Improving Monocular Depth Estimation by Leveraging Structural Awareness and Complementary Datasets – Supplementary Materials

Tian Chen*, Shijie An*, Yuan Zhang, Chongyang Ma,
Huayan Wang, Xiaoyan Guo, and Wen Zheng

Y-tech, Kuaishou Technology

1 Network Architecture

We show the architecture and parameters of our proposed network in Tab. 1, where $\{\text{rblock}_i\}$ represent residual blocks from ResNet101 [3].

2 More Results on NYUDv2 Test Set

Fig. 1 shows more qualitative results from the test split of NYUDv2 Dataset by comparing our approach with several state-of-the-art monocular depth estimation methods [2,4,1]. As can be seen from this figure, our results present better spatial structure, *i.e.*, sharper boundaries and more faithful ordinal relationship between objects. Fig. 2 demonstrates visualization results of reprojected 3D point clouds using different methods.

3 More Results on HC Depth Test Set

We show the summarized six types of hard cases for monocular depth estimation in Figs. 3 (dark lighting), 4 (portrait), 5 (spurious edges), 6 (reflecting surface), 7 (sky), and 8 (tilted shots), respectively. Each of these figures presents qualitative results of several examples from the test set of our HC Depth dataset.

References

1. Alhashim, I., Wonka, P.: High Quality Monocular Depth Estimation via Transfer Learning. arXiv preprint arXiv:1812.11941 (2018)
2. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)

* Joint first authors

Table 1: Our proposed network architecture. **K**: kernel size; **S**: stride in convolution layers; **in** and **out**: spatial resolution of the input and output; **Input**: input of the layer. conv: convolution layer; upconv: upsample and convolution layer. In the last column, + and U_{1-4} denote layer concatenation.

Layer		**K**	**S**	**Channel**	**in**	**out**	**Input**
Encoder							
E_1	conv	7	2	3/64	H	H/2	**input RGB**
	GCB_1	-	1	64/64	H/4	H/4	conv
E_2	maxpool	3	2	64/64	H/2	H/4	GCB_1
	$rblock_1$	1,3	1	64/256	H/4	H/4	maxpool
	GCB_2	-	1	256/256	H/4	H/4	$rblock_1$
E_3	$rblock_2$	1,3	2	256/512	H/4	H/8	GCB_2
	GCB_3	-	1	512/512	H/8	H/8	$rblock_2$
E_4	$rblock_3$	1,3	2	512/1024	H/8	H/16	GCB_3
	GCB_4	-	1	1024/1024	H/16	H/16	$rblock_3$
E_5	$rblock_4$	1,3	1	1024/2048	H/16	H/32	GCB_4
	GCB_5	-	1	2048/2048	H/32	H/32	$rblock_4$
Decoder							
D_5	$upconv_4$	3	2	2048/1024	H/32	H/16	GCB_5
D_4	SAB_4	1	1	2048/2048	H/16	H/16	$D_5 + GCB_4$
	$conv_4$	3	1	2048/1024	H/16	H/16	SAB_4
	$upconv_3$	3	2	1024/512	H/16	H/8	$conv_4$
D_3	SAB_3	1	1	1024/1024	H/8	H/8	$D_4 + GCB_3$
	$conv_3$	3	1	1024/512	H/8	H/8	SAB_3
	$upconv_2$	3	2	512/256	H/8	H/4	$conv_3$
D_2	SAB_2	1	1	512/512	H/4	H/4	$D_3 + GCB_2$
	$conv_2$	3	1	512/256	H/4	H/4	SAB_2
	$upconv_1$	3	2	256/64	H/4	H/2	$conv_2$
D_1	SAB_1	1	1	128/128	H/2	H/2	$D_2 + GCB_1$
	$conv_1$	3	1	128/64	H/2	H/2	SAB_1
Output							
U_4	$conv_5$	3	1	1024/256	H/16	H/16	$conv_4$
	$upshuffle_3$	-	-	256/1	H/16	H	$conv_5$
U_3	$conv_6$	3	1	512/64	H/8	H/8	$conv_3$
	$upshuffle_2$	-	-	64/1	H/8	H	$conv_6$
U_2	$conv_7$	3	1	256/16	H/4	H/4	$conv_2$
	$upshuffle_1$	-	-	16/1	H/4	H	$conv_7$
U_1	upsample	-	-	64/64	H/2	H	$conv_1$
Out	$conv_0$	3	1	67/32	H	H	U_{1-4}
	depth	3	1	32/1	H	H	$conv_0$

