Supplementary Material for "MVSNet: Depth Inference for Unstructured Multi-view Stereo"

1 MVSNet Architecture

While in the main paper we have described the network design in Sec. 3, here we show the detailed architecture of MVSNet in Table 1.

Table 1: Detailed architecture of MVSNet, where K denotes the kernel size, S the kernel stride and F the output channel number. As in the main paper, N, H, W, D represents input view number, image width, height and depth sample number respectively

Ouput	Layer	Input	Output Size
$\{\mathbf{I}_i\}_{i=1}^N$			$N \times H \times W \times 3$
Image Features Extration			
2D_0	Conv2D+BN+ReLU, K=3x3, S=1, F=8	$\mathbf{I}_i$	$H \times W \times 8$
2D_1	Conv2D+BN+ReLU, K=3x3, S=1, F=8	2D_0	$H \times W \times 8$
2D_2	Conv2D+BN+ReLU, K=5x5, S=2, F=16	2D_1	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D_3	Conv2D+BN+ReLU, K=3x3, S=1, F=16	2D_2	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D_4	Conv2D+BN+ReLU, K=3x3, S=1, F=16	2D_3	$\frac{1}{2}H \times \frac{1}{2}W \times 16$
2D_5	Conv2D+BN+ReLU, K=5x5, S=2, F=32	2D_4	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D_6	Conv2D+BN+ReLU, K=3x3, S=1, F=32	2D_5	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
$\mathbf{F}_i$	Conv2D, K = 3x3, S=1, F=32	2D_6	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
Feature Volumes Construction			
$\{\mathbf{V}_i\}_{i=1}^N$	Differentiable Homography Warping	$\{\mathbf{F}_i\}_{i=1}^N$	$N \times \frac{1}{4}H \times \frac{1}{4}W \times D \times 32$
Cost Volume Construction			
$\mathbf{C}$	Variance-based Cost Metric	$\{\mathbf{V}_i\}_{i=1}^N$	$\frac{1}{4}H \times \frac{1}{4}W \times D \times 32$
Cost Volume Regularization			
3D_0	Conv3D+BN+ReLU, K=3x3x3, S=1, F=8	$\mathbf{C}$	$\frac{1}{4}H \times \frac{1}{4}W \times D \times 8$
3D_1	Conv3D+BN+ReLU, K=3x3x3, S=2, F=16	3D_0	$\frac{1}{8}H \times \frac{1}{8}W \times \frac{1}{2}D \times 16$
3D_2	Conv3D+BN+ReLU, K=3x3x3, S=1, F=16	3D_1	$\frac{1}{8}H \times \frac{1}{8}W \times \frac{1}{2}D \times 16$
3D_3	Conv3D+BN+ReLU, K=3x3x3, S=2, F=32	3D_2	$\frac{1}{16}H \times \frac{1}{16}W \times \frac{1}{4}D \times 32$
3D_4	Conv3D+BN+ReLU, K=3x3x3, S=1, F=32	3D_3	$\frac{1}{16}H \times \frac{1}{16}W \times \frac{1}{4}D \times 32$
3D_5	Conv3D+BN+ReLU, K=3x3x3, S=2, F=64	3D_4	$\frac{1}{32}H \times \frac{1}{32}W \times \frac{1}{8}D \times 64$
3D_6	Conv3D+BN+ReLU, K=3x3x3, S=1, F=64	3D_5	$\frac{1}{32}H \times \frac{1}{32}W \times \frac{1}{8}D \times 64$
3D_7	Deconv3D+BN+ReLU, K=3x3x3, S=2, F=32	3D_6	$\frac{1}{16}H \times \frac{1}{16}W \times \frac{1}{4}D \times 32$
3D_8	Addition	3D_7 + 3D_4	$\frac{1}{16}H \times \frac{1}{16}W \times \frac{1}{4}D \times 32$
3D_9	Deconv3D+BN+ReLU, K=3x3x3, S=2, F=16	3D_0	$\frac{1}{8}H \times \frac{1}{8}W \times \frac{1}{2}D \times 16$
3D_10	Addition	3D_9 + 3D_2	$\frac{1}{8}H \times \frac{1}{8}W \times \frac{1}{2}D \times 16$
3D_11	Deconv3D+BN+ReLU, K=3x3x3, S=2, F=8	3D_0	$\frac{1}{4}H \times \frac{1}{4}W \times D \times 8$
3D_12	Addition	3D_11 + 3D_0	$\frac{1}{4}H \times \frac{1}{4}W \times D \times 8$
$\mathbf{P}$	Conv3D, K = 3x3x3, S = 1, F = 1	3D_12	$\frac{1}{4}H \times \frac{1}{4}W \times D$
Depth Map Regression			
$\mathbf{D}_{init}$	Soft Argmin	$\mathbf{P}$	$\frac{1}{4}H \times \frac{1}{4}W \times 1$
Depth Map Refinement			
2D_Cat	Concatenation	$\mathbf{D}_{init}$, $\mathbf{I}_i$	$\frac{1}{4}H \times \frac{1}{4}W \times 4$
2D_7	Conv2D+BN+ReLU, K=3x3, S=1, F=32	2D_Cat	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D_8	Conv2D+BN+ReLU, K=3x3, S=1, F=32	2D_7	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
2D_9	Conv2D+BN+ReLU, K=3x3, S=1, F=32	2D_8	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
$\mathbf{D}_{res}$	Conv2D+BN+ReLU, K=3x3, S=1, F=1	2D_9	$\frac{1}{4}H \times \frac{1}{4}W \times 1$
$\mathbf{D}_{refined}$	Addition	$\mathbf{D}_{init}$, $\mathbf{D}_{res}$	$\frac{1}{4}H \times \frac{1}{4}W \times 1$

