Supplement - Convex Decomposition of Indoor Scenes

Vaibhav Vavilala
UIUC

David Forsyth
UIUC

1. Additional Ablations

In this supplement, we present additional ablations. We place quantitative evaluation in Table 2. First we examine the effects of adjusting the number of parts at training time, $n_{start} \in [12, 24, 36, 48, 60]$, finding that more parts does not necessarily translate to better quality due to bias-variance tradeoff issues (Fig. 10). We also examine how our Manhattan World losses affect the quality of our primitives, finding that they help error metrics across the board. We show in Fig. 9 how the Manhattan World losses help orient the convexes in manner consistent with scene layout. Finally, we examine training and refining without our segmentation loss (entropy on the segmentation labels). Quantitatively, adding this entropy loss had an approximately neutral effect on our error metrics. We think that that this is due to the complexity of our segmentation labels and inability of our procedure to manage larger numbers of convexes (see Fig. 11).

We show a 3D mesh of our primitives from multiple views in Fig. 12. We note that our primitives are represented in normalized coordinates and we preserve scale/shift coefficients in the X/Y/Z directions to raytrace our primitives from the original viewpoint and obtain a depth map in the original camera frame.

In Table 3, we quantitatively compare our method against the most similar work, [4], using their error metric. Our AUC's are better across the authors' reported range $5 - 50cm$, but our mean is worse. That would indicate the presence of a few outlier test scenes that our method performed poorly on.

In Table 4, we evaluate our method, removing one loss at a time. While the guidance loss had a marginal effect, the remaining losses meaningfully improved the error metrics (row C).

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021.

[2] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:2108.10528*, 2021.

[3] Hao Jiang. Finding approximate convex shapes in rgbd images. In *European Conference on Computer Vision*, pages 582–596. Springer, 2014.

[4] Florian Kluger, Hanno Ackermann, Eric Brachmann, Michael Ying Yang, and Bodo Rosenhahn. Cuboids revisited: Learning robust 3d shape fitting to single rgb images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

Cfg.	ManWorld	Entropy	Split	$Depth_{GT}$	Seg_{GT}	n_{init}	n_{used}	AbsRel↓	RMSE↓	Mean↓	Median↓	11.25° ↑	22.5° ↑	30° ↑	Seg_{Acc} ↑
12-NO	✓	✓	×	×	×	12	6.5	0.185	0.735	40.052	36.009	0.122	0.303	0.418	0.500
24-NO	✓	✓	×	×	×	24	14.4	0.144	**0.603**	**38.235**	**33.621**	**0.133**	**0.335**	**0.451**	0.615
24-NS	✓	✓	✓	×	×	24	25.9	0.183	0.851	41.512	37.445	0.107	0.286	0.396	**0.630**
24-NM	×	✓	×	×	×	24	9.5	0.189	0.716	43.003	38.914	0.059	0.223	0.355	0.536
24-NE	✓	×	×	×	×	24	16.1	**0.143**	0.614	38.311	33.872	0.132	0.330	0.445	0.629
36-NO	✓	✓	×	×	×	36	12.6	0.198	0.781	45.546	41.796	0.062	0.201	0.317	0.565
48-NO	✓	✓	×	×	×	48	14.6	0.199	0.833	43.963	39.802	0.074	0.267	0.370	0.584
60-NO	✓	✓	×	×	×	60	16.9	0.175	0.665	42.387	38.418	0.092	0.269	0.383	0.600
12-O	✓	✓	×	✓	✓	12	6.2	0.153	0.581	40.246	36.475	0.120	0.297	0.416	0.492
24-O	✓	✓	×	✓	✓	24	13.9	0.098	*0.514*	*37.355*	*32.395*	*0.144*	*0.353*	*0.469*	0.618
24-OS	✓	✓	✓	✓	✓	24	27.4	0.143	0.819	41.235	36.839	0.111	0.295	0.407	*0.631*
24-OM	×	✓	×	✓	✓	24	9.8	0.151	0.580	43.349	39.117	0.057	0.218	0.349	0.568
24-OE	✓	×	×	✓	✓	24	15.7	*0.096*	0.520	37.513	32.700	0.143	0.347	0.464	0.630
36-O	✓	✓	×	✓	✓	36	13.0	0.165	0.629	44.189	40.012	0.066	0.212	0.335	0.577
48-O	✓	✓	×	✓	✓	48	14.8	0.159	0.657	44.219	40.025	0.069	0.255	0.362	0.597
60-O	✓	✓	×	✓	✓	60	17.4	0.172	0.652	43.771	39.620	0.089	0.258	0.369	0.605
ref	-	-	-	-	-	-	-	0.110	0.357	14.9	7.5	0.622	0.793	0.852	0.719

