Object Instance Segmentation and 3D Reconstruction for Grasping in Cluttered Scenes

Sharath Chandan Reddy Patlolla
School of Computing
University of Utah
Salt Lake City, UT 84111
sharath.patlolla@utah.edu

Dr. Tucker Hermans
School of Computing
University of Utah
Salt Lake City, UT 84111
thermans@cs.utah.edu

Project Committee: Dr. Tucker Hermans, Dr. John Hollerbach, Dr. Alan Kuntz

Abstract:
This research focuses on extending the use case of Continuous 3D reconstruction using Point Sign Distance Function (SDF) to include cluttered scenes and enable robot grasping in unseen multi-object environments. PointSDF doesn't work for cluttered scenes because the occluded object point cloud disables reconstruction. This is solved by leveraging current research on cluttered scene segmentation to get object instance masks and process them to get the individual object partial-view point clouds. The two networks are merged by adding an intermediate processor to extract object instance point clouds. Additionally, RANSAC is added to the network for plane fitting to reconstruct the tabletop to improve the reconstruction in the further scope of the research. The merged pipeline is examined with a real-world dataset and also simulated datapoint which assisted in identifying the limitations and increased the further scope.

I. INTRODUCTION

In recent years, deep learning has become a dominant tool in computer vision which often produces comparable or better results for analytically solvable tasks and its performance is significantly higher for heuristic or stochastic tasks [1]. Much of this recognition is due to the breakthroughs achieved by Convolutional Neural Net (CNN) in solving image recognition challenges and surpassing the state of art by significant metrics on datasets like ImageNet [2] and CIFAR [3]. Object detection in unseen scenes is one of the unsolved tasks which has a wide range of applications in self-driving cars, robotics, the medical field, etc. [1]. Object detection is an essential feature for robots to attain full autonomy because numerous robots are being deployed each day, for various tasks like drones, industrial robots, etc. [1]. The ability of a robot to accurately identify new objects and approximate their geometry is a crucial task to enable robot interaction like grasping in an unseen setting.

Robot grasping has been studied by various researchers [4] [5] and is partially solved for some settings such as single object scene [6]. This project focuses on studying the existing approaches and investigating new strategies to develop the code for extending the reconstruction research to work for cluttered scenes. The key limitation in the current reconstruction research [6] is that the continuous PointSDF 3D reconstruction is nonfunctional for multiple objects, which is the most common use case. On the other hand, cluttered scene segmentation is widely researched with the state of the art being able to satisfy most use cases [7].

Unseen Object Instance Segmentation (UOIS) research [8] is very effective in solving the cluttered scene problem. This project uses UOIS architecture to segment the cluttered scene and extracts individual object point cloud data from the cluttered scene. Continuous 3D reconstruction[6] works for single object partial-view point clouds and this project focuses on leveraging both these architectures to model a network with use cases including cluttered scenes. The primary contribution of this project is to combine both architectures by investigating their use cases and limitations. It also aims to develop the code for Segmented Mask processing to extract point clouds and RANSAC plane estimation to reconstruct tabletop surfaces to further the project scope.

II. RELATED WORK

Implicit or explicit geometric reasoning of the object is the first step in grasping it, but research based on indirect reasoning [5] cannot fully solve the grasping problem because they cannot detect object collisions. Continuous 3D reconstruction using Signed Distance Function (SDF) [6] enables explicit geometric reasoning. SDF 3D reconstruction learns a grasp success classifier by leveraging the reconstruction network and the classifier is used to minimize error in grasping by acting as an objective function for continuous grasp optimization. However, the

continuous 3D reconstruction fails to generalize for multiple object or cluttered scenes because of the occluded point clouds. Object instance segmentation [8] fits perfectly into this pipeline to extract individual objects so that the reconstruction can be performed on single object partial-view point clouds.

Object segmentation can be of two types, one is semantic segmentation, and the other is instance segmentation. In semantic segmentation, identical objects are classified into one group whereas in instance segmentation every object is classified individually even if they are identical. Although semantic segmentation is widely researched [9] [10] [11], it is not particularly useful for the case of instance segmentation because semantic labels don't generalize well for unseen objects. On the other hand, the current research in Object instance segmentation [7] [8] [12] is very useful because individual object reconstructions are relatively accurate compared to multi-object reconstruction. UOIS uses the depth image to produce instance center votes and initial mask of the cluttered scene and the RGBD image to refine the masks. UOIS implementation trains the network on synthetic data [8] and contrary to expected behavior, it generalizes well for the real world images. This expectation is because RGB values of synthetic dataset are unrealistic and the contradictory result is because of the added noise in synthetic data and UOIS only uses RGB for refining the masks which is an easy problem compared to defining the masks.

