Supplementary Material: MosaicMVS

https://anonymous.4open.science/w/MosaicMVS-1B17/

Paper ID 6833

Contents

1

1 Implementation Details

1.1 Details on Input Image Resolution

Original 6000×4000 high resolution RGB images were resized to 4000×3000. As we conducted the experiments without ground truth camera poses and depth maps, we utilized a SfM-based reconstruction framework COLMAP [2] for extracting the pseudo camera poses and depth maps. In addition, the images were undistorted to 4042×3013 during the camera poses estimation by the COLMAP [2]. We used the undistorted images and the camera pose as the inputs of our proposed framework with a 3-stage-cascaded depth inference network. From the first to the third stage, the spatial resolution of feature volume is gradually increased by 1/16, 1/4, and 1 of the original input image size. In this way, the output depth and image resolution are finally set to 1152×832.

1.2 Network Parameters

We use the Cas [1] as the baseline model in our implementation and replace its various components with the proposed depth range adjustment, view selection method, and valid view cost volume considering zero overlapping hypothetical voxels. For all scenes, the interval scale was set at 1.06, following the base model. From the first to the third stage, the number of depth hypotheses is 48, 32, and 8, and the corresponding depth interval is set to 4, 2, and 1 time following the interval of the base model, respectively. Other model configurations and hyperparameters follow the base model, including cascade cost volume formulation of 3 stages. We used an RTX 3090 for evaluation with a batch size of 1.

1.3 Details on Post-processing

For geometric filtering process, we follows [6], the method goes through three masking steps, with the whole process iterated three times. First, we created a mask M_{dist} that filters pixels with a significant reprojection error between estimated depth and reprojected depth. The threshold parameter Th_{dist} limits the displacement between the target viewpoint depth and the reprojected source viewpoint depths. The mask M_{dist} is defined as:

$$M_{dist}(d_p^{tgt}) = \begin{cases} 1, & \text{if } \|d_p^{tgt} - \hat{d}_p^{src}\| < Th_{dist}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where d_p^{tgt} and $\hat{d}_p^{src}$ indicate the target viewpoint depth and source viewpoint depth reprojected to target viewpoint at a pixel p, respectively. Second, the mask M_{depth} filters pixels whose reprojection error ratio is greater than the threshold parameter Th_{depth} such that:

$$M_{depth}(d_p^{tgt}) = \begin{cases} 1, & \text{if } \frac{|d_p^{tgt} - \hat{d}_p^{src}|}{\max(d_p^{tgt}, \hat{d}_p^{src})} < Th_{depth}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Above masks ensure geometric consistency of reprojected depths with the target viewpoint depth. Lastly, the mask M_{geo} was applied to filter pixels such that;

$$M_{geo}(d_p^{tgt}) = \begin{cases} 1, & \text{if } M_{depth} * M_{dist} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We then check the geometric consistency of each pixel by accepting pixels with at least Th_V number of reprojected depths showing a similar depth value with that of the pixels. The above steps of geometric filtering were iterated three times. Using the 3^{rd} mask (M_{final}), we filtered the pixels to form the 3d point cloud generated in the final step of the geometric filtering process. We compared the masks M_{geo} generated at each iteration, and M_{final} made in the final iteration of the geometric filtering process in Fig. 1.

