Generalized Binary Search Network for Highly-Efficient Multi-View Stereo

-Supplementary-

Zhenxing Mi Chang Di Dan Xu
The Hong Kong University of Science and Technology
{zmiaa, dchangac}@connect.ust.hk, danxu@cse.ust.hk

In this supplementary, we introduce details about the depth map fusion procedure and provide more qualitative results regarding the ablation study and the overall performance of the proposed model.

1. Depth map Fusion

As indicated in the main paper, after obtaining the final depth maps of a scene, we filter and fuse depth maps into one point cloud. The final depth maps are generated from center points of the selected bins of the final stage. We consider both the photometric and the geometric consistency for depth map filtering. The geometric consistency is similar to MVSNet [8] measuring the depth consistency among multiple views. The photometric consistency, however, is different. The probability volume $\mathbf{P}$ is considered to construct the photometric consistency, following R-MVSNet [9]. As the probability volume is the classification probabilities for the depth hypotheses, it measures the matching quality of these hypotheses. Since the proposed method consists of K stages, we can obtain K probability volumes, *i.e.* $\{\mathbf{P}_k | k = 1, ..., K\}$. For each pixel $\mathbf{p}$, its photometric consistency from its K probabilities can be calculated as follows:

$$Ph(\mathbf{p}) = \frac{1}{K'} \sum_{k=1}^{K'} \max\{\mathbf{P}_k(j, \mathbf{p}) | j = 1, ..., D\}. \quad (1)$$

Where $Ph(\mathbf{p})$ is the photometric consistency of pixel $\mathbf{p}$; D is depth hypothesis number; The $\max$ operation obtains the classification probability of a selected hypothesis; K' is the maximum stage considered in photometric consistency and $1 \leq K' \leq K$. Equation 1 actually computes an average of the probabilities of the K' stages. In practice, when the maximum stage number $K = 8$, we set $K' = 6$. It means that we take the average probability of the first 6 stages as the score of the photometric consistency. In our multi-stage search pipeline, as the resolutions of probability volumes are different, we upsample them to the maximum resolution of stage K before the computation. After producing the photometric consistency score for each pixel, the depths of pixels are discarded if their consistency scores are below a threshold.

Figure 1a in the supplementary shows the results of each stage of a sample in the DTU dataset [2]. The depth map in each stage consists of the center-point depth values of selected bins. The quality of these depth maps can be improved quickly, demonstrating a fast search convergence of our method. The valid mask maps represent valid pixels in each search stage. Note that these mask maps are combined with the ground-truth mask maps from the dataset, and thus the background pixels are not considered. The photometric consistency (Photo. Consi.) map in stage k is computed using Equation 1 by setting $K' = k$. As shown in the Figure 1a, the photometric consistency maps is an effective measurement of depth map quality. As shown in Figure 1b, the photometric consistency (Photo. Consi.) maps from Stage 6 are used to filter the final depth maps produced from Stage 8. The filtered depth maps are further refined by geometric consistency map, and finally fused into one point cloud. Figure 2 also shows qualitative results of a sample from the Tanks and Temples [4] dataset. The background of this image is far away from the foreground and is out of the depth range, so the MVS methods predict outlier values for the background pixels. Using the photometric consistency maps, we can effectively filter out these outliers.

2. Evaluation on ETH3D dataset

We perform additional experiments on ETH3D [1]. Our model results are shown in Table 1, comparing with the best results of other methods obtained from the leaderboard [1], where our evaluation is named as GBi-Net. Although the ETH3D dataset is challenging, our method can still produce clearly better results compared to both traditional and competitive learning-based MVS methods.

Table 1. Evaluation results on the training and test splits of ETH3D multi-view benchmark (F1 score, higher is better).

Methods	Training ↑	Test ↑
Gipuma [3]	36.38	45.18
COLMAP [5]	67.66	73.01
PVSNet [7]	67.48	72.08
PatchmatchNet [6]	64.21	73.12
Ours	**70.78**	**78.40**

3. More Visualization Results

We show more qualitative results of the proposed model in this section. Figure 3 and Figure 5 show several images and their corresponding depth maps in DTU dataset [2] and Tanks and Temples dataset [4] respectively. The depth maps are filtered by photometric consistency. Figure 4 and Figure 6 shows several point clouds of our method in DTU dataset [2] and Tanks and Temples dataset [4] respectively.

