**Supplementary Material for**

# CostDCNet: Cost Volume based Depth Completion of Single RGB-D Images

Jaewon Kam, Jungeon Kim, Soongjin Kim,
Jaesik Park, and Seungyong Lee

POSTECH
{jwkam95, jungeonkim, kimsj0302, jaesik.park, leesy}@postech.ac.kr

In this supplementary document, we first present the details of our network architecture (Section 1) and then provide additional experiments: analysis for number of depth planes (Section 2), robustness to noisy input (Section 3), stereo matching framework for depth completion (Section 4), and additional qualitative results (Section 5). Please refer to the supplementary video for results of depth sequences.

## 1 Network Architecture

Our framework is composed of three networks: 2D and 3D encoders, and a 3D UNet. All three networks are based on convolutional neural networks. The 2D encoder consists of 2D convolutional layers. For computational efficiency, we apply sparse 3D convolutions [4] and pseudo-3D convolutions [7] to the 3D encoder and 3D UNet, respectively. The network architectures are illustrated in Figure 1.

## 2 Analysis on Number of Depth Planes

We measured RMSE, runtime, and memory footprint according to the number of depth planes $K$ (Table 1). While the memory footprint and runtime increase as $K$ increases, RMSE does not improve anymore when $K$ is larger than 16. Therefore, we set $K$ to 16.

## 3 Robustness to Noisy Input

We compare our methods (Types A and C) and NLSPN [6] (2D CNN-based SOTA) on robustness to input depth noise. We simulate noisy inputs by adding zero-mean Gaussian noise to the input depth samples (the higher the sigma value $\sigma$, the stronger the noise). When the input noise increases, our methods show better performance compared to NLSPN (Table 2). This result indicates that our 3D CNN-based approach is more robust to noise. In addition, Type A achieves better performance than Type C on noisy inputs, because Type C assigns image features only to noisy input depth positions.

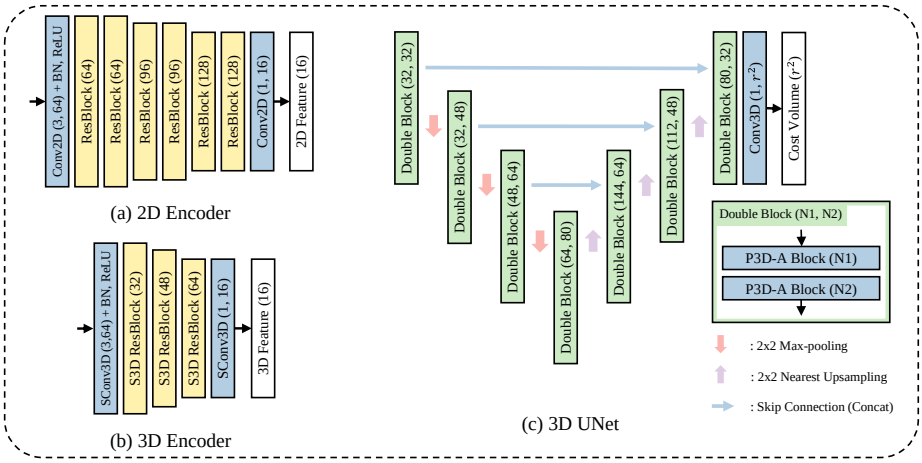



**Fig. 1.** Detailed network architectures for individual components (2D and 3D encoders, and the 3D UNet). Each ResBlock [5] and P3D-A Block [7] are characterized by channel dimensions, and each convolutional layer is characterized by kernel size and channel dimensions.

**Table 1.** Analysis on the number of depth planes on NYUv2 dataset.

| K | 4 | 8 | 16 | 32 | 64 | 96 |
|---|---|---|---|---|---|---|
| RMSE (m) | 0.104 | 0.098 | **0.096** | 0.097 | 0.097 | 0.097 |
| Runtime (ms) | **14** | **14** | 15 | 17 | 22 | 32 |
| Memory (MiB) | **16** | 20 | 26 | 54 | 79 | 104 |

## 4 Stereo Matching Framework for Depth Completion

Our *CostDCNet* infers the cost volume before determining the final completed depth. Therefore, we can naïvely think of direct adaptation of the existing stereo matching frameworks for depth completion task. To be specific, we can remove the building cost volume part of a stereo matching framework and replace it with cost-volume construction of our *CostDCNet*. To compare the merged framework and *CostDCNet*, we chose the BGNet [9] among several stereo matching frameworks because this is one of the most efficient state-of-the-art frameworks.

We initialize our *CostDCNet* part with pre-trained weights, then train the merged framework. Table 3 shows the quantitative results of the merged framework and ours. The merged framework shows improvements on both RMSE and REL metrics. However, these performance gains seem marginal, considering both considerable increase in network parameters and inference time. Consequently, we do not adopt the merged framework as our final model because we focus on designing a fast and lightweight architecture.

