– Supplemental Materials –

Masked Spatial Propagation Network for Sparsity-Adaptive Depth Refinement

Jinyoung Jun
Korea University
jyjun@mcl.korea.ac.kr

Jae-Han Lee
Gauss Labs Inc.
jaehanlee@mcl.korea.ac.kr

Chang-Su Kim
Korea University
changsukim@korea.ac.kr

A. Detailed network architecture

Figure S1 shows the structure of the proposed guidance network. Input signals $\mathbf{I}$, $\mathbf{S}$, and $\mathbf{D}^0$ generate features of 48, 16, and 16 channels, respectively, which are mixed through concatenation and convolution. The mixed feature is fed to an encoder-decoder network. As the encoder, we adopt PVT-Base [67], which processes a $64 \times H \times W$ tensor to yield an encoded feature of size $512 \times H/32 \times W/32$. The decoder consists of five blocks, each of which performs 3×3 transposed convolution, layer normalization [1], ReLU, and the NAF block [5] operation. The number of channels in each decoder block remains unchanged.

Figure S2 shows the structure of the guidance network for conventional depth completion. In this case, $\mathbf{D}^0$ is estimated similarly to $\mathbf{G}$, using the output of the guidance network. The structure of the encoder-decoder network is identical to that of Figure S1.

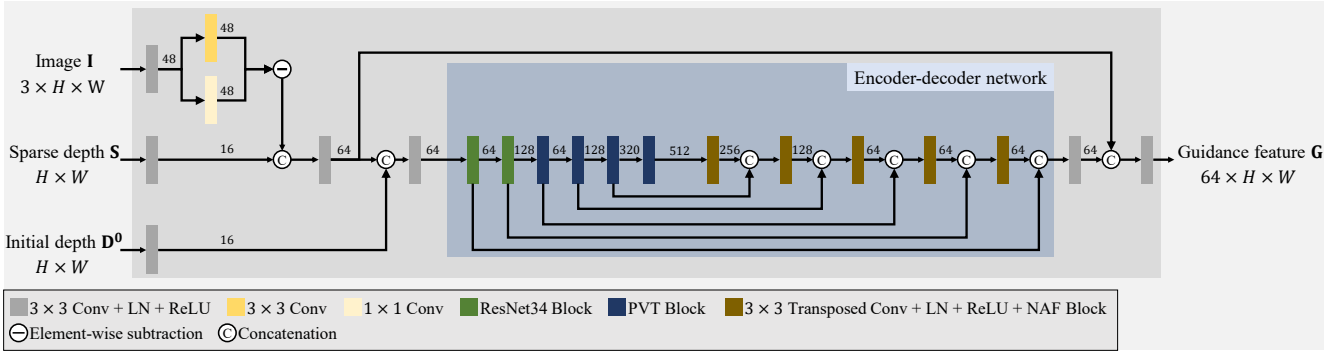

Figure S1. Detailed network architecture of the proposed guidance network.

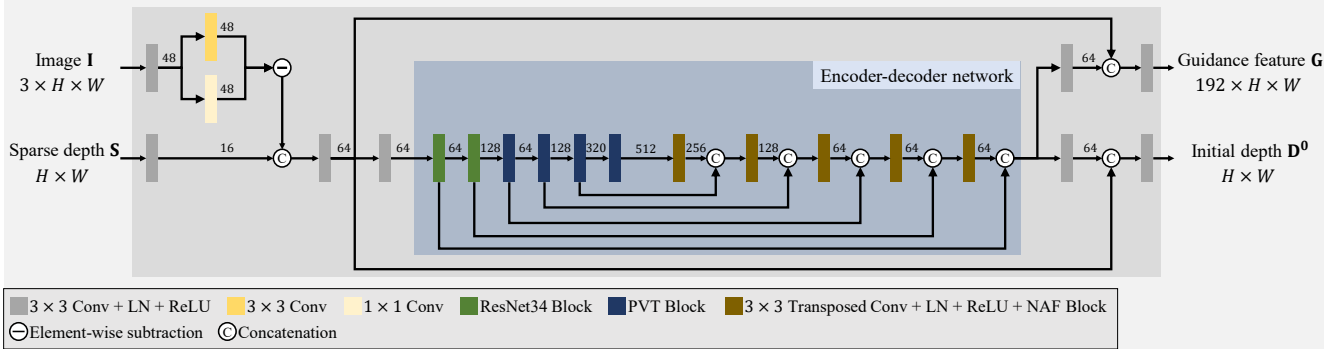

Figure S2. Detailed network architecture of the guidance network in conventional depth completion.

B. Sparsity-based depth refinement

B.1. NYUv2

Figure S3 compares the RMSE performances according to the number of sparse depths on NYUv2. In Figure S3, a solid line indicates that a single model is evaluated for various numbers of sparse depths. On the contrary, each symbol means that a separate model is trained and evaluated for a fixed number of sparse depths. Table S1 and S2 compare the RMSE scores of lined methods and symboled methods on NYUv2, respectively.