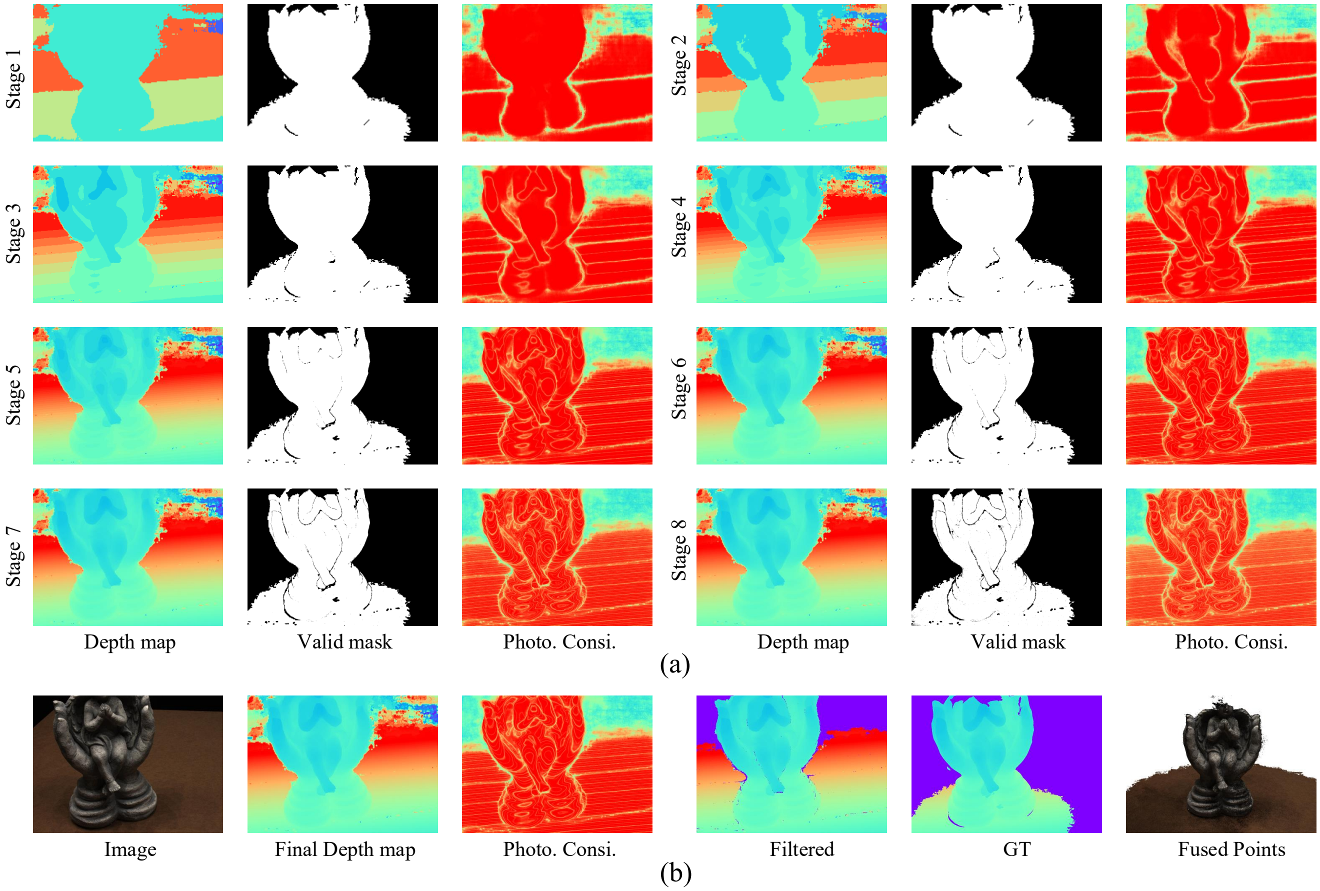

Figure 1. (a) The predicted depth maps, valid mask maps and photometric consistency (Photo. Consi.) maps in all the stages of a sample in DTU [2]. (b) The input image, final predicted depth map of Stage 8, photometric consistency (Photo. Consi.) map of Stage 6, filtered depth map by photometric consistency, ground truth depth map and fused point cloud.

