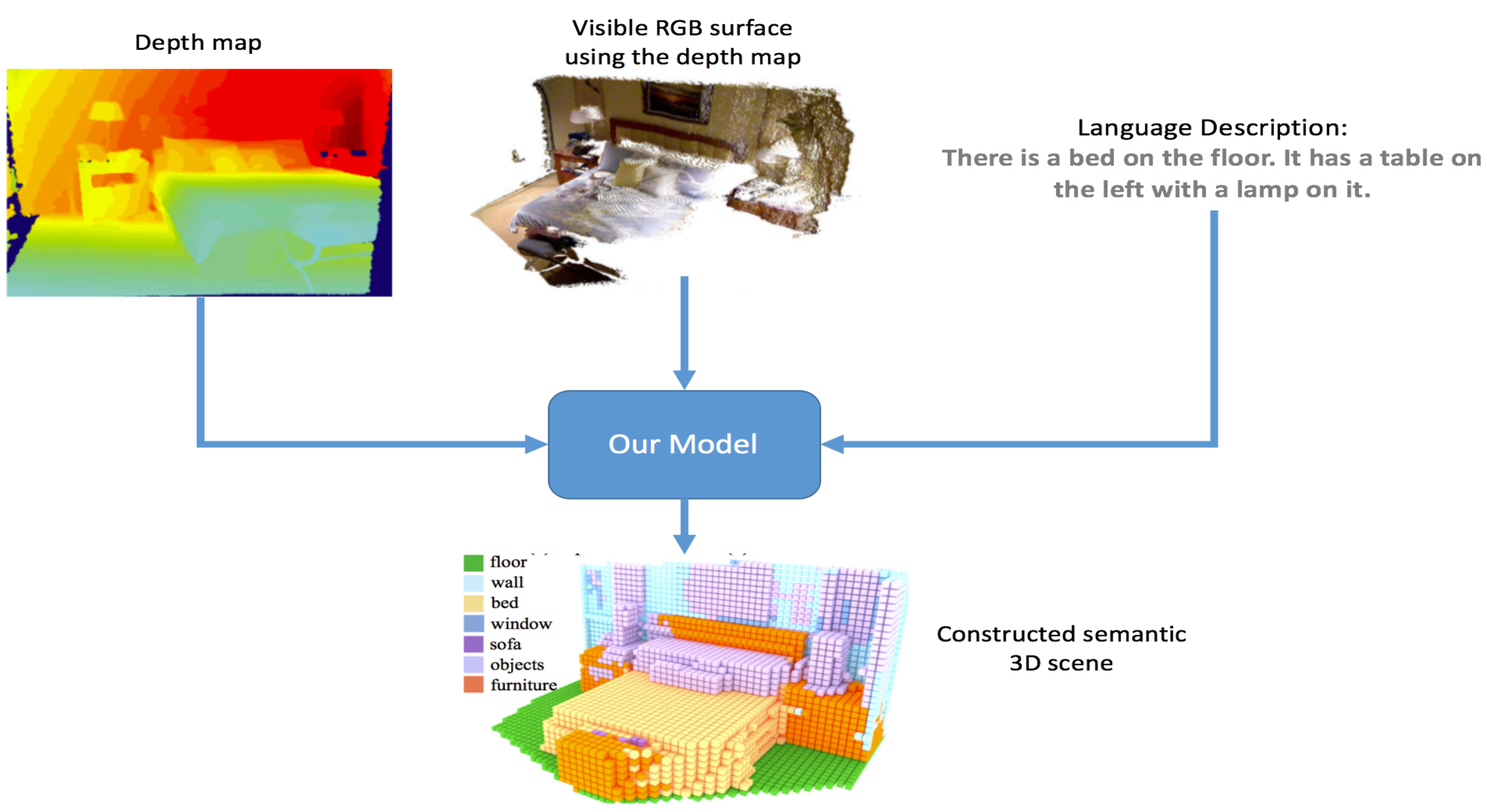
Semantic 3D Scene Generation using RGBD Image and Natural Language Descriptions

Jisan Mahmud

Final Project Comp 776

Problem

- Semantic 3D scene generation from an RGBD image and some natural language descriptions of the scene.
- Producing complete 3D voxelized representation of volumetric occupancy and semantic labels of the scene.