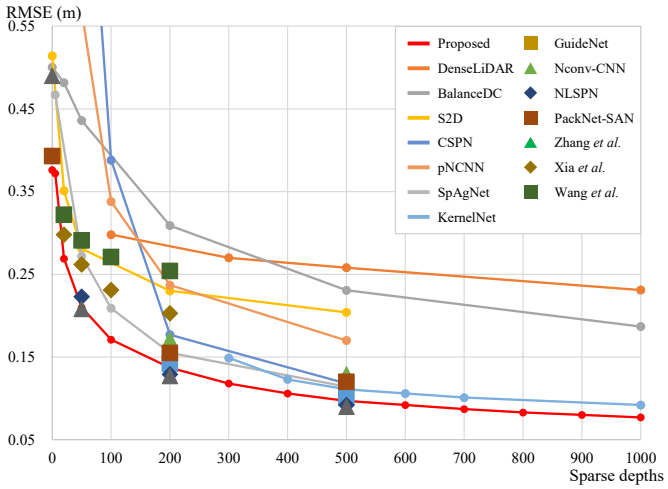

Figure S3. Comparison of the SDR performances on NYUv2.

Table S1. Comparison of RMSE performances (in meters) of the lined methods on NYUv2. In each test, the best result is **boldfaced**.

Method	Number of sparse depths													
	MDE	5	20	50	100	200	300	400	500	600	700	800	900	1000
DenseLiDAR [23]	-	-	-	-	0.298	-	0.270	-	0.258	-	-	-	-	0.231
BalanceDC [77]	0.500	-	0.482	0.436	-	0.309	-	-	0.231	-	-	-	-	0.187
S2D [51]	0.514	-	0.351	0.281	-	0.230	-	-	0.204	-	-	-	-	-
CSPN [9]	-	2.063	-	0.884	0.388	0.177	-	-	0.118	-	-	-	-	-
pNCNN [19]	-	2.412	-	0.568	0.338	0.237	-	-	0.170	-	-	-	-	-
SpAgNet [12]	-	0.467	-	0.272	0.209	0.155	-	-	0.114	-	-	-	-	-
KernelNet [46]	-	-	-	-	-	-	0.149	0.123	0.111	0.106	0.101	-	-	0.092
Proposed	**0.376**	**0.372**	**0.269**	**0.210**	**0.171**	**0.137**	**0.118**	**0.106**	**0.097**	**0.092**	**0.087**	**0.083**	**0.080**	**0.077**

Table S2. Comparison of RMSE performances (in meters) of the symboled methods on NYUv2. In each test, the best result is **boldfaced**.

Method	Number of sparse depths					
	MDE	20	50	100	200	500
GuideNet [64]	-	-	-	-	0.142	0.101
Nconv-CNN [18]	-	-	-	-	0.171	0.129
Xia *et al.* [69]	-	**0.298**	0.262	**0.231**	0.203	-
Wang *et al.* [66]	-	0.322	0.291	0.271	0.254	-
NLSPN [54]	0.562	-	0.223	-	0.129	0.092
PackNet-SAN [24]	**0.393**	-	-	-	0.155	0.120
Zhang *et al.* [79]	0.490	-	**0.208**	-	**0.127**	**0.090**

B.2. KITTI validation set

Figure S4 compares the RMSE performances according to the number of LiDAR lines on the KITTI validation set. In Figure S4, a solid line indicates that a single model is evaluated for various numbers of sparse depths. On the contrary, each symbol means that a separate model is trained and evaluated for a fixed number of LiDAR lines. Table S3 and S4 compare the RMSE scores of the lined methods and the symboled methods on the KITTI validation set, respectively.

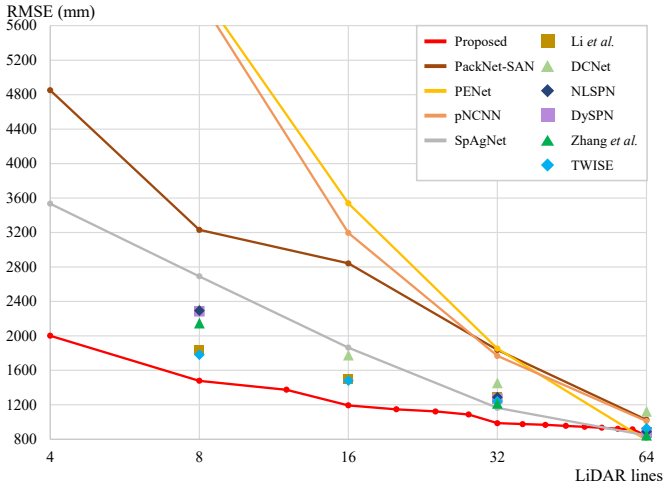

Figure S4. Comparison of the SDR performances on the KITTI validation set.

Table S3. Comparison of RMSE performances (in millimeters) of the lined methods on the KITTI validation set.

