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

Abstract

Our everyday lives are filled with occlusions that we strive to see through. By aggregating desired background information from different viewpoints, we can easily eliminate such occlusions without any external occlusion-free supervision. Though several occlusion removal methods have been proposed to empower machine vision systems with such ability, their performances are still unsatisfactory due to reliance on external supervision. We propose a novel method for occlusion removal by directly building a mapping between position and viewing angles and the corresponding occlusion-free scene details leveraging Neural Radiance Fields (NeRF). We also develop an effective scheme to jointly optimize camera parameters and scene reconstruction when occlusions are present. An additional depth constraint is applied to supervise the entire optimization without labeled external data for training. The experimental results on existing and newly collected datasets validate the effectiveness of our method. Our project page: https://freebutuselesssoul.github.io/occ_nerf

1. Introduction

Neural Radiance Fields (NeRF) are capable of learning the scene representation implicitly from a set of 2D images, yet not every scene is favored by observers. Many undesirable occlusions in our world obscure details that are essential to our understanding of the world. In general, such obstructions range from water droplets and scribbles on a piece of glass, to fences or any objects occluding the desired scenes (*e.g.*, a statue closer to the camera in a landmark scene). How to apply computational methods to exclude them from the scene representation is of great interest.

Occlusion removal (*e.g.*, [29]) is the direct solution to achieve this goal. However, explicit occlusion removal may oversmooth essential details necessary for clearly observing the desired background scenes. In addition, current methods mainly depend on external occlusion-free supervision (*e.g.*, fence removal [4], raindrop removal [22]) to develop the reliable capability in removing certain types of occlusions. Once encountering a new scenario with unseen occlusion types beyond their training data, these methods might show degraded performances. To handle more diverse types of occlusions, generic constraints from multiple viewpoints are widely adopted [4, 10, 12, 14, 19, 27, 29] via mimicking our human vision systems, who can easily piece together the desired background scenes by looking at them from different viewpoints. But the majority of these methods just consider viewpoints as a prior in relation to spatial correlations. Their backbones still rely on external training data with corresponding ground truth for optimization, which still does not fundamentally alleviate the difficulty of handling diverse occlusions in the real world.

An occlusion-free world can be progressively aggregated by seeing its occluded part from different viewing directions to reveal occlusions previously unobservable in each single perspective, as illustrated in the left part of Fig. 1 (the fan is occluded in the target view). Since NeRF [18] employs an implicit representation to map viewpoints to pixels, one may come to the naive solution of directly constructing a NeRF which is optimized across multiple viewpoints. However, the vanilla NeRF [18] representing the scene as a whole is not able to treat occlusion and background scenes distinctively, and as long as the occlusion remains static, NeRF is designed to faithfully reconstruct its presence. Meanwhile, many NeRF variants can decompose the whole scene into different components (*e.g.*, NeRF-W [17], Ha-NeRF [3], NeRFReN [7]), but they cannot handle the real-world static occlusions. This is because NeRF-W [17] and Ha-NeRF [3] rely on the inconsistency of undesired components across different views to achieve such separation, which is difficult to be observed in a continuous 3D world. On the other hand, NeRFReN [7] only works in separating the transmission and reflection compo-

*Corresponding author.

nents caused by semi-transparent planar glass, which is incapable of handling opaque occlusions in the real world.

Another problem comes from NeRF's reliance on camera parameters pre-computed by COLMAP [23], because handcrafted features extracted and matched using COLMAP [23] are for the whole scene, and are incapable in distinguishing between undesired occlusions and the desired background. When the features from occlusion dominate the matching process, the obtained camera parameters cannot faithfully model the spatial correlation of background scenes across multiple viewpoints. Besides, COLMAP [23] is not a stable option for pose estimation in the real world [28]. The existence of occlusions may prevent it from working properly, making the occlusion-free scene representation infeasible.

In this paper, we aim at seeing through the occluded scenes by developing an occlusion-free scene representation without considering specific occlusion types, based on which we can render any occlusion-free images from desired viewpoints. Our method first maps viewing angles and their corresponding scene details by leveraging NeRF. We then introduce a depth constraint to probe the occluded areas by measuring the depth of occlusion and background, by assuming that *occlusions are always in the foreground with closer distance*. During the scene modeling process, a pose refinement scheme is further introduced to refine the camera pose with the features of the background scene. As outlined in Fig. 1, our pipeline contains three modules to achieve the above goals: 1) a scene reconstruction module to represent the whole scene using NeRF (with occlusions), 2) a cost volume construction module to gather information from neighboring views as guidance (to indicate where occlusions are), and 3) a selective supervision scheme to constrain another NeRF on the desired background information (occlusions removed), and our contributions can be summarized as follows:

- an occlusion-free representation without relying on any external prior as supervisory knowledge;
- a joint optimization of pose refinement and scene reconstruction by effective multi-view feature fusion;
- a selective supervision scheme to probe the occluded areas guided by the scene depth information.