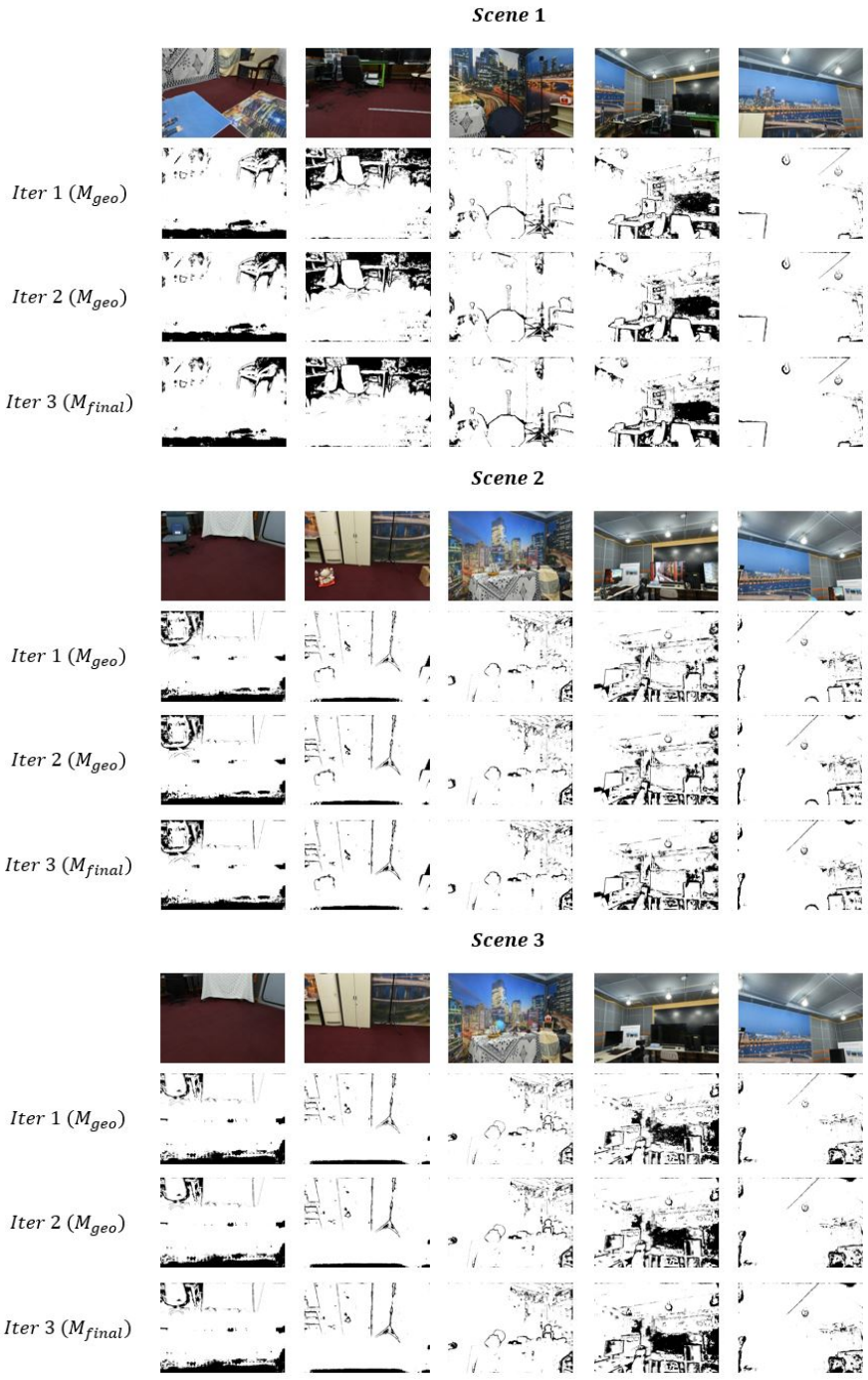


Fig. 1: Qualitative comparison of geometric masks M_{geo} generated at each of the iteration. The M_{final} is the mask made in the last iteration.

2 Experiments

In this section, we described detail explanation of evaluation metrics and additional experimental results.

2.1 Evaluation Metrics on Depth Estimation

In our evaluations, we adapted standard quantitative measures of depth quality. The used metrics are as follows:

Error metrics

$$\text{rmse} = \sqrt{\frac{1}{|\Omega|}\sum_{p\in\Omega}(d_p - d_p^{gt})^2},$$
$$\text{abs\_diff} = \frac{1}{|\Omega|}\sum_{p\in\Omega}|d_p - d_p^{gt}|,$$
$$\text{abs\_rel} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\frac{|d_p - d_p^{gt}|}{d_p^{gt}},$$
$$> P \text{ px} = \% \text{ of } d_p \text{ s.t. } |(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p})| < P \text{ for P} = 1, 3, 5,$$

Accuracy metric

$$\delta < \alpha^t = \% \text{ of } d_p \text{ s.t. } \delta = \max(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p}) < \alpha^t \text{ for } t = 1, 2, 3,$$

where d_p and d_p^{gt} indicate the estimated depth map and ground truth depth map at a pixel p, respectively. Ω represents a set of valid pixels. Following prior works, we set $\alpha = 1.25$ for accuracy metric.

2.2 Qualitative Evaluation on Depth Estimation Results

Fig. 2 shows the qualitative comparisons with depth estimation results from other networks. Our framework accurately predicts even in reflective regions, i.e., the umbrella or the surface of the air conditioner. Also, our framework predicts depth without holes while covering fine object boundaries like the umbrella handle. Furthermore, our framework better predicts depth at challenging areas, including the texture-less regions like the floor and regions with many textures like wallpaper.

2.3 Qualitative Evaluation on Reconstruction Results

Fig.3 shows the qualitative results of reconstructed point clouds of the SAOI dataset scenes which are reconstructed by depths estimated from AACVP [5], CVP [4], SSCVP [3], Vis [6], and ours. From these results, we can confirm that our framework generates a robust indoor point cloud compared to other MVS depth estimation frameworks. The more interactive reconstruction results can be seen at https://anonymous.4open.science/w/MosaicMVS-1B17/.

