CS763 Project
3D Object Detection and Semantic Map Generation

- Parth Shettiwar 170070021
- Siddharth Saha 170100025
- Parikshit Bansal 170050040

Problem Statement

Our project submission is a part of the 2021 Robotic Vision Scene Understanding (RVSU) Challenge organised by the Australian Centre for Robotic Vision to be submitted on 10th May,2021 (https://eval.ai/web/challenges/challenge-page/807/overview). We will be participating in the phase “Semantic SLAM, with passive actuation & ground-truth localisation”. Following is the overview of the problem specification of this phase “”:

- We are provided a robot which will traverse around the environment, and we have to build up an object-based 3D semantic map from the robot's RGBD sensor observations and odometry measurements.
- The robot will follow a fixed-trajectory, and we are given a single method to control the robot: moving to the next pose. Furthermore, at each time step, we are given the robot’s groundtruth pose for each of its components.

Problem Statement (Continued)

- In other words, at the end of the traversal through environment, we should be able to do the following tasks:

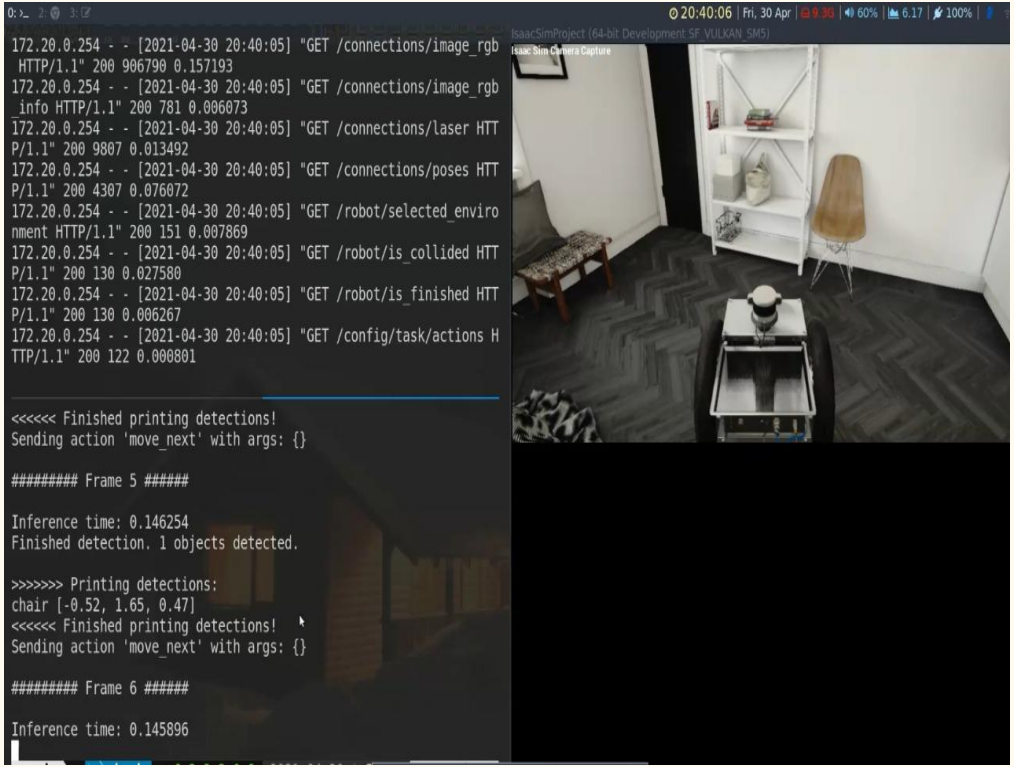
1. Detect all the objects in the environment
2. Classify them as one of the objects from the list given on the website
3. Create a 3D bounding box around all the objects detected
4. Create a 3D semantic map of all the objects detected in the environment from the starting position camera perspective.

At each time step of robot, we would get the the RGB image and a depth image in addition to the bots current position.

On the right is the list of objects which will/might be present in the environment on any run. There can also be multiple instances of same object in the environment

On the right we show a snap of the Bot running in the environment.

Class
Bottle
Cup
Bowl
Spoon
Banana
Apple
Orange
Cake
Plant
Mouse
Keyboard
Laptop
Book
Clock
Chair
Table
Couch
Bed
Toilet
TV
Microwave
Toaster
Fridge
Sink
Person



Motivation

To participate in an international challenge to solve a core computer vision problem is what incentivised us to take up this project. 3D Object detection and semantic map generation is highly used nowadays in Driverless cars or autonomous driving. It becomes very crucial as whether the car detects objects correctly in the run time so that it can predict most optimal action to be taken for next time step. This is complicated by the localisation problem. However, in this challenge we have assumed the trajectory of bot is fixed and hence there is no localisation problem. Even after that, there are many things which we learnt during the process which would be relevant in many other tasks in computer vision field and have real life applicability.

