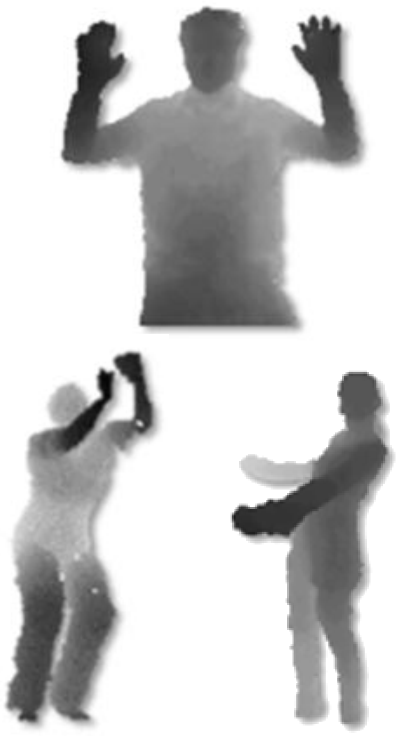
Aron Yu
Nov 2, 2012

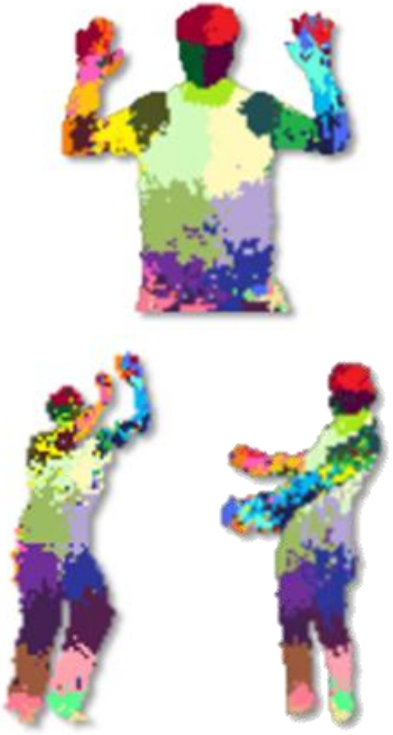
Object Recognition with Single Depth Images

Pose Recognition in Parts

Depth Image



Body Parts



3D Joint Est.

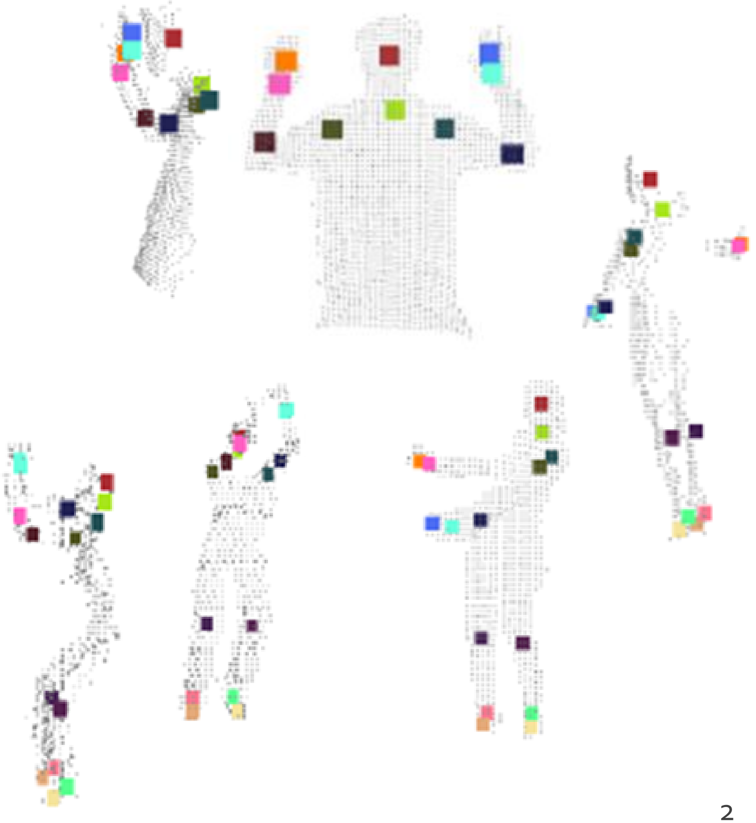
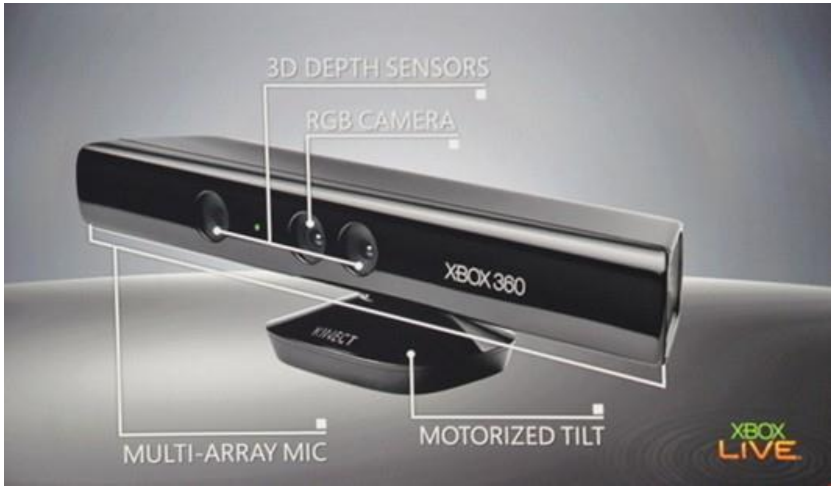


Image Credit: Shotton et al. – Real-Time Human Pose Recognition in Parts from Single Depth Images

Microsoft Kinect

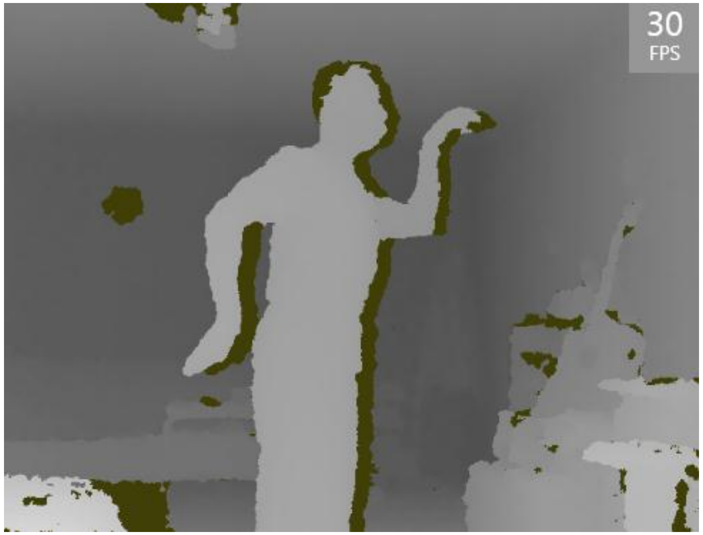


- **Released:** Nov 4, 2010
- **Color:** 640 x 480@ 32 bits
- **Depth:** 640 x 480 @ 16bits
- **Frame Rate:** 30/sec
- **Ideal Range:** 1.2m ~ 3.5m
- **Operational Range:** 0.7m ~ 6.0m
- **Tracking:** Up to 6 people, including 2 active players
- **Method:** 20-point joint tracking per player

- Opened doors to new research (and games)!

Source: www.xbox.com/en-US/kinect

Microsoft Kinect



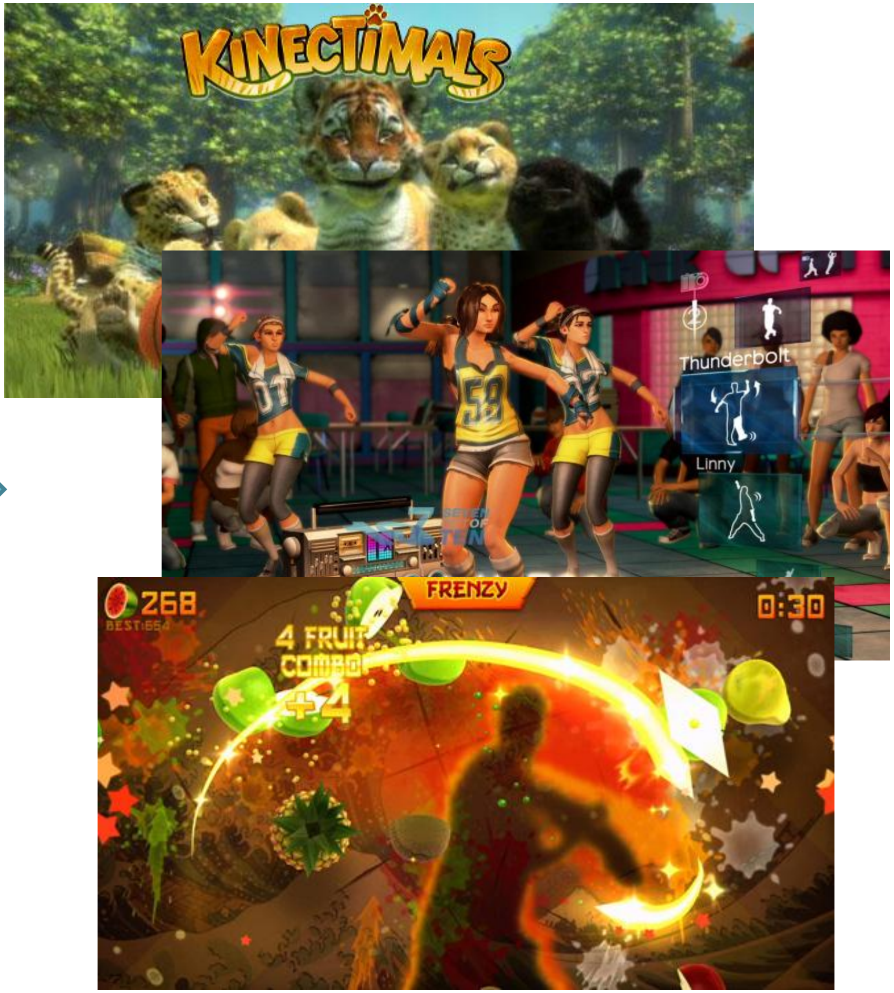
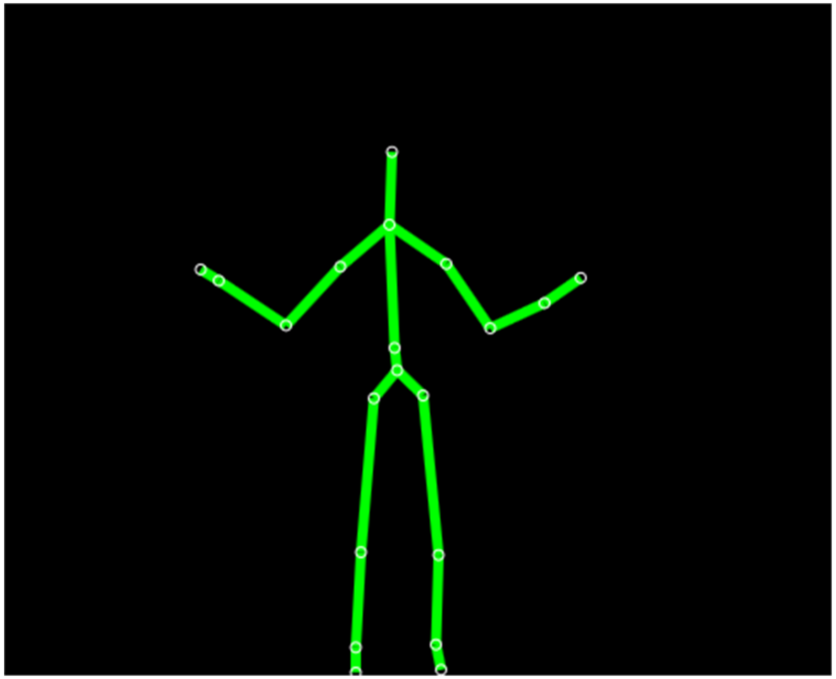


Image Credit: www.gamerant.com

Skeletal Images

Demo Time!

- Windows SDK 1.5 & Toolkit 1.6



Depth Feature

- Depth Comparison Feature
  - weak but efficient
  - offsets in pixel distance

depth invariant

$$f_\theta(I,\mathbf{x}) = d_I\left(\mathbf{x}+\boxed{\frac{\mathbf{u}}{d_I(\mathbf{x})}}\right) - d_I\left(\mathbf{x}+\boxed{\frac{\mathbf{v}}{d_I(\mathbf{x})}}\right)$$

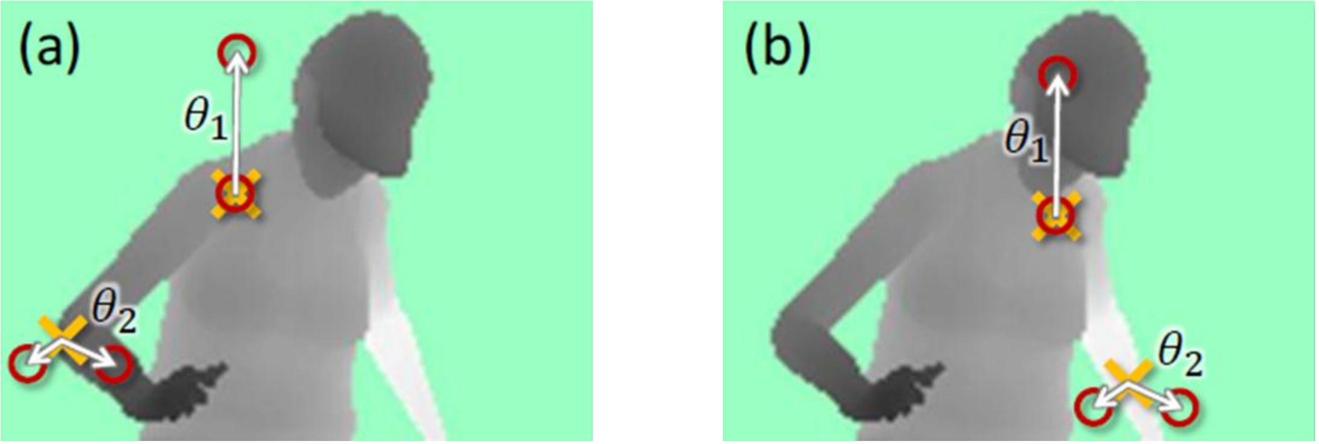
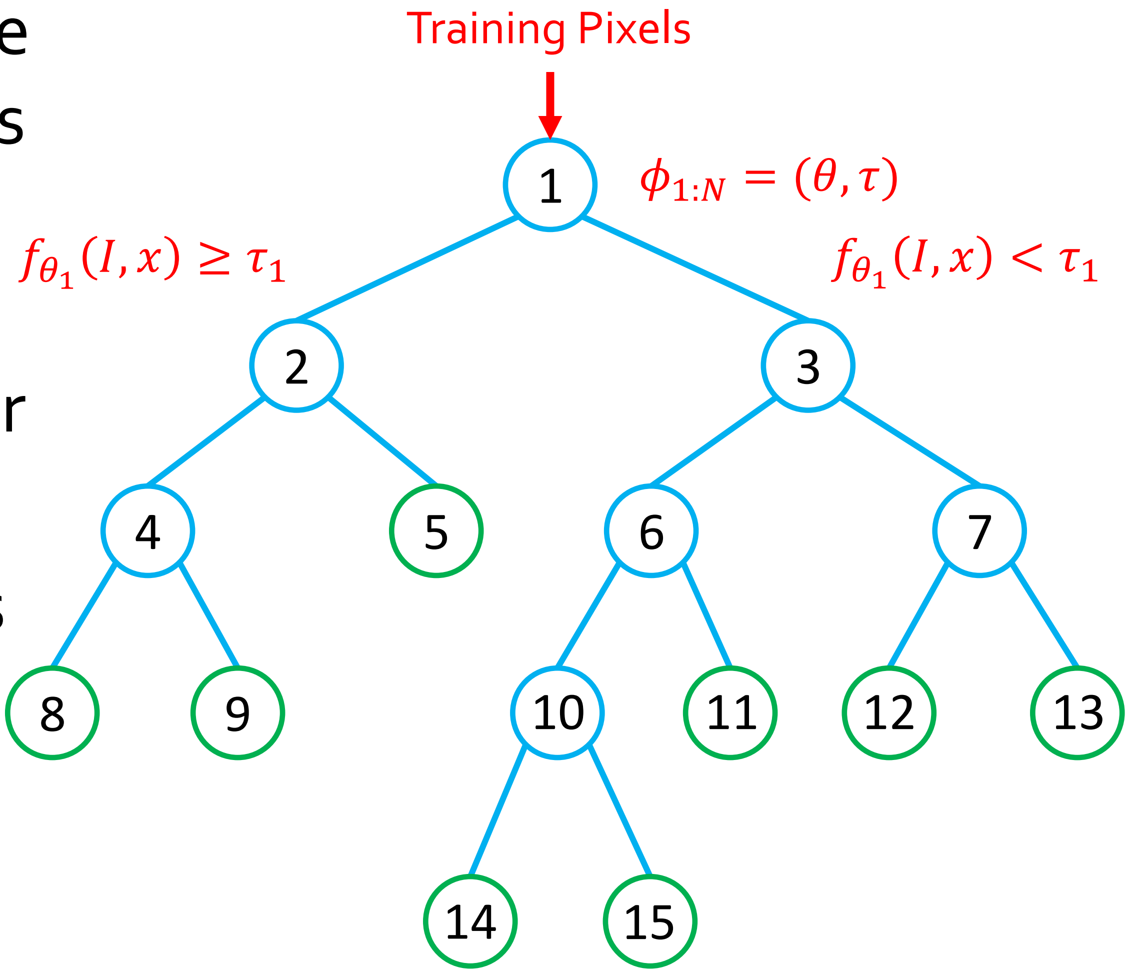


