Supplementary Material: Non-Local Spatial Propagation Network for Depth Completion

Jinsun Park[1], Kyungdon Joo[2], Zhe Hu[3], Chi-Kuei Liu[3], and In So Kweon[1]

[1] Korea Advanced Institute of Science and Technology, Republic of Korea
{zzangjinsun, iskweon77}@kaist.ac.kr
[2] Robotics Institute, Carnegie Mellon University
kjoo@andrew.cmu.edu
[3] Hikvision Research America

1 Overview

In this supplementary material, we present the detailed network architecture and additional ablation study results on loss functions and the number of propagation steps. We also provide additional analysis on the affinity normalization algorithms (*cf.*, Sec. 6.3 in the main paper), and additional results on the NYU Depth V2 (NYUv2) [10] and the KITTI Depth Completion (KITTI DC) [11] datasets.

2 Network Architecture

The proposed non-local spatial propagation network mainly consists of two parts: (1) an encoder-decoder architecture to predict an initial depth map, confidence, non-local neighbors, and raw affinities, and (2) a non-local spatial propagation layer with a confidence-incorporated learnable affinity normalization. Figure A shows the overview and the detailed network architecture of the proposed algorithm. Our encoder-decoder architecture is designed based on the ResNet34 [5] together with the encoder-decoder feature connection strategy [9, 4]. Features extracted from the encoder-decoder network are fed into inference layers which predict initial depth, confidence, non-local neighbors, and affinities. These inference results are fed into the non-local spatial propagation layer, and the initial depth is iteratively refined to generate the final dense depth.

3 Additional Ablation Study Results

In this section, we conduct additional ablation studies on 1) loss functions and 2) the number of propagation iterations, which are not included in the main paper due to the limited space. Unless stated, experimental settings are the same as those of Sec. 6.3 in the main paper.

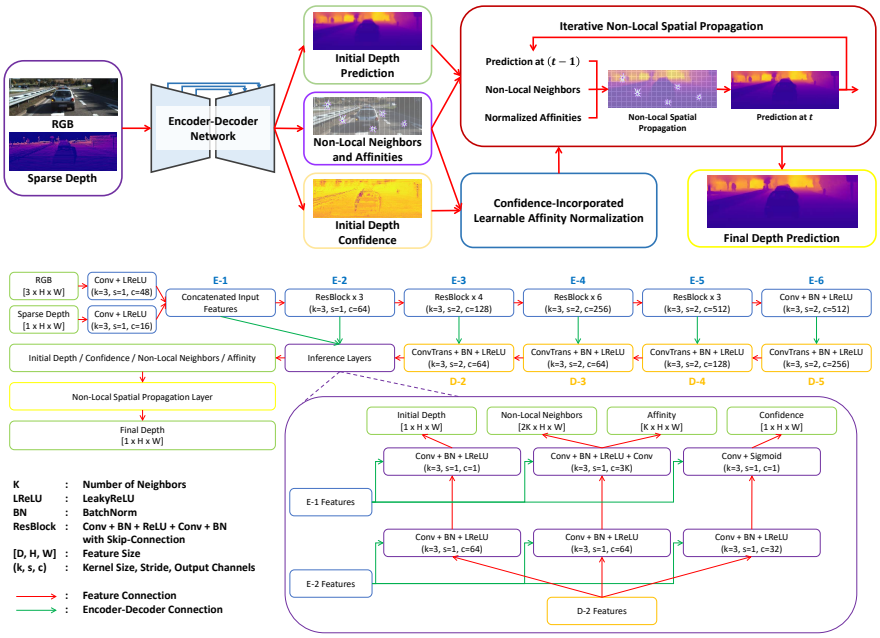


Fig. A. Detailed network architecture of the proposed algorithm. Green boxes: input/output data, blue boxes: encoder layers, orange boxes: decoder layers, purple boxes: inference layers, and yellow box: non-local spatial propagation layer.

3.1 Loss Functions

In order to compare the performance of the proposed algorithm with different loss functions, our network is trained with ℓ_1, ℓ_2, and $\ell_1 + \ell_2$ loss functions, as shown in Tabs. 1 and 2. The network trained with ℓ_1 loss shows superior performance compared to the one trained with ℓ_2 loss in all metrics. Previous works on various computer vision applications such as image super-resolution [6] and depth estimation [1] have demonstrated that the ℓ_1 loss favors results with sharp boundaries compared to the ℓ_2 loss. We also presume that the ℓ_1 loss favors a sharp depth output, and it leads to better performance compared to that of ℓ_2 loss.

