Technical University of Munich TUM

Computer Aided Medical Procedures

Google

ICRA LONDON · 2023

Paper

GitHub

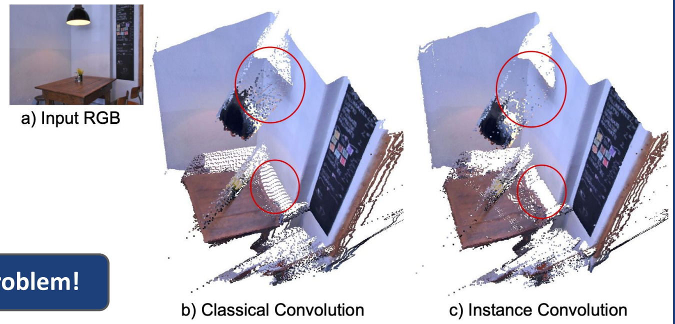
Object-aware Monocular Depth Prediction with Instance Convolutions

Enis Simsar*, Evin Pınar Örnek*, Fabian Manhardt, Helisa Dhamo, Nassir Navab, Federico Tombari

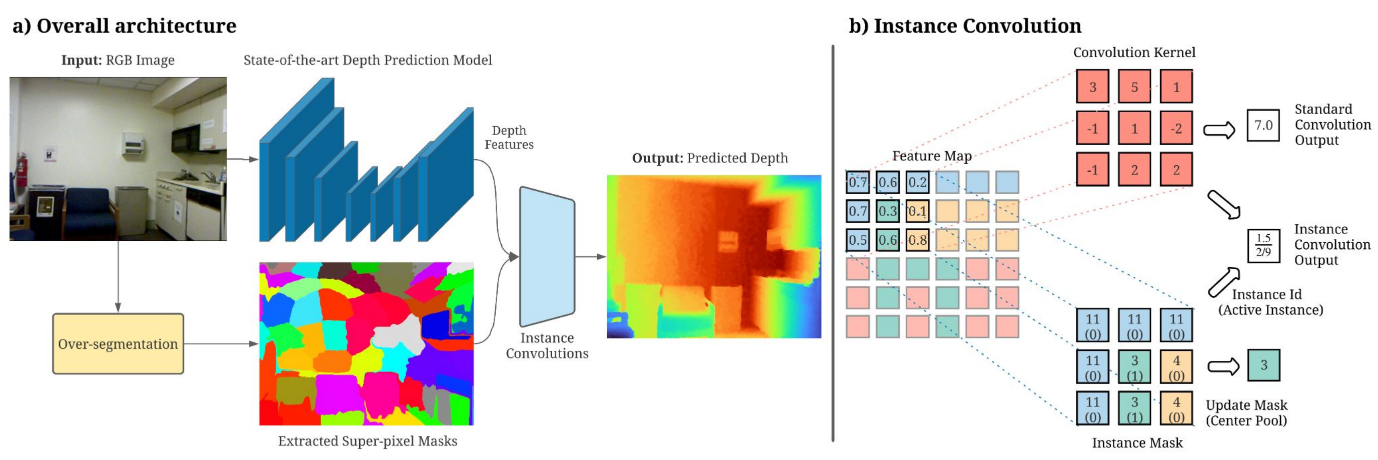
TLDR

- Monocular depth prediction performs poorly on local geometric details (planar surfaces, object boundaries)
- This is often overlooked because not directly visible in 2D depth maps
- Occlusion boundaries are very important for robotic grasping and navigation

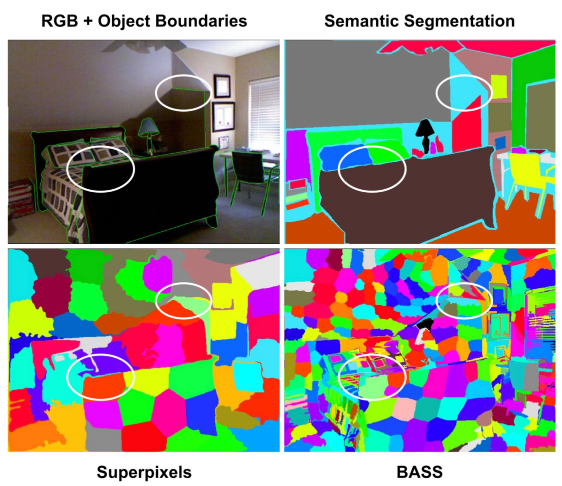
We propose an object-aware MDP method to solve this problem!



