From Points to Multi-Object 3D Reconstruction

Francis Engelmann[1] Konstantinos Rematas[2] Bastian Leibe[1] Vittorio Ferrari[2]

[1]RWTH Aachen University, Germany [2]Google Research, Zurich

https://francisengelmann.github.io/points2objects/

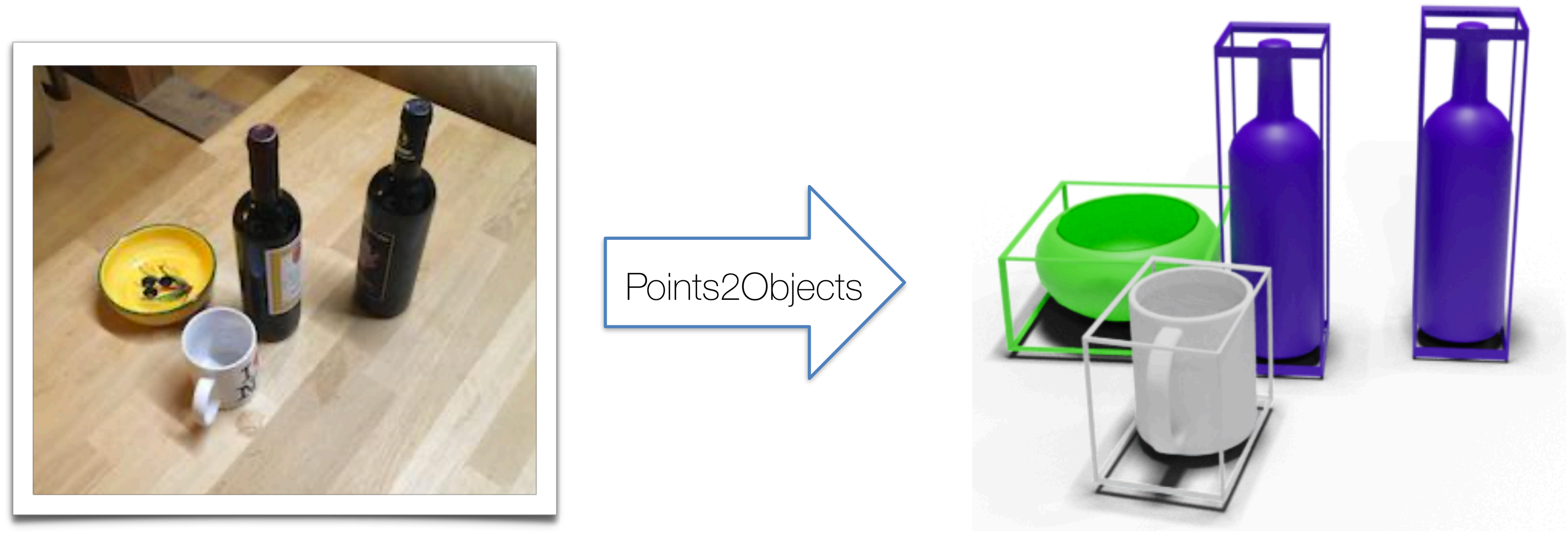
DeeViSe ERC-2017-CoG-773161

CVPR VIRTUAL JUNE 19-25

Abstract

We propose a method to detect and reconstruct multiple 3D objects from a single RGB image. The key idea is to optimize for detection, alignment and shape jointly over all objects in the RGB image, while focusing on realistic and physically plausible reconstructions. To this end, we propose a key-point detector that localizes objects as center points and directly predicts all multi-object properties, including 9-DoF bounding boxes and 3D shapes — all in a single forward pass.

