Reviewer 2

> This paper introduces a novel technique that refines depth maps from Multi-View Stereo (MVS) via iterative optimization of Neural Radiance Fields (NeRF). It synergizes the object surface depth estimation capabilities of MVS with the boundary depth estimation strengths of NeRF, incorporating Huber loss to enhance the precision of depth map refinement. Experiments conducted on the Redwood-3dscan and DTU datasets demonstrate the method's superiority over traditional approaches such as COLMAP, NeRF, and DS-NeRF. Moreover, it utilizes depth scale-invariant metrics alongside standard depth map estimation measures to validate its effectiveness.

Thank you very much for your careful review of our paper. Our responses and revisions to the points you mentioned are as follows.

> However, the study has certain shortcomings and areas that require clarification:
>
> 1. Experimental Design and Dataset Diversity: The research primarily utilizes Redwood-3dscan and DTU datasets for validation. To enhance the universality of the findings, it is recommended that the authors expand the experimental scope to include a wider variety of datasets, assessing the method's adaptability and robustness across different scenarios.

Thank you very much for pointing this out. Redwood consists of video images captured by an RGB-D camera, many of which are difficult to reconstruct for MVS, and the image quality is low because the data was captured by an RGB-D camera on hand. Therefore, we consider that Redwood is sufficient in terms of evaluating the adaptability of the MVS methods. DTU provides multi-view images acquired in a controlled environment, so the differences between the methods are smaller than with Redwood. We provide a supplemental explanation of the datasets as follows.

> (Line 289, Page 8)
>
> Therefore, multi-view images in Redwood are difficult to reconstruct 3D shapes by MVS due to external factors such as motion blur, noise, poor-textured objects, and illumination changes.

> (Line 303, Page 9)
>
> Multi-view images in DTU are acquired under the controlled environment. Therefore, we can evaluate the potential performance of MVS methods themselves since there are few external factors using DTU.

Since the dataset used in this experiment is multi-view images taken indoors, experiments using multi-view images taken outdoors will be required. However, the editor asked us to revise the paper in five days, which made it difficult to conduct the experiment with a new dataset in time. As an alternative, we have added results from multi-view images of outdoor objects taken with a standard camera as follows.

(Line 481, Page 15)

In this experiment, we use "Scan9," "Scan33," and "Scan118" from the DTU dataset, and multi-view images taken outdoors by the authors.

(Line 495, Page 15)

We evaluate the applicability of the proposed method by performing 3D reconstruction from multi-view images taken outdoors using an ordinary camera. The dataset consists of 35 RGB images of Shore to Shore," which is a 14-foot bronze-cast sculpture located in Vancouver's Stanley Park, Canada, taken by the authors in June 2023. Fig. 9 shows examples of images used in this experiment. It is a difficult situation to apply multi-view stereo and NeRF since not only the sculpture but also dynamic objects such as tourists are in the image. Fig. 10 shows the results of 3D reconstruction from multi-view images using COLMAP and the proposed method. COLMAP reconstructs the details of the sculpture, while there are many outliers on the object surface and at the object boundaries. The proposed method reconstructs the sculpture with high accuracy due to few outliers on the object surface. From the above, the proposed method can refine the depth maps estimated by COLMAP under real-world environments.

2. Comparative Analysis: While comparisons have been made with traditional methods like COLMAP, NeRF, and DS-NeRF, there is a lack of direct comparison with the latest related works. The authors are advised to update the literature review and compare their method against the state-of-the-art to highlight its uniqueness and advantages.

As you pointed out, there is a lack of comparison with the recent method. RC-MVSNet [18] is a method that combines MVS and NeRF, similar to the proposed method. We have added an overview of RC-MVSNet to the related work and added RC-MVSNet to the comparison method for our experiments as follows.

(Line 184, Page 5)

Recently, RC-MVSNet [18] has been proposed, which combines CasMVSNet [23] and NeRF to train CasMVSNet by unsupervised learning. Although unsupervised learning reduces the limitation on the number of training data, the depth map cannot always be estimated with high accuracy since NeRF is estimated based on the depth map generated by CasMVSNet.

(Line 315, Page 9)

In our experiments, we compare the accuracy of depth map estimation among COLMAP [4], NeRF [7], DS-NeRF [24], RC-MVSNet [18] and the proposed method to demonstrate the effectiveness of the proposed method.

