3D Object Detection in RGB-D Images

Ye Wang

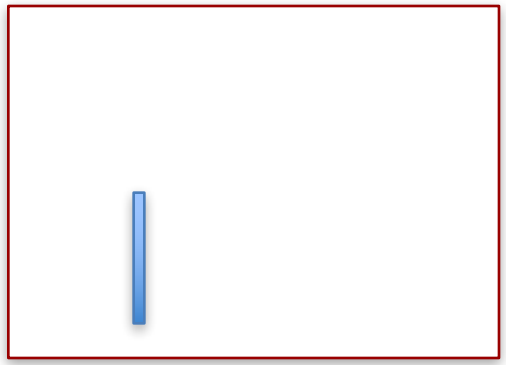
University of Southern California

Outline

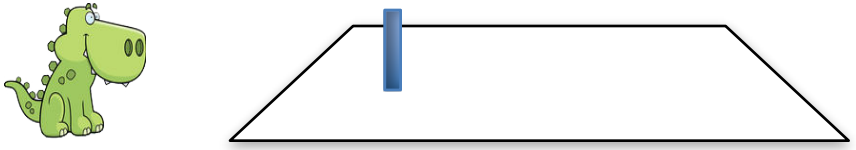
- Motivation
- RGB-D images and applications
- Typical 2D object detection methods
- 3D object detection in RGB-D images

Why 3D?

Viewpoint



2D Image



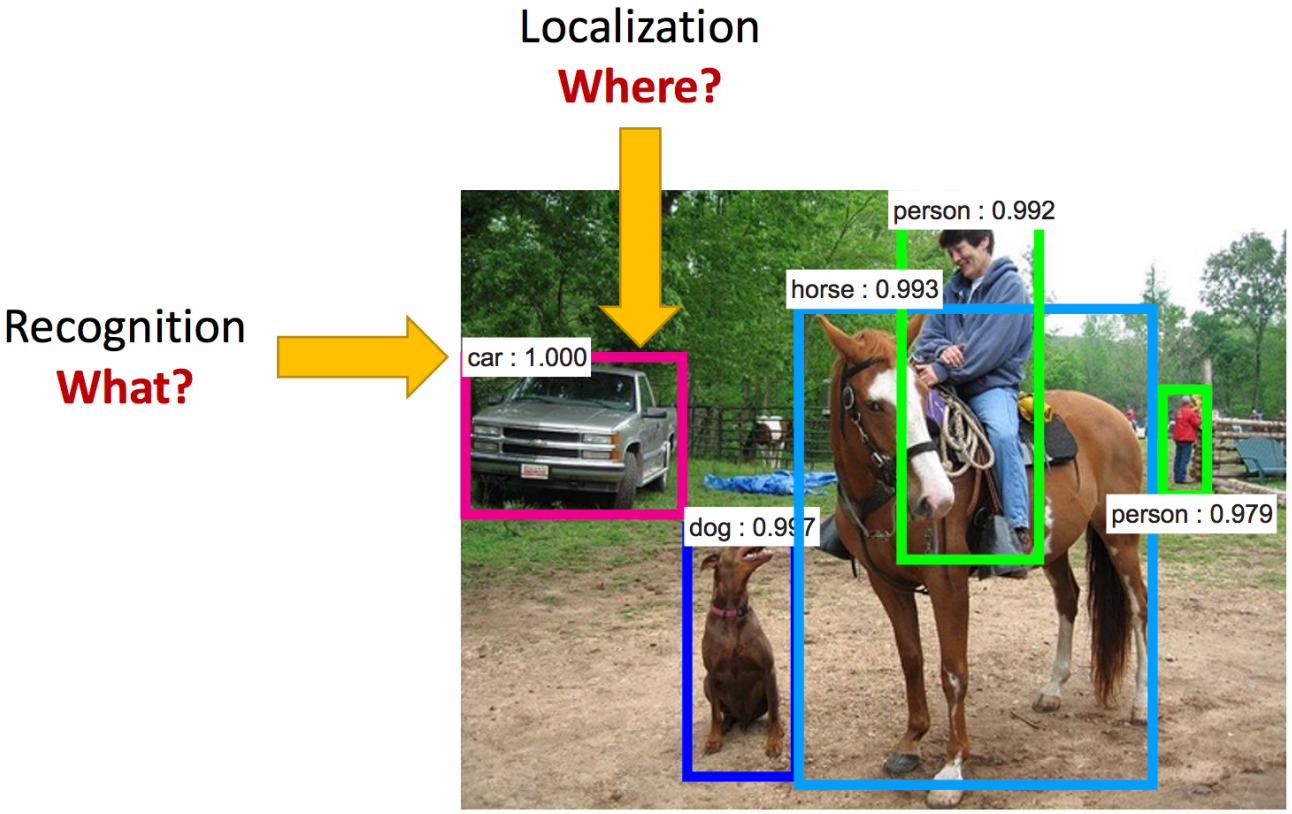
3D world

3D world

3D world

But directly teach computer to see 3D world is extremely hard!

How computers see?



RGB-D Sensors

- Microsoft Kinect
- Intel RealSense
- Google Project Tango
- Apple Primesense
- Asus Xtion
- LEAP Motion
- Structure.io
- Stereo Cameras

Outline

- Motivation
- RGB-D images and applications
- Typical 2D object detection methods
- 3D object detection in RGB-D images

Example of RGBD Images

1. Before the Microsoft Kinect (2010), depth datasets small, captured in the lab
2. Now get RGBD data from dynamic and static scenes from the real world, with a range of labelling and capture conditions

Color image

Depth image

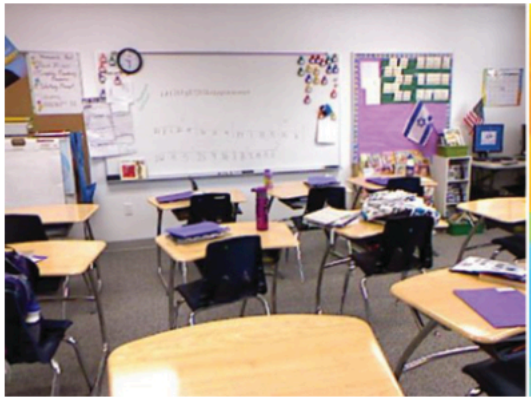
Existing RGBD Datasets

1. NYUv2
2. SUN RGB-D
3. NYU Dataset v1
4. SUN3D
5. B3DO (Berkeley 3-D Object Dataset)

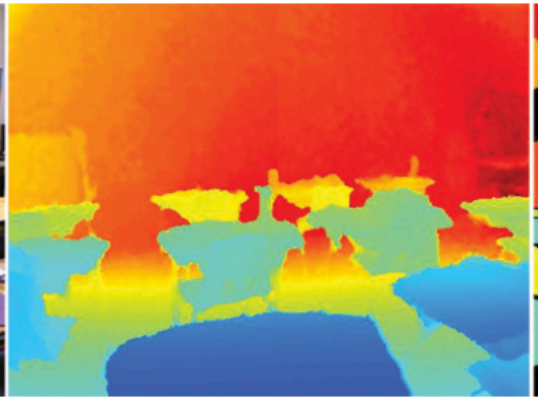
Michael Firman, “RGBD Datasets: Past, Present and Future ”, arXiv, 13 Apr 2016

NYUv2

1. **Introduced**: ECCV 2012
2. **Device**: Kinect v1
3. **Description**: ~408,000 RGBD images from 464 indoor scenes, larger diversity than NYUv1
4. **Labelling**: Dense labelling of objects at a class and instance level for 1449 frames



