CSC2547 Project Midterm Report

Haoping Xu, Tao Wu, Zhoujie Zhao

1 Introduction

3D shape estimation is an area of growing interest in past decades. Recent advances in RGB-D cameras allow us to automate the construction of 3D shapes and scenes modeling the real world. Nevertheless, the performance of standard 3D sensors falls short for transparent materials such as plastic bottles and glass mugs which are very common in everyday life. Many classic stereo vision algorithms fail to fit transparent objects due to their special visual property. Hence, generating accurate depth estimates for transparent objects is a challenging task.

In this project, we propose to extend the ClearGrasp algorithm [1]. We will first train a segmenting model that predicts a mask of transparent objects in a 2D color image. The major procedure then is to predict the correct depth image for masked areas using 3D point cloud completion. We plan to employ state-of-the-art techniques like GRNet [2]. In addition, we propose to collect a novel dataset by capturing RGB-D images and augmenting them with alternative illumination and scenes.

2 Related work

2.1 ClearGrasp

ClearGrasp [1] aims to fix the distorted and incomplete raw depth due to transparent object. It is based on the deep depth completion network [3], where the surface normal and occlusion boundary are predicted from single RGB image. The raw depth map from the RGBD sensor is then combine with the estimated depth to do global optimization. This process is aimed to minimize a weighted error between raw depth, estimated depth, neighboring pixel depth and the predicted normal. On top of that, cleargrasp includes a segment head to predict the mask for transparent objects, and remove their corresponding depth information during global optimization. The masking operation utilize the capability of deep depth completion network's ability to complete missing depth map.

ClearGrasp uses synthetic dataset generated by blender as the training set and collects the validation and testing dataset from the real world using matched transparent and opaque objects and manually place them at same location.

2.2 Keypose

Keypose network [4] can predict a set of key points from a transparent object which can be used to determining its pose for downstream applications, i.e. robot manipulation. Unlike ClearGrasp [1], keypose use stereo images as input, and predict each key point using a 2D image location as well as the disparity between left and right image to obtain the 3D location. This paper proposes two variations to compute them, for early fusion model, it computes the disparity map and one 2D location, whereas the later fusion network estimates a pair of 2D location from each stereo pair and compute the disparity using the matched key points.

Keypose collects its dataset using a eye-in-hand robot arm and tracking system, Apriltag 2[5]. The camera path is calculated using Apriltag, and along the scan trajectory, several frames' keypoints are manually labelled, where the rest are labeled based on the trajectory. Also, authors also collect the depth info using a similar replacing method as CleatGrasp [1].

3 Method

3.1 Depth correction as 3D completion

Due to the optical proprieties of transparent objects, the depth information from common RGBD sensors are usually distorted and incomplete. We propose to segment out the incorrect depth map via the mask predicted from color image, similar to ClearGrasp [1]. But instead of training the segment network only on generated data, the model will be trained on much larger real world dataset, like Trans10K [6] and LabPic [7], and use architecture specifically designed to handle transparent object, i.e. TransLab [6], instead of general purpose network used in ClearGrasp.

With the isolated depth for transparent objects, Raw geometry to true geometry conversion can be treated as 3D point cloud completion task. Where the raw depth will be projected to a distorted and incomplete point cloud of the transparent object, and the actual point cloud sampled from the object's mesh is the target. We propose to use the SOTA model in this task, GRNet [2] and train it using dataset from Keypose [4] as well as new dataset collected by us. With the completed point cloud, we can test its performance by combining it with robot grip pose estimation and actual robot picking transparent objects.

3.2 Automate data collection

We will use an eye-in-hand Franka Emika Panda with Inteal realsense RGBD camera to collect the dataset. To automate the data collection and annotation process, a virtual scene with matched transparent object location, and same eye-in-hand robot setup will be created to replicate the real setup. And the robot's joint state can be replicated to the simulation, which can generate the ground truth point cloud, depth map and segmentation mask. Additionally, the virtual scene allows us to do synthetic data generation using the actual raw depth as input and virtual depth as target. Another method to fully automate the annotation is using the Apriltag 2, and placed the object to a fixed location and pose respect to the tags, which can be used later generate the ground truth with the object mesh.

4 Evaluation

We will evaluate our model's ability to estimate the shape of transparent objects in the following ways. To measure the depth error, we will use Root Mean Squared Error and the percentage of pixels that are close to the ground truth (e.g. the distance is smaller than a threshold). We will test our algorithm on the dataset containing unseen object shapes to examine the generalization ability of our model. ClearGrasp [1] will serve as a baseline for quantitative comparison, and several ablated version of our model will also be evaluated. The completed point cloud from our model will finally be incorporated as part of the robot picking system to check if the robot can grasp the transparent object through our output.