Besides, the combination of ℓ_1 and ℓ_2 losses is tested, however, there is no performance improvement on the NYUv2 dataset [10]. Therefore, we adopt the ℓ_1 loss as our loss function for the training on the NYUv2 dataset [10]. On the contrary, we have empirically found that the combination of the ℓ_1 and ℓ_2 losses shows better performance in RMSE compared to that of the ℓ_1 loss alone for the KITTI DC dataset [11]. Therefore, we adopt the combination of the ℓ_1 and ℓ_2 losses as our loss function to train the proposed approach on the KITTI DC dataset.

Loss	RMSE (m)	REL	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
ℓ_2	0.096	0.014	99.5	**99.9**	**100.0**
ℓ_1	**0.092**	**0.012**	**99.6**	**99.9**	**100.0**
$\ell_1 + \ell_2$	0.093	**0.012**	**99.6**	**99.9**	**100.0**

Table 1. Quantitative evaluation on the NYUv2 dataset [10]. Our network is trained with different loss functions. Note that the training setup for this table is same as that of Sec. 6.1. in the main paper (*i.e.*, training on the full dataset).

Loss	RMSE (mm)	MAE	iRMSE	iMAE
ℓ_2	903.9	240.4	2.9	1.0
ℓ_1	898.4	**197.4**	**2.4**	**0.8**
$\ell_1 + \ell_2$	**884.1**	225.0	2.6	0.9

Table 2. Quantitative evaluation on the KITTI DC validation set [11]. Our network is trained with different loss functions.

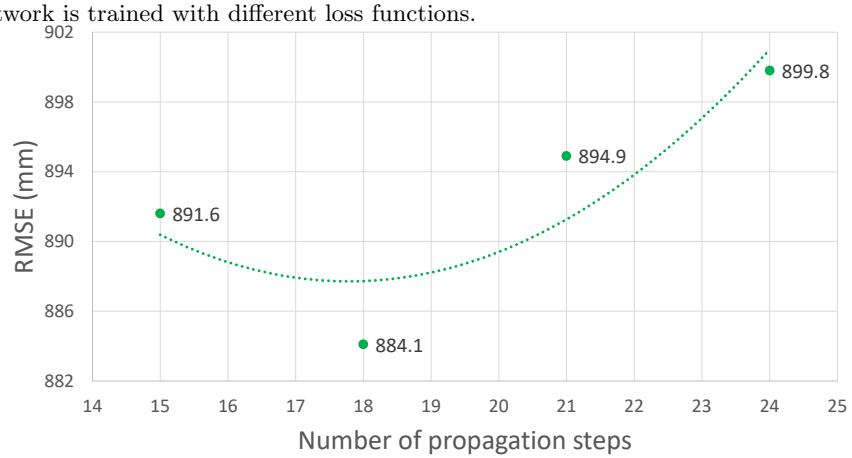


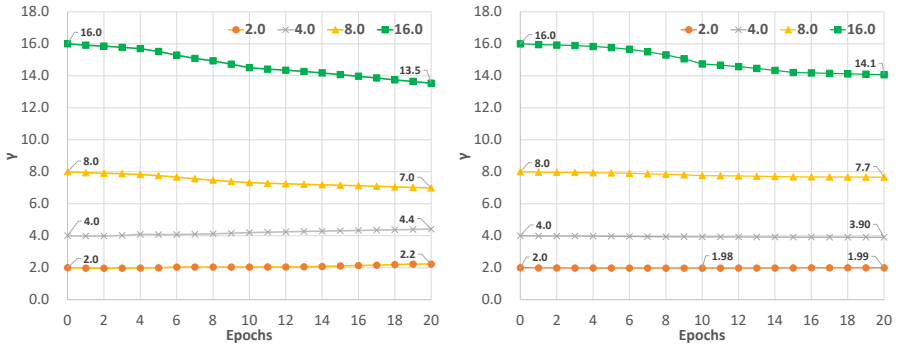
Fig. B. Performance comparison with various number of propagation iterations. A quadratic trend line (dotted green line) is shown together.