RGB image

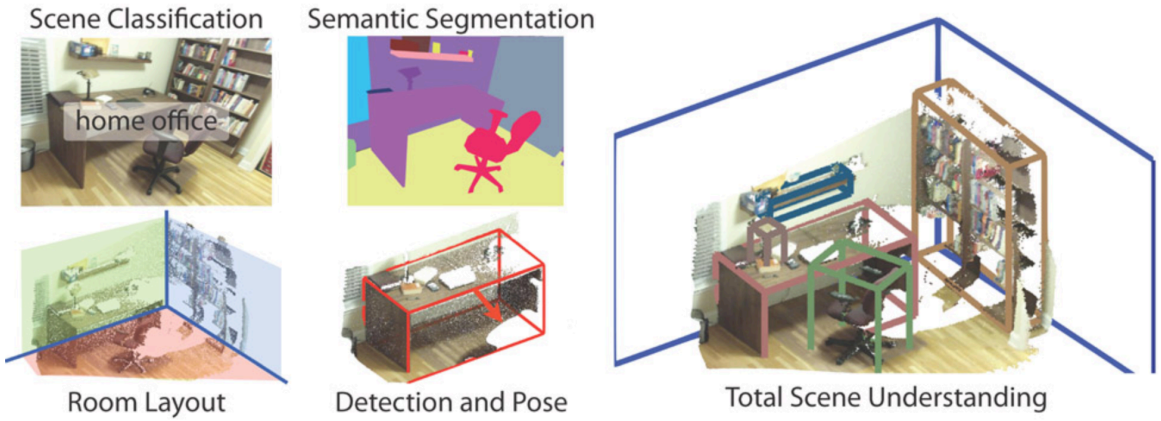


Preprocessed Depth Image

Labels for the image

SUN RGB-D

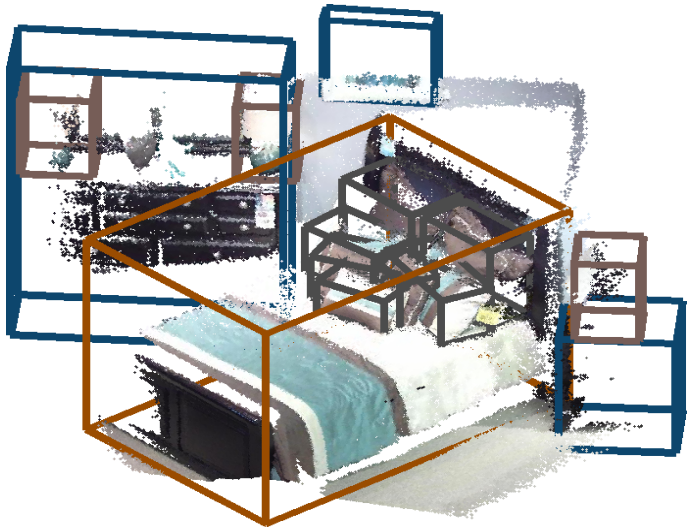
1. **Introduced**: CVPR 2015
2. **Device**: Kinect v1, Kinect v2, Intel RealSense and Asus Xtion Live Pro
3. **Description**: New images, plus images taken from NYUv2, B3DO and SUN3D. All of indoor scenes.
4. **Labelling**: 10,335 images with polygon annotation, and 3D bounding boxes around objects