5 Progress

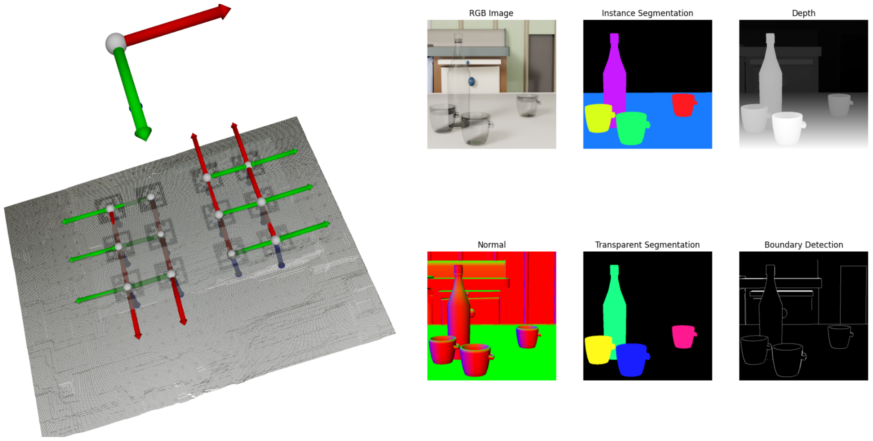
5.1 Dataset Collection

The Franka Panda robot eye-in-hand system is built and controllers for the robot as well as the realsense camera are programmed using FrankaPy [8]. In terms of automating annotation, the Apriltag 2 [5] approach turns to be easier both in calculating the relative 6DoF pose of the transparent object as well as writing the script. The detail dataset collection process is the following:

- An array of Apriltag2 are placed around a transparent object, each tag is from the same family with different index. And the offset from the tag and object is known using a printed template.
- Multiple random positions and viewing angles are selected. For every viewpoint, the robot aligns the camera with it and captures the RGB-D images.

- For automatic annotation, we first detect the Apriltag pose using the color input. The object pose can be extracted via tag's pose and using the 3D model of the object, the segmentation mask, ground truth depth map can be computed.
- Every viewpoint in the dataset will be consists of color image, raw depth map, ground truth depth, segmentation mask as well as the camera 6DoF pose relative to the object. The camera pose will be useful for recreating the scan scene when rendering additional synthetic dataset.

Figure 1: (i)Apriltag2 pose estimation using RGB image from realsense camera; (ii)Synthetic data from Nvidia IsaacSim



(i)

(ii)

Additionally, with the help from my lab colleague, the pipeline for rendering the synthetic dataset in Nvidia IsaacSim has a good progress. Currently the code base is able to render transparent object with high fidelity using ray tracing, and generate range of different annotations including ground truth depth, segmentation map, boundary map and normal map. This can greatly increase the size of our dataset and improve the trained model performance.

5.2 3D Point Cloud Completion

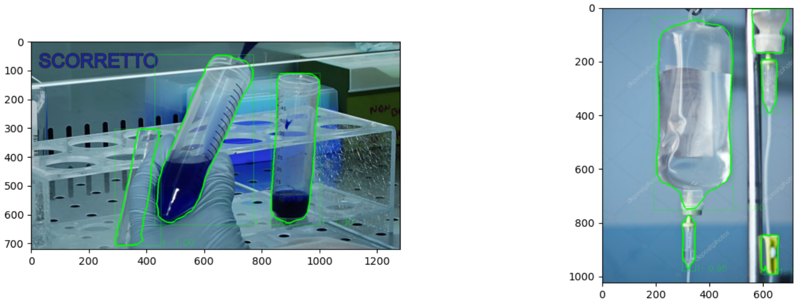
A mask rcnn model is trained to segment out the transparent objects, a new version of LabPic [7] and COCO dataset's [9] subsets including the class with transparent objects are used as training data. Overall the new mask rcnn model can achieve mAP of 0.785 for 0.5 IoU. Whereas the Deeplab3 used in Cleargrasp [1], it can only get average IoU of 0.58 for 5 different objects and only 3 of them are not in their synthetic dataset. In comparison, our mask rcnn network generalizes to hundreds of different transparent objects in multiple real world scenes.

GRNet [2], the state of the art to our best knowledge, is employed as the model for 3D point cloud completion. To test the efficacy of GRNet, we leverage real-world data from ClearGrasp [1]. The input data and the ground truth for GRNet are obtained by projecting the segmented depth images of transparent and opaque objects respectively to point cloud representations. A pre-trained GRNet model provided by the authors is used to generate the auto-completed point clouds for transparent objects. Then the Gridding Loss, which is introduced in the GRNet paper, is implemented to compare the produced point clouds with the ground truth. This part is still in progress currently, though we expect GRNet to present decent point cloud completion results.

6 Future Plan

In terms of data collection, we will try to collect more different types of objects in various lighting and backgrounds in order to generate a more general dataset. Also the rendering pipeline needs to be adapted to taking the scanned dataset and recreate and augmenting the scene, constructing the raw

Figure 2: Segmentation map from LabPic mask rcnn network



depth map by overlaying the scanned one with rendered ground truth.

We are still comparing the effectiveness of original ClearGrasp, substituted ClearGrasp with a better segmentation module as well as GRNet. Once our real and synthetic datasets are ready we will try to train all three networks and compare they performance.

References

[1] Shreeyak S. Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Cleargrasp: 3d shape estimation of transparent objects for manipulation. *CoRR*, abs/1910.02550, 2019.

[2] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion, 2020.

[3] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image, 2018.

[4] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. *CoRR*, abs/1912.02805, 2019.

[5] J. Wang and E. Olson. Apriltag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4193–4198, 2016.

[6] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild, 2020.

[7] Sagi Eppel, Haoping Xu, Mor Bismuth, and Alan Aspuru-Guzik. Computer vision for recognition of materials and vessels in chemistry lab settings and the vector-labpics data set. *ACS Central Science*, 6(10):1743–1752, 2020.

[8] Kevin Zhang, Mohit Sharma, Jacky Liang, and Oliver Kroemer. A modular robotic arm control stack for research: Franka-interface and frankapy, 2020.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.