Neural Radiance Fields Approach to Deep Multi-View Photometric Stereo [Supplementary Material]

Berk Kaya[1] Suryansh Kumar[1] Francesco Sarno[1] Vittorio Ferrari[2] Luc Van Gool[1,3]
Computer Vision Lab, ETH Zürich[1], Google Research[2], KU Leuven[3]

Abstract

Here, we extend the experimental section from the main paper. Our supplementary document is organized as follows: First, we present the visual comparison of the rendered images with our method and other view synthesis methods. We also show the advantage of using surface normals via training and validation curves. The following sections provide further analysis by comprehensive experiments. Concretely, we demonstrate the effect of change in the light source direction between the frames on NeRF [2]. We also investigate the importance of viewing directions and sampling strategy on our method. Later, we provide some additional implementation details related to our method. ***Finally, we reemphasize the main motivation of this work in the concluding remarks section****.*

1. Comparison with View Synthesis Methods

In this section, we compare rendering performances of view-synthesis methods.

Visual Comparison. Fig.1 shows the images rendered by IDR [4], NeRF [2] and our method. All the methods use the DiLiGenT-MV images captured with the same light source. We observed that the IDR method fails to capture the geometry and appearance information with this setting. The NeRF method performs much better than IDR; however, the rendered images are often blurry and lack surface details. We observed that our method performs significantly better than both methods, and it can generate important details of the object. For example, our approach renders the nose of the BEAR and the eyes of the BUDDHA very accurately. These details are not apparent with NeRF due to the missing surface normal information in rendering.

Training and Validation Analysis. Our method combines photometric stereo surface normals in the continuous volume rendering process for better image rendering. The surface normals obtained using the photometric stereo take care of shading in the image formation process. And therefore, it can be observed from the plots presented in Fig.2, that our method clearly shows better convergence behavior. In addition to the BUDDHA scene presented in the main paper, we analyze the loss curves of the remaining DiLiGenT-MV objects. Fig.2 shows the training and validation curves of our method and NeRF for BEAR, COW, POT2, and READING. Our method has a lower loss value in all of the categories, indicating that the image rendering quality is better.

2. Effect of Multiple Light Sources on NeRF

We want to study the effect of change in the light source direction over MVPS images. This setting has not been studied before, where the light is different across images that are captured from multiple views. To simulate this experiment, the subject in the scene should have the appropriate material properties and surface area to show the change of light direction over images. In this paper, to conduct this experiment, we choose the suitable example that has complex surface profile and has significant surface area to really capture the effects of change in the light source direction across images. Hence, we choose COW and BUDDHA as suitable examples from DiLiGenT-MV dataset to simulate this experiment.

NeRF enforces the learned scene representation to be multi-view consistent by learning position-dependent volume density σ, and it renders images by taking view direction into account. For this reason, having constant lighting on the scene is required, and the core idea of the method is questionable if lighting varies across camera viewpoints. In this section, we study the behavior of NeRF under multiple light sources. For our experiment, we assign a different light source for each camera viewpoint. To be precise, we randomly pick 20 light sources from 96. Table 1 compares the 3D reconstruction accuracy achieved using images from the same light source and multiple sources.

Method	BUDDHA	COW
NeRF with single light source	0.99	0.9
NeRF with multiple light sources	1.30	1.06

Table 1. Comparison of reconstruction accuracy achieved by NeRF with single light source and multiple light sources. We report Chamfer-L1 distance for the comparison.

3. Effect of Viewing Direction

Here, we want to study the effect of viewing direction on rendered image quality obtained using our method for this experiment. To that end, we remove view direction information $\gamma(\mathbf{d})$ from our MLP. Table 2 compares the quality of image rendering with and without view dependence. As expected, the image quality sharply decreases without the view direction. So we conclude that similar to surface normals, view direction is also crucial for the rendering.

Method	BEAR	BUDDHA	COW	POT2	READING
Ours without view dependence	31.71	29.76	30.26	30.28	29.41
Ours with view dependence	**37.16**	**33.59**	**34.49**	**30.47**	**30.46**

Table 2. Quantitative image rendering quality measurement with PSNR metric with and without view dependence (The higher the better).

