3D Instances from a single RGB Image

Master Thesis Final Presentation

Peter Mortimer

Advisor: Yida Wang

Supervisor: Federico Tombari

September 6, 2019

Table of Contents

Introduction

Introduction

Problem Statement

We compare different output representations on a deep learning architecture developed for the task of 3D scene understanding from a single RGB image. In particular, we focus on a factored object representation and a point cloud representation.

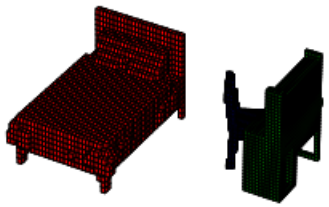

Figure 1 Possible output of a 3D scene understanding model from a single RGB image.

Related Work

Related Work - Factored3D

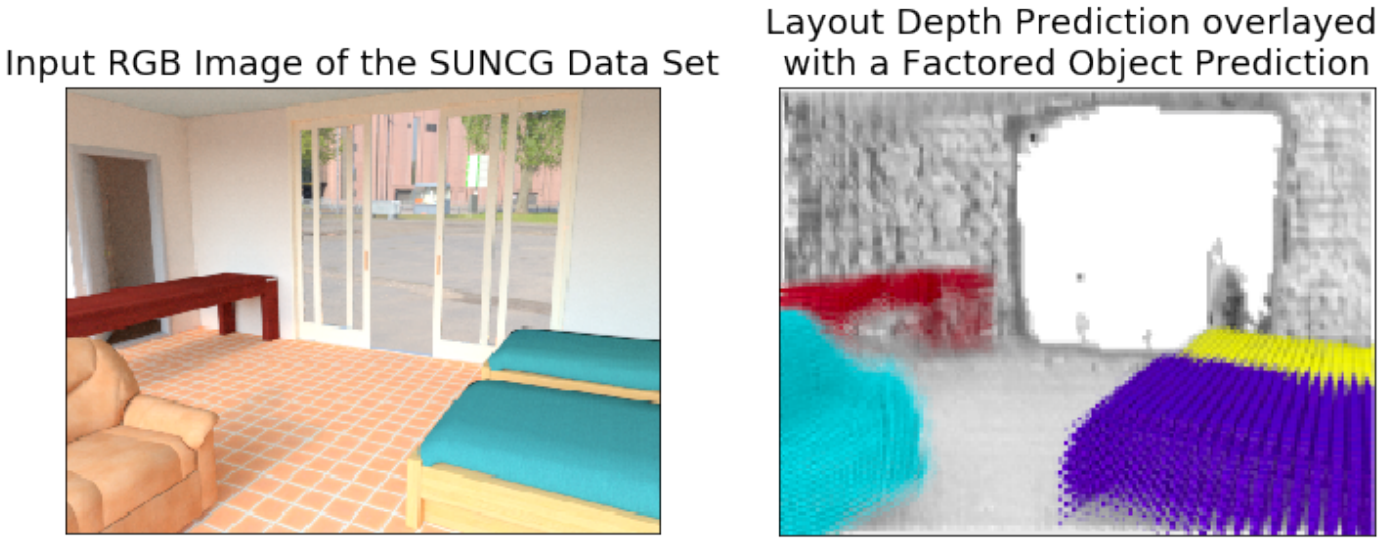

Figure 2 Example Prediction output of Factored3D [1]

- 3D scene prediction based on a single RGB image
- trained on a data set of synthetic indoor scenes (SUNCG)
- master thesis extends on this architecture

