Supplementary Material: Occlusion-Free Scene Recovery via Neural Radiance Fields

Chengxuan Zhu[1,2] Renjie Wan[3] Yunkai Tang[1,2] Boxin Shi[1,2*]
[1]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[3]Department of Computer Science, Hong Kong Baptist University
{peterzhu, shiboxin}@pku.edu.cn, renjiewan@hkbu.edu.hk, tangyunkai@stu.pku.edu.cn

A. Uniqueness of our method

For easy reference and comparison, we categorize our work as "a novel method for occlusion removal" in the main paper. But it has essential differences from existing occlusion removal methods. While existing methods mostly remove occlusions for given viewpoints, our method can synthesize occlusion-free images with novel viewpoints, by building an occlusion-free scene *representation*. Moreover, without relying on external supervision of specific occlusion types for training, our method is capable of handling more diverse occlusions in the real world when their surrounding environment is given. Examples can be found in Fig. 4, Fig. 5, and Fig. 6 (in the main paper), and Fig. S3 in this document. Readers are encouraged to visit our project page[1] to watch the animated results

B. Additional analysis to key modules

Two modules ensures the effectiveness of our method: a joint optimization of camera parameters and scene MLP guaranteeing high-quality scene reconstruction, and a supervision mask enabling the selective training of the background MLP to learn the background scene representation only.

Our method is conducted in a selective manner. In the joint optimization, the pose refinement can select similar features belonging to background or occlusion by building a cost volume, which ensures a high-quality representation of whole scenes. For the subsequent selective supervision, we use a mask MLP to generate a supervision mask to selectively learn the background representation.

We will dig deeper into the each module with more explanations and examples here.

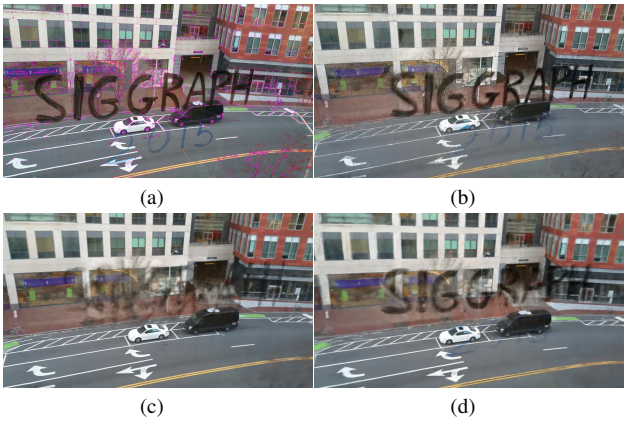


Figure S1. The SCRIBBLE2 scene in our dataset for illustration of (a) the feature extraction and matching process of COLMAP [9]; (b) the rendered novel views by vanilla NeRF [8], based on the pose estimated by COLMAP [9]; (c) the rendered novel views by NeRF−− [10]; (d) the rendered results with occlusion by the scene MLP in our method. The red circles in (a) refer to feature points not matched with any other image, while the magenta ones refer to those matched with some other images (please zoom-in for details). Please use **Adobe Acrobat** or **KDE Okular** to see the animated results in (b), (c) and (d).

B.1. Scene reconstruction with cost volume

An example is displayed in Fig. S1 to show the importance of joint optimization. One motivation for joint optimization is that the camera poses generated by COLMAP [9] are not stable when occlusions are presented. We notice that the features estimated by COLMAP [9] tend to be dominated by one of the layers, *e.g.* the SCRIBBLE2 scene in Fig. S1(a) where background dominates the pose estimation process. This means COLMAP [9] cannot calculate the accurate spatial location of the scene as a whole, including both the background scene and the foreground occlusions. The estimated poses by COLMAP [9] lead to the

*Corresponding author.
[1]https://freebutuselesssoul.github.io/occnerf

(a) (b)

Figure S2. (a) The synthesized novel view by our scene MLP; (b) corresponding supervision mask P. White regions in the supervision mask refer to the regions used to supervise the training of background MLP. Please use **Adobe Acrobat** or **KDE Okular** to see the animated results.

unstable position changes of scribbles in Fig. S1(b), which is far from a reliable representation of the whole scene.

We seek cost volume to achieve a pose refinement. As a commonly-used technique in multi-view stereo methods, cost volume shows robustness in reconstructing the scene [1, 3, 12]. In short, the cost volume is calculated by the variance of the features from multiple viewpoints. Assuming a 3D point appears similar in neighboring views, the cost volume can provide the reconstruction with a hint about the likelihood of an actual point's presence in any given position, which helps the pose refinement select the suitable features to learn camera poses. In Fig. S2(a), we show the synthesized novel views of the whole scene using the scene MLP, to illustrate that we indeed obtain a reliable scene representation and camera poses.

