Efficient Multi-view Stereo by Iterative Dynamic Cost Volume –Supplemental Materials–

Shaoqian Wang♯ Bo Li♯ Yuchao Dai*
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

1. Efficiency Analysis for Initial Depth Prediction Module

In this section, we describe the network architecture of our proposed initial depth prediction module (Table 1) and analyze its time and memory consumption (Table 2). Considering that regularizing a 3D cost volume with 3D CNN is an expensive operation, we run the initial depth prediction module at the coarsest stage to reduce its time and memory consumption. As shown in Table 2, our initial depth prediction module affects the efficiency slightly.

Output	Layer	Input	Output Size
Cost Volume			H/8 × W/8 × D × 32
Conv0	Conv3DBn,S=1,F=8	Cost Volume	H/8 × W/8 × D ×8
Conv1	Conv3DBn,S=1,F=8	Conv0	H/8 × W/8 × D ×8
Conv2	Conv3DBn,S=2,F=16	Conv1	H/16 × W/16 × D/2 ×16
Conv3	Conv3DBn,S=1,F=16	Conv2	H/16 × W/16 × D/2 ×16
Conv4	Conv3DBn,S=2,F=32	Conv3	H/32 × W/32 × D/4 ×32
Conv5	Conv3DBn,S=1,F=32	Conv4	H/32 × W/32 × D/4 ×32
Conv6	DeConv3DBn,S=2,F=16	Conv5	H/16 × W/16 × D/2 ×16
Conv7	DeConv3DBn,S=2,F=8	Conv3+Conv6	H/8 × W/8 × D × 8
P	Conv3DBn,S=1,F=1	Conv1+Conv7	H/8 × W/8 × D ×1

Table 1. The network architecture of our 3D CNN in initial depth prediction module. The Conv3DBn layer consists of a Conv3D module and a BatchNorm module. D represents the number of depth hypotheses.

Method	Time(s)	Memory(GB)
w/o IDP	0.16	2.6
Ours	0.19	3.1

Table 2. Ablation study of our proposed initial depth prediction module (IDP) in terms of time and memory consumption.

2. Depth Filter

After obtaining the depth map corresponding to the input images, we need to filter out the outliers in the depth map and then fuse the filtered depth maps to 3D point clouds. Considering that the depth filter algorithms have a great impact on the quality of the final point clouds, most previous MVS works apply geometric constraints [4] to filter the outliers, and some recent works use a dynamic consistency check algorithm [5] to maintain a more reliable and accurate depth value. An important filter strategy in these algorithms is to compare the reprojected depth map D_s with the reference depth map D_r:

$$||D_s - D_r||/D_r < \gamma, \tag{1}$$

where γ is a constant threshold. However, as shown in Fig. 1, since the range of D_r is between d_{min} and d_{max}, the filter method is unfair to different depth values under a fixed threshold. Specifically, for a same point P in different views, the depth filer method using Eq. 1 will set different sizes of depth ranges that meet the threshold. Meanwhile, for different points P_i in a single view, Eq. 1 will also set different sizes of depth ranges that meet the threshold.

To address the above problems, we make some improvements on the basis filter method [4], so that the sizes of depth ranges that meet the threshold will be not be affected by the change of D_r:

$$||D_s - D_r|| < \tau, \tag{2}$$

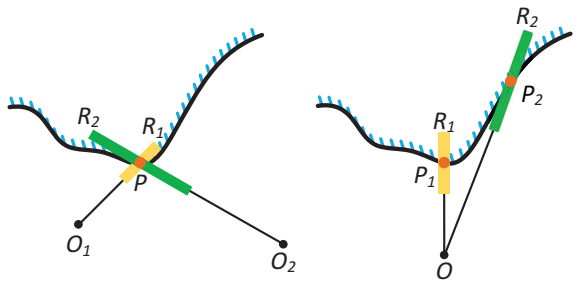

Figure 1. The depth filter strategy using Eq. 1. **Left:** the range R_1 and R_2 of depth values that meet the threshold γ for a same point P in different views. **Right:** the range R_1 and R_2 of depth values that meet the threshold γ for different points P_1 and P_2 in a single view.

*The first two authors contributed equally. Yuchao Dai is the corresponding author (daiyuchao@gmail.com).

where τ is a constant depth determined by the depth range of each image of the scene. For the input image $\{I_0...I_{N-1}\}$ of the scene and the corresponding depth ranges $\{[d^0_{min}, d^0_{max}]...[d^{N-1}_{min}, d^{N-1}_{max}]\}$, we set:

$$\tau = \lambda \frac{\sum_{i=0}^{N-1}(d^i_{min} + d^i_{max})}{2N}, \tag{3}$$

where λ is the weight parameter that can be adjusted. We verify the proposed depth filter method on the DTU's evaluation set [1]. In addition, we also utilize the reprojected coordinate error to filter the depth map [4]. As shown in Table 3, our proposed depth filter achieve improved performance in terms of accuracy and overall.

Honestly, due to the influence of background depth value, calculating threshold based on the depth range of all views may not be an excellent solution. In the future, we plan to study more robust depth filtering methods.

Method	Geometric Constrain [4]			Our Method		
	Acc.	Comp.	Overall.	Acc.	Comp.	Overall.
Ours(Iters: 3 3 3)	0.332	0.312	0.322	0.321	0.313	0.317
Ours(Iters: 1 1 1)	0.328	0.332	0.330	0.314	0.334	0.324

Table 3. Ablation study of our proposed depth filter method in term of distance metric(mm) on DTU's evaluation set [1]

3. Numbers of Depth Hypotheses for Local Cost Volume

The local cost volume is the most important part of dynamic cost, because it can provide local geometric information based on the sampled depth hypotheses. Specifically, the number of depth hypotheses determine the search range when extracting the geometric information. In this section, we conduct ablation experiments with numbers of depth hypotheses for local volume construction on DTU's evaluation set [1]. It is worth noting that the numbers of depth hypotheses are same at different stages. In addition, when the number of depth hypothesis is set to 1, the local cost volume can only represent the geometric information based on the current depth map [2]. As shown in Table 4, we observe a performance drop when the number of depth hypotheses is 1 or 2. Meanwhile, when the number is greater than 4, the performance of our method does not further improve with the increase of the number of depth hypotheses. This shows that our method can extract enough local geometric information from a very thin cost volume, which is one of the reasons why our method is very efficient.

4. Visualization of Point Clouds

We visualize the reconstructed point clouds from DTU's evaluation set [1] and Tanks & Temples dataset [3] in Fig. 2, 3.

Number	Acc.(mm)	Comp.(mm)	Overall(mm)
1	0.318	0.418	0.368
2	**0.316**	0.328	0.322
4	0.321	0.313	**0.317**
6	0.324	**0.310**	0.317
8	0.325	0.311	0.318

Table 4. Ablation study of the number of depth hypotheses on DTU's evaluation set [1].

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*, 120(2):153–168, 2016. 2, 3

[2] Xiaodong Gu, Weihao Yuan, Zuozhuo Dai, Chengzhou Tang, Siyu Zhu, and Ping Tan. Dro: Deep recurrent optimizer for structure-from-motion. *arXiv preprint arXiv:2103.13201*, 2021. 2

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

[4] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 1, 2

[5] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020. 1

Figure 2. Reconstruction results on DTU's evaluation set [1].

Figure 3. Reconstruction results on Tanks & Temples dataset [3].