Related Work - SUNCG Data Set I

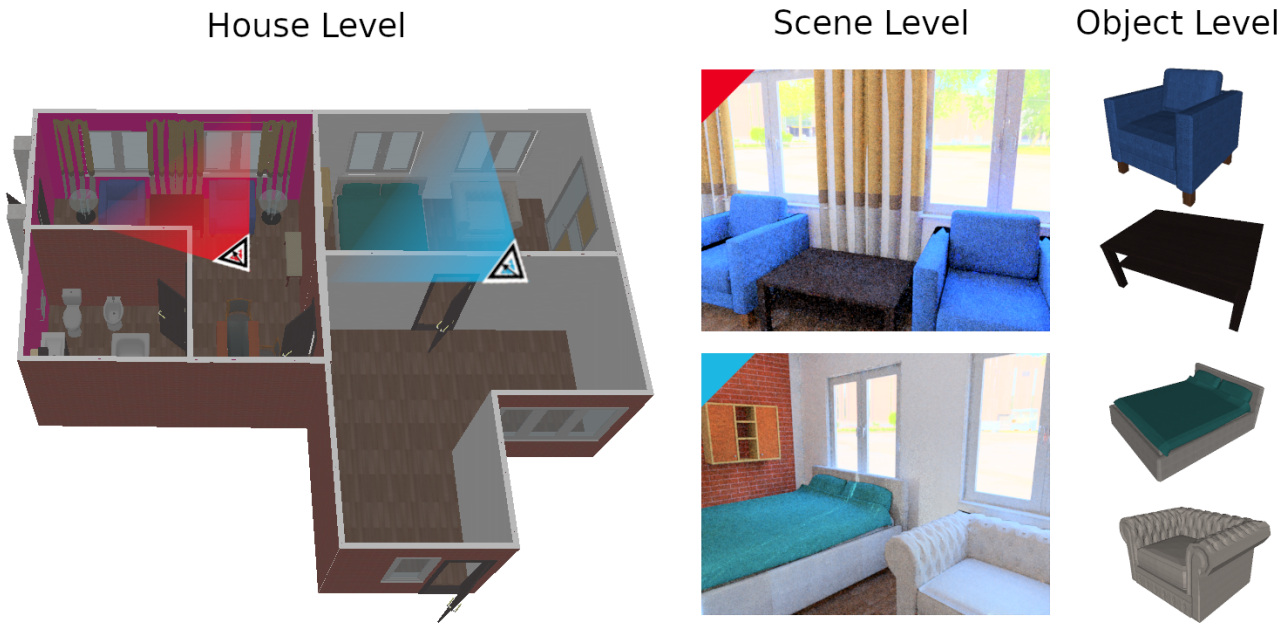

Figure 3 An example house from the SUNCG data set including two camera positions and objects found in each scene [2, 3]

- indoor scenes created by humans on the Planner5D platform
- images are generated using physically-based rendering
- we limit our object prediction to the six object classes: {bed, chair, desk, sofa, table, television}

Related Work - SUNCG Data Set II

SUNCG Data Set	
total number of houses	41 497
total number of images	568 561
image dimensions	640px $\times$ 480px
total number of valid objects	845
average number of valid objects per image	1.6

Table 1 Overview statistics of the SUNCG data set.

- largest data set of indoor scenes with semantic and 3D geometric information to date
- scenes contain an object density close to real world indoor scenes
- currently unavailable due to a lawsuit on copyright infringement

Methodology

Methodology - Network Architecture

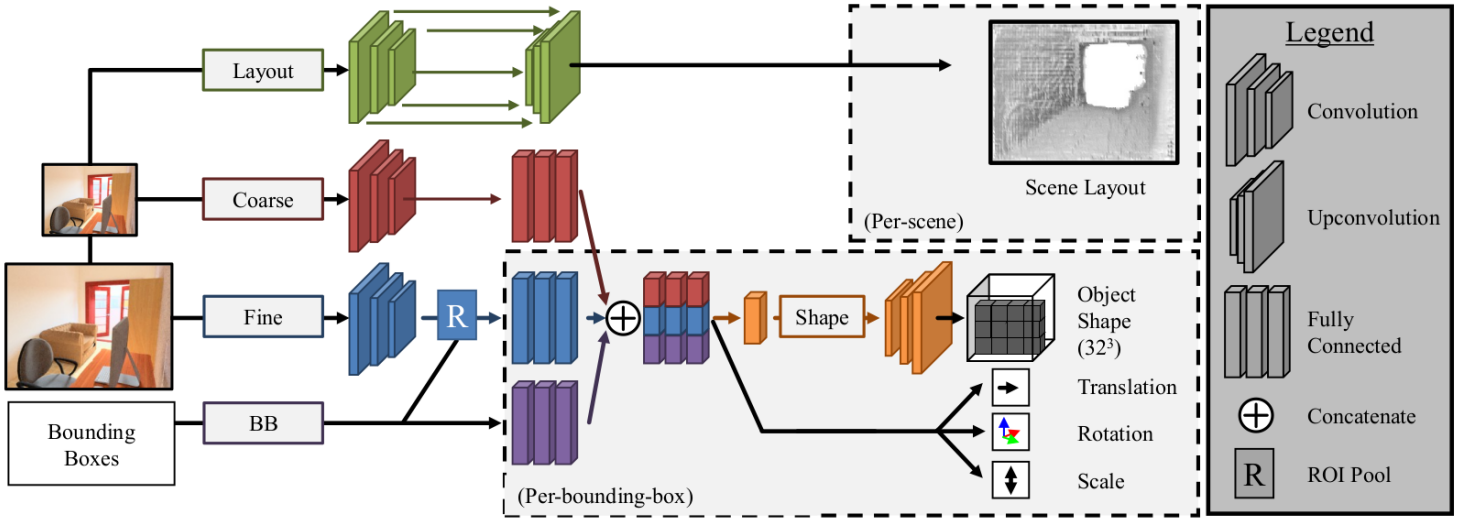

Figure 5 Factored3D network architecture [1]

- The Layout Module predicts the amodal scene layout as a depth map.
- The Coarse Module, Fine Module, and Bounding Box Module are used to produce one instance prediction per detected bounding box.
- The bounding box proposals are produced using EdgeBoxes [4].