Image Credit: Shotton et al. – Real-Time Human Pose Recognition in Parts from Single Depth Images

Random Decision Tree

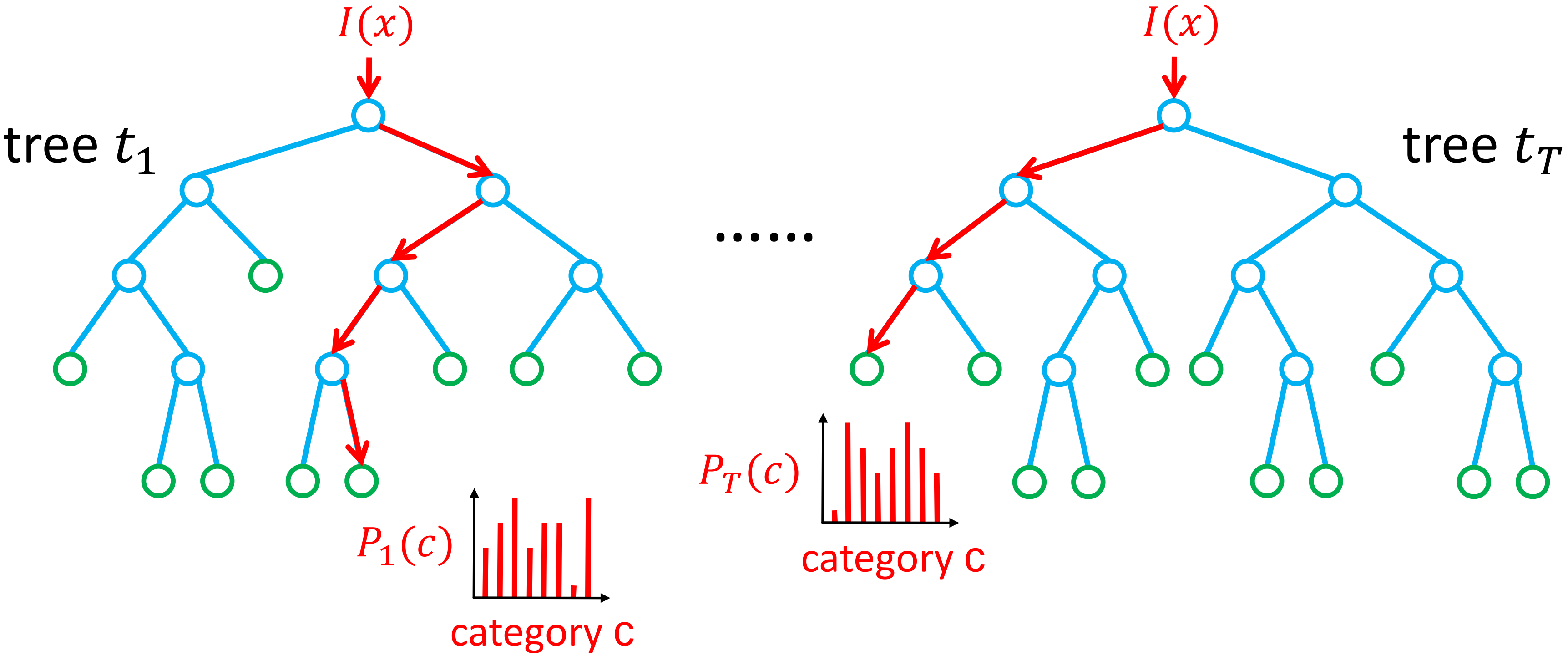
- Randomly generate splitting candidates at each node
- Partition training pixels and check for entropy gain
- Repeat until gain is minimal


Training Pixels
$\phi_{1:N} = (\theta, \tau)$
$f_{\theta_1}(I, x) \geq \tau_1$
$f_{\theta_1}(I, x) < \tau_1$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Source: www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial

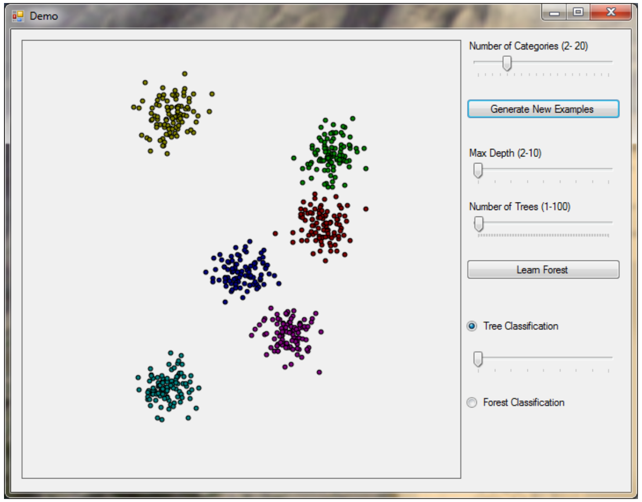
Random Decision Forest

- Ensemble of random decision trees
  - final distributions are averaged



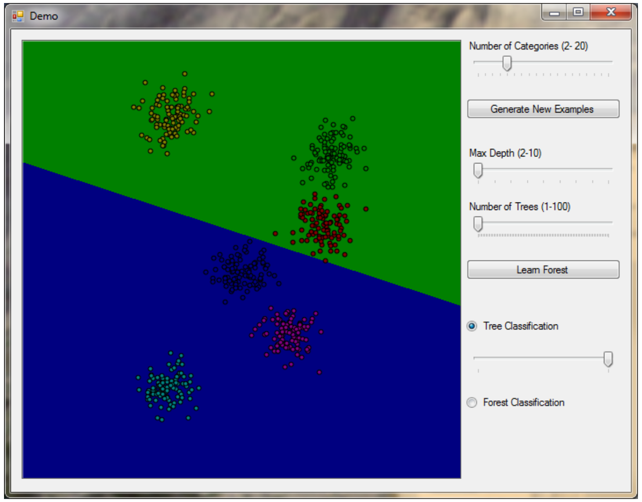
Source: www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial

Toy Demo



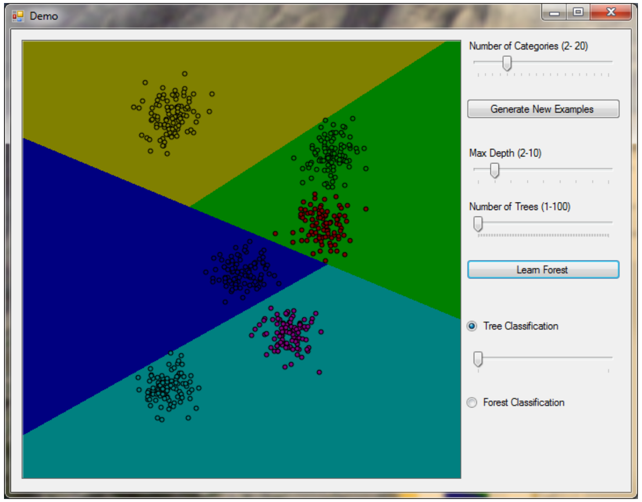
● Tree Classification

○ Forest Classification



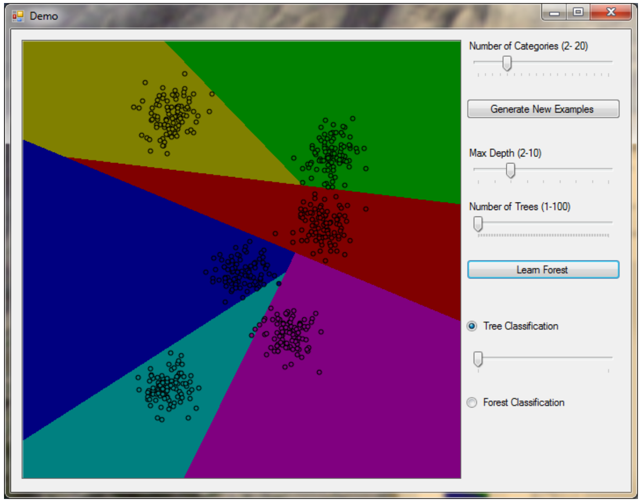
● Tree Classification

○ Forest Classification



● Tree Classification

○ Forest Classification

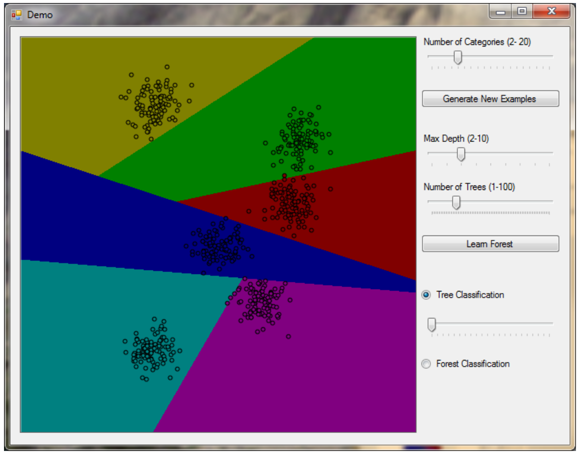


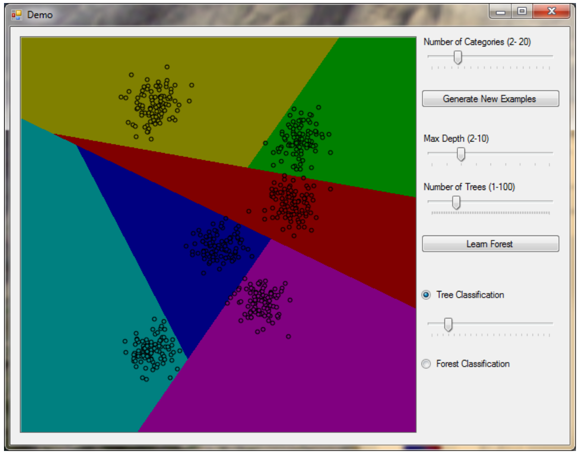
● Tree Classification

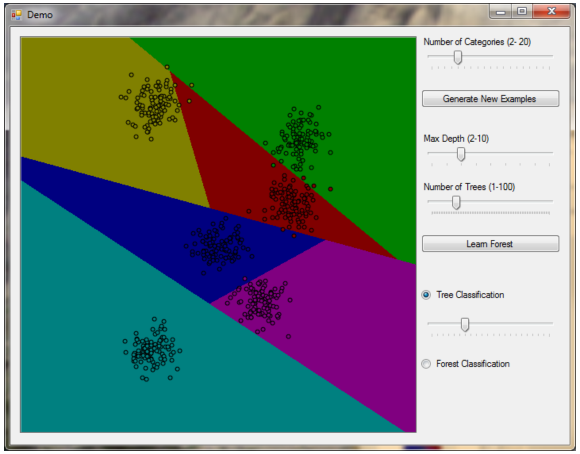
○ Forest Classification

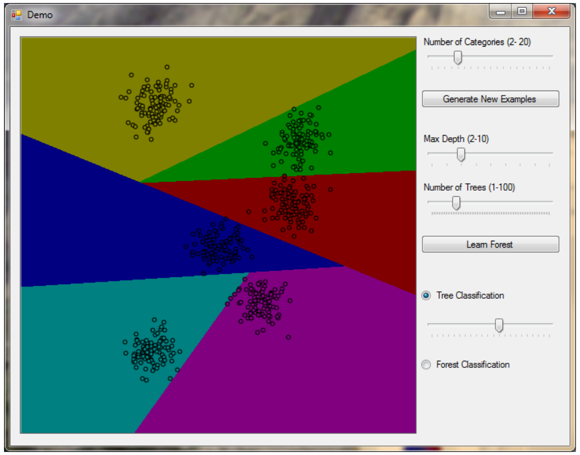
Image Credit: www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial

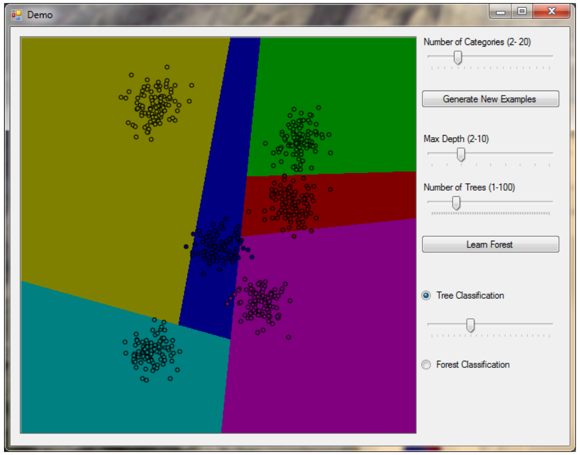
Toy Demo











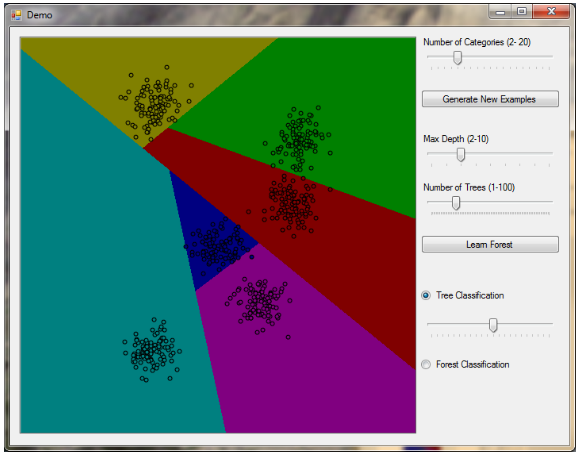
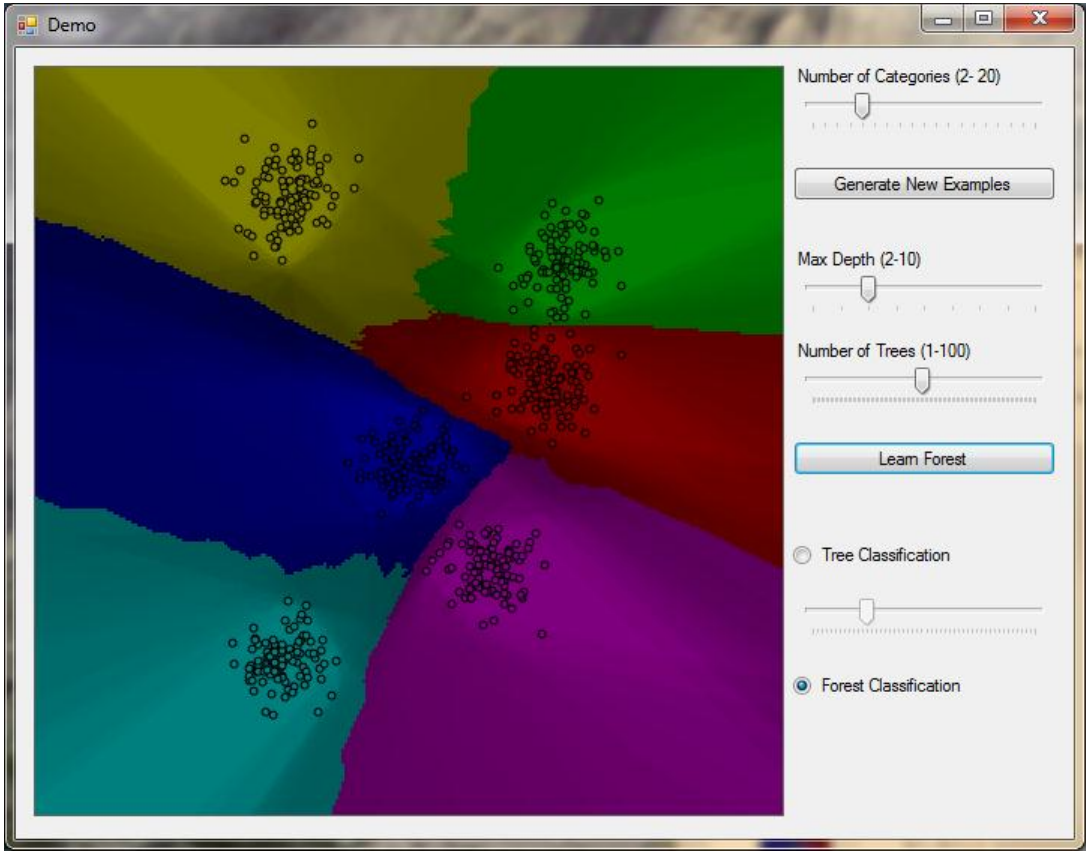