Approach - Instance Convolution

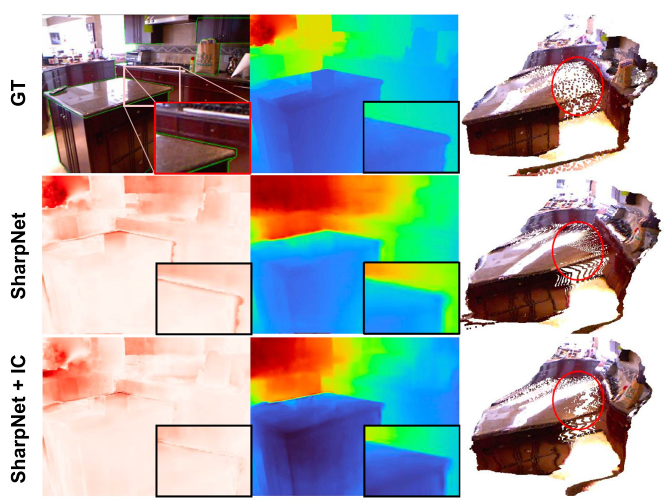


Why superpixels?



- Semantic segmentation does not consider intra object discontinuities (highlighted in white-circles).
- Thus, we leverage super-pixels to account for any discontinuities based on the RGB input.

Error maps



- To avoid aggregation of features appertaining to different image layers, we thus propose to leverage superpixels in an effort to guide convolution operator.

Metrics & Loss functions

Absolute Relative Difference (Abs. Rel.)

$$\frac{1}{|N|}\sum_{i\in N}\frac{|d_i - d_i^*|}{d_i^*}$$

Accuracies

$$\max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = \delta < threshold$$

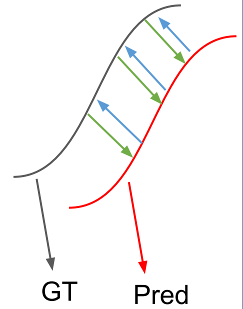
$\delta_i < 1.25^i$

DBE Accuracy

$$\epsilon_{acc} = \frac{1}{\sum_i \hat{Y}_i}\sum_i E_i \cdot \hat{Y}_i$$

DBE Completeness

$$\epsilon_{com} = \frac{1}{\sum_i Y_i}\sum_i \hat{E}_i \cdot Y_i$$



$$L_1(d, d^{GT}) = \frac{1}{N}\sum_{i=1}^{N}|d_i^{GT} - d_i|$$

$$L_{normal}(n, n^{GT}) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\langle n_i, n_i^{GT}\rangle}{||n_i|| \cdot ||n_i^{GT}||}\right)$$

$$L_{grad}(d, d^{GT}) = \frac{1}{N}\sum_{i=1}^{N}|\nabla_h d_i - \nabla_h d_i^{GT}| + |\nabla_v d_i - \nabla_v d_i^{GT}|$$