Methodology - Point Cloud Transformation I

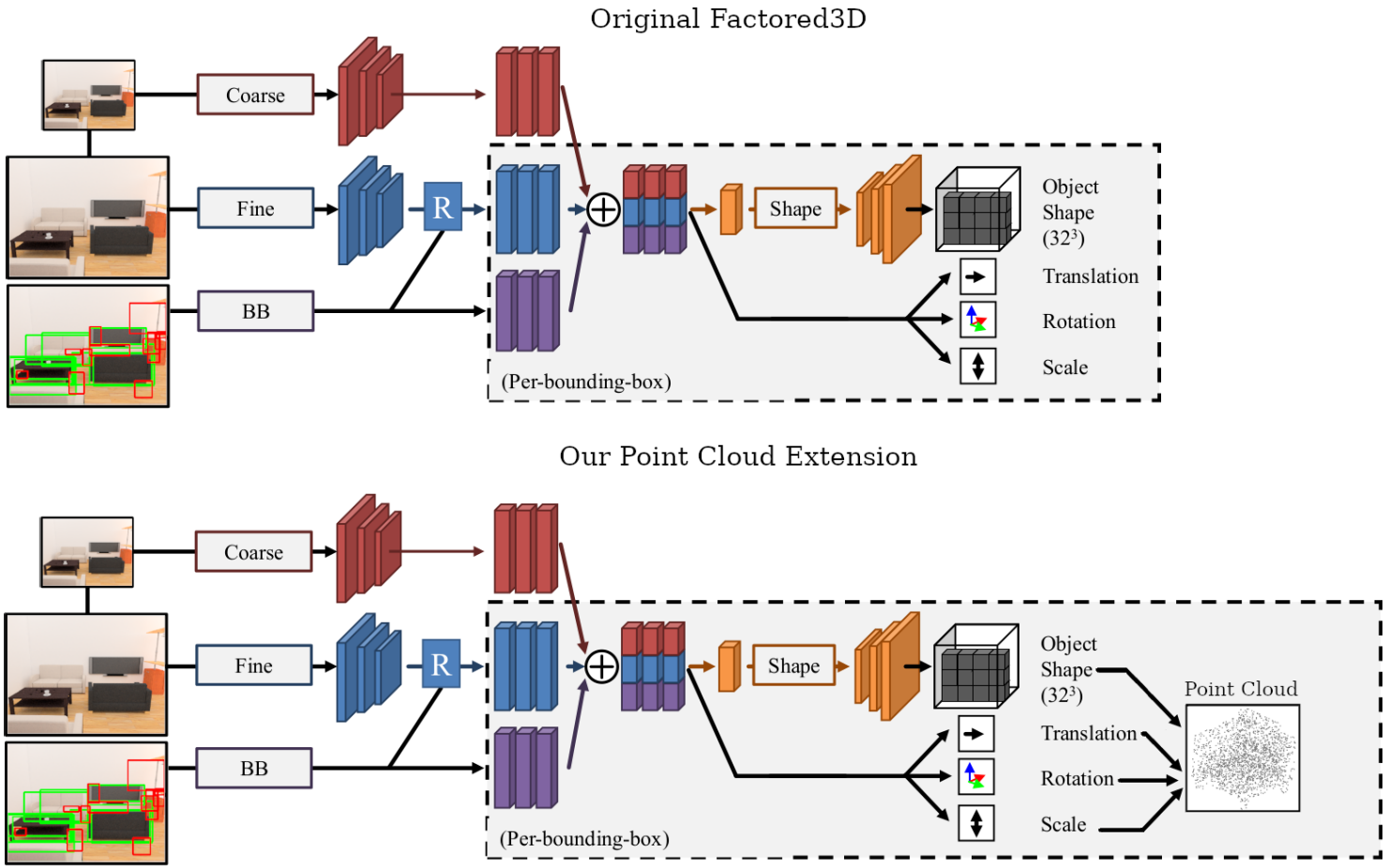

Figure 6 The Factored3D architecture is extended with a point cloud transformation step. It combines the factored outputs to a point cloud representation in camera space.

Methodology - Point Cloud Transformation II

We transform a voxel shape prediction $\hat{\mathbf{V}} \in \mathbb{R}^{32 \times 32 \times 32}$ to a fixed size point cloud $\hat{\mathbf{P}}_{cam} \in \mathbb{R}^{1024 \times 3}$ in camera space.
We define:

- $\mathbf{m}$: selection mask for a random sample of voxel predictions above δ_{vox}
- $\mathbf{I} \in \mathbb{R}^{32^3 \times 3}$: index matrix denoting the point coordinate for each voxel grid entry

We apply the following steps:

1. flatten $\hat{\mathbf{V}}$ and take the ceiling of all voxel predictions
2. apply the selection mask $\mathbf{m}$ on $\hat{\mathbf{V}}_{flat}$ and $\mathbf{I}$
3. apply the element-wise multiplication $\odot$ to construct $\hat{\mathbf{P}}_{vox} = \mathbf{m}(\hat{\mathbf{V}}_{flat}) \odot \mathbf{m}(\mathbf{I})$
4. construct a transformation matrix $\hat{\mathbf{T}}_{cam}$ from the scale, pose, and translation predictions $(\hat{\mathbf{c}}, \hat{\mathbf{q}}, \hat{\mathbf{t}})$ to transform $\hat{\mathbf{P}}_{vox}$ into $\hat{\mathbf{P}}_{cam}$

