RC-MVSNet: Unsupervised Multi-View Stereo with Neural Rendering

-Supplementary-

Di Chang[1], Aljaž Božič[1], Tong Zhang[2], Qingsong Yan[3], Yingcong Chen[3], Sabine Süsstrunk[2], Matthias Nießner[1]

[1] Technical University of Munich
[2] École Polytechnique Fédérale de Lausanne
[3] Hong Kong University of Science and Technology
di.chang@tum.de

In this supplementary material, we present our results on the Tanks&Temples Advanced subset in Sec. 1 and describe depth map fusion in detail in Sec. 2. We provide details about cost volume regularization of the Cas-MVSNet backbone and implicit volume construction of our rendering consistency network in Sec. 3. We also show additional depth map prediction and point cloud reconstruction results in Sec. 4. Furthermore, we provide an ablation study on number of sampled rendering rays in Sec. 5. Finally, we discuss training strategy of pseudo label based multi-stage self-training and end-to-end unsupervised training in Sec. 6.

1 Performance on Tanks&Temples Advanced Benchmark

We train the proposed RC-MVSNet on the DTU training set, and test on the Tanks&Temples Advanced dataset without finetuning. We compare our method to state-of-the-art supervised [3, 6, 8, 11], pseudo-label-based multi-stage self-supervised method U-MVSNet [7]. Table 1 shows the evaluation results on the advanced subset. Our RC-MVSNet is the first end-to-end unsupervised method on advanced subset of Tanks&Temples. We achieve comparable performance to the multi-stage self-supervised method [7] and supervised methods [3, 6]. We also outperform supervised method CIDER [8] by **+7.7 (33%)** and R-MVSNet [11] by

2 Depth Map Fusion

After obtaining depth maps through RC-MVSNet, we need to convert them into 3D point clouds through depth map fusion for further evaluation. Following MVSNet [10] and R-MVSNet [11], we use geometric consistency and photometric consistency to remove occlusions and unreliable regions. Geometric consistency projects depth maps of source images into the reference image and masks out depth-inconsistency regions. To unify point cloud representation, we compute the average depth value of the consistent region. As for photometric consistency, we directly use the confidence maps generated by RC-MVSNet and only keep

the regions with high confidence. Finally, we re-project the pixels that satisfy geometric consistency and photometric consistency to the world coordinate system to generate point clouds.

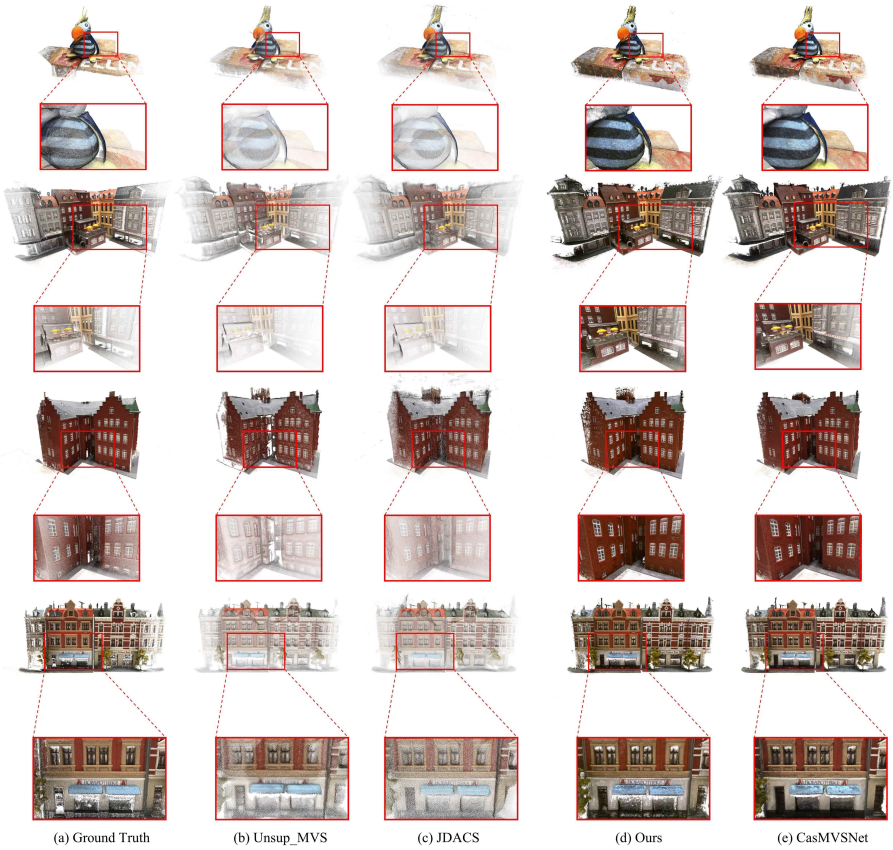

Fig. 1. Qualitative comparison of point cloud reconstructions on DTU.

