Supplementary Material

1. Visualizations of 2D and 3D detection results

Visualizations of both 2D and 3D detection results are shown in Figure 1 and Figure 2 for both the based on RGB-D system and based on depth image only system. From Figure 1 we can see that the 3D detection system works well for both the based on RGB-D and based on Depth only systems. The RGB-D based 3D detection system will generate some false positive 3D detections as it has more false positive detection during 2D detection stage. We can also find out that our system can detect objects which were not labeled during the data annotation. In Figure 2, in the left two images, our system can successfully detects unlabelled objects. On the right two images of Figure 2, we can see we have some false negative detections as there are too few points within the object. Detail explanations are given in captions.

2. Proof of equation 1

Proof. Define the threshold used for positive 3DCB as $threshold_{3D}$, and a 3DCB is positive when $IoI^{3D} = IoI^{XY} * IoI^{Z} \geq threshold_{3D}$. The $recall^{volume}$ $recall^{XY}$, $recall^{Z}$, $threshold_{XY}$ and $threshold_{Z}$ are defined in the main article. We set the $threshold_{3D} = threshold_{XY} * threshold_{Z}$.
As $3DCB^{positive}_{XY} \cap 3DCB^{positive}_{Z}$ implies $IoI^{XY} \geq threshold_{XY}$ and $IoI^{Z} \geq threshold_{Z}$. We can further get $IoI^{3D} = IoI^{XY} * IoI^{Z} \geq threshold_{XY} * threshold_{Z} = threshold_{3D}$, which implies 3DCBs in the set of $3DCB^{positive}_{XY} \cap 3DCB^{positive}_{Z}$ are positive. Meanwhile, we can show from an example that the set of $3DCB^{positive}$ can possibly be obtained from $3DCB^{positive}_{XY} \cap 3DCB^{nonpositive}_{Z}$, where $3DCB^{nonpositive}_{Z}$ is a complement set of $3DCB^{positive}_{Z}$: if $threshold_{XY} = 0.9$, $threshold_{Z} = 2.9$, we can get $threshold_{3D} = 0.81$. A 3DCB with $IoI_{XY} = 1.0$, $IoI_{Z} = 0.82$ will be an element of set $3DCB^{positive}_{XY} \cap 3DCB^{nonpositive}_{Z}$. Also it is a positive 3DCB. From above arguments, we can conclude the following relation:

$$3DCB^{positive} \supseteq \{3DCB^{positive}_{XY} \cap 3DCB^{positive}_{Z}\} \quad (1)$$

From equation 1, we can get:

$$|3DCB^{positive}| \geq |3DCB^{positive}_{XY} \cap 3DCB^{positive}_{Z}| \quad (2)$$

We can also rewrite right part of equation 1 as:

$$\begin{aligned} &\{3DCB^{positive}_{XY} \cap 3DCB^{positive}_{Z}\} \\ &\qquad = 3DCB^{positive}_{XY} \setminus \\ &\qquad\qquad \{3DCB^{positive}_{XY} \cap \{3DCB \setminus 3DCB^{positive}_{Z}\}\} \end{aligned} \quad (3)$$

From equation 3, we can further get:

$$\begin{aligned} &|3DCB^{positive}_{XY} \setminus \\ &\quad \{3DCB^{positive}_{XY} \cap \{3DCB \setminus 3DCB^{positive}_{Z}\}\}| \\ &\quad \geq |\{3DCB^{positive}_{XY} \setminus \{3DCB \setminus 3DCB^{positive}_{Z}\}| \\ &\quad \geq |3DCB^{positive}_{XY}| - |3DCB \setminus 3DCB^{positive}_{Z}| \\ &\quad = |3DCB^{positive}_{XY}| - (|3DCB| - |3DCB^{positive}_{Z}|) \end{aligned} \quad (4)$$