Methodology - Point Cloud Cost Functions

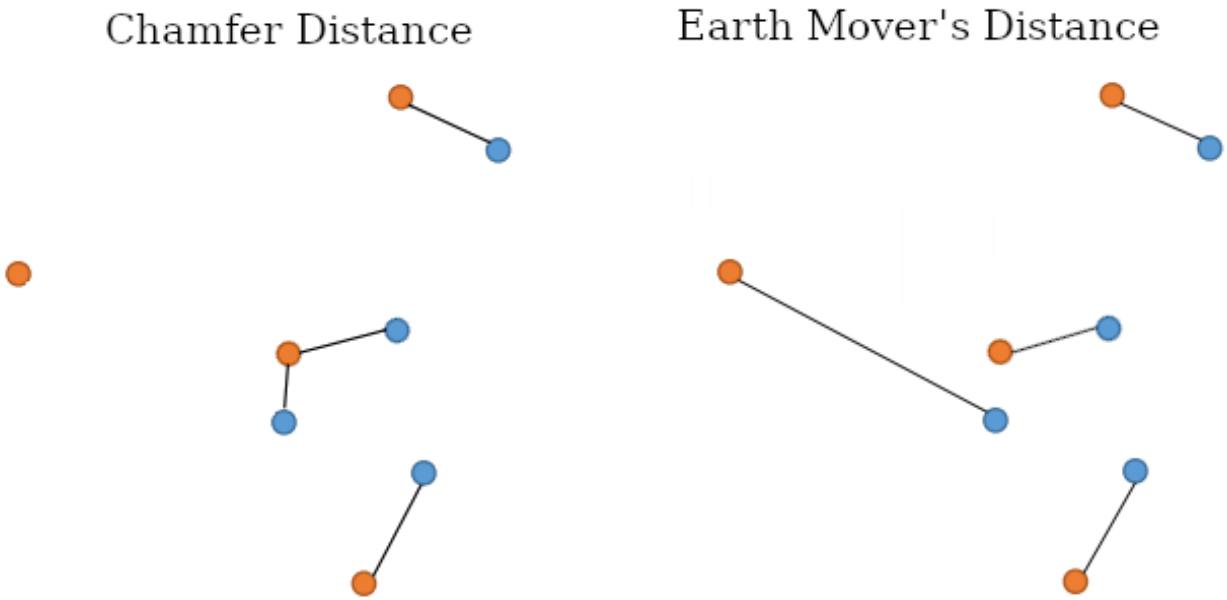

Figure 7 The Chamfer distance (D_{CD}) and the Earth Mover's distance (D_{EMD}) can produce different point matchings.

$$D_{CD}(S_1, S_2) = \max \left\{ \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2, \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2 \right\}$$

$$D_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} \frac{1}{|S_1|} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

Results

Quantitative Comparison - Instance Level

	Shape		Rotation		Translation		Scale	
method	%(IoU $>$ 0.25)	Med-IoU	%(Err $<$ 30°)	Med-Err	%(Err $<$ 1m)	Med-Err	%(Err $<$ 0.5)	Med-Err
Factored3D	47.80	0.19	**78.10**	5.20	**92.50**	0.27	88.30	0.11
Factored3D + D_{CD}	48.00	0.19	77.50	5.27	92.20	0.28	86.00	0.15
Solo D_{CD}	46.30	0.17	76.60	5.60	92.30	0.28	**51.00**	**0.49**
Factored3D + pose-only D_{CD}	47.00	0.16	**68.40**	**9.70**	91.20	0.30	88.50	0.11
Factored3D + pose-only D_{EMD}	47.10	0.18	75.30	5.87	91.40	0.31	**88.80**	0.11
handcrafted baseline	**63.89**	0.35					70.87	0.27

Table 2 Performance predictions on the ground-truth bounding boxes of the SUNCG test set.

Quantitative Comparison - Scene Level

method	%(IoU $>$ 0.20)	Med-IoU
Factored3D	**56.52**	**0.228**
Factored3D + D_{CD}	53.48	0.215
Solo D_{CD}	41.35	0.167
Factored3D + pose-only D_{CD}	51.56	0.207
Factored3D + pose-only D_{EMD}	**55.54**	**0.223**

Table 3 Scene prediction performance on input images from the SUNCG test set.