Image Credit: www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial

Toy Demo

Forest of 50 Trees



Number of Categories (2- 20)

Max Depth (2-10)

Number of Trees (1-100)

○ Tree Classification

● Forest Classification

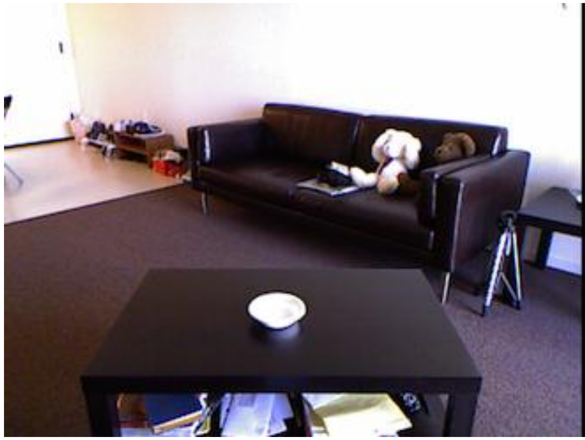
Image Credit: www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial

Experimental Setup

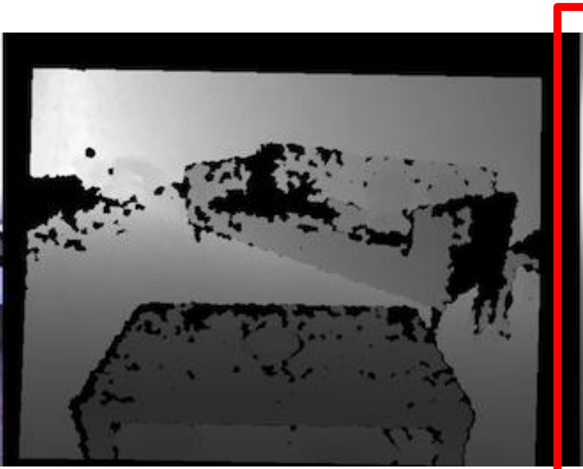
- B3DO dataset with objects (synthetic & real depth data)
  - bounding box ground truth (pixel-level ground truth)
- 300~350 training images (350k~1M images)
  - 2000~3000 pixels per image
- Fixed and random features (uv pairs)
  - 4~16 fixed, 50~150 random (2000 random features)
- TreeBagger function from Matlab
  - 16 trees, 80% of the samples used per tree
  - quad core computer w/ 16GB RAM (1000-core cluster)

B3DO Dataset

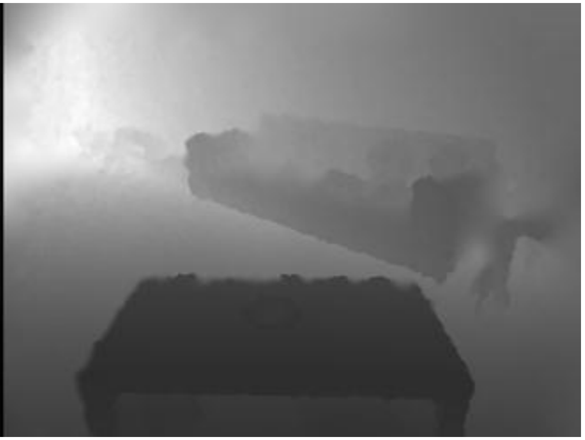
- Berkeley 3D Object Dataset
  - household object detection
  - 849 images (color, raw depth, smoothed)
  - 89 object classes



Color



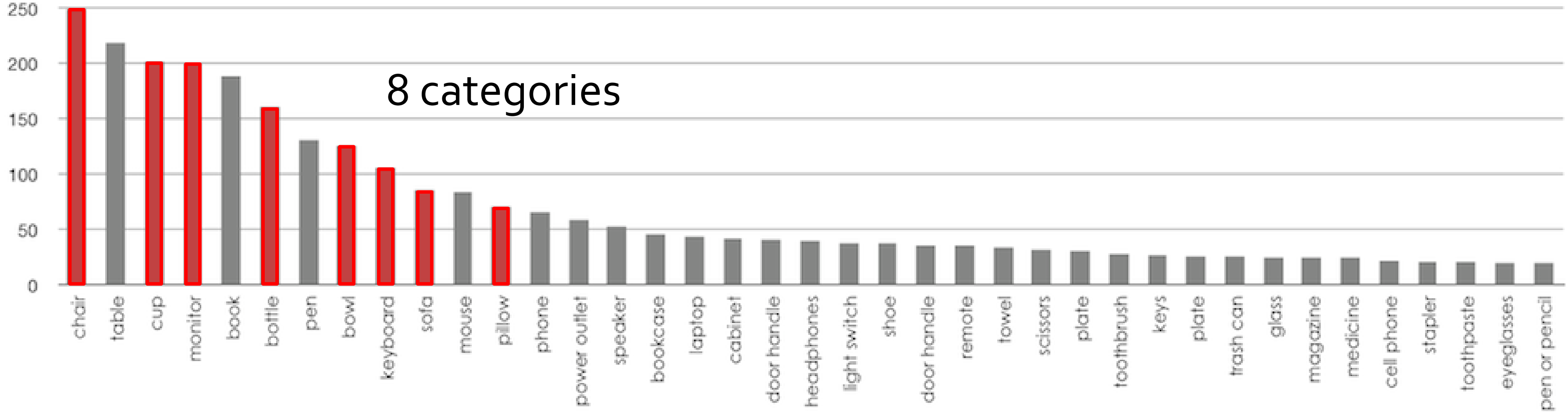
Raw Depth



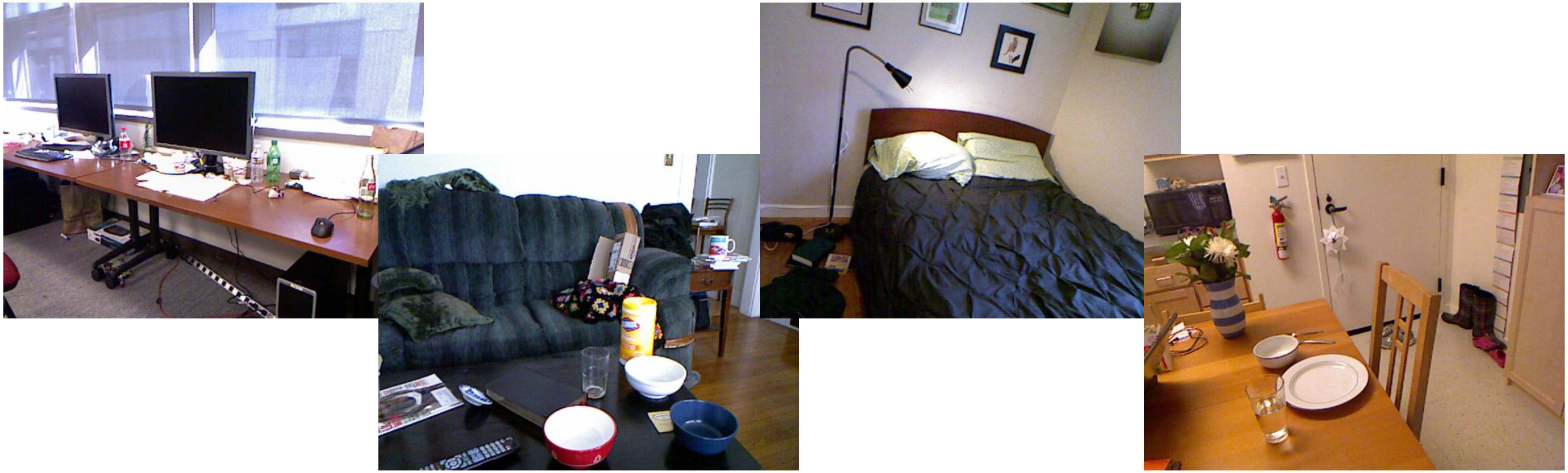
Smoothed

Source: Berkeley 3D Object Dataset (www.kinectdata.com)

B3DO Dataset



8 categories

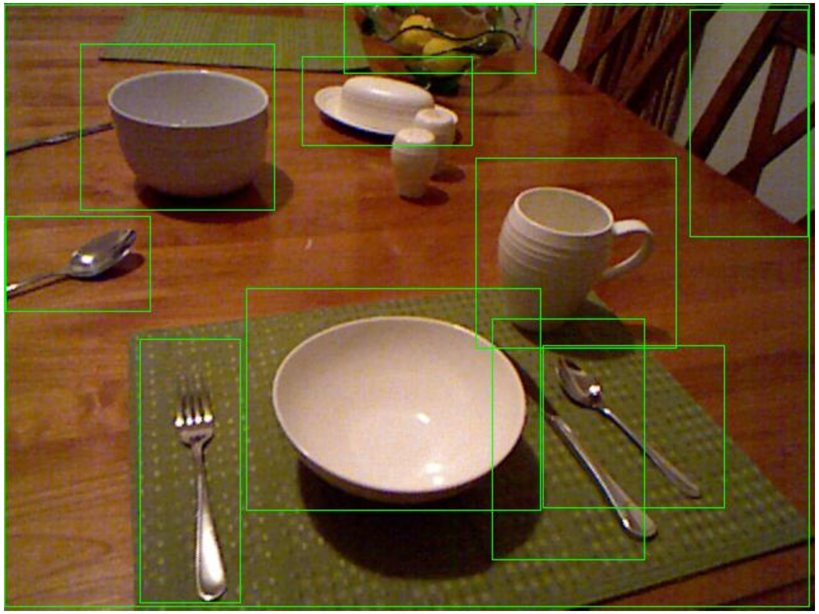


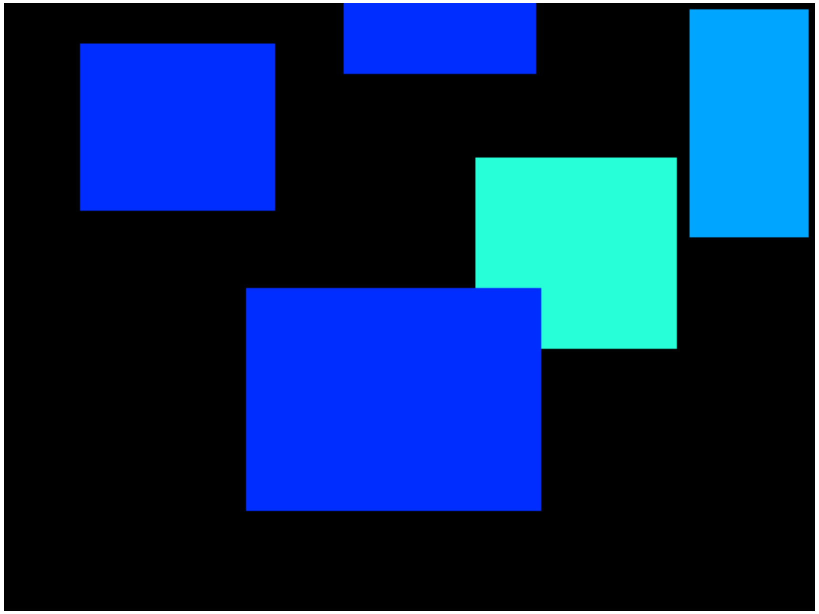
Source: Berkeley 3D Object Dataset (www.kinectdata.com)

Ground Truth

- VOC format bounding box
  - create pixel-level ground truth
  - inevitable overlaps

bottle	🟩	keyboard	🟨
bowl	🟦	monitor	🟧
chair	🟦	pillow	🟥
cup	🟩	sofa	🟥

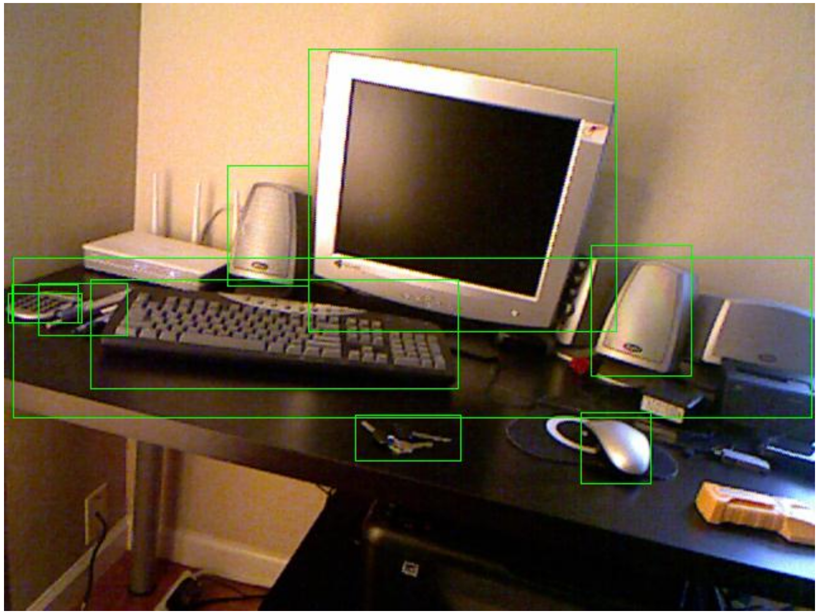


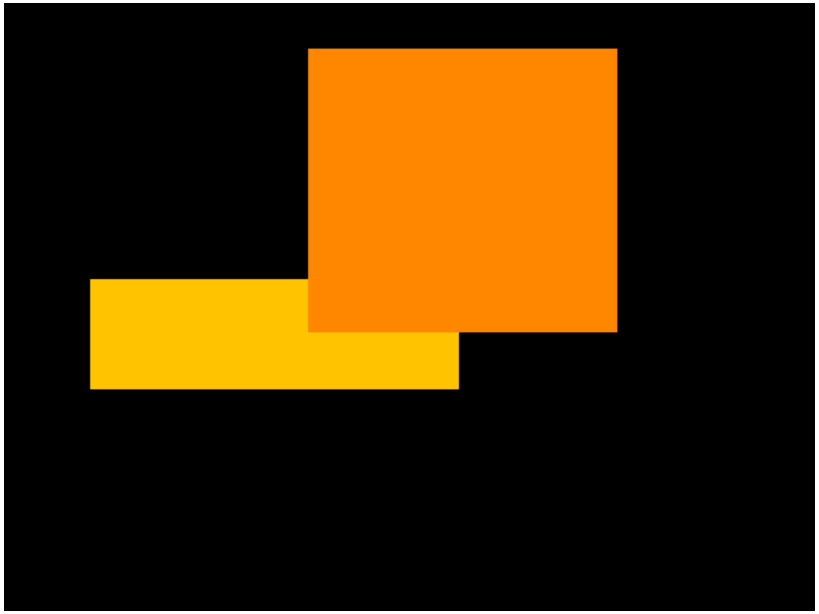


Ground Truth

- VOC format bounding box
  - create pixel-level ground truth
  - inevitable overlaps