Table 2. Additional ablations. All configs have pruning and polishing applied at the refinement stage. The eight experiments between 12-NO through 60-NO refer to not having an oracle available during refinement (i.e. the depth and segmentation are inferred by [5, 2]). The next eight configs have oracles available. The last row ref. shows the reference error of our pretrained depth estimation network, a recent normal estimation work, and our pretrained segmentation network [5, 1, 2]. By testing the number of starting parts at training time between [12-NO,24-NO,36-NO,48-NO,60-NO] as well as [12-O,24-O,36-O,48-O,60-O] and then applying refinement, 24 parts performs best. This is a classic case of bias-variance tradeoff: too few parts biases the decomposition with insufficient capacity; too many parts results in variance problems. We also evaluate the effects of our Manhattan World losses (24-NO vs. 24-NM) and (24-O vs. 24-OM); quantitatively, these harm our error metrics across the board. Clearly, indoor scenes have objects oriented in a similar way, and enforcing that as a loss improves the quality of our decompositions. From there, we examine the effects of no entropy loss (i.e. segmentation loss) during training nor inference, comparing experiments (24-NO vs. 24-NE) and (24-O vs. 24-OE). The depth and normal error metrics are nearly identical with or without this loss, and the segmentation accuracy is slightly better WITHOUT the segmentation loss. This was slightly unexpected, but could be due to the overall noisiness and complexity of the segmentation maps preventing clean segmentation of objects. We illustrate an example in Fig. 11. Also note how the pruning process removed more parts with the segmentation loss, and it is generally to be expected that more parts can lead to better segmentation accuracy. Finally, we examine the effects of splitting each convex into 8 equal volume pieces during refinement before applying refinement/pruning. These are examined quantitatively in configs (24-NO vs. 24-NS) and (24-O vs. 24-OS). Overall, we achieved our best segmentation score with splitting, to be expected due to the increased number of parts. However, depth and normal error metrics suffer. This is due to optimization difficulties - the optimizer struggled to improve the fit of so many parts. Qualitatively, the pruning process resulted in holes in the representation as shown in Fig. 9. We think that additional investigation into splitting and pruning can lead to near-arbitrary resolution convex decompositions, an exciting next step.

	$AUC_{@50cm}$	$AUC_{@20cm}$	$AUC_{@10cm}$	$AUC_{@5cm}$	$mean_{cm}$	$median_{cm}$
Ours - RGB	**77.3**	**47.6**	**26.8**	**13.9**	40.2	26.2
Kluger *et al.*	57.0	33.1	18.9	10.0	**34.5**	-
Ours - Depth	**86.9**	**72.5**	**56.5**	**38.2**	26.6	10.1
Kluger *et al.*	77.2	62.7	49.1	34.3	**20.8**	-

Table 3. Comparison with previous work [4] - Occlusion-Aware distance metric reported in all columns. Our AUC's are better, but mean is worse; medians indicate we suffer because of outliers.