Method	LiDAR lines																
	MDE	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64
PackNet-SAN [24]	-	4850.2	3231.0	-	2841.4	-	-	-	1836.8	-	-	-	-	-	-	-	1027.3
PENet [30]	-	9318.9	6015.0	-	3538.0	-	-	-	1853.1	-	-	-	-	-	-	-	**791.6**
SpAgNet [12]	-	3533.7	2691.3	-	1863.3	-	-	-	1164.2	-	-	-	-	-	-	-	844.8
pNCNN [19]	-	9364.6	5921.9	-	3194.7	-	-	-	1766.8	-	-	-	-	-	-	-	1011.9
Proposed	**2377.1**	**2002.0**	**1478.9**	**1374.9**	**1193.1**	**1147.8**	**1123.2**	**1087.6**	**987.2**	**976.5**	**967.4**	**955.6**	**944.7**	**934.3**	**921.8**	**915.9**	840.0

Table S4. Comparison of RMSE performances (in millimeters) of the symboled methods on the KITTI validation set.

Method	LiDAR lines					
	MDE	4	8	16	32	64
Li *et al.* [43]	4185.1	-	1841.6	1504.3	1288.8	889.7
DCNet [32]	4433.6	-	2311.9	1777.3	1456.2	1125.1
DySPN [44]	-	-	2285.8	-	1274.8	878.5
NLSPN [54]	4362.4	2293.1	-	-	1288.9	889.4
TWISE [33]	**4078.8**	-	**1782.5**	**1481.1**	1242.6	927.6
Zhang *et al.* [79]	-	**2150.0**	-	-	**1218.6**	**848.7**

C. Analysis

C.1. MSPN

Figure S5 (a) shows the RMSE performances of each MSPN layer on NYUv2. Since the role of the second layer is to refine nearby regions of sparse depths, it yields a higher gain as more sparse depths are provided. Figure S5 (b) compares the SDR results of this ablated setting on NYUv2. We see that the ablated setting severely degrades the SDR results, especially with a small number of sparse depths. This indicates that the mask update plays a crucial role in MSPN. Table S5 and S6 compare the RMSE scores of Figure S5 (a) and (b), respectively.

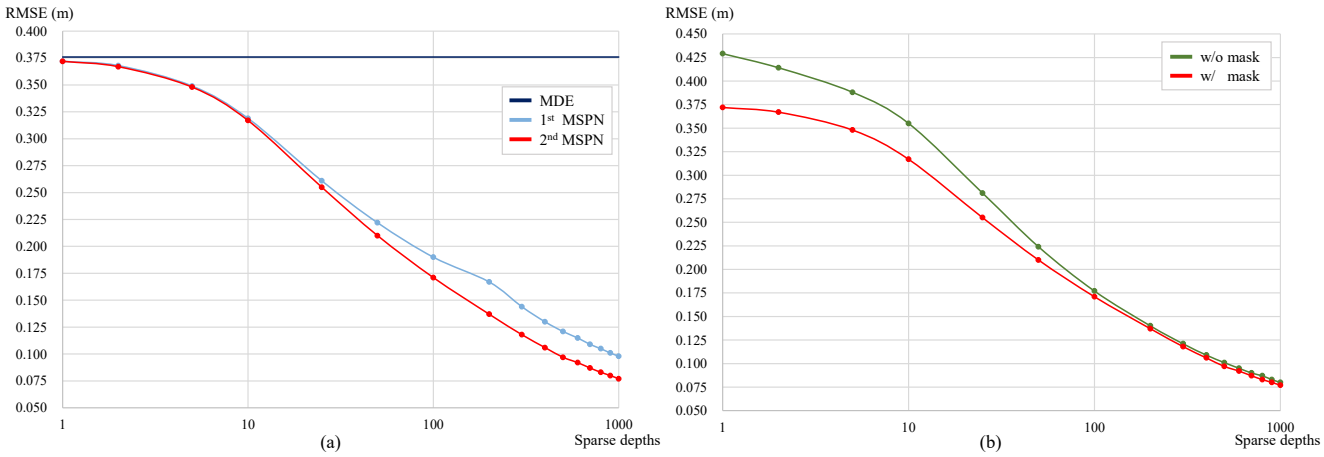

Figure S5. (a) SDR results of each MSPN layer. (b) Ablation study on MSPN.

Table S5. SDR results of each MSPN layer on NYUv2.

Method	Number of sparse depths																
	MDE	1	2	5	10	25	50	100	200	300	400	500	600	700	800	900	1000
MDE	**0.376**	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376	0.376
1st MSPN	**0.376**	**0.372**	0.368	0.349	0.319	0.261	0.222	0.190	0.167	0.144	0.130	0.121	0.115	0.109	0.105	0.101	0.098
2nd MSPN	**0.376**	**0.372**	**0.367**	**0.348**	**0.317**	**0.255**	**0.210**	**0.171**	**0.137**	**0.118**	**0.106**	**0.097**	**0.092**	**0.087**	**0.083**	**0.080**	**0.077**

Table S6. Ablation study of the mask update process in MSPN on NYUv2.

