# Learning 6DoF Grasp-poses in Occluded Scenes Using RGB and 3D Point Cloud Modalities

Abhay Dayal Mathur  Sachin Bhadang  Prem Raj  Laxmidhar Behera  Tushar Sandhan

Indian Institute of Technology, Kanpur

## Introduction

Localisation and pose estimation of the target fruit is central to any automation-based solution for fruit harvesting. These tasks can be challenging due to the occlusion caused by branches and leaves, as it limits the options to approach the target fruit.

The contributions of this paper are as follows:

- We propose an augmentation method, **3D-Copy-Paste (3DCP)**, for augmenting point-cloud data for estimating the Six Degrees of Freedom (6DoF) grasp pose.
- A 6DoF apple grasp-pose dataset and a **two-step deep learning-based baseline** for apple grasp-pose estimation are also analysed.
- Additionally, we present a setup for our customised **semi-automated annotation method** for labelling 6DoF apple grasp-poses, developed in conjunction with the RViz widget.

## Dataset Generation

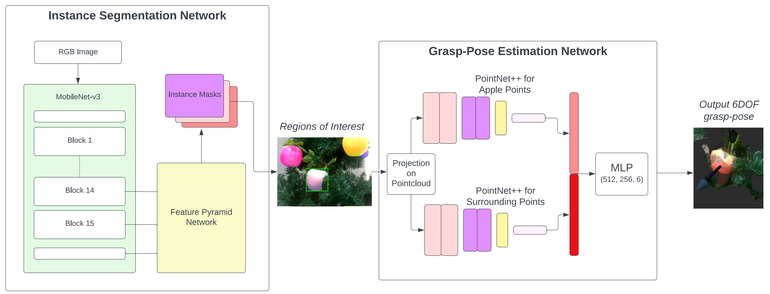
Data has been collected, both in outdoor and laboratory environments, using artificial apples placed on artificial (indoor) as well as real (outdoor) trees.

Data was collected using a **Realsense D435i** camera, with the distance between the camera and the apple trees ranging from 0.15m to 2m. The dataset comprises **225** RGB images and point clouds, spanning **712** instances.

Ground truth segmentation masks have been labelled using the **CVAT** tool from OpenCV, and the ground truth 6DoF grasp-poses were labelled using our **customised semi-automated annotation method**, described in the next section.

## Deep Learning Network design

The two sub-modules of the Network are described below:




### The Instance Segmentation Network (ISN)

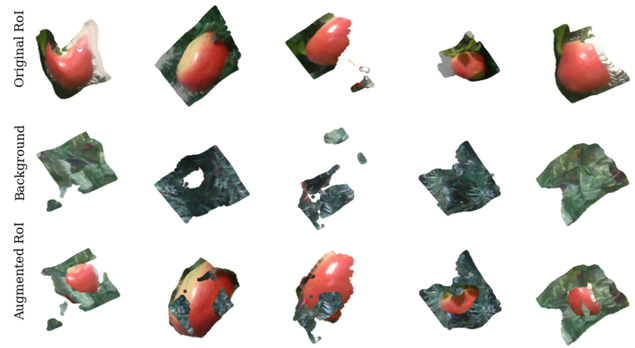
The Instance Segmentation Network (ISN) uses a **Mask-RCNN** based CNN network that consists of a backbone of **MobileNet-V3** mounted with a Feature Pyramid Network (FPN).

### Grasp-pose Estimation Network (GPEN)

Instance masks and bounding boxes from the ISN network are projected onto the point cloud. Two point cloud segments are generated, the projection of the segmentation mask (apple points) and the projection of the background in the bounding box (surrounding points) and are processed separately by two different instances of the **PointNet++**.

## 3DCP : A Point Cloud Augmentation Method

In this method, we first select the data sample that includes the 3D points corresponding to the target fruit (i.e. target points) and the 6DoF grasp-pose label. To augment this data sample with 3DCP, the target points are **pasted** onto a point- cloud region selected randomly from the background scene, which majorly comprises branches and leaves.