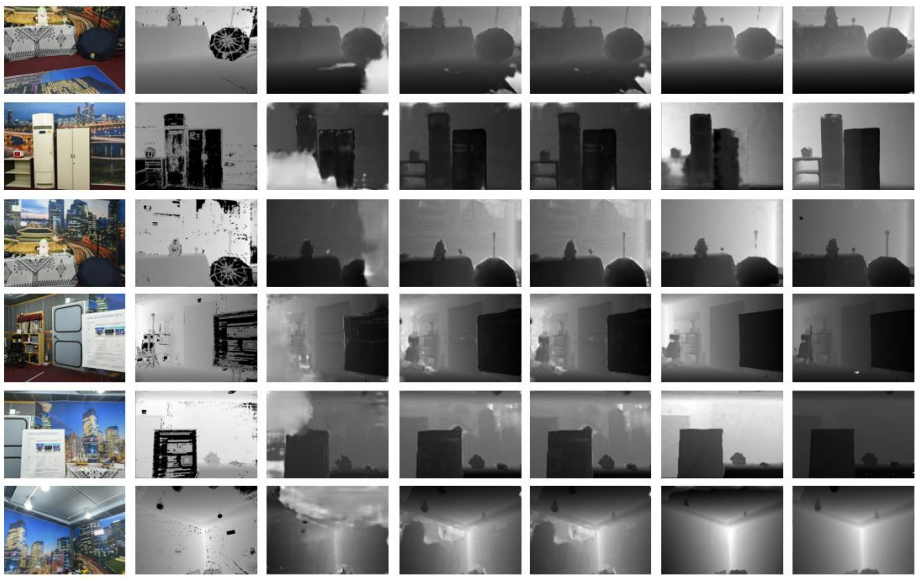


Fig. 2: Qualitative comparison depth estimation results on SAOI scene 1.

2.4 Qualitative Evaluation on View Synthesis Results

Fig. 4 shows the qualitative results of reconstructed surface meshes of the SAOI dataset scenes. All of the meshes were reconstructed based on the final depth maps of the proposed framework. For surface reconstruction, we applied Poisson surface reconstruction method with OPEN3D [7], setting tree depth parameter to 9. We also compared the qualitative results of synthesized view images in Fig. 5. First row of each scene shows comparison between target view images synthesized, and second row shows comparison between target view mesh depth maps rendered from surface meshes reconstructed by COLMAP[2], Vis [6], and the proposed method. Ground truths for target view depth map are blanked as there are no ground truths for target view depths of the SAOI dataset.

2.5 Comparison of Generated Masks

We also compared M_{final} masks generated from different networks. The Fig. 6 shows valid areas as white and invalid as black in terms of geometric consistency between viewpoints. As shown in the figure, the masks from our framework show more valid white areas, which indicate the geometric consistency of our framework compared to other networks. Furthermore, the masks from our framework show even more valid white areas than the COLMAP [2], which we used as pseudo ground truth for depth map evaluation. These results demonstrate that the estimated depth maps from our framework are more geometric consistent than other depth maps, including pseudo ground truth COLMAP depth maps.

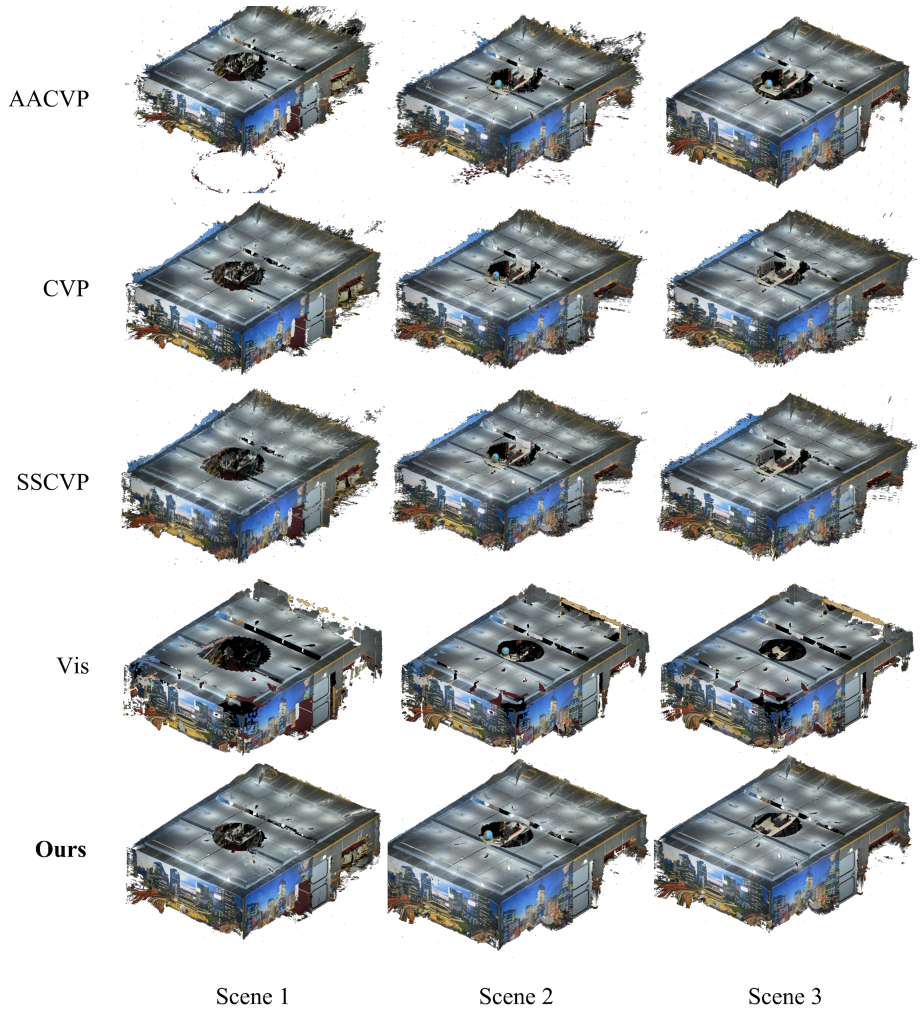
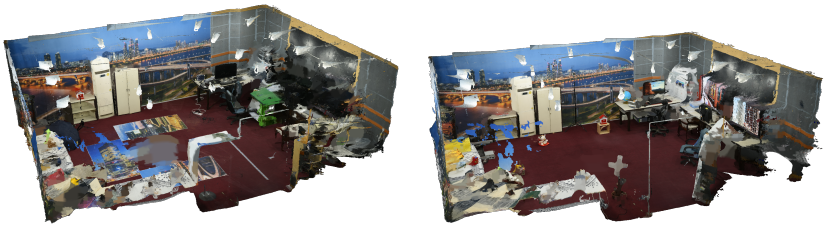