Based on the experiments with a dataset containing diverse types of occlusions, the proposed method can eliminate occlusions including scribbles and water droplets on a piece of glass, fences, and even irregular-shaped statues without relying on any external supervisions.

2. Related Work

2.1. NeRF and its variants

NeRF becomes a popular choice for implicit volumetric scene representation [18]. By mapping the point and current viewing direction to its color and density using a neural network and a differentiable rendering scheme, it provides an effective way to learn from multiple viewing angles. Such ability has been explored to complement the missing depth for grasping transparent objects [9]. Based on NeRF, a series of methods have been proposed to tackle the problems like scene understanding [33] and reasoning [26]. Besides, by considering specific physical priors, NeRF-based frameworks also show promising performance in learning clear representation from challenging scenarios [7, 8, 16]. Recent NeRF variants can even reconstruct the scene from input with various perturbations. For example, NeRF-W [17] and Ha-NeRF [3] can separate the transient objects from the whole scene. However, they target occlusions that are inconsistent throughout the image set and become ineffective when such occlusions remain static and consistent.

Since NeRF relies on pre-computed camera parameters for scene representation, it fails when the pre-computation is not feasible. Some methods (*e.g.*, NeRF−− [28] and BARF [13]) have been proposed to alleviate this issue by optimizing the camera parameters along with the scene representation. However, how to avoid the interference from undesired scene during the camera parameter optimization remains a problem to be solved.

2.2. Occlusion removal

Traditional methods in occlusion removal address this problem by propagating neighboring pixels via anisotropic diffusion [2] or solving differential equations [1]. Recently, there are more approaches solving this problem in a learning-based manner [11, 34]. A recent trend is to employ the edge information [21] or segmentation masks [24] as a stronger prior to alleviate the ill-posedness of this problem.

Researchers in this area have been aware of the importance of removing occlusions from multiple viewpoints. Traditional methods try to take the benefits of multiple viewpoints by computing a disparity map [10], dense flow field [29], or visual parallax [19]. Using a light field camera that records the spatial and angular information of light rays in the space is also helpful in removing occlusions [25]. However, these methods also pose restrictions on their inputs, such as using stereo image pairs [10], requiring the background to be visible in at least one of the input views [25, 29], or requesting obviously relative motion between occlusions and backgrounds [19]. It is also a recent trend to apply deep learning in guiding network optimization with fewer constraints on input. For example, some methods seek to use the temporal information [4] or the optical flow [14] from video frames to guide the occlusion-free recovery process. However, these methods are confined to small movements of the camera, partially due to the difficulty in 3D scene representation.

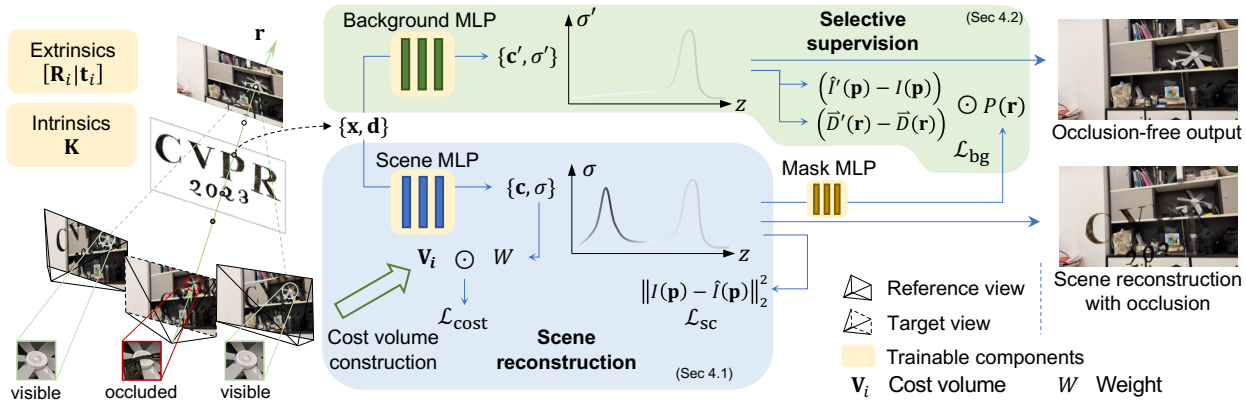


Figure 1. Pipeline of the proposed method. By taking the positional encoding [18] of position and viewing direction, the "scene MLP" (colored in blue) models the scene with occlusion faithfully, and the "background MLP" (colored in green) models the background only. To this end, a cost volume is constructed using images captured from neighboring reference views, guiding the joint optimization of scene MLP and camera parameters, as we discuss in Sec. 4.1. To supervise the training of background MLP and remove occlusions, we propose to aggregate the information from scene MLP about whether the output of background NeRF should be similar to the observed color, by learning a supervision mask from the weights along the ray using a "mask MLP" (colored in amber), as explained in Sec. 4.2. The reference and target views constitute the family of neighboring images whose features are warped to construct the cost volume.