From equations 2, 4, we can get:

$$\begin{aligned} |3DCB^{positive}| &\geq |3DCB^{positive}_{XY}| - (|3DCB| - |3DCB^{positive}_{Z}|) \\ &= |3DCB^{positive}_{XY}| + |3DCB^{positive}_{Z}| - |3DCB| \end{aligned} \quad (5)$$

From equation 5, we can get:

$$\begin{aligned} &\frac{|3DCB^{positive}|}{|3DCB|} \\ &\quad \geq \frac{|3DCB^{positive}_{XY}| + |3DCB^{positive}_{Z}| - |3DCB|}{|3DCB|} \end{aligned} \quad (6)$$

Equation 6 can be rewritten as:

$$recall^{volume} \geq recall^{XY} + recall^{Z} - 1 \quad (7)$$

Since $recall^{volume}$ is supposed to be greater or equal to 0, we get:

$$recall^{volume} \geq \max(0, \quad recall^{XY} + recall^{Z} - 1) \quad (8)$$

□

3. Evaluate Frustum VoxNet Results based on Ground Truth 2D Bounding Box

3.1. Evaluation Metrics

We are using following metrics to evaluate the predictions based on ground truth 2D bounding boxes.

$$D_x = |x^* - x|, \quad D_y = |y^* - y|, \quad D_z = |z^* - z|$$

$$D_w = |w^* - w|, \quad D_d = |d^* - d|, \quad D_h = |h^* - h|$$

$$D_{xyz} = \sqrt{(x^* - x)^2 + (y^* - y)^* + (z^* - z)}|$$

$$D_{wdh} = \sqrt{(w^* - w)^2 + (d^* - d)^* + (h^* - h)}|$$

x^*, y^*, z^* are the predicted center and x, y, z are ground truth. w^*, d^*, h^* are the predicted width/depth/height and w, d, h are ground truth.

3.2. Evaluation Results

We compare the center prediction based on the frustum average center and the prediction from our Frustum VoxNet system. Table 1 provides the average distance between predicted and ground truth centers by using these two methods. As expected, the Frustum VoxNet prediction is better than the average center from frustum. Evaluation results for the performance of Frustum Voxnet based on frustums generated from ground truth bounding boxes are shown in Table 2. Histograms of the dot product between predicted orientations and GT orientations for each category are shown in Figure 3. Histograms of 3D detection IoU for each category are shown in Figure 4.

		$\overline{x - x^*}$	$\overline{y - y^*}$	$\overline{z - z^*}$	$\overline{D_{xyz}}$
Table	Frustum Average Center	-0.005	-0.233	0.075	0.522
	Predicted from Frustum VoxNet	0.014	-0.040	0.030	**0.395**
Desk	Frustum Average Center	-0.010	-0.198	0.109	0.428
	Predicted from Frustum VoxNet	0.028	-0.040	0.048	**0.319**
Sofa	Frustum Average Center	-0.015	-0.168	0.010	0.516
	Predicted from Frustum VoxNet	0.007	0.041	0.013	**0.444**
Bed	Frustum Average Center	0.031	-0.195	0.013	0.573
	Predicted from Frustum VoxNet	-0.009	0.010	-0.012	**0.354**

Table 1: Result comparison predicted between frustum center and predicted center from Frustum VoxNet.

4. ResNetFCN35 network structure

ResNetFCN35 network structure is shown in Figure 5.