III. METHOD OVERVIEW

A. Dataset:

PointSDF reconstruction network is trained with synthetic data which contains 590 meshes from the Grasp Database [13] and 76 meshes from the YCB Database [14] at 200 random orientations each, adding noise to the depth images to reflect Kinect noise.

In the UOIS Network, the model is trained with Google Open Image Dataset (OID) [15] and synthetic data from [8]. Simulated Noise is added to the synthetic dataset to approximate it to the real world and since the dataset isn't yet publicly available, pre-trained weights from [7] were used.

B. SDF Reconstruction:

This project essentially combines the continuous 3D reconstruction using SDF's and object instance segmentation using UOIS network. Signed Distance Function is specific to a set and its value is determined for a point where the value is positive or negative depending on the displacement between the set and the point. In this implementation, the SDF value is negative if a point is inside the set and positive if it's outside the space which means for points that form the boundary of the set, SDF is zero.

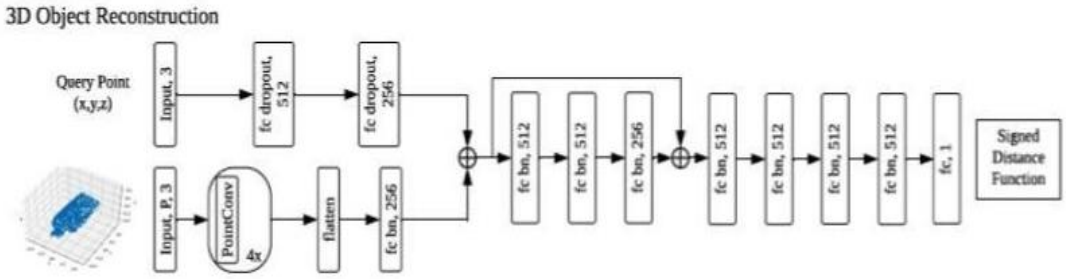

Figure 1:3D Reconstruction Network from [6] that takes a point cloud as input and answers SDF queries for points on the reconstructed object surface

Given a point cloud view of the object, o, and a query point in 3D space, x, the PointSDF function [6] is defined as,

$$f_{SDF}(o, x, \theta) = s;\ o \in R^{Px3}, x \in R^3, s \in R \qquad \text{--- (1)}$$

Here, θ represents the reconstruction network hyperparameters. It is to be noted here that the set is the object itself, which means the boundary points are the outer surface of the fully reconstructed object. The network is trained with the query point as input, true SDF values as expected output and the mean-squared error between predicted SDF and true SDF is minimized. The network diagram in Fig. 1 is taken from [6] but only the reconstruction part of the network that is specific to this project is mentioned.

The key feature of reconstruction with SDF function is the learned classifier for grasp success prediction, which is also the objective function for continuous grasp optimization. The helps in enabling collision avoidance in grasp planning.

C. UOIS Network:

UOIS architecture [8] separately leverages synthetic depth in Depth Seeding Network DSN and synthetic RGB in Region Refinement Network (RRN) for unseen object instance segmentation. DSN builds masks by reasoning in 3D space and introduces a novel loss function, separation loss which significantly improves accuracy in cluttered scenes. The DSN is essentially trained on four loss functions, which are Foreground loss, Center Offset Loss, Clustering Loss

and Separation Loss. The loss functions are briefly discussed to add context to the merged network,

Foreground loss is the weighted cross entropy which can effectively detect object boundaries in imbalanced images [16]

$$l_{fg} = \sum_i w_i l_{ce}\,(F_i, \overline{F_i}) \qquad \text{--- (2)}$$

Where $F_i, \overline{F_i}$ are predicted and ground truth probabilities and l_{ce} is cross-entropy loss, while i ranges through all the pixels

Center Offset Loss is the Huber loss ρ or the smooth L1 loss that is applied to the center offsets V' to minimize the error between ground truth object centers and center votes.

$$l_{co} = \sum_{i \epsilon \Omega} w_i \rho\,(D_i + V_i' - c_i) \qquad \text{--- (3)}$$