3. Preliminary

Our goal is to generate an occlusion-free scene representation given a collection of N images $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$ captured from N different viewpoints, where the i-th viewpoint is defined by the extrinsic matrix $[\mathbf{R}_i|\mathbf{t}_i]$ and the camera is defined by a shared intrinsic matrix $\mathbf{K}$. Together they form the camera matrices $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_N\}$, where $\mathbf{P}_i = \mathbf{K}[\mathbf{R}_i|\mathbf{t}_i]$, as illustrated in the left part of Fig. 1.

As we follow the approach of NeRF [18], we briefly introduce it in this section for self-contained purposes. NeRF uses a neural network F_Θ to construct a radiance field, which maps the 3D coordinate of a point $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$ into the point's radiance color $\mathbf{c} = (r, g, b)$ and density σ. Then, the volume rendering scheme is applied to render an image. For each training image and every pixel $\mathbf{p} = (u, v)$, a ray is emitted from the camera position $\mathbf{o}$ at direction

$$\mathbf{d}_i(u, v) = \mathbf{R}_i \left[\frac{u - W/2}{f}, -\frac{v - H/2}{f}, -1 \right]^\top, \quad (1)$$

where W, H, and f are the width, height, and focal length of the input image, respectively. Marching through the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ and aggregating the colors and densities provide us the final color

$$\hat{I}_i(\mathbf{p}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d}_i)dt, \quad (2)$$

where $T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds)$ stands for the accumulated transmittance along the ray, or intuitively, the probability of the emitted radiance of $\mathbf{r}(t)$ to reach the camera without being blocked by points in between. t_n and t_f are the far end and the near end of rendering respectively.

To make the integral tractable, points are sampled along the ray, and Eq. (2) can be rewritten into

$$\hat{I}_i(\mathbf{p}) = \sum_{k}^{N_s} T_k(1 - \exp(-\sigma_k\delta_k))\mathbf{c}(\mathbf{r}_k, \mathbf{d}), \quad (3)$$

$$T_k = \exp(-\sum_{j=1}^{k-1} \sigma_j\delta_j). \quad (4)$$

For the convenience of notation, $\sigma(\mathbf{r}_k)$ is abbreviated to σ_k, and δ_k is defined by $t_{k+1} - t_k$; N_s is the number of points sampled on the ray, $\mathbf{r}_k = \mathbf{o} + t_k\mathbf{d}$ denotes the position of k-th sampling point, and t_k refers to the depth of the k-th sampling point seen from the target view.

NeRF optimizes the scene representation neural network by minimizing the photometric loss $\mathcal{L} = \sum_{i=1}^{N} \|I_i - \hat{I}_i\|_2$, which is simply the difference between input images $\mathcal{I}$ and corresponding predicted images $\hat{\mathcal{I}}$. The optimization process can be formulated as:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\hat{\mathcal{I}}|\mathcal{I}, \mathcal{P}). \quad (5)$$

Despite NeRF's strong capability in representing a scene from multiple viewpoints, the occlusions pose challenges to it from two aspects: 1) How to selectively aggregate background information from multiple viewpoints and 2) how to accurately estimate the camera pose of desired background in a continuous space.

4. Proposed approach

We focus on how to address the two challenges above when using NeRF to see through occlusion from multiple viewpoints and our pipeline is illustrated in Fig. 1. In Sec. 4.1, we first describe our joint optimization framework to accurately estimate the camera pose and reconstruct the scene with occlusions using a scene MLP. Then, in Sec. 4.2, we introduce how to selectively aggregate background information from multiple viewpoints based on the rectified background pose using a mask MLP. Under the guidance of mask MLP, we finally train a background MLP to reconstruct occlusion-free scenes. Note that both the scene MLP and the background MLP map the positional encoding of position and viewing direction into color and density, as described in Sec. 3.

4.1. Scene reconstruction with cost volume

NeRF and most of its variants use COLMAP [23] as a silver bullet to estimate camera poses for the scene representation, while COLMAP [23] is not an ideal setting for building an occlusion-free representation. The extracted local features cannot be distinguished between undesired occlusion and the desired background, so the matching process can lead to an erroneous correspondence, or even directly fail to estimate the camera parameters. Some recent works try to address the dependency on COLMAP [23] by jointly optimizing camera parameters and the neural radiance fields with the photometric loss [28], achieving comparable performance to COLMAP [23].

However, only using photometric loss does not solve the fundamental problem, because during training, it only uses one view at a time, namely the target view itself. This makes the system ignorant of the 3D scene as a whole, and thus still cannot tell occlusions from the desired background. Besides, in the early stage of training when NeRF can hardly output meaningful information, the photometric loss cannot constrain camera extrinsics to converge to the correct rotation and translation. This can lead to a sub-optimal solution.