2 Benchmarks

2.1 DTU Dataset

Table 2: Evaluation results on the 22 evaluation scans. MVSNet with smaller input image size even perform better in accuracy and overall quality as the given ground truth point clouds are only partially complete

	Mean Acc.	Med. Acc.	Mean Comp.	Med. Comp.	Overall
Camp [2]	0.853	0.496	0.559	0.196	0.706
Furu [3]	0.613	0.324	0.941	0.464	0.777
Tola [8]	0.342	0.21	1.19	0.492	0.766
Gipuma [4]	0.283	0.201	0.873	0.313	0.578
SurfaceNet [5]	0.45	0.294	1.04	0.285	0.745
MVSNet (1600x1184)	0.396	0.267	**0.527**	0.282	0.462
MVSNet (1280x1024)	**0.371**	**0.248**	0.534	**0.277**	**0.443**

Quantitative Results The Matlab evaluation script provided by *DTU* dataset [1] measures the mean accuracy, medium accuracy, mean completeness and medium completeness. Here we list the full results on the evaluation set in Table 2. We notice that SurfaceNet use their own script for the evaluation, in this table we stick to the original evaluation code from *DTU* dataset.

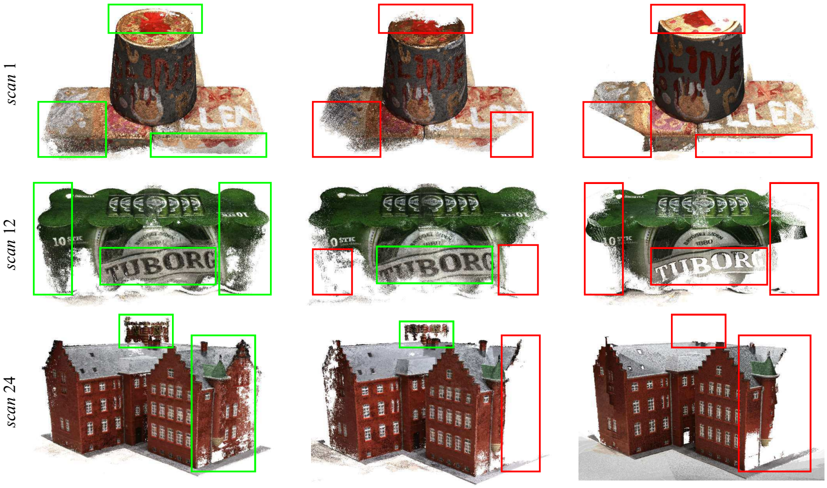
Accuracy vs. Completeness There is always a trade-off between the accuracy and the completeness in MVS reconstruction. As for our MVSNet, we have achieved a significantly improvement in reconstruction completeness than other methods. However, this also brings a “problem” in the quantitative evaluation: our point clouds are even more complete than the ground truth point clouds (Fig. 1), which reduces the reconstruction accuracy in the evaluation.

