Supplementary Material: Learning Mutual Modulation for Self-Supervised Cross-Modal Super-Resolution

Xiaoyu Dong[1,2], Naoto Yokoya[1,2](✉), Longguang Wang[3], and Tatsumi Uezato[4]

[1] The University of Tokyo, Tokyo, Japan
[2] RIKEN AIP, Tokyo, Japan
[3] National University of Defense Technology, Changsha, China
[4] Hitachi, Ltd, Tokyo, Japan
dong@ms.k.u-tokyo.ac.jp, yokoya@k.u-tokyo.ac.jp
https://github.com/palmdong/MMSR

Section I analyzes mutual modulation with asymmetric neighborhood sizes. Section II studies different feature fusion approaches. Section III compares the time cost of different self-supervised cross-modal super-resolution (SR) methods, and further compares their performance under noisy guidance. Section IV provides more discussions. Section V provides more qualitative results.

I Modulation with Asymmetric Neighborhood Sizes

In Section 4.3 Ablation Study, we have discussed the effect of the asymmetric neighborhood sizes in our mutual modulation.

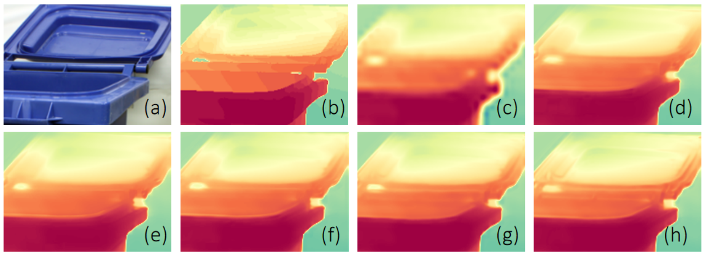

Fig. I. (a) Guide. (b) Ground truth. (c) Bicubic source. Results from models of which the neighborhood sizes for the guide-to-source modulation are (d) 11 × 11, (e) 9 × 9, (f) 7 × 7, (g) 5 × 5, and (h) 3 × 3, respectively

Fig. I provides visual examples of the case in which we fixed the neighborhood size in the source-to-guide modulation as 11 × 11 and reduced that in the guide-to-source modulation. When the neighborhood size is reduced to 9 × 9, the result (Fig. I(e)) is lack of details, because the spatial suppression to the guide is strong. When it is further reduced to 3 × 3, the result (Fig. I(h)) has extraneous structures, because the suppression to the spatial discrepancy in the guide is

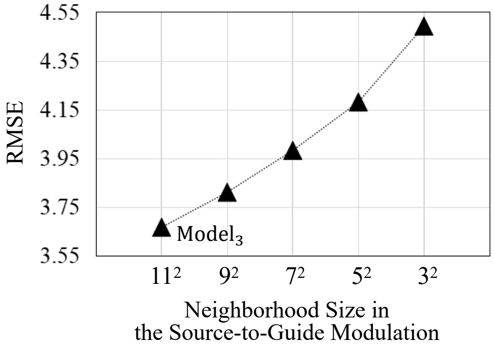

Fig. II. Effect of asymmetric neighborhood sizes. The neighborhood in the guide-to-source modulation is fixed as 11 × 11

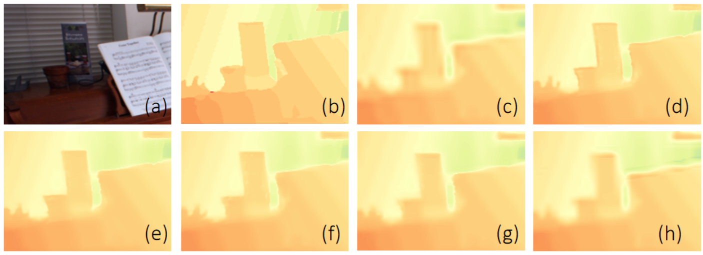

Fig. III. (a) Guide. (b) Ground truth. (c) Bicubic source. Results from models of which the neighborhood sizes for the source-to-guide modulation are (d) 11 × 11, (e) 9 × 9, (f) 7 × 7, (g) 5 × 5, and (h) 3 × 3, respectively

weak. When it is set as 5 × 5, the result (Fig. I(g)) is close to the ground truth (Fig. I(b)) with respect to both spatial resolution and modality characteristics.

