Supplementary Materials for
Fast-MVSNet: Sparse-to-Dense Multi-View Stereo With Learned Propagation and Gauss-Newton Refinement

Zehao Yu
ShanghaiTech University
yuzh@shanghaitech.edu.cn

Shenghua Gao
ShanghaiTech University
gaoshh@shanghaitech.edu.cn

1. Architecture

As presented in the main paper, our Fast-MVSNet has three parts: sparse high-resolution depth map prediction, depth map propagation, and Gauss-Newton refinement. For the sparse high-resolution depth map prediction, our network is similar to MVSNet [4] except that we build a sparse cost volume in spatial domain and use fewer virtual depth planes (*e.g.*, 96). Therefore, we can obtain a sparse high-resolution depth map at much lower cost. For the depth map propagation module, we use a 10-layer convolutional network to prediction the weights W. We show the details of this network in Table 1. For the Gauss-Newton refinement, we use a similar network architecture as propagation module to extract deep feature representations of the input images $\{I_i\}_{i=0}^N$. In particular, Conv_4 and Conv_7 as in Table 1 are first interpolated to the same size and then are concatenated as the deep feature representation.

Name	Layer	Output Size
Input		H×W×3
Conv_0	ConvBR,K=3x3,S=1,F=8	H×W×8
Conv_1	ConvBR,K=3x3,S=1,F=8	H×W× 8
Conv_2	ConvBR,K=5x5,S=2,F=16	½H×½W×16
Conv_3	ConvBR,K=3x3,S=1,F=16	½H×½W×16
Conv_4	ConvBR,K=3x3,S=1,F=16	½H×½W×16
Conv_5	ConvBR,K=5x5,S=2,F=32	¼H×¼W×32
Conv_6	ConvBR,K=3x3,S=1,F=32	¼H×¼W×32
Conv_7	Conv,K=3x3,S=1,F=32	¼H×¼W×32
Conv_8	Conv,K=3x3,S=1,F=16	¼H×¼W×16
W	Conv,K=3x3,S=1,F=k^2	¼H×¼W× k^2

Table 1: Weights prediction network in the propagation module. We denote the 2D convolution as Conv and use BR to abbreviate the batch normalization and the Relu. K is the kernel size, S the kernel stride and F the output channel number. H, W denote image height and width, respectively.

2. Depth maps fusion

The fusion has three steps: photometric filtering, geometric consistency, and depth fusion. For photometric filtering, we first interpolate the predicted probability of the sparse high-resolution depth map to a high-resolution probability map and filter out points whose probability is below a threshold. The filtering threshold is set to 0.5. For geometric consistency, we compute the discrepancy of each depth map and filter out points whose discrepancy is larger than a threshold η. Specifically, a point p in reference dpeth map D is first projected to p' in the neighboring depth map $\hat{D}$, then the discrepancy is defined as $f \cdot baseline \cdot \|\frac{1}{D(p)} - \frac{1}{\hat{D}(p')}\|$, where f is the focal length of reference image and $baseline$ is the baseline of two images. The threshold η is set to 0.12 pixels. For depth fusion, we require each point to be visible in $V = 3$ views and take the average value of all reprojected depths.

In the main paper, for a fair comparison, we use the same parameters for depth map fusion as that in Point-MVSNet [2]. However, we find that the fusion parameters η and V have a significant impact on reconstruction results. We show the quantitative comparison of reconstructions with different η and V in Table 2. The comparison of visualization results are shown in Figure 1. From the comparison results, we can see the trade off between *Accuracy* and *Completeness*. Increasing η, the reconstructed points gets less accurate but more complete. Increasing V, the reconstructions become more accurate while become incomplete. As the fusion has significant impact on the final reconstruction results, integrating a learnable fusion module [3] into the overall pipeline will be an interesting direction in future work.

3. Gauss-Newton refinement with more iterations

In this section, we conduct ablation study for Gauss-Newton refinement with more iterations. As shown in