To find out how the incomplete ground truth data would affect the evaluation metrics, we use smaller images that we set $W = 1280$, $H = 1024$ (in the main paper $W = 1600$, $H = 1184$) to reconstruct the point cloud. In this setting MVSNet will generate results with roughly the same complete level to the ground truth point clouds. It is reported in Table 2 that, although the point cloud is visually less complete than reconstruction with full resolution images (Fig. 1), smaller images produces better accuracy and overall quality performance.

Qualitative Results Finally, Fig. 2 shows the point cloud results of the remaining 16 evaluation scans that have not been shown in the paper or Fig. 1.

2.2 ETH3D Dataset

We also test MVSNet on the ETH3D benchmark [7], low-res dataset. Similar to the evaluation on *Tanks and Temples* dataset, we use the model trained on *DTU* without fine-tuning. $W \times H \times D$ are set to $928 \times 480 \times 320$ for the testing. For quantitative evaluation, MVSNet ranks 5^{th} on the benchmark. The qualitative



(a) W = 1600, H = 1184 (b) W = 1280, H = 1024 (c) Ground Truth

Fig. 1: Comparison on MVSNet reconstructions on (a) full resolution setting (main paper), (b) small resolution setting and (c) ground truth point clouds. The full resolution setting produces point clouds even more complete than the ground truth point clouds

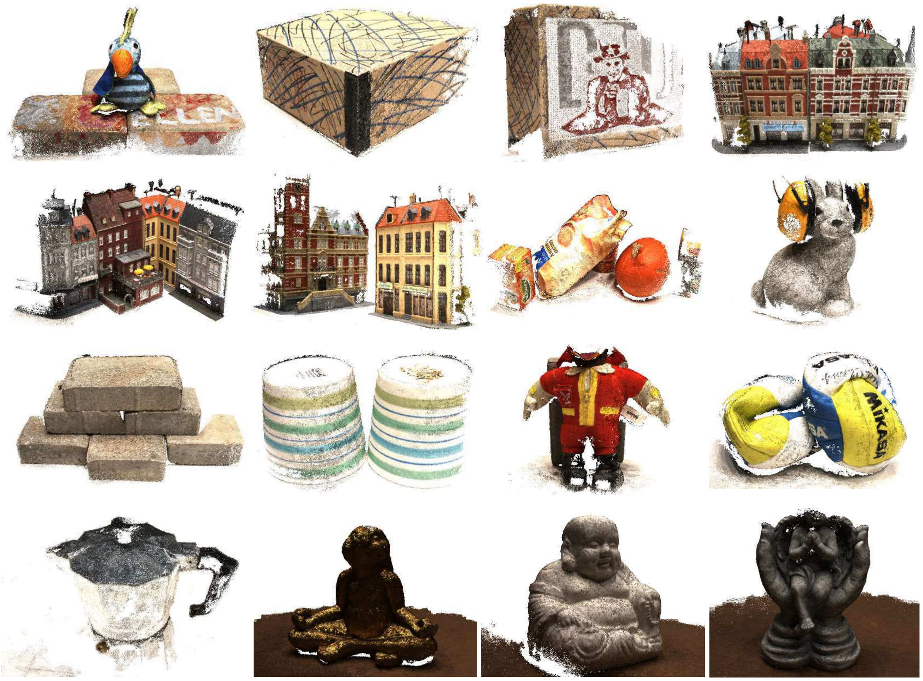
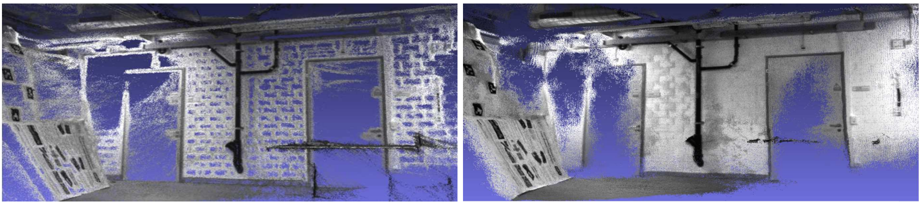