To make our method aware of occlusions and background, we design a cost volume loss inspired by multi-view stereo solutions [6, 30]. It encourages the system to model the location corresponding to each pixel at the correct depth, by utilizing the feature consistency of the same point seen from neighboring reference views. Next, we will introduce how our cost volume loss is constructed.

Feature extraction. A 2D map F_i is extracted for every input image I_i by using a pre-trained VGG-19 network η as a feature encoder as $F_i = \eta(I_i)$. For the convenience of notation, we define $F_i(\mathbf{p})$ as the feature vector at 2D location $\mathbf{p}$. When $\mathbf{p}$ falls off the grid, bilinear interpolation is performed to acquire the feature vector. The extracted 2D features represent local appearance, so intuitively the feature of the same point of interest across neighboring views are assumed to be similar. On the other hand, if two feature vectors vary greatly, it is unlikely that they describe the same area.

Feature volume construction. To construct a feature volume at the j-th viewpoint (target view), the image features from neighboring reference views should be warped with respect to the depth. The warping from the i-th view to the j-th view can be formulated as a 3×3 matrix

$$\mathbf{H}_{ij}(z) = \mathbf{K}\mathbf{R}_i(\mathbf{Id} - \frac{(\mathbf{R}_j^\top \mathbf{t}_j - \mathbf{R}_i^\top \mathbf{t}_i)\mathbf{n}_j^\top}{z} \cdot \mathbf{R}_j)\mathbf{R}_j^\top \mathbf{K}^{-1}, \tag{6}$$

where $\mathbf{Id}$ is the identity matrix, and $\mathbf{n}_j$ is the principle axis of the j-th view. This matrix describes a mapping that the point at the 2D coordinate of $\mathbf{p}$ and the depth of z in warped feature map corresponds to the point at $\mathbf{H}_{ij}(z)\mathbf{p}$ in the original feature map. Applying the warping matrix to the feature map F_i derives a family of warped image feature maps:

$$F_{i,j,z}(\mathbf{p}) = F_i(\mathbf{H}_{ij}(z)\mathbf{p}), \tag{7}$$

which is warped from F_i, seen from the j-th view, and at the depth of z.

We then construct a cost volume at the points which we sample using the warped feature maps. Based on the intuition that similar features mean more likelihood of the same area, a large variance of warped features indicates that features belonging to different areas are warped to the current point. Thus, the variance metric can be used as a clue for scene reconstruction as

$$\mathbf{V}_j(u, v, z) = \mathrm{Var}_i(F_{i,j,z}([u, v, 1]^\top)), \tag{8}$$

where $\mathrm{Var}_i(\cdot)$ denotes the variance across M selected neighboring views.

Joint optimization. To apply the cost volume as supervision for NeRF training, we consider the weight of the k-th sampled point

$$W_k = T_k(1 - \exp(-\sigma_k)), \tag{9}$$

which refers to the contribution of the point at $\mathbf{r}_t$ to the result $\hat{I}_j(u, v)$. We build the scene MLP to reconstruct the scene with occlusion, to provide an intermediate scene representation for further occlusion removal. For a high-quality reconstruction, the scene MLP should be discouraged to assign the points that are inconsistent across neighboring views with a large weight. Based on the observations above, we construct the cost volume loss as

$$\mathcal{L}_{\text{cost}} = \sum_{\mathbf{p}=(u,v)} \sum_{i=1}^{N_s} \mathbf{V}_j(u, v, t_i) W(u, v, t_i). \tag{10}$$

Meanwhile, we also apply the self-consistency loss aiming at scene reconstruction as

$$\mathcal{L}_{\text{sc}} = \sum_{\mathbf{p}} \|\hat{I}_i(\mathbf{p}) - I_i(\mathbf{p})\|_2^2. \tag{11}$$

So far, we have explained the design of scene reconstruction scheme where occlusions exist, including the key component of cost volume, as depicted in the blue box of Fig. 1. Together, they provide a more robust estimation of the scene and the camera poses when occlusions are present.

4.2. Selective supervision scheme

Existing image-based occlusion removal methods [14, 27] can function properly when a large amount of labeled data is available, by considering the occlusion removal as a mapping from samples with occlusions to their occlusion-free counterparts. Our method is grounded on a totally different underlying logic. By simulating human eyes to observe a specified scenes, our method can aggregate a continuous occlusion-free scene representation from different viewpoints. Then, we can render any occlusion-free viewpoints from this learned scene representation. Such mechanism enables a novel way to remove occlusion in an unsupervised manner. The key problem then becomes how to focus on the desired background only during scene aggregation.

We achieve this goal based on an intuitive observation: *a majority of undesired occlusions always appear in the foreground.* However, directly applying depth as supervision is not feasible, since the depth does not necessarily correlate with the presence of occlusion. To this end, we propose to use bidirectional depth inconsistency as a clue. If the expected ray termination depth seen from the camera and that from the other end of the ray show great difference, the ray should pass through some foreground occlusion. Note that *"termination"* in rendering refers to the point that no point behind it contributes to the rendered ray.This assumption follows the physical law, and does not rely on any external prior or excessive training data. Note that this prior is opposite to the shell-shaped geometry assumption by NeRFReN [7], which assumes that the density along the ray peaks out at the surface.

