Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images

Shuran Song [1] Jianxiong Xiaor[1]

[1]Princeton University

Presented by : Melissa Mozifian
May 17, 2017

Overview

Motivation

- Deep ConvNets have revolutionized 2D object detection
  - RCNN, Fast RCNN, Faster RCNN are three iterations of the most successful state of the art object detectors
- More research focus on 3D object detection

Dataset

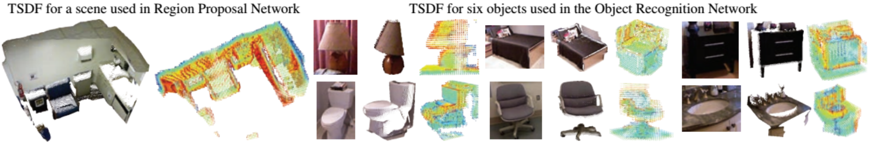
- SUN RGB-D: A RGB-D Scene Understanding Benchmark
- NYU Depth Dataset
- Evaluation:
    - Average Precision (AP) per class
    - mean Average Precision

3D Object Detection

- 3D formulation to learn object proposals and classifiers using 3D convolutional neural networks (ConvNets)
- Challenges:
  - Need to come up with a way to encode 3D representation
  - 3D volumetric representation requires more memory and computation

Encoding 3D Representation

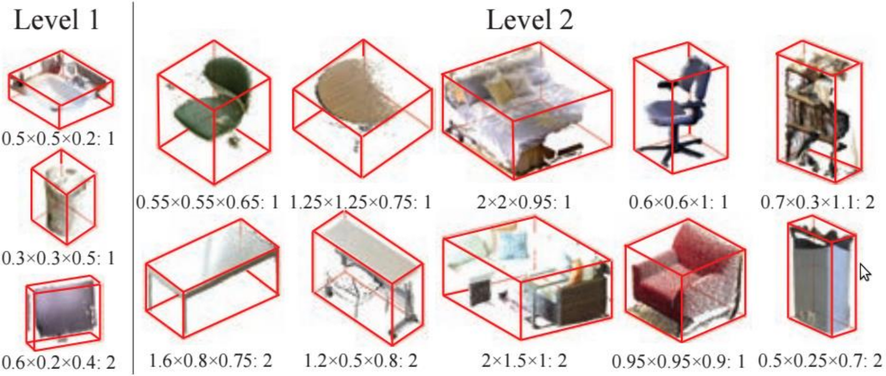
- Encode the geometric shapes in 3D while preserving spatial locality
  - Using directional Truncated Signed Distance Function (TSDF)
  - The resolution is 208x208x100 for the Region Proposal Network, and 30x30x30 for the Object Recognition Network



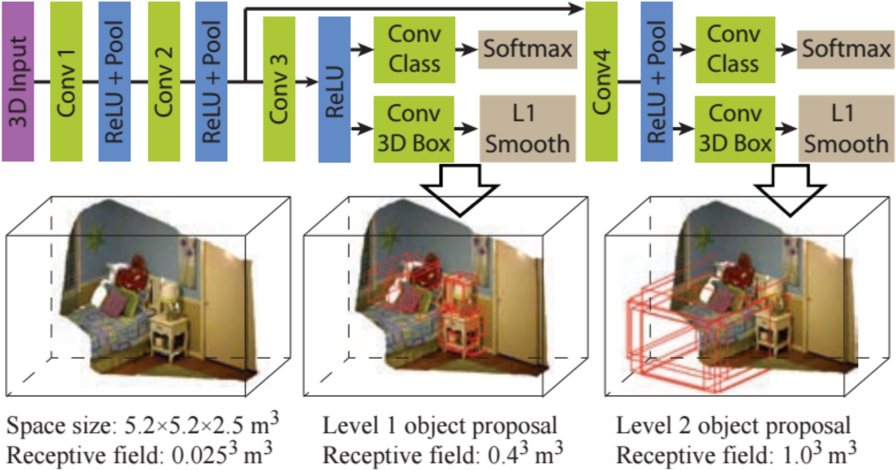
Multi-scale 3D Region Proposal Network

- Extra dimension, increases the possible location for an object by 30 times (45 thousand windows per image in 2D vs 1.4 million in 3D)
- Variation in pixel areas of similar objects with different 3D physical sizes, e.g. bed and a chair