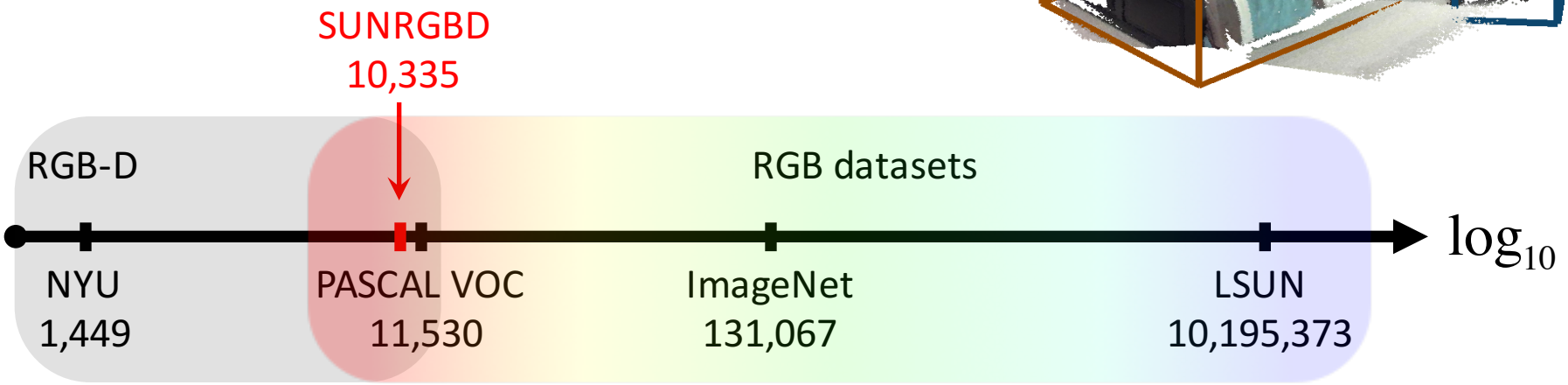


Dataset Size

RGB-D data is hard to get

PASCAL-scale size till CVPR 2015





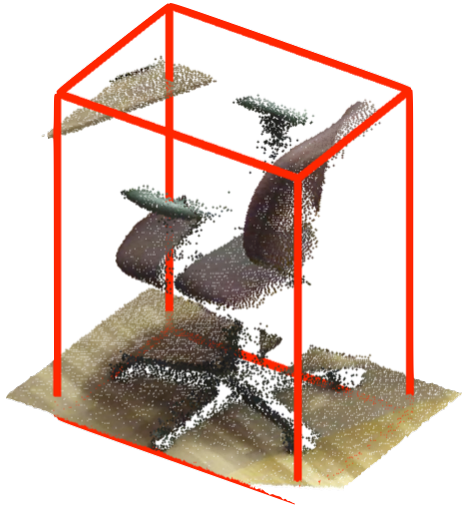
Object Annotation

1. **NYUv2**

- ❑ 2D segmentation

2. **Sun RGB-D**

- ❑ 2D segmentation + 3D object



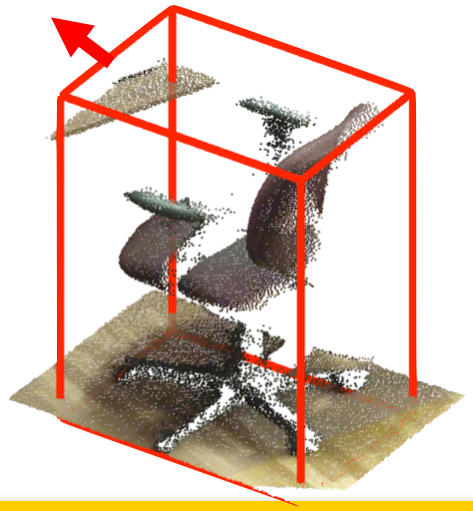
Object Pose

1. **NYUv2**

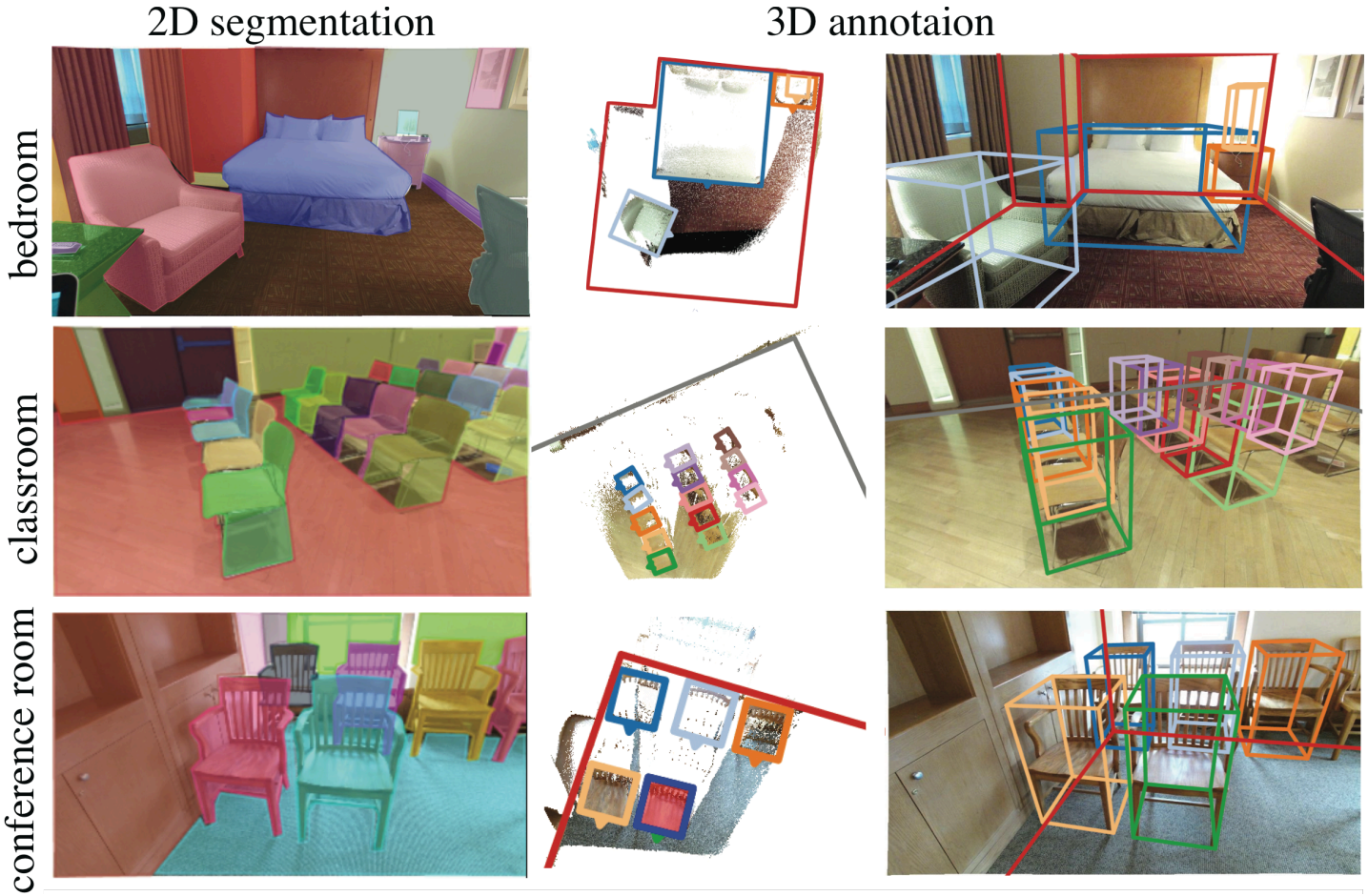
- ❑ None

2. **Sun RGB-D**

- ❑ Yes

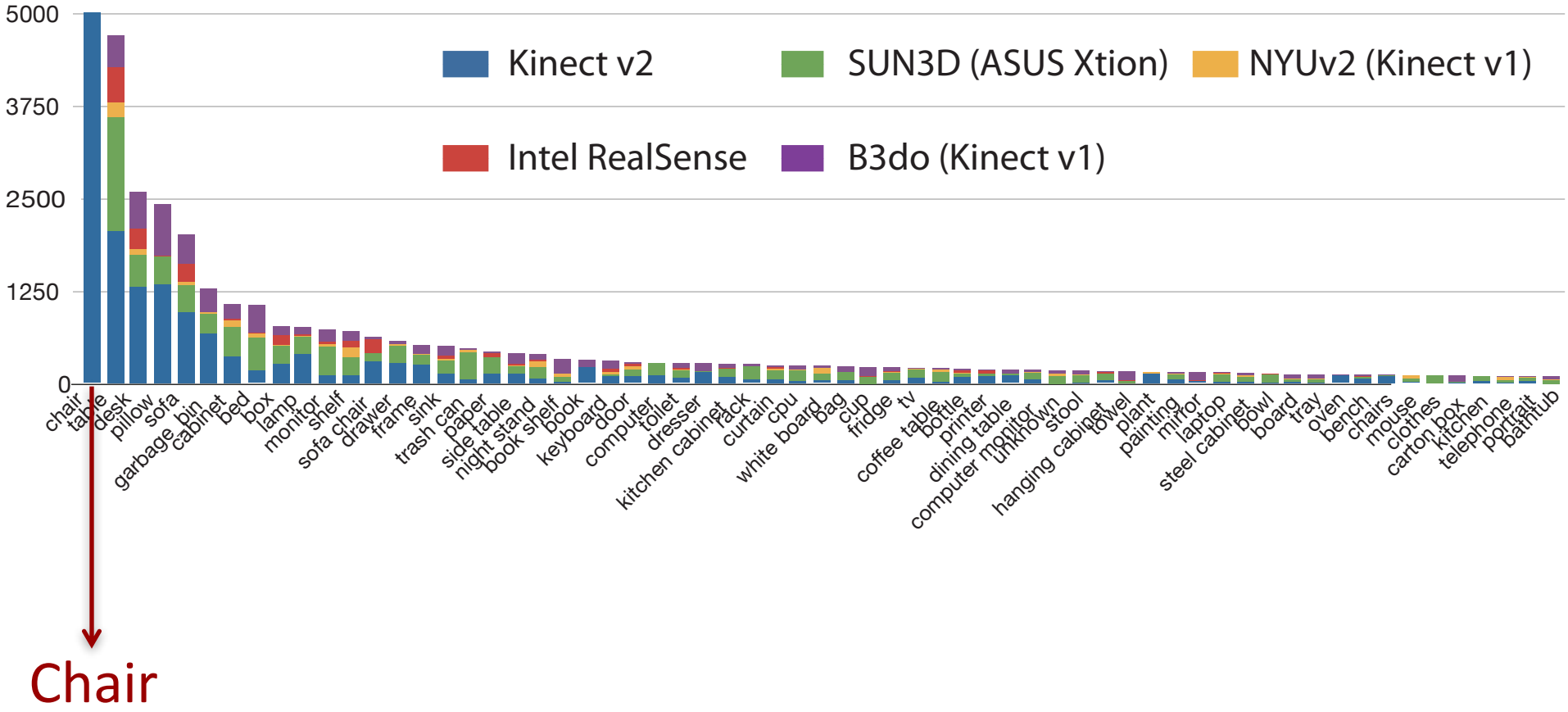


More examples about SUN RGB-D



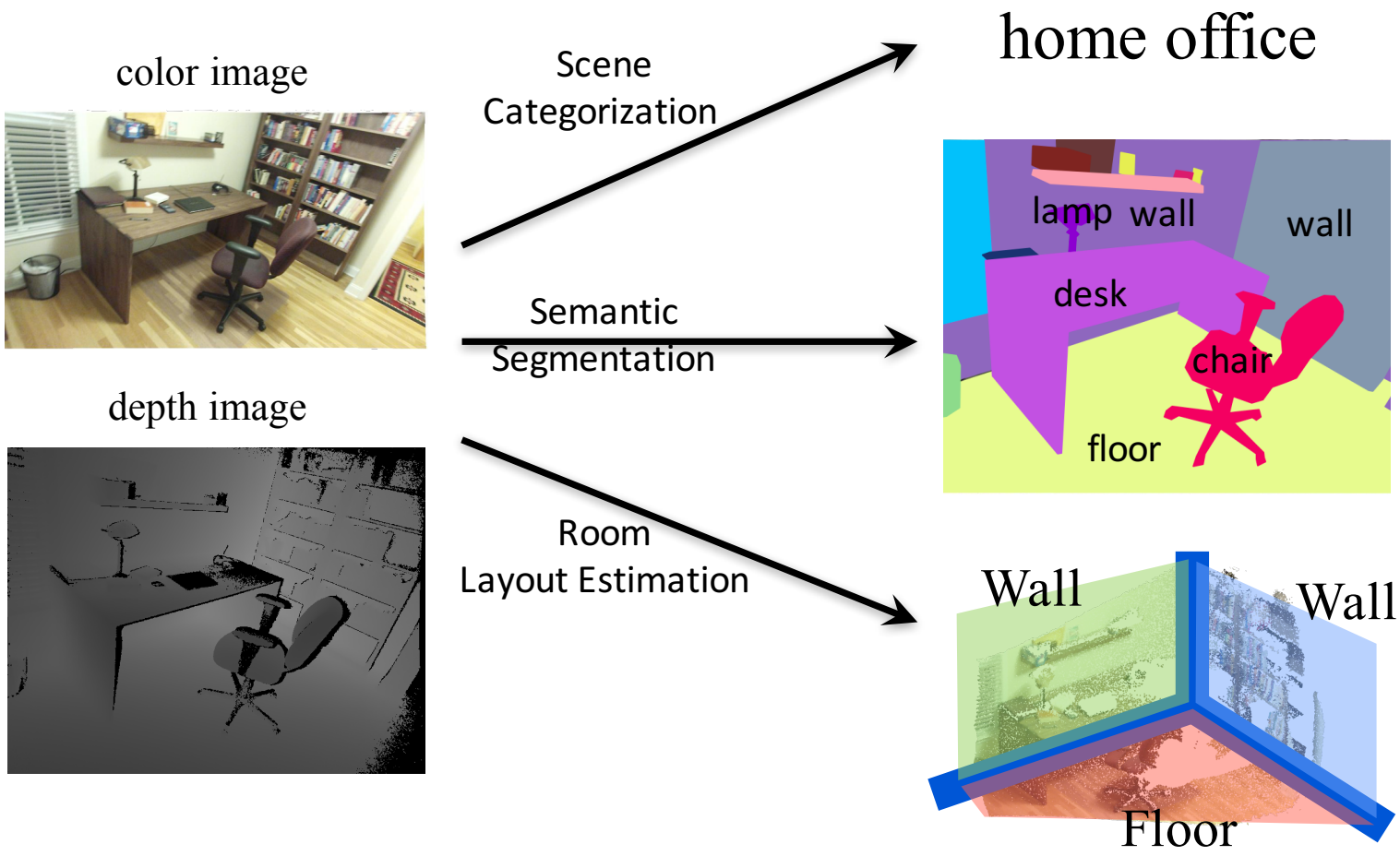
Song, Shuran, Samuel P. Lichtenberg, and Jianxiong Xiao. "Sun rgb-d: A rgb-d scene understanding benchmark suite." CVPR 2015.