Fig. II reports the quantitative results of the case in which we fixed the neighborhood size in the guide-to-source modulation as 11 × 11 and reduced that in the source-to-guide modulation. Fig. III provides visual examples. As the neighborhood size reduces, the results become blurry. This is because the strength to increase the resolution of the source is reduced.

In summary, our mutual modulation allows to handle different types of multi-modal data flexibly. With setting a large neighborhood size for the source-to-guide modulation and a properly small neighborhood size for the guide-to-source modulation, models can optimally increase the resolution of the source and capture and suppress the spatial discrepancy in the guide.

II Different Feature Fusion Approaches

In MMSR, the modulated features $\mathbf{F}_{s2g}$ and $\mathbf{F}_{g2s}$ are fused by a 1×1 convolution. We additionally studied other fusion approaches, including naive summation and

Table I. ×4 depth SR on the Middlebury 2003 dataset

	Sum.	Att. + Sum.	Att. + $\text{Conv}_{1\times1}$	$\text{Conv}_{1\times1}$
RMSE	1.88	1.92	1.84	**1.78**
Training Time	140s	150s	147s	**137s**

attentional fusion (spatial attention and channel attention [9] were performed before summation or 1×1 convolution), as reported in Table I. The model with only a 1×1 convolution as fusion approach achieves the best performance and the shortest training time. Therefore, a 1×1 convolution is adopt to fuse the modulated features in our MMSR.

III Comparisons with Other Self-Supervised Methods

Time Cost. As introduced in Section 1, self-supervised cross-modal SR methods, including CMSR [8], P2P [4], and our MMSR, perform online learning on each combination of low-resolution (LR) source and high-resolution (HR) guide. Table II compares their training time cost. Note that, the time cost of our MMSR is influenced by the modulation neighborhood sizes (i.e., modulation with larger neighborhood sizes results in higher time cost). For depth SR, the neighborhood sizes for the source-to-guide modulation and the guide-to-source modulation in our MMSR were set as 11×11 and 5×5, respectively. The training time of MMSR/P2P/CMSR on each depth-visible input (of size 320×320) is 137s/131s/90s. Our MMSR runs slightly slower yet shows obvious performance superiority, as shown in Table II. For digital elevation model (DEM) SR, the neighborhood sizes for the source-to-guide modulation and the guide-to-source modulation in our MMSR were set as 5×5 and 3×3, respectively. The training time of MMSR/P2P/CMSR on each DEM-visible input (of size 320×320) is 49s/131s/90s. Our MMSR requires much less time and still obtains obvious performance superiority.

Table II. t shows training time on an NVIDIA RTX 3090 GPU. RMSE_{2003}, RMSE_{2005}, and RMSE_{2014} denote the average RMSE on the Middlebury 2003 [7], 2005 [6], and 2014 [5] datasets, respectively. Numbers in brackets show the performance improvement achieved by our MMSR

	×4 Depth SR				×4 DEM SR	
	t	RMSE_{2003}	RMSE_{2005}	RMSE_{2014}	t	RMSE
P2P [4]	131s	2.94 (↑ 39.5%)	3.78 (↑ 34.7%)	3.90 (↑ 41.0%)	131s	1.57 (↑ 53.5%)
CMSR [8]	90s	2.52 (↑ 29.4%)	3.51 (↑ 29.6%)	2.87 (↑ 19.9%)	90s	0.78 (↑ 6.4%)
Ours	137s	1.78 (-)	2.47 (-)	2.30 (-)	49s	0.73 (-)

Performance under Noisy Guidance. Fig. IV further compares our MMSR with CMSR [8] and P2P [4] under noisy guidance. As we can see, under even heavy noise, our MMSR still outperforms CMSR and P2P by a large margin and can produce results that are closer to ground truth. This demonstrates the robustness of our MMSR and the effectiveness of our mutual modulation with cross-domain adaptive filtering.