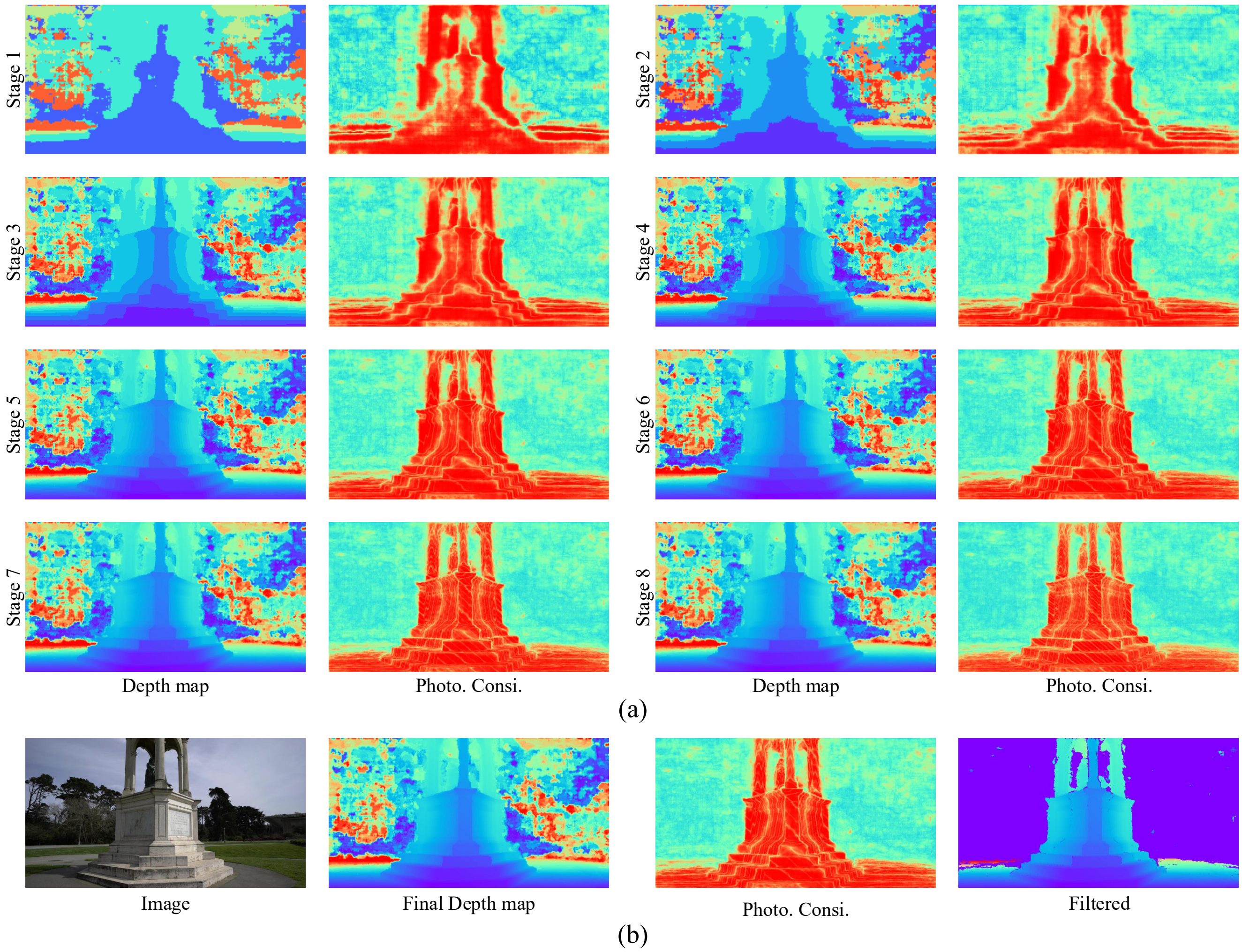

Figure 2. (a) The predicted depth maps and photometric consistency (Photo. Consi.) maps in all the stages of a sample in Tanks and Temples [4]. (b) The input image, final predicted depth map of Stage 8, photometric consistency (Photo. Consi.) map of Stage 6 and filtered depth map by photometric consistency. The background of this image is far away from the foreground and is out of the depth range so MVS methods will predicted outlier values for background pixels. With the photometric consistency, we can effectively filter out these outliers.