Task Overview

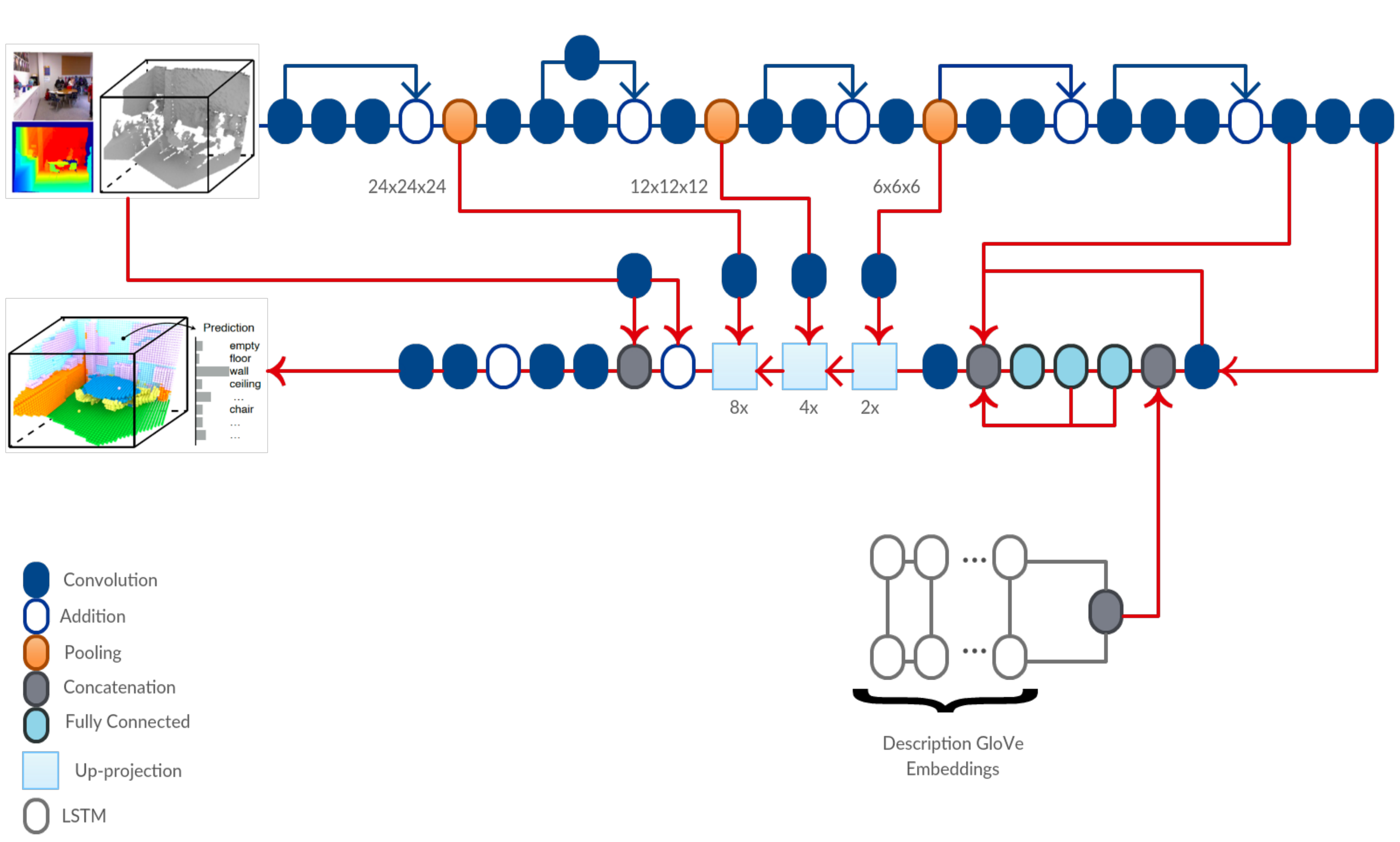
Principle

- Jointly formulate the sub-problems of task scene completion and semantic labeling.
- An end-to-end LSTM-fully convolutional neural network that utilizes the encoded information from the RGBD image and the textual descriptions to generate the semantically labeled scene.

