REF^2-NeRF: Reflection and Refraction aware Neural Radiance Field

Wooseok KIM
The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, Japan
kim@cvl.iis.u-tokyo.ac.jp

Taiki FUKIAGE
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi, Kanagawa-ken
taiki.fukiage@ntt.com

Takeshi OISHI
The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, Japan
oishi@cvl.iis.u-tokyo.ac.jp

Abstract

Recently, significant progress has been made in the study of methods for 3D reconstruction from multiple images using implicit neural representations, exemplified by the neural radiance field (NeRF) method. Such methods, which are based on volume rendering, can model various light phenomena, and various extended methods have been proposed to accommodate different scenes and situations. However, when handling scenes with multiple glass objects, e.g., objects in a glass showcase, modeling the target scene accurately has been challenging due to the presence of multiple reflection and refraction effects. Thus, this paper proposes a NeRF-based modeling method for scenes containing a glass case. In the proposed method, refraction and reflection are modeled using elements that are dependent and independent of the viewer's perspective. This approach allows us to estimate the surfaces where refraction occurs, i.e., glass surfaces, and enables the separation and modeling of both direct and reflected light components. Compared to existing methods, the proposed method enables more accurate modeling of both glass refraction and the overall scene.

1. Introduction

3D reconstruction from 2D images is a well-established technique; however, there is still room for improvement in terms of modeling scenes that involve transparent objects, e.g., glass. Glass is commonly used in the real world; thus, there is a great demand for modeling scenes containing transparent objects. Unfortunately, common image sensors cannot observe transparent objects directly, and such objects produce various effects, e.g., reflections and refractions, which prevent conventional photogrammetry methods from modeling such scenes correctly.

Recent advances in implicit neural representations [22] of 3D scenes have made it possible to model and synthesize novel views of scenes including reflection and refraction effects. Prior to the introduction of the neural radiance field (NeRF) technique, researchers studied these photometric effects as physical occurrences and developed techniques to replicate scenes based on the physical principles. However, modeling real-world scenes based on physical models from images is an ill-posed problem that requires various constraints. In contrast, NeRF-based methods allow neural networks to learn these complex phenomena to synthesize new views effectively and enable geometry modeling. The ability of these methods to model transparent objects, metallic objects, objects in transparent medias, and objects in liquids without the need for complex models or constraints was a major development in the field.

However, even with these techniques, modeling scenes with multiple transparent surfaces remains a challenge. Scenes frequently feature glass objects, as demonstrated by the showcase depicted in Fig. 1. When such objects appear in a scene, they generate multiple reflections and refractions along the viewer's line of sight, which increases the task complexity of generating neural fields that capture the corresponding objects accurately.

Figure 1. Example images of a scene including a glass case and objects. Images of those scenes contain effects of light ray reflection and refraction, which vary depending on viewpoint.

Thus, in this paper, we propose a neural modeling method that considers the characteristics of scenes containing multiple glass surfaces, particularly objects enclosed in a glass case. The proposed method introduces two networks to handle refraction and reflection independently, and it decomposes the view-independent and view-dependent components of these effects. In the case of refraction, the view-independent component is the refraction point, and the view-dependent component is modification of the ray's direction. Accordingly, the proposed method learns to synthesize new viewpoint images and estimates the position and magnitude of the refraction within the given scene. An additional network decomposes the direct and reflection components in geometric and photometric aspects.

The contributions of this paper can be summarised as follows:

- We propose a neural network-based approach for modeling scenes with multiple glass surfaces, focusing particularly on objects inside a glass case. This approach involves using two separate networks to handle the effects of refraction and reflection.
- We introduce a framework that handles refraction and reflection efficiently by learning the view-dependent and view-independent components separately.
- The proposed method decomposes direct and reflection components in geometric and photometric terms and estimates refraction position and magnitude in the scene.

2. Related work

In the following, we briefly review previous studies related to the multiview reconstruction and learning of scenes that include transparent objects, e.g., glass.

2.1. Multi-view method for transparent object

Common image-based 3D reconstruction methods based on the structure from motion and multiview stereo (MVS) [9, 10, 29, 30] techniques estimate the surface geometry using triangulation from feature points or patches. MVS technology has matured recently, and its use is common in various practical applications. In addition, the development of deep learning has made them more robust and accurate. However, strong reflection affects feature and texture matching, and refraction distorts the estimated surface shape, which makes it difficult to reconstruct such scenes using conventional MVS methods.

Methods have also been proposed to model transparent objects and objects behind or within them. For example, several methods operate under a controlled environment, e.g., using background patterns [18, 39] and polarization [19, 23]. However, these methods attempt to estimate the surface shape of the target transparent object; thus, they are unsuitable for modeling regions or objects behind transparent objects. A popular scene where it is difficult to control the environment and target behind or within transparent objects is modeling underwater environments. For underwater scenes, the refraction of light rays occurs at the surface between the lens and the water [4, 5, 13, 27]. In other words, multiple reflections and refractions are not considered or require prior ray calibration to model more complex scenes.

Some methods employ supervised learning approaches [16]; however, the results are dependent on the available training data, and such methods do not handle complex light effects as well as other model-based methods.

2.2. Neural Radiance Fields (NeRF)

The NeRF technique [22] can represent various optical phenomena because the trained neural network data, density, and color fields are based on volume rendering [21]. NeRF and its variants optimize the fields represented by implicit neural functions to reduce the similarity error between the input and rendered images. A well-trained NeRF model allows us to reconstruct 3D scenes or synthesize novel views from the neural models.

However, the original NeRF has some drawbacks; thus, various methods have been proposed to improve NeRF in terms of acceleration [7, 24, 42], synthesized image quality enhancement [2, 3, 33], and robustness against various conditions [17, 20, 26, 32, 37, 43]. Originally, NeRF was designed to realize novel view synthesis; however, more accurate shape reconstruction is achievable using, for example, signed distance function (SDF)-based approaches [34, 35].

NeRF has shown promising results; however, similar to common 3D reconstruction methods, it assumes a straight light ray and generates poor results when refraction is present in the input images.

In addition, reflections in a scene appear as if another scene exists behind the observable reflective and transparent objects, and those reflections may not be visible depending

on the viewpoint from which the scene is observed. Note that this phenomenon contradicts the view-consistency assumption of NeRF.

