Supplemental Materials
RayMVSNet: Learning Ray-based 1D Implicit Fields for Accurate Multi-View Stereo

Junhua Xi* Yifei Shi* Yijie Wang Yulan Guo Kai Xu†
National University of Defense Technology

1. More implementation details

We provide more details of the network architecture and experimental setting. Figure 1 shows the detailed network architecture of epipolar transformer. Four transformer layers are adopted to compute the cross-view feature correlation and aggregate the feature. In the figure, B represents the batch size, i.e. the number of sampled points in one batch. N denotes the number of input views. $C1$ and $C2$ are the channel dimension of the features. We set $C1 = 8$ and $C2 = 8$ in all the experiments.

Our method can take an arbitrary number of source images and does not require the number of source images to be the same for training and testing. This is because the features of multi-view images are first fed into a self-attention module and then processed by the mean/variance operation, outputting a fixed-length feature vector that is independent of the number of input views. For *DTU*, the network takes 3 and 4 images as input respectively during the training and testing. For the testing on *Tanks & temples* and *BlendedMVS*, the network takes 7 images as input.

2. Challenging test set of DTU

We create three challenging test subsets focusing on regions with *Specular reflection*, *Shadow* and *Occlusion*, respectively, from the DTU test set. To annotate these regions, we first select a reference image and two source images. For the specular reflection and shadow subsets, we select the regions whose appearance is significantly different in the reference image and the source images due to the influences caused by specular reflection and shadow, respectively. For the occlusion subset, we select the regions that are visible in the reference image and invisible in any of the two source images. The challenging test subsets contain 602 reference images and $\sim 100,000,000$ pixels of the challenging regions in total. Examples of the annotated regions are visualized in Figure 2.

*Joint first authors

†Corresponding author: kevin.kai.xu@gmail.com

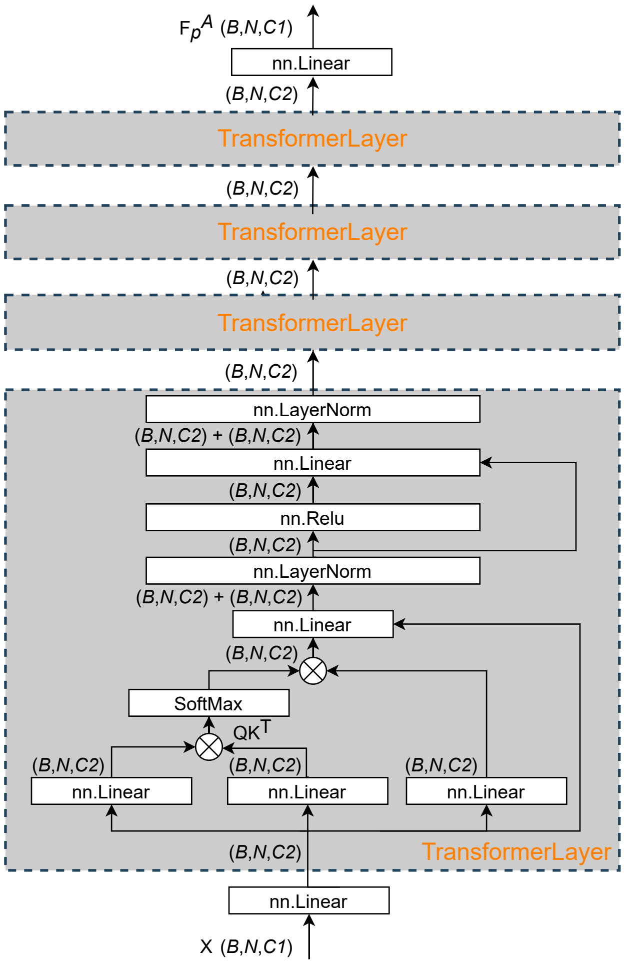

Figure 1. Network architecture of epipolar transformer.

3. More results of RayMVSNet

We provide more results of RayMVSNet in this section. Figure 3 shows visual comparisons against the baselines in terms of the depth estimation. Figure 5 contains additional qualitative results of RayMVSNet on DTU, Tanks & Temples, and BlendedMVS. In general, we see that RayMVS-