Where $D_i + V_i'$ is the center votes, c_i is the 3D coordinate of the ground truth object center for all pixels i. *Clustering Loss*, l_{cl} unrolls Gauss Mean Shift clustering every few iterations and its specifics are mentioned in [8]

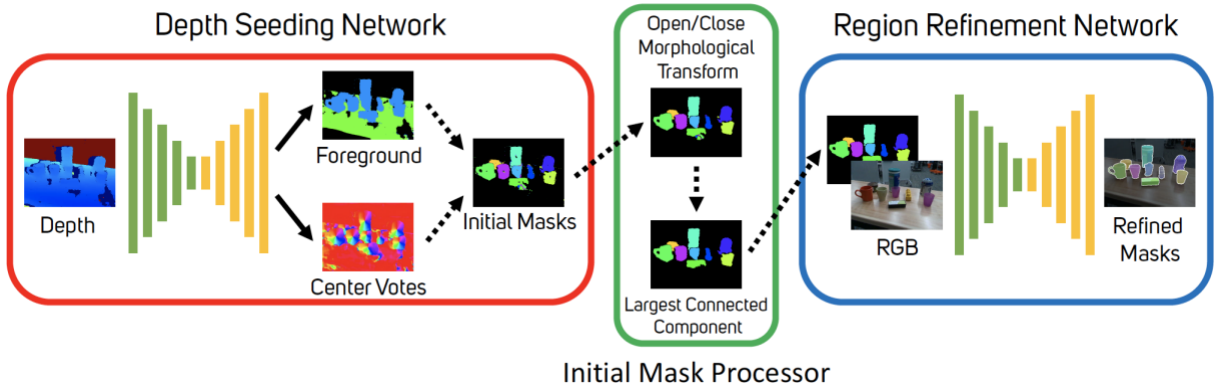

Figure 2: Overall UOIS Architecture taken from [7] where the red box is DSN and the blue box is RNN. The green box in between is the Initial Mask Processor

Separation Loss is the novel loss that was introduced to allow the center votes to be at locations other than the object center as long as it is far away from other object center votes. Separation loss if formulated as tensor M_{ij}, maximization problem given by

$$M_{ij} = \frac{\exp\left(-\tau\, d(c_j, D_i + V_i')\right)}{\sum_{j'=1}^{J} \exp\left(-\tau\, d(c_{j'}, D_i + V_i')\right)} \qquad \text{--- (3)}$$

Where c_j is the j[th] ground truth object center, $i \epsilon$ Ò and $\tau > 0$ is the hyperparameter. The total loss used in the DSN is the sum of all four losses is

$$L = \lambda_{fg} l_{fg} + \lambda_{co} l_{co} + \lambda_{cl} l_{cl} + \lambda_{sep} l_{sep} \qquad \text{--- (4)}$$

After DSN, an open and close operation which are mask erosion and dilation respectively, are applied to remove the salt and pepper noise and close small holes in the mask. Finally, RGB is used in RRN to output refined segmentation masks.

D. Merged Network:

The output of the UOIS network is processed with standard image processing techniques and individual objects point clouds are obtained. The point clouds are sequentially passed through the 3D reconstruction network to obtain the fully reconstructed objects. Although these reconstructed objects mostly resemble the true object, they are not perfect, but it's shown in [6] that the grasping works with a high rate of success, even for the imperfect reconstructions. The network pipeline is shown below,