2.3. NeRF for reflective scenes

Several NeRF variants have been proposed to handle scenes that include reflective surfaces. For example, NeRFReN [11] handles reflections by assuming reflective plane surfaces and separating the scene into transmitted and reflected components utilizing two rendering paths. In addition, NeuS-HSR [28] also estimates an auxiliary plane that separates the reflection components, which facilitate reconstruction of an object inside a glass case. The Neural Transmitted Radiance Fields method [46] detects recurring edges in the input images to optimize transmission and reflection components independently.

The MS-NeRF [41] technique separates the input scene into multiple spaces and estimates density fields and feature fields for these spaces. Here, the feature fields reduce the number of estimated parameters while estimating the color map and weights of the spaces.

The neural point catacaustics method [15] introduces a pointwise neural warp field that represents the reflection from curved surfaces, which makes it possible to render the reflected points that are separated from the primary point cloud.

The aforementioned methods can model scenes containing reflective objects effectively; however, they do not consider refraction by transparent objects, and they do not handle scenes containing multiple transparent objects.

2.4. NeRF for refractive scenes

The LB-NeRF method [8] addresses scenes in which objects are present inside a refractive medium. The LB-NeRF method handles refraction by simplifying it as an offset from a straight light ray. By adding the offsets to each sampled point's position prior to training NeRF's MLP, the LB-NeRF method models canonical space without refraction effects. Other methods based on the physical properties of reflective and refractive medium [25, 31, 38, 44] require additional information, e.g., a known image pattern, the refractive index, or a mask image of the refraction, as a clue to detect and estimate the refraction present in an image.

A number of NeRF-based methods have been proposed for reflective or transparent objects [6, 12, 14, 32]; however, it is difficult to apply these methods to the target scene considered in this study for the reasons described above.

3. Preliminary and overview

In this section, we summarize the refraction and reflection effects and provide an overview of the proposed framework, which is based on these effects.

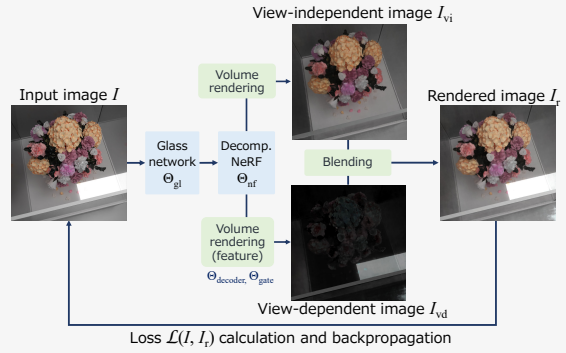


Figure 2. Overview of the proposed framework. Glass network MLP models refraction occurred by transparent object and adjusted each sampled position. Then, we decompose the scene into view-dependent and view-independent components to separate reflection from input images and model both.

3.1. Refraction effect

Refraction is a phenomenon whereby the direction of light changes due to differences in the refractive index between mediums. Note that refraction also changes according to the incident angle, and it follows Snell's law. Thus, the glass surface where refraction occurs is independent of the viewpoint; however, the amount of refraction depends on the viewpoint. In addition, when light passes through a glass plate, it passes through the parallel boundary in the air-to-glass and glass-to-air order; thus, the ray is parallel to the original ray and is shifted according to the incident angle and the thickness of the glass.

3.2. Reflection effect

Reflection occurs on the surface of the glass and, similar to refraction, is a phenomenon whereby the path of the light changes. Here, the reflected light intensity is largely distributed in the direction opposite to the incident angle relative to the surface normal, which means that the reflected light intensity is strongly dependent on the viewpoint. Reflections, like mirrors, create a mirror object, i.e., it appears as if the same object is behind the glass. However, it differs from a mirror in that the mirror object is semitransparent because some of the light is reflected on the glass surface, and the remaining light is transmitted into the glass. In other words, in addition to the view-independent objects, view-dependent semitransparent objects can be assumed to exist in the scene.

3.3. Overview of proposed framework

Based on the above considerations, the proposed framework is designed to estimate the refraction effect and separate the view-independent and view-dependent components present in the given scene. Figure 2 shows an overview of the pro-

posed method. In the proposed method, we assume that the input is multiple images $\{I_k\}(k = 1, 2, ..., n)$ taken from different viewpoints, similar to existing NeRF variants. Here, n is the number of input images, and we omit the index k in this section for simplicity.

The proposed framework primarily comprises two independent MLPs, i.e., $\Theta_{\rm gl}$ and $\Theta_{\rm nf}$, to handle the refractive and reflective components independently, respectively. The former is referred to as the glass network. The glass network represents the refraction points and the amount of refraction, which give the parallel shift of the ray for sampling points in the latter network. The latter is the main NeRF network, which decomposes the direct and reflection components. Here, $\Theta_{\rm nf}$ represents the fields to render an image $I_{\rm vi}$ of a direct component and an image $I_{\rm vd}$ of a reflection component. The density and feature fields of the view-dependent component generate the image through the decoder and gate MLPs: $\Theta_{\rm dc}$, $\Theta_{\rm gt}$ [41]. The blended image of $I_{\rm vi}$ and $I_{\rm vd}$ is the output image $I_{\rm r} = I_{\rm vi} \oplus \alpha I_{\rm vd}$, where α is the blending parameter derived from the network.

The training process optimizes both networks while minimizing the loss $\mathcal{L}$ between the input and rendered images as follows:

$$\hat{\Theta}_{\rm gl}, \hat{\Theta}_{\rm nf}, \hat{\Theta}_{\rm dc}, \hat{\Theta}_{\rm gt} = \underset{\Theta_{\rm gl}, \Theta_{\rm nf}, \Theta_{\rm dc}, \Theta_{\rm gt}}{\arg\min} \mathcal{L}(I, I_{\rm r}). \quad (1)$$

4. Proposed framework

This section describes the proposed framework and its implementation in detail. Figure 3 shows the network architecture of the proposed framework and the rendering flow.

4.1. Glass network

The glass network is employed to estimate the location of the glass surface where the refraction occurs and the amount of refraction. The method used to simplify and express the refraction as an offset is similar that utilized in the LB-NeRF technique [8]. In contrast to LB-NeRF, the proposed method introduces the view-independent density field and view-dependent offset field to estimate the refraction surfaces and the offsets simultaneously.