	instance										average	3D recall
category	number	$\overline{D_x}$	$\overline{D_y}$	$\overline{D_z}$	$\overline{D_{xyz}}$	$\overline{D_w}$	$\overline{D_d}$	$\overline{D_h}$	$\overline{D_{wdh}}$	$\overline{\lvert o^* \cdot o\rvert}$	3D IoU	(IoU@0.25)
table	1269	0.201	0.280	0.070	0.395	0.206	0.132	0.042	0.287	0.747	0.319	0.656
desk	457	0.158	0.220	0.080	0.319	0.180	0.122	0.052	0.258	0.752	0.329	0.674
dresser	91	0.248	0.298	0.135	0.489	0.126	0.064	0.107	0.209	0.758	0.241	0.451
sofa	381	0.213	0.320	0.075	0.444	0.210	0.099	0.048	0.264	0.847	0.459	0.796
bed	441	0.195	0.220	0.096	0.354	0.154	0.125	0.083	0.246	0.746	0.462	0.898
night stand	220	0.156	0.226	0.069	0.314	0.050	0.037	0.044	0.087	0.830	0.329	0.655
bathtub	37	0.162	0.114	0.067	0.226	0.134	0.071	0.040	0.173	0.805	0.383	0.811
chair	4777	0.118	0.217	0.067	0.286	0.038	0.048	0.047	0.089	0.886	0.369	0.708
sofa chair	575	0.109	0.168	0.070	0.242	0.058	0.051	0.045	0.103	0.840	0.466	0.849
garbage bin	248	0.065	0.098	0.050	0.145	0.043	0.035	0.042	0.082	0.760	0.384	0.782
toilet	87	0.051	0.093	0.073	0.148	0.028	0.039	0.047	0.076	0.825	0.498	0.929
bookshelf	106	0.183	0.303	0.130	0.433	0.410	0.063	0.149	0.474	0.880	0.345	0.679

Table 2: **Detail evaluation results.** Frustum VoxNet is evaluated based on SUN-RGBD validation set. Frustums used to finalize detection are generated from ground truth 2D bounding boxes. The 3D IoU threshold used for 3D recall is 0.25.