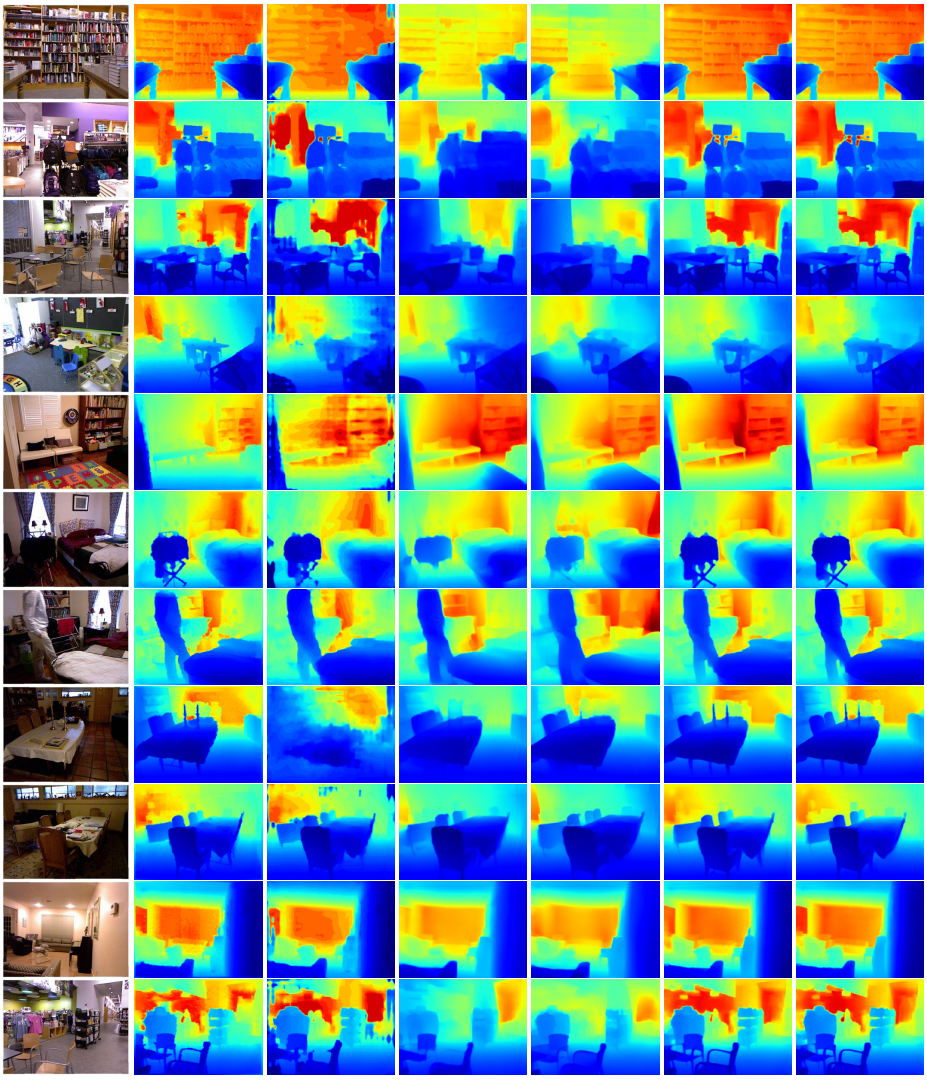

Fig. 1: Qualitative results on NYUDv2. From left to right: input images, ground-truth depth maps, Fu *et al.* [2], Hu *et al.* [4], Alhashim *et al.* [1], our full model trained on NYUDv2, and our full model trained on all the six datasets. All the depth maps are visualized using the same scale.

