Radar-Camera Pixel Depth Association for Depth Completion – Supplementary Material –

Abstract

In this supplementary material, we show the details of the U-Net we use in our experiment. As ablation studies, we investigate how neighborhood size and PDA thresholds influence depth estimation performance. We also provide visualization of MER for three samples.

1. Network Details

Fig. 1 shows a typical structure of U-Net [2] we use for both Stages 1 and architecture [1] for Stage 2. It has five levels of resolutions where downsampling and upsampling are achieved by max-pooling and nearest neighbor interpolation. Except the input and output, all tensors in the network have the same number of channels N_c. All convolutions use 3×3 filters. B_0 represents the input block, a convolution layer that increases the number of channels to N_c. B_1 denotes residual blocks connected in series and B_2 is a sequence of convolutional blocks where each block consists of batch normalization, ReLU and convolution. B_3 is the output block which includes batch normalization, ReLU and a convolutional layer changing the number of channels from N_c to that of output channels.

In our experiment, for Stage 1 network, we set N_c to 80. There are two residual blocks in B_1 and four convolutional blocks in B_2, respectively. For Stage 2 network, N_c is set to 64. B_1 has four residual blocks and B_2 has 8 convolutional blocks. We use RMSProp as optimizer with a learning rate of 5×10^{-5}.

Inference time When we use Intel Core i7-8700 CPUs and an NVIDIA GeForce RTX 2080 Ti GPU, the inference times of Network 1 and Network 2 are 4.5 ms and 8.2 ms per frame, respectively.

2. Ablation Studies

2.1. Neighborhood Sizes

The neighborhood is a rectangular region around a radar pixel, which is defined as the neighborhood center. A neighborhood is designed to let the radar pixel have more space to expand upwards, and thus, typically, the neighborhood center does not correspond to the geometric center. As shown in Fig. 2, we use (w_1, w_2, h_1, h_2) to represent the shape of a neighborhood, where w_1 and w_2 are the number of horizontal neighboring pixels to the left and right of the neighborhood center and h_3 and h_4 denote the number of vertical neighboring pixels above and below the center.

Neighborhood	MAE	Abs Rel	RMSE	RMSE log
$(2, 2, 30, 5)$	**1.487**	**0.085**	**3.210**	**0.145**
$(3, 3, 20, 5)$	1.497	0.086	**3.210**	0.146
$(1, 1, 8, 1)$	1.536	0.088	3.274	0.148

Table 1: Full-image depth estimation/completion errors (m).

PDA Thresholds	MAE	Abs Rel	RMSE	RMSE log
$(0.6 : 0.1 : 0.9, 0.95)$	1.487	0.085	3.210	0.145
$(0.5 : 0.1 : 0.9, 0.95)$	**1.472**	0.085	**3.179**	**0.144**
$(0.6 : 0.05 : 0.9, 0.95)$	1.494	**0.084**	3.229	0.146

Table 2: Full-image depth estimation/completion errors with MER created by different thresholds (m). In the first column, "a:s:b" indicates a set of thresholds $a, a + s, a + 2s, \cdots, b$.

As shown in Table 1, we test different neighborhood sizes for Stage 1 network and compute their resultant depth error after using them in Stage 2 network [1]. The T_l to create MER in Stage 2 are $(0.6, 0.7, 0.8, 0.9, 0.95)$.

2.2. PDA Thresholds for MER

MER can be generated with different PDA threshold T_l where $l = 1, 2, ..., N_e$. With fixed neighorbood size $(2, 2, 30, 5)$, we test different PDA thresholds and compute their resultant depth error. Results are shown in Table 2.

3. Visualization of MER

Fig. 3 shows MER created by PDA thresholds of 0.6, 0.7, 0.8, 0.9 and 0.95, respectively. We can see the expanded region decreases as PDA threshold increases.

References

[1] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation*, pages 4796–4803, 2018. 1

