Supplementary Material: Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong,
Chunjie Zhang, Chunyu Lin, Meiqin Liu, Yao Zhao*
Institute of Information Science, Beijing Jiaotong University
Beijing Key Laboratory of Advanced Information Science and Network, Beijing, 100044, China
{lingzhihe, hongguang, l1feng, hhbai, rmcong, cjzhang, cylin, mqliu, yzhao}@bjtu.edu.cn

1. Network Architectures

As a supplement to the content of Section 4 and Section 5, Table 1 depicts the whole architecture and parameter settings of FDSR, which consists of initial operations, high-frequency guidance branch (HFGB) and multi-scale reconstruction branch (MSRB).

2. Experimental Settings

To evaluate the global and local depth map SR accuracy, we have introduced two quantitative indicators: depth value errros and edge errors. In this part, we will illustrate how we calculate the depth value errors and edge errors.

Dataset split and Training Strategy. As for RGB-D-D dataset, we randomly split 1586 portraits, 380 plants, 249 models for training and 297 portraits, 68 plants, 40 models for testing.What's more, the 430 pairs of lights data are all used to test when we evaluate methods in more challenge scenes. The model for every scaling factor is optimized using the Adam optimizer[3] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a batch size of 1.

Depth Value Errors. As for each value in the depth map, we calculate the percentage of absolute value changes over 10% between ground truth and output. To focus on the foreground subjects, in RGB-D-D dataset, we use binary masks to calculate the value changes in 3 meters. And in NYU v2 [5], we use binary masks to calculate the value changes in 10 meters.

Edge Errors. To calculate the edge errors, we fisrt use Sobel operator to get a fine edge map of RGB image and normalize it to $[0, 255]$. Then, we make edge expansion by using a 3×3 median filter on the edge map. After that, we make binary edge masks from the edge map through the threshold $[10, 255]$. In the area of the binary edge masks, we calculate the percentage of absolute value changes over 1.2% between ground truth and output. Using RGB edges to calculate the edge errors can reflect the edge characteristics of the depth SR effect to a certain extent. There are two reasons: (1) Most depth SR methods are RGB guided and the RGB edges may be introduced into HR depth maps, which makes it not accurate. Therefore, it is meaningful to calculate the edge errors brought by RGB guidance. (2) Most depth map edges are included in RGB edges. With the influence of illumination in the practical scenes, most discontinuity areas in depth maps also exist edges in RGB images. Thus, calculating the RGB edge errors can to some extent reflect the edge errors of depth maps.

3. Additional Experimental Results

We conduct more experiments on our proposed RGB-D-D dataset to demonstrate the effectiveness of our dataset and method. We retrain $\times 8$ depth map SR models of DKN [2], FDKN [2], DJFR [4] and our FDSR on the training set of RGB-D-D in downsample manner and test them on the testing set. We first downsample the ground truth to obtain LR depth maps by utilizing traditional bicubic interpolation operation. When training on RGB-D-D, we use the same training strategy as the author mentioned in their work. The quantitative results of $\times 8$ depth map SR are shown in Table 2. It can be found that, the performance of all the methods have smaller RMSE than they trained on NYU v2 and our method FDSR achieves the best performance. Observing Figure 1, after retraining, all the methods can produce HR depth maps with clearer and sharper boundaries than before. Besides, noticing the background color in Figure 1, DKN and FDKN may have bigger global errors in some caces when they were trained on NYU v2 [5]. Benifited by the RGB-D-D, it was improved after retraining. Moreover, our model can to some extent correct the depth value errors