Here, for a point $\boldsymbol{x}_i \in \mathbb{R}^3, (i = 1, 2, ..., N)$ sampled along a ray $\boldsymbol{r}$, the glass network $\Theta_{\rm gl}$ estimates the glass density $\sigma_{\rm gl}$, which indicates the degree to which that point is involved in the refraction. N is the number of sampled points. $\Theta_{\rm gl}$ also outputs the offset vector $\Delta\boldsymbol{x}_i \in \mathbb{R}^3$, which represents the magnitude and direction of the refraction arising from the view direction $\boldsymbol{d}_i \in \mathbb{R}^3$ and position $\boldsymbol{x}_i$. In other words, the glass network takes the encoded $\boldsymbol{x}$ and $\boldsymbol{d}$ as inputs and outputs $\sigma_{\rm gl}$ and $\Delta\boldsymbol{x}$. This process is expressed as follows:

$$\mathcal{F}_{\Theta_{\rm gl}} : \Gamma(\boldsymbol{x}), \Gamma(\boldsymbol{d}) \rightarrow \sigma_{\rm gl}(\boldsymbol{x}), \Delta\boldsymbol{x}(\boldsymbol{x}, \boldsymbol{d}), \quad (2)$$

where Γ represents the positional encoding.

The sampling points are adjusted after refraction using the offset vectors, as shown in Fig. 4. Similar to NeRF's volume rendering [21], in the proposed method, the refraction weight of each sampling point is calculated using the glass density $\sigma_{{\rm gl},i}$ and the distance between adjacent sampling points δ_i as follows:

$$w_i = T_i(1 - exp(-\sigma_{{\rm gl},i}\delta_{{\rm gl},i})), \quad (3)$$

$$T_i = exp(-\sum_{j=1}^{i-1} \sigma_{{\rm gl},j}\delta_{{\rm gl},j}). \quad (4)$$

As a result, we obtain the amount of ray shifting that represents the distance each sampling point moves from its original coordinate by adding the weighted offset cumulatively along the ray, and we estimate the adjusted position of the sampling points $\boldsymbol{x}'$ as follows:

$$\boldsymbol{x}'_i = \boldsymbol{x}_i + \sum_{j=1}^{i} w_j \Delta\boldsymbol{x}_j. \quad (5)$$

4.2. Decomposition NeRF

We assume that an input scene can be separated into the view-independent component, which does not change based on the viewpoint, and the view-dependent component, which does change based on the viewpoint, as shown in Fig. 2. We then define two NeRF-like fields representing each component.

4.2.1 View independent NeRF

The view-independent components are represented using density $\sigma_{\rm vi} \in \mathbb{R}$ and color $\boldsymbol{c}_{\rm vi} \in \mathbb{R}^3$, similar to the conventional NeRF method. However, the view-independent components do not require the view direction; thus, in the proposed method, we only use a former part of $\Theta_{\rm nf}$, which takes the position $\boldsymbol{x}'$ as input. By performing volume rendering with the $(\sigma_{\rm vi}, \boldsymbol{c}_{\rm vi})$ of the points sampled on a ray $\boldsymbol{r}$, we obtain the view-independent color $\boldsymbol{C}_{\rm vi}$ of a pixel corresponding to that ray as follows:

$$\boldsymbol{C}_{\rm vi}(\boldsymbol{r}) = \sum_{i=1}^{N} T_{{\rm vi},i}(1 - exp(-\sigma_{{\rm vi},i}\delta_{{\rm vi},i}))\boldsymbol{c}_{{\rm vi},i}. \quad (6)$$

Note that the calculation of $T_{{\rm vi},i}$ is the same as given in Eq. 4.

4.2.2 View dependent NeRF

The view-dependent components are separated using the feature field approach [41]. MS-NeRF introduced the feature field, which extracts multiple spaces explicitly as the

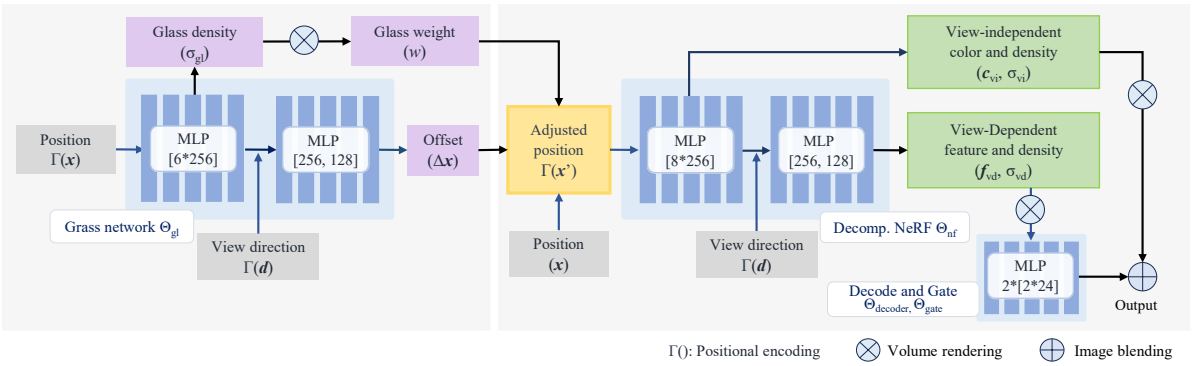


Figure 3. Network architecture of the proposed method. The glass network outputs the glass density and offset, which modify the ray by the refraction effect through glass walls. Here, the glass density is view-independent, and the offset is view-dependent. The NeRF network takes the adjusted position as input and outputs the view-independent and view-dependent densities and color or feature. The feature renderer provides the corresponding feature map, and the decoder and gate MLPs convert the rendered feature map to a view-dependent image with a blending weight. Finally, the image blending module generates the image by composing the rendered view-independent and view-dependent images. The training process minimizes the loss calculated from the composed image and the input image while optimizing the MLPs.

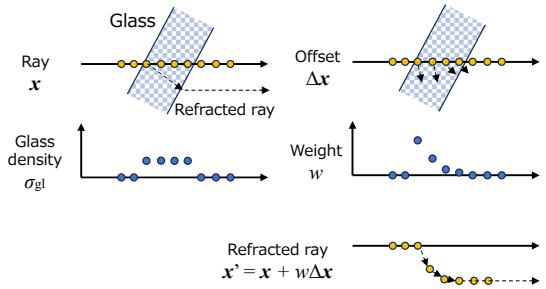


Figure 4. Structure of the proposed method to express light refraction as volume rendering using glass density to estimate the offset. Here, refraction is simplified as a parallel translation in 3D space occurring on the glass surface. We estimate the path of the light considering refraction by accumulating the vectors of this translation.