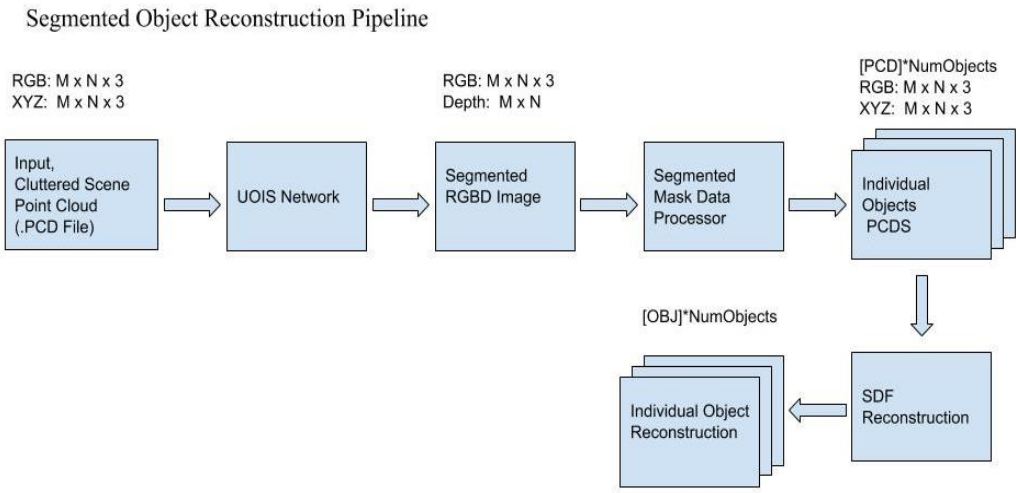

Figure 3:Merged UOIS and 3D Reconstruction Pipeline which is called the Segmented Object Reconstruction Pipeline because the object instance point clouds are initially segmented and passed to PointSDF 3D reconstruction to get object reconstructions

Reconstruction performance can be improved by leveraging the physical shape stability of the segmented objects in the scene. The project is currently at this phase and the first step here would be to recreate a table surface where the objects can be assembled by estimating their poses. The table surface can be estimated by finding the segmenting plane using analytical methods like least square or RANSAC [17] Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers when outliers are to be accorded no influence on the values of the estimates

Segmenting plane found using least-squares would be prone to outliers which is very common for the segmented point clouds. So, RANSAC which is also known as the outlier detection method [17] is used. The cartesian equation of a plane is,

$$ax + by + cz + d = 0 \quad \text{--- (5)}$$

The above equation has three unknowns and for this reason, RANSAC starts with three random points and calculates the coefficients a, b, c along with intercept d using those points. The number of inliers I, are estimated by setting a threshold, T_H and checking if the deviation for a point (x_p, y_p, z_p) lies below T_H,

$$\left(\delta = \frac{ax_p + by_p + cz_p + d}{\sqrt{a^2+b^2+c^2}} < T_H\right) \epsilon\, I \quad \text{--- (6)}$$

This iterative process stops after a set number of iterations, which is 1000 for the results in the first row, third and fourth column of Figure 4 and Figure 5 respectively

IV. RESULTS

A. Metrics:

The metrics defined in [18] are used to evaluate the UOIS network classification accuracy. Precision/Recall/F-measure are the common classification metrics which are calculated by recording False Positives (FP), True Positives (TP), False Negatives (FN), True Negatives (TN),

$$Precision, P = \frac{TP}{TP+FP};\quad Recall, R = \frac{TP}{TP+FN};\quad F1 = 2 * \frac{R*P}{R+P} \quad \text{--- (7)}$$

Although the idea of P, R, F1 remains the same, these metrics are calculated using the Hungarian method. In the Hungarian method, these measures are computed between all pairs of predicted objects and ground truth objects. Then the final P/R/F is calculated for the boundary and overlap condition [18]

$$Overlap\ P = \frac{\sum_i |c_i \cap g(c_i)|}{\sum_i |c_i|};\ Overlap\ R = \frac{\sum_i |c_i \cap g(c_i)|}{\sum_i |g_i|};\ Overlap\ R = \frac{2PR}{P+R} \quad \text{--- (8)}$$

$$Bound.P = \frac{\sum_i |c_i \cap D[g(c_i)]|}{\sum_i |c_i|}; Bound.R = \frac{\sum_i |D[c_i] \cap g(c_i)|}{\sum_i |g_i|}; Bound.R = \frac{2PR}{P+R} \quad \text{--- (9)}$$

where c_i denotes the set of pixels belonging to predicted object i, $g(c_i)$ is the set of pixels of the matched ground-truth object of c_i, and g_j is the set of pixels for ground truth object j. D[.] denotes the dilation operation, which allows for minor errors in prediction [7]

Training Dataset	OCID Dataset - Overlap			OCID Dataset - Boundary		
	P	R	F	P	R	F
OID [15]	48.5	41.2	44.6	60.6	43.9	50.9

Table 1: Average Overlap and Boundary P/R/F evaluation metrics for OCID dataset containing 1780 cluttered scenes when tested on the Project UOIS implementation

Training Dataset	OCID Dataset - Overlap			OCID Dataset - Boundary		
	P	R	F	P	R	F
OID [15]	87.9	79.6	81.7	84.0	69.1	74.1

Table 2: Average Overlap and Boundary P/R/F evaluation metrics for OCID dataset containing 1780 cluttered scenes when tested on the Original UOIS implementation

The P/R/F measures are in the range [0; 100] (P/R/F X 100). The difference in the metric evaluations in Tables 1 and 2 for the same dataset between the project results and the original results is drastic is assumed to be caused by unsound preprocessing of the input data.