3 Cost Volume and Implicit Neural Volume

Arbitrary fully-supervised MVS network could be used as our backbone in our framework – we use CasMVSNet [3] as default backbone. The 2D CNN extract latent features from N input views. Then the features from the $N-1$ source views are reprojected into the reference view via differential homography warping, as following:

$$H_j(d) = K_j \cdot R_j \cdot \left(I - \frac{(t_1 - t_j) \cdot n_1^T}{d} \right) \cdot R_1^T \cdot K_1^{-1} \tag{1}$$

Table 1. Point cloud evaluation results on the advanced subset of Tanks&Temples dataset [4]. Higher scores are better. The Mean is the average score of all scenes. The sections are partitioned into supervised, multi-stage self-supervised and end-to-end unsupervised, respectively. The best result is highlighted in bold for each section.

		Tanks&Temples advanced						
	Method	Mean↑	Auditorium↑	Ballroom↑	Courtroom↑	Museum↑	Palace↑	Temple↑
Supervised	CIDER [8]	23.12	12.77	24.94	25.01	33.64	19.18	23.15
	R-MVSNet [11]	24.91	12.55	29.09	25.06	38.68	19.14	24.96
	CasMVSNet [3]	31.12	19.81	**38.46**	29.10	**43.87**	27.36	28.11
	PatchmatchNet [6]	**32.31**	**23.69**	37.73	**30.04**	41.80	**28.31**	**32.29**
Multi-Stage Self-sup.	U-MVSNet [7]	**30.97**	**22.79**	**35.39**	**28.90**	**36.70**	**28.77**	**33.25**
E2E Unsup.	**RC-MVSNet**	**30.82**	**21.72**	**37.22**	**28.62**	**37.37**	**27.88**	**32.09**

where $H_j(d)$ denote the homography between the feature maps of the $j^{th}(2 \leq j \leq N)$ view and the reference feature map at depth d. The camera intrinsics K_j, rotations R_j and t_j are also given according to j^{th} view respectively. n_1 refers to the principle axis of the reference camera. The variance of these feature maps are calculated to construct a cost volume, which is regularized by 3D CNNs in each stage of the cascade structure. After the 3D convolutions, a pixel-wise depth map is regressed with soft-argmax upon the depth dimension of the probability volume.

For the reference volume V_1 of reference image I_1 and warped volume $\{V_j\}_{j=2}^N$ of source images $\{I_j\}_{j=2}^N$, the cost volume C in the backbone for depth estimation was constructed by Eq. 1. in our paper. This variance volume contains image appearance information and camera poses across all input views. However this volume is used for geometry reconstruction of depth inference, specifically for the reference view. We expect to use the information from *only* source views to synthesize the reference view. Hence, we calculate the variance volume C' from warped volume $\{V_j\}_{j=2}^N$ by:

$$C' = Var\left(V_2, \cdots, V_N\right) = \frac{\sum_{j=2}^{N}\left(V_j - \overline{V}'_j\right)^2}{N-1} \tag{2}$$

where Var denotes the same calculation and $\overline{V}'_j$ is the mean of warped volumes. In this way, we aggregate the information across $N-1$ source views construct the implicit neural volume.

4 Additional Qualitative Results

4.1 Depth Map Visualization on DTU Benchmark

Fig. 2 provides visualization of depth map of scans 4, 9, 10, 29 and 75 of DTU benchmark [1].