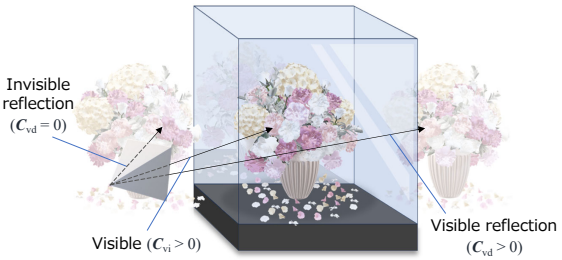


Figure 5. A proposal method structure that divides the scene into two fields. Elements that do not change depending on the viewpoint, e.g., objects and backgrounds in the scene, are represented in the view-independent field. Elements that do change depending on the viewpoint, e.g., reflections caused by glass and reflections from light sources, are represented in the view-dependent field, where the density changes based on the viewpoint.

density and feature fields, and it functions effectively for scenes with several mirrors. However, multiple glass plates generate reciprocal reflections of objects and light sources; thus, representing all spaces individually with a number of spaces is a highly complex task.

In the proposed method, we address this problem using a single view-dependent feature field in addition to the previously describe view-independent NeRF. The feature field is represented using Θ_{nf}, which estimates the view-dependent density σ_{vd} and the θ-dimensional feature vector $\boldsymbol{f}_{\mathrm{vd}}$ from an adjusted position $\boldsymbol{x}'$ and view direction $\boldsymbol{d}$ as follows:

$$\mathcal{F}_{\Theta_{\mathrm{nf}}} : \Gamma(\boldsymbol{x}'), \Gamma(\boldsymbol{d}) \rightarrow \sigma_{\mathrm{vi}}(\boldsymbol{x}'), \boldsymbol{c}_{\mathrm{vi}}(\boldsymbol{x}'), \sigma_{\mathrm{vd}}(\boldsymbol{x}', \boldsymbol{d}), \boldsymbol{f}_{\mathrm{vd}}(\boldsymbol{x}', \boldsymbol{d}). \quad (7)$$

We obtain the feature vector $\boldsymbol{F}_{\mathrm{vd}}$ corresponding to a ray $\boldsymbol{r}$ by volume rendering along the ray for σ_{vd} and $\boldsymbol{f}_{\mathrm{vd}}$, In addition, we estimate the color $\boldsymbol{C}_{\mathrm{vd}}$ using a decoder MLP Θ_{dc} and determine the blending parameter α using a gate MLP Θ_{gt} [41]. Here, $\boldsymbol{C}_{\mathrm{vd}}(\boldsymbol{r})$ represents the reflection component corresponding to the pixel of the ray.

$$\boldsymbol{F}_{\mathrm{vd}}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i (1 - exp(-\sigma_{\mathrm{vd},i}\delta_{\mathrm{vd},i}))\boldsymbol{f}_i, \quad (8)$$

$$\mathcal{F}_{\Theta_{\mathrm{dc}}} : \boldsymbol{F}_{\mathrm{vd}}(\boldsymbol{r}) \rightarrow \boldsymbol{C}_{\mathrm{vd}}(\boldsymbol{r}), \quad (9)$$

$$\mathcal{F}_{\Theta_{\mathrm{gt}}} : \boldsymbol{F}_{\mathrm{vd}}(\boldsymbol{r}) \rightarrow \alpha(\boldsymbol{r}). \quad (10)$$

4.3. Optimization

Finally, we find the color $\boldsymbol{C}(\boldsymbol{r})$ of the pixel corresponding to the ray as follows:

$$\boldsymbol{C}(\boldsymbol{r}) = \boldsymbol{C}_{\mathrm{vi}}(\boldsymbol{r}) + \alpha(\boldsymbol{r})\boldsymbol{C}_{\mathrm{vd}}(\boldsymbol{r}). \tag{11}$$

We train the MLPs $\{\Theta_{\mathrm{gl}}$,Θ_{nf}, Θ_{dc}, and $\Theta_{\mathrm{gt}}\}$ by evaluating the rendered pixel color $\boldsymbol{C}(\boldsymbol{r})$ with that of the input image $\bar{\boldsymbol{C}}(\boldsymbol{r})$. Here, we utilize the same loss function utilized in the conventional NeRF method, i.e., the summation of L2 distances, which is expressed as follows:

$$\mathcal{L}_{\mathrm{render}} = \sum_{\boldsymbol{r}\in\boldsymbol{R}} \|\bar{\boldsymbol{C}}(\boldsymbol{r}) - \boldsymbol{C}(\boldsymbol{r})\|^2, \tag{12}$$

where $\boldsymbol{R}$ is a batch of sampled rays. In addition, we introduce L2 regularization loss $\mathcal{L}_{\mathrm{offset}}$ to train the glass network in a stable manner. This regularization loss prevents the neural network from learning biased offset and parallelly shifted scenes.

$$\mathcal{L}_{\mathrm{offset}} = \sqrt{\sum_{\boldsymbol{x},\boldsymbol{d}} (\Delta\boldsymbol{x})^2} \tag{13}$$

The entire loss $\mathcal{L}$ is an integration of these two loss functions:

$$\mathcal{L} = \mathcal{L}_{\mathrm{render}} + \epsilon\mathcal{L}_{\mathrm{offset}}, \tag{14}$$

where ϵ is a small number, which was set to $\epsilon = 10^{-5}$ in our experiments.

5. Experiment

5.1. Experimental dataset

In this study, we generated a simulation dataset using Blender [1]. The experimental dataset includes several scenes containing a glass showcase with the size of 50cm $\times$ 50cm $\times$ 50cm surrounded by walls with textures. Here, the thickness of the glass is 1cm, and the refractive index is 1.45. Inside this showcase is an object in {Lego, House, Color Ball, Flower}. We also generated a dataset of Lego object placed in art gallery, which is surrounded by wall with paintings. We placed a ceiling light with area in the scene and utilized Blender's BSDF model to render the scene with physical simulations of both the light and the glass. One set of a scene comprised 200 training images, and both the test and validation sets contained 25 images each with their respective intrinsic and extrinsic camera parameters.