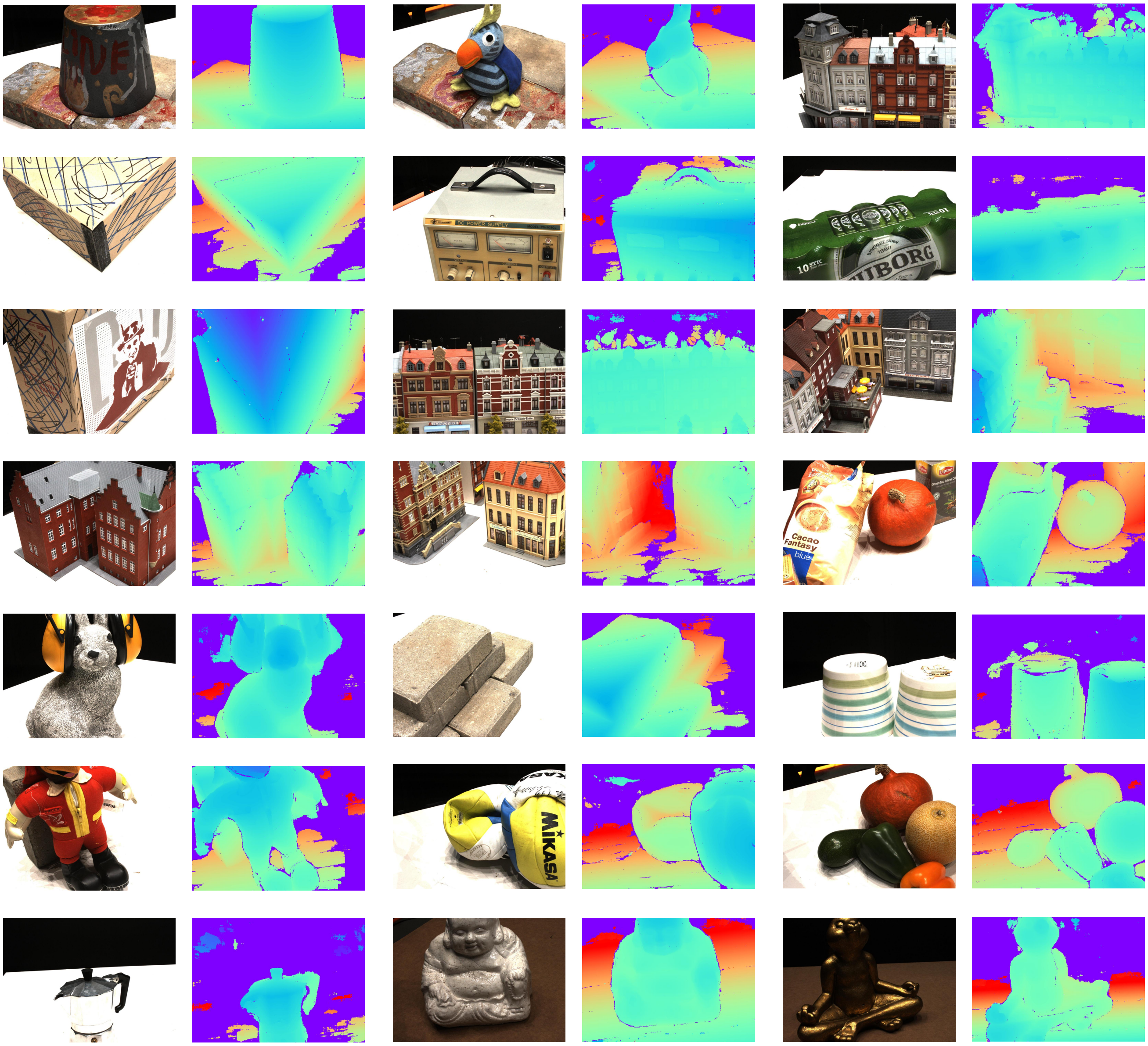

Figure 3. Examples of images and their corresponding depth maps in DTU dataset [2]. The depth maps are filtered by photometric consistency.

