Spherical Space Feature Decomposition for Guided Depth Map Super-Resolution
SUPPLEMENTARY MATERIALS

Zixiang Zhao[1,2] Jiangshe Zhang[1*] Xiang Gu[1] Chengli Tan[1]
Shuang Xu[3] Yulun Zhang[2] Radu Timofte[2,4] Luc Van Gool[2]

[1]Xi'an Jiaotong University [2]Computer Vision Lab, ETH Zürich
[3]Northwestern Polytechnical University [4]University of Würzburg

zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn

Abstract

In this document, we provide the additional supplementary information for the paper "Spherical Space Feature Decomposition for Guided Depth Map Super-Resolution". This file contains:
(I) The detail architecture for Restormer Block and the Depth Encoder ($\mathcal{E}_{\mathcal{D}}$) in Sec. 3.1.2.
(II) Training details for the Defect Patches Classifier of Spherical Contrast Refinement (SCR) module in Sec. 3.1.4.
(III) Detailed illustration for the training&testing datasets in Sec. 4.1.
(IV) Detailed introduction for the selection and analysis of hyperparameters in Sec. 4.1.
(V) More qualitative comparison fusion results in Sec. 4.2.

1. Detailed introduction for $\mathcal{E}_{\mathcal{D}}$ and $\mathcal{E}_{\mathcal{R}}$

The detailed architecture for *Depth Encoder* $\mathcal{E}_{\mathcal{D}}$ (similar to the RGB Encoder $\mathcal{E}_{\mathcal{R}}$) and the Restormer block [10] in $\mathcal{E}_{\mathcal{D}}$ are illustrated in Fig. 1.

2. Training details for the Defect Patches Classifier of spherical contrast refinement (SCR)

We train the *Defect Patches Classifier* (DPC) on our synthetic "imperfect image dataset" and we utilize the following settings. We use inputs of size 64×64 with a batch size of 256. We train the model for 300 epochs with an initial learning rate of 0.001 (decrease by 10× every 100 epochs), momentum of 0.9 and weight decay of 10^{-6}. The standard multi-class cross-entropy loss is used to train the network. We achieve a training accuracy of 97% and a validation accuracy of 92% in the validation set.

3. Detailed introduction to datasets

We adopt four widely-used benchmarks (NYU v2 [8], Middlebury [2, 7], Lu [6] and RGBDD dataset [1]) for the guided depth super-resolution task. The preprocessing and separation of NYU v2, Middlebury, and Lu datasets follow [4, 5, 9, 3, 12], and that of RGBDD dataset follows [1, 12].

- NYU v2 dataset[1] [8]: it consists of 1449 RGBD image pairs captured by the Microsoft Kinect [11]. We use the first 1,000 images in this dataset for training, and the rest 449 images for testing.
- Middlebury dataset[2] [2, 7]: we use 30 image pairs from 2001-2006 datasets provided by Lu *et al*. [6] for testing.

*Corresponding author.
[1]https://cs.nyu.edu/silberman/datasets/nyu_depth_v2.html
[2]https://vision.middlebury.edu/stereo/data/