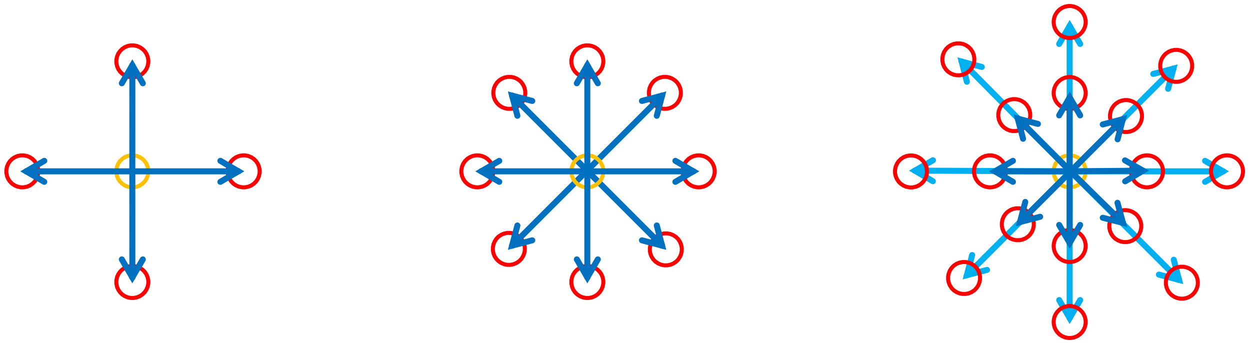
bottle	🟩	keyboard	🟨
bowl	🟦	monitor	🟧
chair	🟦	pillow	🟥
cup	🟩	sofa	🟥



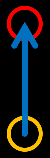


Feature Selection

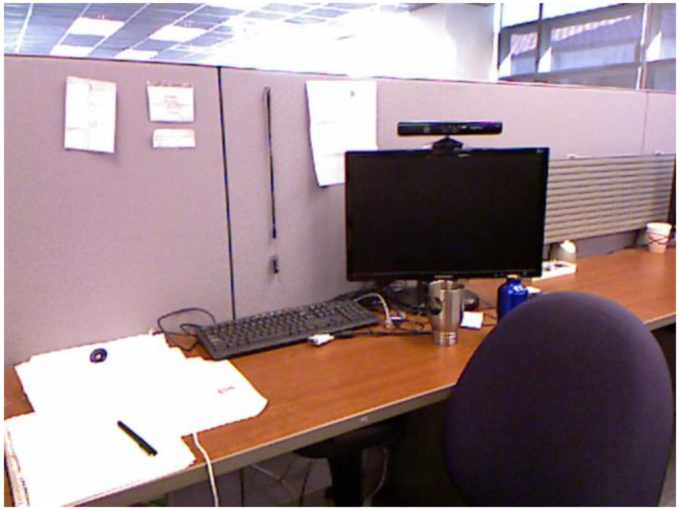
- Random features
  - body parts are deformable, each with unique shapes
  - find the best from large samples of random features
- Fixed features
  - household objects are rigid with defined shapes
  - might be sufficient with few known features