3.2 Number of the Propagation Iterations

To verify the relationship between the number of propagation iterations and performance, our network is trained with different number of propagation iterations, denoted by N_p. Figure B shows the performance with various $N_p \in \{15, 18, 21, 24\}$. The network trained with $N_p = 18$ shows the best performance. Therefore, we adopt $N_p = 18$ for our network empirically.

4 Additional Analyses on the Affinity Normalization

In order to analyze the proposed `Tanh`$-\gamma-$`Abs`$-$`Sum`*, we have trained our network on NYUv2 [10] and KITTI DC [11] datasets with various initial values of γ. For this ablation study on the NYUv2 dataset, we randomly sampled 10K images from the NYUv2 dataset for training, and evaluate the performance on the full validation dataset.



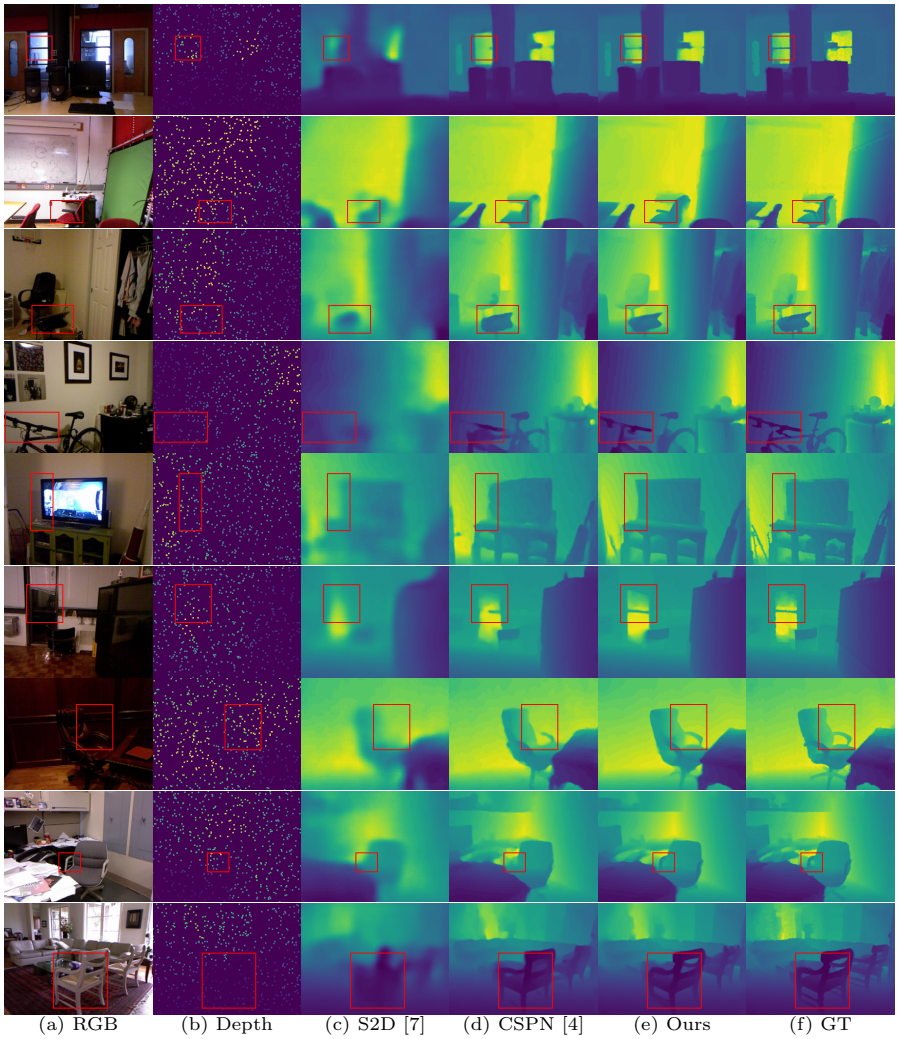
(a) NYUv2 [10] (b) KITTI DC [11]

Fig. C. The learned γ values during training with different initial values on the NYUv2 and KITTI DC datasets . Note that K is set to 8.

Figure C shows the γ values during training with different initial values. Note that if $\gamma < 1$, the normalization is close to the `Abs−Sum`*, and if $\gamma = K$, the normalization is same as `Tanh−C`.