Method	Number of sparse depths																
	MDE	1	2	5	10	25	50	100	200	300	400	500	600	700	800	900	1000
w/o mask	**0.376**	0.450	0.439	0.417	0.385	0.302	0.236	0.182	0.144	0.124	0.112	0.103	0.097	0.092	0.088	0.085	0.081
w/ mask	**0.376**	**0.372**	**0.367**	**0.348**	**0.317**	**0.255**	**0.210**	**0.171**	**0.137**	**0.118**	**0.106**	**0.097**	**0.092**	**0.087**	**0.083**	**0.080**	**0.077**

C.2. Generalization

Figure S6 and Table S7 compare the generalization performance of the proposed SDR on MDEs. In this experiment, we use other off-the-shelf networks, [7, 21, 36, 41] as monocular depth estimators and evaluate the SDR performances. Also, the guidance network and MSPN, trained for Jun *et al*. [37], are not fine-tuned. We observe similar trends for various MDEs, which indicates that the proposed SDR framework provides robust results without being sensitive to the adopted monocular depth estimators.

Next, to simulate MDE with better performance, we blend the monocular depth estimate $\mathbf{D}^0$ of Yuan *et al*. [78] with a ground-truth depth map as follows,

$$\mathbf{D}^0_{\text{blend}} = c \cdot \mathbf{D}_{\text{GT}} + (1 - c) \cdot \mathbf{D}^0. \tag{S1}$$

In Figure S6 and Table S7, we can see that there is room for further development with better MDEs.

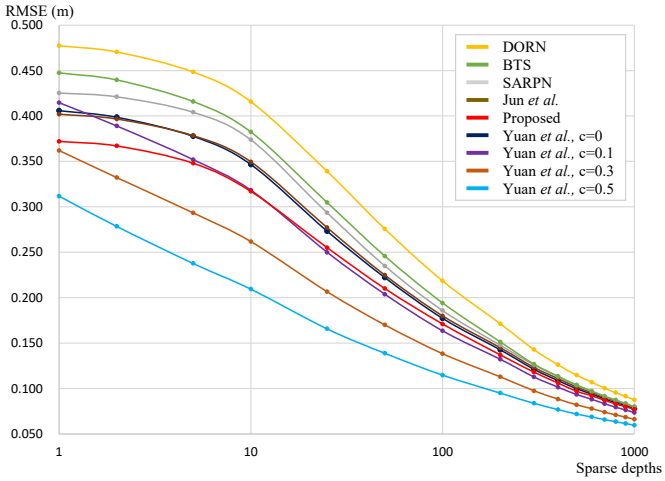

Figure S6. SDR results using different monocular depth estimators NYUv2.

Table S7. SDR results using different monocular depth estimators on NYUv2. In each test, the best result is **boldfaced**.

Method	Number of sparse depths																
	MDE	1	2	5	10	25	50	100	200	300	400	500	600	700	800	900	1000
DORN [21]	0.493	0.477	0.470	0.448	0.416	0.339	0.276	0.218	0.171	0.143	0.126	0.115	0.107	0.100	0.095	0.092	0.088
BTS [41]	0.486	0.447	0.440	0.416	0.382	0.305	0.246	0.194	0.151	0.127	0.113	0.104	0.097	0.091	0.087	0.083	0.080
SARPN [7]	0.420	0.425	0.421	0.404	0.374	0.293	0.235	0.186	0.148	0.126	0.113	0.104	0.097	0.092	0.087	0.084	0.080
Jun *et al*. [36]	0.402	0.402	0.397	0.378	0.350	0.277	0.225	0.180	0.145	0.123	0.111	0.102	0.095	0.090	0.086	0.083	0.079
Jun *et al*. [37]	**0.376**	**0.372**	**0.367**	**0.348**	**0.317**	**0.255**	**0.210**	**0.171**	**0.137**	**0.118**	**0.106**	**0.097**	**0.092**	**0.087**	**0.083**	**0.080**	**0.077**
Yuan *et al*. [78]	0.417	0.406	0.399	0.378	0.347	0.273	0.222	0.177	0.143	0.121	0.109	0.100	0.094	0.089	0.084	0.081	0.078
Yuan *et al*. [78], c=0.1	0.376	0.415	0.389	0.352	0.318	0.250	0.204	0.163	0.132	0.113	0.102	0.093	0.088	0.083	0.079	0.076	0.074
Yuan *et al*. [78], c=0.3	0.292	0.362	0.332	0.293	0.262	0.207	0.170	0.138	0.113	0.097	0.088	0.082	0.078	0.074	0.071	0.068	0.066
Yuan *et al*. [78], c=0.5	**0.209**	**0.312**	**0.278**	**0.238**	**0.209**	**0.166**	**0.139**	**0.115**	**0.095**	**0.084**	**0.077**	**0.072**	**0.069**	**0.066**	**0.064**	**0.062**	**0.060**

C.3. Cross-dataset evaluation

Figure S7 and Table S8 compare the cross-dataset evaluation performance of SDR on the RealSense split of SUN RGB-D [63] with NLSPN [54] and Zhang *et al*. [79]. We observe that the proposed SDR provides robust performance on unseen cameras. Figure S8 qualitatively compares the cross-dataset evaluation results using 100 sparse depth points. We can see that conventional methods provide less reliable results in areas with larger holes, where sparse depths cannot be provided as input. Similar trends can be observed in Table 4 and Figure 11 of the main paper.