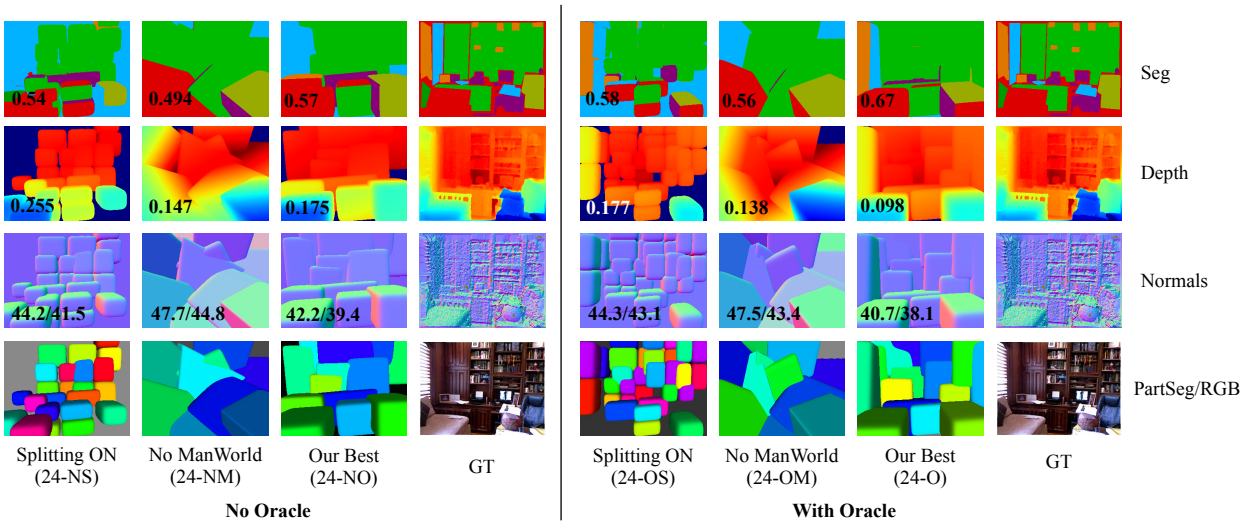

Figure 9. Qualitative ablation study on the effects of splitting during refinement and our Manhattan World losses. Results shown with and without ground truth depth and segmentation during refinement. Corresponding quantitative evaluation in Table 2. **Convex splitting** Applying convex splitting during the refinement step increases the granularity of our representation; however our optimization procedure struggles to improve the fit and convex pruning results in holes (comparing 24-NS against 24-NO and 24-OS against 24-O). **Manhattan world losses** Removing the Manhattan World losses during training and refinement results in qualitatively cluttered representations. The depth is approximately correct, but the convexes are not organized in a manner representing the scene (e.g. maintaining parallel lines where possible). Our decompositions are quantitatively and qualitatively worse with the Manhattan World losses removed (comparing 24-NM against 24-NO and 24-OM against 24-O).

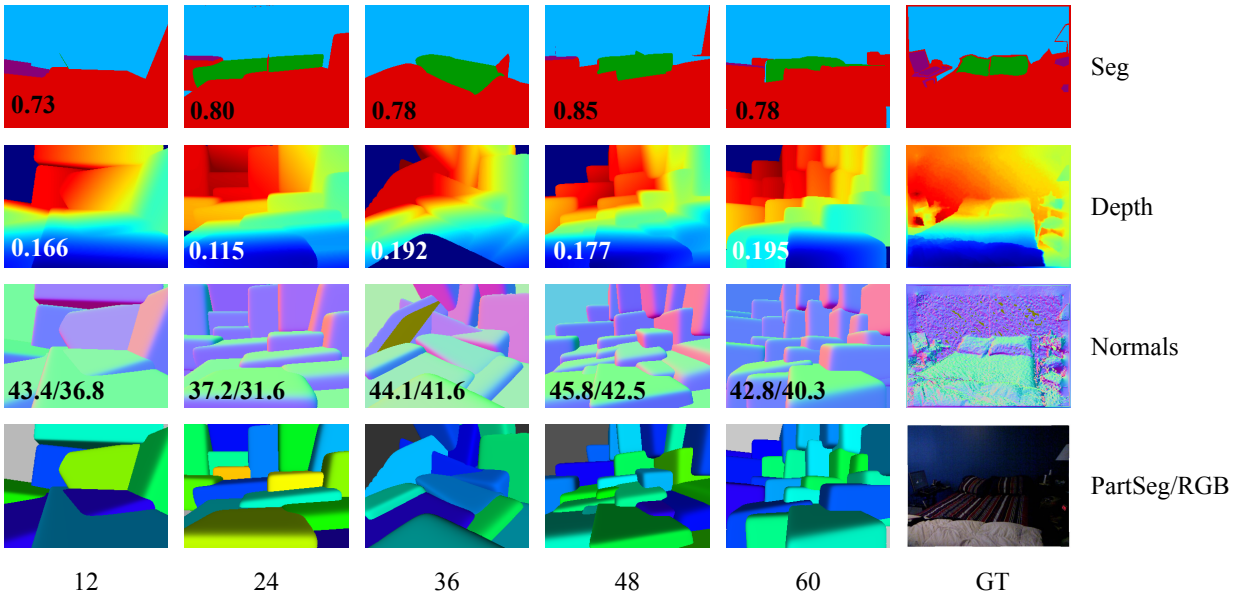

Figure 10. Ablation study on number of parts at training time. We examine training with $n_{parts} \in [12, 24, 36, 48, 60]$ on a random NYUv2 test image. We refine without GT depth/seg. The optimal number of parts is this experiment was 24. Notice how less than this value, we run into bias issues, and above this value, we see variance problems. Quantitative results shown in Table 2.