- all objects are transformed into a scene voxel occupancy grid with $64 \times 32 \times 64$ voxels, where each voxel represents a $8cm \times 8cm \times 8cm$ volume
- objects are all transformed to world space before voxelization
- only consider voxels in the scene as occupied or unoccupied by an object

Qualitative Comparison - Scene Level I

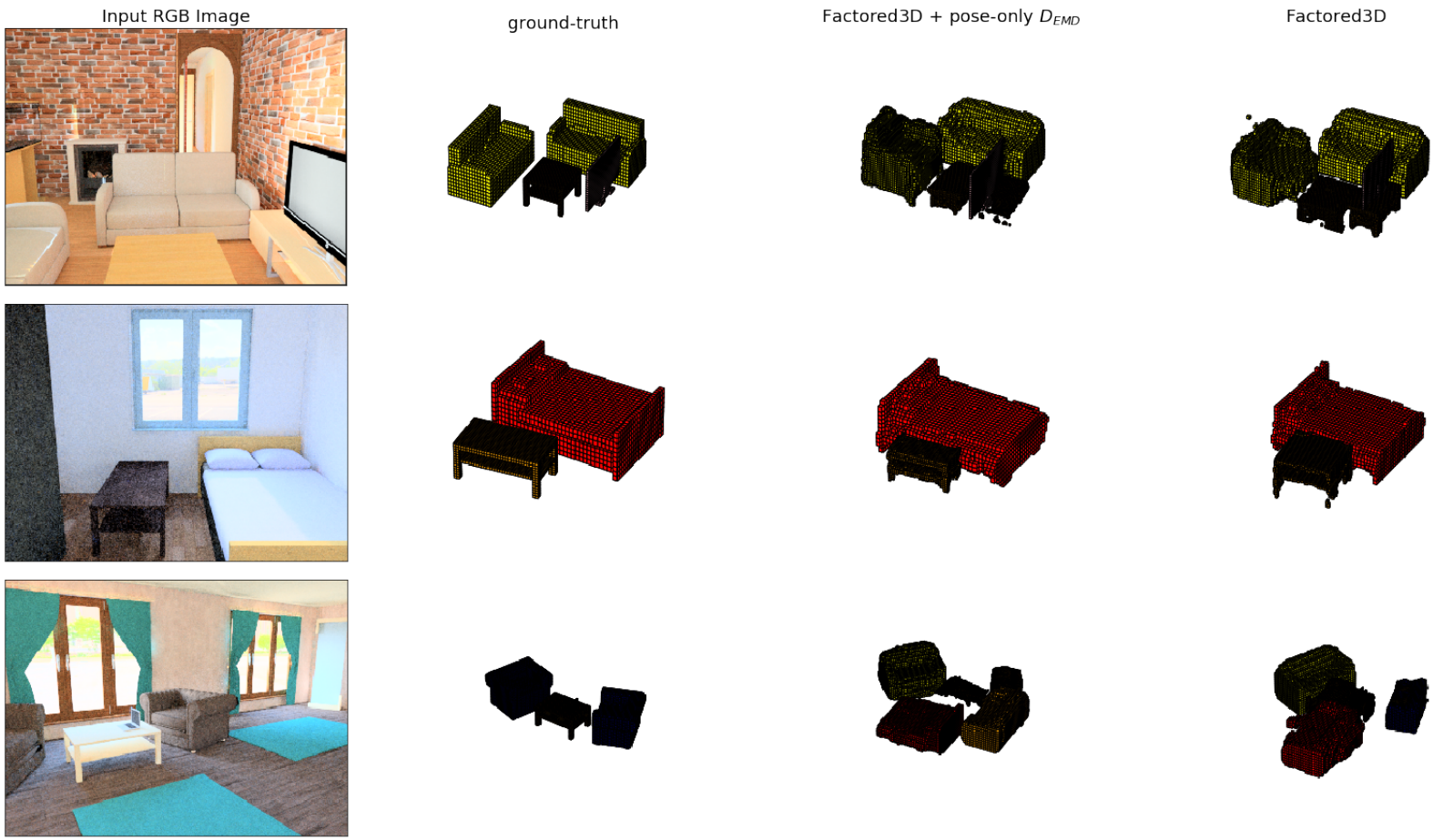

Figure 8 Comparison of the object prediction of the original Factored3D network to the extension using the point cloud based D_{EMD} as loss function on the rotated shape prediction.