5.2. Implementation and training

We implemented the proposed method based on NeRF-Pytorch [40]. The training process sampled the rays corresponding to 1,024 pixels from a randomly selected training image in each iteration. In addition, each ray initially sampled 128 points at uniform intervals. In the coarse-to-fine strategy of NeRF, an additional 64 points are sampled in a hierarchical manner for segments with higher densities inferred from the coarse model. The proposed method samples an additional 32 points using the glass density and 32 points using the view-independent density. This results in a total of 192 sampling points being input into the fine model. We set the feature vector's dimension θ to 64. We performed a total of 200,000 training iterations on an Nvidia RTX 4080 graphics processing unit, which took approximately 10 hours for a single scene.

5.3. Evaluation

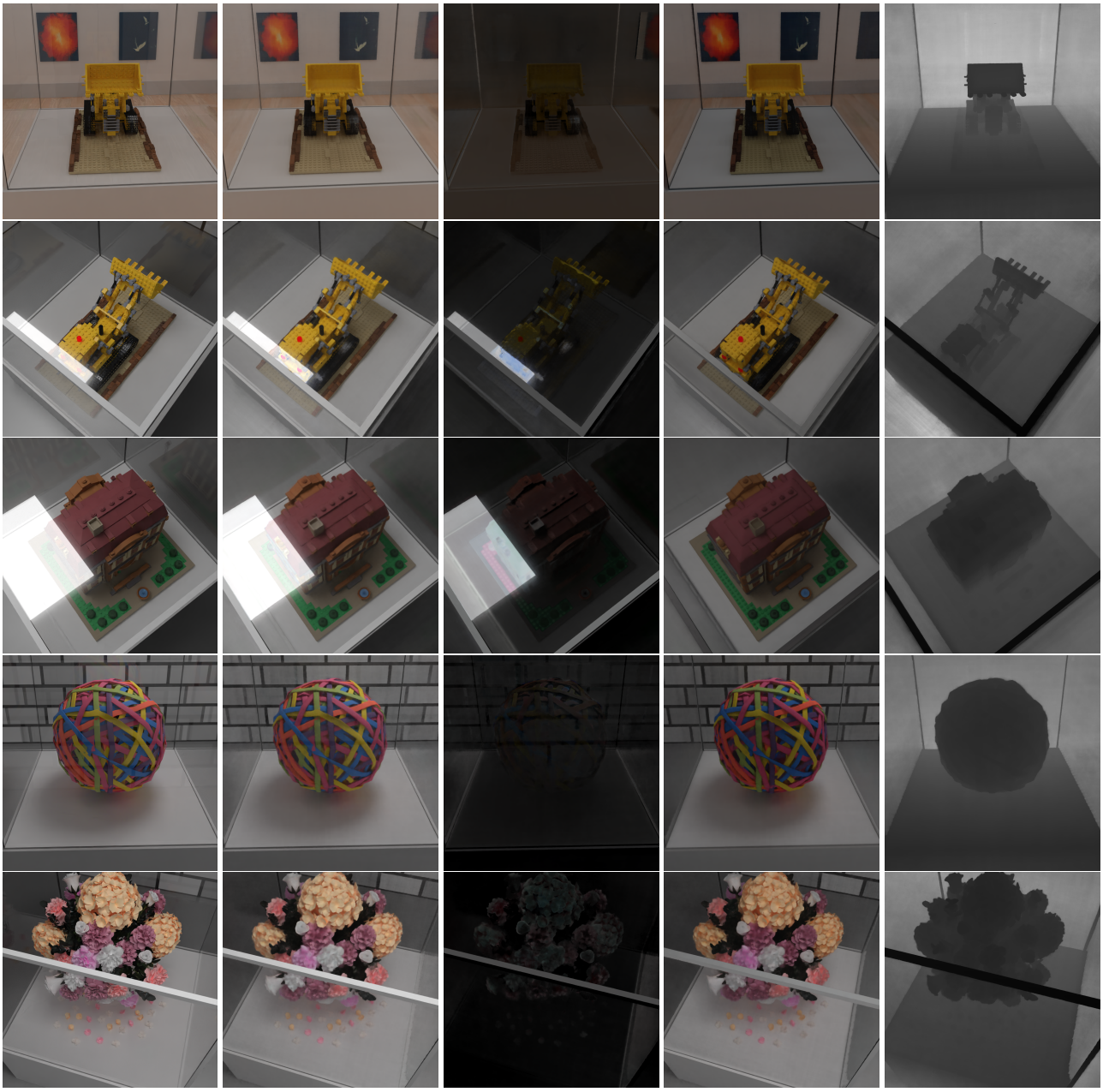
Figure 6 shows examples of the images obtained by the proposed method using test images and camera poses. Here, each row (from top to bottom) shows the results for Lego (Gallery), Lego, House, Color Ball, and Flower, respectively. In addition, each row (from left to right) shows the ground-truth image (i.e., the test image), the rendered image, the view-dependent component, the view-independent component, and the corresponding depth image.

The view-dependent images with specular reflection components demonstrate that particularly strong reflections were extracted correctly by the proposed method. In addition, the view-dependent image also shows that the reflection component was removed effectively. However, closer observation indicates that a small amount of the direct component remains in the view-dependent images and the the reflective component remains on the glass at the back in the view-independent images. The depth image was estimated correctly, which indicates that the refraction was estimated well, and the reflective component, which is problematic for shape estimation, was removed effectively.

The proposed method estimates the points where refraction occurs explicitly; thus, the accuracy of the estimated refractive surface can be evaluated according to how close these points are to the original glass surface. We determined that the estimated glass surface point where the offset $\Delta x \times w$ is greater than a threshold value (0.01 cm in this experiment) when rendering test images. An example of the estimated glass surface is shown in Fig. 7, and Table 1 shows the average error of the estimated glass surfaces. Although there are some outliers, it can be seen that reasonably good results were obtained by the proposed method.

5.4. Comparative evaluation

In this evaluation, the proposed method was compared with several NeRF-based methods. Here, we applied the original NeRF [22], Mip-NeRF [2], Ref-NeRF [32], LB-NeRF [8], and MS-NeRF [41] methods to the experimental dataset constructed in this study. Note that no open source code is available for LB-NeRF; thus, we implemented its structure by adding a 3D offset estimated from a concatenation



(a) Ground truth (b) Rendered image (c) View-dependent component (d) View-independent component (e) View-independent depth

Figure 6. Result images of our proposal method modeling test datasets. (a) Ground-truth image of scene viewed from a test pose. (b) The image rendered by the proposed method. (c) The view-dependent component and (d) view-independent component modeled by the proposed method. (e) Depth map image of the view-independent component.

of the 3D point's positions and view direction using a fully-connected neural network with seven layers containing 256 nodes. In this evaluation, evaluation metrics commonly used for image comparison were used to assess the compared methods, i.e., the peak signal-to-noise Ratio (PSNR), the structural similarity index measure (SSIM) [36], and the learned perceptual image patch similarity (LPIPS) [45].