In implementing this, we design a mask MLP P_Ψ to predict the occluding likelihood $\tilde{P}(\mathbf{r})$ based on the weights along the ray, formulated as:

$$\tilde{P}(\mathbf{r}) = P_\Psi\left(\{W_k\}_{k=1}^{N_s}\right). \tag{12}$$

Based on the intuition discussed above, the mask MLP is trained based on the bidirectional depth inconsistency. The expected termination depth $\overrightarrow{D}$ and reversed termination depth $\overleftarrow{D}$ is defined by

(a) $\overrightarrow{D}$ (b) Δ_D (c) P

Figure 2. An illustration of the depth and mask used for selective supervision (without losing generality, we show the example for fence, and this applies to other types of occlusions as well): (a) estimated depth ($\overrightarrow{D}$) of the scene with occlusion, (b) the bidirectional depth difference (Δ_D), and (c) the supervision mask (P). Note that the white regions in (c) denote the desired background details that need supervision and undesired foreground occlusions are marked in black. Please visit our project page for animated results.

$$\overrightarrow{D}(\mathbf{r}) = \sum_{k=1}^{N_s} W_k t_k = \sum_{k=1}^{N_s} T_k(1-\exp(-\sigma_k\delta_k))t_k, \tag{13}$$

and

$$\overleftarrow{D}(\mathbf{r}) = \sum_{k=1}^{N_s} \overleftarrow{W}_k t_k = \sum_{k=1}^{N_s} \overleftarrow{T}_k(1-\exp(-\sigma_k\delta_k))t_k, \tag{14}$$

respectively, where $\overleftarrow{T}_k = \exp(-\sum_{j>k}\sigma_j\delta_j)$ is the reversed accumulated transmittance. An example of $\overrightarrow{D}$ can be found in Fig. 2(a). We train the mask MLP based on the prior described before, that $P(\mathbf{r})$ correlates with $\Delta_D(\mathbf{r}) = \overleftarrow{D} - \overrightarrow{D}$, by

$$\mathcal{L}_{\text{mask}} = -\cos(\{P(\mathbf{r})-\alpha\}_{\mathbf{r}\in\mathcal{R}}, \{\Delta_D(\mathbf{r}) - \overline{\Delta_D}\}_{\mathbf{r}\in\mathcal{R}}), \tag{15}$$

where $\mathcal{R}$ is the patch of rays for training, $\overline{\Delta_D}$ is the mean value of Δ_D across the entire input view, and $\cos(\mathbf{a},\mathbf{b})$ is the cosine similarity between $\mathbf{a}$ and $\mathbf{b}$. We show an example of Δ_D in Fig. 2(b). α is a pre-defined threshold for the binary classification of $\tilde{P}$, deriving a binary supervision mask P by

$$P(\mathbf{r}) = \begin{cases} 0, & \tilde{P}(\mathbf{r}) < \alpha \\ 1, & \tilde{P}(\mathbf{r}) \geq \alpha \end{cases}. \tag{16}$$

As shown in Fig. 2(c), the obtained binary mask P can clearly distinguish the desired background to be kept and the undesired occlusion to be removed.

We further train a background MLP which reconstructs the scene belonging to the background only. Specifically, we only supervise the rendered color of background MLP to be similar to the input image in the background region

where $P(\mathbf{r}) = 1$.

$$\mathcal{L}_{\text{bg}} = \sum_{\mathbf{p}} \left\| ((\hat{I}_i'(\mathbf{p}) - I_i(\mathbf{p})) \odot P(\mathbf{r}) \right\|_2^2 + \sum_{\mathbf{p}} \left\| (\overrightarrow{D}'(\mathbf{r}) - \overrightarrow{D}(\mathbf{r})) \odot P(\mathbf{r}) \right\|_1, \tag{17}$$

where $\hat{I}_i'(\mathbf{p})$ and $\overrightarrow{D}'(\mathbf{r})$ denote the rendered color and expected ray termination depth of the background MLP.

With background MLP holding the occlusion-free scene representation and the scene MLP containing all the information including occlusions, we have finally gathered all the pieces of the puzzle.

4.3. Implementation details

We implement our method using PyTorch. The scene MLP and the background MLP, each having 256 channels, consist of 6 layers and 8 layers respectively, which predict the colors and densities corresponding to the whole scene and the background separately. The mask MLP is a 2-layer MLP with 64 channels. When constructing cost volume, $M = 4$ neighboring views are involved to calculate variance in Eq. (8). In the training phase, the scene MLP is jointly optimized with camera parameters of each scene, after being trained coarsely to fit the scene from COLMAP [23] camera pose or identity matrix initialization. All of the trainable parameters are optimized per scene. We train scene MLP for the first 20% of iterations, and then train the system end-to-end.