Anchor Types



3D Amodal Region Proposal Network



RPN Multi-task Loss

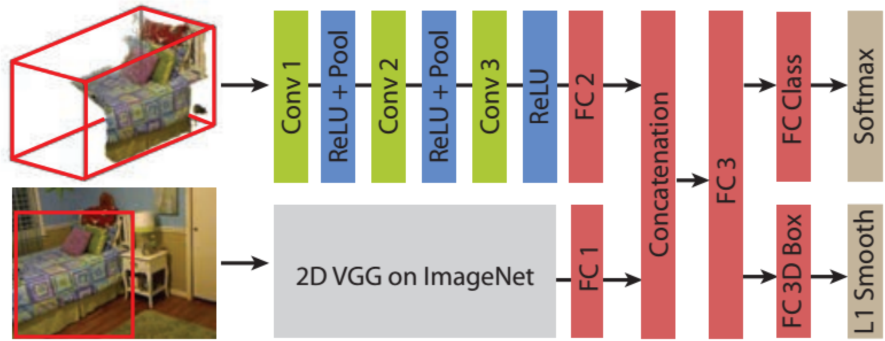
$$L(p, p^*, \mathbf{t}, \mathbf{t}^*) = L_{cls}(p, p^*) + \lambda p^* L_{reg}(\mathbf{t}, \mathbf{t}^*)$$

- First term is objectness score
- Second term is for the box regression
- p is the predicted probability of anchor being an object
- p^* is the ground truth
- L_{cls} is log loss over two categories (object vs non-object)
- L_{reg} is smooth $\mathbf{L}_1$ loss

3D NMS

- Adopt 3D Non-Maximum Suppression (NMS) to remove redundant proposals
- Thresholding IOU in 3D to pick the top 2000 boxes
- Key factor to speed up the algorithm

Joint Object Recognition Network



3D ConvNet Features

2D t-SNE embedding of 5000 foreground volumes using their the last layer features learned from the 3D ConvNet

2D Object Detection

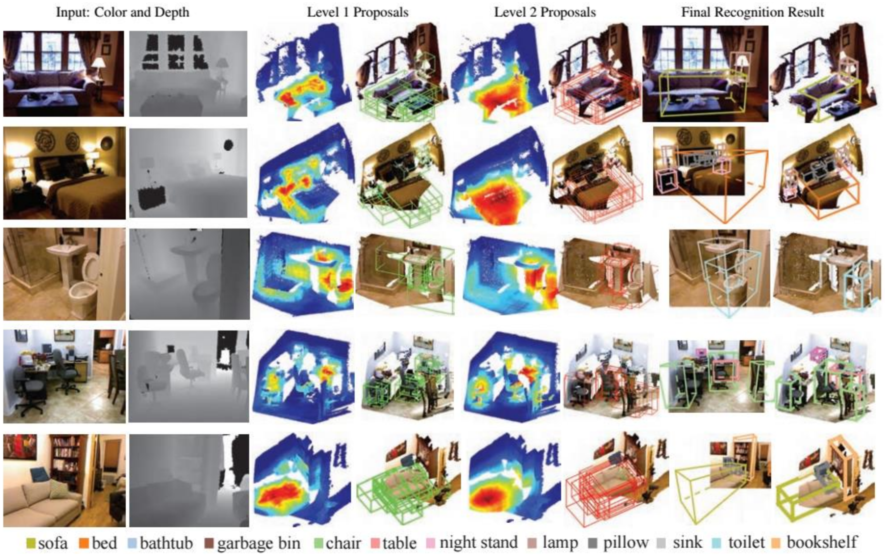
- Project the 3D points inside the proposal box to 2D image plane
- VGGnet pre-trained on ImageNet to extract color features
- Region-of-Interest Pooling Layer from Fast RCNN

ORN Multi-task Loss

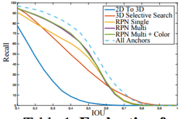
$$L(p, p^*, \mathbf{t}, \mathbf{t}^*) = L_{cls}(p, p^*) + \lambda^{'}[p^* > 0]L_{reg}(\mathbf{t}, \mathbf{t}^*)$$

- p is the predicted probability over 20 categories (negative non-objects is labeled as class 0)

Results


Input: Color and Depth
Level 1 Proposals
Level 2 Proposals
Final Recognition Result
sofa
bed
bathtub
garbage bin
chair
table
night stand
lamp
pillow
sink
toilet
bookshelf

Results Table



																				Recall	ABO	#Box
2D To 3D	41.7	53.5	37.9	22.0	26.9	46.2	42.2	11.8	47.3	33.9	41.8	12.5	45.8	20.7	49.4	55.8	54.1	15.2	50.0	34.4	0.210	2000
3D Selective Search	79.2	80.6	74.7	**66.0**	66.5	92.3	80.9	**53.9**	89.1	**89.8**	**83.6**	**45.8**	85.4	75.9	83.1	85.5	80.9	**69.7**	83.3	74.2	0.409	2000
RPN Single	87.5	98.7	70.1	15.6	95.0	100.0	93.0	20.6	94.5	49.2	49.1	12.5	100.0	34.2	81.8	94.9	93.3	57.6	96.7	75.2	0.425	2000
RPN Multi	100.0	**98.7**	**73.6**	42.6	94.7	**100.0**	92.5	21.6	**96.4**	78.0	69.1	37.5	**100.0**	75.2	97.4	97.1	96.4	66.7	100.0	84.4	0.460	2000
RPN Multi Color	**100.0**	98.1	72.4	42.6	**95.0**	**100.0**	**93.0**	19.6	**96.4**	79.7	76.4	37.5	**100.0**	**79.0**	**97.4**	**97.1**	**95.4**	57.6	**100.0**	**84.9**	**0.461**	2000
All Anchors	100.0	98.7	75.9	50.4	97.2	100.0	97.0	45.1	100.0	94.9	96.4	83.3	100.0	91.2	100.0	97.8	96.9	84.8	100.0	91.0	0.511	107674