(Line 342, Page 10)

For RC-MVSNet, we use the trained model and evaluation code available in the official GitHub repository[1]. The threshold of reprojection error used in depth map filtering is set to 0.5 pixels.

(Line 434, Page 12)

Tables 5 and 6 show the quantitative results for Redwood. COLMAP and NeRF have larger errors and lower accuracy than the other methods, indicating that the depths contain large errors. RC-MVSNet and the proposed method exhibit better results than other methods in most evaluation metrics. In particular, SILog for the proposed method is smaller than that for COLMAP, NeRF, DS-NeRF, and RC-MVSNet in most cases. This result indicates that the depth map refined by the proposed method contains fewer errors. Fig. 6 shows the depth maps estimated by each method. RC-MVSNet shows comparable results to the proposed method in the quantitative evaluation, however, it has more missing regins in the depth map compared to the other methods. The reason is that RC-MVSNet uses filtering of the depth map based on reprojection errors. Therefore, the estimated depths are highly accurate, while the depth maps include missing regions. The proposed method estimates the depth map more smoothly than the conventional methods. For example, the proposed method can estimate accurate and smooth depths of flat surfaces such as the floor and the ground in "amp#05668" and "childseat#04134." This is because the proposed method optimizes the radiance field based on the depth map estimated by COLMAP, unlike NeRF and DS-NeRF. These results indicate that the depth map estimated by COLMAP can be refined through iterative optimization of MLP representing the radiance field since the proposed method has fewer missing regions than the depth map estimated by COLMAP. On the other hand, neither COLMAP nor the proposed method can estimate the depth on the surface of the trashcan with poor texture in "trashcontainer#07226." In "travelingbag#01991," COLMAP has missing depths on the surface of the traveling bag, while the proposed method can smoothly estimate their depths. The difference between "trashcontainer#07226" and "travelingbag#01991" is the size of the missing region in the depth map estimated by COLMAP. If the missing regions in the input depth map are large, the proposed method cannot interpolate the depth map.

(Line 459, Page 14)

Table 7 shows the quantitative results for DTU. The proposed method exhibits better results than the conventional methods in most evaluation metrics. In particular, SILog for the proposed method is smaller or equal to that for COLMAP, NeRF, DS-NeRF, and RC-MVSNet. The proposed method has few large outliers in the depths since the errors are small and the accuracy is high as shown in Table 7. Fig. 7 shows the depth maps estimated by each method. All the methods

[1] `https://github.com/Boese0601/RC-MVSNet`

estimate depth maps with high accuracy in DTU. As mentioned in the experimental results for Redwood, RC-MVSNet stands out as missing regions compared to the other methods. NeRF, DS-NeRF, and the proposed method estimate depth maps even in poor-texture regions compared to COLMAP since the depth maps are synthesized from the radiance field.

3. Methodological Details and Transparency: To improve reproducibility, more detailed descriptions of the method, including parameter settings, network architecture, and optimization strategies, are suggested. Moreover, open-sourcing the code and datasets, where possible, would significantly enhance the impact of the work.

There was a lack of explanation about the training of the proposed method, so we have added the following.

(Line 341, Page 10)

In the experiments, Adam [**?**] is used as the optimizer. The learning rate begins at 5.0×10^{-4} and decays exponentially to 5.0×10^{-5} during the optimization process.

The network architecture of MLP used in the proposed method was shown in Fig. 2. This figure has been modified for clarity. Although the source code of the method proposed in this paper can be made publicly available, we did not have time to clean up the source code because we had only five days to revise the paper. We will release the source code in the future.

4. Limitations in Depth Map Refinement: The research focuses on depth map refinement but inadequately addresses challenges such as occlusions and complex textures. It is suggested that the authors delve deeper into these challenges and propose potential solutions or directions for improvement.

Occulusion can be addressed if the number of input viewpoints is large. Complex textures work well for MVS based on image matching because the textures are rich. Reviewer 1 gave us comments on the 3D reconstruction of transparent and translucent objects as a challenging task for MVS. The combination of photometric stereo and the proposed method has the potential to improve the accuracy of 3D reconstruction of transparent and translucent objects. We have added this to the conclusion as a challenging task to be investigated in the future.