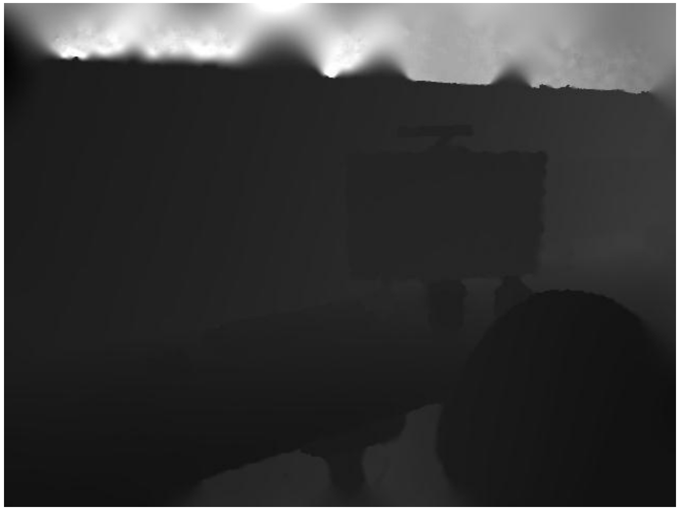
Feature Map



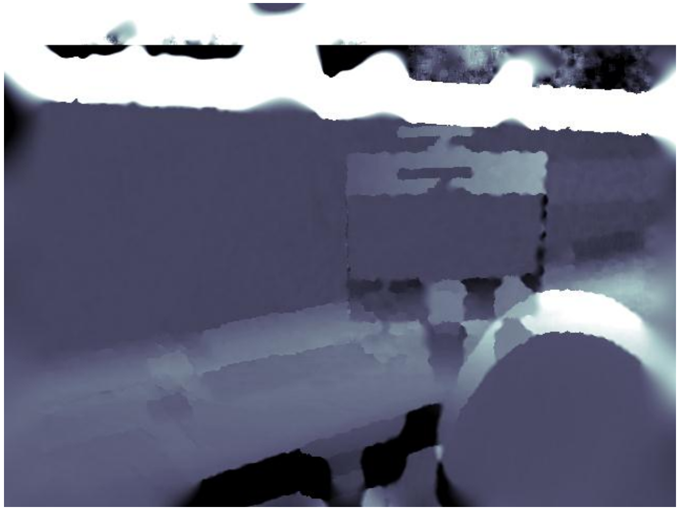
Color Image



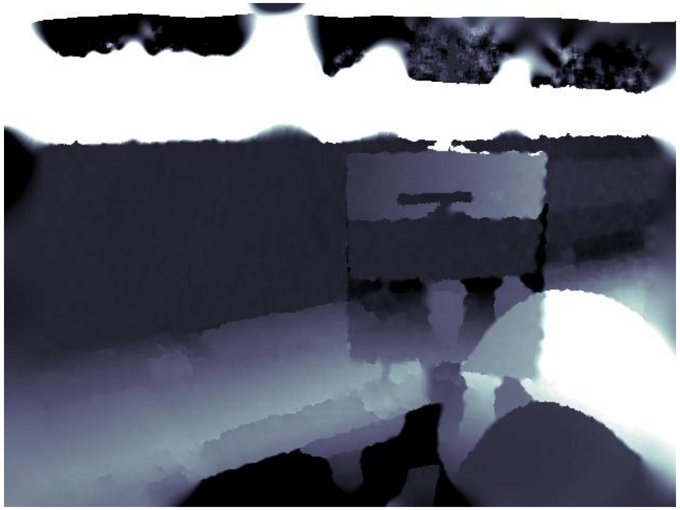
Depth Image



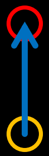
Not Normalized

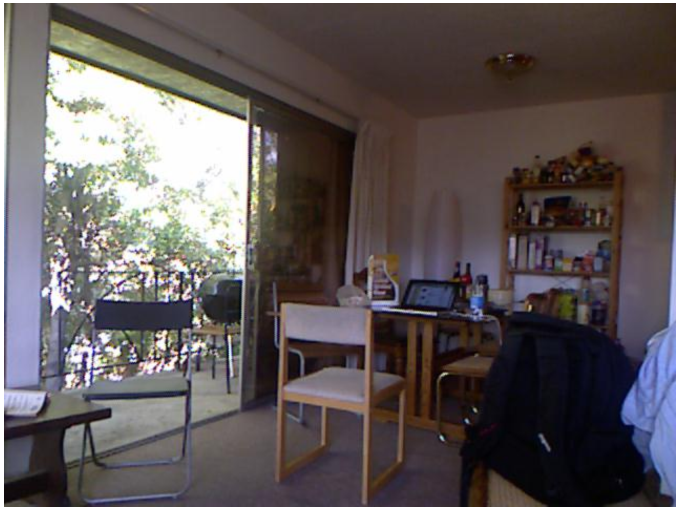


Normalized

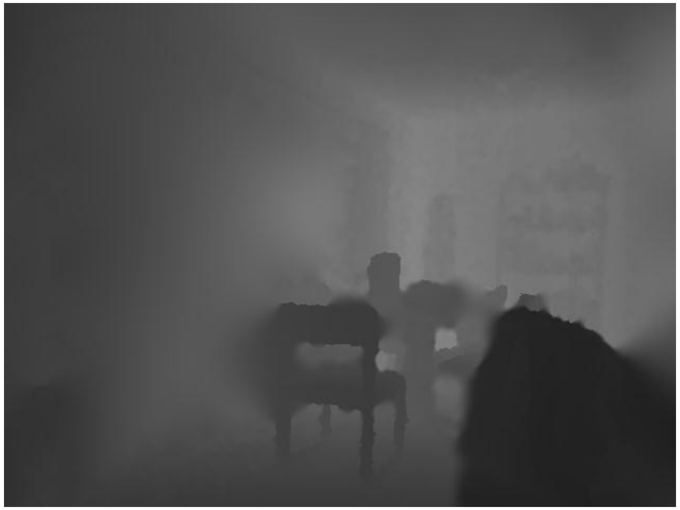


Feature Map

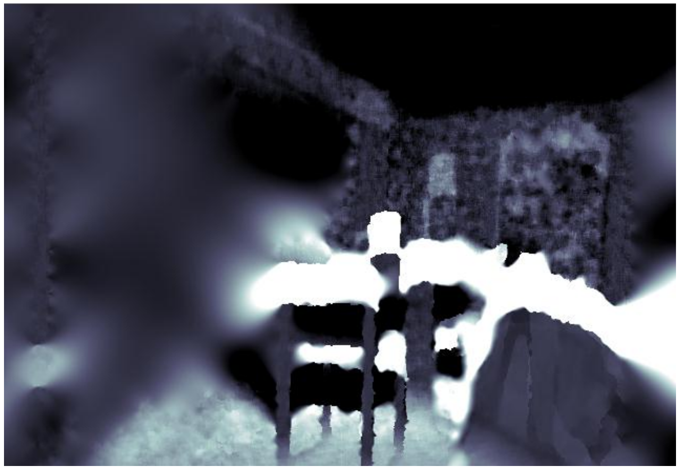




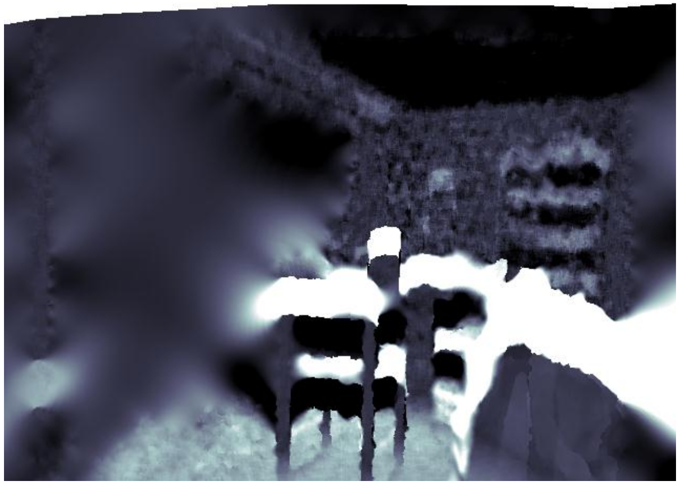
Color Image



Depth Image



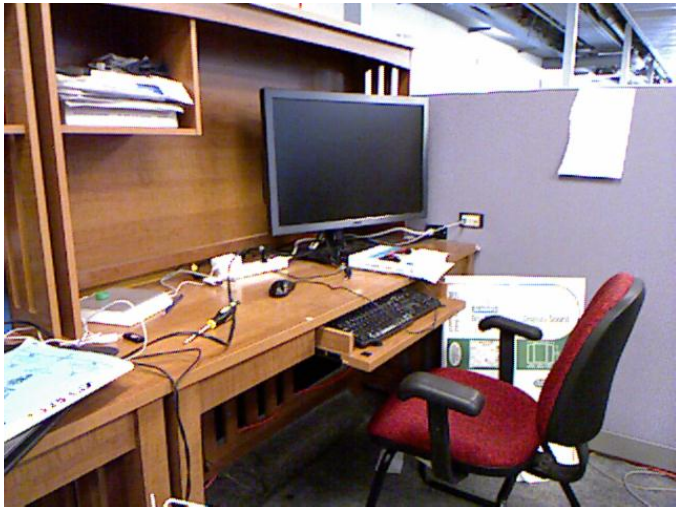
Not Normalized

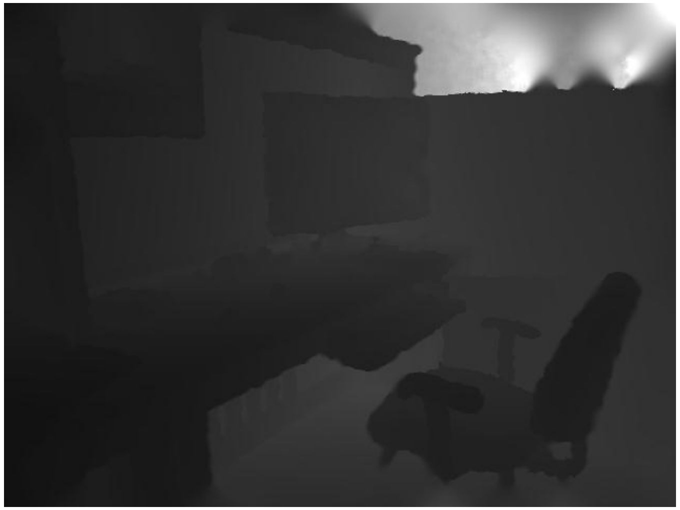


Normalized

Feature Map

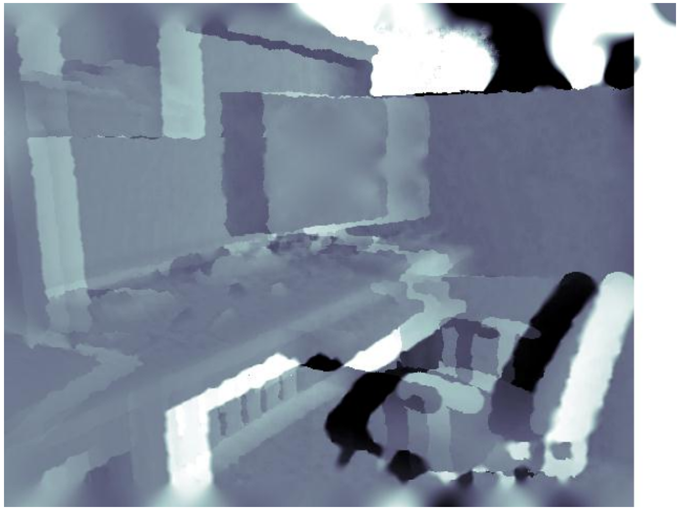
Color Image

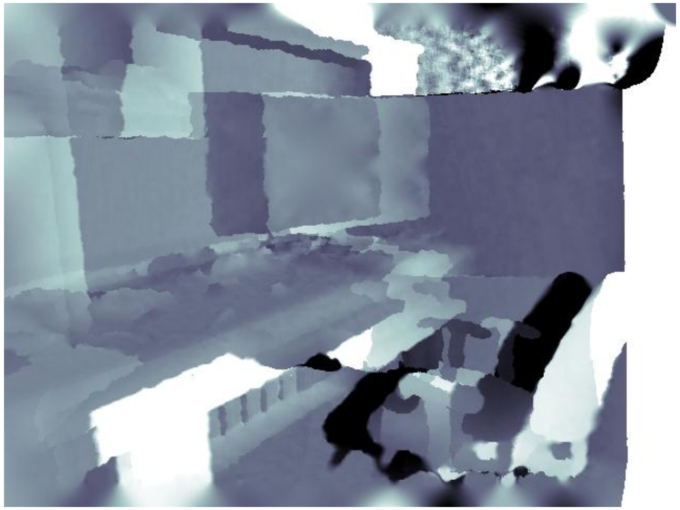




Depth Image

Not Normalized



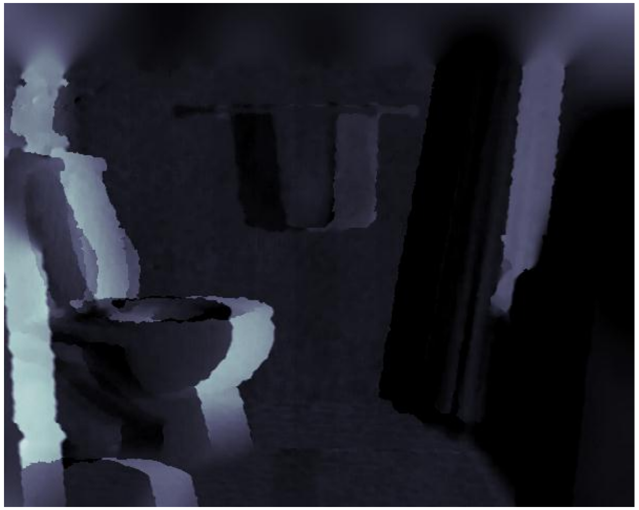


Normalized

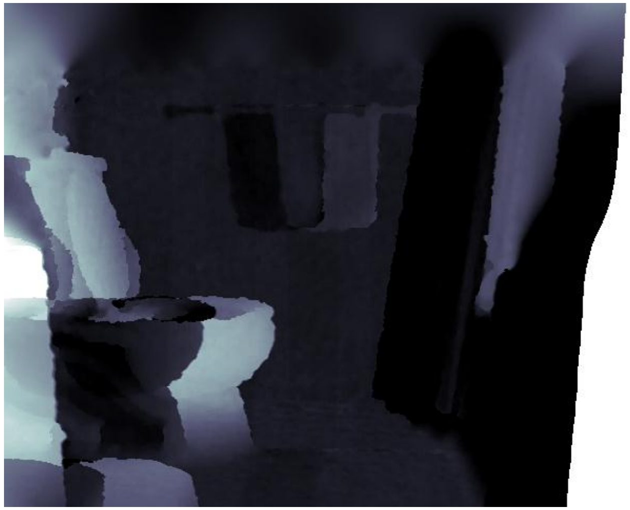
Feature Map

Color Image

Depth Image

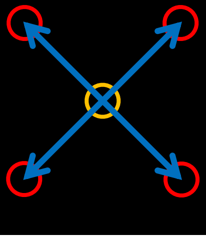


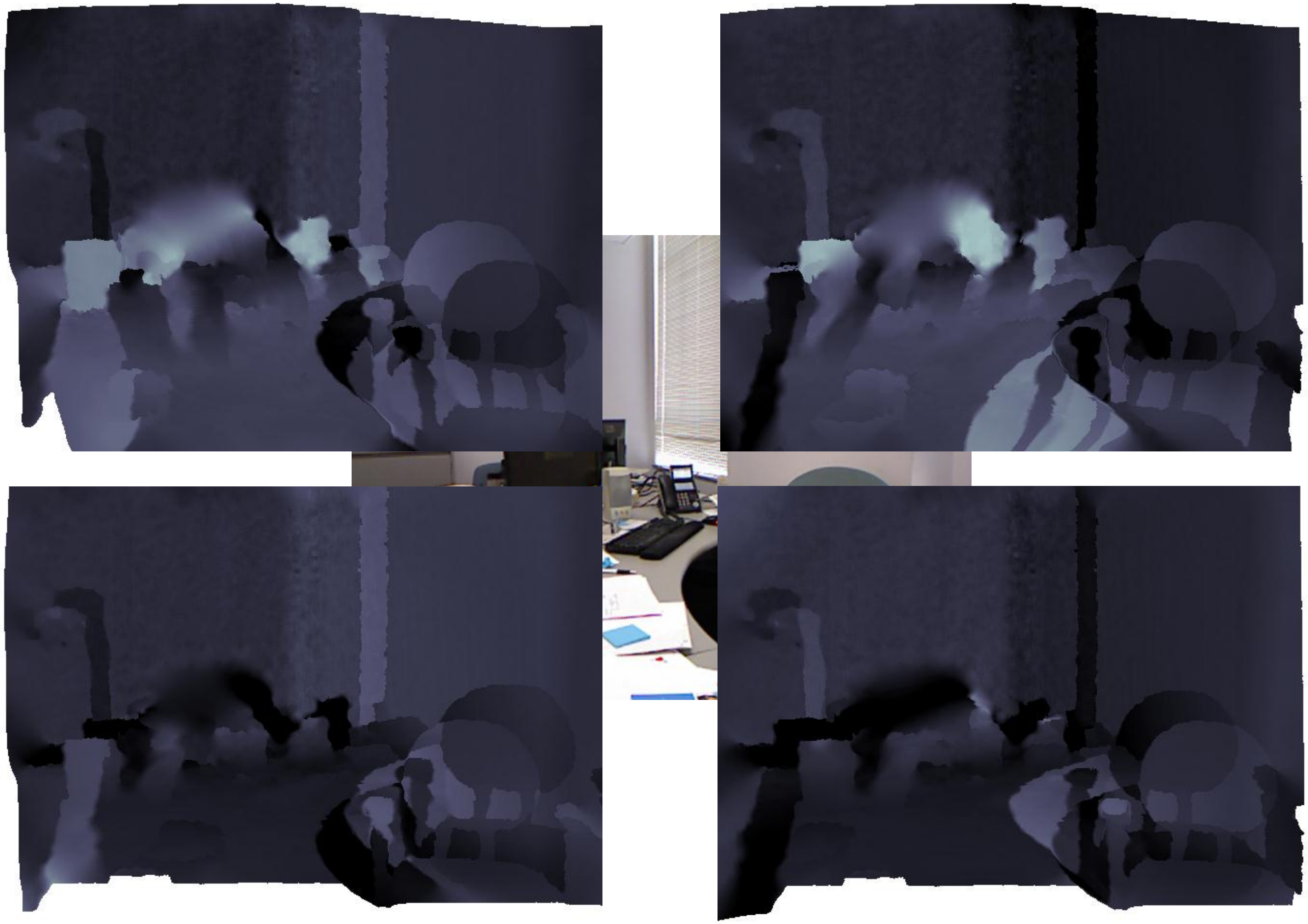
Not Normalized



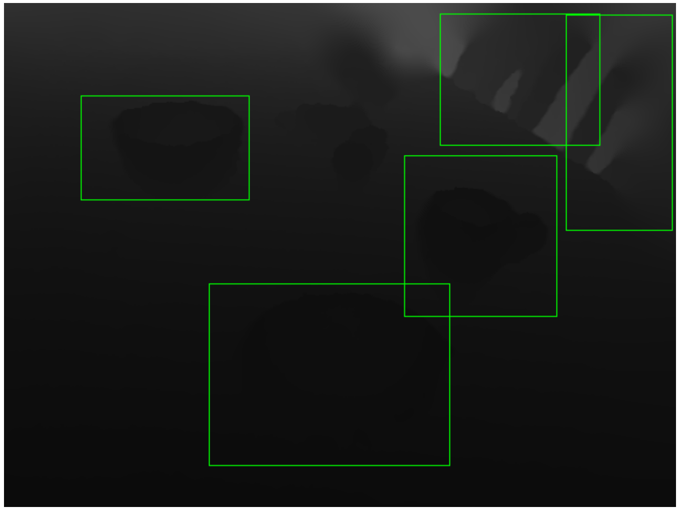
Normalized

Feature Map

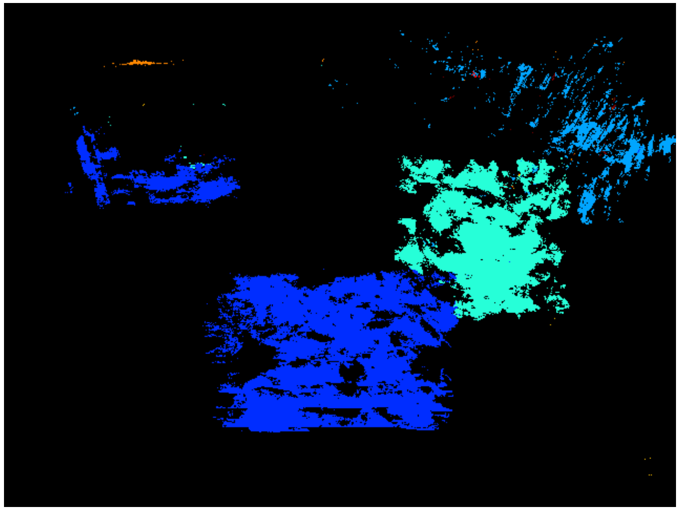




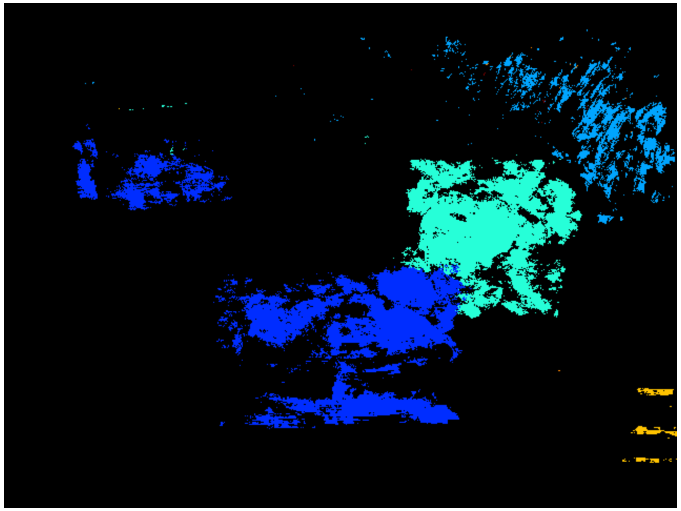
Varying # of Random Features