Fig. 3: Qualitative results on reconstruction results. Comparison between point clouds reconstructed from depth estimated by AACVP [5], CVP [4], SSCVP [3], Vis [6], and ours.

2.6 Experiment Results of Extra Hemisphere Setup Images

Furthermore, we experimented our framework on other images in order to prove flexibility that our framework can be applied well even in the hemisphere, limited mosaic-based omnidirectional camera setup. The images were taken at 10-degree intervals in the gym. The total number of images N_{total} is 133. Fig. 7 shows the estimated camera pose of the images and the qualitative results of the depth



(a) Scene 1 reconstructed mesh. (b) Scene 2 reconstructed mesh.

(c) Scene 3 reconstructed mesh.

Fig. 4: Qualitative results on view synthesis, which show reconstructed surface meshes of the SAOI dataset scene 1, 2, and 3 using the results of the proposed method.

estimation and post-processing. The quality of each depth map is good enough, despite the setup where the viewpoints at both ends of the horizontal direction are not connected to each other, creating a situation where the source view is likely to be biased in one direction. Additionally, as shown in Fig. 8, the estimated depth map generated robust 3D point cloud and mesh reconstructions even under intense lighting conditions. According to the reconstruction results, the synthesized images were also derived very similar to the ground truth target image.

Table 1: Comparison qualitative results of depth range non-fixed and fixed on the SAOI dataset scene 1

Methods	Geometric validation metric (E_{geo})	Error metrics ↓			Accuray metrics ↑ ($\delta < \alpha^t$)		
		rmse	abs_diff	abs_rel	α	α^2	α^3
Before	67.95	6.021	3.293	0.258	0.911	0.945	0.966
After	85.04	1.575	0.532	0.037	0.947	0.969	0.980

3 Ablation Studies

We conducted more ablation studies to demonstrate the effectiveness of our framework.

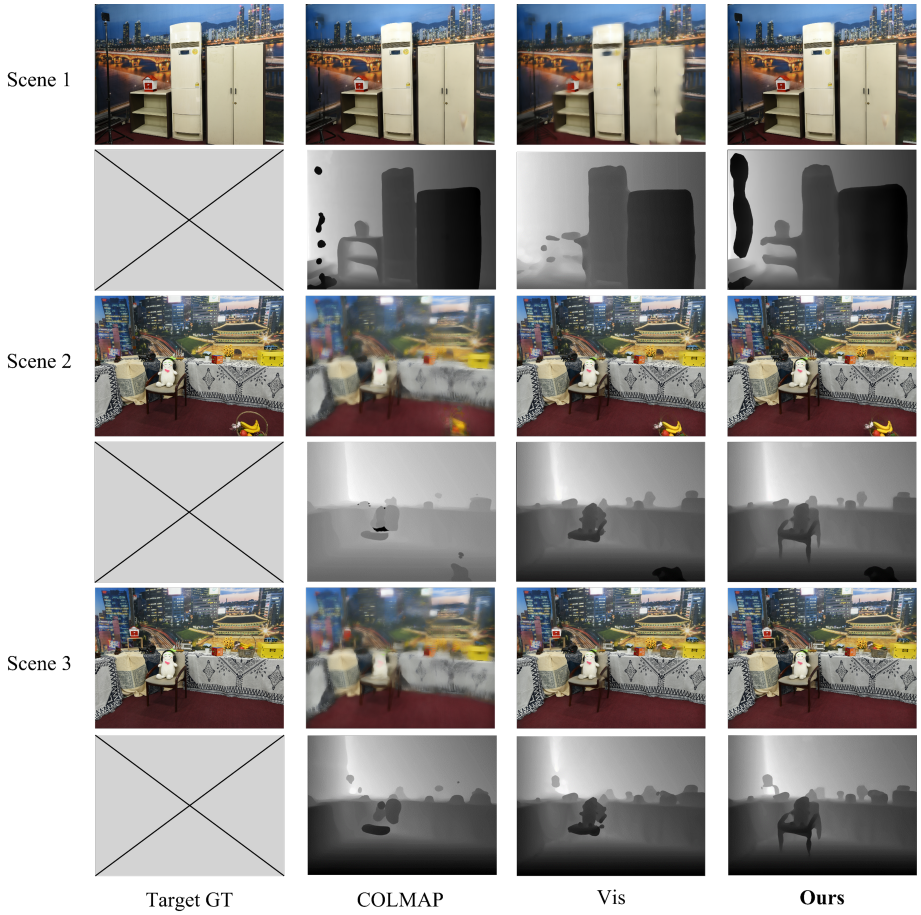


Fig. 5: Qualitative results on view synthesis. Comparison between target view images synthesized and mesh depth maps rendered from surface meshes on reconstructed by COLMAP[2], Vis [6], and the proposed method.