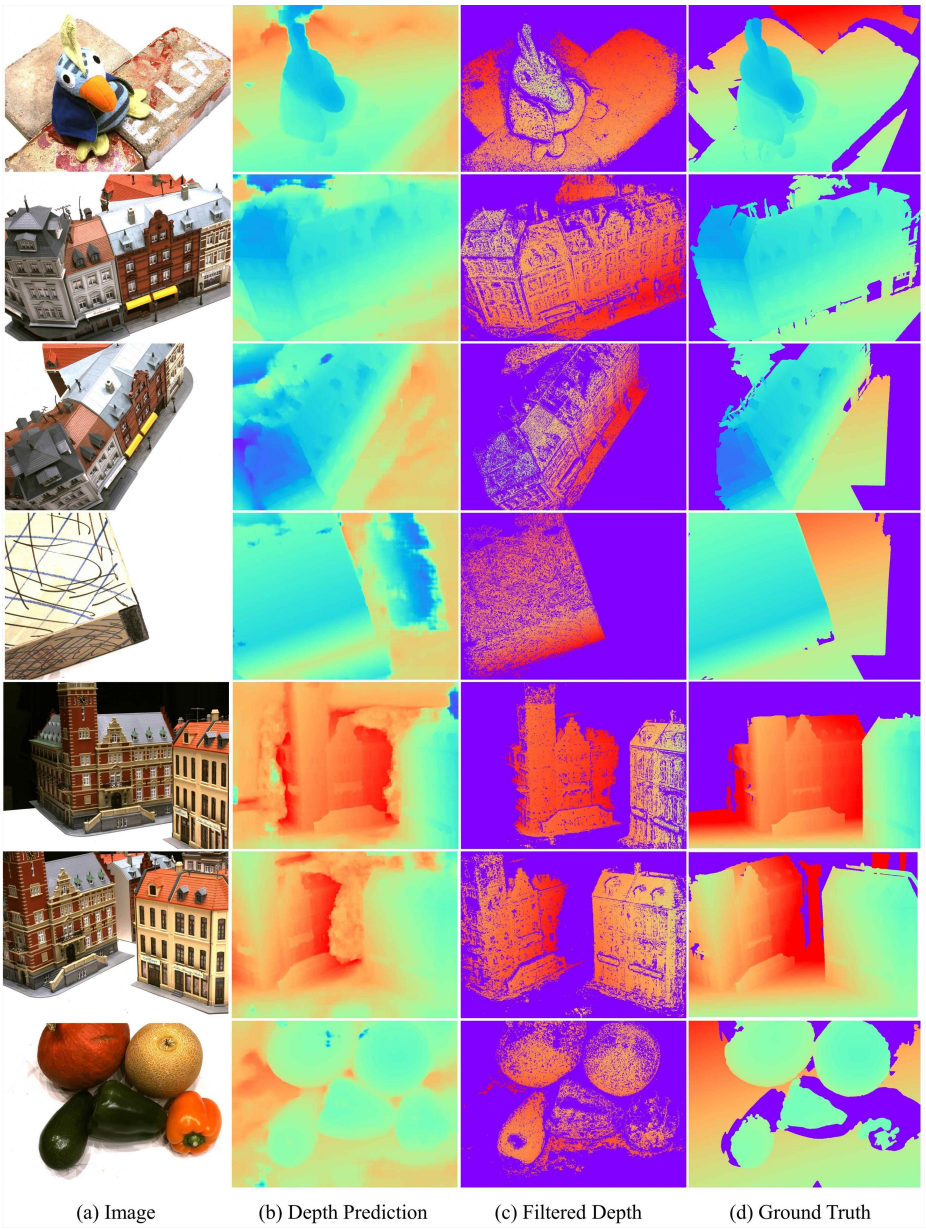

Fig. 2. Visualization of the prediction and filtered depth maps.

4.2 Point Cloud Visualization on DTU Benchmark

Fig. 1 provides additional reconstruction visualization on DTU benchmark[1]. Our unsupervised model shows significant improvement compared to previous state-of-the-arts, and achieves comparable reconstruction results to the supervised approach Cas-MVSNet[3].

4.3 Point Cloud Visualization on Tanks&Temples Benchmark

Fig. 3 visualizes additional point cloud reconstruction results on Tanks&Temples benchmark. Our method produces accurate and complete reconstructions. **+5.91 (24%)**.

Fig. 3. Qualitative results of point cloud reconstructed on Tanks&Temples.

5 Additional Ablation Study

Number of sampled rays for reference view synthesis Due to limited memory usage, we're not able to render complete depth maps and images during training times. Following common setting [2, 5], we only sampled a subset of

Table 2. Performance at different number of sampled rays during volumetric rendering

Num_rays	Acc↓	Comp↓	Overall↓	Train Mem	Train Img Size	Test Mem	Test Img Size
256	0.404	0.296	0.350	14.6 GB	640×512	7.5 GB	1600×1152
1024	0.396	**0.295**	**0.345**	15.5 GB	640×512	7.5 GB	1600×1152
4096	0.400	0.299	0.350	18.3 GB	640×512	7.5 GB	1600×1152
8192	**0.395**	0.300	0.348	23.3 GB	640×512	7.5 GB	1600×1152

rays during the volumetric rendering process. The performance of using different number of sampled rays is shown in Table. 2.

6 Discussion

As we described in the paper, multi-stage self-supervised methods suffer from complicated pre-training and pre-processing. The limitation of training time makes it difficult for these methods to be applied in practical scenarios. According to U-MVSNet [7], the pretraining of PWC-Net on DTU[1] and whole self-supervision training stage take 16 epochs in total. Then the post-training based on generated pseudo label takes further 16 epochs. We use the same backbone as them and it only takes 15 epochs to converge with 6 hours per epoch on NVIDIA RTX 3090. For self-supervised CVP-MVSNet [9], the self-training takes 15 hours per epoch on NVIDIA RTX 2080Ti. Hence, improving the efficiency of previous learning-based methods, both running time and memory usage, while maintaining comparable performance with self-supervised and supervised ones, can be regarded as one of our method's advantages.

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 **120**(2), 153–168 (2016)
2. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. arXiv preprint arXiv:2103.15595 (2021)
3. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: CVPR (2020)
4. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM ToG **36**(4), 1–13 (2017)
5. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
6. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: CVPR (2021)
7. Xu, H., Zhou, Z., Wang, Y., Kang, W., Sun, B., Li, H., Qiao, Y.: Digging into uncertainty in self-supervised multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6078–6087 (2021)
8. Xu, Q., Tao, W.: Learning inverse depth regression for multi-view stereo with correlation cost volume. In: AAAI. vol. 34 (2020)
9. Yang, J., Alvarez, J.M., Liu, M.: Self-supervised learning of depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7526–7534 (2021)
10. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
11. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: CVPR (2019)