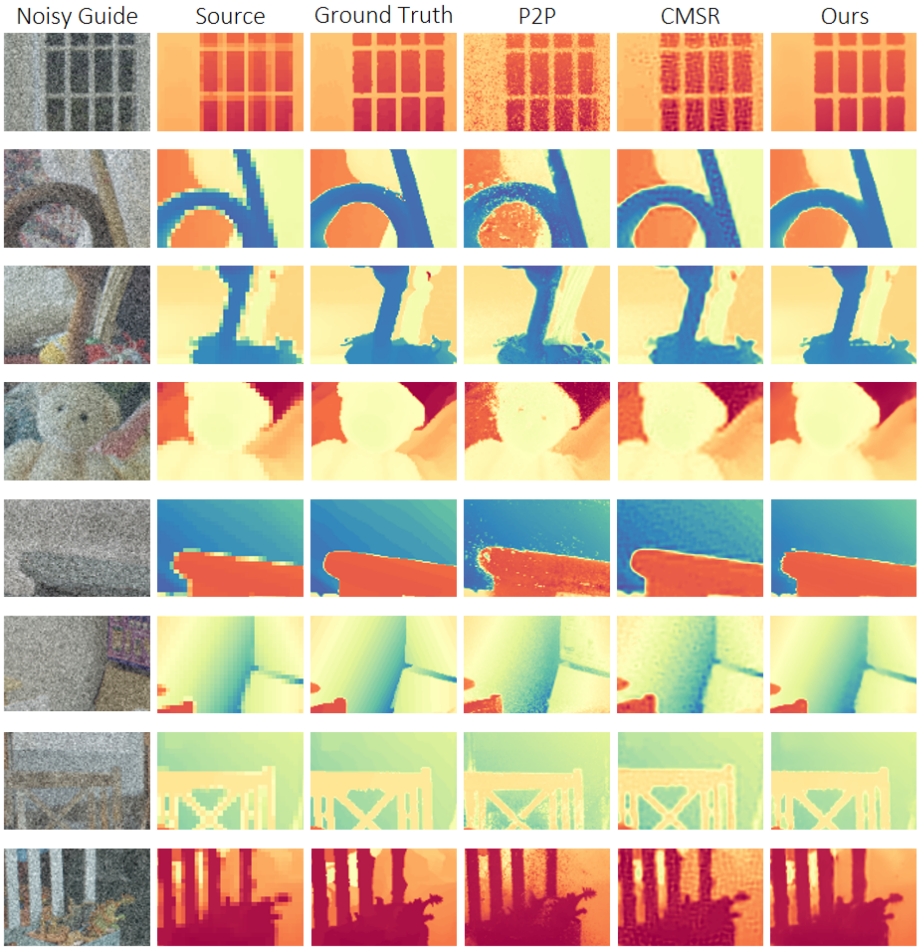

Fig. IV. ×4 depth SR under noisy guidance. The first four and the second four rows show results on the Middlebury 2005 and 2014 datasets, respectively. ′Noisy Guide′ is generated by adding Gaussian noise with noise level 50

IV More Discussions

What Is Important in Cross-Modal SR? Given an LR source and an HR guide from different modalities, cross-modal SR aims at achieving an image product that has spatial resolution comparable with the guide and modality characteristics faithful to the source. We argue both the structural cues from the HR guide and the modality constraint from the LR source are important in the task. Thus we develop a mutual modulation strategy and adopt cycle consistency constraint to fully exploit the guide and also the source, enabling a robust self-supervised MMSR model.

Why Can MMSR Outperform Supervised Methods? Supervised cross-modal SR methods have shown promising performance. However, they have two problems: **(1)** They suffer limited performance in real-world scenes because large-scale paired training data is hard to acquire. **(2)** They cannot easily generalize well to test data that is not in the same domain as the training data. The reasons of our superior performance are twofold: **(1)** Our mutual modulation strategy and cycle-consistent self-supervised learning effectively facilitate our MMSR to achieve state-of-the-art performance. **(2)** The employed online learning scheme allows our MMSR a strong generalization capability to any given input. With robust performance and strong generalizability, MMSR can outperform even supervised methods.

Contributions beyond Superior Performance. Our MMSR outperforms previous supervised and self-supervised methods on various tasks. Moreover, our work also has the following three major contributions: **(1)** The state-of-the-art performance of our MMSR bridges the gap of robust self-supervised cross-modal SR. **(2)** For the first time, our mutual modulation effectively overcomes the spatial discrepancy and resolution gap of multi-modal images, and show correlation-based filtering provides an effective inductive bias for deep cross-modal SR. This benefits further progress in research fields. **(3)** Our MMSR shows superior generalization capability to diverse modalities, robustness to noise, and applicability to real-world scenarios. This is beneficial to real-world applications.