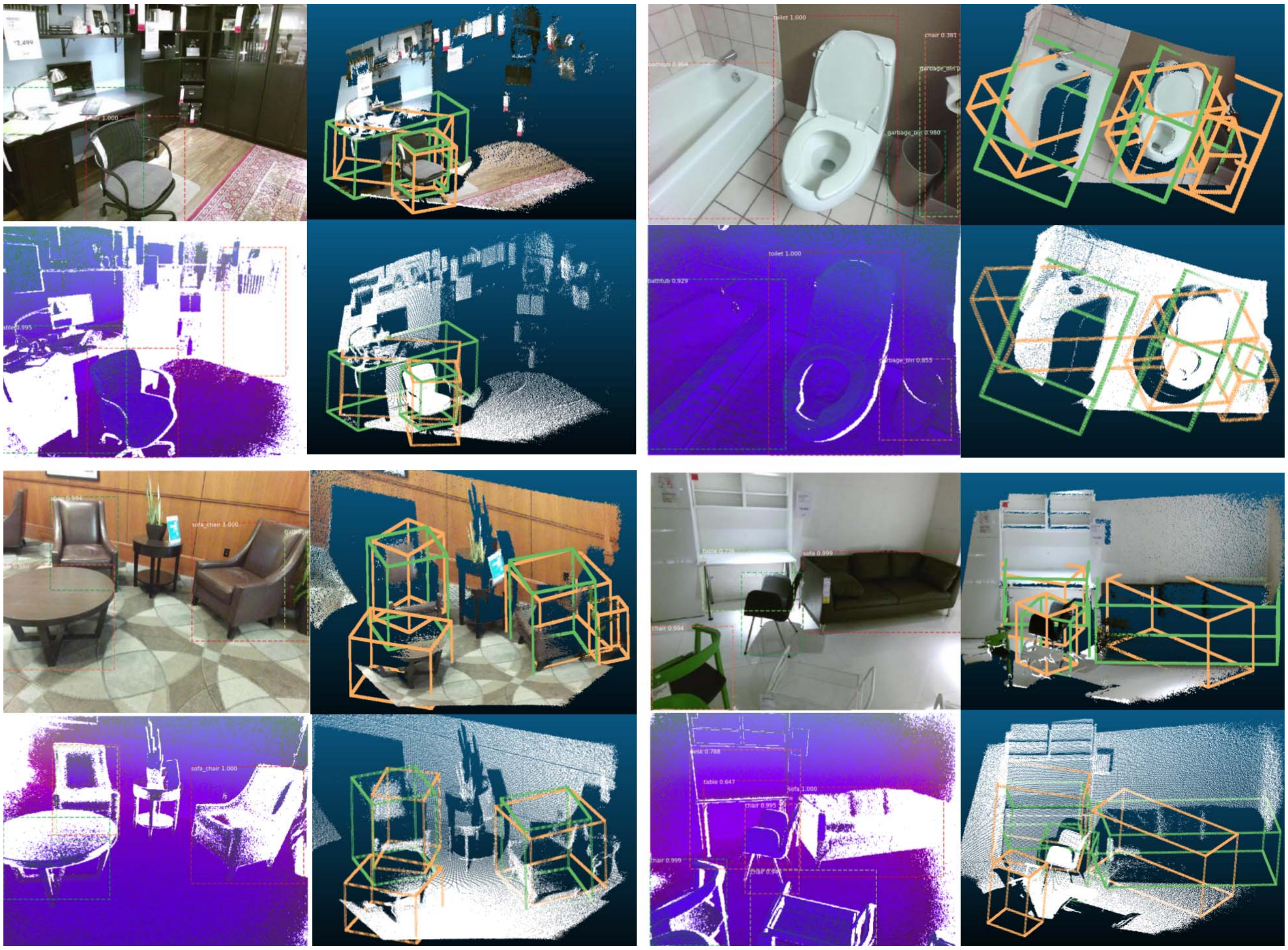

Figure 1. Visualizations of 2D and 3D detection results part 1. This visualization contains four images. For each image, upper left shows 2D detection based on RGB image. Upper right shows the corresponding 3D detection results (light green ones are the 3D ground truth boxes and orange-colored boxes are predictions) based on frustums generated from RGB image 2D detections (to have a better visualization, RGB colors are projected back to the cloud points). Lower left shows 2D detection based on DHS image. Lower right shows the corresponding 3D detection results (light green ones are the 3D ground truth boxes and orange-colored boxes are predictions) based on frustums generated from DHS image 2D detections. For the first image in the first row, our system can perfectly detect the chair. For the desk, the orientation is off as the frustum generated by the 2D bounding box contains some cloud points from the chair. For the second image, we can see that the based on RGB image system detect more false positive objects in the 2D stage and hence more 3D false positive objects will be detected. For the first image of the second row, our system successfully detect the unlabelled table. For the last image, the sofa's orientation is off as there are too many points are missing for the sofa.