The reconstructions are evaluated visually because the grasp planning and optimization for the segmented point cloud reconstruction haven't yet been incorporated. A representative example of the one synthetic and one real-world input and output are discussed below

B. Synthetic Scene:

This synthetic scene is created in ROS-Gazebo by using 3 YCB objects, Apple, Banana and Cereal Box. The object instance segmentation works perfectly for this data point, but the point clouds recorded by the simulated Kinect camera are not properly cleaned. Its effect can be seen in the failed reconstruction of the cereal box which is caused due to the line extending perpendicular to the cereal front point cloud and also the taper angle which causes the reconstruction network to assume that the point cloud corresponds to a tapered bat (like a baseball bat). This can be handled either in the reconstruction network or in the Segmented Mask Data Processor.

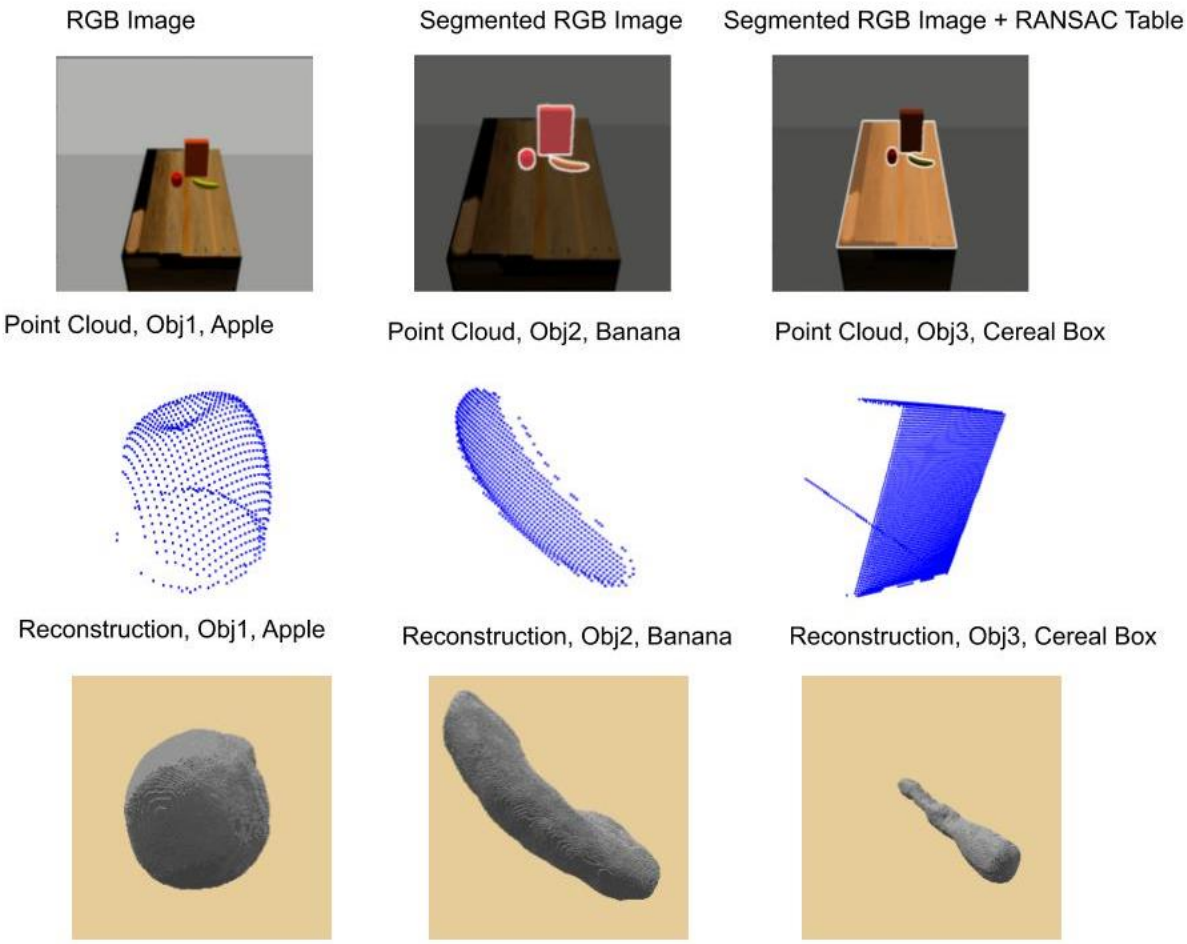