Limitation. Like other methods, MMSR produces ghosting artifacts on some samples. In Fig. 5, ghosting artifacts can be observed around the antlers in the results of FDSR [1], FDKN [2], DKN [2], and MMSR. This is caused by the bicubic/bilinear upsampled source input. Since P2P [4] inputs only the guide image, it does not suffer from ghosting artifacts but produces discrepancy artifacts. Likewise, in Fig. 9, in feature $\mathbf{F}_{g2s}$, the ghosting along the antler is because the guide-to-source modulation induces $\mathbf{F}_g$ to mimic $\mathbf{F}_s$ which has bilinear ghosting. However, compared with previous state-of-the-art methods [1,8,2,4], our MMSR achieves final predictions that are closer to ground truth. Exploring the upperbound performance of self-supervised cross-modal SR models would be an interesting and challenging research problem.

V More Qualitative Results

We provide more visual comparisons between our MMSR and the five cross-modal SR methods [8,1,2,4]. Fig. V, Fig. VI, and Fig. VII show SR results on the depth-visible data from the Middlebury 2003 [7], 2005 [6], and 2014 [5] benchmarks, respectively. Fig. VIII shows SR results on the real-world DEM-visible data from [3]. For depth SR, error maps are provided for better visual comparison. As we can see, our MMSR produces lower errors and finer edge details. Overall, as a self-supervised method, our MMSR achieves state-of-the-art performance on various tasks, and outperforms fully supervised methods (FDSR [1], DKN [2], and FDKN [2]) and previous self-supervised methods (CMSR [8] and P2P [4]) consistently.

References

1. He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., Lin, C., Liu, M., Zhao, Y.: Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In: CVPR (2021)
2. Kim, B., Ponce, J., Ham, B.: Deformable kernel networks for joint image filtering. International Journal of Computer Vision **129**(4), 579–600 (2021)
3. Kunwar, S., Chen, H., Lin, M., Zhang, H., D'Angelo, P., Cerra, D., Azimi, S.M., Brown, M., Hager, G., Yokoya, N., Hänsch, R., Le Saux, B.: Large-scale semantic 3-d reconstruction: Outcome of the 2019 ieee grss data fusion contest—part a. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 922–935 (2020)
4. Lutio, R.d., D'Aronco, S., Wegner, J.D., Schindler, K.: Guided super-resolution as pixel-to-pixel transformation. In: ICCV. pp. 8828–8836 (2019)
5. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: GCPR (2014)
6. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: CVPR (2007)
7. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR. pp. 195–202 (2003)
8. Shacht, G., Danon, D., Fogel, S., Cohen-Or, D.: Single pair cross-modality super resolution. In: CVPR (2021)
9. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: ECCV (2018)