The proposed normalization converges to neither `Abs−Sum`* (*i.e.*, $\gamma > 1$) nor `Tanh−C` (*i.e.*, $\gamma \neq K$) for both of datasets with various initial γ values. In addition, we can observe that the convergence range of γ value is different for each dataset, and this supports our assumption that the optimal γ value for the given task should be determined by the learning process.

5 Additional Results on the NYUv2 Dataset [10]



(a) RGB (b) Depth (c) S2D [7] (d) CSPN [4] (e) Ours (f) GT

Fig. D. Additional depth completion results on the NYUv2 dataset [10]. Note that sparse depth images are dilated for visualization.

Figure D shows additional depth completion results on the NYUv2 dataset [10]. Compared to the others (Fig. D(c) and (d)), our algorithm generates sharp depth results near the object boundaries.

6 Additional Results on the KITTI DC Dataset [11]

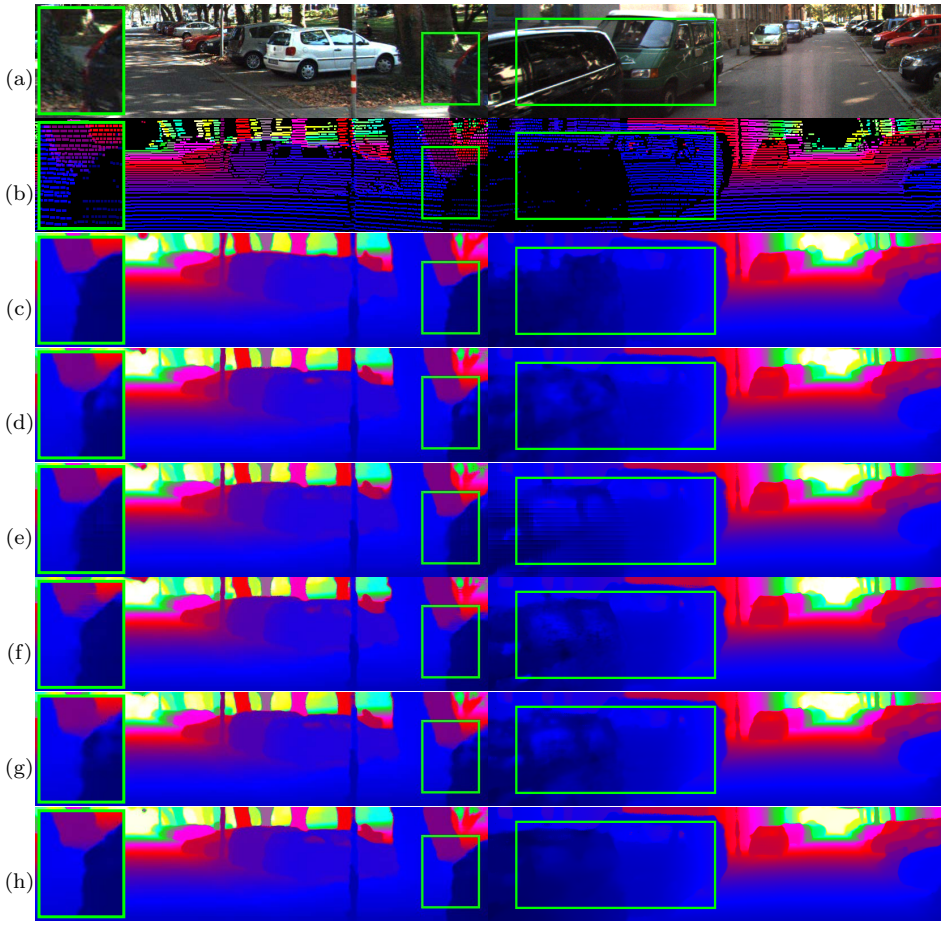


Fig. E. Additional depth completion results on the KITTI DC dataset [11]. (a) RGB, (b) Sparse depth, (c) CSPN [4], (d) DepthNormal [12], (e) DeepLiDAR [8], (f) FuseNet [2], (g) CSPN++ [3], (h) Ours. Note that sparse depth images are dilated for visualization.

Figures E and F show additional depth completion results on the KITTI DC dataset [11]. Results from the proposed algorithm (Figs. E(h) and F(h)) show better results especially on small or tiny objects compared to those of the others.