Related work

Much of the research till now has been done in 3D Object detection given a point cloud. However, there is no particular research done on how to integrate the results from these 3D object detection algorithms into creating a 3D semantic map of environment from a series of RGBD inputs. We list the 3D object detection networks which have recently produced state of art results (all the networks take a pointcloud as input and produce 3D Bounding boxes and labels classification for each object):

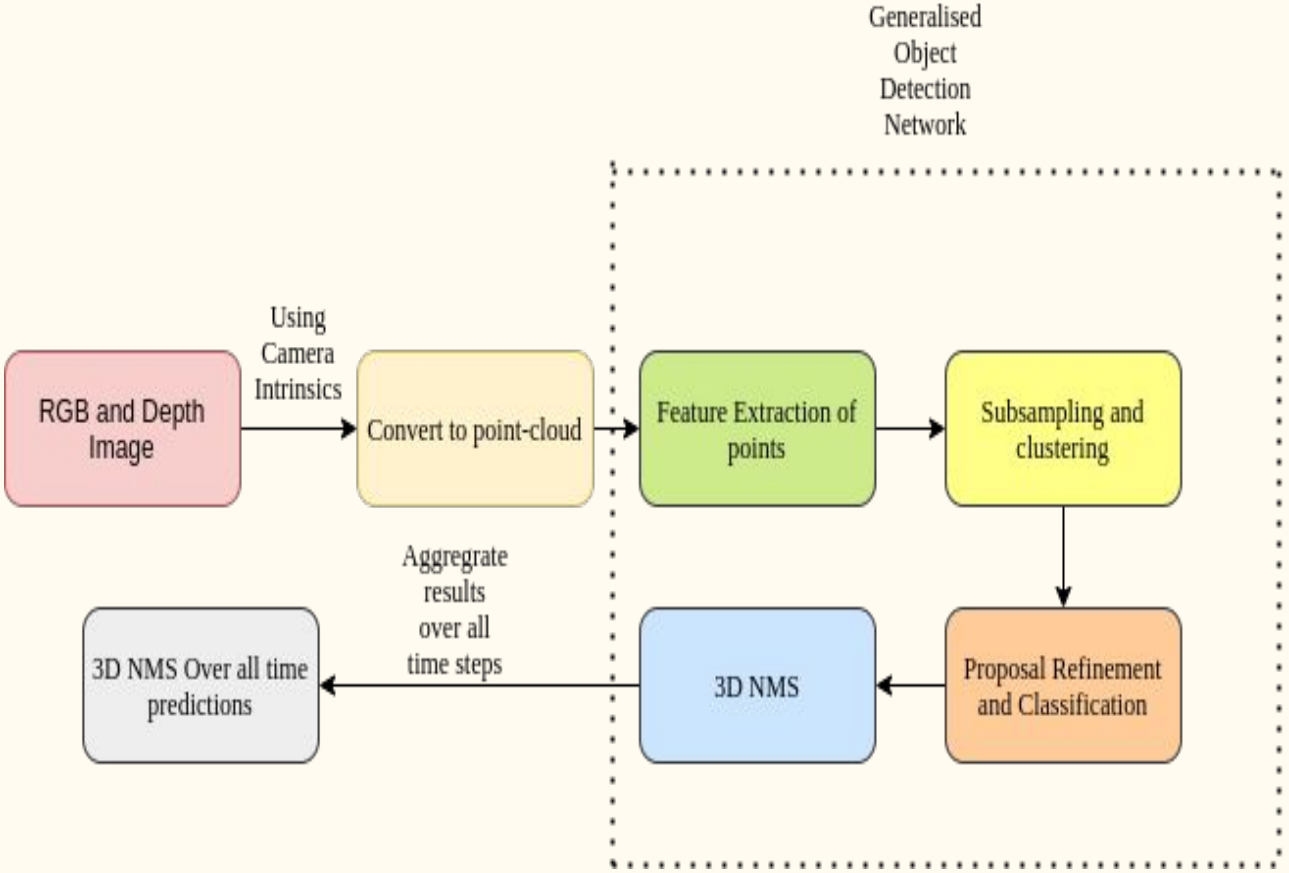
1) Group-Free 3D Object Detection via Transformers: This network produces state of art results on the SUNRGBD Dataset leveraging the recent Transformer network used in NLP Tasks.
2) Votenet: Next to Group-Free 3D, the architecture relies on voting mechanism to subsample and take appropriate points from the pointcloud for Object detection and 3D Box generation
3) ImVoteNet: Similar to Votenet, but have additional feature of including information from RGB data of image in addition to conventional prediction from point cloud lacking color information
4) There are various other networks: BRNet, H3DNet, HGNet, all of them vary in some sort of mechanism in sampling subset of points from the input pointcloud.