Qualitative Comparison - Scene Level II

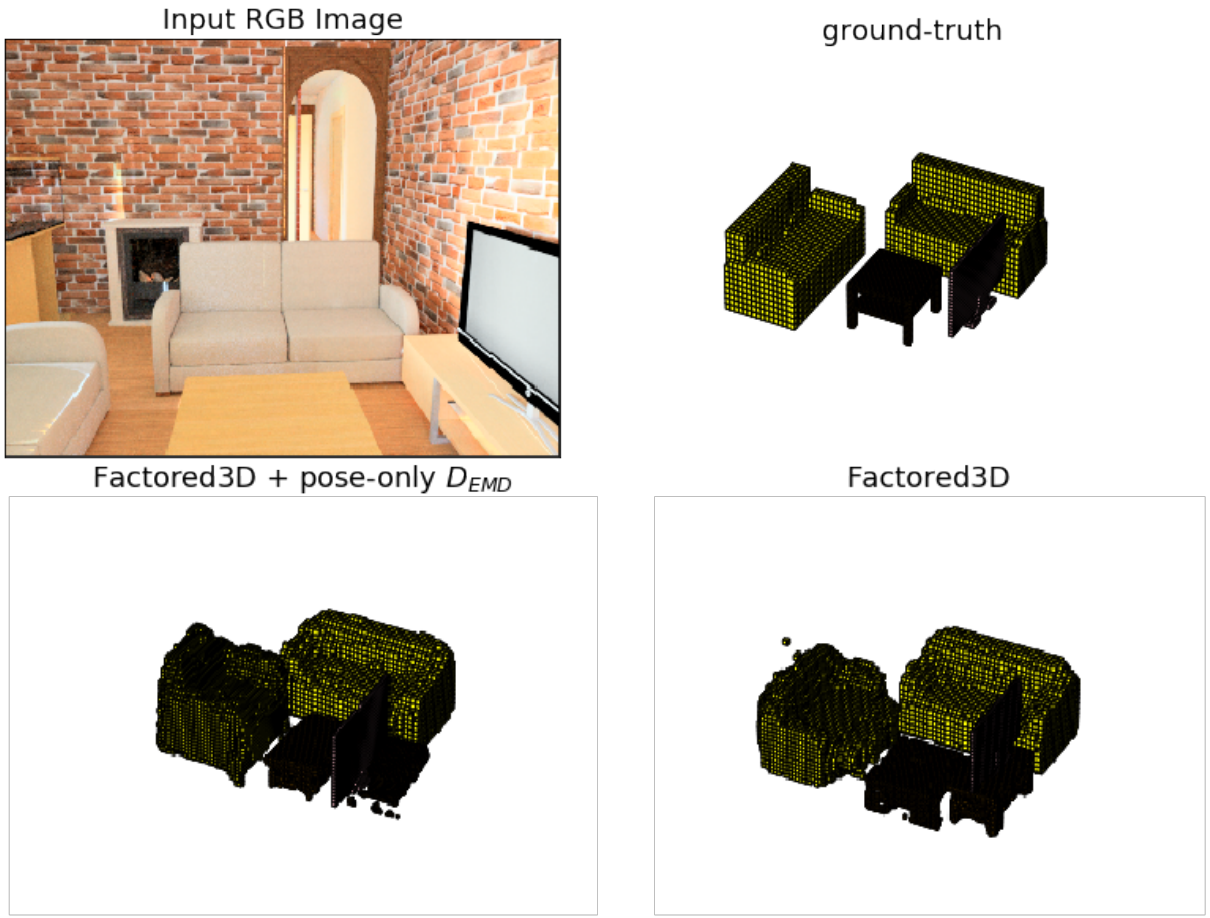

Figure 8a Comparison of the object prediction of the original Factored3D network to the extension using the point cloud based D_{EMD} as loss function on the rotated shape prediction.