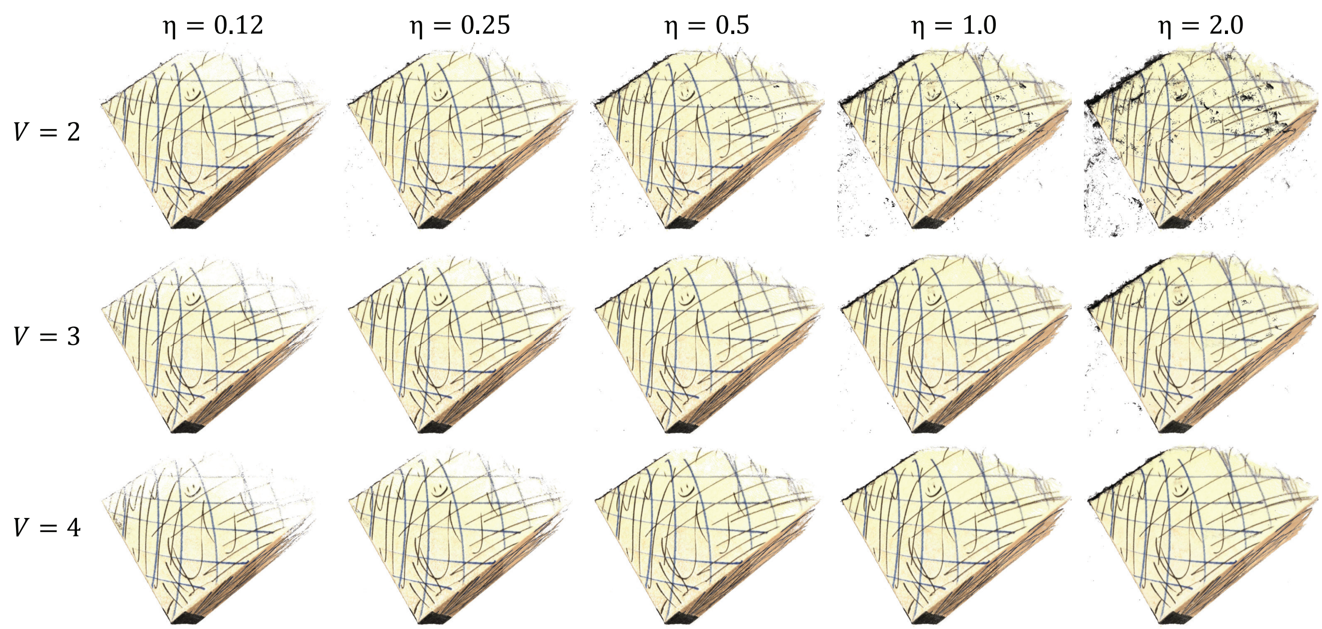

Figure 1: Reconstruction results of *scan10* on the DTU dataset [1] with different fusion parameters. η is the threshold of geometric consistency check. V is the number of views that a point should be visible. As η increases, the reconstruction becomes denser while has more noise. As V increases, the reconstruction becomes cleaner while also becomes sparser.

η	V	Acc. (mm)	Comp. (mm)	Overall (mm)
0.12	2	0.3969	0.3140	**0.3555**
0.12	3	0.3360	0.4030	0.3695
0.12	4	**0.3007**	0.5212	0.4109
0.25	2	0.4663	0.2843	0.3753
0.25	3	0.3951	0.3341	0.3646
0.25	4	0.3542	0.3959	0.3750
0.5	2	0.5480	**0.2773**	0.4127
0.5	3	0.4614	0.3076	0.3845
0.5	4	0.4128	0.3447	0.3788
1.0	2	0.6655	0.2888	0.4772
1.0	3	0.5555	0.3091	0.4323
1.0	4	0.4923	0.3330	0.4126
2.0	2	0.8381	0.3187	0.5784
2.0	3	0.7002	0.3323	0.5163
2.0	4	0.6152	0.3500	0.4826

Table 2: Quantitative results of reconstruction quality on the DTU evaluation dataset [1]. Increasing the geometric consistency threshold η, the reconstruted points become less accurate but also become more complete. Increasing the number of visible views V, the reconstruction becomes accurate while also becomes incomplete.

Table 3, Gauss-Newton refinement can significantly improves the reconstruction quality. However, the performance improvements of applying Gauss-Newton refinement with more interations are marginal. Therefore, we only use one iteration in Gauss-Newton refinement.

# iterations	Acc. (mm)	Comp. (mm)	Overall (mm)
0	0.3679	0.4475	0.4077
1	**0.3360**	0.4030	0.3695
2	0.3391	0.3956	0.3673
3	0.3420	0.3902	0.3662
4	0.3435	0.3885	0.3660
5	0.3443	**0.3875**	**0.3659**

Table 3: Quantitative results of reconstruction quality on the DTU evaluation dataset [1] with different iteration number in Gauss-Newton refinement.

4. Reconstruction results

We show more reconstruction results on the DTU dataset [1] in Figure 2. Our reconstruction is dense and accurate for all scenes.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 2, 3

[2] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[3] Simon Donne and Andreas Geiger. Learning non-volumetric depth fusion using successive reprojections. In *The IEEE Con-*

Figure 2: Reconstruction results on the DTU dataset [1].

ference on Computer Vision and Pattern Recognition (CVPR), June 2019. 1

[4] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1