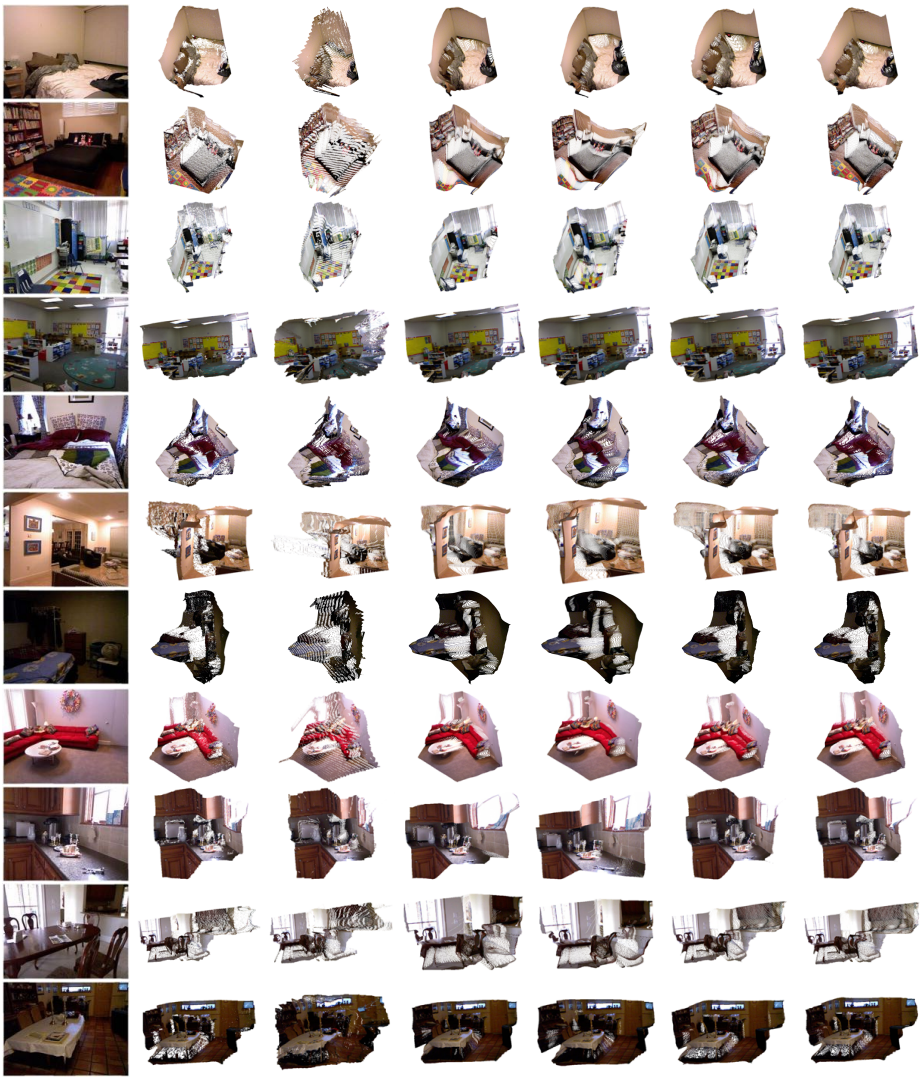

Fig. 2: Qualitative results on NYUDv2. From left to right, we show input images, depth maps visualized as point clouds using Fu *et al.* [2], Hu *et al.* [4], Alhashim *et al.* [1], and our proposed method. The last column is the results of our full model trained on all the six datasets. The 3D point clouds of each example are rendered from the same viewpoint.