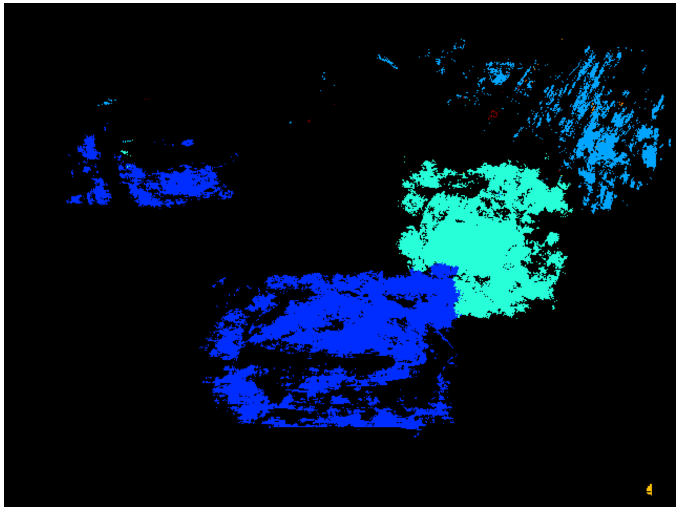
Ground Truth



50 Features

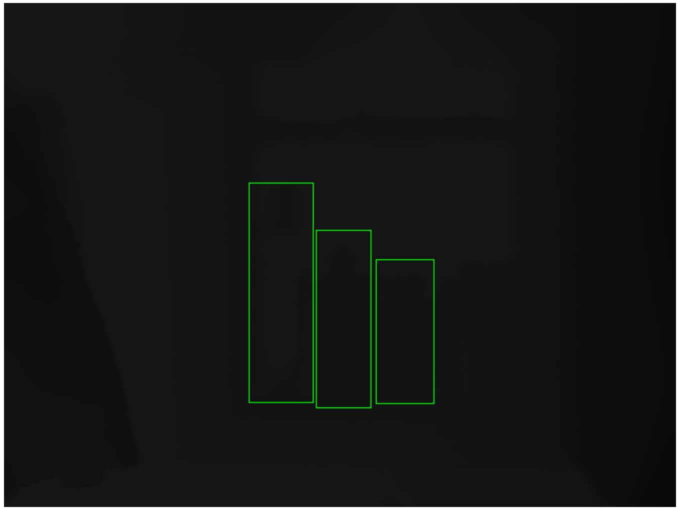


100 Features

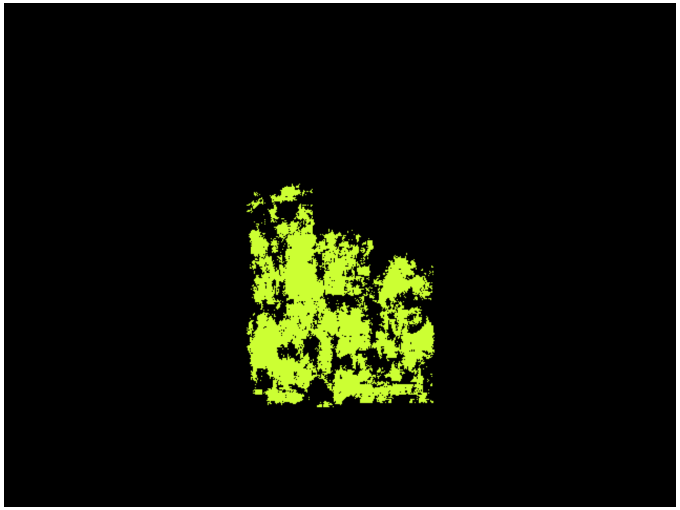


150 Features

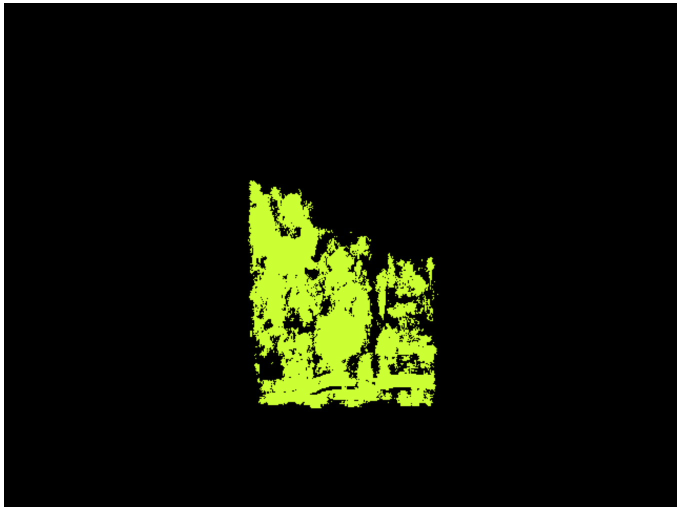
Varying # of Random Features



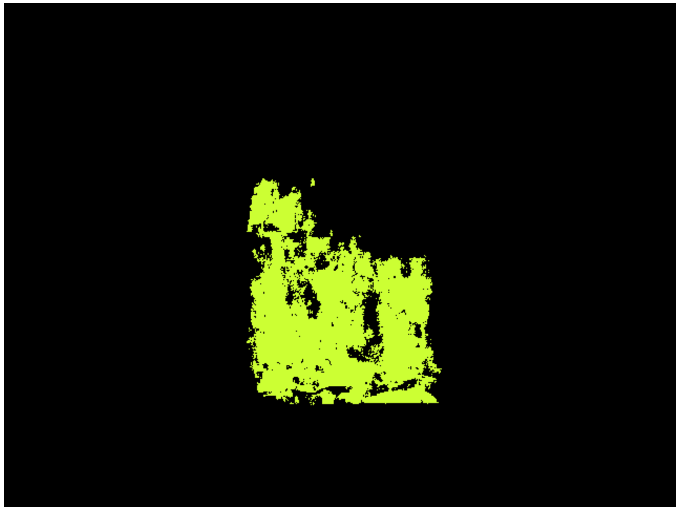
Ground Truth



50 Features

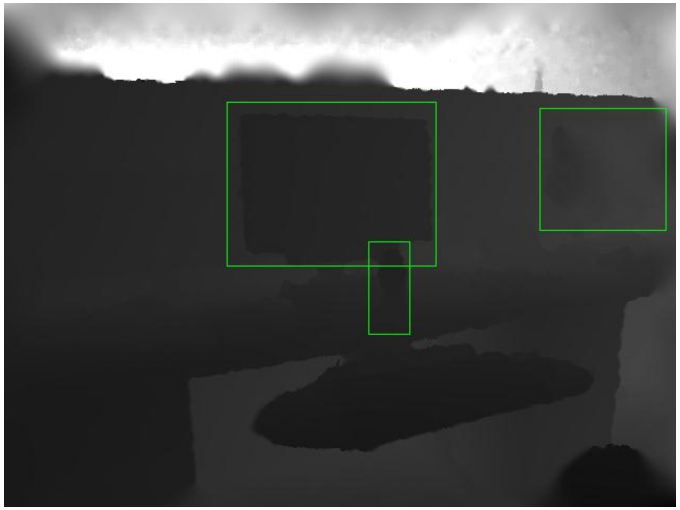


100 Features

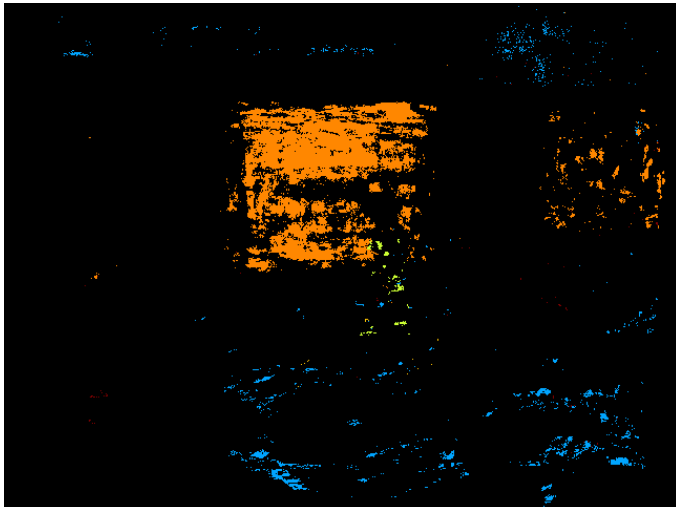


150 Features

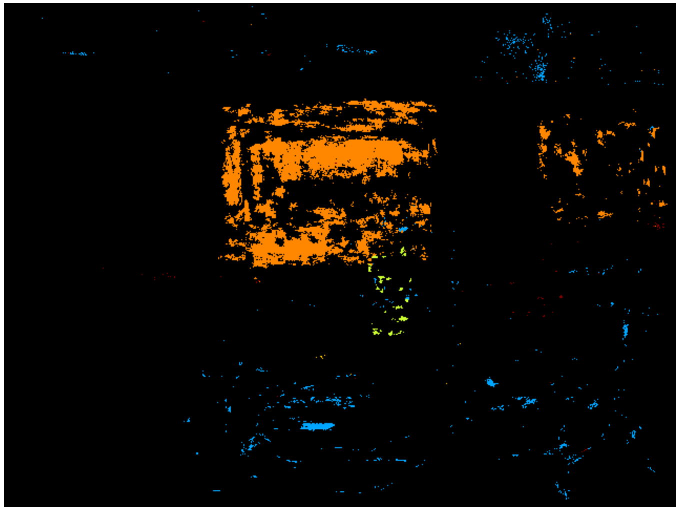
Varying # of Random Features



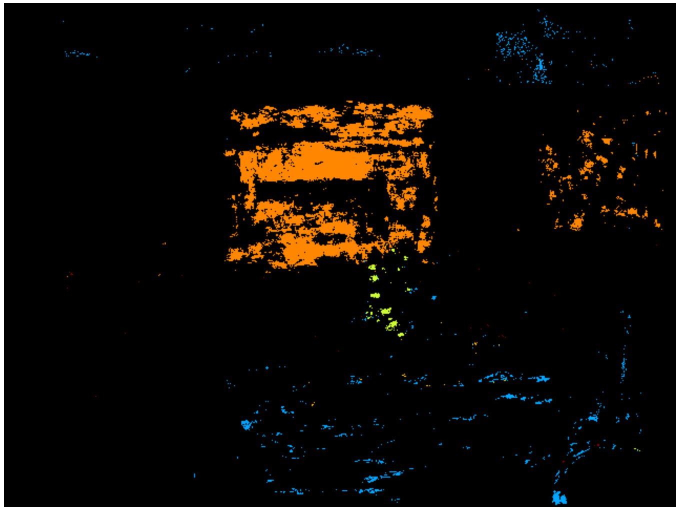
Ground Truth



50 Features

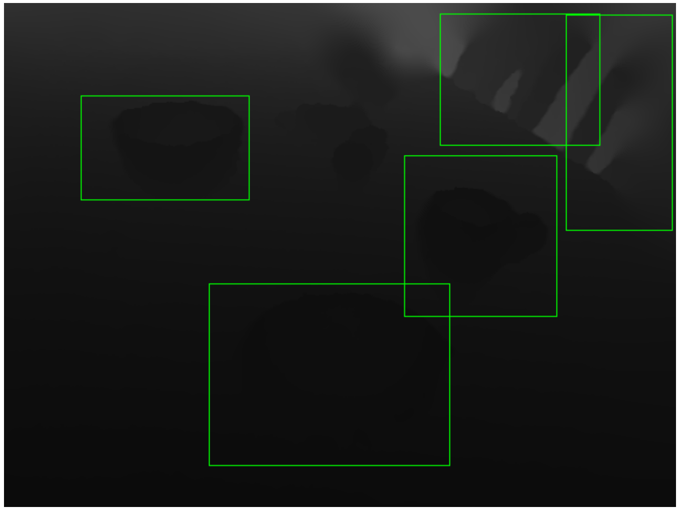


100 Features

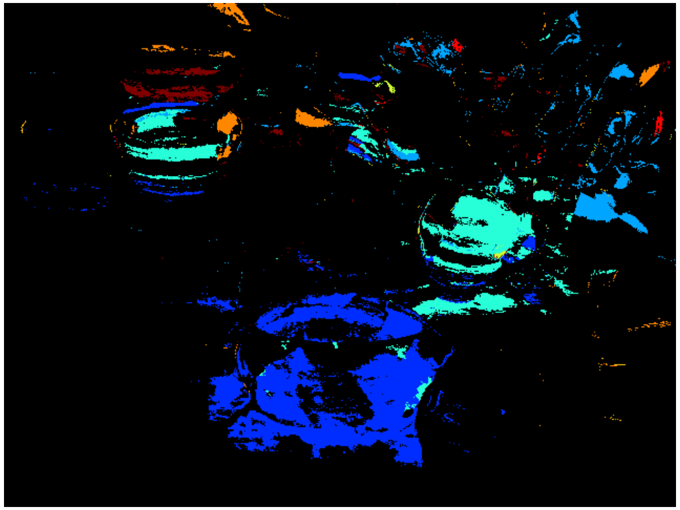


150 Features

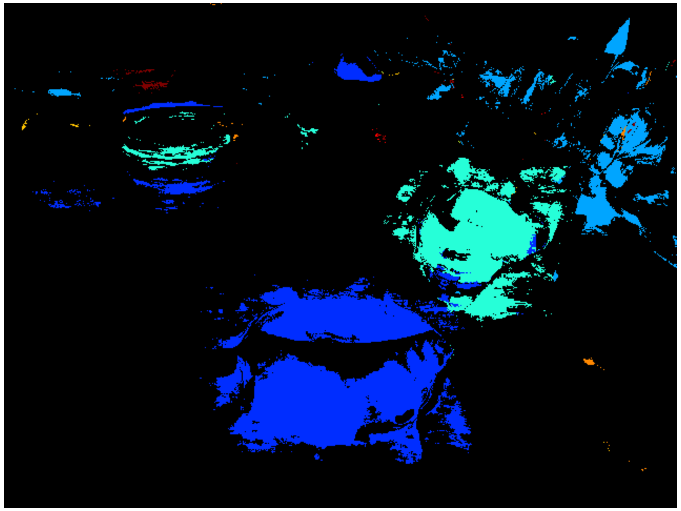
Varying # of Fixed Features



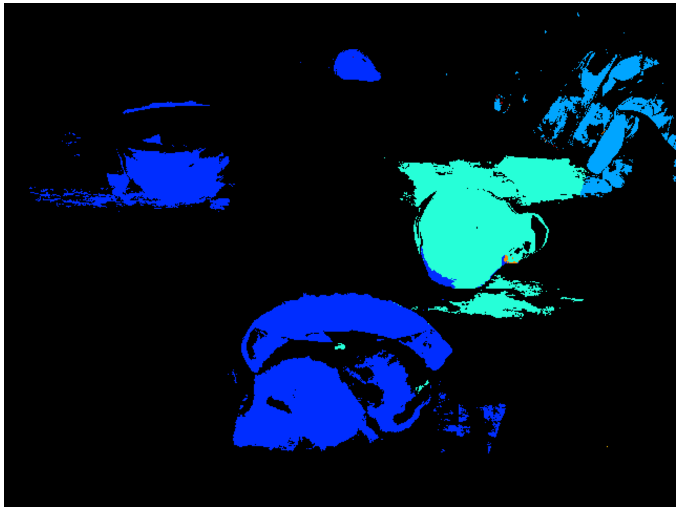
Ground Truth



4 Features

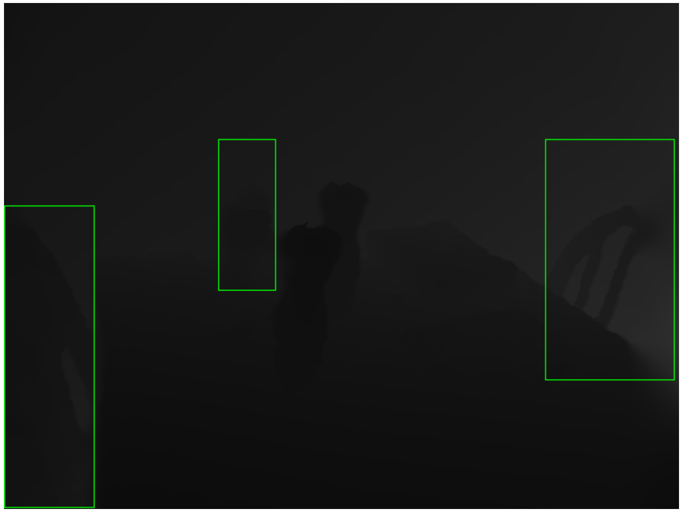


8 Features

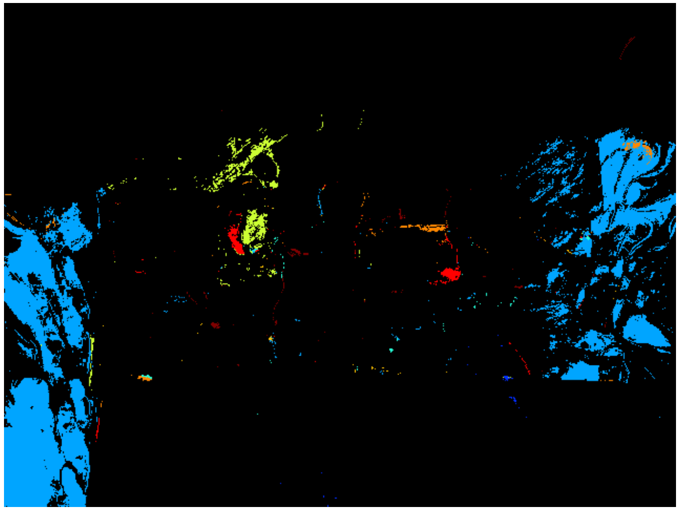


16 Features

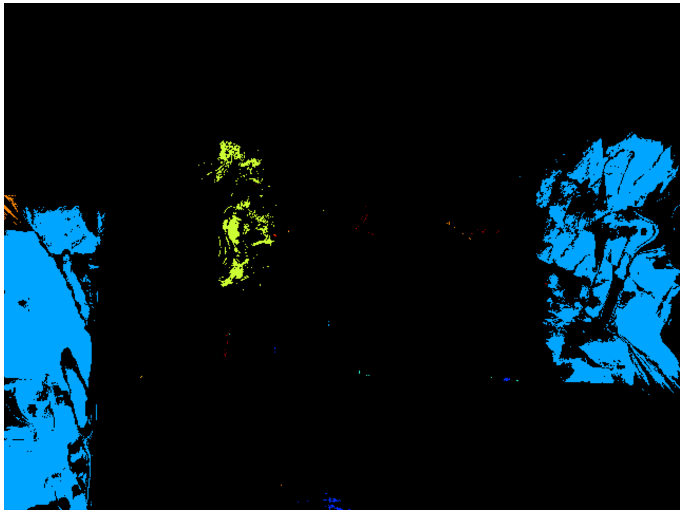
Varying # of Fixed Features



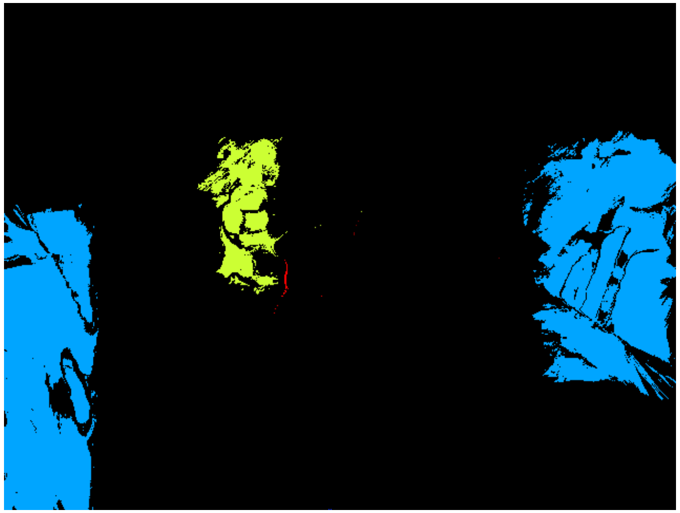
Ground Truth



4 Features

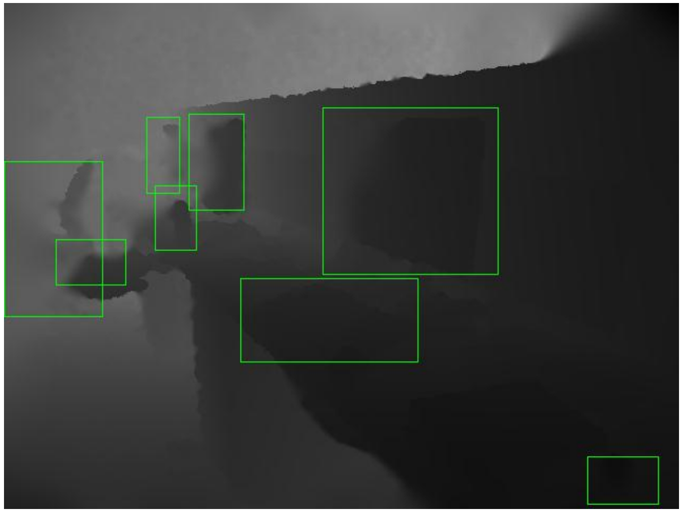