(Line 516, Page 16)

One of the challenging tasks in MVS is to reconstruct the 3D shapes of transparent and translucent objects [33]. The method described in this paper cannot reconstruct the 3D shapes of transparent and translucent objects since the depth map estimated by COLMAP is used. The 3D shapes of transparent and translucent objects can be reconstructed by using photometric stereo, which estimates surface normals from images taken by a camera under varying lightings [34]. NeRF can

also consider the degree of transparency on the rays to take into account transparent and translucent objects. We expect that the combination of photometric stereo and the proposed method will be effective in addressing this task. Thus, we will consider refining the depth maps obtained by other MVS using the proposed method and also optimizing the camera parameters by NeRF in our framework.

5. Parameter Sensitivity Analysis: Including a series of charts that show how main parameters (e.g., iteration numbers, Huber loss thresholds) affect model performance would help understand the model's sensitivity to parameters and guide the setting of parameters in practical applications.

To confirm the parameter dependency of the proposed method, an ablation study was conducted on the ϵ in the Huber loss and the weight parameter λ_d of the loss functions as follows.

(Line 385, Page 11)

4.4 Ablation Study of Depth Loss

In this subsection, we describe an ablation study on the depth loss of the proposed method to confirm the dependence of the proposed method on the parameters. In this experiment, we use amp#05668 in Redwood.

4.4.1 Threshold of Huber Loss

The depth loss used in the proposed method is designed based on the Huber loss as described in Sect. 3.3.2. Huber loss uses L2 loss if the difference between the estimated depth and the true value is less than or equal to the threshold ϵ, otherwise L1 loss is used. Therefore, ϵ has an impact on the accuracy of the depth map estimation. Table 2 shows the accuracy of depth map estimation using the proposed method when Huber loss ϵ is multiplied by the scale factor s, where the numbers in bold and underlined indicate the best and second best in each evaluation metric, respectively. In the case of ϵ multiplied by 0.5, i.e., $s = 0.5$, AbsRel, SqRel, and RMSE, the accuracy of the depth estimation is the best, while SILog and $\delta < 1.25$ are the third most accurate. In the case of ϵ multiplied by 0.1 and 2, i.e., $s = 0.1, 2.0$, the accuracy of the depth estimation is degraded for most of the evaluation metrics. On the other hand, when ϵ is used, i.e., $s = 1.0$, the accuracy of depth estimation is within the top 2 in all the evaluation metrics. From the above, the proposed method employs $s = 1.0$ as a scale factor for the threshold ϵ for depth loss.

4.4.2 Hyper Parameter of Objective Function

The objective function used in the proposed method has a hyperparameter λ_D that adjusts the balance between the color reconstruction loss $\mathcal{L}_{Color}$ and the depth loss $\mathcal{L}_{Depth}$. In this experiment, we perform an ablation study on λ_D. Table 3 shows the accuracy of the depth map estimation of the proposed method when λ_D is changed. The accuracy of the depth map estimation of the proposed method is the highest when $\lambda_D = 0.1$. Therefore, $\lambda_D = 0.1$ is used in the following experiments.

6. Case Studies and Visual Results: Adding visual result comparisons for specific case studies, especially in complex scenarios (e.g., varying lighting conditions, occlusions), would visually demonstrate the method's advantages and limitations.

As mentioned in the response to comment 1, Redwood consists of multi-view images taken under uncontrolled environments. Therefore, it is a complex condition in the datasets used to evaluate the accuracy of MVS. We could add experiments for evaluation on other datasets, but time did not allow us to conduct the experiments since the paper revision period is only 5 days. As mentioned in our response to comment 1, we have added the reconstruction results from the multi-view images we took outdoors as an alternative.

7. Practical Applications and Case Studies: Although the method performs well on selected datasets, there is a lack of case analysis in real-world application scenarios. Providing examples of specific applications, such as cultural heritage reconstruction or virtual reality, would help readers better understand the method's potential value and application range.

As mentioned in our response to comment 1, we have added the reconstruction results from the multi-view images we took outdoors to demonstrate the potential of the proposed methon in real-world applications.

These suggestions aim to further enhance the study's academic value and practicality, providing valuable reference for future research.

As mentioned in our responses to Reviewer 1 and comment 6, we have added a challenging task to be investigated in the future to the conclusion.

The English quality is satisfactory, adhering to the standard conventions of academic writing. However, minor improvements could be made in terms of proofreading to correct occasional grammatical errors and enhance clarity in certain sections.

Thank you very much for your kind attention. We have carefully checked the paper and corrected English errors

Thank you very much for your kind comments.