Table 2 shows the results. As can be seen, the proposed method outperformed all compared methods on each dataset. Among the compared method, the MS-NeRF method obtained comparatively superior results. Note that the MS-NeRF method does not clearly determine the de-

Table 1. Evaluation of glass surface estimation

Dataset	Distance (cm)
Lego	0.4657
House	0.4728
Color Ball	0.3553
Flower	0.2693
Lego (Gallery)	0.9662

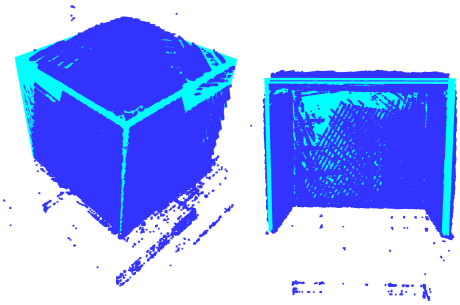


Figure 7. Diagram representing points in the Lego scene where refraction greater than 0.01 cm occurred as a point cloud. The results confirm that the surfaces of the glass in the scene were estimated to have a high glass density.

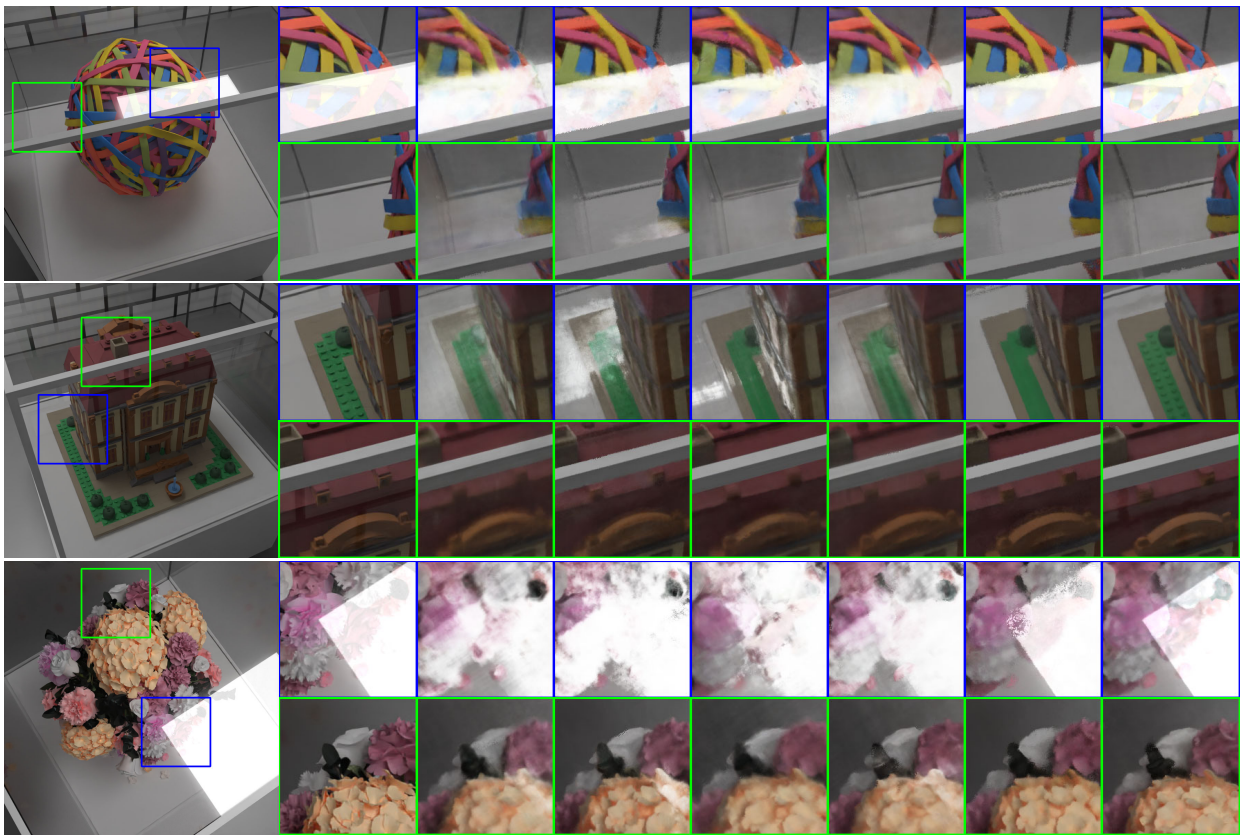
composed spaces for nonmirror surfaces; thus, it enhances the image synthesis precision by separating the intense reflective components associated with glass.

The images in Fig. 8 show the view synthesis results obtained by all compared methods. As can be seen, it is evident that the existing methods struggled to accurately model scenes in complex situations where objects are surrounded by glass surfaces, e.g., showcases, with complicated reflections and refraction effects. In contrast, by separating and modeling the view-independent and view-dependent components of each effect, the proposed method obtained more accurate estimations than the compared existing methods.

6. Conclusion

In this paper, we have proposed a method that utilizes implicit neural representations to model scenes with objects enclosed in a glass case. The proposed method distinguishes between the refraction and reflection effects by learning them with view-independent and view-dependent components, and by determining the refraction points indicative of the glass surfaces simultaneously. The proposed method was evaluated experimentally, and the experimental results demonstrate that the proposed method is proficient in terms of separating the components that vary with viewpoint from those that do not. In addition, compared to several existing NeRF-based methods, the proposed method demonstrated clear superiority in terms of synthesizing novel views.

On the other hand, as explained in Sec. 5.3, the separation of view-dependent and view-independent components may not be sufficient. Improving the accuracy of component separation is one of the future works.



(a) Ground truth (b) Ground truth (c) NeRF (d) Mip-NeRF (e) Ref-NeRF (f) LB-NeRF (g) MS-NeRF (h) Proposed method

Figure 8. Qualitative evaluation of comparative methods.