3.1 Depth Range Adjustment

Fig. 9 shows the qualitative comparison of the proposed depth range adjustment. Applying the proposed depth range adjustment shows overall improved depth maps. Table 1 shows the results of quantitative comparison before and after fixing the depth range in Scene 1. We experimented with our setting 3 except for the depth range adjustment. For comparing the average of the pixel occupancy used for reconstruction and depth estimation metric, fixing the depth range shows improved performances.

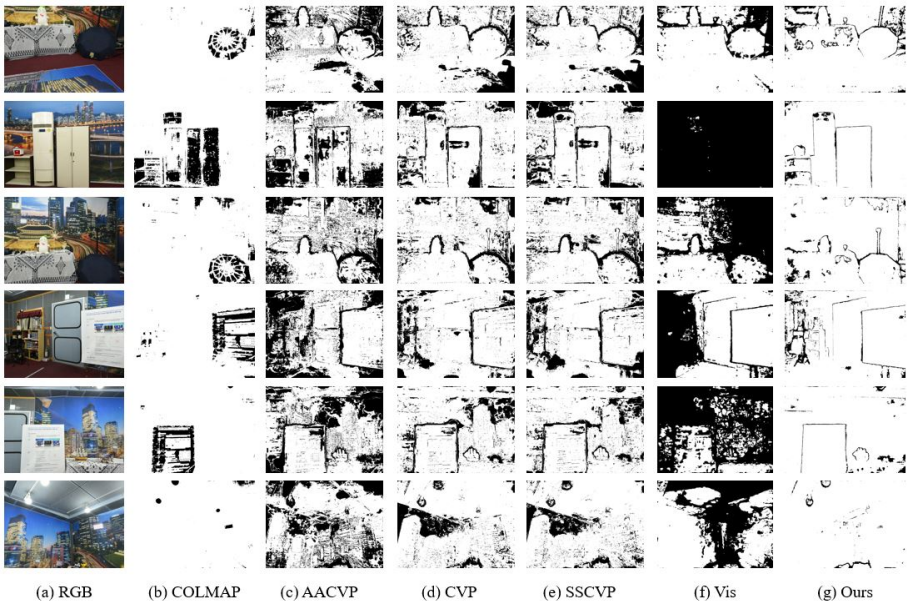


Fig. 6: Comparison of post-processed output masks.

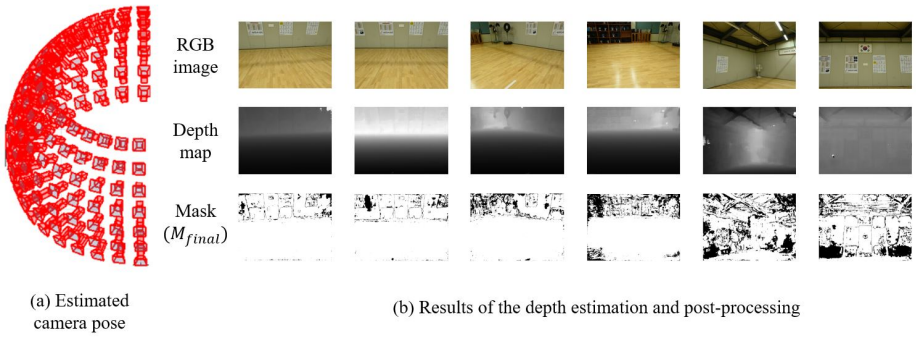
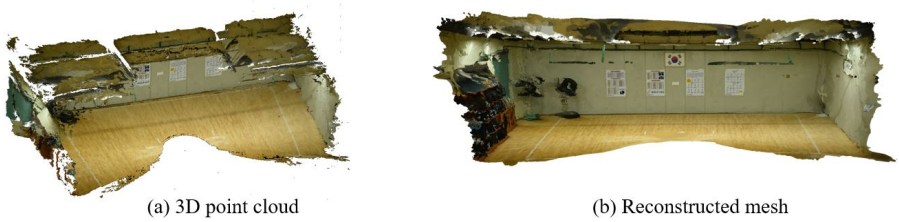


Fig. 7: Depth map estimation results and post-processed output masks for the hemisphere setup.