Fig. 2: Point cloud results of the remaining 16 evaluation scans, *DTU* dataset [1]. From top left to bottom right: scans {4, 10, 13, 15, 23, 29, 32, 33, 34, 48, 49, 62, 77, 110, 114, 118}

result of *storage_room_2* is shown in Figure 3, where MVSNet produces complete result especially in those white wall areas.

Fig. 3: Results of storage_room_2, ETH3D. Left: Colmap[6]. Right: MVSNet.



3 More Ablation Studies

3.1 Depth Sample Number

We conduct an ablation study on *DTU* evaluation set that we fix W, H, D_{min}, D_{max} to 1280, 1024, $425mm$, $937mm$, and vary the depth sample number $D = 128, 192, 256, 320$ for reconstructions (resolution $= 4mm, 2.67mm, 2mm$, $1.6mm$ respectively). Table 3 shows that sufficiently high sample resolution ($D = 256$) results in better reconstruction quality than low resolutions ($D = 128, 192$), but higher than that ($D = 320$) reaches to a plateau. We believe $2mm$ is the effective depth resolution for *DTU* dataset.

3.2 Lighting Condition

We demonstrate in this experiment that MVSNet is robust to lighting changes. In the main paper, MVSNet is trained on consistent lighting conditions (the N images of one training sample are selected from the same lighting condition), and we perform all evaluations using the uniform lighting. We conducted one more ablation study on *DTU* evaluation set that we replaced the 49 uniform lighting images in the scan with the random lighting images (randomly chosen from 7 lighting conditions) for evaluation. Table 3 shows that the random lighting only results in little performance drop.

3.3 Baseline Angle

In the main paper we choose $\theta_0 = 5, \sigma_1 = 1, \sigma_2 = 10$ and select the best $N - 1$ views for each reference image. To determine the generalization w.r.t. the baseline length/angle, we changed θ_0 to $10, 20$ and 30. Table 3 shows MVSNet still produces high quality results for large baseline angle $\theta_0 = 20$.

Table 3: Generalization w.r.t. depth sample number, lighting condition and baseline angle.

Setting	<1mm			<2mm		
	Accu.	Comp.	f-score	Accu.	Comp.	f-score
MVSNet	86.46	72.13	75.69	91.06	75.31	80.25
w. r. t. depth sample number ($W = 1280, H = 1024$)						
$D = 128$	75.31	61.97	66.44	91.32	69.30	77.00
$D = 192$	89.80	68.44	75.80	94.63	72.55	80.21
$D = 256$	91.02	69.68	77.07	94.75	73.82	81.21
$D = 320$	89.86	69.64	76.72	94.42	74.08	81.32
w. r. t. lighting condition						
random lighting	85.68	70.49	75.26	90.45	74.68	79.90
w. r. t. baseline angle						
$\theta_0 = 10$	85.83	70.96	75.57	90.43	74.78	79.86
$\theta_0 = 20$	87.70	64.16	72.01	93.77	69.07	77.37
$\theta_0 = 30$	84.63	57.02	66.35	91.12	63.21	72.79

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision (IJCV) (2016)
2. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. European Conference on Computer Vision (ECCV) (2008)
3. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2010)
4. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. International Conference on Computer Vision (ICCV) (2015)
5. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. International Conference on Computer Vision (ICCV) (2017)
6. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. European Conference on Computer Vision (ECCV) (2016)
7. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. Computer Vision and Pattern Recognition (CVPR) (2017)
8. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications (MVA) (2012)