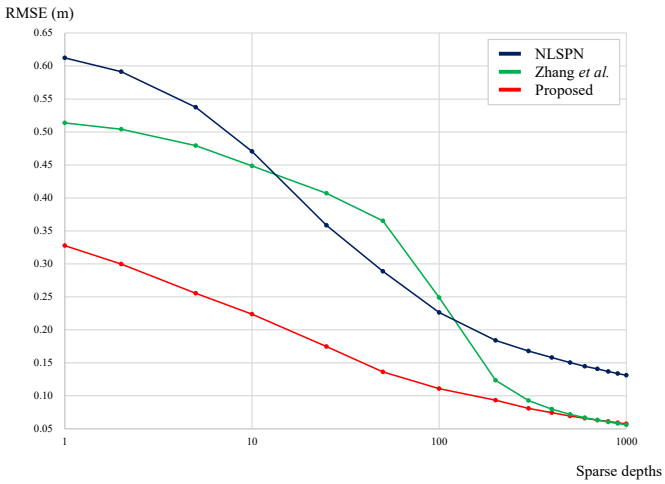

Figure S7. Cross-dataset evaluation results on SUN RGB-D.

Table S8. Comparison of SDR results on the RealSense split of SUN RGB-D.

Method	Number of sparse depths																
	MDE	1	2	5	10	25	50	100	200	300	400	500	600	700	800	900	1000
NLSPN [54]	0.633	0.612	0.591	0.537	0.470	0.359	0.289	0.226	0.184	0.168	0.158	0.150	0.145	0.141	0.137	0.134	0.131
Zhang *et al*. [79]	0.523	0.514	0.504	0.479	0.448	0.407	0.365	0.249	0.124	0.093	0.080	0.072	0.067	**0.063**	**0.060**	**0.058**	**0.056**
Proposed	**0.378**	**0.328**	**0.300**	**0.255**	**0.224**	**0.175**	**0.136**	**0.111**	**0.093**	**0.081**	**0.074**	**0.069**	**0.066**	**0.063**	0.061	0.059	0.057

C.4. Network parameter and inference speed

Table S9 compares the number of network parameters and the inference speed of the proposed algorithm. In this test, we use a Ryzen 5900x CPU and an RTX 3090 GPU. Since the number of iterations changes with the number of sparse depths, we report the speed of MSPN per iteration. It can be observed that the proposed algorithm has a reasonable runtime applicable to real-world applications.

Table S9. Network parameter and inference speed of the proposed algorithm.

Component	Parameters (M)	Speed (seconds-per-frame)
Guidance Network	82.49	0.036
MSPN (per iteration)	0.018	0.007