3.2 Mosaic Array View Selection

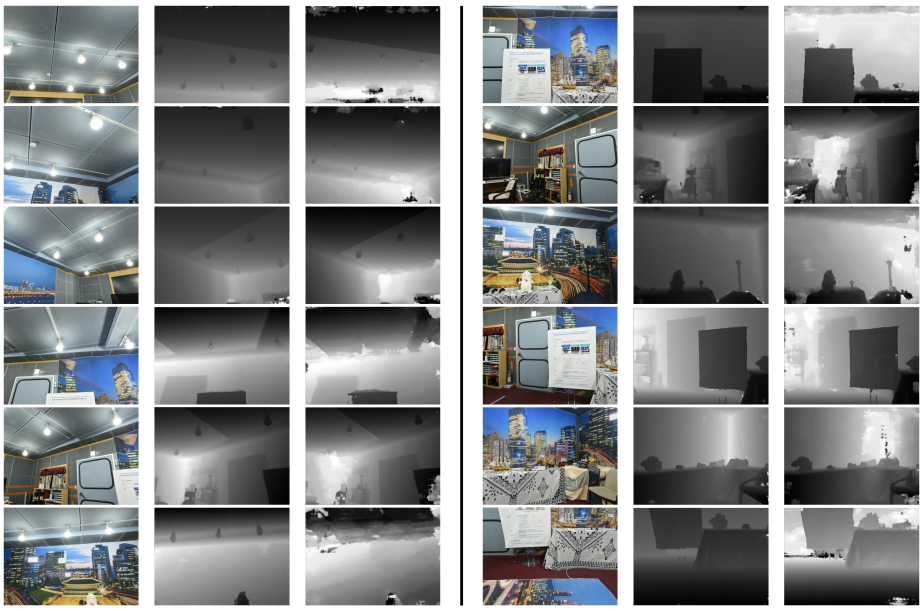
The proposed mosaic array view selection can be applied without any accurate point cloud reconstruction as described in the main paper. Fig. 10 shows an example comparison of selected source views between the COLMAP-based method [2] and our method. The number of source views N are 3, 5, 8, 14, and 24, respectively. In addition, the example illustrates the source views selected by using the mosaic array view selection method are positioned uniformly and close to the target view. Fig. 11 shows the evaluation results of reconstruction for the



(a) 3D point cloud (b) Reconstructed mesh

(c) Comparison between target and synthesized target images

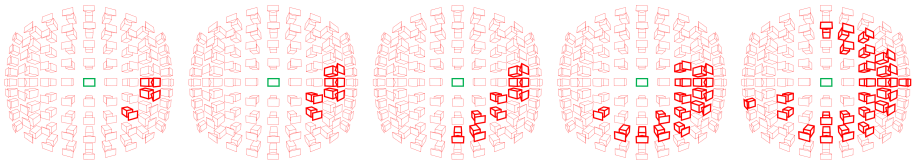
Fig. 8: Reconstructions and view synthesis images of the hemisphere setup.



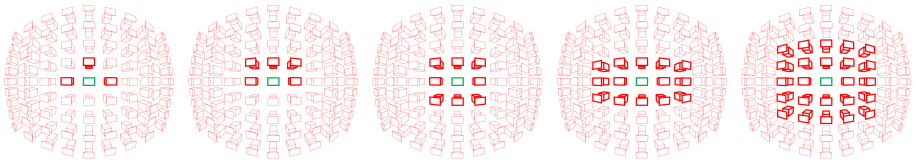
(a) RGB (b) w/ DR (c) w/o DR (a) RGB (b) w/ DR (c) w/o DR

Fig. 9: Qualitative comparison of the proposed depth range (DR) adjustment.

different number of source views in each view selection.The reconstruction performance was evaluated using the geometric validation metric E_{geo}. Also, the same $N = 24$ was used for both methods. The proposed view selection method



(a) View selection using SfM-based method (COLMAP).



(b) View selection using proposed mosaic array view selection.

Fig. 10: Comparison of selected source views between COLMAP-based method and ours. For each figure from the right, the source views are 3, 5, 8, 14, and 24, respectively.

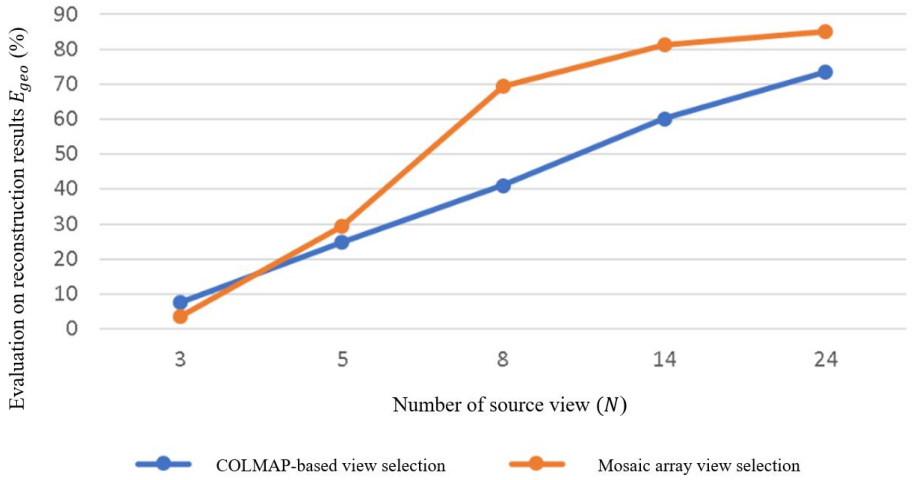


Fig. 11: The results of the evaluation E_{geo} on scene 1 as the number of source view N changes.

constantly outperformed the COLMAP-based view selection method [2] on all number of source views except for the $N = 3$. Fig. 12 shows the comparison of depth map results due to the change in the number of source views, denoted by N. As the number of source views increases, the depth maps show less visible artifacts and more consistent depth among viewpoints.