Object categories distribution

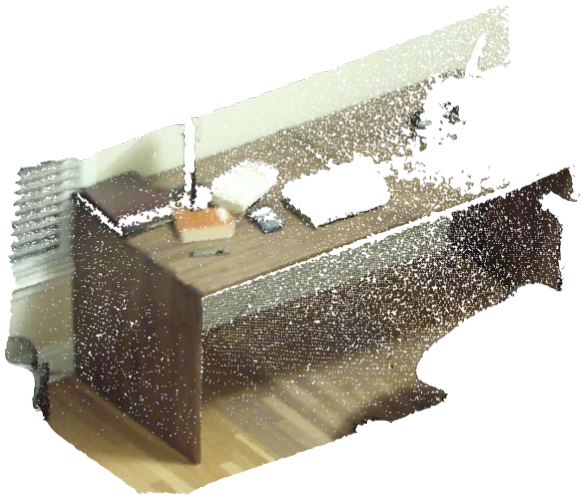


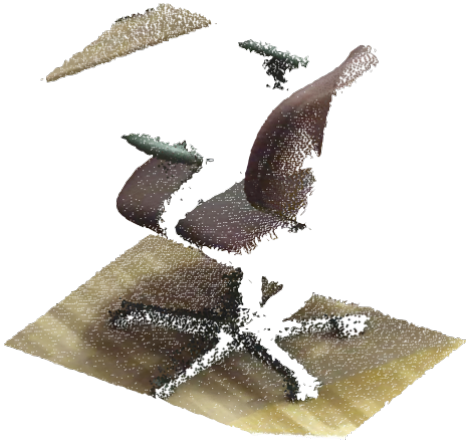
Chair

Applications about SUN RGB-D



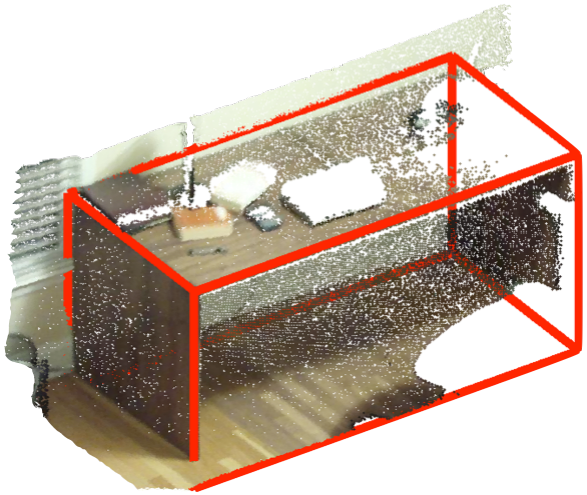
Applications about SUN RGB-D



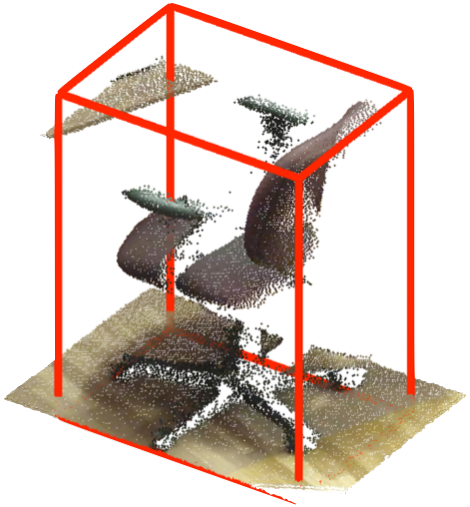


3D Object Detection

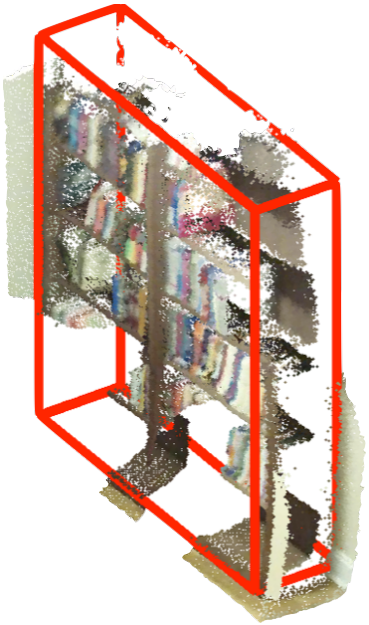
Applications about SUN RGB-D



table



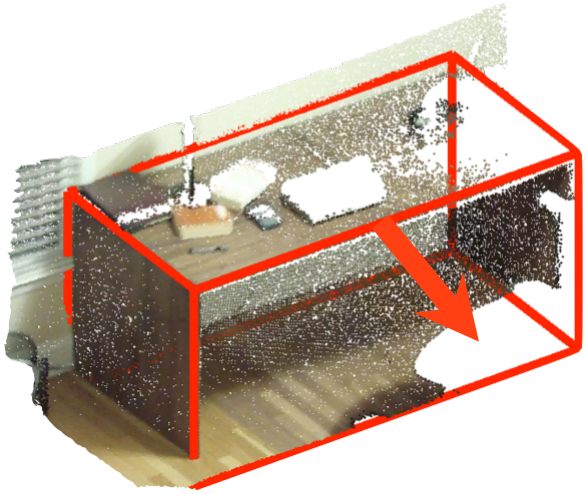
chair



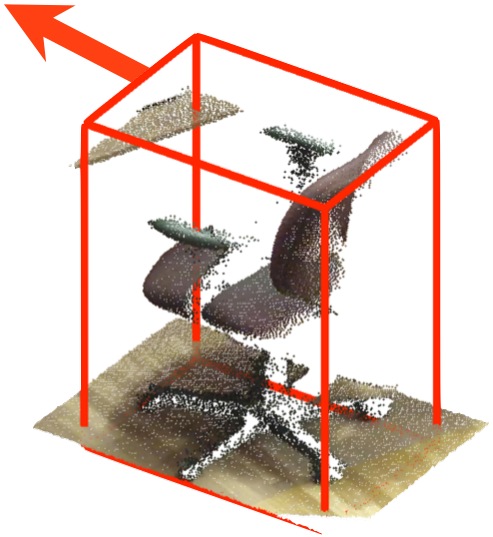
bookshelf

3D Object Detection

Applications about SUN RGB-D



table

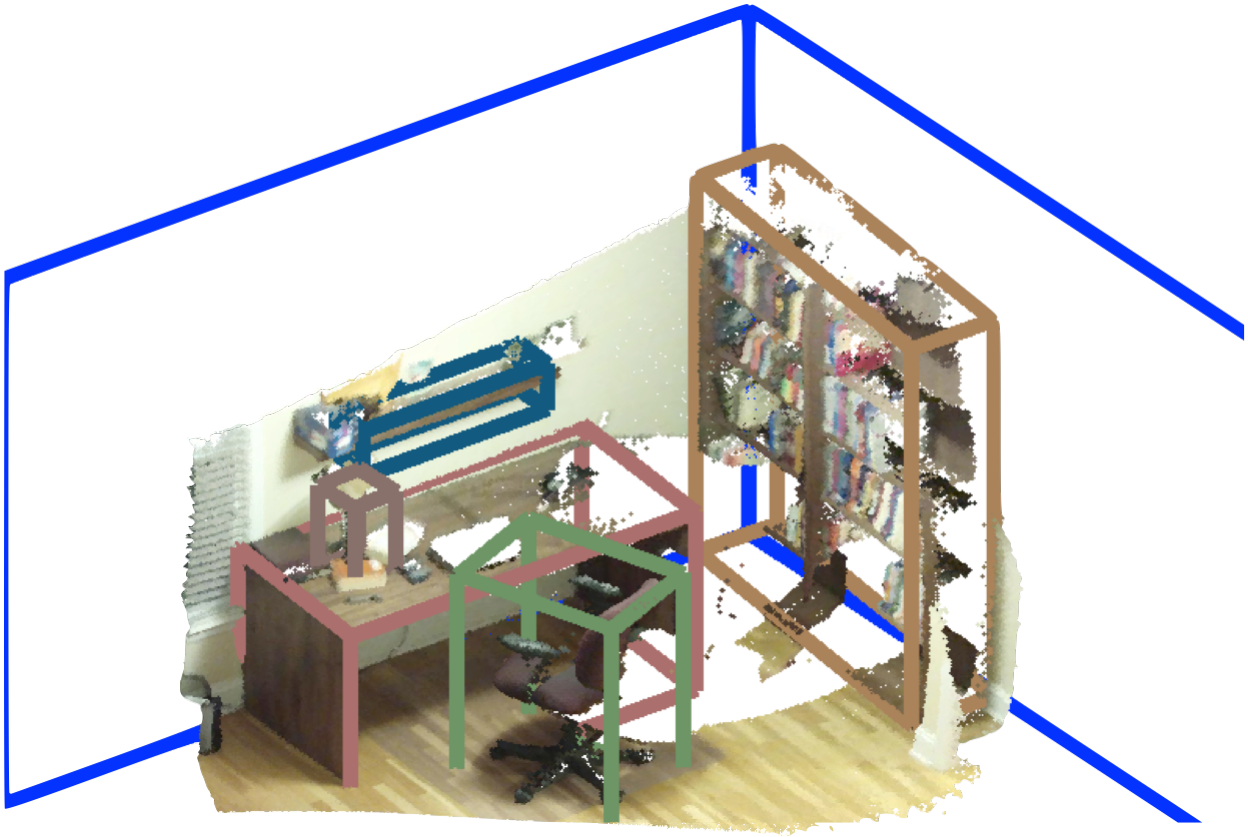


chair

bookshelf

3D Object Orientation

Applications about SUN RGB-D



Total Scene Understanding

Other RGBD datasets application

1. Semantics
2. Object pose estimation
3. Camera tracking
4. Scene reconstruction
5. Object tracking
6. Human actions, faces and identification

~100 RGB-D Datasets

Outline

- Motivation
- RGB-D images and applications