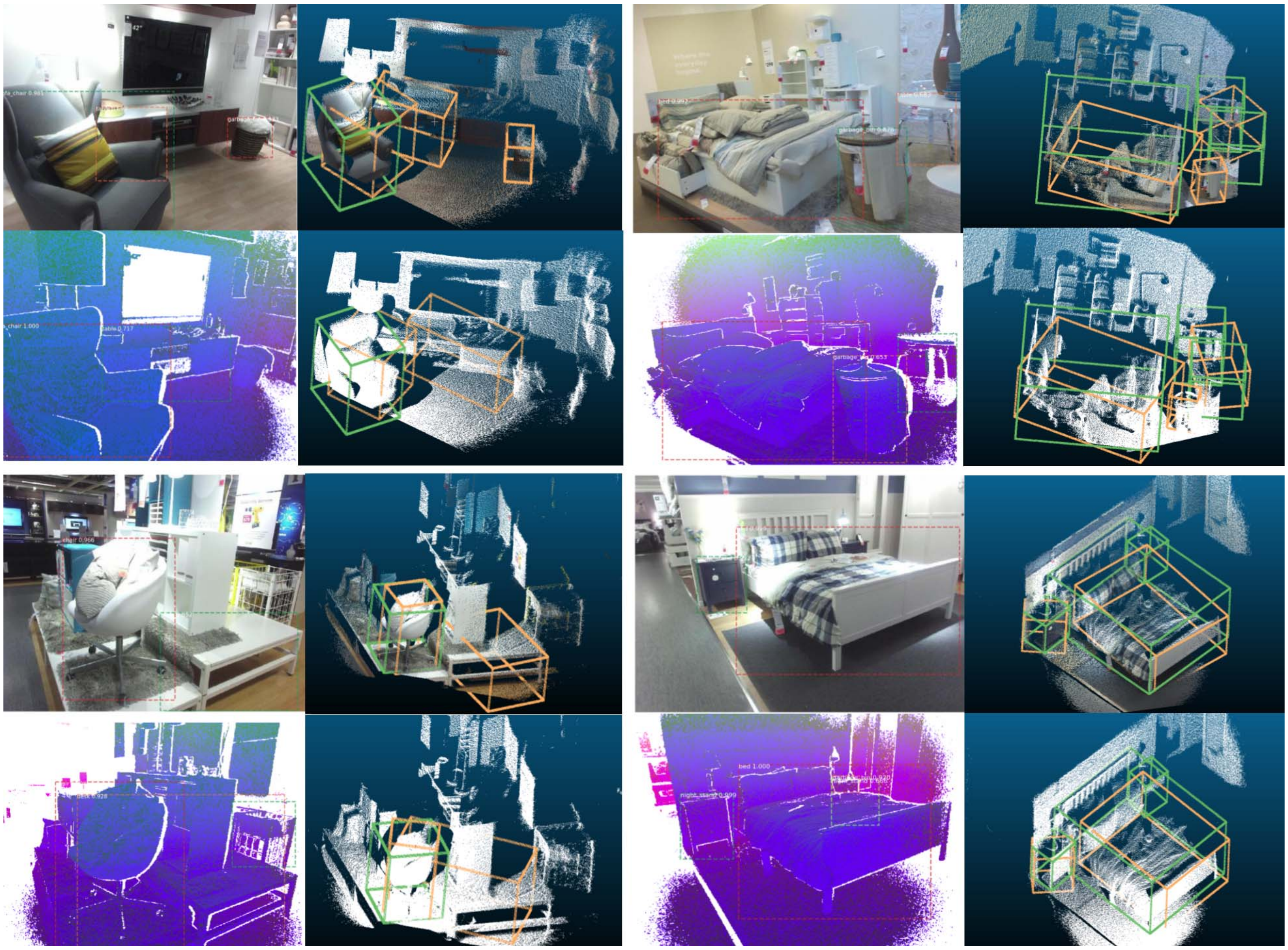

Figure 2. Visualizations of 2D and 3D detection results part 2. This visualization contains four images. Please read the caption of Figure 1 to get an explanation about how to understand the visualization. On the left part, our system can successfully detect unlabelled object such as garbage bin and table. On the top right image, our system fails to detect one table in 2D detection stage as it is partially observed. For the last one, one night stand is undetected as it is blocked by bed.