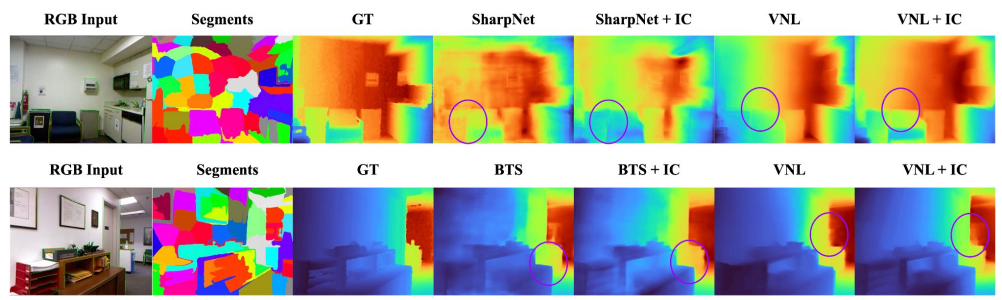
$$L = L_1 + L_{grad} + L_{normal}$$

Comparison on iBims

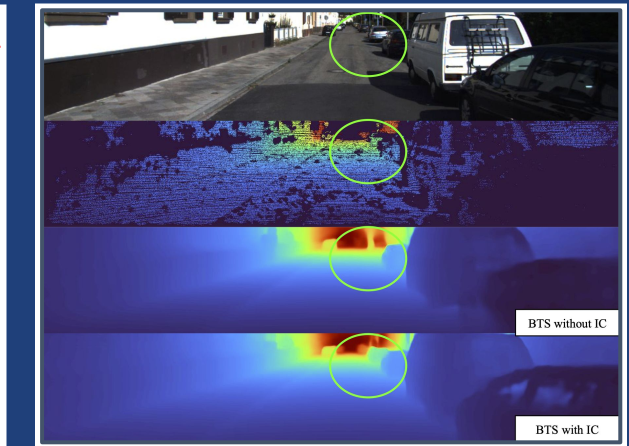
Method	Error ↓			Accuracy ↑			PE (in $cm/°$) ↓		DBE (in px) ↓		DDE (in %)		
	$absrel$	log_{10}	$rmse$	δ_1	δ_2	δ_3	ϵ^{plan}	ϵ^{orie}	ϵ^{acc}	ϵ^{comp}	ϵ^0 ↑	ϵ^- ↓	ϵ^+ ↓
Eigen [21]	0.32	0.17	1.55	0.36	0.65	0.84	7.70	24.91	9.97	9.99	70.37	27.42	2.22
Laina [23]	0.26	0.13	1.20	0.50	0.78	0.91	6.46	19.13	6.19	9.17	81.02	17.01	1.97
Liu [50]	0.30	0.13	1.26	0.48	0.78	0.91	8.45	28.69	2.42	7.11	79.70	14.16	6.14
Li [52]	0.22	0.11	1.09	0.58	0.85	0.94	7.82	22.20	3.90	8.17	83.71	13.20	3.09
Liu [53]	0.29	0.17	1.45	0.41	0.70	0.86	7.26	17.24	4.84	8.86	71.24	28.36	0.40
SharpNet [25]	0.26	0.11	1.07	0.59	0.84	0.94	9.95	25.67	3.52	7.61	84.03	9.48	6.49
with Instance Conv.	0.29	0.12	1.14	0.55	0.82	0.92	9.83	25.88	3.11	7.83	81.84	8.27	9.88
BTS [27]	0.24	0.12	1.08	0.53	0.84	0.94	7.24	20.51	2.50	5.81	82.24	15.50	2.27
with Instance Conv.	0.22	0.11	1.11	0.57	0.86	0.94	6.76	19.39	3.71	8.01	84.04	13.3	2.67
VNL [13]	0.24	0.11	1.06	0.54	0.84	0.93	5.73	16.91	3.64	7.06	82.72	13.91	3.36
with Instance Conv.	0.23	0.10	1.06	0.58	0.85	0.93	5.62	16.53	3.03	7.68	83.85	13.26	2.87

Results on NYU

Method	Error ↓		Accuracy ↑			DBE ↓	
	$absrel$	$rmse$	δ_1	δ_2	δ_3	ϵ_{acc}	ϵ_{com}
SharpNet	0.116	0.448	0.853	0.970	0.993	3.041	8.692
+ Instance Conv.	0.124	0.456	0.847	0.971	0.993	1.961	6.489
VNL	0.112	0.417	0.880	0.975	0.994	1.854	7.188
+ Instance Conv.	0.117	0.425	0.863	0.970	0.991	1.780	6.059
BTS	0.110	0.392	0.885	0.978	0.994	2.090	5.820
+ Instance Conv.	0.121	0.467	0.848	0.964	0.993	1.817	6.197



Qualitatives on KITTI



- Our method also works for outdoor scenes.
- It provides sharper edges for the objects and finds hidden objects (highlighted in green).

Ablation Study

Method	Error ↓		DBE ↓		Runtime	
	$absrel$	$rmse$	ϵ_{acc}	ϵ_{com}	FPS	FPS*
SharpNet [25]	0.12	0.45	3.04	8.69	16.7	16.7
GT Masks	0.12	0.46	2.05	6.49	13.5	13.5
PointRend [55]	0.13	0.45	2.21	6.76	13.5	3.64
BASS [40]	0.12	0.46	2.19	6.63	13.2	0.59
IC 16	0.14	0.47	2.07	6.59	13.5	3.08
IC 32	0.14	0.47	2.09	6.66	13.6	3.04
IC 64	0.12	0.46	1.96	6.48	13.4	2.97
SC 64	0.12	0.45	2.18	6.63	15.2	3.05
IC 128	0.13	0.46	1.92	6.57	13.3	2.89

Ablation study on NYUv2

- comparing usage of different masks (ground truth, PointRend, and BASS)
- super-pixels with Standard Convolutions (SC)
- different # of segments (16-32-64-128) with Instance Convolutions (IC)

Conclusion

- We propose InstanceConv, provides sharp depth values around object boundaries.
- We show comprehensive evaluation on NYU depth v2, iBims, and KITTI, demonstrating the method's effectiveness without compromising the quality in edges and remaining regions.
- InstanceConv can be incorporated into other domains such as semantic segmentation to similarly improve sharpness.