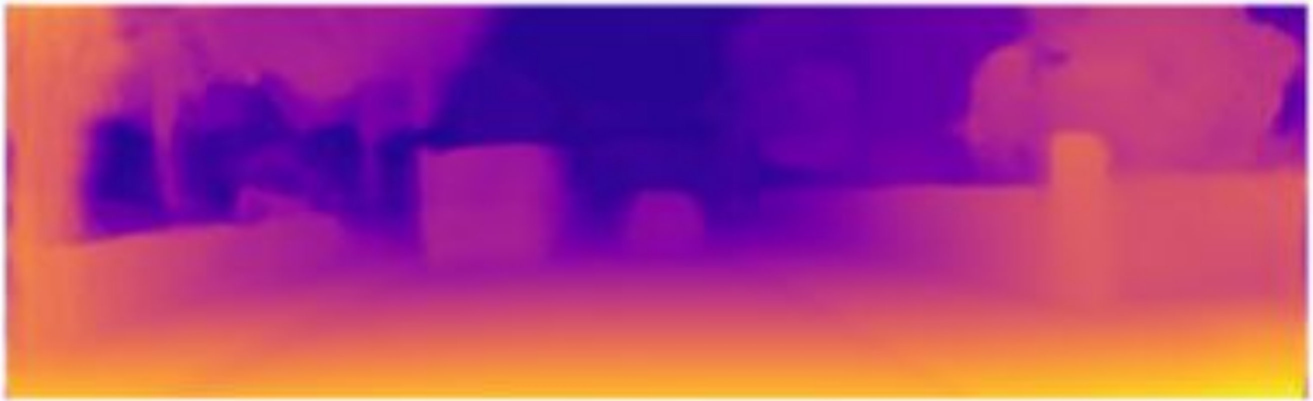
MONOCULAR DEPTH ESTIMATION FOR VISUAL SLAM

Nitin Nataraj
Siddharth Sharma

University at Buffalo
School of Engineering and Applied Sciences

Introduction

- Depth estimation refers to the set of techniques and algorithms aiming to obtain a representation of the spatial structure of a scene.
- In other terms, to obtain a measure of the distance of, ideally, each point of the visible scene.
- This information will help in estimation of landmark locations
- Has applications in vision-based SLAM