4. Effect of Volume Sampling

Our method uniformly samples points along the ray between near and far bounds t_n, t_f. Increasing the number of these query points enables a denser evaluation of the network. Still, it is computationally not feasible to sample a lot of points uniformly. To make the process more efficient, we use a two-stage hierarchical volume sampling strategy by optimizing coarse and fine networks simultaneously. For that, we first consider the coarse network rendering:

$$\tilde{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} \mathbf{w}_i(\mathbf{x}_i)\mathbf{c}_i(\mathbf{x}_i, \mathbf{n}_i^{ps}, \mathbf{d}), \text{where } \mathbf{w}_i(\mathbf{x}_i) = T_i\Big(1 - \exp\big(\sigma(\mathbf{x}_i)\delta_i\big)\Big) \tag{1}$$

We calculate weights $\hat{\mathbf{w}}_i(\mathbf{x}_i) = \mathbf{w}_i(\mathbf{x}_i)/\sum_{j=1}^{N_c}\mathbf{w}_j(\mathbf{x}_j)$ to have a probability density function on the ray. Then, we sample fine points from this distribution using inverse transform sampling. For the coarse network, we sample $N_c = 64$ points uniformly. For the fine network, we sample $N_f = 128$ points by taking the coarse network weights into account.

To show the effectiveness of our two-stage sampling strategy, we simulate an experiment. To that end, we remove fine sampling from our approach and evaluate the performance by using only the uniformly sampled points. Table 3 reports the 3D reconstruction accuracy using 64 and 256 coarse samples only, as well as the two-stage approach of using both coarse and fine sampling. The results suggest that choosing N_c and N_f introduces a better trade-off between computation time and accuracy.

Volume Sampling	BEAR	BUDDHA	COW	POT2	READING
$N_c = 64, N_f = 0$	0.85	1.45	0.86	0.63	1.38
$N_c = 256, N_f = 0$	0.68	0.92	1.01	0.64	1.64
$N_c = 64, N_f = 128$	0.66	1.00	0.71	0.63	0.82

Table 3. Reconstruction accuracy achieved with different number of points. We provide scores using Chamfer-L1 distance metric.

5. Additional Implementation Details

Deep Photometric Stereo Network: The deep photometric stereo network is trained on the synthetic CyclesPS dataset [1]. CyclesPS has 15 different shapes, and for each shape, three sets of images are rendered with varying material properties. For training, the effect of different light sources on each pixel is represented using an observation map. The data is further augmented by applying ten different rotations. Concretely, the same rotation is applied to the observation maps and corresponding ground-truth normals. The idea is that rotation of surface normals, and light directions around the view direction of photometric stereo setup do not alter the image value for the isotropic surfaces. For more information, we refer the readers to CNN-PS work [1].

The observation maps are rotated ten times by uniformly picking angles in the range $[0, 360]$ during testing. We run inference on each observation map separately and aggregate the results by simply applying inverse rotations. We then average and normalize these vectors to get the per-pixel normal estimate. This strategy improves the accuracy of the normal estimations.

MLP Optimization for MVPS: Our MLP is implemented using PyTorch [3]. While comparing our method against other standalone MVS methods, we picked the images coming from the same light source in the DiLiGenT-MV dataset. We observed that the fourth light source provides a consistent surface profile image throughout the image sequence. And therefore, we use it for evaluating all MVS methods. Our method and NeRF also require a common threshold value for getting the 3D. For this reason, we extracted meshes using density thresholds of 1, 5, 10, 20, 50, and 100 using NeRF method. We observed that choosing the density threshold of 10 results in the best performance for NeRF, and therefore we applied the same threshold to our method during mesh extraction.

6. Concluding Remarks

Firstly, we want to indicate that MVPS is not an ordinary 3D data acquisition setup that can be realized with common commodity cameras. It requires sophisticated hardware, and special care must be taken to calibrate cameras and lights. Only then, it becomes possible to acquire 3D and render scenes accurately. Secondly, we want to emphasize again that MVPS is generally solved using a sequence of involved steps. Hence, the main motivation of this work is to utilize the modern approach for the classical MVPS problem and explore how far we can go with it (with a framework that is as simple as possible). This work shows that we can get closer to state-of-the-art multi-stage MVPS methods with a much simpler framework by leveraging the continuous volumetric rendering approach. All in all, this work provides a new way to solve MVPS, and maybe working on such ideas can help us come up with a better and a simpler way to recover 3D from MVPS images.