**Table 2.** Robustness comparison on NYUv2 dataset (Measure: RMSE(m)).

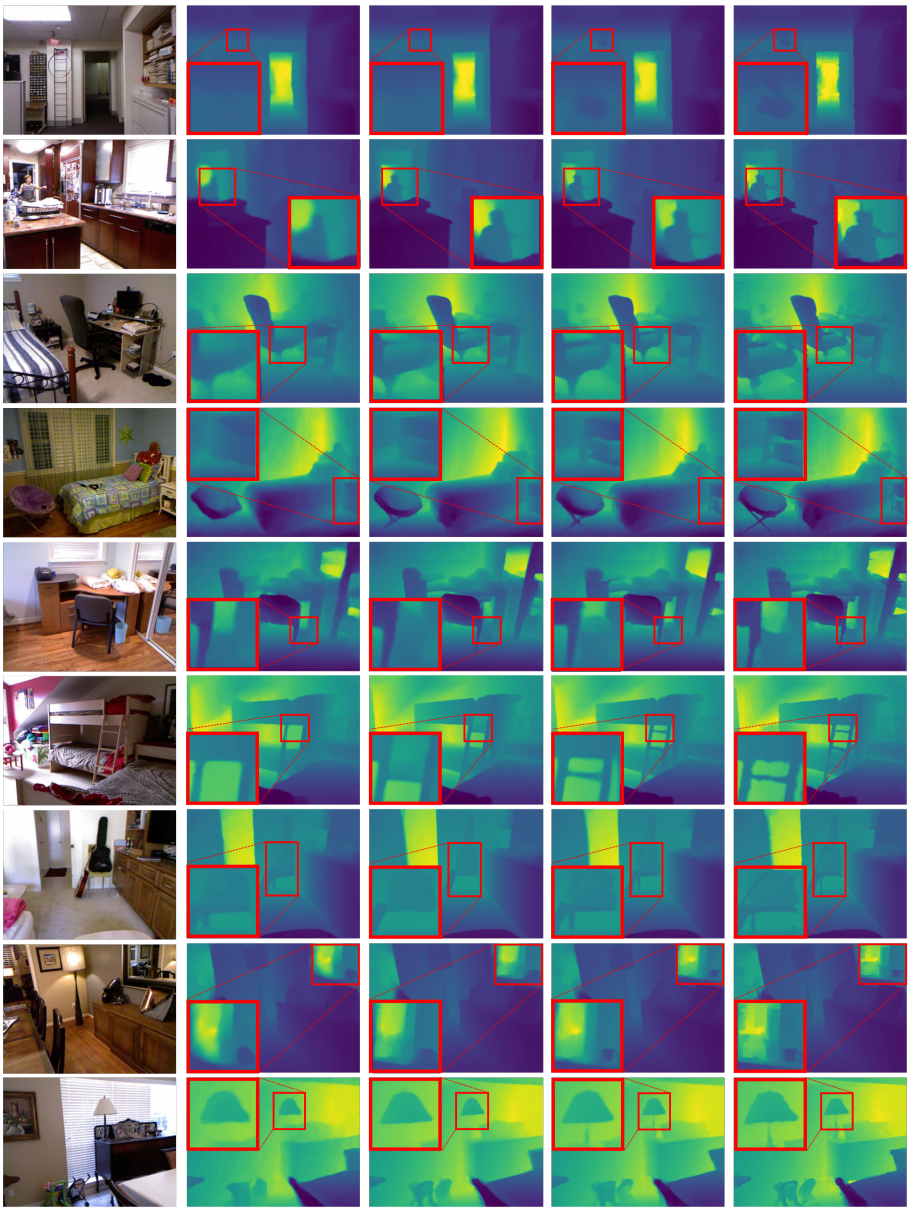
| Method | $\sigma$ (m) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| NLSPN [6] | **0.092** | **0.132** | 0.202 | 0.273 | 0.345 | 0.416 |
| Ours (Type A) | 0.097 | 0.134 | **0.197** | **0.265** | **0.333** | **0.405** |
| Ours (Type C) | 0.096 | 0.133 | 0.198 | 0.269 | 0.338 | 0.407 |

**Table 3.** Quantitative comparisons with the merged framework for sparse depth completion on the NYUv2 dataset.

| Method | #params↓ | Infer. time↓ | RMSE(m) ↓ | Rel(m)↓ | $\delta_{1.25}$ ↑ | $\delta_{1.25^2}$ ↑ | $\delta_{1.05^3}$ ↑ |
|---|---|---|---|---|---|---|---|
| Ours | **1.8M** | **16ms** | 0.96 | 0.013 | 99.5 | 99.9 | 100 |
| Merged | 5.4M | 30ms | **0.93** | **0.012** | 99.5 | 99.9 | 100 |