On each iteration, we evenly sample 128 points along each ray. A batch contains random 1024 pixels in one of the images. We use the Adam optimizer with defaults values $\beta_1 = 0.999$, $\beta_2 = 0.9$, $\epsilon = 10^{-8}$, and a learning rate 10^{-3} that decays following the cosine scheduler [15] during the optimization. If initialized with identity matrix, the model is trained for about 150K iterations. If a coarse COLMAP [23] estimation is available, our model converges after about 50K iterations.

5. Experiments

Dataset. To validate the effectiveness of our method, we collect an evaluation dataset containing 10 different scenes, covering various types of occlusions. The dataset not only contains multiple sparse viewpoints of occlusion scenes collected by ourselves, but also samples selected from existing datasets [29,31], to increase the diversity from different capturing conditions. Specifically, FENCE1, SCRIBBLE2 and RAINDROP are adopted from [29], and WIRE2 is adopted from [31]. The rest 6 of the scenes are captured by ourselves using a Sony α7 III camera or an iPhone 12. For easy reference, we name each set of data using the occlusion type, as shown in Fig. 3.

(a) FENCE1 (b) FENCE2 (c) FENCE3 (d) SCRIBBLE1 (e) SCRIBBLE2
(f) SCRIBBLE3 (g) RAINDROP (h) STATUE (i) WIRE1 (j) WIRE2

Figure 3. From (a) to (j), we provide one sample image for each scene in our evaluation dataset. The names of the scenes are listed below the image indicating occlusion type.

For our newly captured samples, we take 20 to 60 images with different poses for each scene, where 85% of the images are used for training our NeRF-based method, and the rest 15% for testing. This is a commonly adopted setting by many NeRF-based methods [7, 8, 18]. To facilitate the quantitative evaluation, one of newly captured scene is with ground truth, namely SCRIBBLE1. For this scene, we place a piece of glass with scribbles between the desired background scene and the camera. Then, a multi-view sparse image set is acquired by moving cameras in front of the glass. To obtain the ground truth for validation, we remove the glass while taking the first and last view of the scene, so that these two views can come with a ground truth.

Baselines. To our knowledge, there is no similar work that construct an occlusion-free scene based on NeRF representation. We design three baselines for comparison by considering both the contributions from occlusion removal and NeRF to verify our performance advantage. 1) A state-of-the-art occlusion removal method for image sequences or videos (referred to as PWC-Net [14] based on their description) + NeRF [18]: In this setting, PWC-Net [14] serves as a pre-processing tool to eliminate occlusion before training the NeRF [18]. 2) NeRF-W [17]: This is a method designed for separating a complete scene into two components. 3) Ha-NeRF [3]: This is another method that decomposes the scenes into distinct parts. We compare those settings with our method on our evaluation dataset, as listed in Fig. 3. For fairness of comparison, we evenly sample 128 points along the ray without applying importance sampling strategy. We also keep the trainable parameters in each baseline approximately the same with the proposed method.

Evaluation methodology. We conduct both qualitative and quantitative evaluations on our dataset. For quantitative evaluation, we compare reconstruction quality of the estimated occlusion-free images for viewpoints with corresponding ground truth by PSNR, SSIM, and LPIPS [32].

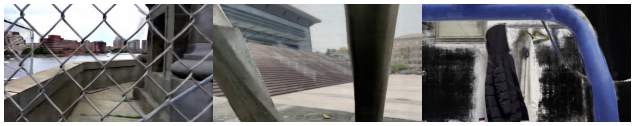
5.1. Qualitative evaluation

The qualitative results of the rendered novel occlusion-free views are presented in Figs. 4 and 5. In Fig. 4, we use

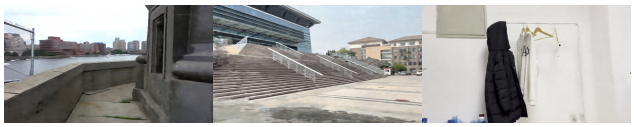
(a) Vanilla NeRF [18]

(b) PWC-Net+NeRF [14, 18]

(c) Ha-NeRF [3]



(d) NeRF-W [17]



(e) Our method

Figure 4. Visual quality comparisons for occlusion removal. From left to right, we show animated novel view synthesis on FENCE1, STATUE, and WIRE1. From (a) to (e), we show results obtained by (a) vanilla NeRF [18], (b) PWC-Net [14] + NeRF [18], (c) Ha-NeRF [3], (d) NeRF-W [17], and (e) our method. Please visit our project page for animated results. Some results show stronger variation over the animation due to failure in extracting consistent background.

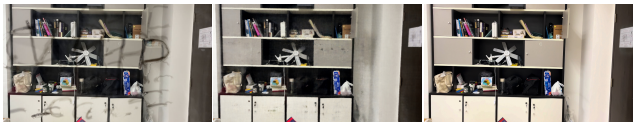
Table 1. Quantitative comparison results. ↑ (↓) indicates larger (smaller) values are better, and **bold** font indicates the best results.