Qualitative Comparison - Scene Level III

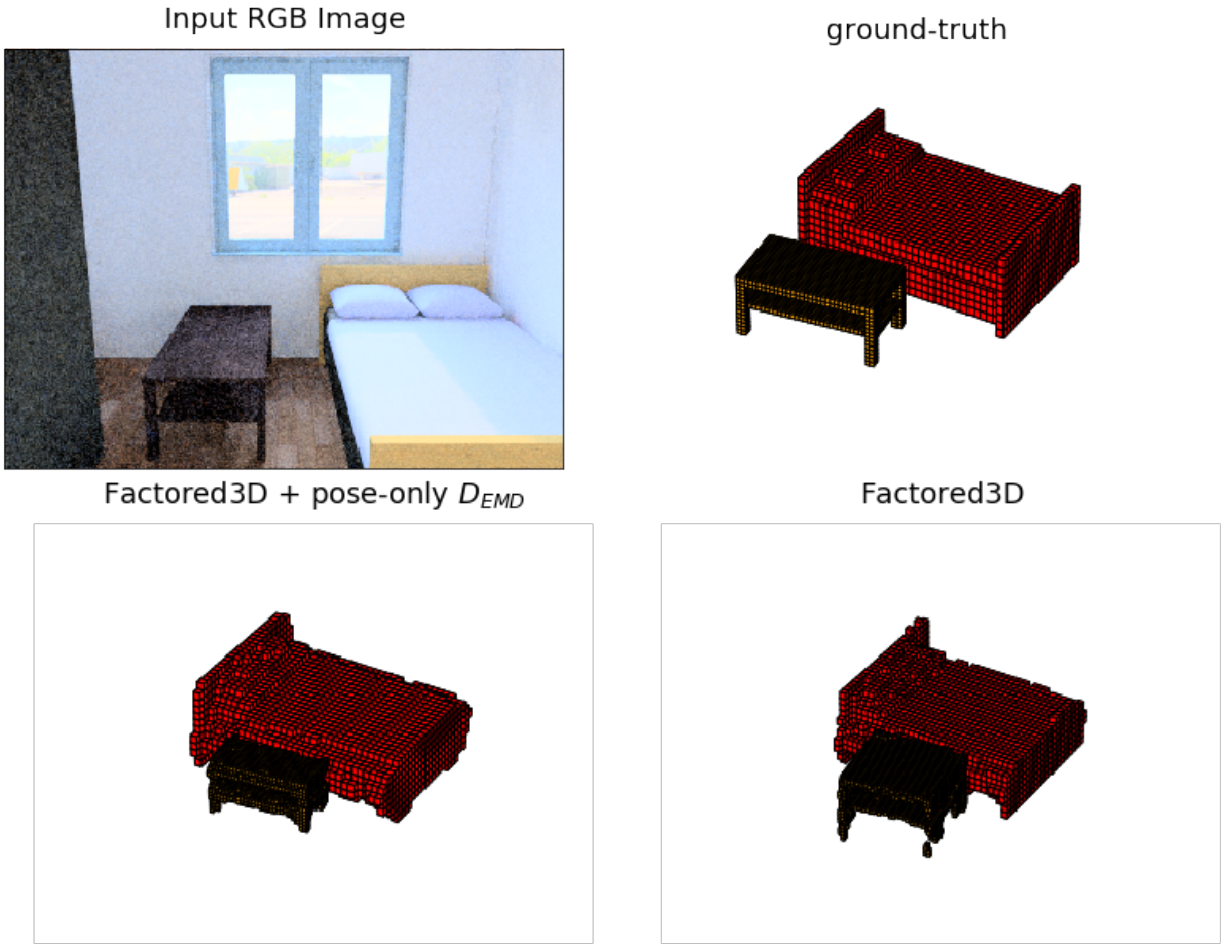

Figure 8b Comparison of the object prediction of the original Factored3D network to the extension using the point cloud based D_{EMD} as loss function on the rotated shape prediction.