- Typical 2D object detection methods
- 3D object detection in RGB-D images

2D object detection methods

1. **RCNN**

- ❑ Region Proposal + CNN (mAP 58% in VOC2007, DPM 43%)
- ❑ Transfer detection problem to classification problem

2. **SPP-Net**

- ❑ Feature map region proposal represent original image region proposal

3. **Fast RCNN**

- ❑ Output bounding box and label together (mAP 58%)

4. **Faster RCNN**

- ❑ RPN represent selective search (mAP 68%)

5. **YOLO** Real time (mAP 63.4%)

6. **Region based FCN** (Deep Residual Network) (mAP 83.4%)

RCNN Pipeline

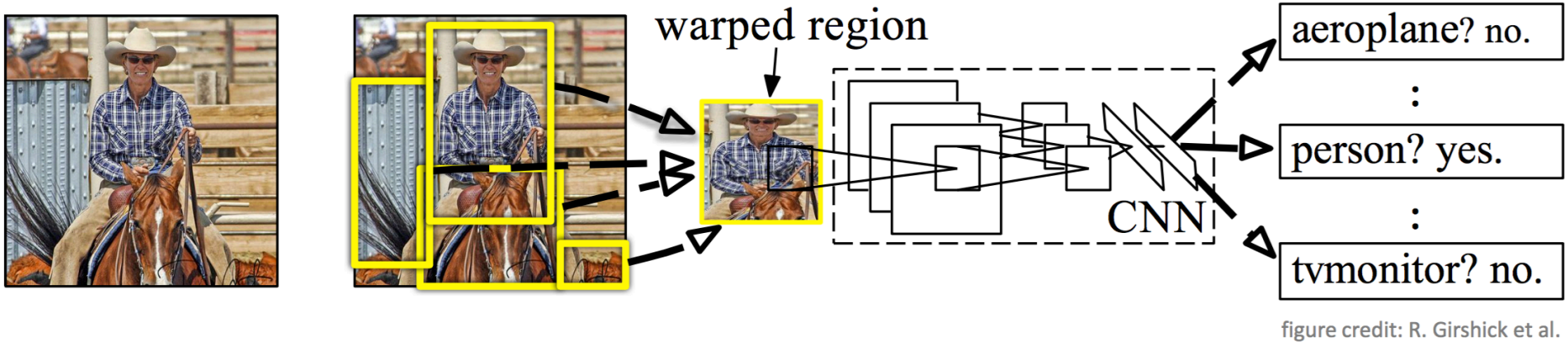


figure credit: R. Girshick et al.

input image

region proposals
~2,000

Region of Interest from selective search

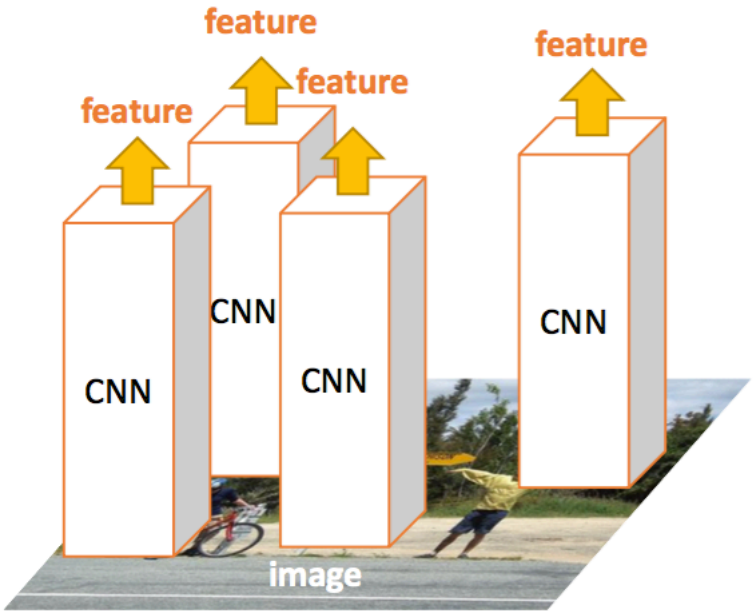
1 CNN for each region

classify regions

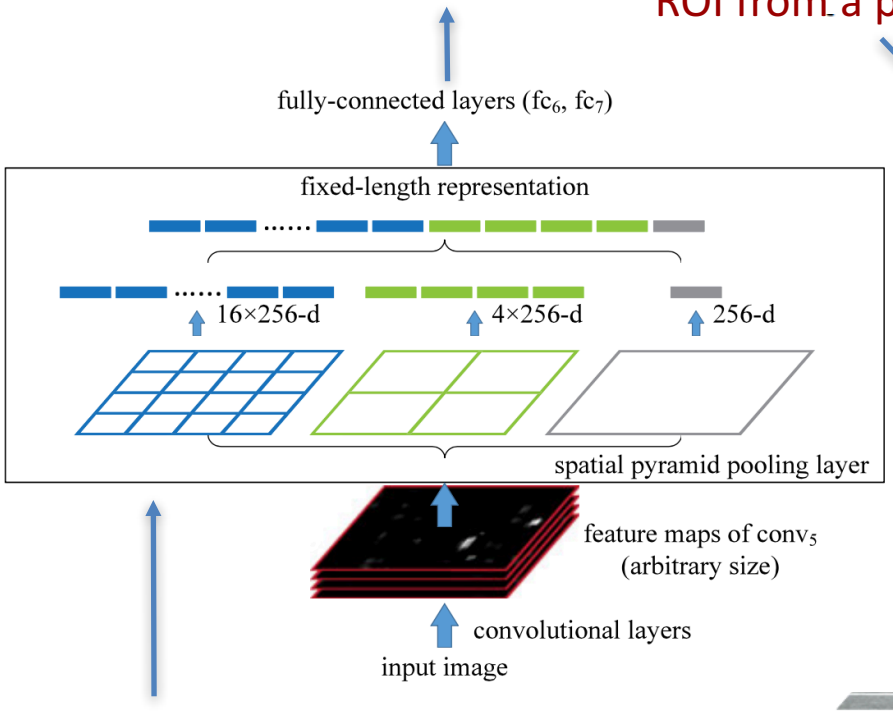
SVM + Bbox reg

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." CVPR 2014.