Figure 11. Four example ground truth segmentation maps of NYUv2. The entropy loss encouraging our primitive decomposition to roughly obey segmentation boundaries quantitatively showed neutral results on this dataset - though it helped in very simple toy models we tried. Observe how the segmentations are quite cluttered due to occlusions or objects on top of one another. Say - posters on the fridge/cabinet or objects on a desk. The signal to the entropy loss can confuse the part decomposition in areas we want just one primitive representing that area. An obvious approach to dealing with scene complexity is more primitives. However, training with more primitives resulted in significant variance issues and a drop in quality beyond 24 as Table 2 shows. Splitting the parts during refinement followed by polishing also failed to yield helpful results due to optimization difficulty. Finally, one could tackle this from the perspective of data: we could pre-process the segmentation with the intent to simplify them (similar to [3]).

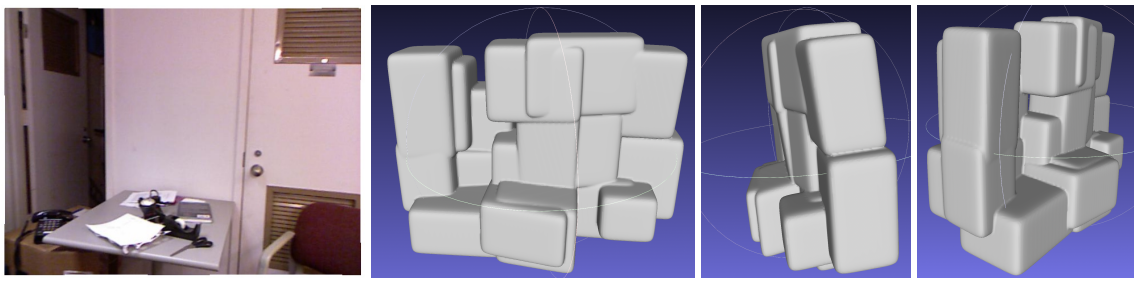

Figure 12. Input image, front of primitives, right side, left side.

Cfg.	n_{used}	AbsRel↓	RMSE↓	Mean↓	Median↓	$11.25^\circ \uparrow$	$22.5^\circ \uparrow$	$30^\circ \uparrow$	$Seg_{Acc} \uparrow$
no-unique	16.7	0.157	0.695	38.092	34.501	**0.153**	0.328	0.437	0.613
no-guidance	14.4	0.145	0.624	**37.422**	33.471	0.149	**0.339**	0.453	0.611
overlap-10	17.3	0.148	0.646	38.567	34.711	0.149	0.325	0.435	**0.630**
no-overlap	4.3	0.179	0.673	38.457	34.431	0.107	0.300	0.431	0.466
no-volume	16.2	**0.144**	0.618	37.462	**33.292**	0.141	0.338	**0.454**	0.623
no-local	1.0	0.499	1.856	87.673	82.785	0.006	0.019	0.034	0.212
C	14.4	**0.144**	**0.603**	38.235	33.621	0.133	0.335	0.451	0.615

Table 4. Ablations with one CVXnet loss removed at a time. All are non-oracle. C is our final non-oracle model (Table 1 main text). Depth metrics worse without unique param. loss (eqn. 5 of CVXnet). Guidance loss (eqn. 6) has an approx. neutral effect. Removing the overlap loss (decomp. loss eqn. 4 in CVXnet) harms quality; scaling it up to 10, beyond the default 0.1, slightly hurts error metrics, likely because convexes need to move freely during the training process. It remains future work to completely eliminate overlaps while preserving quality. Removing the volume loss has an approx. neutral effect on the error metrics though parsimony is harmed. The localization loss (eqn. 7) is critical to the method.