Qualitative Comparison - Scene Level IV

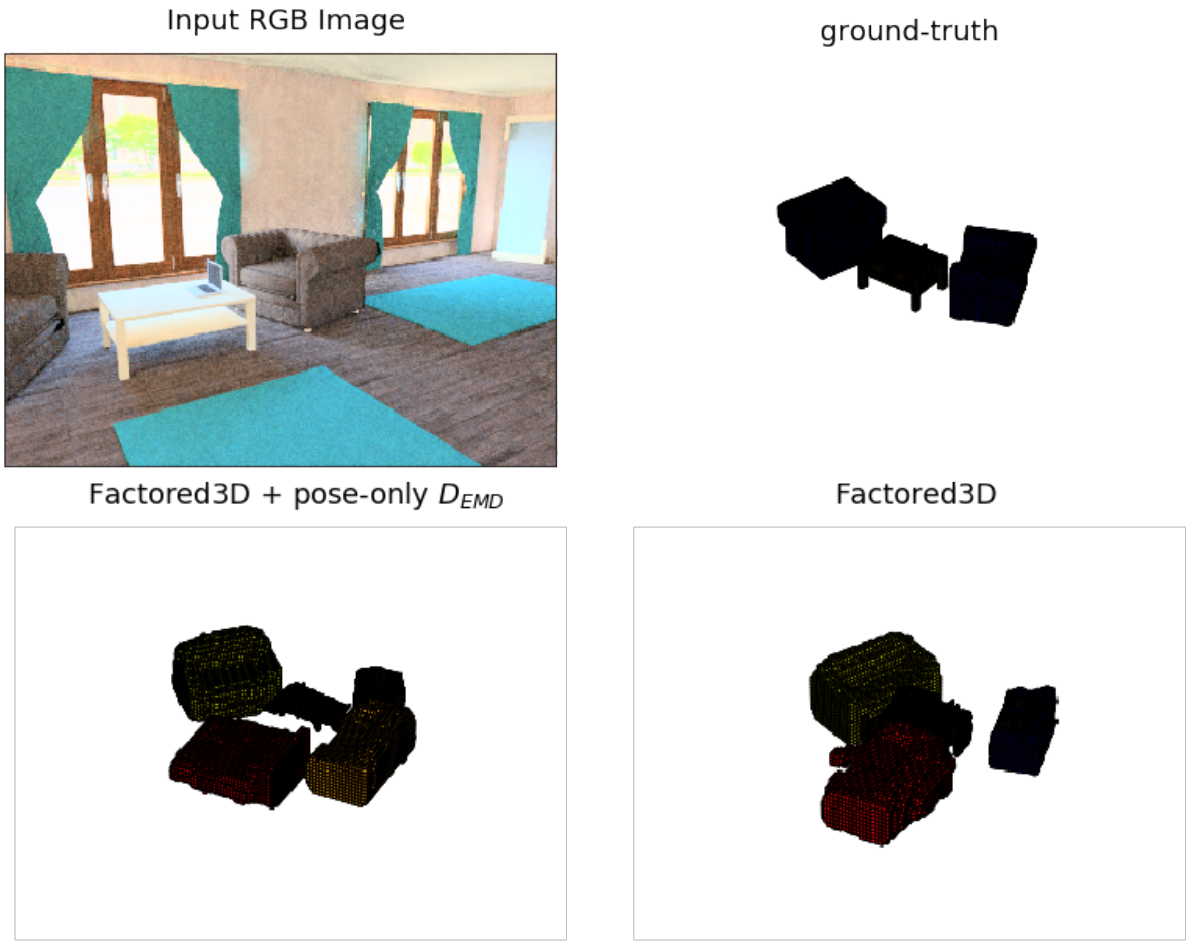

Figure 8c Comparison of the object prediction of the original Factored3D network to the extension using the point cloud based D_{EMD} as loss function on the rotated shape prediction.

Conclusions

Conclusions

- Factored3D does not improve by fine-tuning it solely with a point cloud based loss function.
- Applying a point cloud loss on the shape and rotation only marginally changes the performance to the original Factored3D network.
- The coarse shape prediction bottlenecks the point cloud based approaches from improving the pose estimation.

Bibliography I

[1] Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A. Efros, and Jitendra Malik.

Factoring shape, pose, and layout from the 2d image of a 3d scene.

In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser.

Semantic scene completion from a single depth image.

Proceedings of 29th IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[3] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser.

Physically-based rendering for indoor scene understanding using convolutional neural networks.

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[4] C. Lawrence Zitnick and Piotr Dollár.

Edge boxes: Locating object proposals from edges.

In *ECCV*, 2014.

Bibliography II

[5] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta.

Learning a predictable and generative vector representation for objects.

In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 484–499, 2016.

Thank you for your attention.

Backup

Related Work - TL-Embedding Network

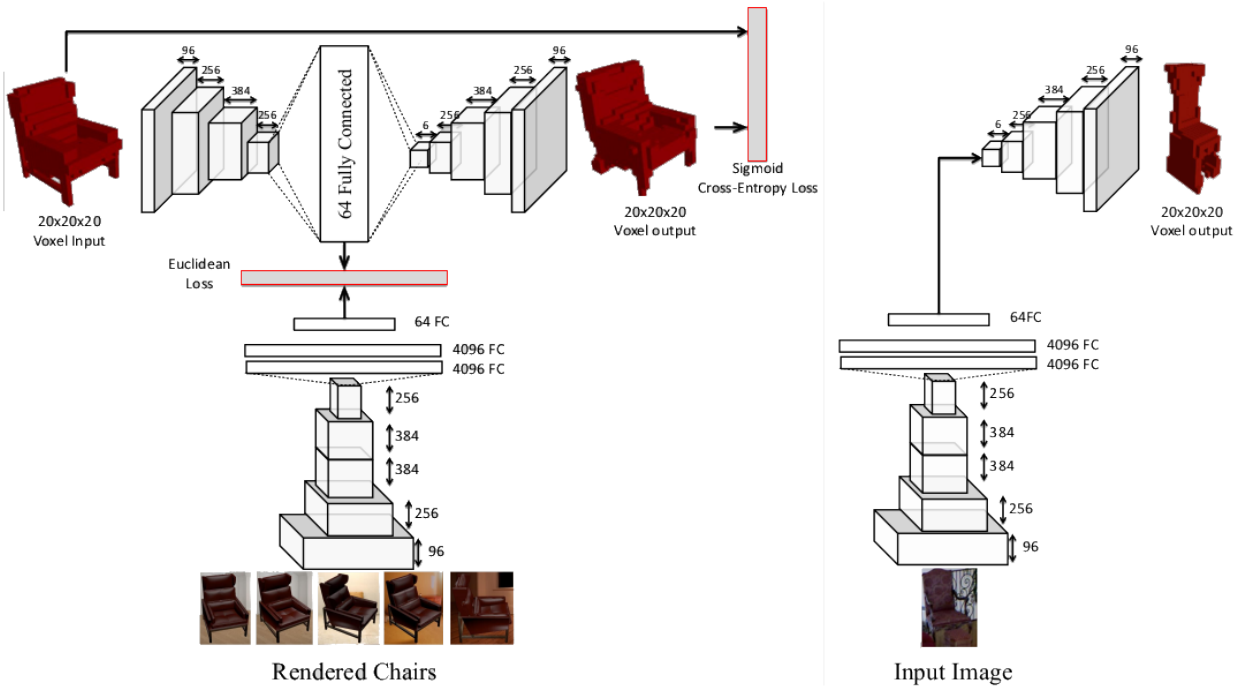

Figure 4 TL-embedding network architecture during training time and test time [5].

- Factored3D incorporates this approach to learn a latent representation good for 3D shape prediction from 2D image input
- Shape encoder and decoder weights are pretrained on the 3D voxel representation of all valid objects in the SUNCG data set.