B.2. Selective supervision with depth constraints

With a faithful representation of the whole scene, we only need a reasonable way to selectively learn the representation of desired background scenes from it. In our method, we leverage depth to achieve this goal. By definition, occlusion is something that blocks the view of the object, which means the ray comes across other objects before it reaches the background object in rendering.

We, therefore, use the depth constraints to describe such spatial correlation by assuming the greater bidirectional depth difference of occlusions, which is, intuitively, the depth difference between the most likely foreground object and the most likely background object. Under our assumption, the smaller the bidirectional depth difference, the less likely there are occlusions along the ray.

In this way, our method can judge the presence foreground occlusion without being trained on excessive data to learn what an occlusion may look like. In Fig. 2 of the main paper, we show the supervision mask for FENCE1, and in Fig. S2(b) we show the supervision mask of more scenes with irregularly-shaped occlusions in our dataset, namely the scenes of STATUE, WIRE1, and WIRE2.

C. More results

We choose PWC-Net [6] as a baseline to compare with the classical occlusion removal methods. Though several occlusion removal methods based on image sequences have been proposed [4, 5, 11], none of them comes with available codes for faithful evaluation. Also, PWC-Net [6] is the latest work to our knowledge, that claims to show occlusion removal ability in various scenes [6]. We demonstrate more results of occlusion-free novel views on the remaining scenes in our dataset (in addition to those already displayed in Fig. 4 of the main paper). Specifically, we test in the scenes of SCRIBBLE2, WIRE2, SCRIBBLE3, FENCE2, and FENCE3. As shown in Fig. S3, for each scene, we show the recovered background scene using background MLP as well as some baseline methods. Note that all of the COLMAP-based methods fail with COLMAP [9] in the scene of WIRE2. For FENCE2, the results of Ha-NeRF [2] seem to recover a clearer background, but it also wrongly removes white textures on the playground. For another example FENCE3 in our dataset, the completeness of background is still the best among all compared methods, albeit some remaining part of occlusions in the result at top left corner.

Besides, we also show animated results with larger camera motion in Fig. S4.

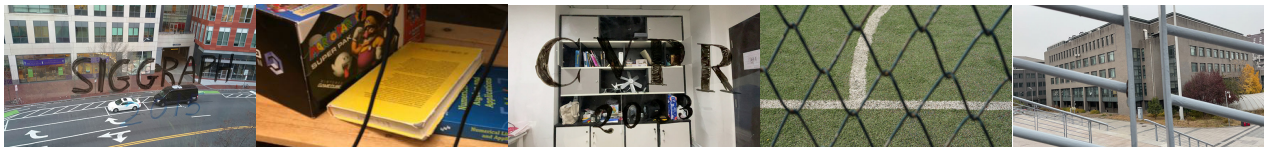
References

[1] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proc. of International Conference on Computer Vision*, 2021. 2

[2] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *CVPR*, 2022. 2, 3

[3] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 2

[4] Monika Kwiatkowski and Olaf Hellwich. Specularity, shadow, and occlusion removal from image sequences using deep residual sets. In *Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2022. 2



(a) A sample view of input scene

(b) Vanilla NeRF [8]

(c) PWC-Net+NeRF [6, 8]

(d) Ha-NeRF [2]

(e) NeRF-W [7]

(f) Our method

Figure S3. From left to right, we show animated novel view synthesis on SCRIBBLE2, WIRE2, SCRIBBLE3, FENCE2, and FENCE3. A sample view of the input scene is shown in (a) for reference. From (b) to (f), we show results obtained by (b) vanilla NeRF [8], (c) PWC-Net [6] + NeRF [8], (d) Ha-NeRF [2], (e) NeRF-W [7], and (f) our method. Please use **Adobe Acrobat** or **KDE Okular** to see the animated results. Some results show stronger variation over the animation due to failure in extracting consistent background.

(a) FENCE1 (b) RAINDROP

Figure S4. The synthesized novel views of the proposed method with larger camera motion. Please use **Adobe Acrobat** or **KDE Okular** to see the animated results.

[5] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Let's see clearly: Contaminant artifact removal for moving cameras. In *Proc. of International Conference on Computer Vision*, 2021. 2

[6] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *CVPR*, 2020. 2, 3

[7] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3

[8] 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. In *Proc. of European Conference on Computer Vision*, 2020. 1, 3

[9] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2

[10] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 1

[11] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *TOG*, 2015. 2

[12] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proc. of European Conference on Computer Vision*, 2018. 2