Table 2. Results of each Dataset and methods

Method	Lego			House			Color Ball			Flower			Lego (Gallery)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [22]	29.0111	0.8893	0.2984	30.9108	0.9029	0.2951	29.2773	0.9054	0.2766	27.8217	0.8557	0.3351	31.7363	0.9184	0.2772
Mip-NeRF [2]	28.9192	0.8923	0.2979	30.6958	0.9100	0.2838	29.9715	0.9211	0.2346	27.9082	0.8711	0.3055	32.1508	0.9267	0.2565
Ref-NeRF [32]	28.2375	0.8679	0.3357	29.6852	0.8836	0.3262	29.0698	0.8944	0.2880	27.3446	0.8396	0.3642	31.3128	0.9104	0.3011
LB-NeRF [8]	29.8683	0.9094	0.2711	30.9116	0.9028	0.2980	30.0377	0.9151	0.2525	28.4501	0.8683	0.3198	32.4674	0.9324	0.2564
MS-NeRF [41]	31.0450	0.9060	0.2817	33.3059	0.9231	0.2822	31.6074	0.9230	0.2594	29.1936	0.8740	0.3197	33.0800	0.9274	0.2665
Proposal Method	**33.1834**	**0.9357**	**0.2073**	**35.3665**	**0.9483**	**0.1929**	**33.0759**	**0.9453**	**0.1842**	**30.6254**	**0.8958**	**0.2637**	**35.3476**	**0.9507**	**0.1913**

References

[1] blender.org - home of the blender project - free and open 3d creation software. 6

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2, 6, 9

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[4] Chris Beall, Brian J Lawrence, Viorela Ila, and Frank Dellaert. 3d reconstruction of underwater structures. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4418–4423. IEEE, 2010. 2

[5] F. Chadebecq, F. Vasconcelos, G. Dwyer, R. Lacher, S. Ourselin, T. Vercauteren, and D. Stoyanov. Refractive structure-from-motion through a flat refractive interface. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5325–5333, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2

[6] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE,

2023. 3

[7] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[8] Taku Fujitomi, Ken Sakurada, Ryuhei Hamaguchi, Hidehiko Shishido, Masaki Onishi, and Yoshinari Kameda. Lb-nerf: light bending neural radiance fields for transparent medium. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2142–2146. IEEE, 2022. 3, 4, 6, 9

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[10] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maujean, Gregoire De Lillo, and Yann Lanthony. Alicevision meshroom: An open-source 3d reconstruction pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 241–247, 2021. 2

[11] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 3

[12] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-nerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*, 2021. 3

[13] Anne Jordt, Kevin Köser, and Reinhard Koch. Refractive 3d reconstruction on underwater images. *Methods in Oceanography*, 15:90–113, 2016. 2

[14] Justin Kerr, Letian Fu, Huang Huang, Yahav Avigal, Matthew Tancik, Jeffrey Ichnowski, Angjoo Kanazawa, and Ken Goldberg. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *6th Annual Conference on Robot Learning*, 2022. 3

[15] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *ACM Trans. Graph.*, 41(6), 2022. 3

[16] Zhengqin Li, Yu-Ying Yeh, and Manmohan Chandraker. Through the looking glass: neural 3d reconstruction of transparent shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1262–1271, 2020. 2

[17] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[18] Jiahui Lyu, Bojian Wu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Differentiable refraction-tracing for mesh reconstruction of transparent objects. *ACM Trans. Graph.*, 39(6), 2020. 2

[19] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[21] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 2, 4

[22] 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 *ECCV*, 2020. 1, 2, 6, 9

[23] D. Miyazaki, M. Kagesawa, and K. Ikeuchi. Transparent surface modeling from a pair of polarization images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):73–82, 2004. 2

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[25] Jen-I Pan, Jheng-Wei Su, Kai-Wen Hsiao, Ting-Yu Yen, and Hung-Kuo Chu. Sampling neural radiance fields for refractive objects. In *SIGGRAPH Asia 2022 Technical Communications*, pages 1–4. 2022. 3

[26] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[27] Xiaorui Qiao, Atsushi Yamashita, and Hajime Asama. Underwater structure from motion for cameras under refractive surfaces. *Journal of Robotics and Mechatronics*, 31(4):603–611, 2019. 2

[28] Jiaxiong Qiu, Peng-Tao Jiang, Yifan Zhu, Ze-Xin Yin, Ming-Ming Cheng, and Bo Ren. Looking through the glass: Neural surface reconstruction against high specular reflections, 2023. 3

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[30] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 519–528, 2006. 2

[31] Jinguang Tong, Sundaram Muthu, Fahira Afzal Maken, Chuong Nguyen, and Hongdong Li. Seeing through the glass: Neural 3d reconstruction of object inside a transparent container. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12555–12564, 2023. 3

[32] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition (CVPR), pages 5481–5490. IEEE, 2022. 2, 3, 6, 9

[33] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 2

[34] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[35] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[37] 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

[38] Ziyu Wang, Wei Yang, Junming Cao, Qiang Hu, Lan Xu, Junqing Yu, and Jingyi Yu. Neref: Neural refractive field for fluid surface reconstruction and rendering. In *2023 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2023. 3

[39] Bojian Wu, Yang Zhou, Yiming Qian, Minglun Cong, and Hui Huang. Full 3d reconstruction of transparent objects. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018. 2

[40] Lin Yen-Chen. Nerf-pytorch. https://github.com/yenchenlin/nerf-pytorch/, 2020. 6

[41] Ze-Xin Yin, Jiaxiong Qiu, Ming-Ming Cheng, and Bo Ren. Multi-space neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12407–12416, 2023. 3, 4, 5, 6, 9

[42] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[43] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[44] Yifan Zhan, Shohei Nobuhara, Ko Nishino, and Yinqiang Zheng. Nerfrac: Neural radiance fields through refractive surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18402–18412, 2023. 3

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[46] Chengxuan Zhu, Renjie Wan, and Boxin Shi. Neural transmitted radiance fields. In *Advances in Neural Information Processing Systems*, pages 38994–39006. Curran Associates, Inc., 2022. 3

Supplementary material: REF^2-NeRF: Reflection and Refraction aware Neural Radiance Field

Anonymous CVPR submission

Paper ID 11276

1. Implementation details

To ensure verifiability, we provide detailed network configurations in Figs. 1 and 2. These networks are structured using Multilayer Perceptrons (MLPs). The "+" symbols in the figures represent data concatenation. We will make the code openly available once it has been organized.