Table 2 is the result of a comparison of the correlation between each method by applying different view selection methods to depth fusion and depth inference.

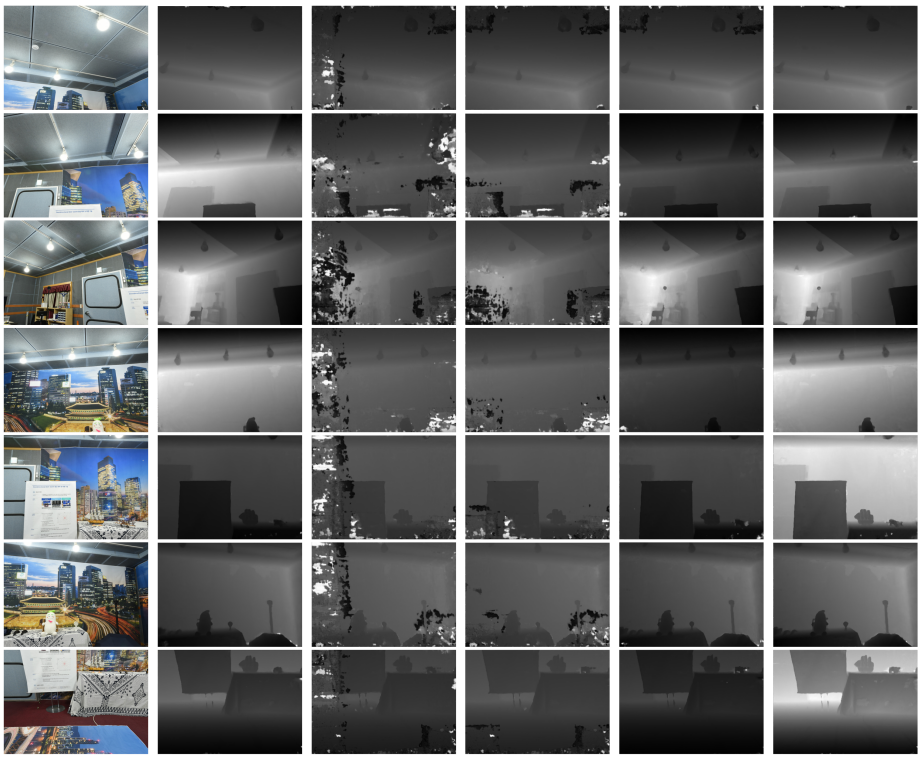


Fig. 12: Comparison of depth results due to the change in the number of views.

N of the proposed view selection method and COLMAP [2] were both set to 24, 10, respectively. Moreover, even when using the same estimated depths from our framework, the accuracy of depth fusion decreased when the COLMAP-based source views were used during depth fusion. In addition, the COLMAP-based source views also show the decreased depth accuracy when used for depth estimation. The worst results were shown when the COLMAP-based view selection was used for both depth estimation and fusion. Therefore, the proposed mosaic array view selection method is shown to be more effective for depth estimation and fusion than the conventional method, and the reconstruction result was the best when using our method at each depth estimation and fusion process.

Table 2: Evaluation on reconstruction using geometric validation metric on different view selection methods applied in each of the depth estimation and depth fusion on the SAOI scene 1

Methods		Depth fusion methods			
		Mosaic array view selection		SfM-based view selection	
		Th_v=3	Th_v=4	Th_v=3	Th_v=4
Depth estimation methods	Mosaic array view selection	85.04	81.07	76.82	70.54
	SfM-based view selection	58.94	0.08	0.53	0.17