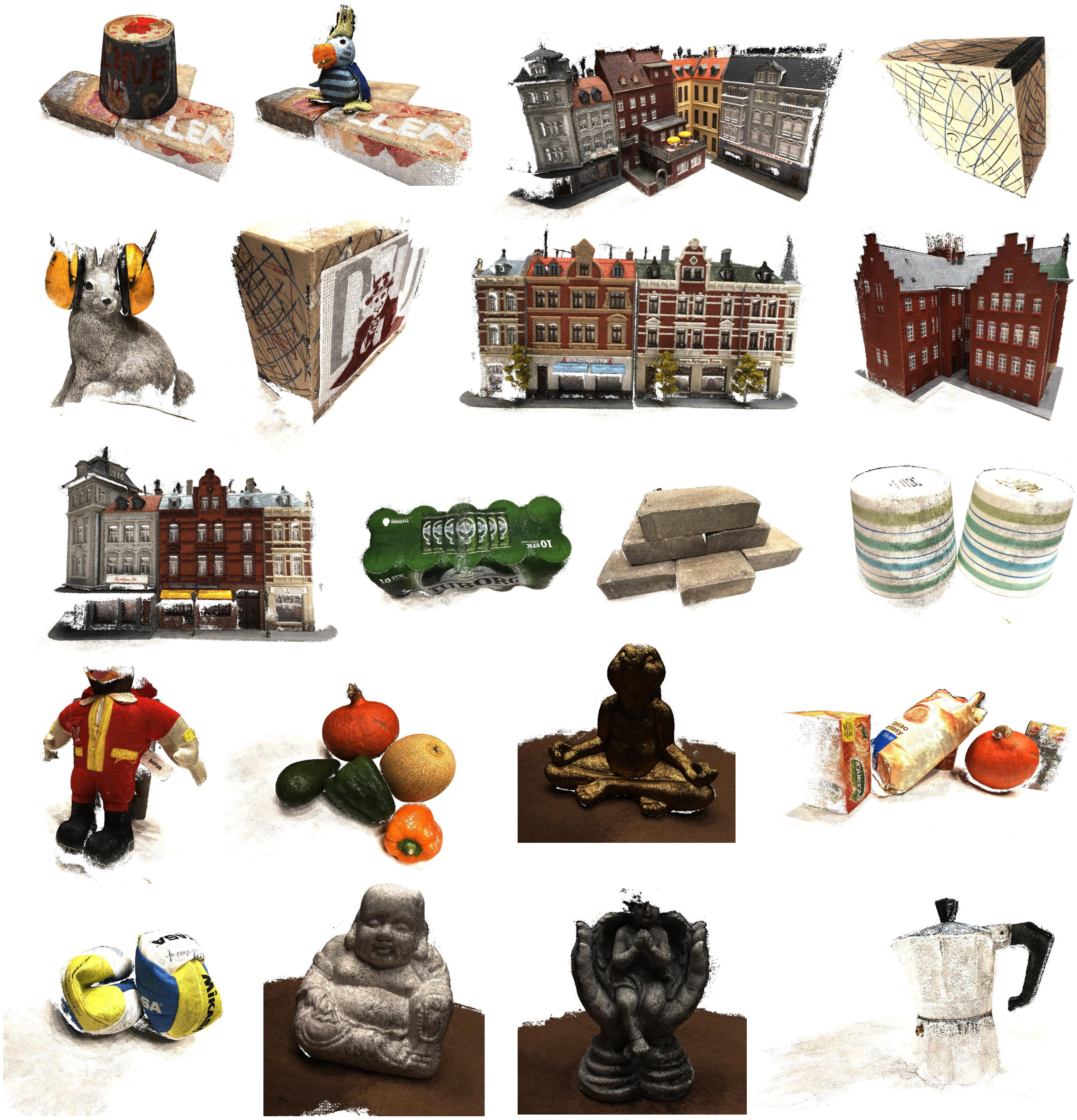

Figure 4. Point clouds of our method on DTU dataset [2].