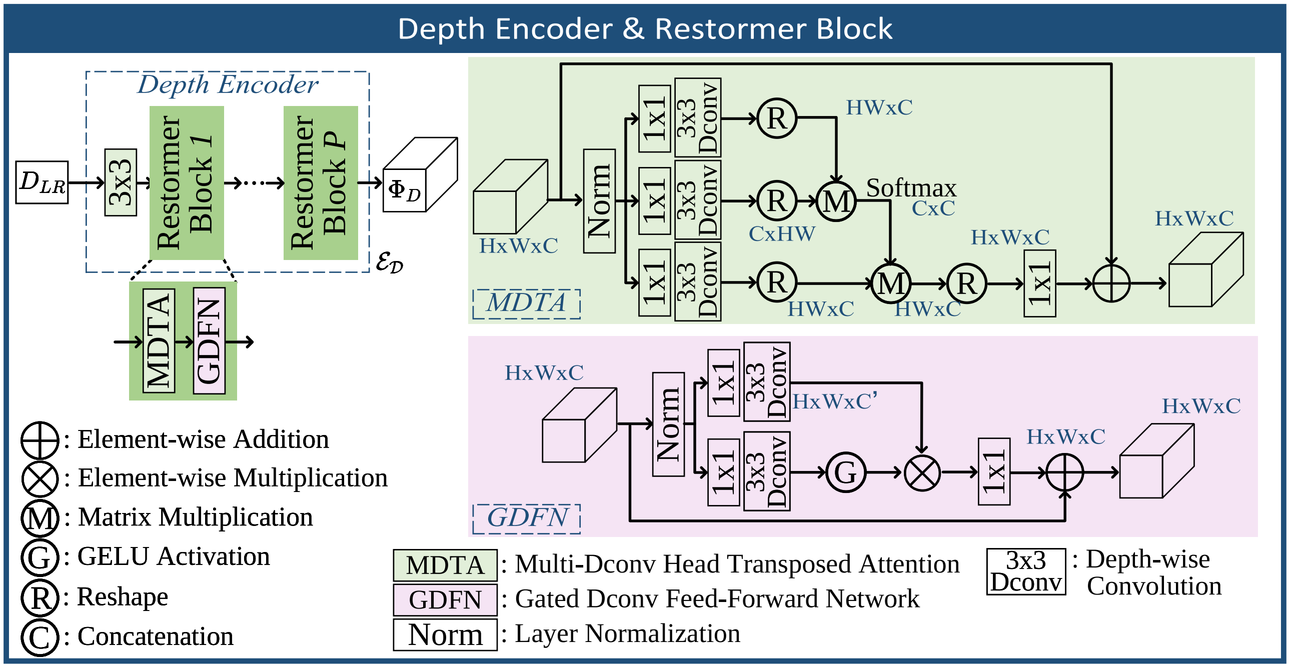

Figure 1: Detail architecture for *Depth Encoder* and the Restormer block of SSDNet.

Interval training epoch for SCR fine-tuning						
Scaling factor	1	2	5	10	20	50
$\times 4$	1.57	1.58	1.60	1.61	1.72	1.83
$\times 8$	3.08	3.08	3.10	3.11	3.22	3.38
$\times 16$	5.81	5.81	5.82	5.84	6.23	6.72

Table 1: The impacts of interval training epoch M for performing SCR fine-tuning on the SSDNet.

- Lu dataset[3] [6]: this dataset consists of 6 RGBD image pairs acquired by ASUS Xtion Pro camera. We use it for testing.
- RGBDD dataset[4] [1]: a new RGBD dataset benchmark proposed in CVPR 2021 [1] with four main categories: portraits, models, plants, and lights. The RGB images and LR depth maps are collected by Huawei P30 Pro and the HR depth maps are captured by Helios ToF camera[5] produced by LUCID vision labs. In our experiments, 297 portraits, 68 plants, and 40 models are utilized for testing. For the *real-world branch*, 1586 portraits, 380 plants, and 249 models are for training, and the test set is the same as above.

4. Selection for the hyperparameters

In this section, we determine the interval training epoch M for performing *Spherical Contrast Refinement* (SCR) fine-tuning. For our proposed SSDNet, SCR fine-tuning is important in addressing the detail issue and further improving the effectiveness of GDSR. We show the results for performing SCR fine-tuning once at every different epoch in training, and the performance in the validation set is shown in Tab. 1.

Obviously, when M is below 10, there is no significant improvement in performance, but there is an increase in training time. Finally, to have a good balance of model performance and training cost, we set $M = 10$ for the other experiments.

[3] http://web.cecs.pdx.edu/ fliu/project/depth-enhance/
[4] http://mepro.bjtu.edu.cn/resource.html
[5] https://thinklucid.com/helios-time-of-flight-tof-camera/