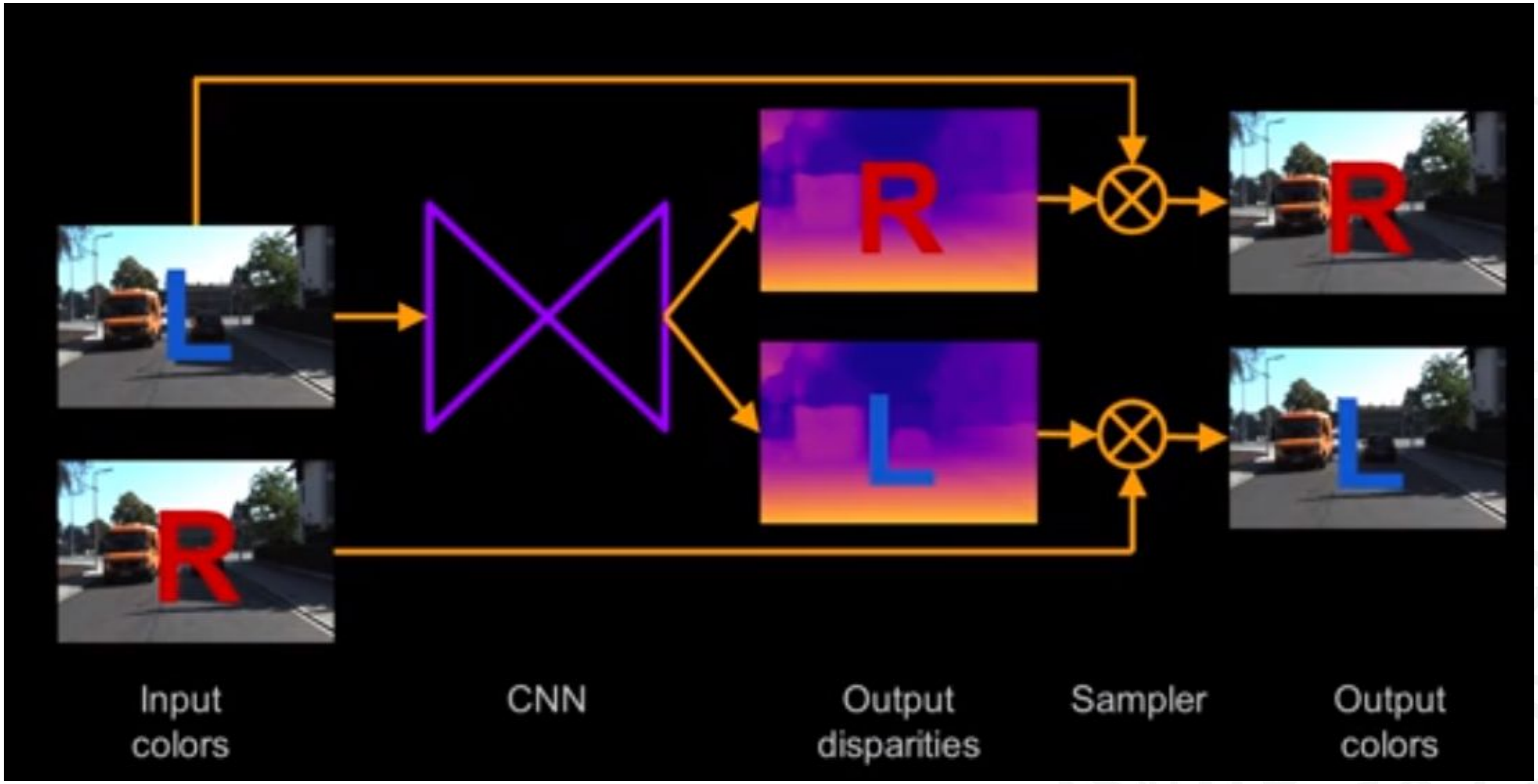
Current Issues

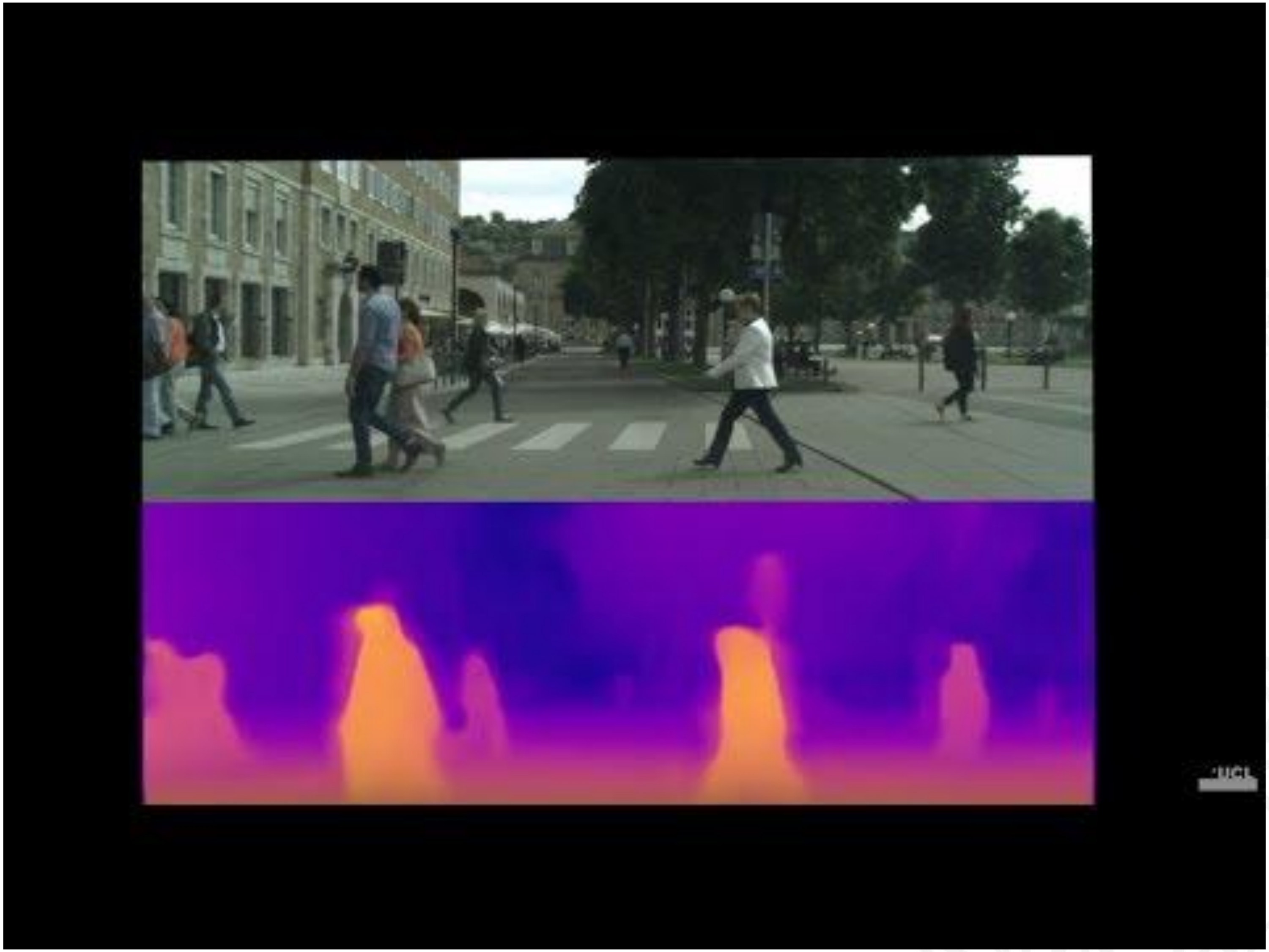
- Lots of problems with existing methods
- LIDAR sensors can be expensive
- Kinect throws problems in sunlight
- Problems with ORB SLAM
- Cheap and efficient to use monocular vision for depth estimation