Method	PSNR↑	SSIM↑	LPIPS↓
PWC-Net+NeRF [14, 18]	18.25	0.79	0.45
NeRF-W [17]	11.06	0.45	0.61
Ha-NeRF [3]	15.41	0.70	0.41
Proposed	**19.39**	**0.83**	**0.37**

the scenes of FENCE1, STATUE and WIRE1 in our dataset to show that our method can reliably reconstruct novel background views, and it works no matter whether the occlusions cover a large area or have an irregular shape. In Fig. 5, we take one frame in SCRIBBLE1 with ground truth to show the fidelity of our results.

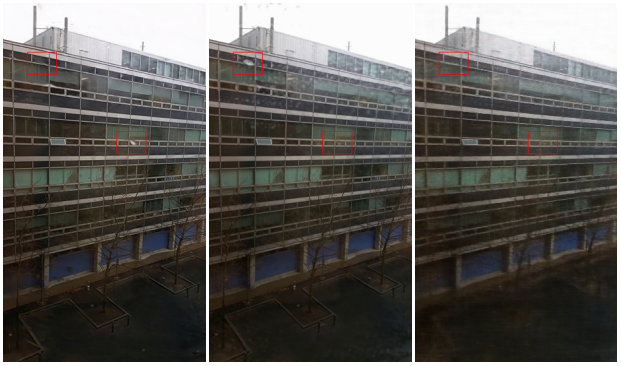
PWC-Net + NeRF [14, 18] only achieves comparable results in FENCE1 where the occlusions are regularly shaped, but shows poor artifacts on unseen types of occlusions, such as the irregularly-shaped STATUE and WIRE1, as shown in Fig. 4(b). This shows their reliance on training data. Besides, the occlusion removal method can lead to an over-

(a) (b) (c)



(d) (e) (f)

Figure 5. Visual quality comparisons for occlusion removal on SCRIBBLE1 scene. (a) is one image in the validation set with occlusion and (f) is ground truth of this view) with results from: (b) results obtained by Ha-NeRF [3], (c) results obtained by NeRF-W [17], (d) results obtained by occlusion removal [14] + NeRF [18], and (e) our method.



(a) (b) (c)

Figure 6. Visual quality comparisons for occlusion removal on RAINDROP scene, with results from (a) an image-based raindrop removal method [22], (b) an image-based occlusion removal method [14], and (c) our method. Our method outperforms the baselines in the highlighted red boxes.

smoothing issue which undermines the reconstruction accuracy. Meanwhile, the results from Ha-NeRF [3] and NeRF-W [17] in Figs. 4(c) and 4(d) either shows degenerated details or still preserves occlusions, which verifies our claim that they cannot handle the consistent occlusions observed in a continuous image sequence. The results from our method in Fig. 4(e) clearly remove these occlusions with different shapes and distributions.

Our method relies on an accurate separation between foreground and background, but the assumption may be violated if obstructions are semi-transparent. We conduct experiments on the RAINDROP scene in our validation dataset to evaluate the robustness of our method under such scenario. With the results displayed in Fig. 6, our method can also effectively remove the raindrop in this example. It achieves competitive or even better results in some regions when compared with a method specifically designed for raindrop removal [22] as well as PWC-Net [14] designed to remove occlusions including raindrop.

Table 2. Ablation study results aggregated by the types of occlusions. We compare the rendered results of the scene MLP with NeRF [18] using COLMAP [23] as camera parameter estimation and NeRF−− [28] initialized with COLMAP camera parameters.

Metrics		PSNR ↑ / SSIM ↑ / LPIPS↓		
Method		COLMAP [23]	NeRF−− [28]	Scene MLP
Dataset	FENCE{1,2,3}	24.01 / **0.73** / 0.38	18.67 / 0.54 / 0.51	**25.46** / 0.72 / **0.35**
	SCRIBBLE{1,2,3}	22.70 / 0.78 / **0.31**	17.98 / 0.52 / 0.46	**23.85** / **0.84** / 0.34
	WIRES{1,2}	26.21 / 0.93 / **0.20**	22.28 / 0.88 / 0.25	**34.44** / **0.95** / 0.23
	RAINDROP	27.69 / **0.87** / 0.26	27.78 / 0.83 / 0.25	**28.69** / 0.86 / **0.25**
	STATUE	**27.53** / **0.84** / **0.33**	18.49 / 0.56 / 0.49	27.08 / 0.80 / 0.37

5.2. Quantitative evaluation

The quantitative results of the occlusion-free novel view in Tab. 1 validate the observation in Figs. 4 and 5. Higher PSNR values show that our method can render occlusion-free novel views and recover the color information with higher accuracy. Higher SSIM values indicate that our method can preserve the structural information with high-frequency details. Lower LPIPS values show that recovered images by our method better aligns with human perception.