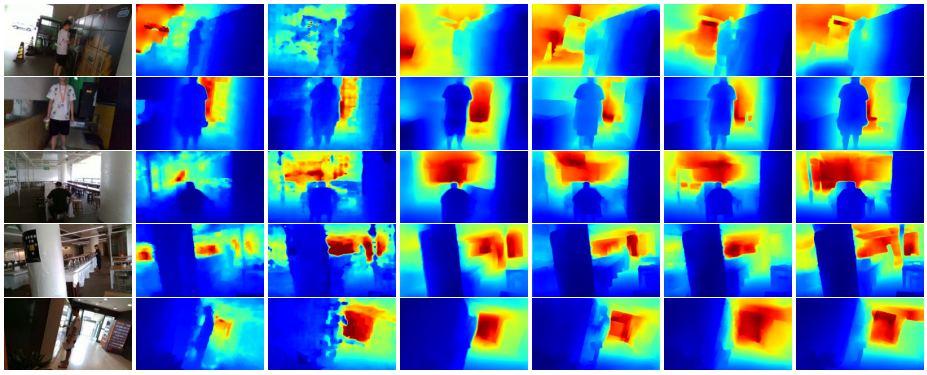

Fig. 3: Qualitative results on the cases of *dark lighting* from HC Depth dataset.

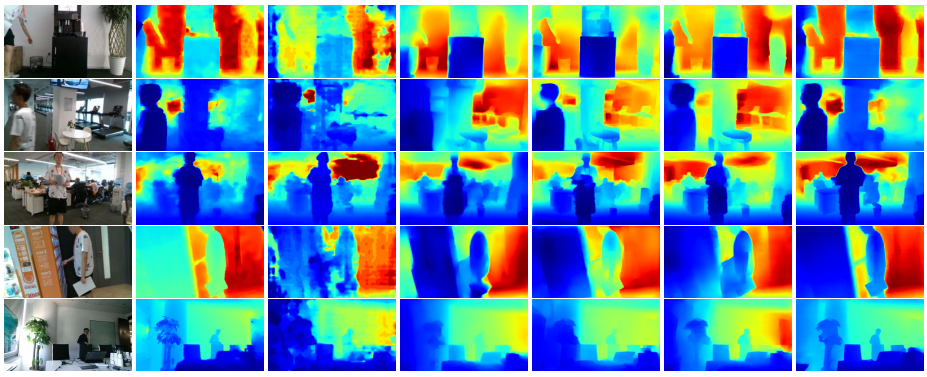

Fig. 4: Qualitative results on the cases of *portrait* from HC Depth dataset.

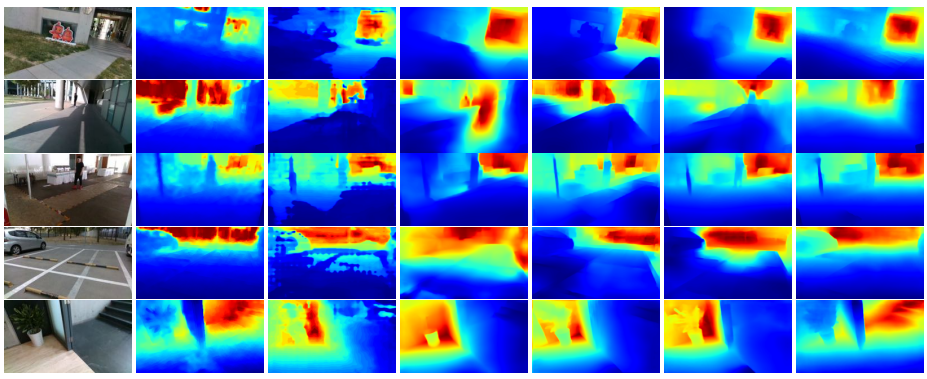

Fig. 5: Qualitative results on the cases of *spurious edges* from HC Depth dataset.

