Supplementary Material for Bi-directional Recurrent MVSNet

Taku Fujitomi, Seiya Ito, Naoshi Kaneko and Kazuhiko Sumi
Aoyama Gakuin University
{fujitomi.taku, ito.seiya}@vss.it.aoyama.ac.jp, {kaneko, sumi}@it.aoyama.ac.jp

1 Evaluation Metrics

In our paper, distance metric and percentage metric are used as evaluation metrics.

The distance metric consists of accuracy, completeness, and overall. The reconstructed point cloud is the set $\mathcal{R}$ and the ground truth point cloud is the set $\mathcal{G}$. Accuracy, completeness, and overall are formulated as:

$$Accuracy = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} e_{r \to \mathcal{G}} \tag{1}$$

$$Completeness = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} e_{g \to \mathcal{R}} \tag{2}$$

$$Overall = \frac{Accuracy + Completeness}{2} \tag{3}$$

where $e_{r \to \mathcal{G}}$ and $e_{g \to \mathcal{R}}$ are the distance from the reconstructed point r to the ground truth point cloud $\mathcal{G}$ and the distance from the ground truth point g to the reconstructed point cloud $\mathcal{R}$, respectively. Each distance are defined as:

$$e_{r \to \mathcal{G}} = \min_{g \in \mathcal{G}} \|r - g\| \tag{4}$$

$$e_{g \to \mathcal{R}} = \min_{r \in \mathcal{R}} \|g - r\| \tag{5}$$

The percentage metric consists of precision, recall, and F-score, which are defined as:

$$Precision = \frac{100}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} [e_{r \to \mathcal{G}} < d] \tag{6}$$

$$Recall = \frac{100}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} [e_{g \to \mathcal{R}} < d] \tag{7}$$

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

where $[\cdot]$ is the Iverson bracket, which is 1 if the proposition is true, and 0 if it is false. d is a pre-defined threshold. The percentage metric evaluates the percentage of points with errors less than d.

2 Experiments

2.1 Trainning on BlendedMVS

The BlendedMVS [1] is a multi-view stereo dataset consisting of 113 scenes and about 17,000 images in total. In the process of creating the dataset, a mesh model is first generated from multi-view images, and then a rendered image and its depth map are generated from the same viewpoint as the images. The high-frequency components of the generated images are combined with the low-frequency components of the original images to produce a training image that incorporates the effects of ambient light while maintaining the accurate depth maps and camera parameters.

2.2 Implementation

The number of input images N is 3, and the image resolution is 768×576. The depth sample number D is set to 128. The learning rate is set to 0.001, which is multiplied by 0.9 for every $10,000$ steps. The batch size is set to 1. We trained our model on an NVIDIA TITAN RTX up to 6 epochs.

2.3 Tanks and Temples Benchmark

The evaluation results on the Tanks and Temples Benchmark [2] are shown in Table 1. The proposed method outperformed the F-score of the comparison method in all scenes.

3 Network Architecture

The detailed architecture of our proposed network is shown in Table 2. We used the same feature extractor for our method and R-MVSNet+DC to fairly compare the regularization of both methods.

References

[1] Y. Yao, et al.: "BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks" *Computer Vision and Pattern Recognition*, pp.1787–1796, 2020.

[2] K. Arno, et al.: "Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction" *ACM Transactions on Graphics*, vol.36, no.4, pp.1–13, 2017.

[3] J. Yan, et al.: "Dense Hybrid Recurrent Multi-view Stereo Net with Dynamic Consistency Checking" *European Conference on Computer Vision*, vol.12349, pp.674–689, 2020.

[4] Y. Yao, et al.: "Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference" *Computer Vision and Pattern Recognition*, pp.5525–5534, 2019.

Table 1. Quantitative results on the Tanks and Temples Benchmark [2] for models trained with the BlendedMVS [1]. We implemented R-MVSNet [4] using Dynamic Consistency Checking [3] and denoted it as R-MVSNet+DC. L.H. and P.G. are abbreviations for Lighthouse and Playground, respectively.

	Percentage Metric (%)								
	Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
R-MVSNet+DC	55.22	70.71	49.56	42.98	58.90	57.09	54.69	59.43	48.44
Ours	**56.89**	**72.42**	**50.93**	**44.58**	**60.78**	**58.95**	**56.43**	**59.84**	**51.20**

Table 2. The detailed architecture of our proposed network. We denote the 2D convolution as Conv and the 2D deconvolution as Deconv. GR is an abbreviation for the Group normalization and the Relu. K is the kernel size, S is the stride, and N, H, W, and D are the number of input images, image height, width, and depth sample number, respectively.