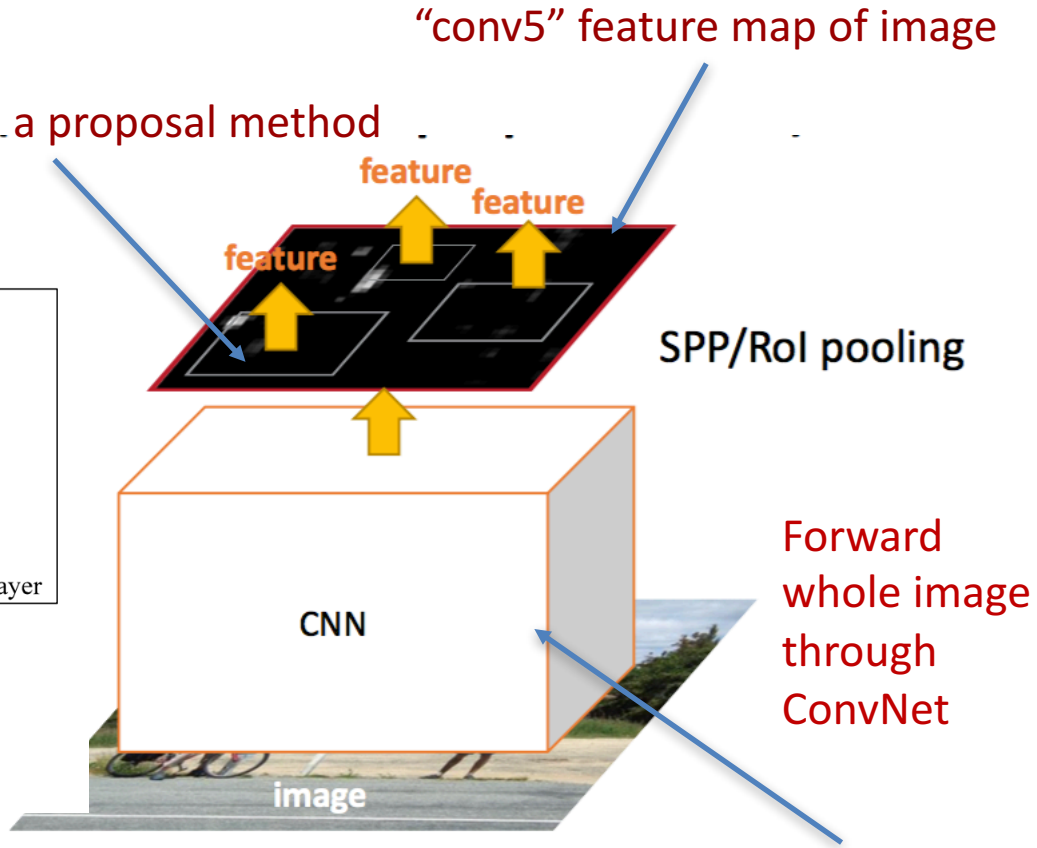
RCNN



SPP-Net & Fast RCNN



Deal with Different scale regions



End to End training and testing

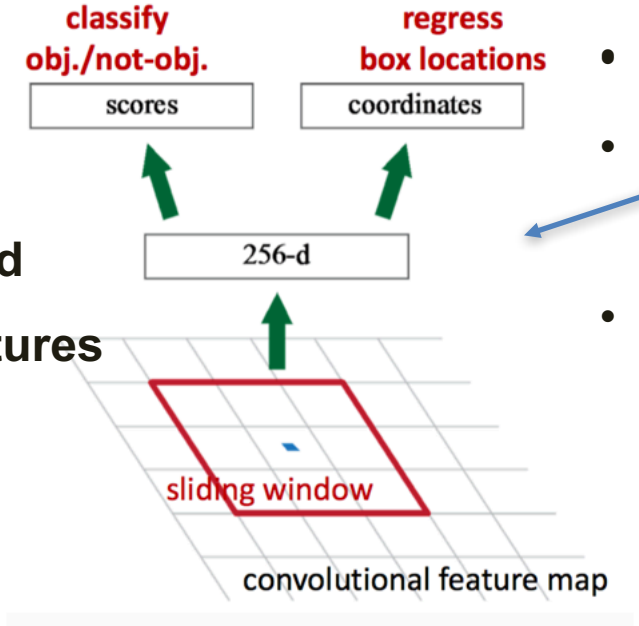
He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." ECCV 2014

Girshick, Ross. "Fast r-cnn." CVPR 2015

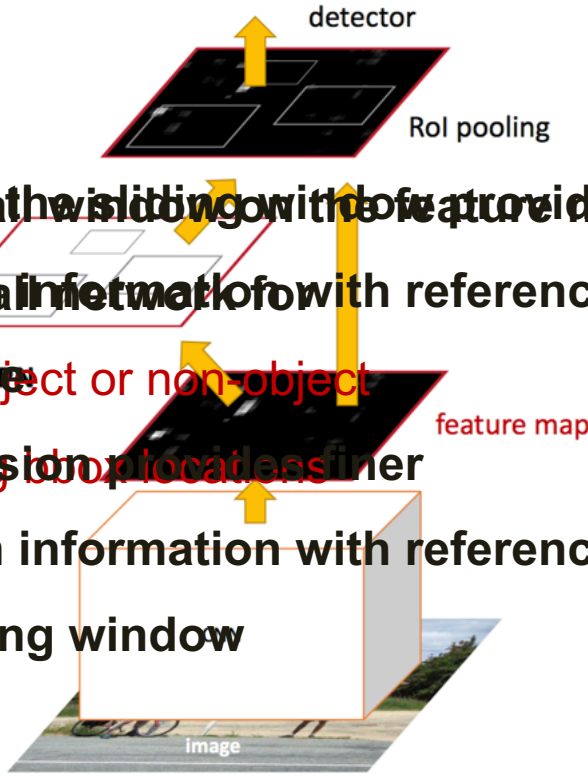
Faster RCNN

1. The speed limitation is the region proposal method

1. FC
2. End to End
3. Share features



- Slide a small window on the feature map
- Build a small network for:
  - Classify object or non-object
  - Regressing box locations
- Box regression provides finer localization information with reference to this sliding window



~300 region proposals instead of ~2000 proposals

Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." NIPS 2015

YOLO

1. Real-time bounding box
2. Using regression instead of region based CNN
3. Faster but relatively lower mAP than faster RCNN

YOLO Video

Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." arXiv preprint arXiv:1506.02640 (2015).

Outline

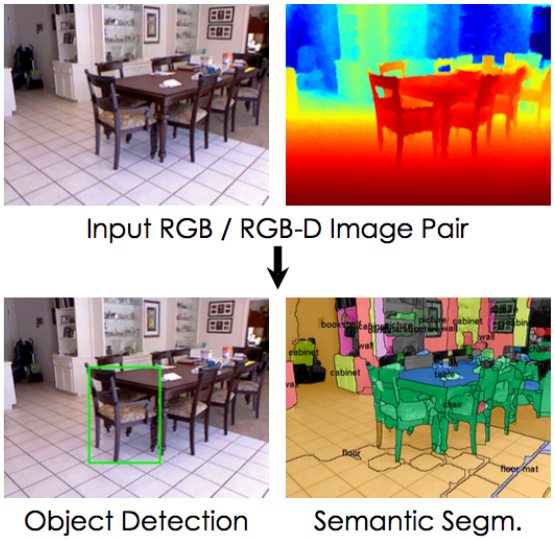
- Motivation
- RGB-D images and applications
- Typical 2D object detection methods
- 3D object detection in RGB-D images

2D approach for 3D object detection

Depth RCNN:

detect objects in 2D image plane by treating depth as extra channels of RGB image

then fit a model to the points inside the 2D detected window by using ICP alignment.