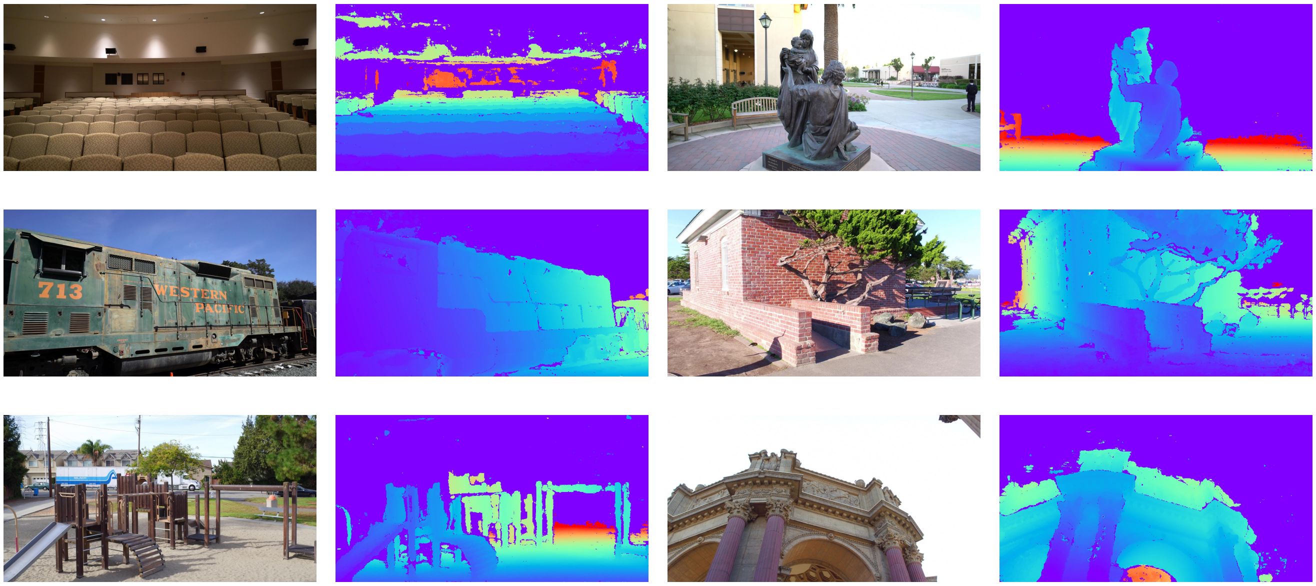

Figure 5. Examples of images and their corresponding depth maps in Tanks and Temples dataset [4]. The depth maps are filtered by photometric consistency.

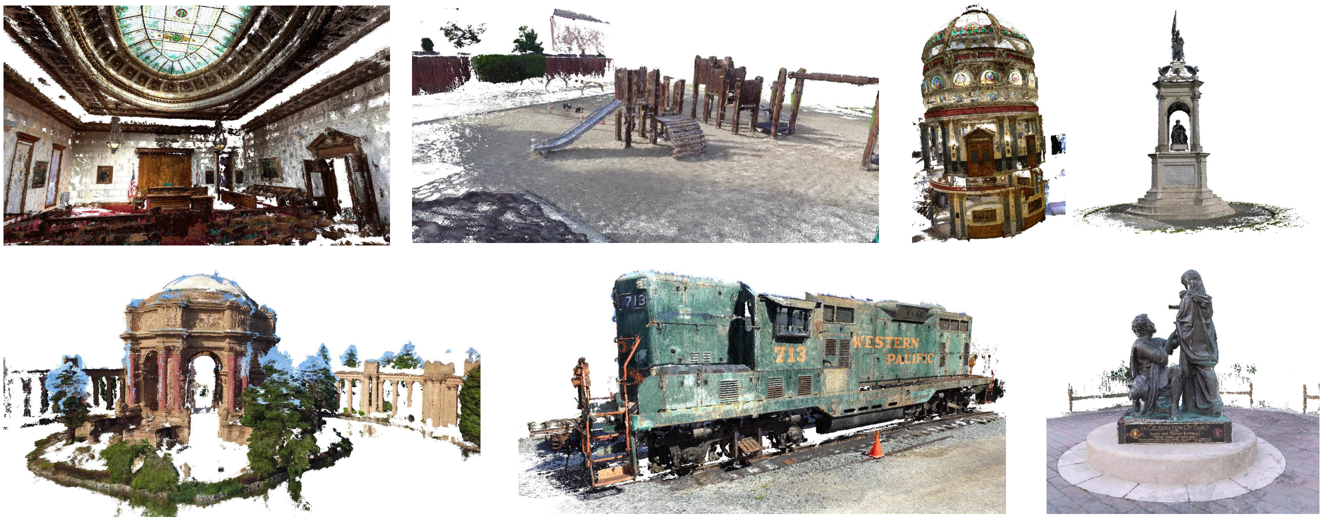

Figure 6. Point clouds of our method on Tanks and Temples dataset [4].

References

[1] Eth3d high-resolution multi-view leaderboard. https://www.eth3d.net/high_res_multi_view. 1

[2] 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*, 120(2):153–168, 2016. 1, 2, 3, 5, 6

[3] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 2

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM ToG*, 36(4):1–13, 2017. 1, 2, 4, 7

[5] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*. Springer, 2016. 2

[6] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021. 2

[7] Qingshan Xu and Wenbing Tao. Pvsnet: Pixelwise visibility-aware multi-view stereo network, 2020. 2

[8] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 1

[9] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 1