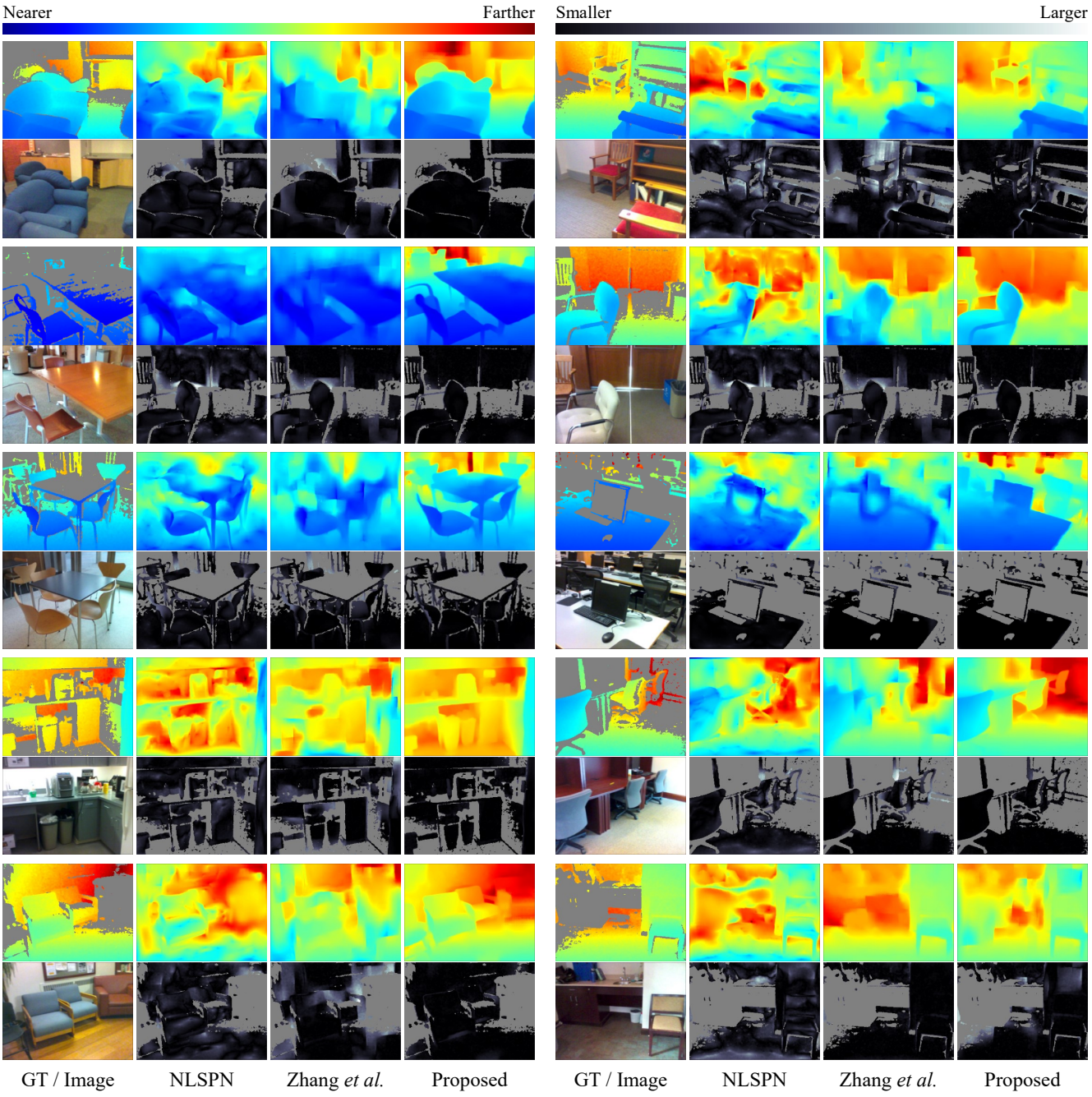

Figure S8. Qualitative comparison of SDR results using 100 sparse depth points on the RealSense split of SUN RGB-D. For each depth map, the corresponding error map is provided below, in which brighter pixels represent larger errors.

D. Qualitative results

Figures S9 and S10 compare SDR results of the proposed algorithm. Figures S11 compares the proposed algorithm with conventional algorithms [52, 54, 79] qualitatively on NYUv2.