Input representations

- **HHA (depth, height above ground, angle with gravity)**
- **Normal Images**

S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. "Learning rich features from RGB-D images for object detection and segmentation". ECCV 2014

2D approach for 3D object detection

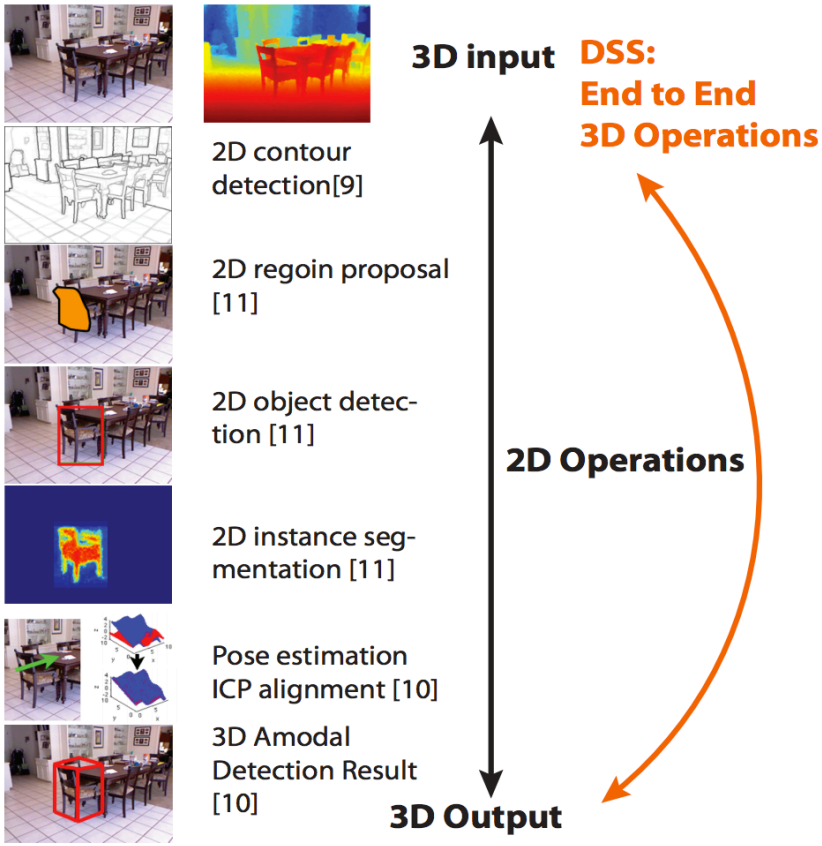
Improved contours from RGB-D

Much better region proposals

Adapting CNNs trained on RGB images to Depth Images

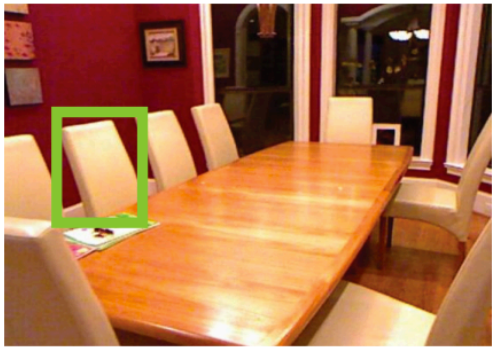
A geocentric embedding for Depth images (HHA)

Extend this for task of instance segmentation



3D approach for 3D object detection

They propose 3D Region Proposal Network to learn object from geometric shapes and using Object Recognition Network to extract geometric features in 3D and color features in 2D