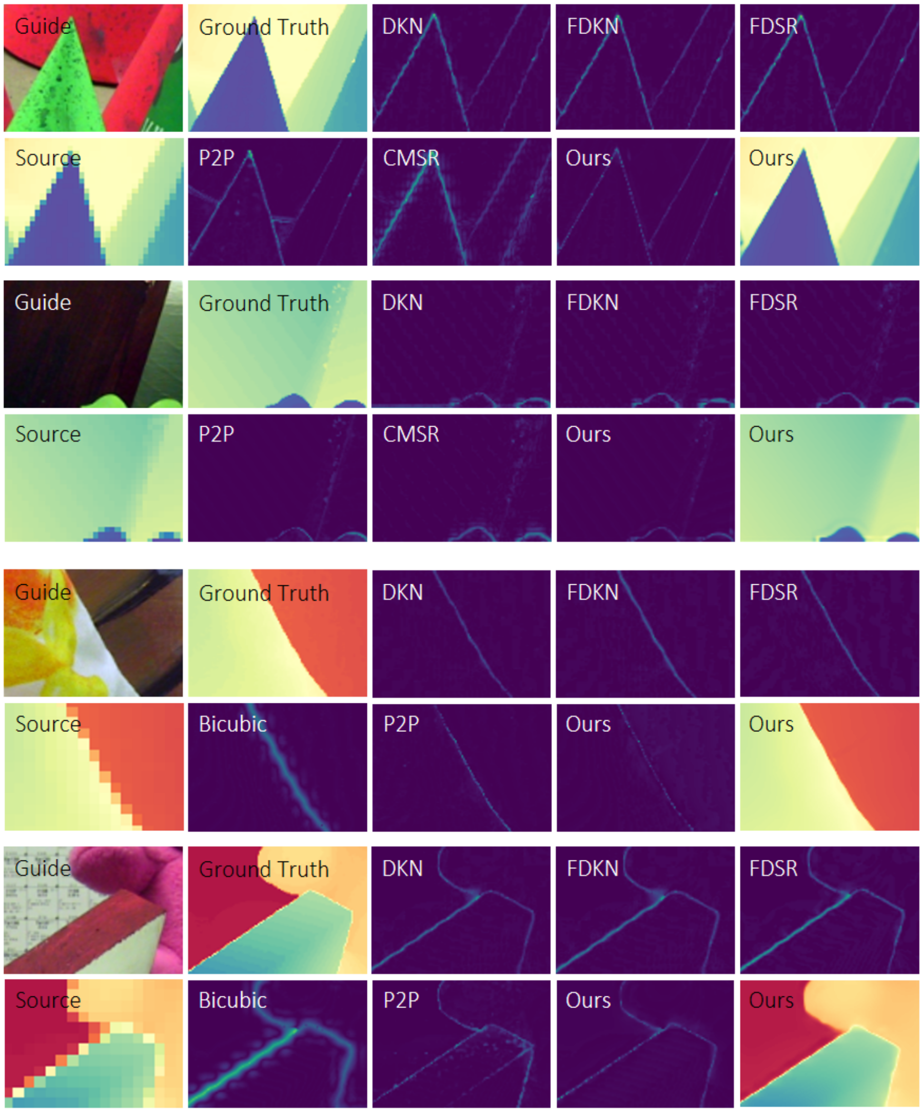

Fig. V. Depth SR on the Middlebury 2003 dataset. The first and second rows show ×4 SR results. The third and fourth rows show ×8 SR results.

