap Architecture

★ We employ an end-to-end learnable pipeline

1. Objects instances are detected as **heatmaps** in a per-pixel manner
2. Joint **shape, pose, and size code** denoted is predicted for detected object centers using specialized heads
3. **Shape auto-encoder** pretraining on collection of CAD models
4. Jointly optimizing 3D and 2D heads to predict shapes, pose and sizes in a single-forward pass
5. Artificact-free depth prediction **aids sim2real transfer**

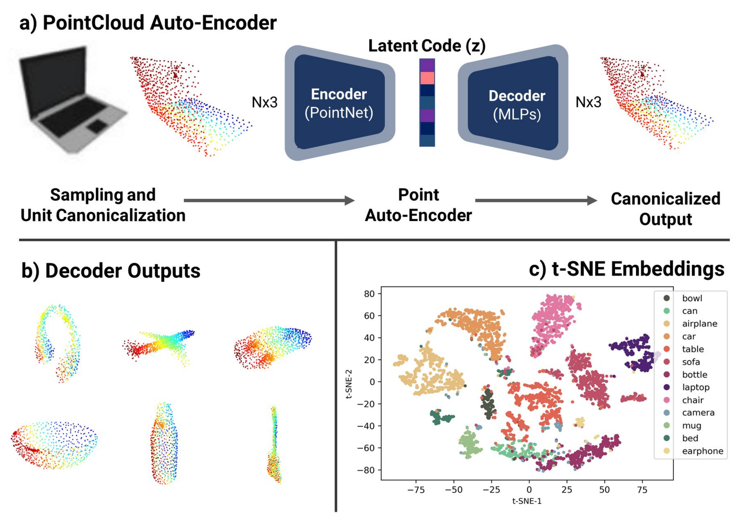
$$\mathcal{L} = \lambda_l \mathcal{L}_{\text{inst}} + \lambda_{O_{3d}} \mathcal{L}_{O_{3d}} + \lambda_d \mathcal{L}_D$$



Shape Prior

★ Shape, Pose and Size Codes

- Design an auto-encoder
- Encoder (**PointNet** [3]), Decoder (**MLP**)
- Learn a Shape-code (z)
- Shape-code space finds a distinctive 3D space for semantically similar objects



$$D_{cd}(\mathbf{P}_i, \hat{\mathbf{P}}_i) = \frac{1}{|\mathbf{P}_i|} \sum_{x \in \mathbf{P}_i} \min_{y \in \hat{\mathbf{P}}_i} \|x - y\|_2^2 + \frac{1}{|\hat{\mathbf{P}}_i|} \sum_{\mathbf{y} \in \hat{\mathbf{P}}_i} \min_{x \in \mathbf{P}_i} \|x - y\|_2^2$$

★ Inference

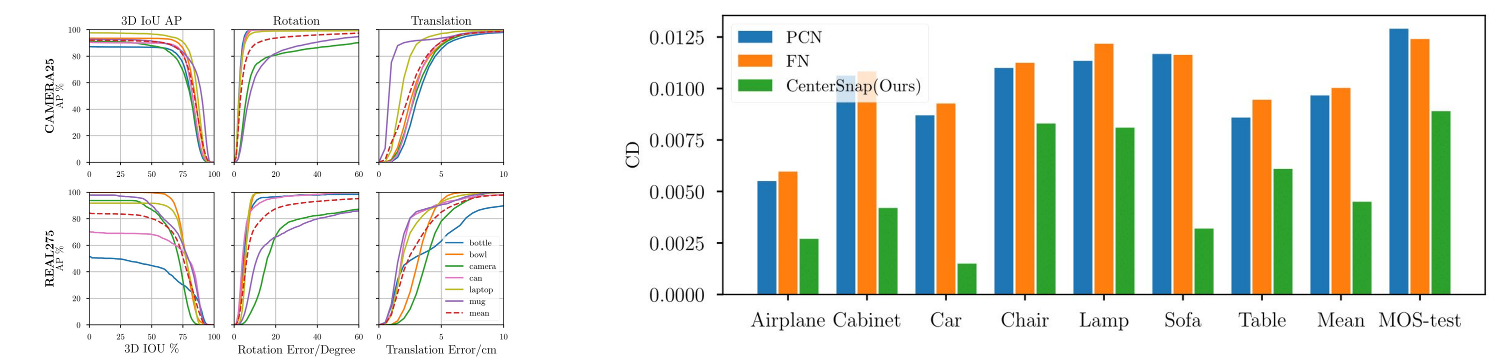
- Perform peak detection to get detection centerpoints
- Decode shape latent codes using frozen point decoder
- Decode pose by sampling directly in object-centric 3D maps

$$\hat{P}_i^{recon} = [\hat{\mathcal{R}}_i^p | \hat{t}_i^p] * \hat{s}_i^p * \hat{P}_i$$