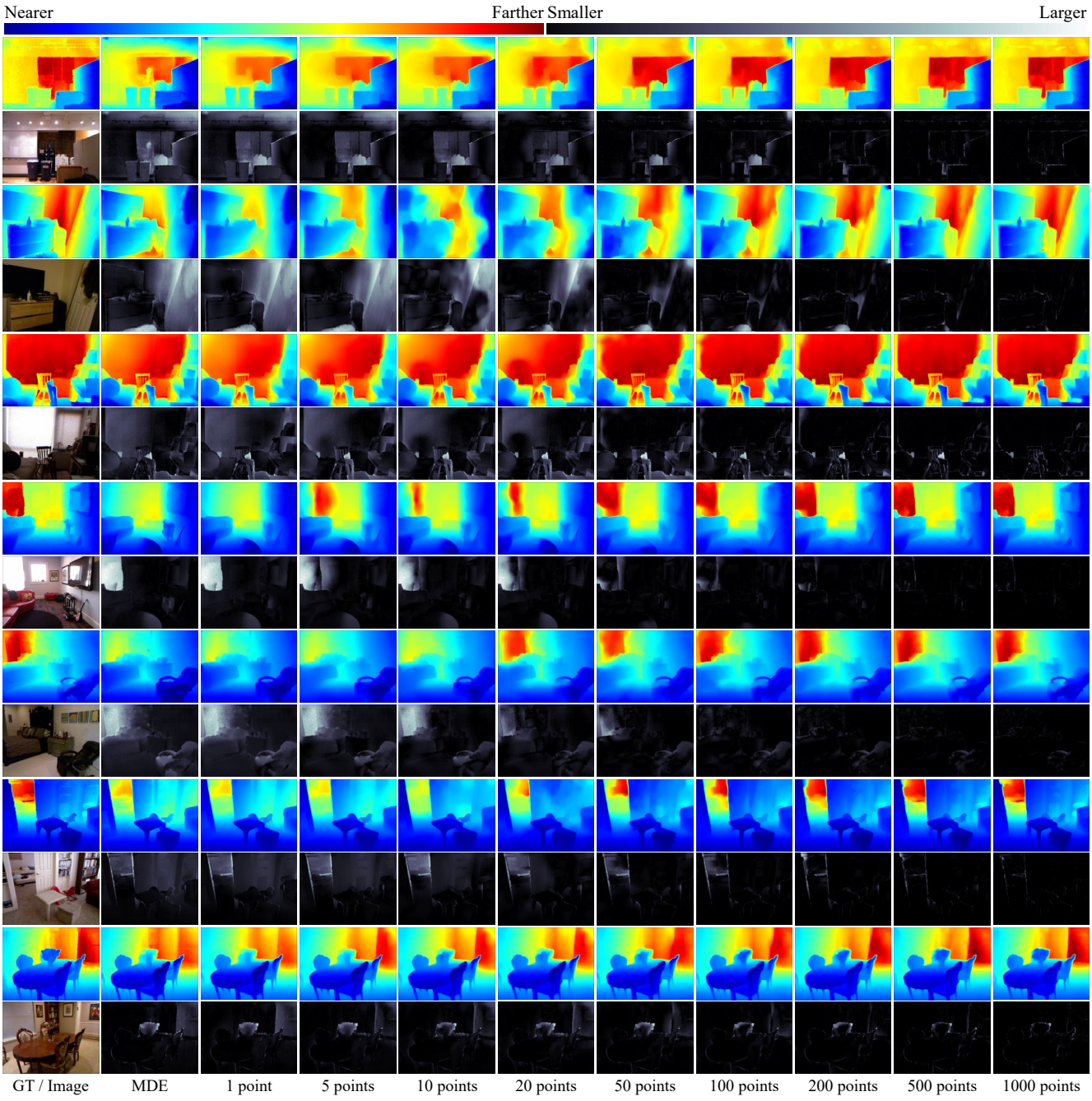

Figure S9. Qualitative comparison of SDR results on NYUv2. For each depth map, the corresponding error map is provided below, in which brighter pixels represent larger errors.