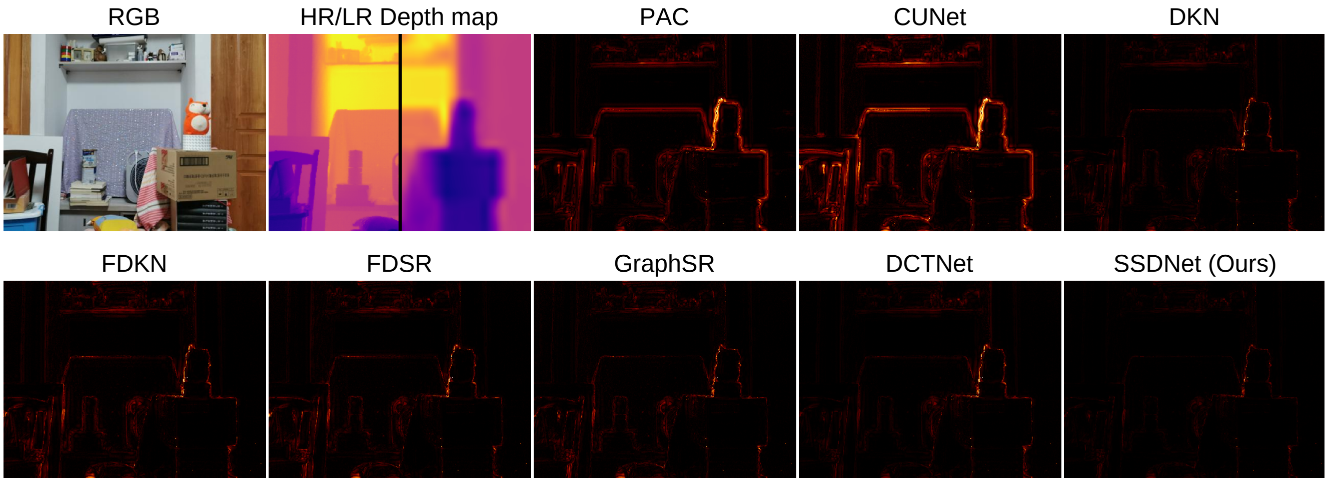

Figure 2: Error maps for visual comparisons in $8\times$ upscaling.

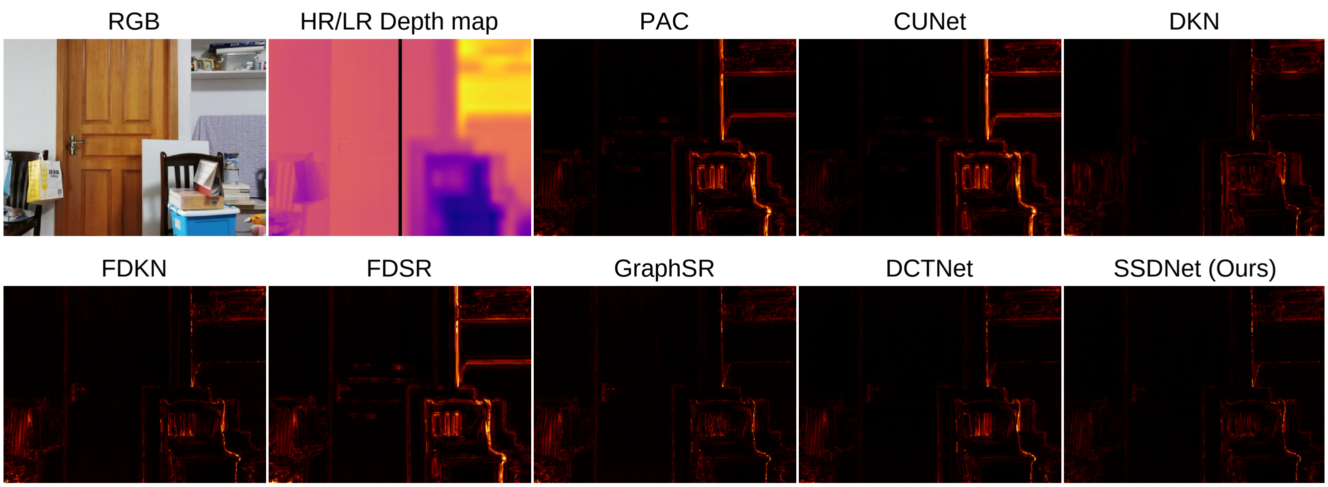

Figure 3: Error maps for visual comparisons in $8\times$ upscaling.

5. More qualitative comparison results

More qualitative comparison results are displayed in Figs. 2 to 5. Our method has excellent performance under multiple datasets and different downsampling scales, showing that our method is suitable for different objects and imaging conditions, and can outperform SOTA methods.

References

[1] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, pages 9229–9238, 2021. 1, 2

[2] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007. 1

[3] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *Int. J. Comput. Vis.*, 129(2):579–600, 2021. 1

[4] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, pages 154–169. Springer, 2016. 1