Figure 4: The first row is Object Instance Segmentation on one synthetic cluttered scene. The second row is the individual object point clouds that are extracted from the segmented mask. The third row is the corresponding reconstructed objects.

C. Real-World Scene:

This real-world scene is taken from the OCID-table-bottom dataset. The object instance segmentation doesn't fully work because the two stacked objects are considered as one object. But the stacked boxes case worked on the training dataset in [7] which shows that this is a testing error and can be improved by making the Segmented Mask Data Processor more robust. Although the segmentation hasn't fully worked here, the reconstructions from the point clouds are similar to the real-world scene, even for the stacked boxes

.

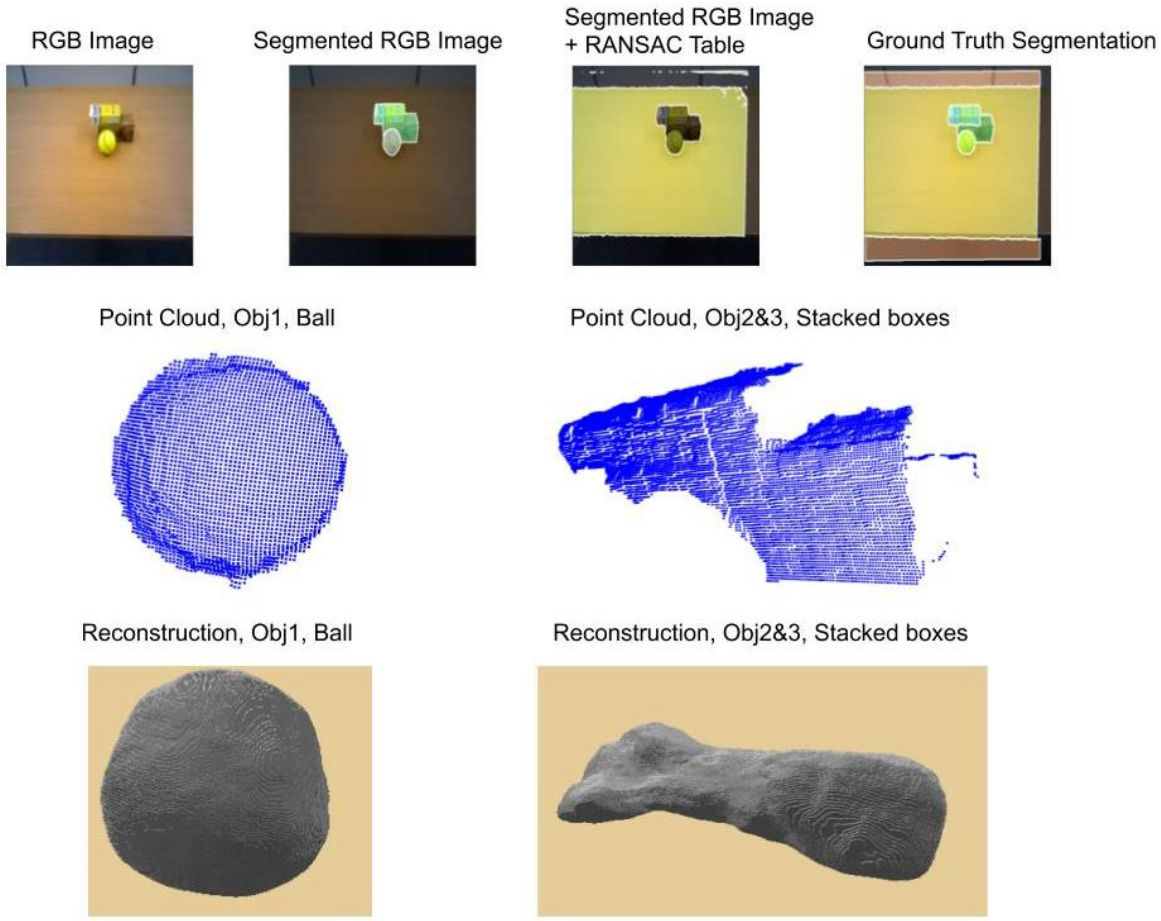

Figure 5: The first row is Object Instance Segmentation on one real world cluttered scene. The second row is the individual object point clouds that are extracted from the segmented mask (which is inaccurate). The third row is the corresponding reconstructed objects.