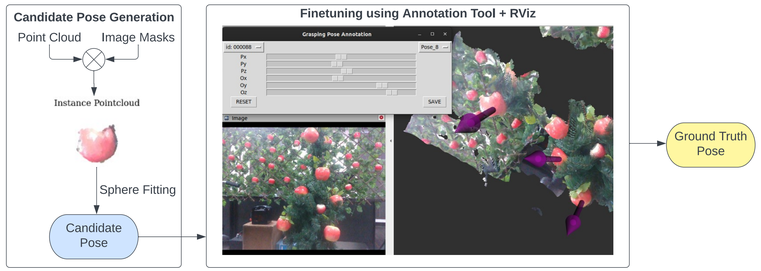



The important constraint in the process is that the target pose must remain valid. This is equivalent to ensuring that the occlusion created by the surrounding (background) points in the new point cloud, about the target pose, remains comparable both in magnitude and distribution to that in the original case.

To this end, we define the **region-of-unobstructed** approach as the region in space which must remain devoid of any occlusions for the apple to be successfully grasped. It is modelled as a conical region with its apex $\alpha$ at the centre of the apple and with its axis along the target pose.

## Annotation method for 6DoF grasp-pose

A two-step process is followed to generate ground-truth annotations for 6DoF grasp poses.




- First, given the segmented points corresponding to the target apple, a candidate grasp pose is generated automatically based solely on the geometric properties of the point cloud.
- The grasp pose is manually fine-tuned in the second step using a customised interactive GUI applet. The grasp-pose label consists of the position of the apple $(x, y, z)$ and the desired orientation of the robot arm's end effector in Euler-angle notation $(\psi, \phi, \theta)$.

## Results and Analysis

### Evaluation of the ISN Network

The small and large configurations of MobileNet-v3 differ in the number of hidden layer parameters and the output resolutions. All the models have been trained on our dataset with backbones adopting pre-trained weights from **ImageNet**. The proposed network (**MobileNet-v3 Large + FPN**) performs comparably against the ResNet backbone

| Backbone | F1-Score | Time |
|---|---|---|
| MobileNetv3-Small | 0.84 | 139ms |
| MobileNetv3-Small + FPN | 0.79 | 115ms |
| MobileNetv3-Large | 0.86 | **83ms** |
| MobileNetv3-Large + FPN | 0.88 | 98ms |
| ResNet50 + FPN | **0.93** | 122ms |

### Evaluation of the GPEN network

The GPEN network has been evaluated for two backbones - **PointNet** and **PointNet++**. PointNet++ applies a PointNet network recursively, which captures local structures and patterns better, achieving better performance but leading to a longer inference time. Both configurations were pre-trained on the **ScanNet dataset**. The inference time and RMSE-loss for position (meter) and orientation (radian) are reported in the table below.

| Backbone | Time | $\text{RMSE}_{position}$ | $\text{RMSE}_{orientation}$ |
|---|---|---|---|
| PointNet | **15** ms | 0.081 | 0.689 |
| PointNet++ | 82 ms | **0.064** | **0.563** |

## Conclusions

We have developed a framework to learn 6DoF apple grasp-pose using the RGB and point cloud data in an orchard-like scenario. We presented a dataset consisting of 2D instances masks and 6DoF grasp poses. We have proposed an augmentation technique for point clouds to improve training with relatively little training data. We have also provided a baseline method with substantial accuracy over the test set to be used as a benchmark for comparison of future research in this domain. Overall, our efforts in this paper would help research progress in autonomous fruit harvesting and related areas.

## Acknowledgements

We are thankful to the Ministry of Electronics and Information Technology, Government of India (MEITY) for funding this work under the project titled "Robotics and Automation in Agriculture" reference number 4(16)/2019-ITEA.