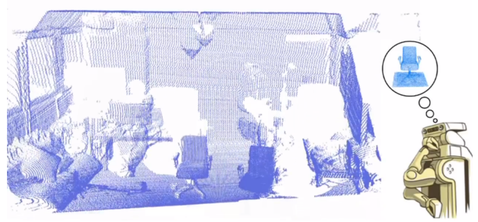
# Object Detection in 3D Scenes using CNNs in Multi-view Images

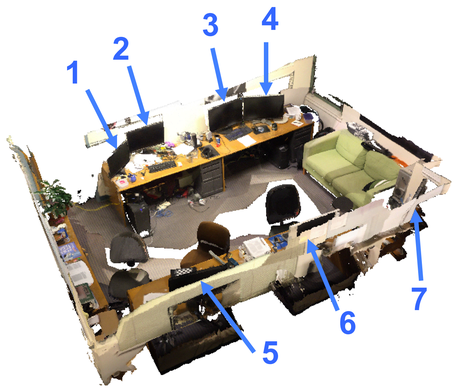
Ruizhongtai (Charles) Qi
Department of Electrical Engineering, Stanford University

## Motivation

**Semantic understanding** of 3D environment is important for VR content creation, AR reality mixture, robotics.



Given a 3D environment as point clouds or reconstructed surfaces, We want to achieve **visual Q&A** like the following:




**Q1:** What objects and where are they in the scene?

**A1:** There are chairs, monitors, desks, sofa and potted plant etc. in the scene.

**Q2:** How many monitors are there in the scene?

**A2:** There are 7 monitors.

## Related Work

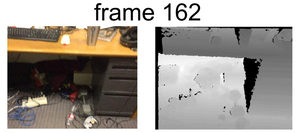
**Object Recognition in 2D and 2.5D:** faster-rcnn, rcnn-depth which focus on object detection from *a single image* (we utilize integrated inference from a series of images)
**Mesh Classification:** MVCNN, 3D ShapeNets, which focus on *mesh/3d model* classification (we work with real point clouds)
**Object Detection in 3D:** SLAM++, Monocular SLAM detection, which assumes models are known or use *hand crafted features*.

## System Pipeline

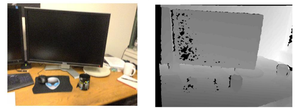
*From range sensor:*

frame 119

frame 162



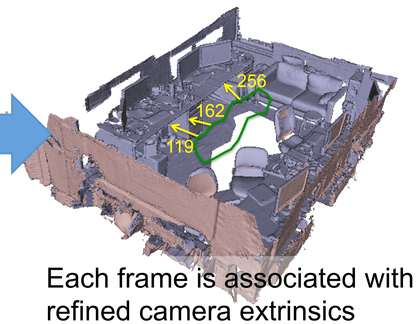
frame 265



…

4000 frames in total

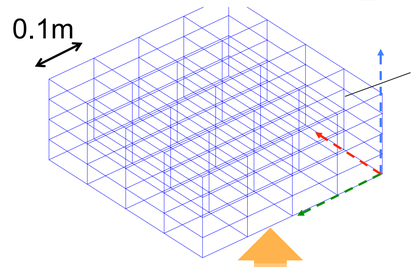
*Registered point clouds:*




Each frame is associated with refined camera extrinsics

Object detection on 2D images by faster rcnn with RPN and ORN

$$VoxelScore(i,j,k,c) = \frac{\sum_P Score(Image(p),c) I[Voxel(p)=(i,j,k)]}{\sum_P I[Voxel(p)=(i,j,k)]}$$




"aggregated" scores for voxel grid at (1,1,3)

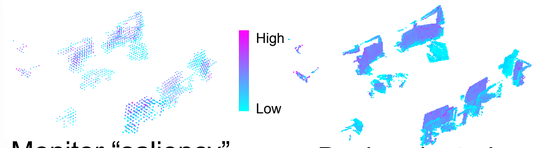
Scores back projected into 3d voxel grids by voxel hashing


…

## Experimental Results

How many  are there?

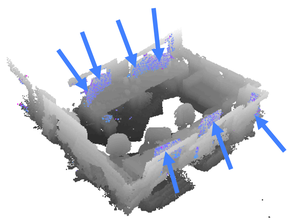
200 frames used, point cloud subsampled by 50x.
Negative voting successfully filters out bkg points




Monitor "saliency" (voxel grid)

Backprojected monitor points

Backprojected points with negative voting



Overlay onto entire scene

## Next Steps

Technical Problems

1. Choice of key frames, based on frame quality and camera poses.
2. Shape prior for instance segmentation and 2D/3D alignment.
3. Joint-optimization among categories e.g. space occupancy exclusions.