Extension to Original work

As mentioned in the problem statement slide, we intent to do 4 tasks. We leverage the existing object detection networks to solve points 1,2,3 from the problem statement slide. Our main extension over the past original work, is to aggregate results from the object detection networks and generate a semantic map of the environment. This was primarily done using a 3D Non maximal suppression approach to discard predictions which are repetitive and are predicted with low confidence in the collection of predictions from all frames. Furthermore, we did appropriate pre-processing of the data feed to convert it into point cloud required for the input to 3D object detection networks. Lastly, we have done a comparison of the top 2 object detection networks in this setup based on various metrics.

Approach

As mentioned in previous slide, we leverage the current state of art object detection networks to produce 3D bounding boxes for each frame. This is followed by a 3D NMS algorithm to appropriately take the necessary predictions across all frames to generate a 3D semantic map. For object detection networks, we used the top 2 networks for SUNRGBD Dataset: Group Free 3D and VoteNet. The following flowchart illustrates our pipeline of the work:



Datasets

A major challenge faced is that existing object detection networks are trained on datasets which have little overlap with the Robotic Vision Scene Understanding Challenge environment objects. The environment has objects from COCO dataset which is a 2D image dataset and hence we can't leverage it. The existing RGBD Image datasets are:

1. SUNRGBD
2. ScanNet
3. Diode Dataset
4. KITTI

Approach Continued

Due to time and hardware constraints, our Goal was to give maximum accuracy using minimal training time. The following approach was used to leverage the current object detection networks:

- Firstly, it was observed that the only SUNRGBD and ScanNet datasets had overlap with some objects in the RVSU Challenge environment. Additionally, even though ScanNet dataset hada better overlap with envoironemtn objects, the dataset had to be requested from the authors and would have taken a week's time for the same. Hence we have done all the work using SUNRGBD dataset.

Dataset	Number of Objects overlap
SUNRGBD	15 out of 25
Pre Trained SUNRGBD Weights	5 out of 25
ScanNet	19 out of 25

- As we can see from table, The pretrained sunrgbd weights of current 3D Object Detection networks have only 5 objects overlapping with the environment objects. Training a full fledge network on whole dataset was highly infeasible due to the time and hardware constraints as mentioned before.
- We tried to mitigate the problem using Ensemble Learning.

1. The pretrained VoteNet network gave bounding boxes for 5 objects as seen in last table. So we have remaining 15-5 = 10 objects which can be trained on .
2. Of these , we took most frequent 8 objects and created the SUNRGBD Dataset again. The dataset reduced by 10 fold.
3. On these reduced dataset we trained the VoteNet Network using previous hyperparameters and saved the weights after 180 epochs. This took about 4 hours as compared to 24 hours required on conventional approach.
4. We integrated this model in additional to pre trained model to give the final prediction. As objects classified by both approaches were disjoint, we had to ensure that predictions from the 2 networks don’t overlap.

3D NMS Algorihtm

We have used the conventional Non Maximum Supression approach to discard irrelevant predictions gathered from all time frames. We have used the 3D IOU Metric to do this. Following is the overview of the algorithm:

- Given list of Proposals K (bounding boxes), Confidence scores S and a threshold X

1. Take the proposal from K with highest confidence score.
2. Find the 3D IOU of this proposal with every proposal in K. If this is greater than X, discard that proposal from B. Do this over all proposals.
3. Take the next proposal from B with highest score S.
4. Do this until the list B empties out.

Quantitative and Qualitative Results

Evaluation Metrics

As mentioned on RVSU Challenge site (https://eval.ai/web/challenges/challenge-page/807/evaluation), we did our comparison of networks on the following metrics:

1. Object map quality (OMQ): Considers a geometric mean of spatial and label quality. Spatial quality is calculated by finding 3D IOU with the groundtruth and label quality by comparing probability assigned with groundtruth. The final OMQ Score is calculated by dividing True positives by total false positive , false negative and true positives where each label of true or false positive is determined by above geometric means for each object.
2. Average overall pairwise quality: Averages above geometric means
3. Additionally we also compare on Average false positive quality , Average pairwise label quality and Average pairwise spatial quality.