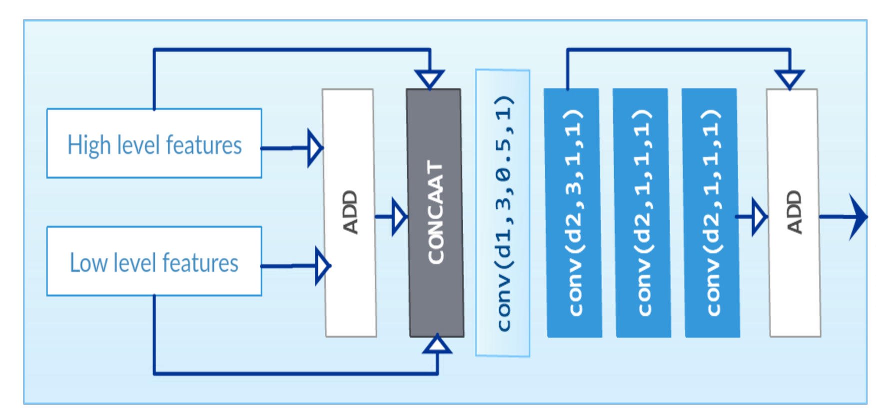
Challenges

- Adapting the dataset created for an entirely different purpose.
- Training the neural network with extreme class imbalance.
- High computational complexity.

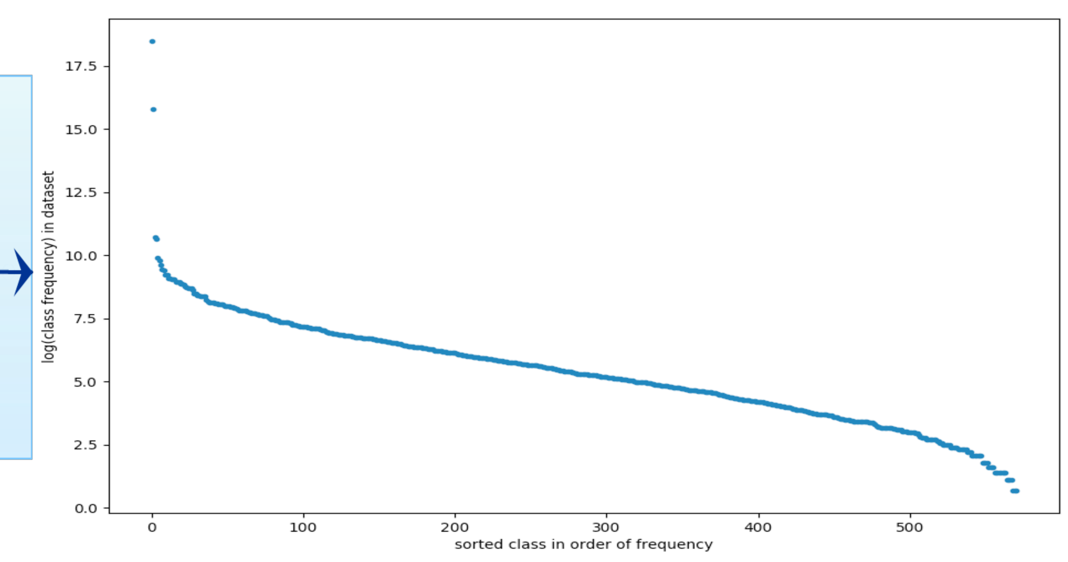
Network Architecture



- Scene descriptions are tokenized and fed into 2 layer stacked LSTM encoder using the tokens' GloVe word vectors.
- Visible voxelized RGB surface passed into a CNN network to have a obtain a high level encoded representation. Reduces spatial resolution of the scene from $(48 \times 48 \times 48)$ down to $(6 \times 6 \times 6)$.
- Encoded information from RGBD image and description are concatenated, followed by some FC layers.
- Successive up-projections that fuse together high level encoded information and some low level scene representations.