Evaluation

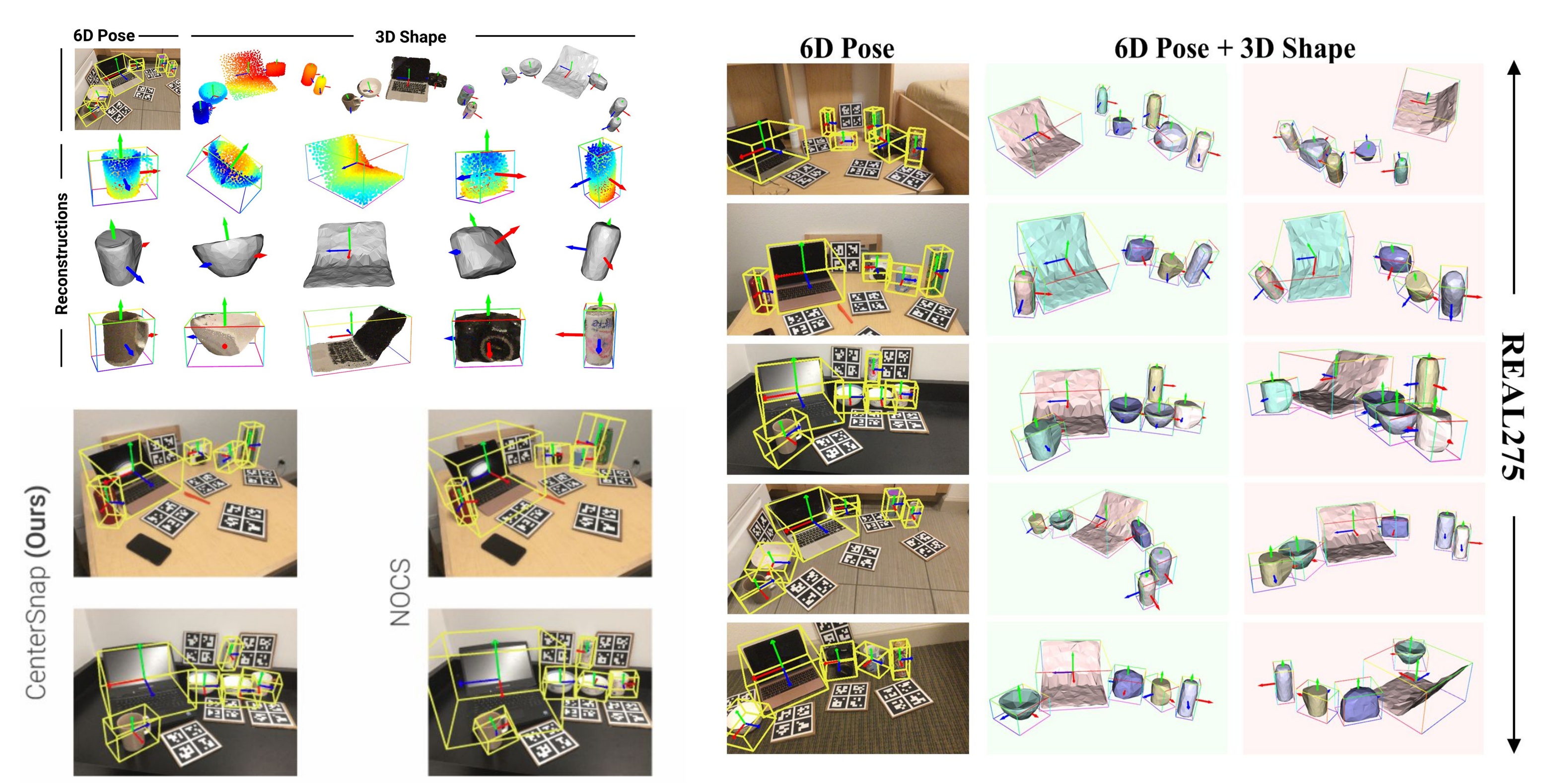
★ Metrics: $\boldsymbol{IOU}25, \boldsymbol{IOU}50, 5^\circ 5\ \mathbf{cm}, 5^\circ 10\ \mathbf{cm}$ and $10^\circ 10\ \mathbf{cm}$

		CAMERA25						REAL275					
	Method	IOU25	IOU50	5°5 cm	5°10 cm	10°5 cm	10°10 cm	IOU25	IOU50	5°5 cm	5°10 cm	10°5 cm	10°10 cm
1	NOCS [22]	91.1	83.9	40.9	38.6	64.6	65.1	**84.8**	78.0	10.0	9.8	25.2	25.8
2	Synthesis* [59]	-	-	-	-	-	-	-	-	0.9	1.4	2.4	5.5
3	Metric Scale [60]	**93.8**	90.7	20.2	28.2	55.4	58.9	81.6	68.1	5.3	5.5	24.7	26.5
4	ShapePrior [21]	81.6	72.4	59.0	59.6	81.0	81.3	81.2	77.3	21.4	21.4	54.1	54.1
5	CASS [44]	-	-	-	-	-	-	84.2	77.7	23.5	23.8	58.0	58.3
6	**CenterSnap (Ours)**	93.2	92.3	63.0	69.5	79.5	87.9	83.5	80.2	27.2	29.2	58.8	64.4
7	**CenterSnap-R (Ours)**	93.2	**92.5**	**66.2**	**71.7**	**81.3**	**87.9**	83.5	**80.2**	**29.1**	**31.6**	**64.3**	**70.9**



Qualitative Results

★ Qualitative pose estimation and shape reconstruction



★ Shape Reconstruction with Texture

★ Future Work

- Articulated Objects, Articulated scene reconstruction
- Category-Level Real World Manipulation
- Various shape representations i.e. SDF, NeRFs

ICRA 2022 IEEE International Conference on Robotics and Automation

Available Material

Project Webpage: https://zubair-irshad.github.io/projects/CenterSnap.html

CenterSnap Github: https://github.com/zubair-irshad/CenterSna

Short Video: https://youtu.be/Bg5vi6DSMdM

References

[1] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," CVPR, 2019

[2] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in European Conference on Computer Vision. Springer, 2020

[3] Charles R. Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017