Table 1. **Evaluation for Amodal 3D Object Proposal.** [All Anchors] shows the performance upper bound when using all anchors.

poposal	algorithm																				mAP
3D SS	dxdydz no bbreg	43.3	55.0	16.2	23.1	3.4	10.4	17.1	30.7	10.9	35.4	20.3	41.2	47.2	25.2	43.9	1.9	1.6	0.1	9.9	23.0
	dxdydz	52.1	60.5	19.0	30.9	2.2	15.4	23.1	36.4	19.7	36.2	18.9	52.5	53.7	32.7	56.9	1.9	0.5	0.3	8.1	27.4
RPN	dxdydz no bbreg	51.4	74.8	7.1	51.5	15.5	22.8	24.9	11.4	12.5	39.6	15.4	43.4	58.0	40.7	61.6	0.2	0.0	1.5	2.8	28.2
	dxdydz no size	59.9	78.9	12.0	51.5	15.6	24.6	27.7	12.5	18.6	42.3	15.1	59.4	59.6	44.7	62.5	0.3	0.0	1.1	12.9	31.5
	dxdydz	59.0	80.7	12.0	59.3	15.7	25.5	28.6	12.6	18.6	42.5	15.3	59.5	59.9	45.3	64.8	0.3	0.0	1.4	13.0	32.3
	tsdf dis	61.2	78.6	10.3	61.1	2.7	23.8	21.1	25.9	12.1	34.8	13.9	49.5	61.2	45.6	70.8	0.3	0.0	0.1	1.7	30.2
	dxdydz+rgb	58.3	79.3	9.9	57.2	8.3	27.0	22.7	4.8	18.8	46.5	14.4	51.6	56.7	45.3	65.1	0.2	0.0	4.2	0.9	30.1
	proj dxdydz+img	58.4	81.4	20.6	53.4	1.3	32.2	36.5	18.3	17.5	40.8	19.2	51.0	58.7	47.9	71.4	0.5	0.2	0.3	1.8	32.2
	dxdydz+img+hha	55.9	83.0	18.8	63.0	17.0	33.4	43.0	33.8	16.5	54.7	22.6	53.5	58.0	49.7	75.0	2.6	0.0	1.6	6.2	36.2
	dxdydz+img	62.8	82.5	20.1	60.1	11.9	29.2	38.6	31.4	23.7	49.6	21.9	58.5	60.3	49.7	76.1	4.2	0.0	0.5	9.7	36.4

Table 2. **Control Experiments on NYUv2 Test Set.** Not working: box (too much variance), door (planar), monitor and tv (no depth).

Results continued

																				Recall	ABO	#Box
3D SS	78.8	87.2	72.8	72.2	65.5	86.1	75.1	65.0	70.0	87.1	67.5	53.1	68.1	82.8	86.8	84.4	85.0	69.2	94.0	72.0	0.394	2000
RPN	98.1	99.1	79.5	51.5	93.3	89.2	94.9	24.0	87.0	79.6	62.0	41.2	96.2	77.9	96.7	97.3	96.7	63.3	100.0	88.7	0.485	2000

Table 4. **Evaluation for regoin proposal generation on SUN RGB-D test set.**

																				mAP
Sliding Shapes [25]	-	42.09	-		33.42	-	-	-	-	-	-	-	-	-	-	23.28	25.78	-	61.86	-
Deep Sliding Shapes	44.2	78.8	11.9	1.5	61.2	4.1	20.5	0.0	6.4	20.4	18.4	0.2	15.4	13.3	32.3	53.5	50.3	0.5	78.9	26.9

Table 5. **Evaluation for 3D amodal object detection on SUN RGB-D test set.**

Limitations

bookshelf chair dresser garbage bin sofa box lamp door door tv

Misses

(1) chair (2) tv (3) bookshelf (4) sofa (5) bed (6) monitor (7) desk (8) night stand (9) garbage bin (10) box

False Positives

Summary

- Great potential of learning 3D shape representation
- Main contribution : Encoding 3D Representation to preseve most important features in both 3D and 2D
- Limitations/Future work
  - Detection still limited by the two level sizes proposed
  - Improving detection of smaller objects
  - Improve speed : RPN takes 5.62s and ORN takes 13.93s per image during testing