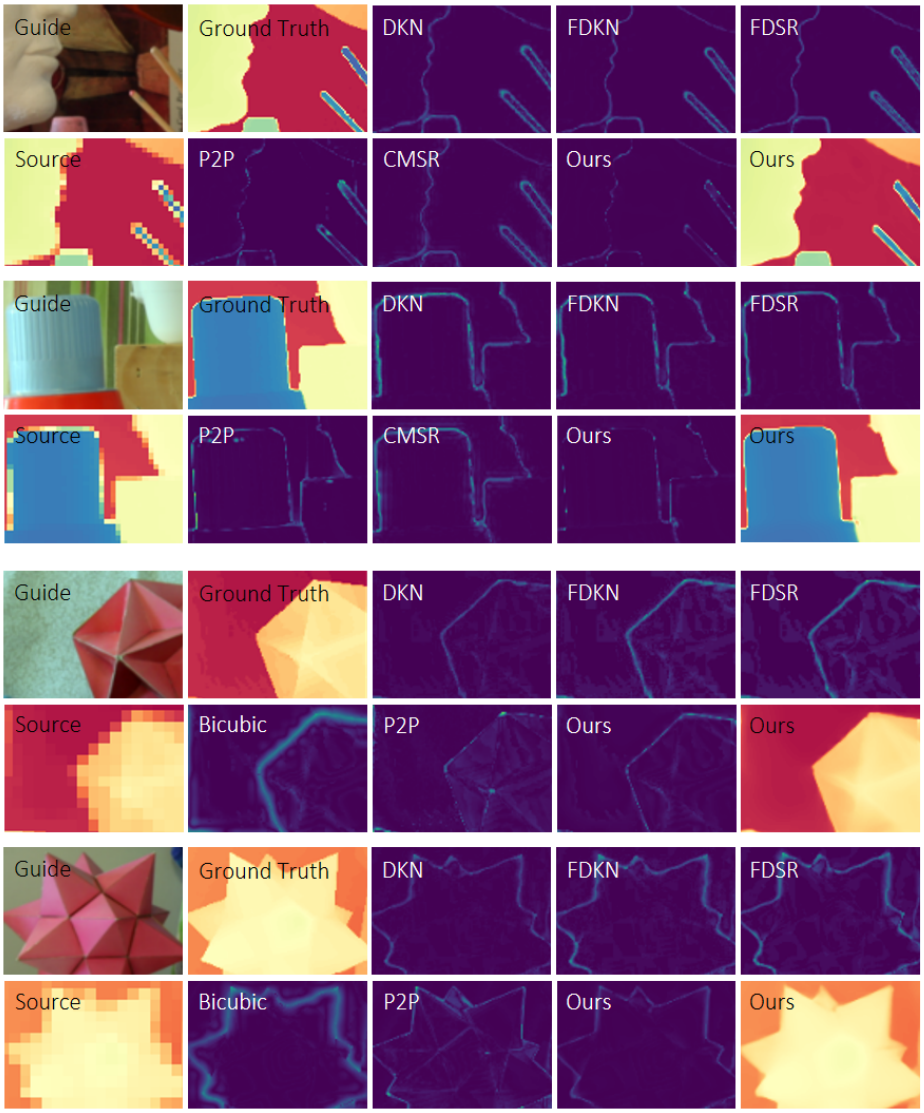

Fig. VI. Depth SR on the Middlebury 2005 dataset. The first and second rows show ×4 SR results. The third and fourth rows show ×8 SR results