What has been done?

1. A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. PAMI, 2009
2. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovit- ¨ skiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In CVPR, 2016
3. **Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." CVPR. Vol. 2. No. 6. 2017.**

Architecture




UCL

RGBD-SLAM - RTABMap

- Odometry estimated from consecutive images
- Point clouds calculated
- RTABMap comes with a GUI
- Front-end
  - Extract visual features in the RGB image, get the depth of these features using the Depth image, than do a RANSAC rigid transformation estimation with the previous image using the corresponding 3D features (correspondences are found by matching 2D visual features between the RGB images).
- Back-end:
  - The graph is created here, where each node contains RBG and depth images with corresponding odometry pose. The links are a transformation between each node. When the graph is updated, RTAB-Map compares the new image with all previous ones in the graph to find a loop closure. When a loop closure is found, graph optimization is done to correct the poses in the graph.
- Visualization:
  - For each node in the graph, a point cloud is generated from the RGB and depth images. This point cloud is transformed using the pose in the node. The 3D map is then created.

Depth from disparity

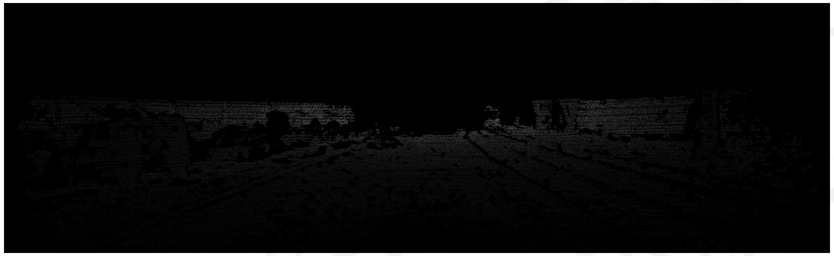
- depth = baseline * f/disparity
  - baseline - distance between stereo camera centers
  - f - focal length
- Problem - we may not have a baseline distance available as test images are not in a stereo setting
- Experimental methods
  - Compare with ground truth values to obtain a depth scale
  - Apply thresholding and scaling
    - Simple thresholding
    - Stacked thresholding
    - Average ground truth scaling
    - Temporal scaling

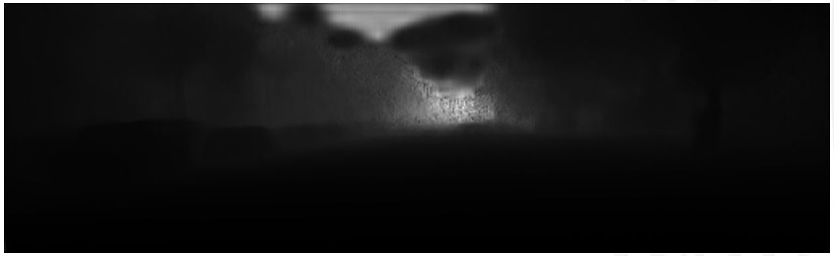
.