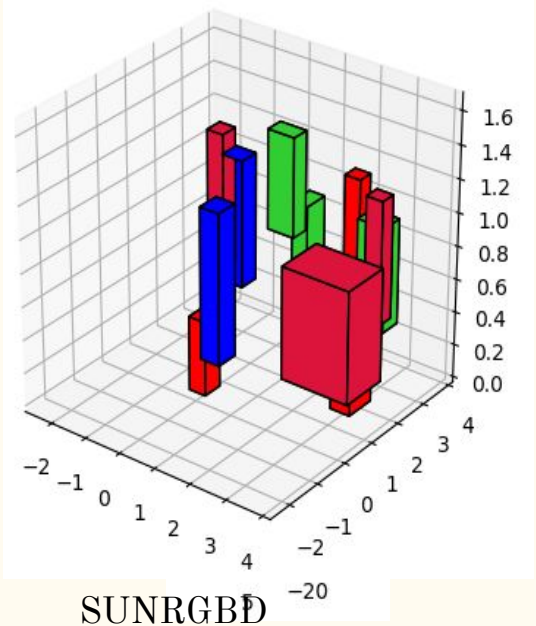
Quantitative Results

Following are the results of our 3 setups on the previous metrics:

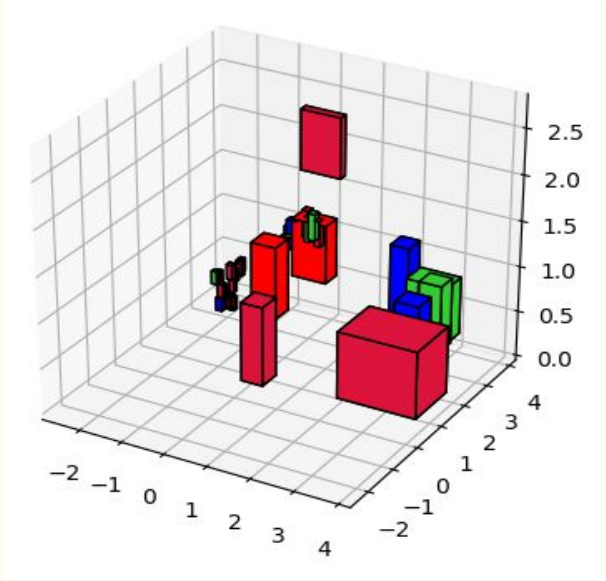
Metric	SUNRGBD Pretrained weights	SUNRGBD (Pretrained + Trained network) (Ensemble learning)	Group-Free 3D (Pre trained weights)
OMQ	0.22	0.072	0.008
Average fp quality	0.095	0.032	0.213
Average label quality	0.98	0.99	0.895
Average pairwise quality	0.71	0.67	0.089
Average spatial quality	0.52	0.48	0.0088

Qualitative Results

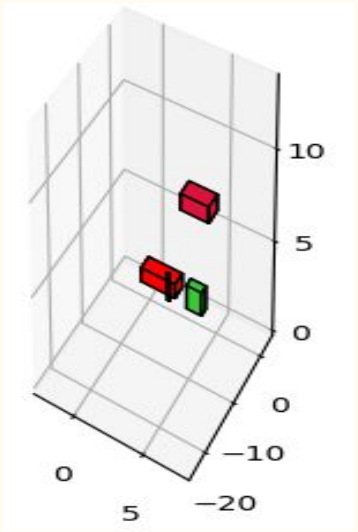
We created the 3D semantic map of the room using the 3 setups mentioned in previous slide. Following are the outputs:



SUNRGBD
Pretrained weights



Ensemble Learning



Group
Free 3D

Observations

We note that votenet with pre trained weights gives best prediction both Qualitative and Quantitative . We speculate that other networks of group free 3D and ensemble learning need further hyper parameter tuning. Furthermore, the SUNRGBD pretrained setup gives good qualitative results by giving a good semantic map whereas other setups , Group free 3d gives very less objects and Ensemble learning predicts too many objects. Point to note, group free 3D gives good fp quality scores, implying less false positives compared to other setups.

Future Work

We intend to perform on our setups to give a good score for our submission to challenge for 10th May. Following are the things we would work on:

1. Better the performances of the setup by hyper parameter tuning. We hope our ensemble learning approach gets a better score over other 2 methods.
2. Better the post processing 3D NMS approach to discard objects as ensemble learning would produce lot of objects as compared to single network.
3. Improve the training time, so that we can conduct many experiments. This we intent by training on sub sample of objects everytime to speed up training.

Load Factor and 16th April submission comparison

In the project, we all three have worked on every aspect equally. The various ideas were brought up in the discussions. Mainly, each one of us worked on the setups individually and aggregated the results later.

As compared to 16th april, we have achieved our objective on 2 different networks. The idea of 3 networks wasnt achieved due to time constraints. However we tried to produce results on a novel ensemble approach which we fell would be very useful for the competition finally. Furthermore, we did a Qualitative and Quantitative as mentioned in 16th April submission. Furthermore, a novel analysis of the dataset to optimally subsample and achieve our objective was done. We hope to speed up training and achieve better results using this approach.

Thank You