V. LIMITATIONS

The reconstruction and segmentation seem to be working but the effect of the segmented point cloud 3D reconstruction would only be clear after experiments with the robot. This encourages the next step to be, incorporating the grasp optimization and evaluating the grasp metrics.

Failed reconstruction for the cereal box in the synthetic dataset due to unsound processing of segmented masks. Although this scenario is specific to the synthetic data point, the issue has to be resolved because synthetic data is critical for training when real world data isn't available**.** This would also fix the inconsistencies in metrics for the same data between the original UOIS implementation and the project implementation and reduce the inconsistencies between Table 1 and 2.

VI. CONCLUSION AND FUTURE WORK

This project investigated different methods for object instance segmentation and developed code for merging UOIS network with PointSDF 3D reconstruction network and also the code for running new datasets quickly. The next step of this research would be to merge the grasp planning and continuous grasp optimization pipeline to run experiments and evaluate reconstruction metrics. Considering the other limitations that were identified, the further scope of the project would be to improve the input data preprocessing and investigate new methods to improve reconstruction. One direction is to reassemble the objects in a physics simulator to leverage the stability information and continuously improve reconstructions for unstable scenes. One recent research presented at CoRL [12] on Amodal reconstruction in Cluttered Scenes explores a similar

idea of leveraging physics information to improve reconstruction.

REFERENCES

[1] N. Doulamis, A. Doulamis and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, 2009.

[4] Y. Zhou and K. Hauser, "6DOF Grasp Planning by Optimizing a Deep Learning Scoring Function," in *RSS Workshop on Revisiting Contact - Turning a Problem into a Solution*, 2017.

[5] Q. Lu, K. Chenna, B. Sundaralingam and T. Hermans, "Planning Multi-Fingered Grasps as Probabilistic Inference in a Learned Deep Network," in *Conference: International Symposium on Robotics Research*, 2017.

[6] M. V. d. Merwe, Q. Lu, B. Sundaralingam, M. Matak and T. Hermans, "Learning Continuous 3D Reconstructions for Geometrically Aware Grasping," in *International Conference on Robotics and Automation (ICRA)*, 2020.

[7] C. Xie, Y. Xiang, A. Mousavian and D. Fox, "Unseen Object Instance Segmentation for Robotic Environments," in *arXiv:2007.08073*, 2020.

[8] C. Xie, Y. Xiang, A. Mousavian and D. Fox, "The Best of Both Modes: Separately Leveraging RGB and Depth for Unseen Object Instance Segmentation," in *Conference on Robot Learning (CoRL)*, 2019.

[9] Y. Zhang, X. Chen, J. Li, C. Wang, C. Xia and J. Li, "Semantic Object Segmentation in Tagged Videos via Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1741 - 1754, 2017.

[10] R. Q. Charles, H. Su, M. Kaichun and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] C. R. Qi, L. Yi, H. Su and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Conference on Neural Information Processing Systems (NIPS)* , 2017.

[12] W. Agnew, C. Xie, A. Walsman, O. Murad, C. Wang, P. Domingos and S. Srinivasa, "Amodal 3D Reconstruction for Robotic Manipulation via Stability and Connectivity," in *Conference on Robot Learning (CoRL)*, 2020.

[13] D. Kappler, J. Bohg and S. Schaal, "Leveraging big data for grasp planning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[14] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel and A. M. Dollar, "The YCB object and Model set: Towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics (ICAR)*, 2015.

[15] H. R. N. A. J. U. I. K. J. P.-T. S. K. S. P. M. M. A. K. T. D. V. F. Alina Kuznetsova, "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," in *International Journal of Computer Vision*, 2020.

[16] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," in *International Conference on Computer Vision (ICCV)* , 2015.

[17] T. Strutz, Data Fitting and Uncertainty, Springer, 2016.

[18] A. Dave, P. Tokmakov and D. Ramanan, "Towards Segmenting Anything That Moves," in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.