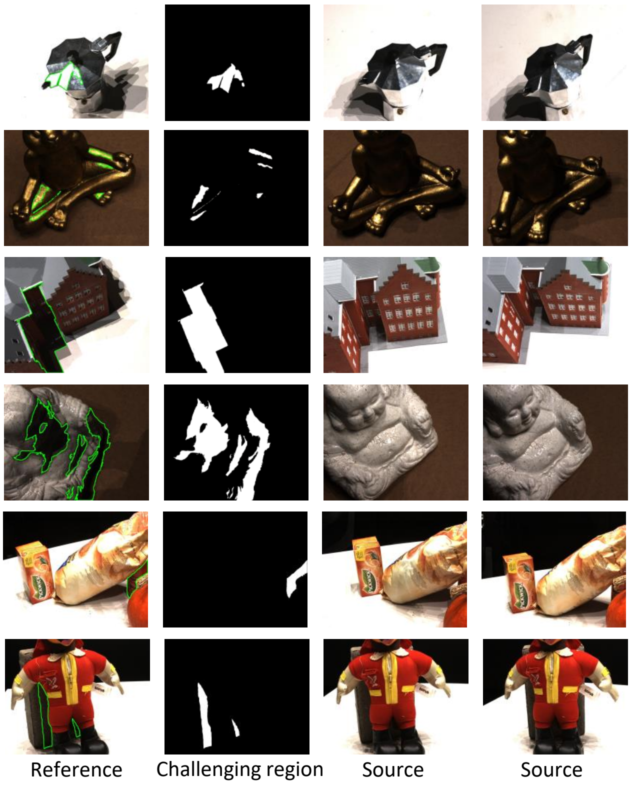

Figure 2. Examples of the annotated regions in the challenging subsets of DTU: *Specular reflection* (row 1-2), *Shadow* (row 3-4), and *Occlusion* (row 5-6).

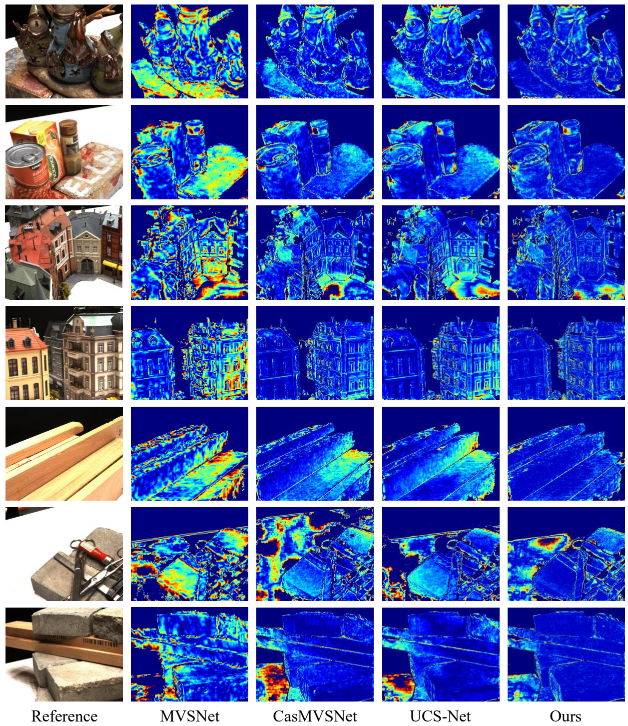

Figure 3. Visual comparison of the estimated depth map by RayMVSNet and the baselines.

Net achieves high-quality reconstruction in various scenes.

To further verify the efficiency, we compare RayMVSNet against the baselines by visualizing the relationship between the overall accuracy of the reconstructed point cloud and the GPU memory consumption. As shown in Figure 4, RayMVSNet achieves state-of-the-art performance and requires less GPU memory compared to most of the baselines. This demonstrates that RayMVSNet is light weight, thanks to the mechanism of ray-based representation.

Last, we conduct experiments of replacing the MVSNet with other MVSNet variants, e.g., UCS-MVSNet, Fast-MVSNet, and CVP-MVSNet, for coarse depth estimation. We found consistent improvement of depth estimation for the alternative backbones. In particular, our method with a UCS-MVSNet backbone achieves a 0.326 overall score on the DTU dataset.

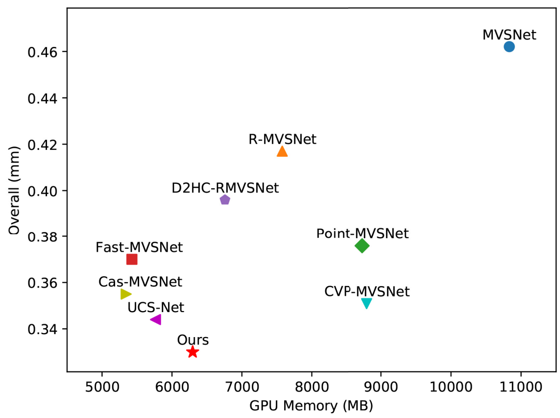

Figure 4. Visualizations of the overall reconstruction score and the GPU memory consumption. RayMVSNet achieves state-of-the-art performance and is light weight compared to most of the baselines.

4. Explanation of quantitative comparison on Tanks & Temples

We compared RayMVSNet against the baselines on the Tanks & Temples dataset. Since the experiment is designed for evaluating the generality of the proposed method, we only compare RayMVSNet to existing learning-based methods that trained on the DTU dataset. Methods that have been fine-tuned on other datasets (e.g. BlendedMVS) are not considered. Those methods include Att-MVSnet [2], Vis-MVSNet [1], AA-RMVSNet [4], and EPP-MVSNet [3].

Figure 5. Visualizations of the reconstructed point cloud on DTU, Tanks & temples, and BlendedMVS.

References

[1] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Visibility-aware point-based multi-view stereo network. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3695–3708, 2020. 2

[2] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020. 2

[3] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5732–5740, 2021. 2

[4] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6187–6196, 2021. 2