5.3. Ablation Study

Our method consists of two major components: joint optimization for pose refinement and a selective supervision strategy to eliminate occlusions. Since the selective supervision scheme largely bases on the intermediate information of scene reconstruction MLP, a high-quality reconstruction of the scene with occlusion is crucial to our performance.

In the ablation study, we first replace pose refinement with COLMAP [23]. From the results shown in Fig. 7(b), the occlusions largely remain, because the fixed erroneous camera parameters makes distinguishing scene components from scene MLP infeasible.

We also explore the results without the selective supervision scheme, while maintaining pose refinement, namely only output of the scene MLP. In this setting, our method cannot remove occlusions, but it can reconstruct the scene containing occlusions with a high fidelity, as shown in Fig. 7(c). It can also reconstruct the scene with higher fidelity than NeRF based on COLMAP [23], as shown quantitatively in Tab. 2. Our complete model shows the best results for majority of scenes in our dataset.

(a) (b) (c)

Figure 7. Visual comparisons for the ablation study on the FENCE1 scene with results from: (a) our complete method, (b) the method without joint optimization, and (c) the method without selective supervisions. Please visit our project page for animated results.

6. Conclusions

In this paper, grounded on Neural Radiance Fields, we construct an occlusion-free scene representation, which is capable of rendering occlusion-free images with desired viewpoints. By directly using the viewpoints as a part of the input, our method is able to aggregate occluded details from multiple viewpoints without relying on any external guidance. We propose a pose refinement scheme to ensure robust training when the camera poses cannot be accurately estimated by COLMAP [23]. To focus on modeling of desired background scenes within a specific viewpoint, we further introduce a depth constraint to probe the occluded areas by measuring the depth of occlusion and background. Experimental results show that our method is capable of achieving promising results under a variety of scenarios.

Limitations. Despite the promising performance of our proposed method, several limitations are still to be addressed in our future study. First, our method clearly assumes the undesired layers are in the foreground. It cannot disambiguate the cases where details from different layers are intertwined, such as scenes with reflection, which is especially challenging due to its semi-transparency and widespread presence. We show a failure case of the proposed method in a scene with reflection as occlusion in Fig. 8. Besides, since several variants of NeRF have been proposed for faster convergence and more accurate reconstruction [5,20], advanced NeRF variants may boost the performance of the proposed method. In addition, we also hope to reduce the number of images required for joint optimization of camera parameters and NeRF in our future work.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant No. 62136001, 62088102.

(a) (b) (c)

Figure 8. A failure scene of the proposed method, when reflection is present as occlusion. (a) input sample, (b) results of the scene MLP, (c) failed reflection removal with the background MLP.

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 2001. 2

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proc. of ACM SIGGRAPH*, 2000. 2

[3] 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. 1, 2, 6, 7

[4] Chen Du, Byeongkeun Kang, Zheng Xu, Ji Dai, and Truong Nguyen. Accurate and efficient video de-fencing using convolutional neural networks and temporal information. In *Proc. of International Conference on Multimedia and Expo*, 2018. 1, 2

[5] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 8

[6] 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. 4

[7] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. NeRFReN: Neural radiance fields with reflections. In *CVPR*, 2022. 1, 2, 5, 6

[8] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. HDR-NeRF: High dynamic range neural radiance fields. In *CVPR*, 2022. 2, 6

[9] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In *Proc. of Conference on Robot Learning*, 2021. 2

[10] Sankaraganesh Jonna, Sukla Satapathy, and Rajiv R Sahay. Stereo image de-fencing using smartphones. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2017. 1, 2

[11] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *CVPR*, 2020. 2

[12] 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. 1

[13] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *CVPR*, 2021. 2

[14] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *CVPR*, 2020. 1, 2, 5, 6, 7

[15] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[16] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-NeRF: Neural radiance fields from blurry images. In *CVPR*, 2022. 2

[17] 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. 1, 2, 6, 7

[18] 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, 2, 3, 6, 7, 8

[19] Yadong Mu, Wei Liu, and Shuicheng Yan. Video de-fencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014. 1, 2

[20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 2022. 8

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. EdgeConnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[22] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018. 1, 7

[23] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 4, 6, 8

[24] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. SPG-Net: Segmentation prediction and guidance network for image inpainting. In *Proc. of British Machine Vision Conference*, 2018. 2

[25] V. Vaish, M. Levoy, R. Szeliski, C.L. Zitnick, and Sing Bing Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *CVPR*, 2006. 2

[26] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. NeSF: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 2

[27] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C Kot. Face image reflection removal. *International Journal of Computer Vision*, 2021. 1, 5

[28] 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. 2, 4, 8

[29] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *TOG*, 2015. 1, 2, 6

[30] 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. 4

[31] Renjiao Yi, Jue Wang, and Ping Tan. Automatic fence segmentation in videos of dynamic scenes. In *CVPR*, 2016. 6

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[33] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proc. of International Conference on Computer Vision*, 2021. 2

[34] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 2021. 2