*Corresponding author: yzhao@bjtu.edu.cn

The Architecture of FDSR Network		
Layers	Operation	Out
Initial1	RGB2GRAY	$\{N \times 1 \times (H \times 4) \times (W \times 4)\}$
Initial2	Resample	$\{N \times 16 \times H \times W\}$
HFGB1	RGB_Resampled	$\{N \times 16 \times H \times W\}$
HFGB2	Conv(3×3) - LReLU	$\{N \times 32 \times H \times W\}$
HFGB3	HFL(3×3) - LReLU	$\{Y_1^H = N \times 24 \times H \times W; Y_1^L = N \times 8 \times H \times W\}$
HFGB4	HFL(3×3) - LReLU	$\{Y_2^H = N \times 24 \times H \times W; Y_2^L = N \times 8 \times H \times W\}$
HFGB5	HFL(3×3) - LReLU	$\{Y_3^H = N \times 24 \times H \times W; Y_3^L = N \times 8 \times H \times W\}$
MSRB1	Depth	$\{N \times 16 \times H \times W\}$
MSRB2	Conv(3×3) - LReLU	$\{N \times 32 \times H \times W\}$
MSRB3	MSDB(3×3) - LReLU	$\{N \times 32 \times H \times W\}$
MSRB4	Concatenate(MSRB3, HFGB3: Y_1^H)	$\{N \times 56 \times H \times W\}$
MSRB5	MSDB(3×3) - LReLU	$\{N \times 56 \times H \times W\}$
MSRB6	Concatenate(MSRB5, HFGB4: Y_2^H)	$\{N \times 80 \times H \times W\}$
MSRB7	MSDB(3×3) - LReLU	$\{N \times 80 \times H \times W\}$
MSRB8	Concatenate(MSRB7, HFGB5: Y_3^H)	$\{N \times 104 \times H \times W\}$
MSRB9	MSDB(3×3) - LReLU	$\{N \times 104 \times H \times W\}$
MSRB10	MSDB(3×3) - LReLU	$\{N \times 128 \times (H \times 2) \times (W \times 2)\}$
MSRB11	MSDB(3×3) - LReLU	$\{N \times 32 \times (H \times 4) \times (W \times 4)\}$
MSRB12	MSDB(3×3) - LReLU	$\{N \times 1 \times (H \times 4) \times (W \times 4)\}$
MSRB13	Add(MSRB12, Initial1)	$\{N \times 1 \times (H \times 4) \times (W \times 4)\}$

Table 1. The archetecture of the FDSR network. Denote that, the LR depth map is of size $1 \times H \times W$, the HR RGB image is of size $3 \times (H \times 4) \times (W \times 4)$. HFL denotes high-frequency layer, $Y_i^H, i = 1, 2, 3$ and $Y_i^L, i = 1, 2, 3$represents high-frequency components and low-frequency components respectively, MSDB is multi-scale dilated block and LReLU represents leaky ReLU with the slop of 0.2.

	DJFR / DJFR$^+$	FDKN / FDKN$^+$	DKN / DKN$^+$	FDSR / FDSR$^+$
RMSE	5.57 / 2.16	1.91 / 1.83	1.96 / 1.93	1.82 / **1.71**
Value Errors	2.15 / 0.32	0.28 / 0.22	0.33 / 0.23	0.26 / **0.21**
Edge Errors	15.66 / 5.22	3.41 / 3.60	3.55 / 4.20	3.09 / **2.79**

Table 2. RMSE, value errors and edge errors of $\times 8$ depth map SR results. The DJFR [4], FDKN [2], DKN [2] and FDSR are trained on NYU v2 [5]. The DFFR$^+$, FDKN$^+$, DKN$^+$ and FDSR$^+$ are trained on RGB-D-D.

	DJFR	FDKN	DKN	FDSR
RMSE	3.61	2.17	2.12	**1.95**

Table 3. RMSE of $\times 8$ depth map SR results.

which makes our results more accuracy.

We also trained more baselines in the real-world manner on our RGB-D-D dataset.We have trained DJFR, SVLRM, and our FDSR on the LR images and the corresponding RMSE are 6.12, 6.23, 5.49, respectively which demonstrates the performance of our our approach.

In addition, We conducted more experiments on another public dataset: Middlebury dataset [6, 1]. We use 30 pairs data from the 2001-2006 datasets provided by Lu [7]. As same as what DKN did, the DJFR [4], FDKN [2], DKN [2] and FDSR was trained on NYU v2 [5] and tested on Middlebuty dataset [6, 1]. The results of $\times 8$ depth map SR shown in Table 3 futher demonstrate the effectiveness of our baseline.