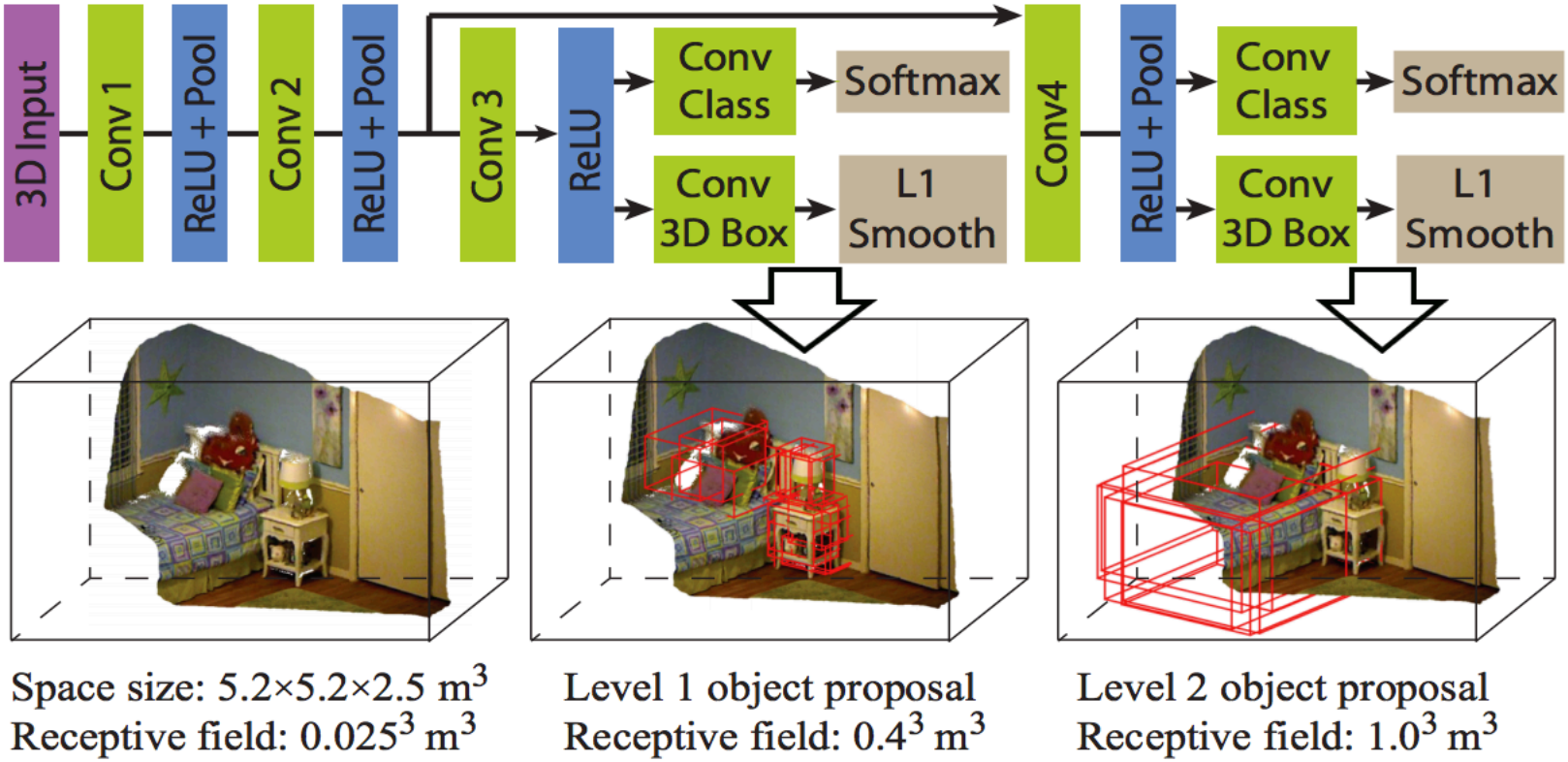
2D Modal

2D Amodal

3D Amodal

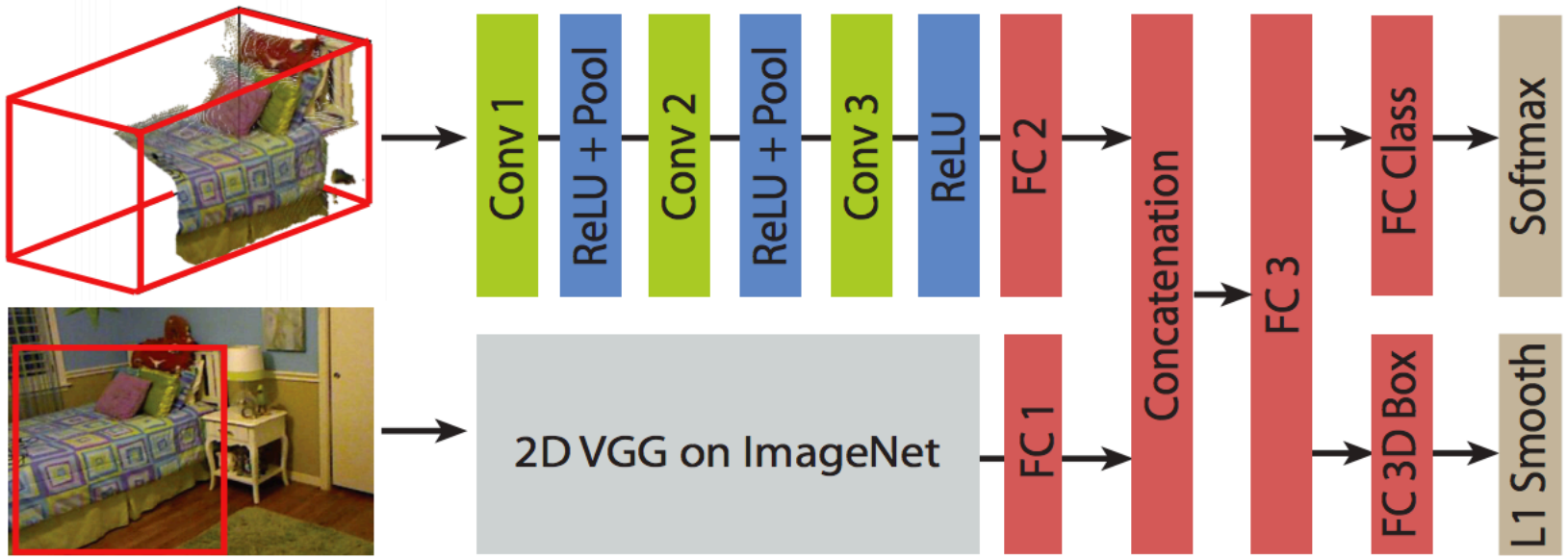
Shuran Song, Jianxiong Xiao "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images ", CVPR2016

3D Amodal Region Proposal Network



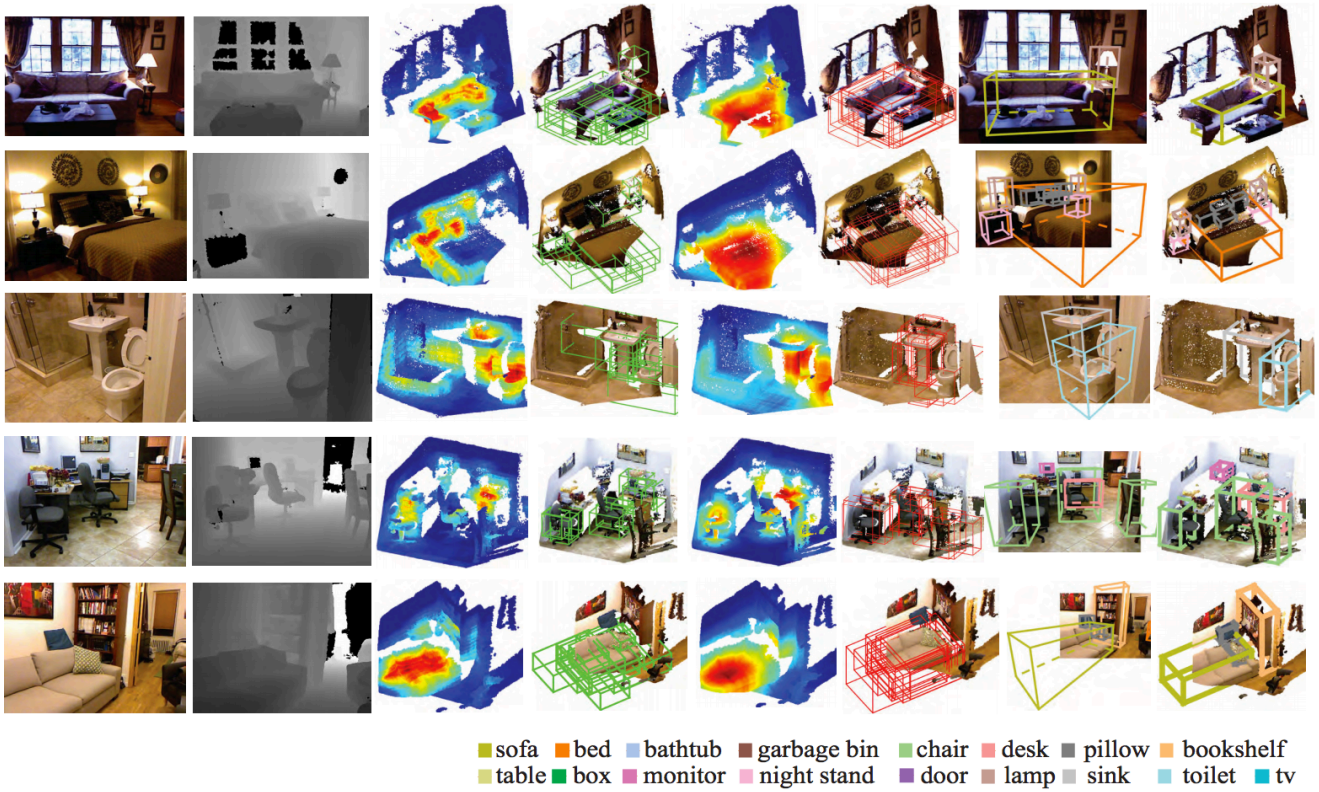
Taking a 3D volume from depth as input, the fully convolutional 3D network extracts 3D proposals at two scales with different receptive fields.

Joint Object Recognition Network



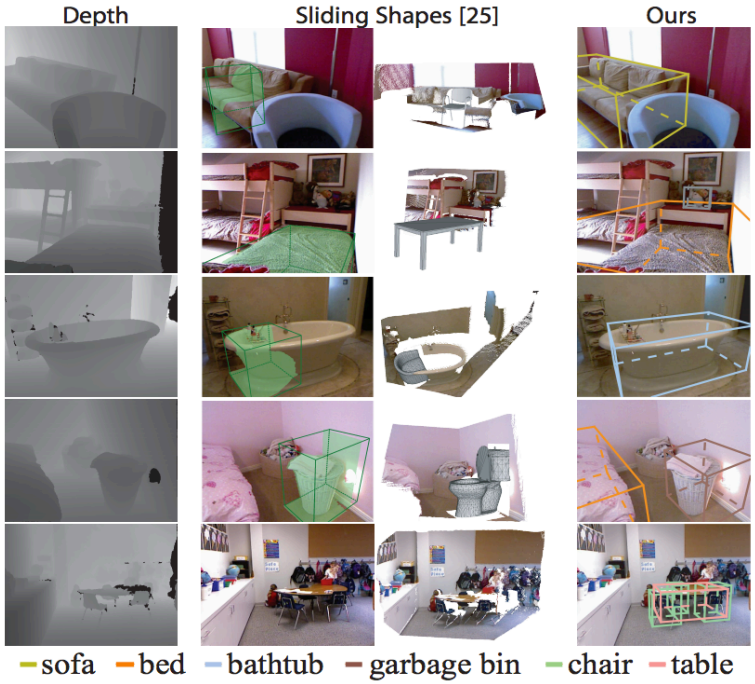
For each 3D proposal, they feed the 3D volume from depth to a 3D ConvNet, and feed the 2D color patch to a 2D ConvNet, to jointly learn object categories and 3D box regression.

Results



Examples for Detection Results. For the proposal results, they show the heat map for the distribution of the top proposals (red is the area with more concentration), and a few top boxes after NMS. For the recognition results, their amodal 3D detection can estimate the full extent of 3D both vertically and horizontally.

Comparison



Comparison with Sliding Shapes. Deep Sliding Shapes is able to better use shape, color, and contextual information to handle more object categories, resolve ambiguous cases, and detect objects with atypical sizes.

S. Song and J. Xiao. "Sliding Shapes for 3D object detection in depth images." ECCV, 2014

Evaluations

Algorithm	input						mAP
Sliding Shapes [25]	d	33.5	29	34.5	33.8	67.3	39.6
[10] on instance seg	d	71.0	18.2	30.4	49.6	63.4	46.5
[10] on instance seg	rgbd	74.7	18.6	28.6	50.3	69.7	48.4
[10] on estimated model	d	72.7	47.5	40.6	54.6	72.7	57.6
[10] on estimated model	rgbd	73.4	44.2	33.4	**57.2**	84.5	58.5
ours [depth only]	d	83.0	58.8	68.6	49.5	79.2	67.8
ours [depth + img]	rgbd	**84.7**	**61.1**	**70.5**	55.4	**89.9**	**72.3**

Comparison on 3D object detection in SUN RGB-D dataset, first line: hand-craft feature, second-fifth lines: depth RCNN, 2D approach, last two lines: 3D ConvNets.

Shuran Song, Jianxiong Xiao "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images ", CVPR2016

Thank you