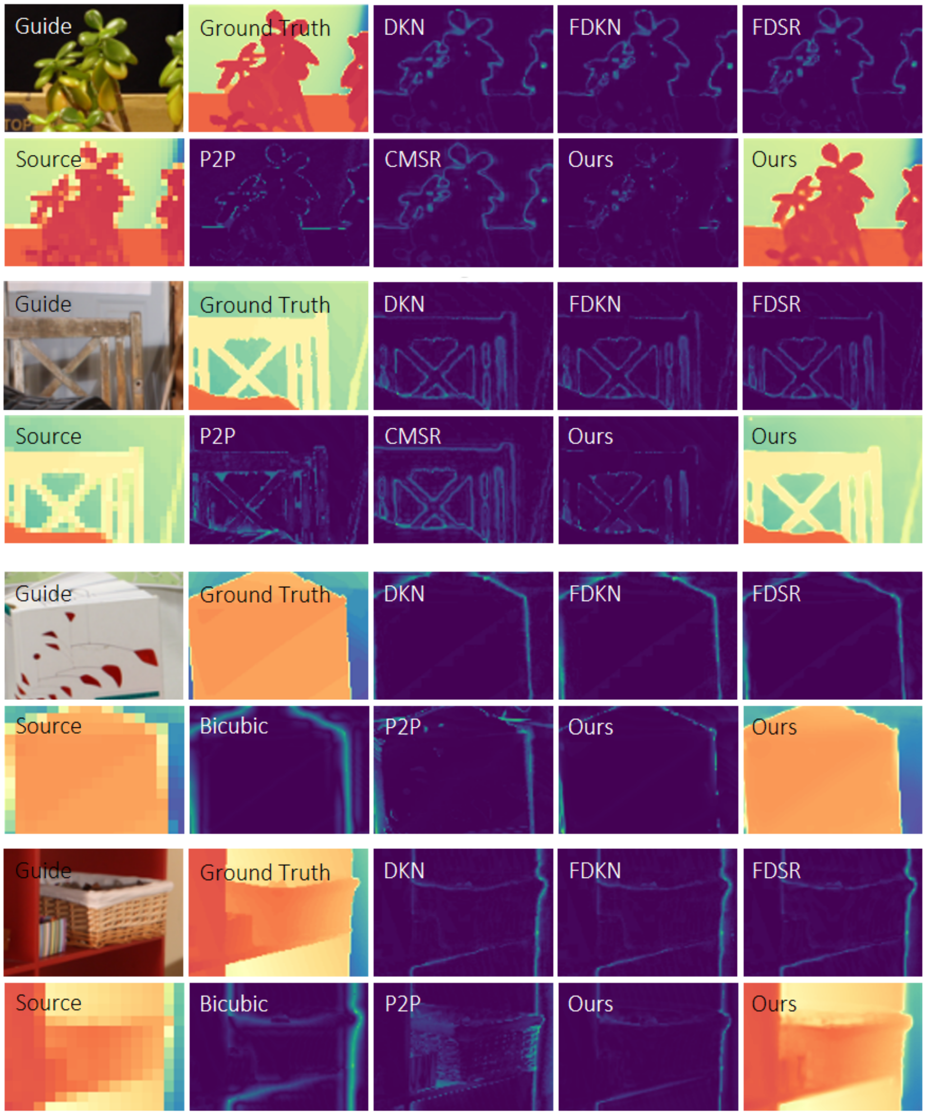

Fig. VII. Depth SR on the Middlebury 2014 dataset. The first and second rows show ×4 SR results. The third and fourth rows show ×8 SR results.

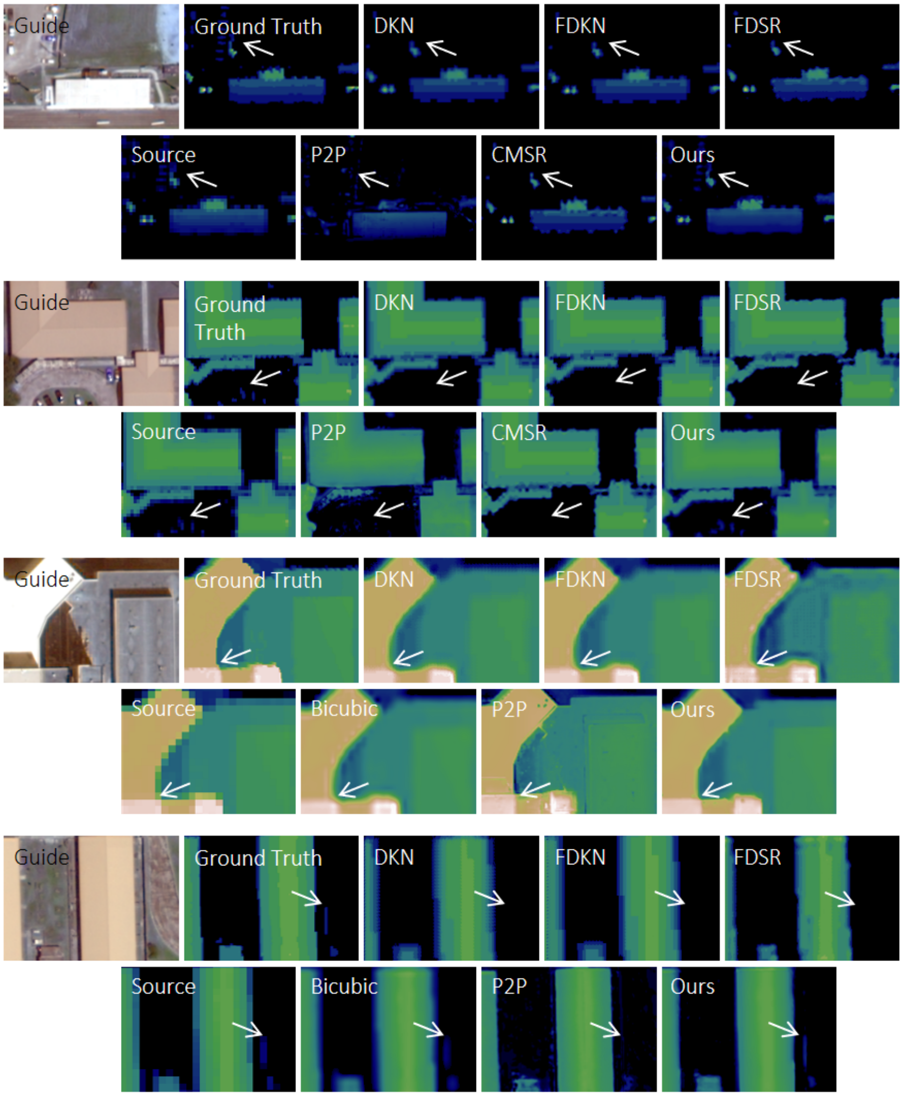

Fig. VIII. DEM SR. The first and second rows show ×4 SR results. The third and fourth rows show ×8 SR results