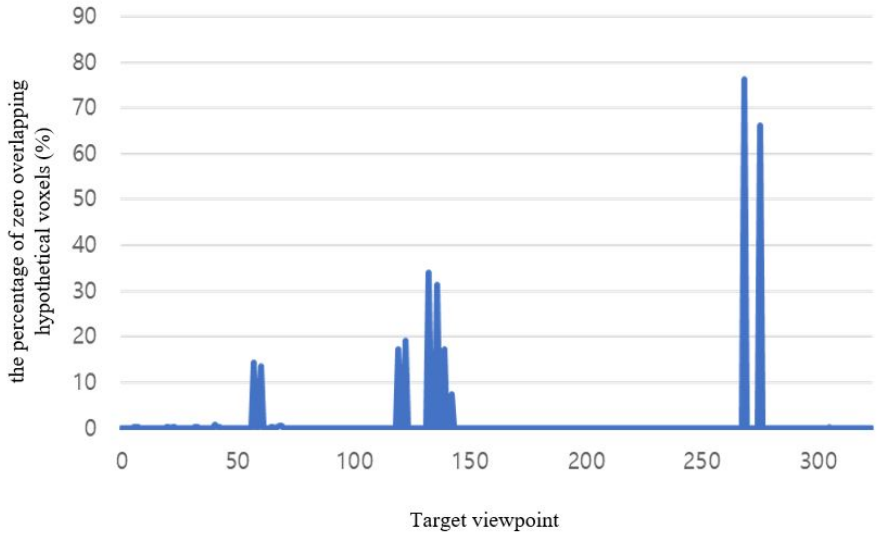
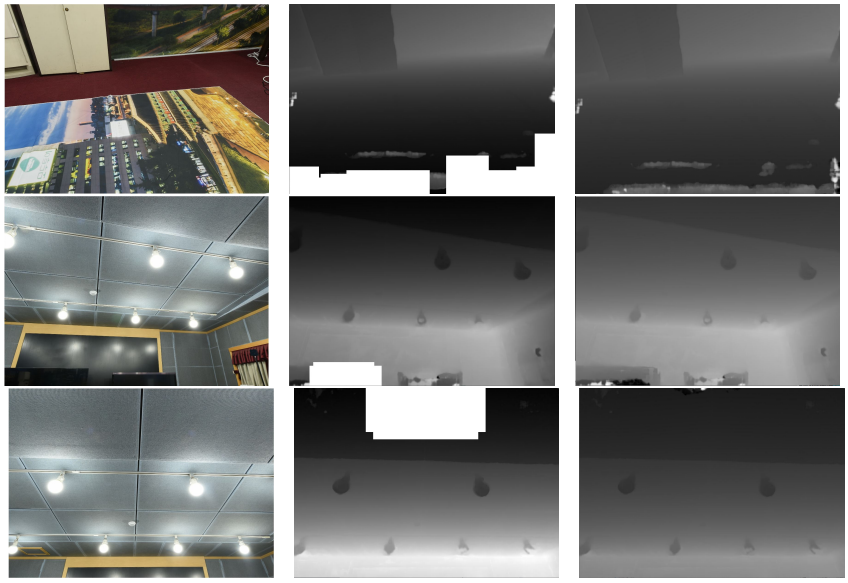


Fig. 13: The percentage of zero overlapping hypothetical voxels in the total valid view cost volume C_v for each viewpoint in scene 1.

3.3 Zero Overlapping Hypothetical Voxels

Fig. 13 shows the percentage of zero overlapping hypothetical voxel in the total valid view cost volume C_v for each viewpoint in scene 1. The figure shows that the viewpoints (target viewpoint images index 50-60, 125-145, 270-280) with large artifacts have a high proportion of zero overlapping-hypothetical voxels, which induced the saturated area in the corresponding depth map. Fig. 14 (**middle**) shows the impact of zero overlapping hypothetical voxels for those viewpoints



(**left**) (**middle**) (**right**)

Fig. 14: Depth map results before and after applying zero overlapping hypothetical voxel effect minimization. (**left**) is RGB images, and depth maps before and after applying zero overlapping hypothetical voxel effect minimization are (**middle**), (**right**), respectively.

with visibly saturated areas. As Fig. 14 (**right**) demonstrates, the saturated areas can be minimized by applying zero overlapping hypothetical voxel minimization. Also, the proposed method can be applied without affecting overall depth since the zero overlapping hypothetical voxels only occupy a small number of voxels on average. Out of the total $1152 \times 832 \times 8 \times 8 = 61{,}341{,}696$ voxels in the valid view cost volume $\mathbf{C}_V$, on average, the number of zero overlapping hypothetical zero voxels is about $580,000$, which is only 0.95% in the $\mathbf{C}_V$.

4 Code descriptions

Our code is based on cascade cost volume for high-resolution multi-view stereo and stereo matching [1]. The change logs are as follows. The main change part is building cost volume in casmvsnet.py (L46-48). The modified version in our code is in mosaicmvsnet.py (L36, L48-51, L65-72). Furthermore, we used post-processing code based on visibility-aware multi-view stereo network [6]. We modified the fusion.py to evaluate the performance, and made post-processing mask in fusioncas.py (by inserting L22-32, L153-154, L194, L276). Code is available at https://anonymous.4open.science/r/MosaicMVS-5EE6/.

References

1. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2492–2501 (2020)
2. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4104–4113 (2016)
3. Yang, J., Álvarez, J.M., Liu, M.: Self-supervised learning of depth inference for multi-view stereo. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7522–7530 (2021)
4. Yang, J., Mao, W., Álvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4876–4885 (2020)
5. Yu, A., Guo, W., Liu, B., Chen, X., Wang, X., Cao, X., Jiang, B.: Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction. ArXiv **abs/2011.12722** (2020)
6. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. British Machine Vision Conference (BMVC) (2020)
7. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018)