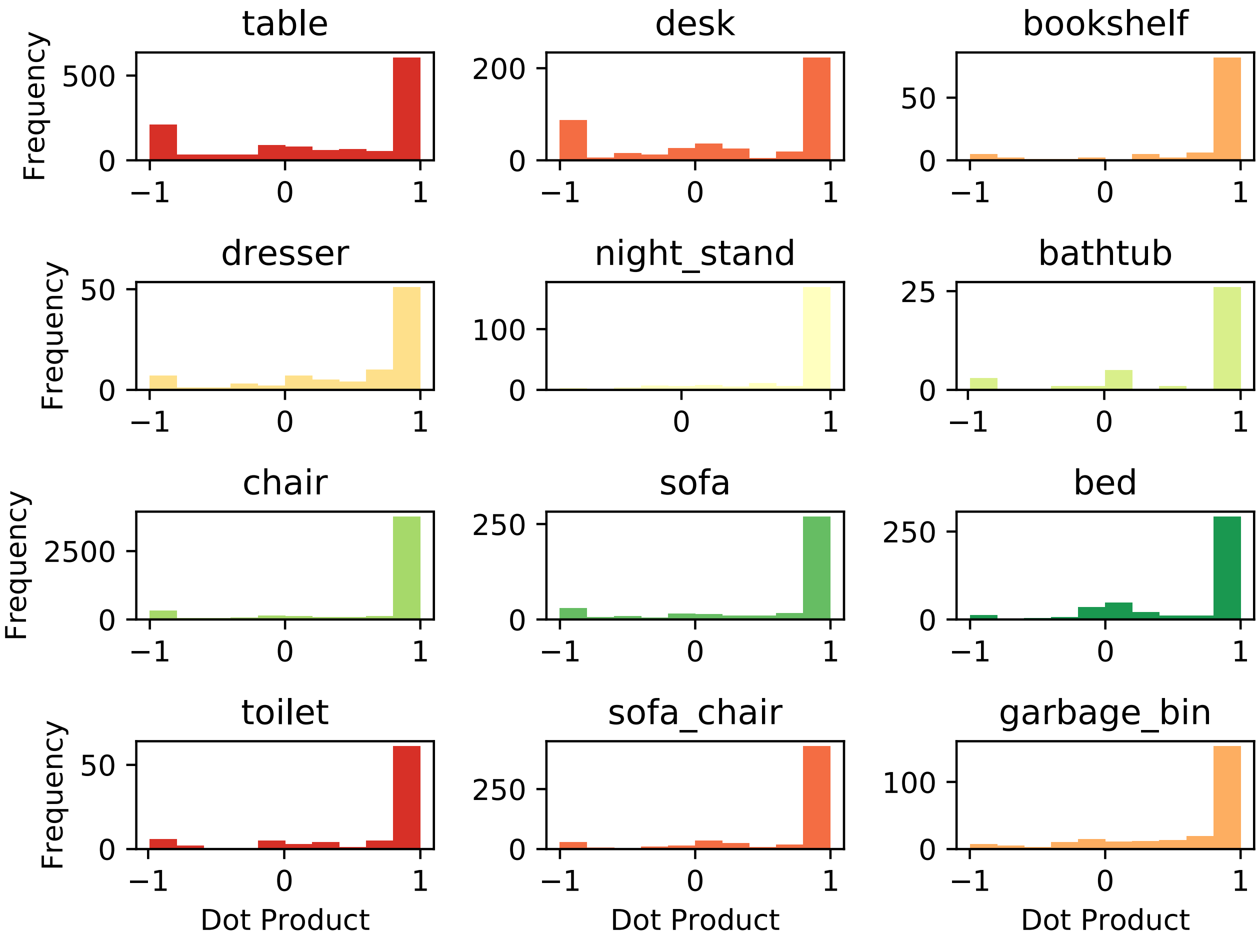

Figure 3. Histogram of the dot product between predicted orientation and GT orientation. Histograms are not normalized.