Up-projection Module



Class frequency Distribution

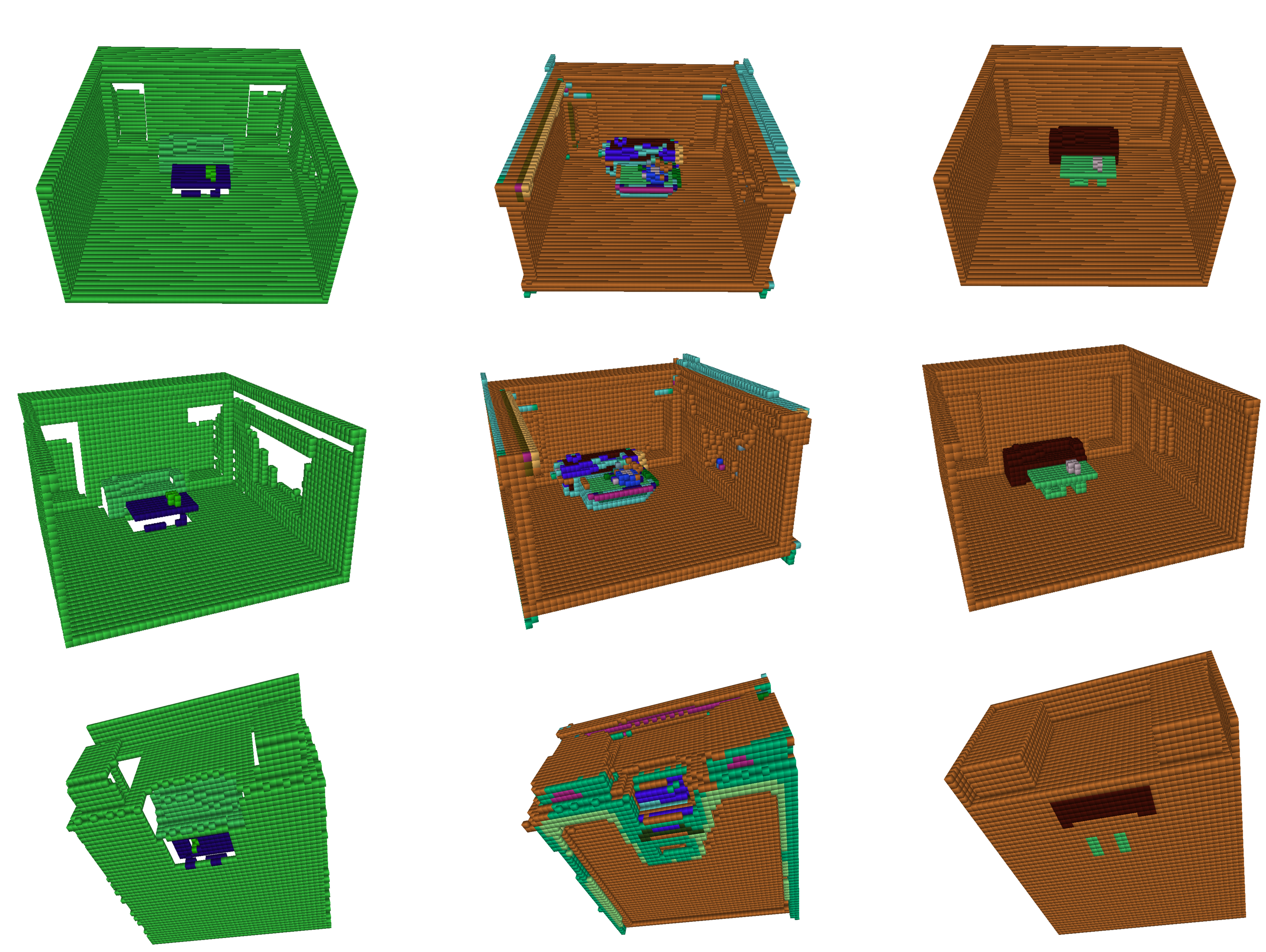
Training the Network

- Weighted softmax cross entropy as loss function due to class distribution imbalance. Each scene is a mini-batch.
- Weight of class i, $w_i^{(c)} = \frac{k_1}{f_i^{(1)}}$; $f_i^{(1)}$ is mini-batch frequency, k_1 is a normalizing constant.
- Weight each voxel j as $w_j^{(v)}$ so that voxels invisible from camera view are weighted higher.

$$\boldsymbol{loss}_{(class\ i, voxel\ j)} = -\boldsymbol{w}_i^{(c)} \times \boldsymbol{w}_j^{(v)} \times \boldsymbol{true\ one\ hot}_{(i,j)} \times \boldsymbol{log}(\boldsymbol{prediction}_{(i,j)})$$

Evaluation

Average Accuracy(%)	Median Accuracy(%)	Average IoU(%) (Over Voxels)
95.82	96.32	92.04



Voxelized RGBD Scene

Predicted 3D Scene

Ground Truth

UNC
DEPARTMENT OF COMPUTER SCIENCE