References

[1] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[2] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, pages 1–22, 2020. 1, 2, 4

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1909–1923, 2019. 1, 2, 4

[5] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 1, 2, 4

[6] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[7] L. Si, X. Ren, and L. Feng. Depth enhancement via low-rank matrix completion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

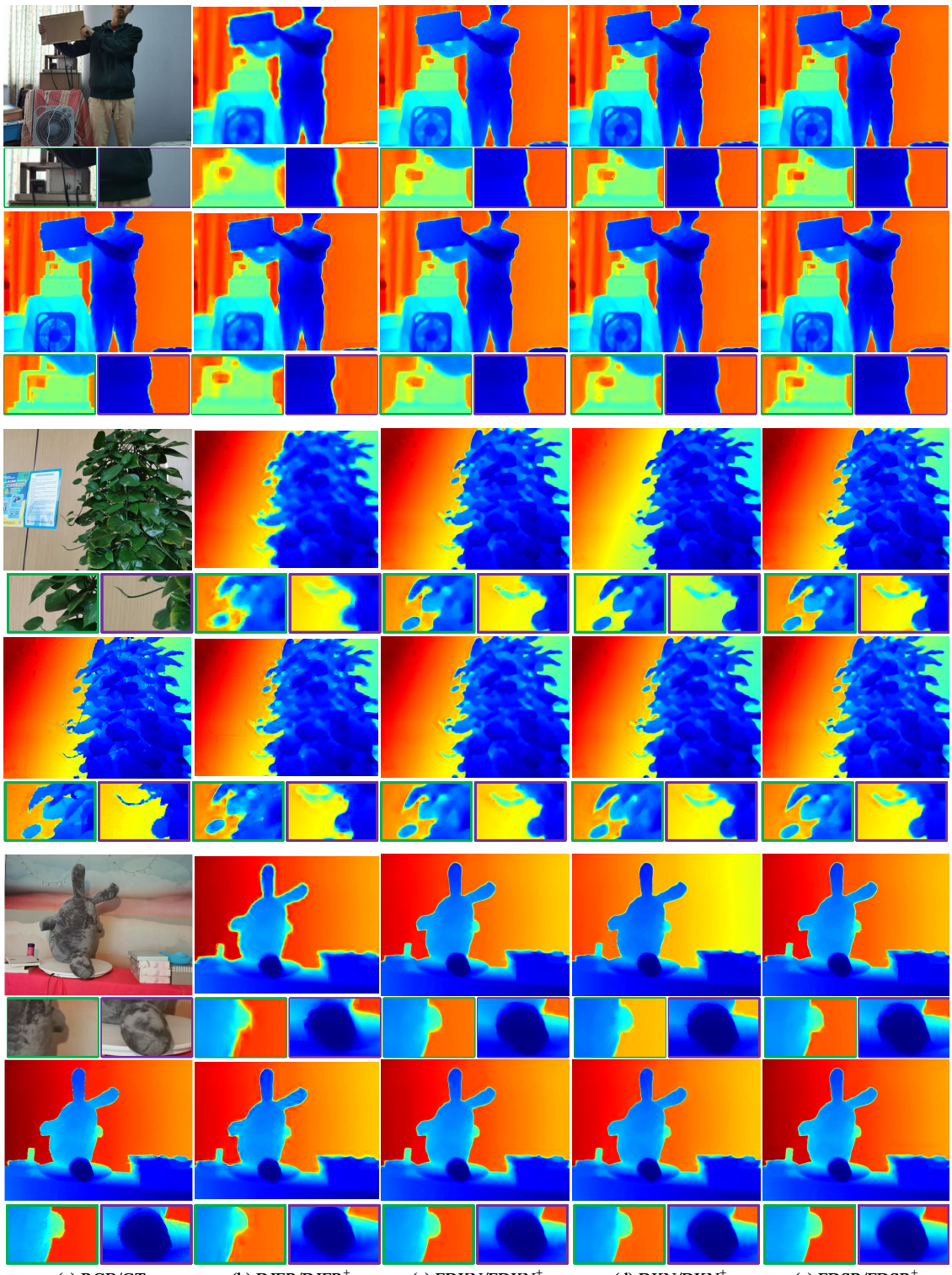

Figure 1. Visual comparison of $\times 8$ depth map SR results. (a) RGB image and ground truth. (b) DJFR [4] trained on NYU v2 [5] and DFFR$^+$ trained on RGB-D-D. (c) FDKN [2] trained on NYU v2 and FDKN$^+$ trained on RGB-D-D. (d) DKN [2] trained on NYU v2 and DKN$^+$ trained on RGB-D-D. (e) FDSR trained on NYU v2 and FDSR$^+$ trained on RGB-D-D. The even and odd rows represent results of models trained on NYU v2 and results of models trained on RGB-D-D respectively.