Input	Layer	Output	Output Size
Image Feature Extraction			
		$\{\mathbf{I}_i\}_{i=0}^{N-1}$	$H \times W \times 3$
$\mathbf{I}_i$	ConvGR, K=3 × 3, S=1	2D0_1	$H \times W \times 8$
2D0_1	ConvGN, K=3 × 3, S=1	2D0_2	$H \times W \times 8$
$\mathbf{I}_i$	ConvGR, K=3 × 3, S=2	2D1_0	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D1_0	ConvGR, K=3 × 3, S=1	2D1_1	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D1_1	ConvGR, K=3 × 3, S=1	2D1_2	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D1_0	ConvGR, K=3 × 3, S=2	2D2_0	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D2_0	ConvGR, K=3 × 3, S=1	2D2_1	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D2_1	ConvGR, K=3 × 3, S=1	2D2_2	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D2_0	ConvGR, K=3 × 3, S=2	2D3_0	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
2D3_0	ConvGR, K=3 × 3, S=1	2D3_1	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
2D3_1	ConvGR, K=3 × 3, S=1	2D3_2	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
2D3_0	ConvGR, K=3 × 3, S=2	2D4_0	$\frac{1}{16}H \times \frac{1}{16}W \times 128$
2D4_0	ConvGR, K=3 × 3, S=1	2D4_1	$\frac{1}{16}H \times \frac{1}{16}W \times 128$
2D4_1	ConvGR, K=3 × 3, S=1	2D4_2	$\frac{1}{16}H \times \frac{1}{16}W \times 128$
2D4_2	DeconvGR, K=3 × 3, S=2	2D5_0	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
[2D3_2, 2D5_0]	ConvGR, K=3 × 3, S=1	2D5_1	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
2D5_1	ConvGR, K=3 × 3, S=1	2D5_2	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
2D5_2	DeconvGR, K=3 × 3, S=2	2D6_0	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
[2D2_2, 2D6_0]	ConvGR, K=3 × 3, S=1	2D6_1	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D6_1	ConvGR, K=3 × 3, S=1	2D6_2	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D6_2	DeconvGR, K=3 × 3, S=2	2D7_0	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
[2D1_2, 2D7_0]	ConvGR, K=3 × 3, S=1	2D7_1	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D7_1	ConvGR, K=3 × 3, S=1	2D7_2	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D7_2	DeconvGR, K=3 × 3, S=2	2D8_0	$H \times W \times 8$
[2D0_2, 2D8_0]	ConvGR, K=3 × 3, S=1	2D8_1	$H \times W \times 8$
2D8_1	ConvGR, K=3 × 3, S=1	2D8_2	$H \times W \times 8$
2D8_2	ConvGR, K=3 × 3, S=2	2D9_0	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D9_0	ConvGR, K=3 × 3, S=1	2D9_1	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D9_1	ConvGR, K=3 × 3, S=1	2D9_2	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D9_2	ConvGR, K=3 × 3, S=2	2D10_0	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D10_0	ConvGR, K=3 × 3, S=1	2D10_1	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D10_1	Conv, K=3 × 3, S=1	$\mathbf{F}_i$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
Homography Warping & Cost Metric			
$\{\mathbf{F}_i\}_{i=0}^{N-1}, d$	Differentiable Homography Warping	$\{\mathbf{V}_i(d)\}_{i=0}^{N-1}$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
$\{\mathbf{V}_i(d)\}_{i=0}^{N-1}$	Variance Cost Metric	$\mathbf{C}(d)$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
Bi-directional GRUs Regularization			
$\mathbf{C}(d)$	GRU, K=3 × 3, S=1	$\mathbf{C}_0^{f,b}(d)$	$\frac{1}{4}H \times \frac{1}{4}W \times 16$
$\mathbf{C}_1^{f,b}(d)$	GRU, K=3 × 3, S=1	$\mathbf{C}_2^{f,b}(d)$	$\frac{1}{4}H \times \frac{1}{4}W \times 4$
$\mathbf{C}_2^{f,b}(d)$	GRU, K=3 × 3, S=1	$\mathbf{C}_r^{f,b}(d)$	$\frac{1}{4}H \times \frac{1}{4}W \times 2$
$[\mathbf{C}_r^f(d), \mathbf{C}_r^b(d)]$	Conv, K=3 × 3, S=1	$\mathbf{C}_r(d)$	$\frac{1}{4}H \times \frac{1}{4}W \times 1$
Probability Volume Construction			
$\mathbf{C}_r(d)\}_{d=0}^{D-1}$	Softmax	$\mathbf{P}$	$D \times \frac{1}{4}H \times \frac{1}{4}W \times 1$