8 Features

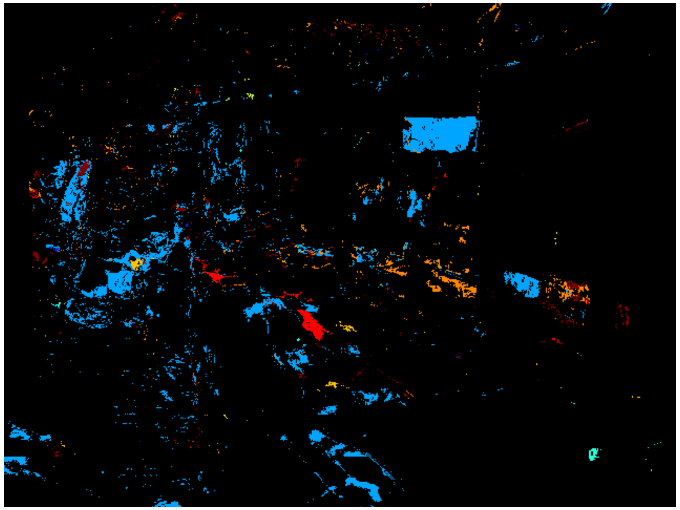


16 Features

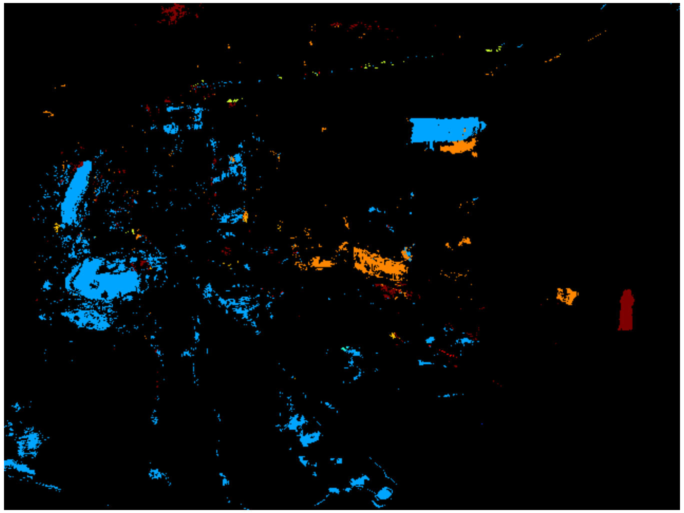
Varying # of Fixed Features



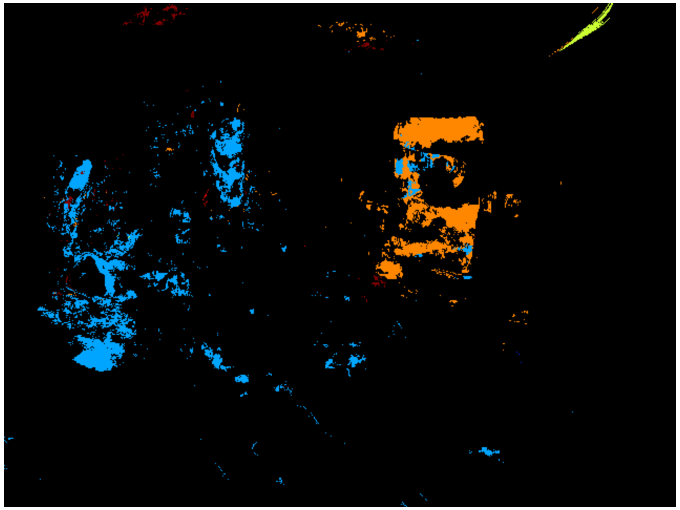
Ground Truth



4 Features



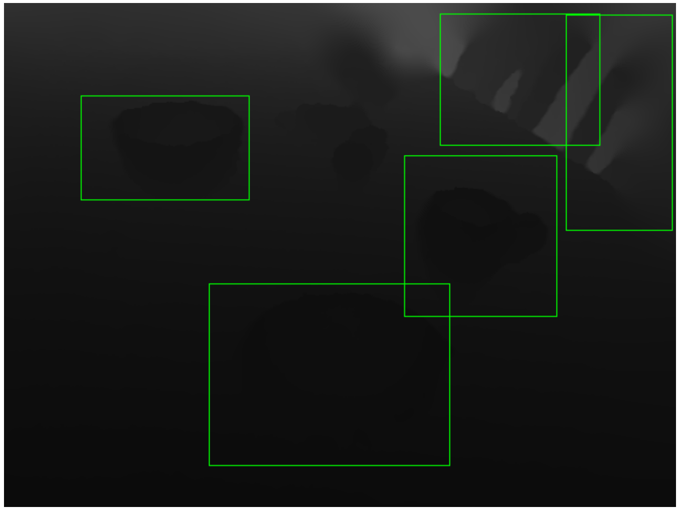
8 Features

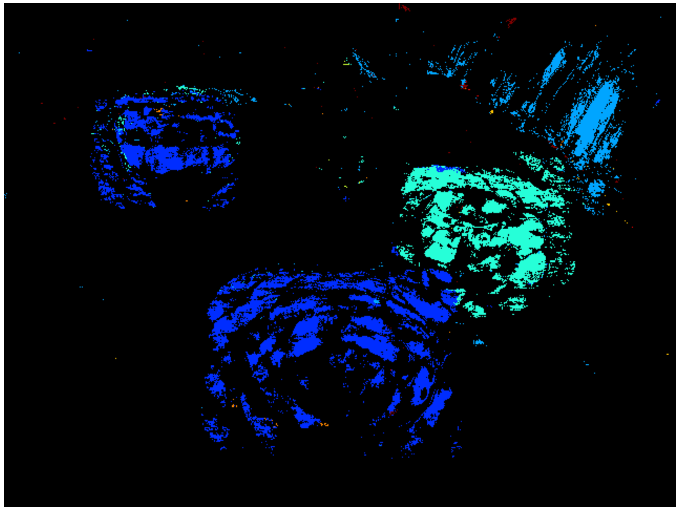


16 Features

Varying Offset

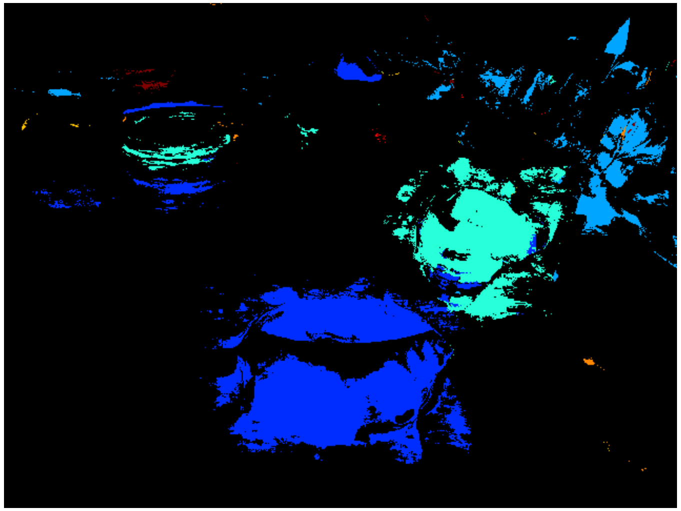
Ground Truth

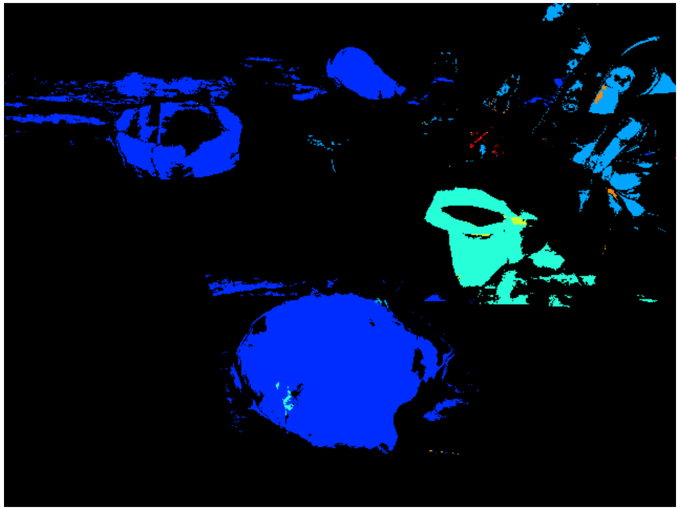




10 Pixel Meters

40 Pixel Meters

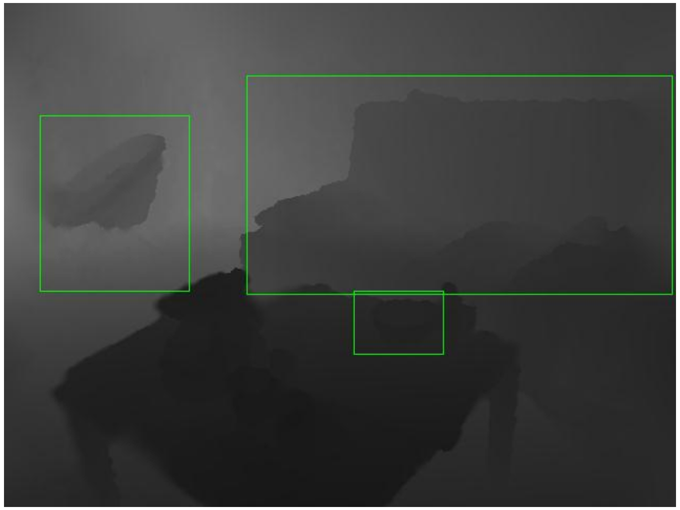


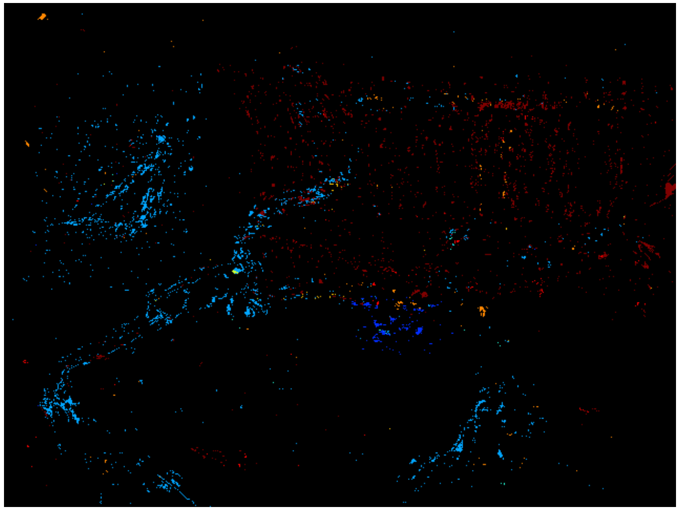


60 Pixel Meters

Varying Offset

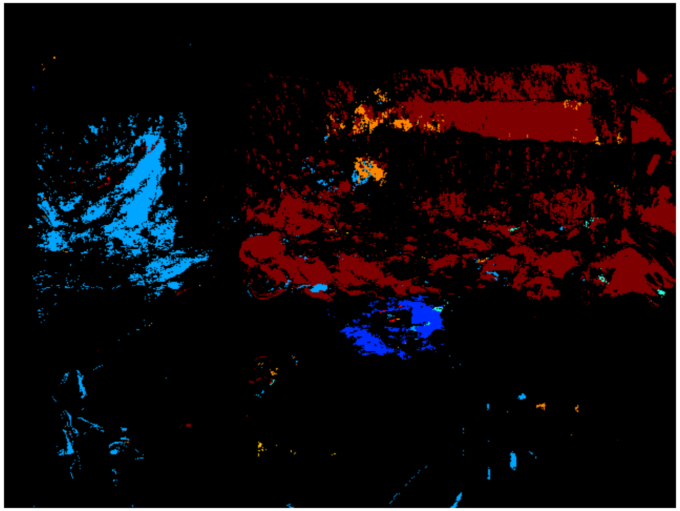
Ground Truth

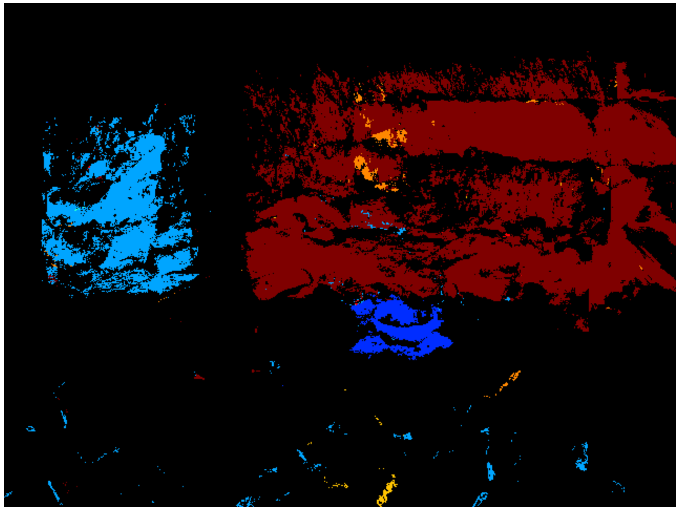




10 Pixel Meters

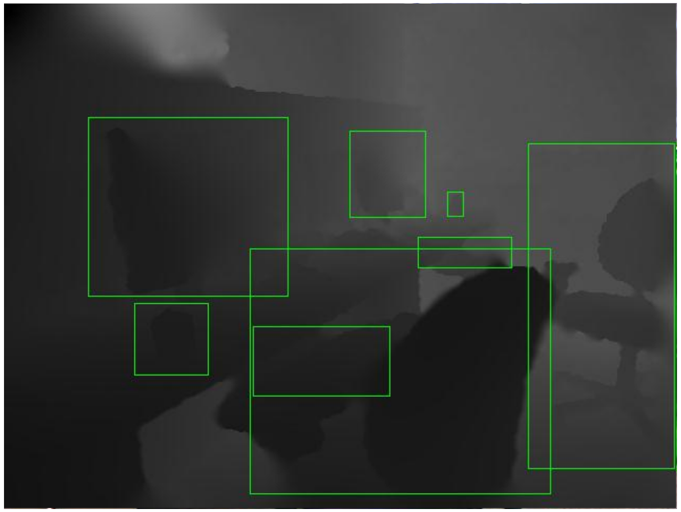
40 Pixel Meters



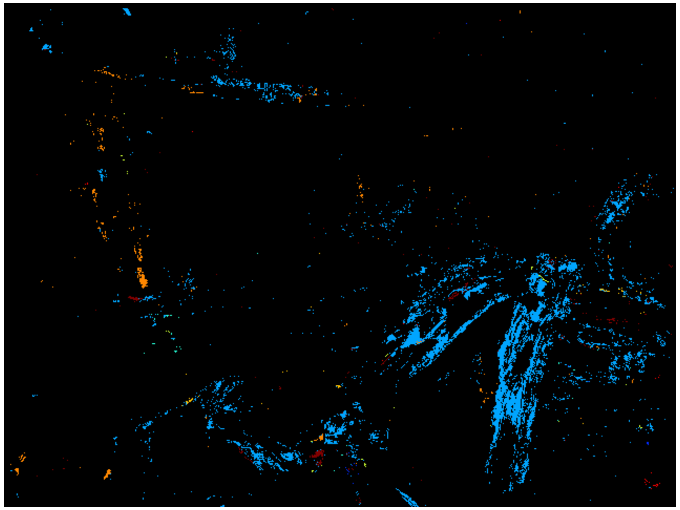


60 Pixel Meters

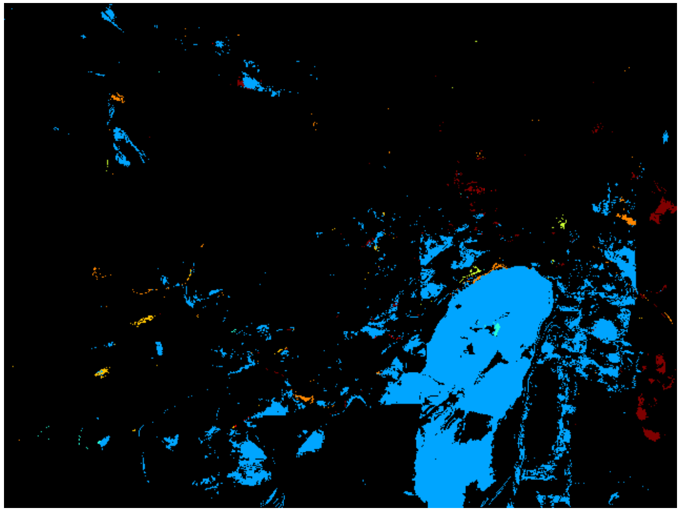
Varying Offset



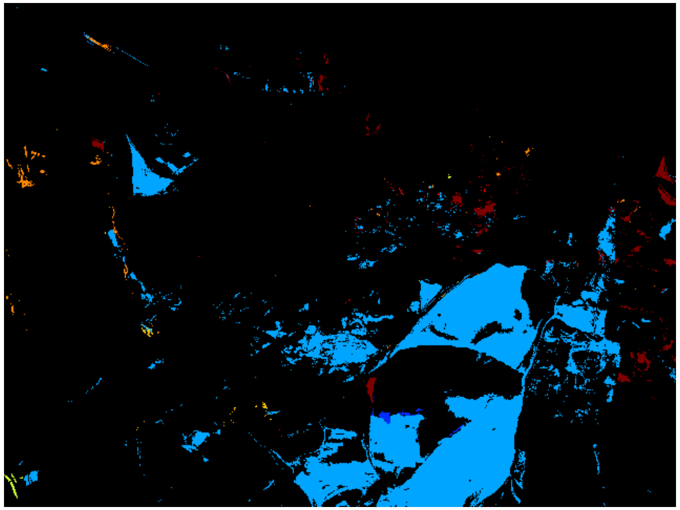
Ground Truth



10 Pixel Meters

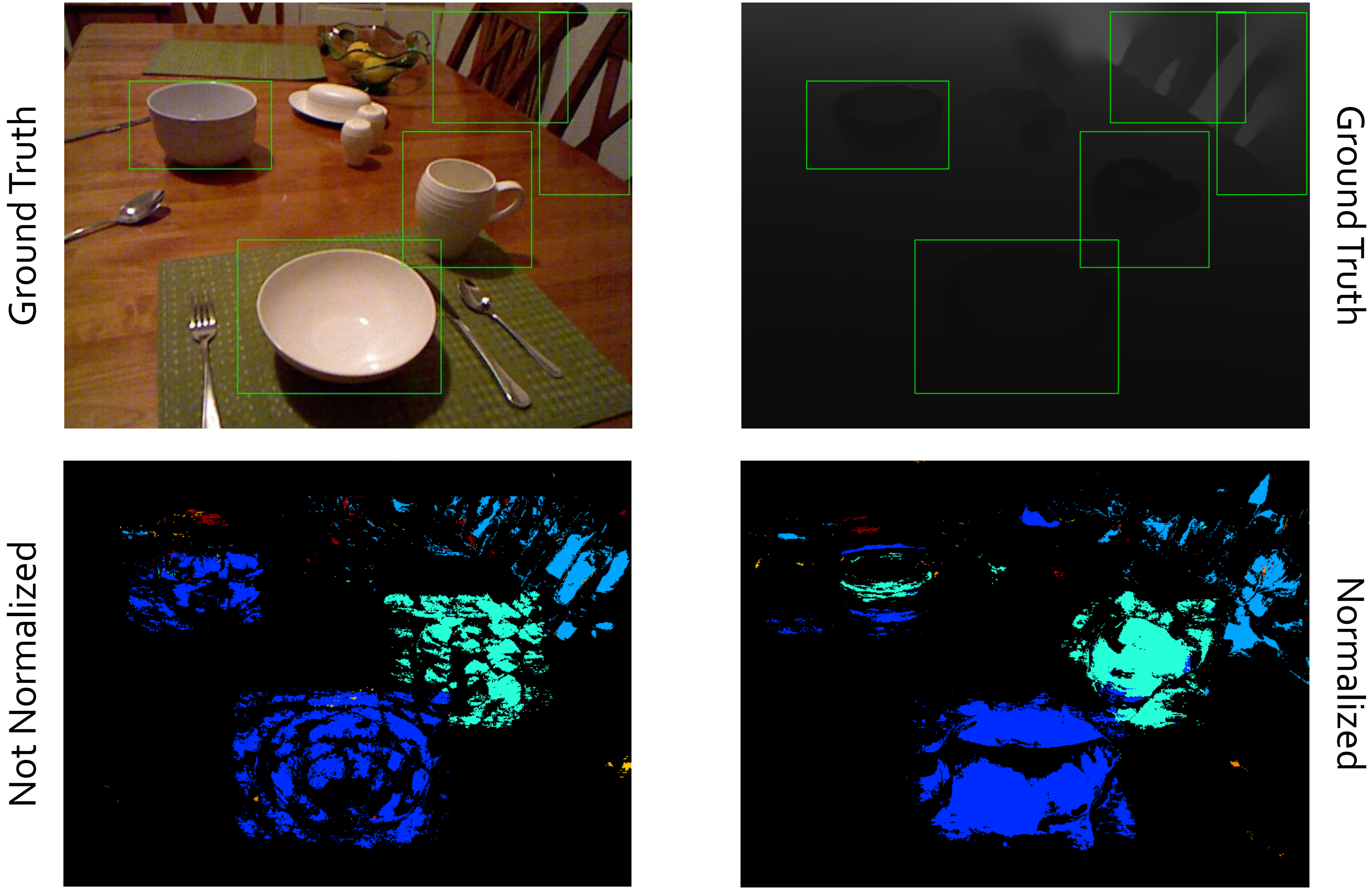


40 Pixel Meters



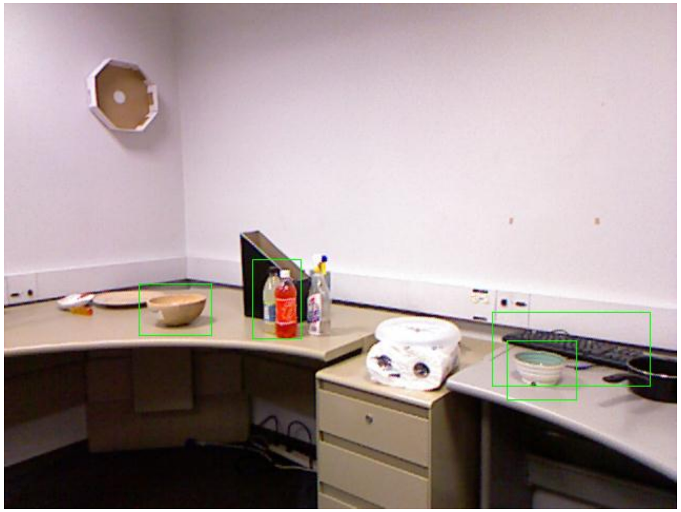
60 Pixel Meters

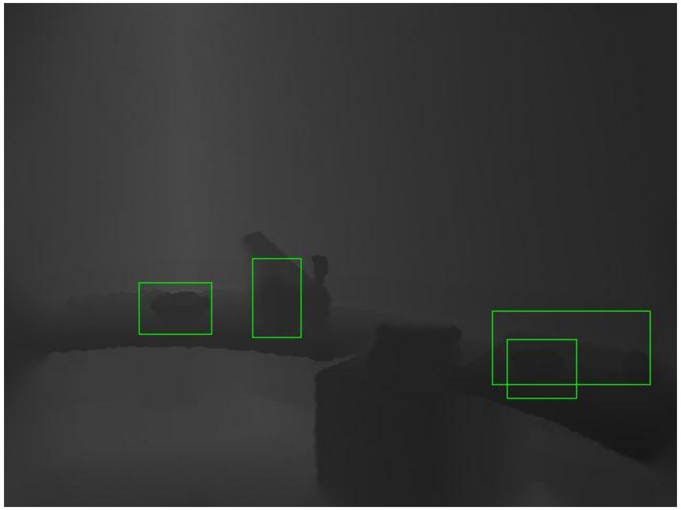
