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.
4 Stereo Matching Framework for Depth Completion

Our CostDCNet infers the cost volume before determining the final completed depth. Therefore, we can naively 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)).

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLSPN [6]</td>
<td>0.092</td>
<td>0.132</td>
<td>0.202</td>
<td>0.273</td>
<td>0.345</td>
<td>0.416</td>
</tr>
<tr>
<td>Ours (Type A)</td>
<td>0.097</td>
<td>0.134</td>
<td>0.197</td>
<td>0.265</td>
<td>0.333</td>
<td>0.405</td>
</tr>
<tr>
<td>Ours (Type C)</td>
<td>0.096</td>
<td>0.133</td>
<td>0.198</td>
<td>0.269</td>
<td>0.338</td>
<td>0.407</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>#params</th>
<th>Infer. time</th>
<th>RMSE(m)</th>
<th>Rel(m)</th>
<th>δ_1.25↑</th>
<th>δ_1.25↓</th>
<th>δ_1.05↑</th>
<th>δ_1.05↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>1.8M</td>
<td>16ms</td>
<td>0.96</td>
<td>0.013</td>
<td>99.5</td>
<td>99.7</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Merged</td>
<td>5.4M</td>
<td>30ms</td>
<td>0.93</td>
<td>0.012</td>
<td>99.5</td>
<td>99.9</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

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.
Fig. 2. Additional qualitative comparisons with other methods on the NYUv2 dataset.
Fig. 3. Additional qualitative comparisons with Cao et al. [1] on the Matterport3D dataset.
References