[5] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1909–1923, 2019. 1

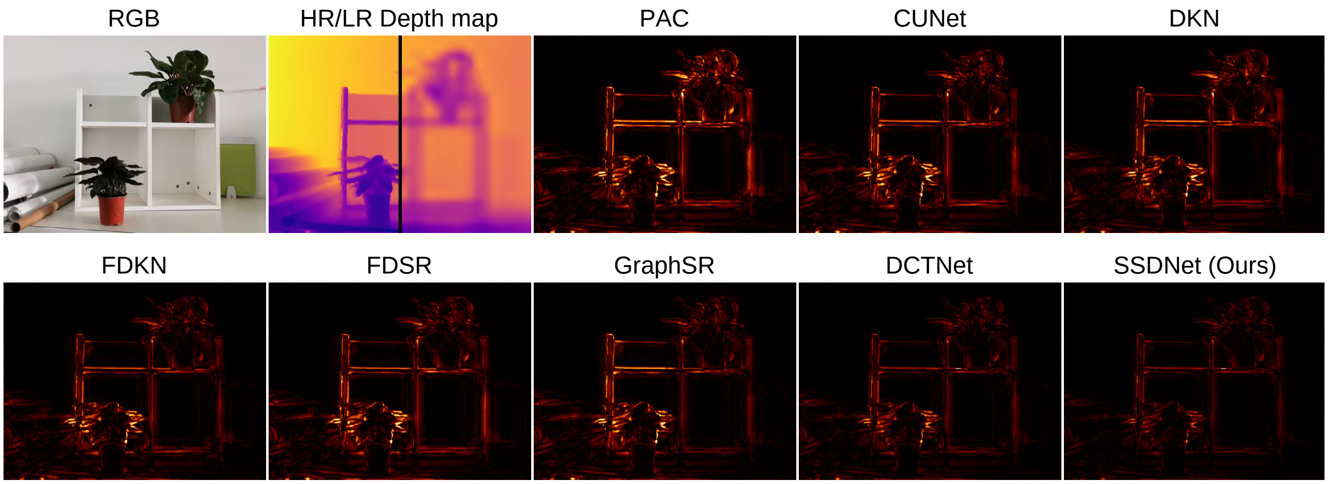

Figure 4: Error maps for visual comparisons in 16× upscaling.

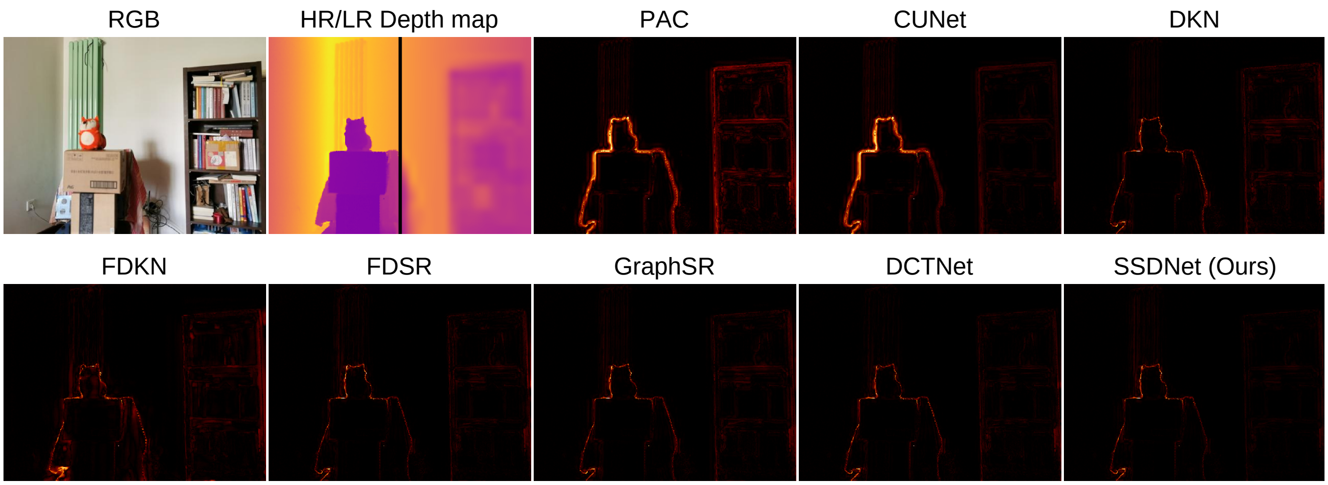

Figure 5: Error maps for visual comparisons in 16× upscaling.

[6] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, pages 3390–3397, 2014. 1, 2

[7] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *CVPR*, 2007. 1

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760. Springer, 2012. 1

[9] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pages 11166–11175, 2019. 1

[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729. IEEE, 2022. 1

[11] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multim.*, 19(2):4–10, 2012. 1

[12] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5687–5697. IEEE, 2022. 1