

NeRFReN: Neural Radiance Fields with Reflections

Yuan-Chen Guo^{1*} Di Kang² Linchao Bao² Yu He^{3†} Song-Hai Zhang¹

¹BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing

²Tencent AI Lab ³BIMSA

guoyc19@mails.tsinghua.edu.cn {dkang, linchaobao}@tencent.com
hooyeeevan2511@gmail.com shz@tsinghua.edu.cn

Abstract

Neural Radiance Fields (NeRF) has achieved unprecedented view synthesis quality using coordinate-based neural scene representations. However, NeRF’s view dependency can only handle simple reflections like highlights but cannot deal with complex reflections such as those from glass and mirrors. In these scenarios, NeRF models the virtual image as real geometries which leads to inaccurate depth estimation, and produces blurry renderings when the multi-view consistency is violated as the reflected objects may only be seen under some of the viewpoints. To overcome these issues, we introduce NeRFReN, which is built upon NeRF to model scenes with reflections. Specifically, we propose to split a scene into transmitted and reflected components, and model the two components with separate neural radiance fields. Considering that this decomposition is highly under-constrained, we exploit geometric priors and apply carefully-designed training strategies to achieve reasonable decomposition results. Experiments on various self-captured scenes show that our method achieves high-quality novel view synthesis and physically sound depth estimation results while enabling scene editing applications. Project webpage: <https://bennyguo.github.io/nerfren/>.

1. Introduction

Photorealistic novel view synthesis (NVS) from unstructured image collections is crucial to creating immersive virtual experiences. Despite having achieved significant progress in controlled settings, challenges still exist in handling light transport at the surface of different materials. For example, reflections caused by glass or mirrors commonly exist in real-world scenes, posing great difficulties for novel view synthesis due to the severe view-dependent effects.

Neural Radiance Fields (NeRF) [13], as an emerging

*Research done when Yuan-Chen Guo was an intern at Tencent AI Lab.

†Corresponding author.

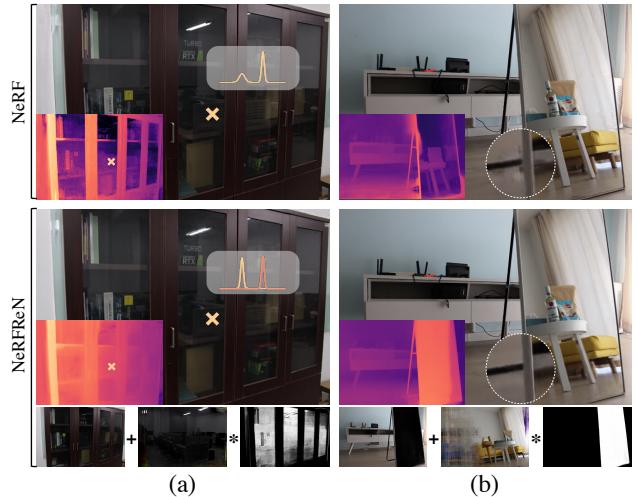


Figure 1. NeRF models the stable reflection image as real geometries rather than view-dependent effect of the points on the reflective surface. We illustrate two problems NeRF may encounter when modeling scenes with complex reflections: (1) inaccurate depth estimation in reflective regions, as in (a) and (b); (2) inaccurate rendering when the multi-view consistency is violated, as in the magnified area in (b). NeRFReN tackles these problems by modeling the transmitted and reflected component of the scene with separate NeRFs, and synthesizing views in the image domain (bottom). See Sec. 3.1 for more detailed analysis.

technique in this field, has achieved impressive view synthesis quality by adopting volumetric representations with coordinate-based neural networks. By conditioning radiance on both coordinates and viewing directions, NeRF is able to faithfully handle view-dependent effects like highlights. However, in the presence of severe reflections that contain a stable virtual image other than highlights, NeRF tends to model the scene geometry behind the reflective surface as translucent (like fogs) and regard the reflected geometries as appearing at some virtual depth. The reason is that NeRF does not explain the view-dependency of a reflecting surface point by changing the color of this very

point when viewed from different directions, but explain it by utilizing all the spatial points behind it along the camera ray to get the correct color through volume rendering as in Eq. (2), resulting in foggy geometries and physically wrong depth. Although this works well for view synthesis in certain scenarios, there are two inherent limitations: (1) it is hard to estimate the correct scene geometry, preventing it from further explanations or editing; (2) inaccurate renderings are generated when the multi-view consistency does not hold since the reflected objects may only be seen in some of the images, which limits the views NeRF can correctly synthesize.