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmen-

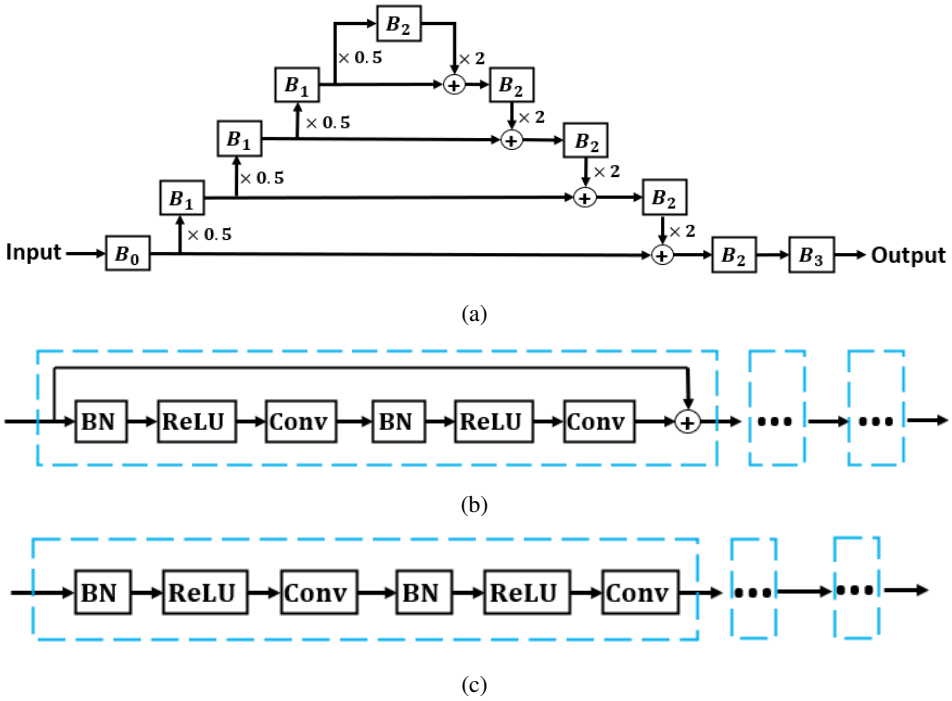

Figure 1: (a) U-Net used in our experiment. (b) Diagram of Block B_1, a series of residual blocks (shown as blue boxes). (c)Diagram of block B_2, a sequence of convolutional blocks (shown as blue box). BN and $Conv$ denote batch normalization and a convolutional layer, respectively.

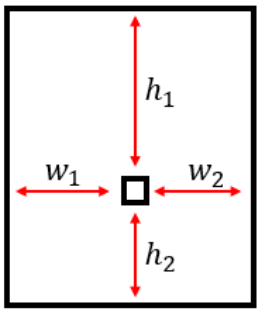

Figure 2: A neighborhood with shape (w_1, w_2, h_1, h_2). The small square represents a radar pixel.

tation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. 1

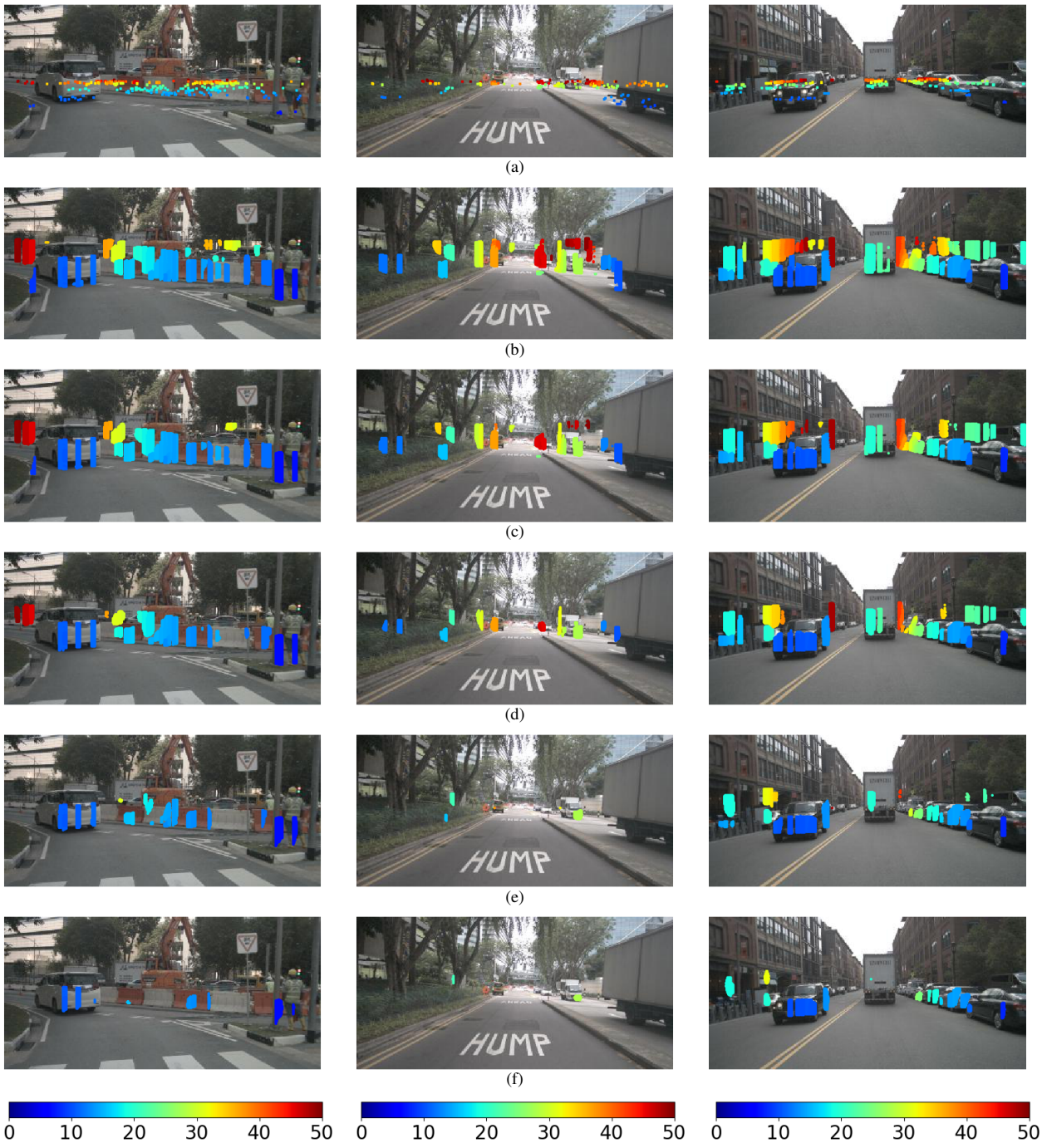

Figure 3: (a) Raw radar depth. MER with PDA thresholds (b) 0.6, (c) 0.7, (d) 0.8, (e) 0.9 and (f) 0.95. Occluded radar points are filtered out and corrected depth are expanded in MER.