Results

Averaged Stacked Scaling

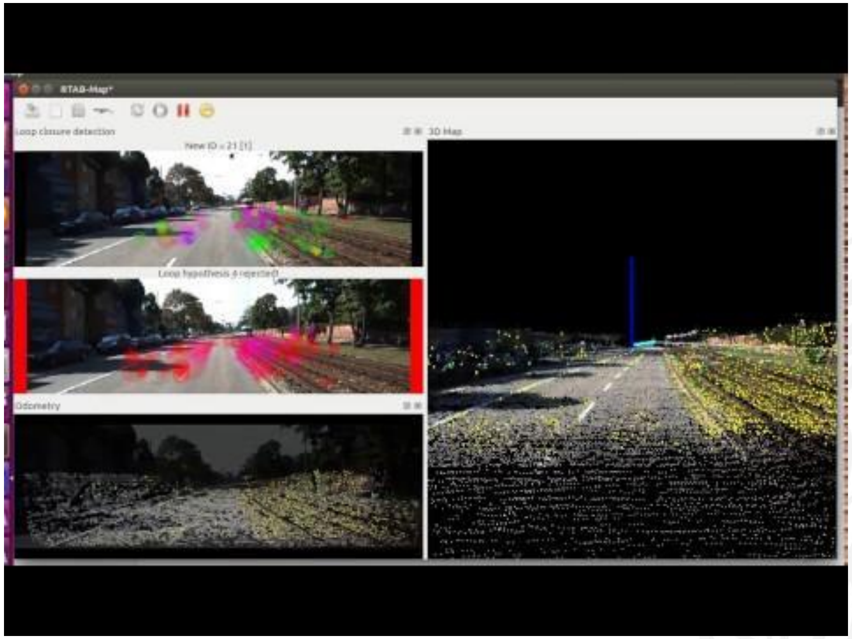
Original





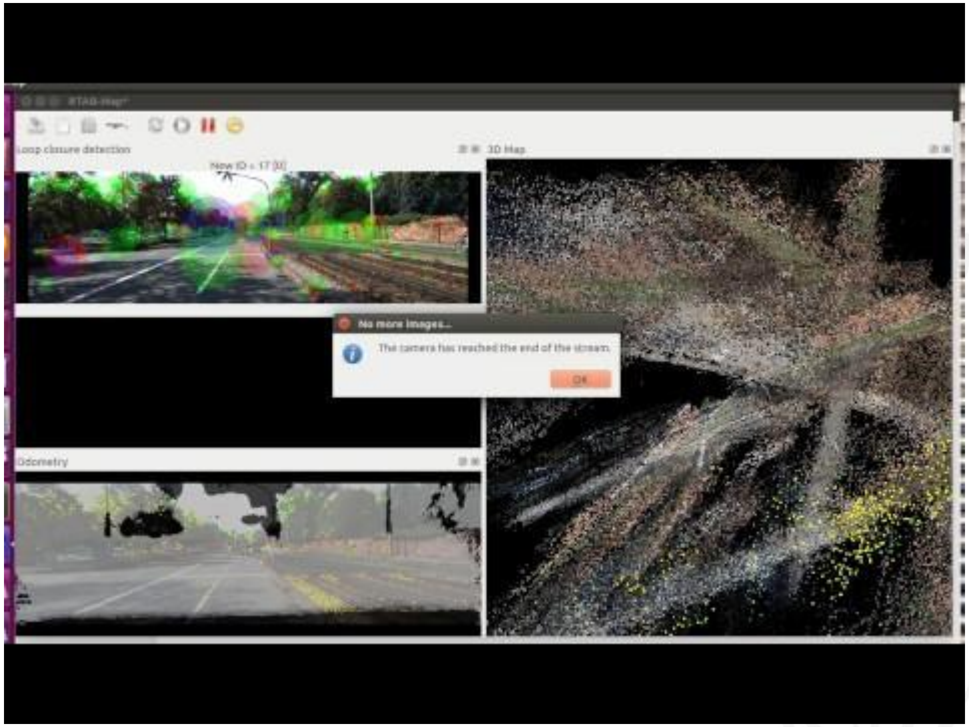
With formula

Ground Truth



With Formula

Averaged Stacked Scaling

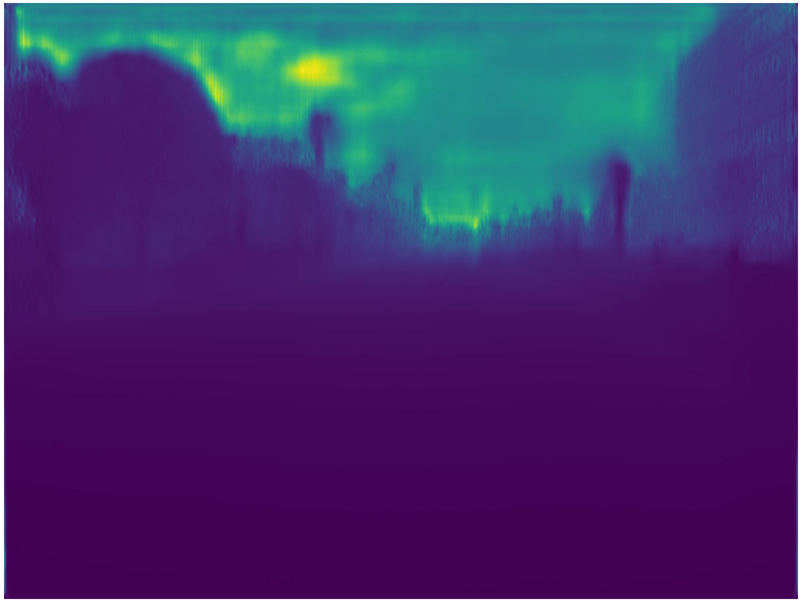


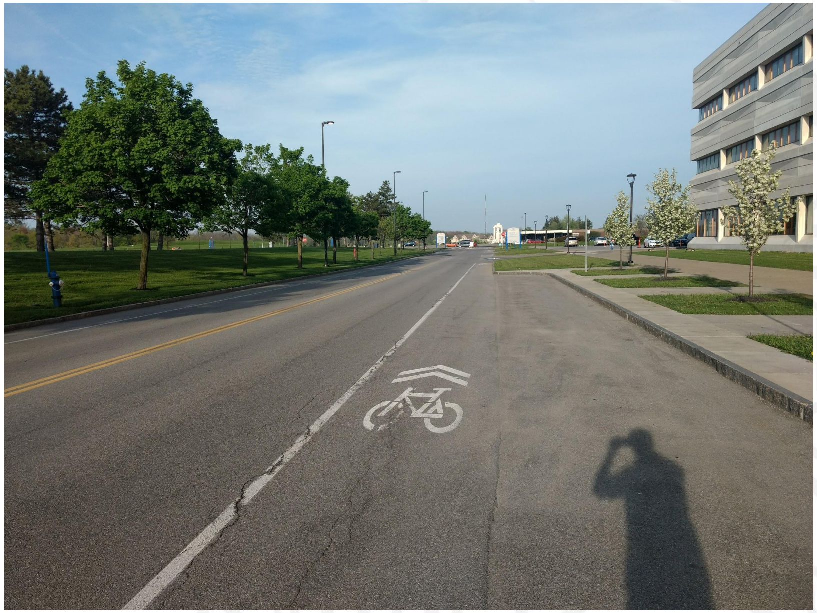
No more images...

The camera has reached the end of the stream.

Results

- Results on photos taken on own:





Progress

- Reviewed existing literature on depth estimation from monocular images
- Downloaded existing pre-trained models for both KITTI and Cityscapes dataset
- Installed RTABMAP SLAM (A visual slam library that works with stereo and depth images)
- Ran RTABMap mapping on readily available KITTI RGB and Depth Images
- Used Monodepth model to obtain disparity images from RGB images and ran RTABMap mapping on these images.
- Cityscapes model worked better with the KITTI dataset.
- Wrote a ROS publisher node to publish camera_info, depth images and RGB Images to the topics required by RTABMap.
- Experimented with simple and stacked thresholding with smoothing

Challenges

- KITTI images much larger than what the network wants. Lost visual odometry with resizing.
- Determining camera parameters
- Obtaining depth images from disparity images (scaling issues)
- Model was slow to run on the CPU.
- GPU issues on Ubuntu.
- Syncing the images from the three ROS topics that RTABMap subscribed to.
- Generalizing model to unseen data.

Future Work

- Obtaining real data
- Scaling estimation using Kinect. Using autoencoders to estimate the scaling function from given ground truth samples.
- Real-time with GPU integration
- ROS image syncing

References

- Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." CVPR. Vol. 2. No. 6. 2017.
- Labbe, Mathieu, and Francois Michaud. "Appearance-based loop closure detection for online large-scale and long-term operation." IEEE Transactions on Robotics 29.3 (2013): 734-745.
- Labbé, Mathieu, and François Michaud. "Memory management for real-time appearance-based loop closure detection." Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, 2011.
- Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. "Make3d: Learning 3d scene structure from a single still image." IEEE transactions on pattern analysis and machine intelligence 31.5 (2009): 824-840.
- Geiger, Andreas, et al. "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research 32.11 (2013): 1231-1237.

Questions?