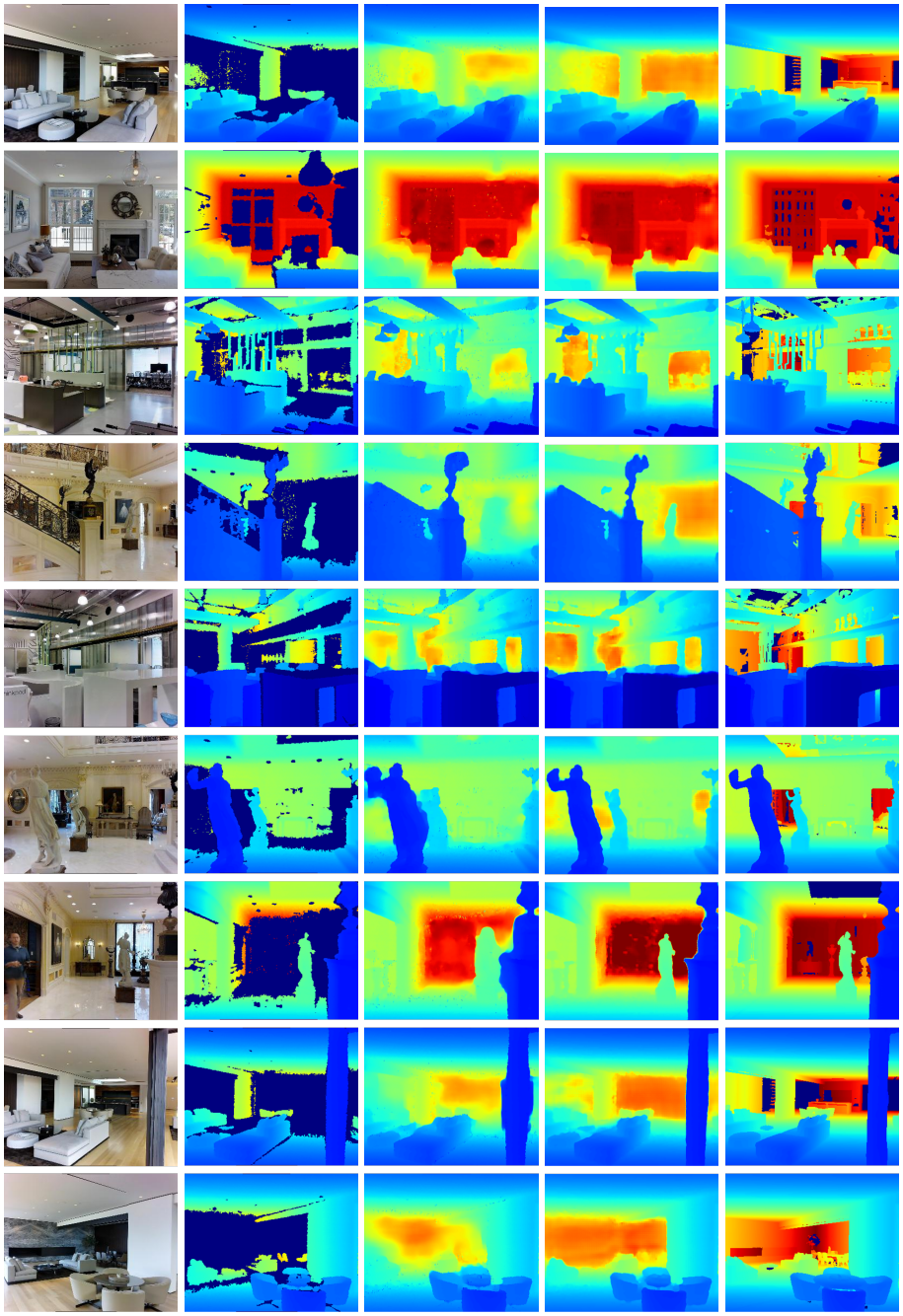
## 5 Additional Qualitative Results

We present additional qualitative results on the NYUv2 dataset [8] (Figure 2) and the Matterport3D dataset [2] (Figure 3). For the NYUv2 dataset, we use the sparse depth setting while the semi-dense one is used for the Matterport dataset. As shown in Figures 2 and 3, our *CostDCNet* overall shows better reconstruct structures from input RGB-D images on various scenes.



(a) RGB (b) CSPN [3] (c) NLSPN [6] (d) Ours (e) GT

**Fig. 2.** Additional qualitative comparisons with other methods on the NYUv2 dataset.



(a) RGB (b) Raw Depth (c) Cao [1] (d) Ours (e) GT

**Fig. 3.** Additional qualitative comparisons with Cao et al. [1] on the Matterport3D dataset.

## References


1. Cao, Z., Li, A., Yuan, Z.: Self-supervised depth completion via adaptive sampling and relative consistency. In: Proceedings of IEEE International Conference on Image Processing (ICIP). pp. 3263–3267 (2021) 5
2. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. arXiv preprint arXiv:1709.06158 (2017) 3
3. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(10), 2361–2379 (2019) 4
4. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR). pp. 3075–3084 (2019) 1
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) 2
6. Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 120–136. Springer (2020) 1, 3, 4
7. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). pp. 5533–5541 (2017) 1, 2
8. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 746–760. Springer (2012) 3
9. Xu, B., Xu, Y., Yang, X., Jia, W., Guo, Y.: Bilateral grid learning for stereo matching networks. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR). pp. 12497–12506 (2021) 2