Task



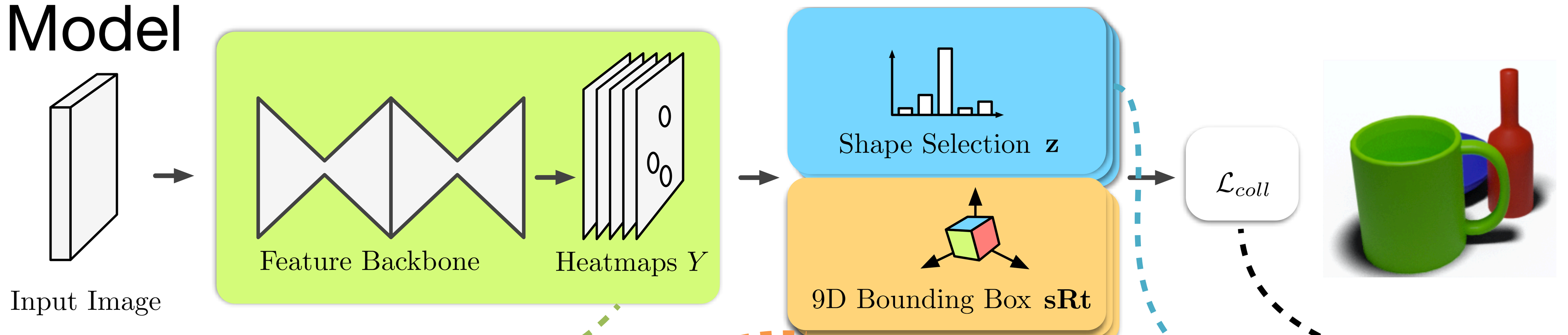
Input: Single Image

Output: 3D Reconstructions of Multiple Objects

Contributions

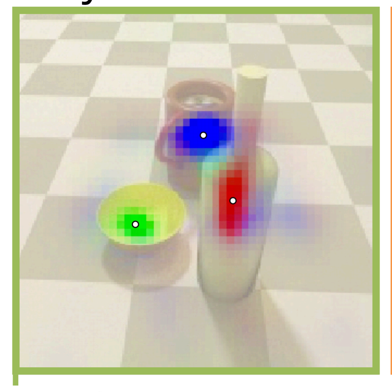
- Fully holistic multi-object 3D scene reconstruction based on CenterNet [1] in a single-stage network from a single input RGB image.
- Our reconstruction is formulated as a shape-selection problem (1-of-K classification) implemented using our novel "*soft target labels*" relying on geometric similarities between exemplar 3D shapes.
- Our collision loss encourages non-intersecting reconstructions and CAD representations guarantee physically plausible and realistic shapes.
- We present a 9-DoF pose estimation study showing that jointly optimizing rotation and translation improves over individual optimization in our setup.

Model



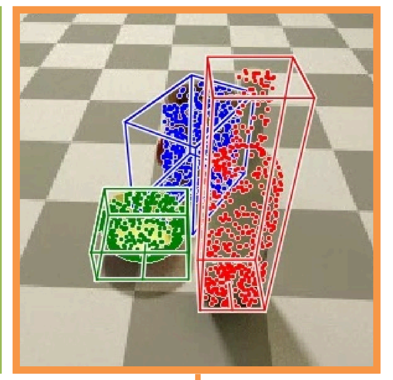
Losses

KeyPoint Loss



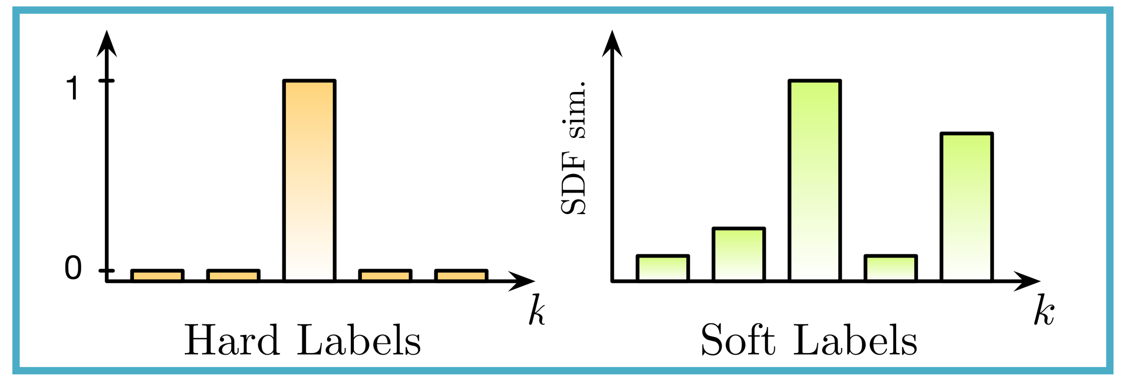
$\mathcal{L}_{key}$ as in CenterNet [1]