Varying Normalization



Varying Normalization

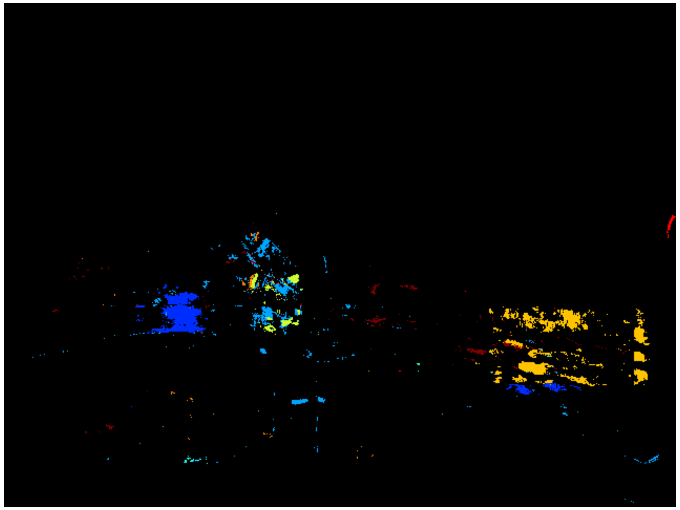
Ground Truth

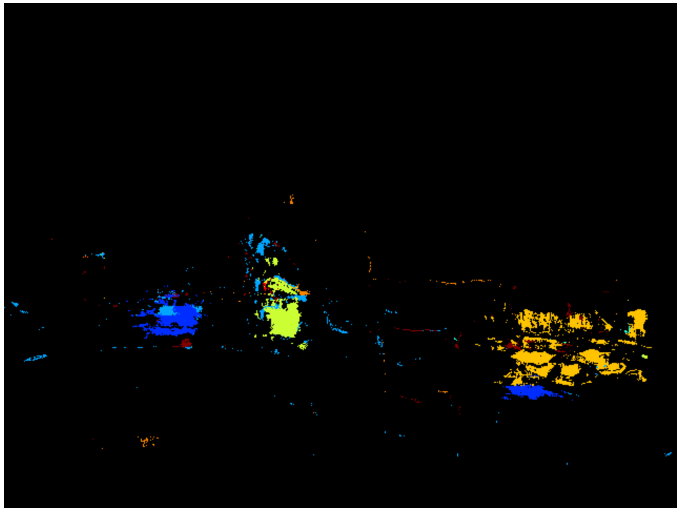




Ground Truth

Not Normalized

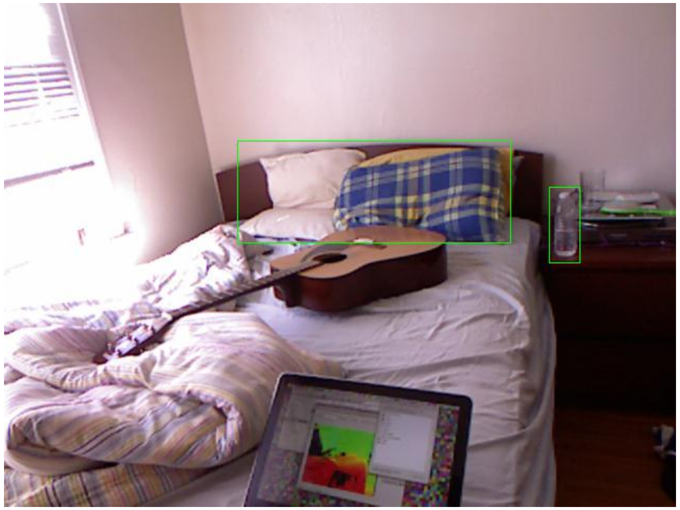


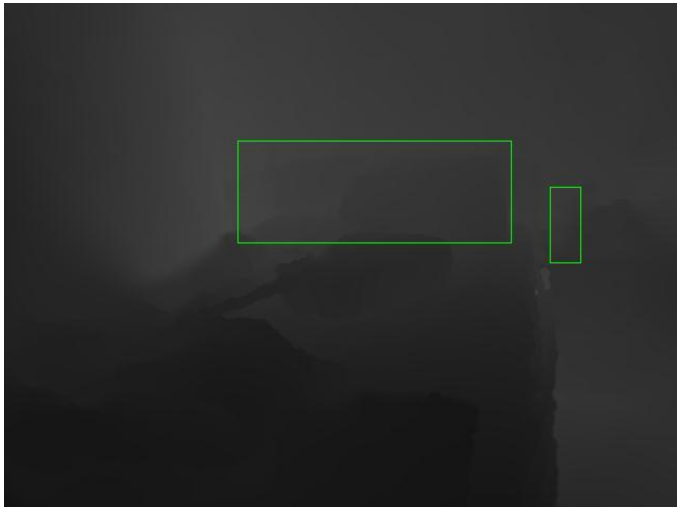


Normalized

Varying Normalization

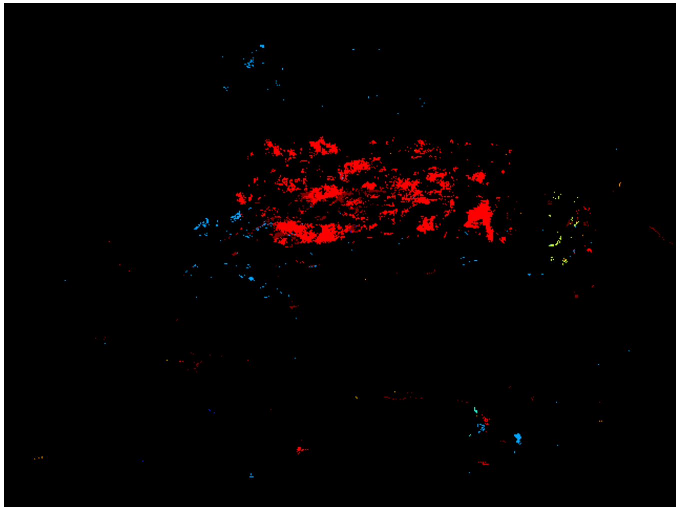
Ground Truth

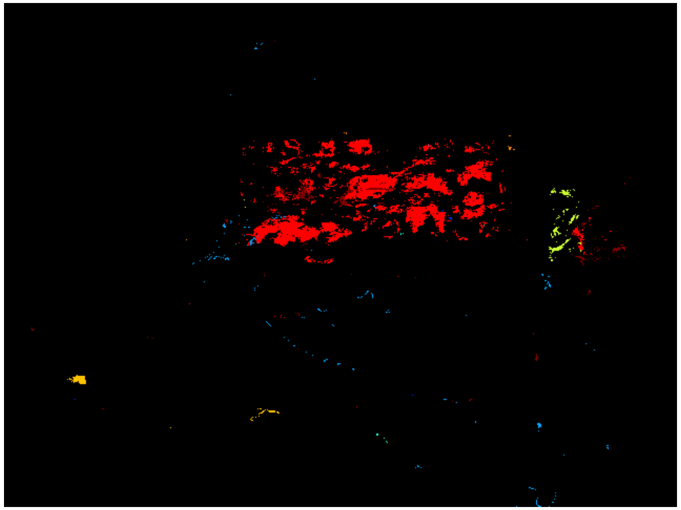




Ground Truth

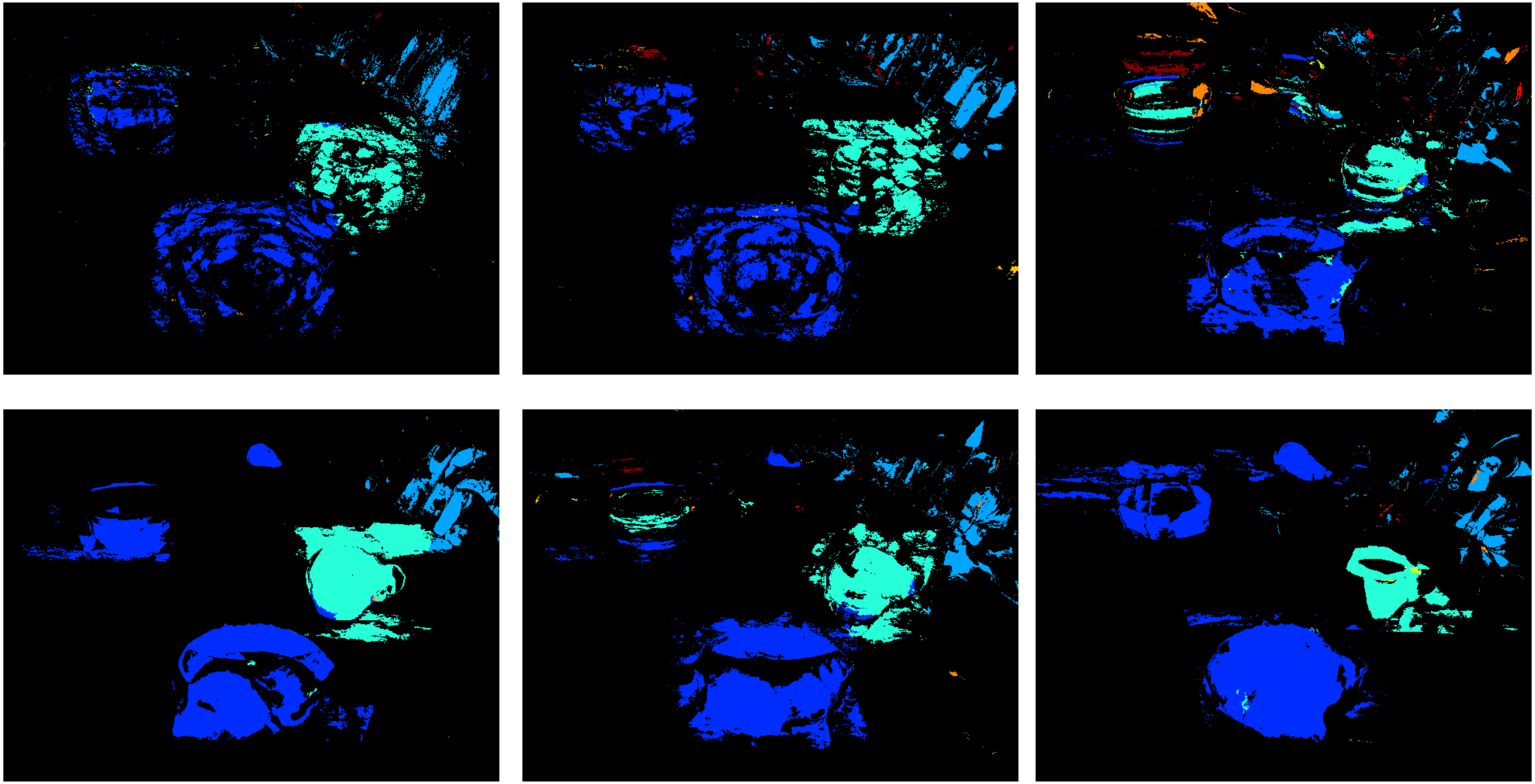
Not Normalized





Normalized

Conclusion



References

[1] Microsoft Kinect SDK & Toolkit (www.microsoft.com/en-us/kinectforwindows/develop)

[2] “Real-Time Human Pose Recognition in Parts from Single Depth Images” J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake (CVPR 2011)

[3] “Randomized Trees for Real-Time Keypoint Recognition” V.Lepetit, P. Lagger, P. Fua (CVPR 2005)

[4] “Boosting & Randomized Forests for Visual Recognition” J. Shotton (www.iis.ee.ic.ac.uk/~tkkim/iccv09_tutorial)

[5] “A Category-Level 3D Object Dataset: Putting the Kinect to Work” A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, T. Darrell (www.kinectdata.com)