References

[1] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020.

[3] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[4] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction with implicit lighting and material. *arXiv preprint arXiv:2003.09852*, 2020.

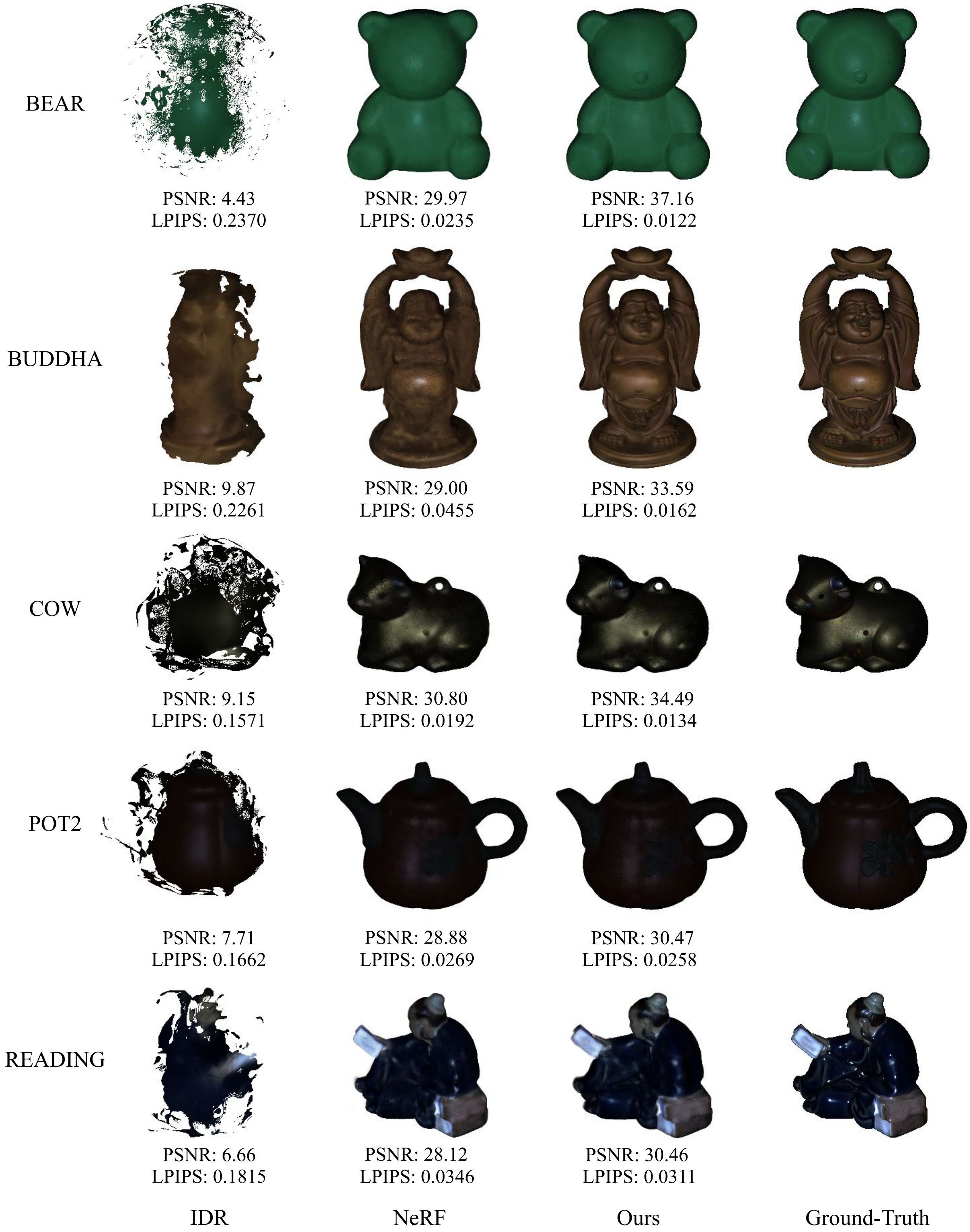
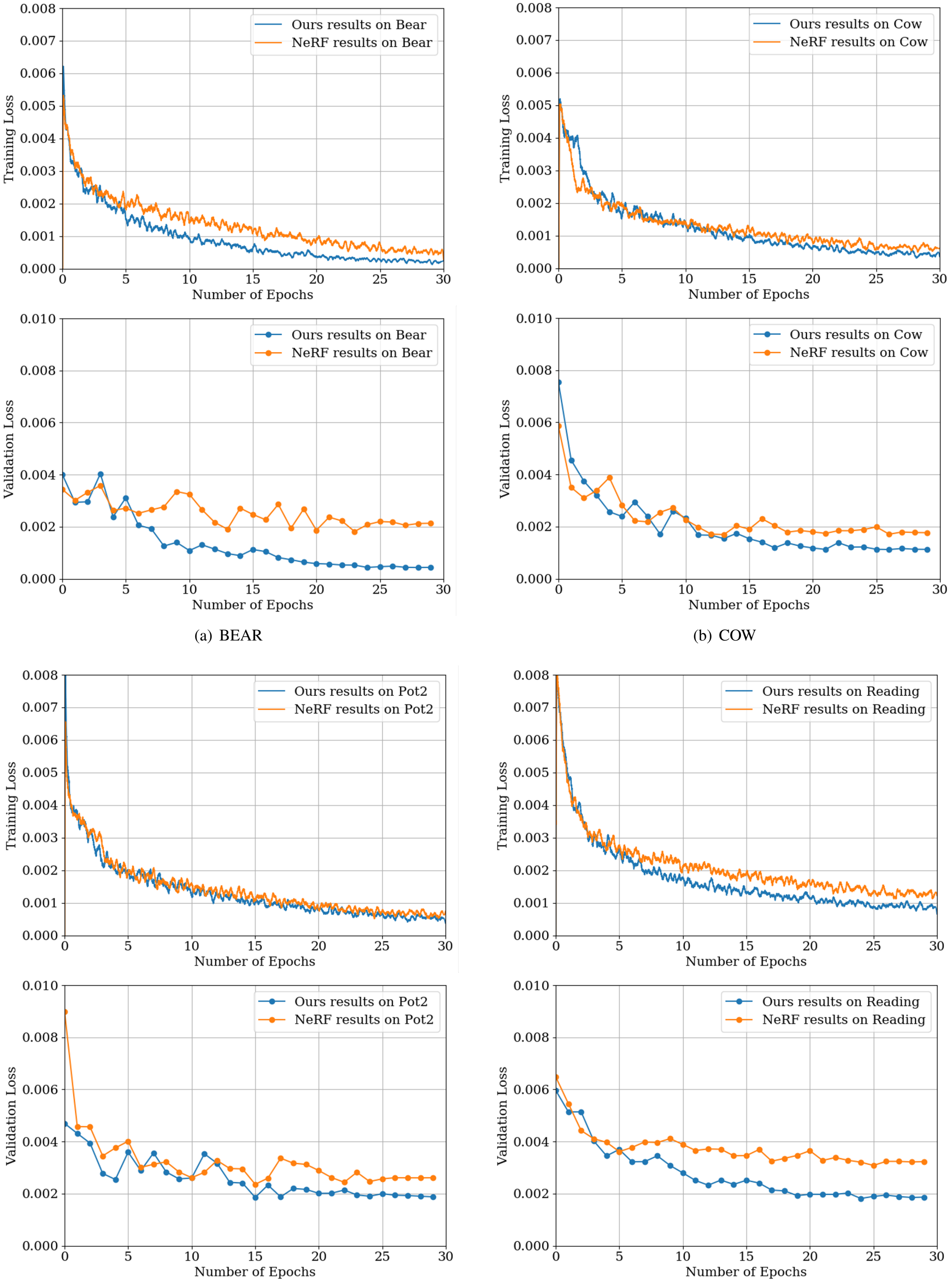


Figure 1. Visual comparison on DiLiGenT-MV renderings achieved by IDR [4], NeRF [2] and our method. Without surface normals, NeRF lacks in details and produces blurry renderings. On the other hand, our method is able to recover fine details and render accurate images by blending surface normal information in volume rendering process. We observed that IDR framework cannot recover the geometry and appearance on this benchmark.



(a) BEAR

(b) COW

(c) POT2

(d) READING

Figure 2. Training and validation curves of BEAR (a), COW(b), POT2(c) and READING(d) using our method and NeRF. Our method consistently shows better convergence behavior with the contribution of the surface normal information.