The full name of NeRF is neural radiance fields, which is an originally design to synthesize novel views from several pictures only. The output of this model is a stream of pictures from different angles. However, this algorithm doesn’t construct any surface. So here, our method combines the NeRF model and a 3d-reconstruction network together to generate the surface model. The model also takes a 5-D positional data consisting of the 3-d coordinate and 2 angular values representing the view direction. And there are two MLP used to learn the mesh and texture separately. ~~While mesh network uses signed distance field to represent the surface implicitly and the other perceptron uses radiance for synthesizing the output.~~

And the benefit of this method is that it is not restricted by the dataset, that’s because the dataset is only used for training. And it uses global Perceptrons, so it doesn’t need to model each object separately, therefore, it can handle scenarios with unseen objects and objects that are hard to distinguish. And because it uses radiance network, the output colour and texture can be more realistic.

However, this method can’t guarantee flawless output.

Please notice that there is a extra hole on the table. ~~Which is apparently not expected by a robot.~~

~~In the picture in the right hand side, it compares the result of the ground truth and the generated model. However, the other two method doesn’t have this problem, because they will learn the objects at first. And they will know that there is no hole on a table~~

And another drawback is that the quality of result is heavily influenced by the quality of input data. It requires a sequence of pictures from different angles to generate a satisfying result.

In this project, we plan to use two datasets: Scan net and 3d-front. Scans of 707 labelled real-world indoor scenes. 2.5 Million RGB-D images​. Which is sufficient for both training and validating. Of course, it contains flaws and noises that may happen in real life. It is useful for us to validate the performance because the robot would be used in real life.

And about 3d-front, it contains around 19 thousands synthetic rooms. The large amount of data is very useful for refinement stage. However, there is no images provided, so we must manually shoot pictures using tools like blender. So, we decide to use it to augment the final model if we have time.

According to the literature review we’ve done before, there are two major approaches of 3d reconstruction. The classical ICP/mathematical method and ai method. As the application of deep learning is evolving rapidly. we plan to explore several methods with different design concept and different underlying intuition and compare then horizontally. Also, we are going to pick one classic method which has been proven efficient and correct as the benchmark for vertical comparison. So, we chose Bundle Fusion in this project.

The three ai methods we have picked are: Learning-based ICP, Scene understanding algorithm, and NeRF.

Due to the nature of their unique architectural design, they all have strengths and weaknesses in different aspect.

Bundle Fusion as a classical one, can provide fast high quality surface modelling with great scalability consistency. It is also worth to mention that, this algorithm can handle tasks to model complicated large indoor scenes. However, to get an accurate result, users must feed sufficient input data that covers each object. So, this method is not a time-efficient choice for most applications. Especially for a robot, it can’t obtain enough input data at start.

Learning-based ICP extends the task of reconstruction to recomposition, which means that the machine is not creating a new model, rather fitting the model into a model in mind which is seen before. This method can generate a fast model from video. Even we don’t give it sufficient time to estimate, it can still give a satisfying result of room layout estimation which is especially useful for robots. However, the model should be trained for each type of scene and object so that it can recompose the real world correctly. It may not work well for unseen scenarios.

The pro&cons of scene understanding is similar to the previous one, but it can model each object separately so it is not strictly restricted by the knowledge base problem. After partition, most object can be modelled with correct shape and position. But for a complicated scene, the information contained in one single image may not be sufficient.

Un like the previous two methods, the NeRF algo doesn’t need to understand the semantic of the picture, so it works well with novel scenarios. However, it needs sequence of pictures instead of one to get he result. Also, the quality is heavily determined by the pose of the camera, which means that the model may not be complete if the scene is not observable from certain angles. This is also a considerable problem for our task.