2. Computational efficiency

We haven't attempted to optimize the proposed method for efficient computation, but for reference, we will provide the execution times. We performed 200,000 training iterations on an NVIDIA RTX 4080 for a single scene. Table 1 shows the computation times for each method when trained with the same number of iterations as mentioned above. In the current implementation, our method needs to train two networks: the glass network and the decomposition network. Additionally, we haven't employed optimizations like hash encoding for faster computation. As a result, it requires more computational time compared to other methods. Improving computational efficiency is one of the future works.

Table 1. Training time of each method (hours)

Method	Lego scene
NeRF [**?**]	5.3
Mip-NeRF [**?**]	8.2
Ref-NeRF [**?**]	6.6
LB-NeRF [**?**]	5.4
MS-NeRF [**?**]	3.6
Proposal Method	10.8

Table 2. Evaluation of glass surface estimation in gallery dataset

Dataset	Distance (cm)
Lego	0.9662
House	0.7361
Color ball	0.6812
Flower	-

3. Additional experimental results

The proposed method is thought to be affected not just by the objects inside the glass case but also by the surrounding background. We've included the experimental result with the LEGO and gallery background dataset in the main paper. We include the results for other objects in this document due to the page limitation in the main paper. We compared the proposed method with the original NeRF [**?**], Mip-NeRF [**?**], Ref-NeRF [**?**], LB-NeRF [**?**], and MS-NeRF [**?**]. The evaluation metrics were the same as the main paper, i.e., PSNR, SSIM [**?**], and LPIPS [**?**].

Table 3 shows the results. Similar to the results with the brick structure background, the proposed method outperformed all compared methods on each dataset. On the other hand, when looking at the results in Fig. 3, we can observe that in the Flower dataset, the method cannot effectively separate the view-dependent and view-independent components. As mentioned in the main paper, the method does not guarantee the clear separation of these components, which poses a significant challenge for future research.

Table 2 shows the average distance between the estimated points by the glass network and the original data's glass surface in the gallery dataset. In the case of House and Color ball, the method could accurately estimate the glass surface. However, as mentioned above, in the case of Flower where components' decomposition does not work well, we have found that the glass surface is also not effectively estimated.

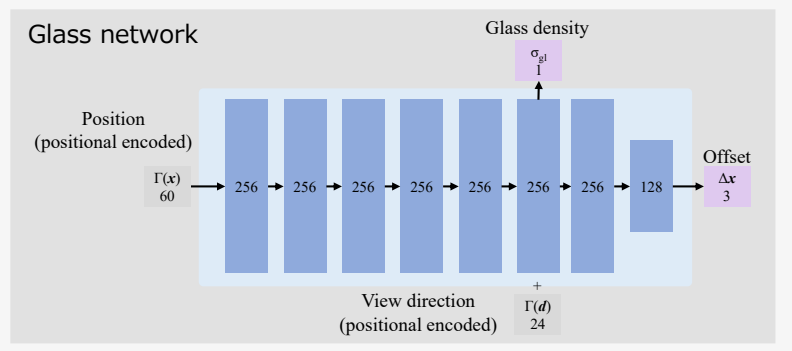


Figure 1. Network architecture (Glass network).

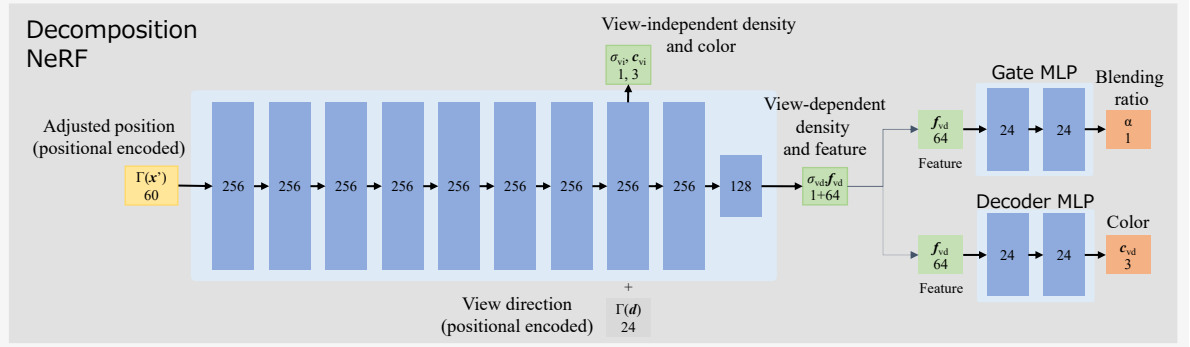


Figure 2. Network architecture (Decomp. NeRF).

Table 3. Results of Gallery background dataset

Method	Lego			House			Color Ball			Flower		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [**?**]	31.7363	0.9184	0.2772	31.2738	0.9162	0.3154	29.5172	0.9031	0.3229	29.4582	0.8813	0.3283
Mip-NeRF [**?**]	32.1508	0.9267	0.2565	31.8706	0.9274	0.2771	30.4037	0.9261	0.2518	29.4503	0.9023	0.2955
Ref-NeRF [**?**]	31.3128	0.9104	0.3011	32.3337	0.9198	0.3092	30.2798	0.9111	0.3162	30.0246	0.8830	0.3348
LB-NeRF [**?**]	32.4674	0.9324	0.2564	30.8119	0.9187	0.2873	31.2125	0.9126	0.3292	29.2704	0.8778	0.3388
MS-NeRF [**?**]	33.0800	0.9274	0.2665	36.1140	0.9475	0.2442	33.0972	0.9348	0.2523	**31.3365**	**0.9108**	**0.2942**
Proposal Method	**35.3476**	**0.9507**	**0.1913**	**36.8355**	**0.9517**	**0.2167**	**34.8838**	**0.9529**	**0.1930**	28.7231	0.8512	0.4743

(a) Ground truth (b) Rendered image (c) View-dependent component (d) View-independent component (e) View-independent Depth

Figure 3. Result images of our proposal method modeling test datasets. (a) Ground-truth image of scene viewed from a test pose. (b) The image rendered by the proposed method. (c) The view-dependent component and (d) view-independent component modeled by the proposed method. (e) Depth map image of the view-independent component.