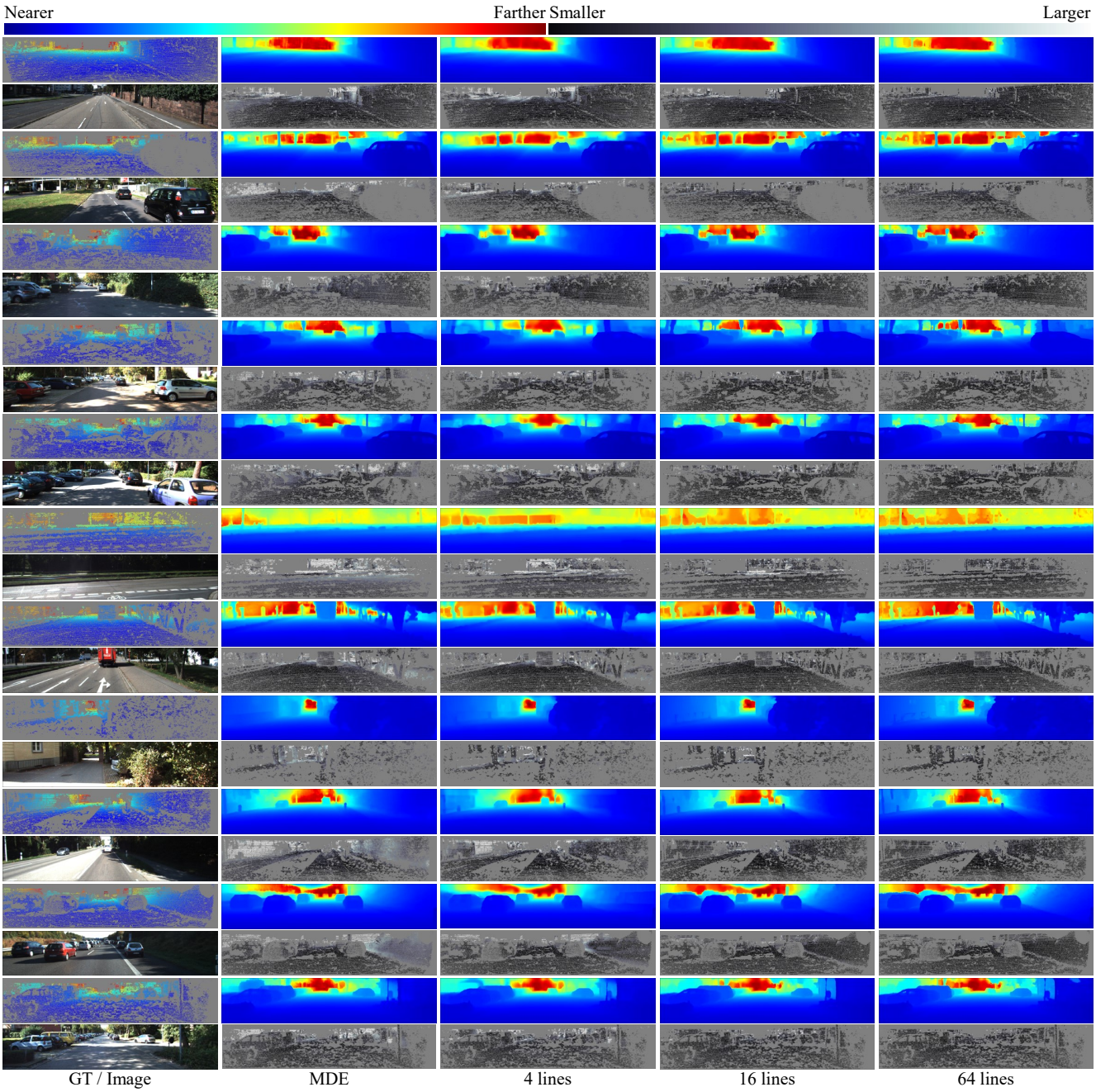

Figure S10. Qualitative comparison of SDR results on the KITTI validation set. For each depth map, the corresponding error map is provided below, in which brighter pixels represent larger errors.

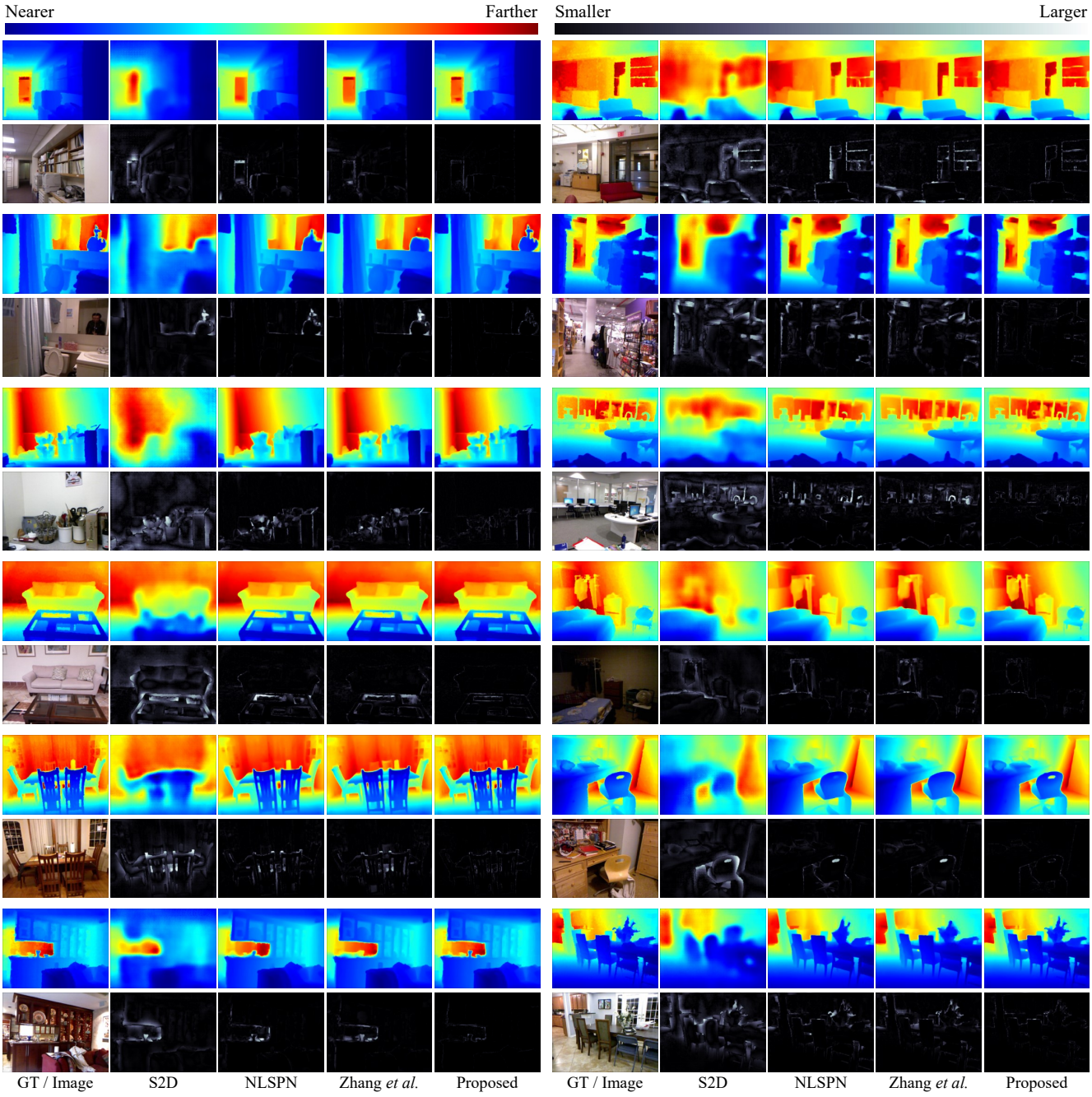

Figure S11. Qualitative comparison of ordinary depth completion results on NYUv2. For each depth map, the corresponding error map is provided below, in which brighter pixels represent larger errors.