3D Box Loss



$$\mathcal{L}_{\mathbf{Rt}} = \sum_{i=1}^{M} \sum_{\mathbf{x}\in\mathcal{P}^i} \| [\mathbf{R}|\mathbf{t}]^i \mathbf{x} - [\hat{\mathbf{R}}|\hat{\mathbf{t}}]^i \mathbf{x} \|_2^2$$

Shape-Selection Loss

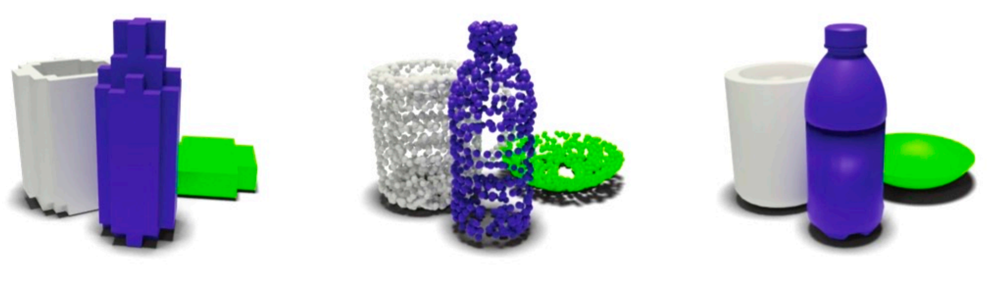


$$\mathcal{L}'_{\mathbf{z}} = \frac{1}{M} \sum_{i=1}^{M} CE(\mathbf{z}^i, \sigma(\hat{\mathbf{z}}^i))$$

$$\mathcal{L}_{\mathbf{z}} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} d(i,k) \cdot \log(S(\hat{z}_k))$$

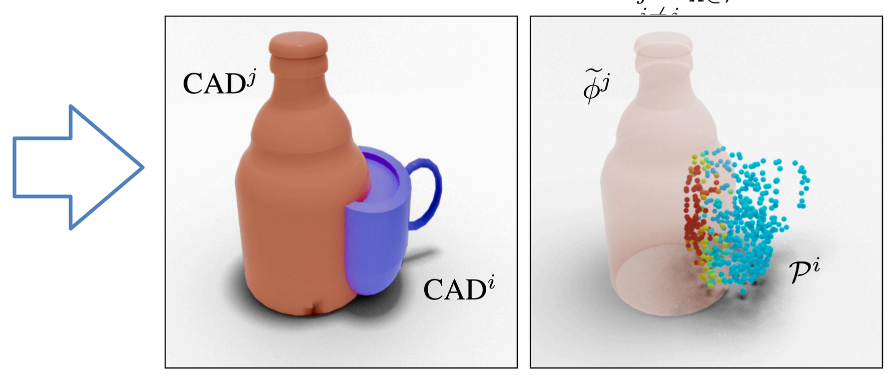
$$d(i,k) = [1 - \|\phi^i - \phi^k\|_2]_+$$

Representation-agnostic Shapes



Collision Loss

$$\mathcal{L}^i_{coll} = \sum_{j=1}^{M} \sum_{\mathbf{x}\in\mathcal{P}^i} \widetilde{\phi}^j(\mathbf{T}^{ij}\mathbf{x})$$