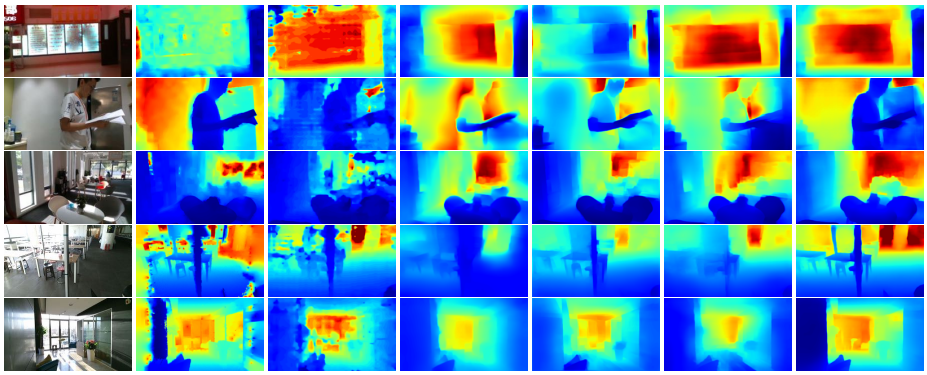

Fig. 6: Qualitative results on the cases of *reflecting surface* from HC Depth dataset.

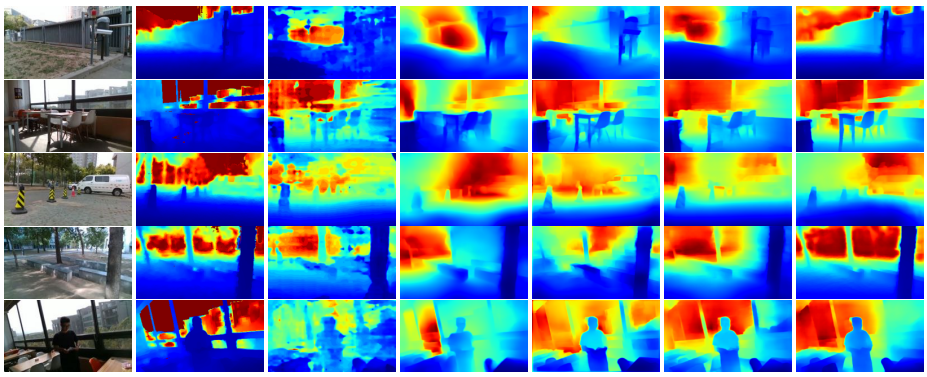

Fig. 7: Qualitative results on the cases of *sky* from HC Depth dataset.

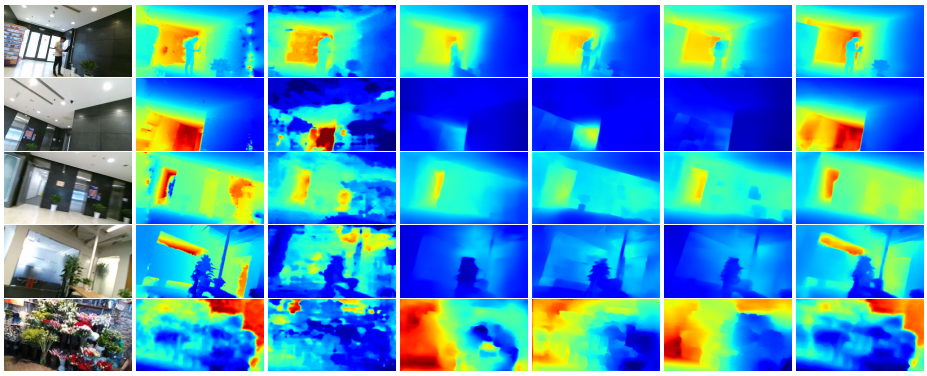

Fig. 8: Qualitative results on the cases of *tilted shot* from HC Depth dataset.

3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
4. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. In: WACV. pp. 1043–1051 (2019)