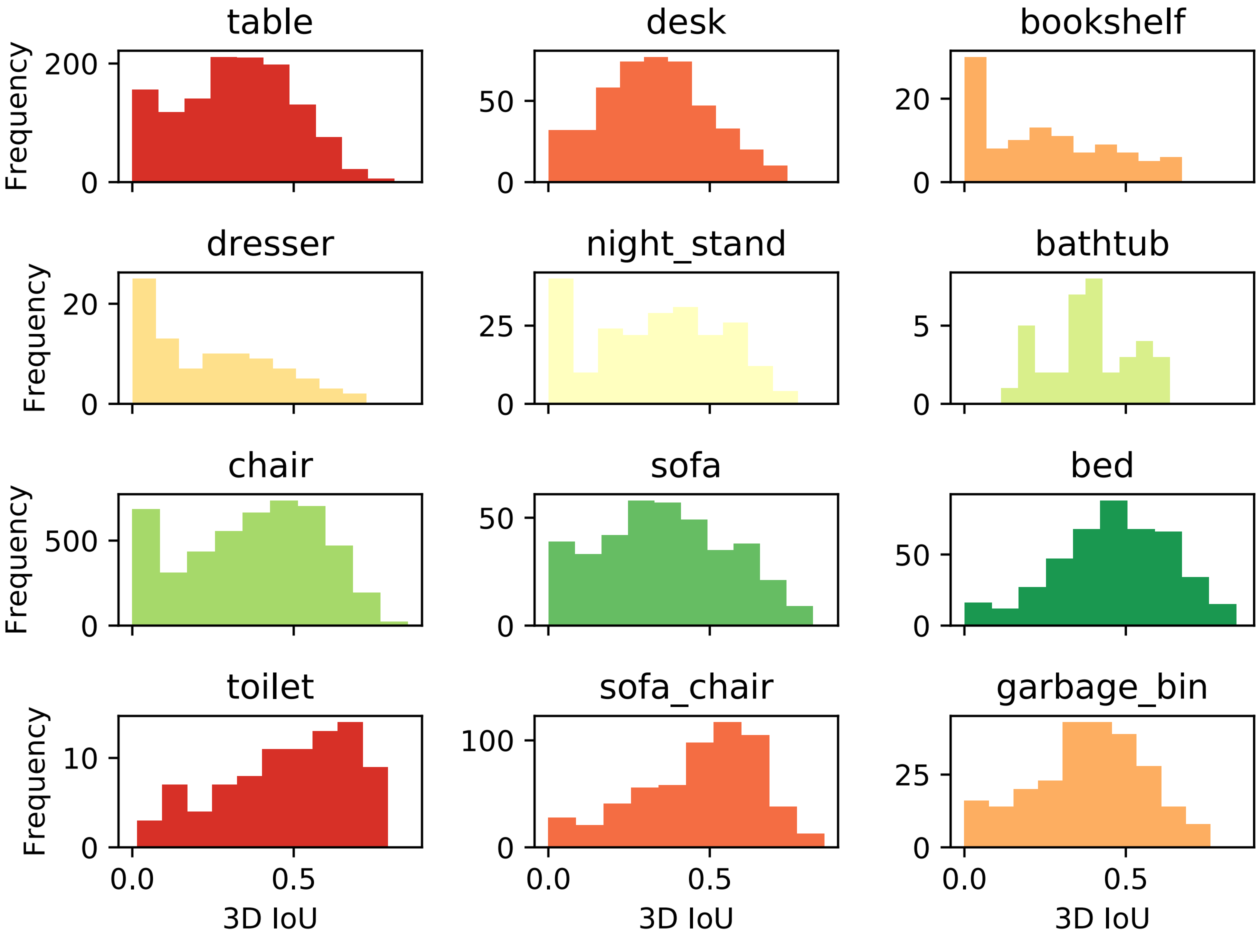

Figure 4. Histogram of 3D IoU. Histograms are not normalized.

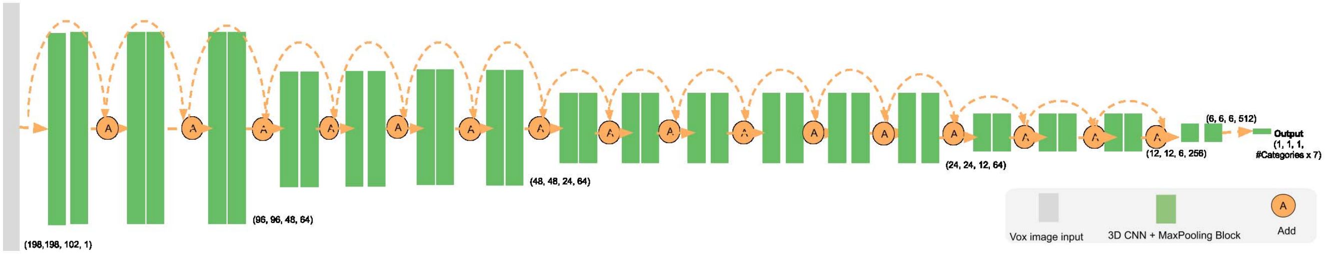

Figure 5. ResNetFCN35 network structure.