Evaluation

Estimating 9-DoF Poses - Study

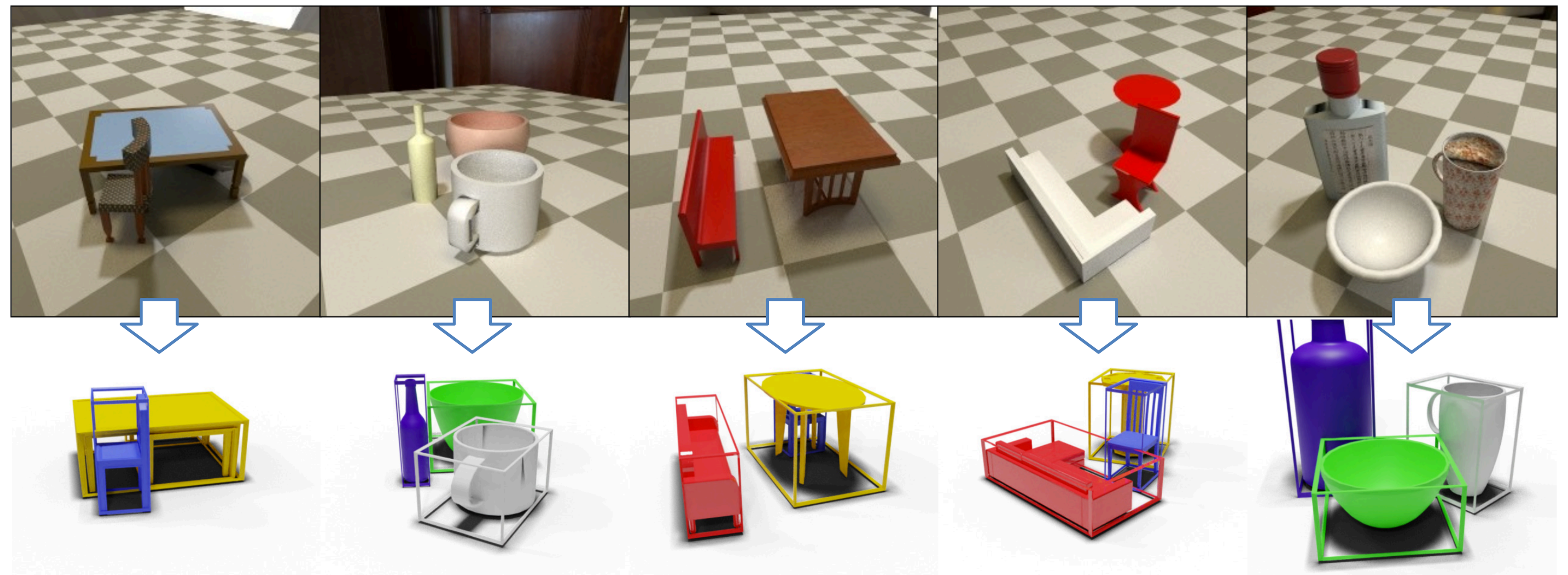
9-DoF Bounding Box	3D mAP: @ 0.5	@ 0.25
$\mathcal{L}_{\text{bin}\mathbf{R}} + \mathcal{L}_{\text{off}\mathbf{R}} + \mathcal{L}_{\mathbf{t}}$ (as in [1])	43.3	75.0
$\mathcal{L}_{\mathbf{M}} + \mathcal{L}_{\mathbf{t}}$ (directly regress rotation matrix M)	44.8	77.0
$\mathcal{L}_{\mathbf{R}} + \mathcal{L}_{\mathbf{t}}$ (add SVD for orthogonal rotation R [5])	46.8	77.2
$\mathcal{L}_{\mathbf{Rt}}$ (ours)	**48.6**	**77.2**

Effect of collision loss	Num. Collisions
Without collision loss	4116
With collision loss	**1627** (−60.5%)

Shape Estimation: Hard vs. Soft Labels

Shape Estimation	Abs. 3D IoU: mean	global
$\mathcal{L}'_{\mathbf{z}}$ Hard-Labels (as in [2])	32.2	40.3
$\mathcal{L}_{\mathbf{z}}$ Soft-Labels (ours)	**36.4**	**44.7**

Results on synthetic images



Results on real images

Casual photos from mobile phone

Generalization from synthetic to real data

Pix3D [4] Single object dataset

Splits	S_1	S_2
CoReNet [3]	33.3%	23.6%
Points2Objects (ours)	**34.1%**	**26.3%**

References

[1] Xingyi Zhou, Dequan Wang, Philipp Krähenbühl. "Objects as Points" ArXiv 2019.

[2] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, Thomas Brox "What Do Single-view 3D Reconstruction Networks Learn?" In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] Stefan Popov, Pablo Bauszat, Vittorio Ferrari "CoReNet: Coherent 3D Scene Reconstruction from a Single RGB Image" In IEEE European Conference on Computer Vision (ECCV), 2020.

[4] Xingyuan Sun*, Jiajun Wu*, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman "Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling" In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[5] J. Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, A. Kanazawa, Afshin Rostamizadeh, and A. Makadia. "An Analysis of SVD for Deep Rotation Estimation". In Neural Information Processing Systems (NeurIPS), 2020.