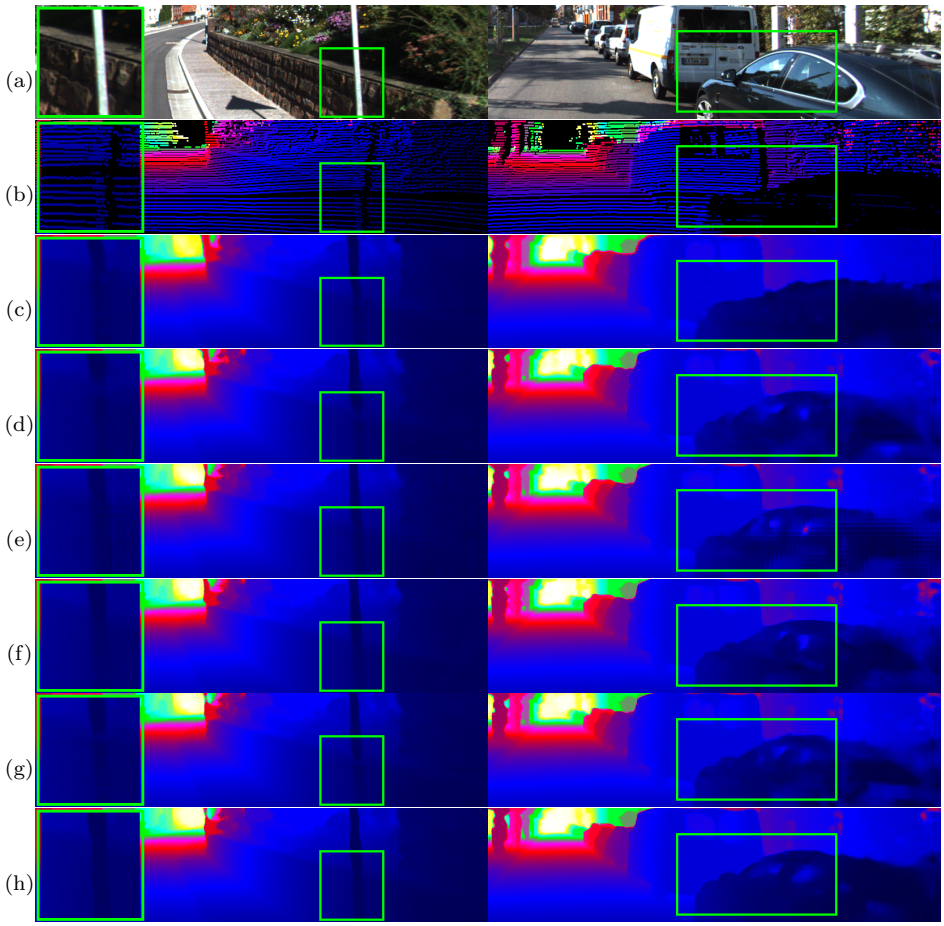


Fig. F. Additional depth completion results on the KITTI DC dataset [11]. (a) RGB, (b) Sparse depth, (c) CSPN [4], (d) DepthNormal [12], (e) DeepLiDAR [8], (f) FuseNet [2], (g) CSPN++ [3], (h) Ours. Note that sparse depth images are dilated for visualization.

References

1. Carvalho, M., Le Saux, B., Trouvé-Peloux, P., Almansa, A., Champagnat, F.: On regression losses for deep depth estimation. In: Proc. of IEEE Int'l Conf. on Image Processing (ICIP) (2018)
2. Chen, Y., Yang, B., Liang, M., Urtasun, R.: Learning joint 2d-3d representations for depth completion. In: Proc. of IEEE Int'l Conf. on Computer Vision (ICCV) (2019)
3. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: Proc. of AAAI Conf. on Artificial Intelligence (AAAI) (2020)
4. Cheng, X., Wang, P., Yang, R.: Depth estimation via affinity learned with convolutional spatial propagation network. In: Proc. of European Conf. on Computer Vision (ECCV) (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
7. Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: Proc. of IEEE Int'l Conf. on Robotics and Automation (ICRA) (2018)
8. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. of Int'l Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2015)
10. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proc. of European Conf. on Computer Vision (ECCV) (2012)
11. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant CNNs. In: Int'l Conf. on 3D Vision (3DV) (2017)
12. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: Proc. of IEEE Int'l Conf. on Computer Vision (ICCV) (2019)