To this end, we propose NeRFReN, which enhances NeRF to handle scenes with ReflectioNs. Instead of representing the whole scene with a single neural radiance field, we propose to model the *transmitted* and *reflected* parts of the scene with separate neural radiance fields. To synthesize novel views, the transmitted image I_t and reflected image I_r rendered by the corresponding radiance fields are composed in an additive way, where the reflected image I_r is weighted by a learned *reflection fraction map* β :

$$I = I_t + \beta I_r \quad (1)$$

However, it is highly under-constrained to decompose the scene into transmitted and reflected components in an unsupervised way. There are infinitely many solutions that could lead to correct renderings on training images. Common degradations are (1) having one of the components explain the full scene while leaving the other empty; (2) two components both model the full scene; (3) somewhere between (1) and (2), where part of the one component is mixed into the other one, resulting in incomplete transmitted/reflected geometry. To solve the ambiguities, we made the following assumptions:

- **Assumption1:** the reflection fraction is only related to the transmitted component, since it indicates the material of the reflecting surface;
- **Assumption2:** the transmitted component has a locally smooth depth map, since most of the reflectors in real world scenes are almost planar;
- **Assumption3:** the reflected component requires only simple geometry to present correct renderings since most of the time we can only see the reflection image from very limited viewing directions.

These assumptions generally hold true for planar or almost planar reflective surfaces commonly seen in real-world scenes. Accordingly, we design a specific network architecture for Assumption1, apply a depth smoothness prior for Assumption2, and a novel bidirectional depth consistency constraint for Assumption3. Together with our warm-up training strategy, we are able to faithfully decompose the

transmitted and reflected components, achieving promising novel synthesis results on various real-world scenes. As for even more challenging cases (such as mirrors) in which the ambiguities cannot be solved by the specific design, we exploit an interactive setting where a small number of user-provided reflection masks can be utilized to get the correct decomposition (Fig. 4, tv and mirror).

To summarize, our contributions are:

- We analyze NeRF’s behavior and limitations in modeling reflective regions, and propose to use separate transmitted and reflected neural radiance fields for scenes with complex reflections.
- We devise specific network architectures, exploit geometric priors, and apply carefully-designed warm-up training strategy, to achieve physically sound decomposition results with competitive novel view synthesis quality compared to vanilla NeRF on several real world scenes, including challenging cases like scenes with mirrors.
- We also investigate depth estimation and scene editing as further applications based on the decomposed results.

2. Related Work

Neural scene representations for novel view synthesis. Novel view synthesis aims to generate images observed from novel views based on a set of captured images of the scene. In order to get correct synthesis results, the underlying 3D geometry of the scene must be taken into consideration, for which various scene representations are proposed. Image-based rendering [3, 7, 16, 17, 25, 27] mainly utilizes a mesh model of the scene, which is usually reconstructed from offline Structure-from-Motion (SfM) and Multi-View-Stereo (MVS) methods [19, 20]. Other commonly adopted representations include voxel grid [23], volume [10], point cloud [1] and multi-plane images [5, 30].

In recent years, coordinate-based neural representations [4, 12–14, 24] have shown remarkable capability of modeling the 3D world. Among them, Neural Radiance Fields (NeRF) [13], as an emerging technique in the novel view synthesis domain, models the scene as a continuous volumetric field parameterized by neural networks. NeRF produces astonishing novel view synthesis results in several casually-captured real scenes, even when view-dependent effects like highlights present. Our work extends NeRF to support view synthesis in scenes with reflective surfaces like glass and mirrors, which are common in real world scenes and strongly affect the immersion of virtual experiences.

Reflections in rendering. Reflections are common in real world, but can be hard to correctly model without simulating the light paths using techniques like ray-tracing. In computer graphics, Screen Space Reflection (SSR) technique is developed to simulate the reflection effects with

low costs. Synthesizing novel views in scenes with reflections has long been a challenging problem, but is rarely explored [21, 31]. Sinha *et al.* [21] for the first time targets at novel view synthesis in scenes with reflections using image-based rendering. They propose to decompose each image into a transmitted layer and a reflected layer combined with a binary reflection mask. They approximate the geometry of each layer as piece-wise planar surfaces, and estimate their depth by two-layer stereo matching. The reflection mask is optimized using graph cut based on depth uncertainty.

Our work adopts similar image formulation as Sinha *et al.* [21], but models the scene with NeRF for its state-of-the-art performance on the novel view synthesis task. NeRF resolves view-dependency by taking the viewing direction as network input, but is only able to model low-frequency view-dependent effects. We extend NeRF by modeling the transmitted and reflected components with two separate neural radiance fields, and achieve visually appealing results in scenes with complex reflections.

Our work is also different from existing NeRF-based inverse rendering methods [2, 35], which assuming opaque surfaces with simple BRDF, distant environment lighting and local illumination models. These assumptions are not suitable in our case with complex reflections. In contrast, we assume composition of transmitted and reflected components in an additive way.

3. Method

Here we present NeRFReN, a neural radiance field approach for novel view synthesis of scenes with reflections. We first briefly recap Neural Radiance Fields (NeRF) [13] in Sec. 3.1, and present our scene formulation and network architecture in Sec. 3.2. To faithfully decompose the scene into transmitted and reflected components, we exploit geometric priors (Sec. 3.3) and apply specifically designed training strategies (Sec. 3.4). Training with minimum user input to handle hard cases is described in Sec. 3.5.

3.1. Neural Radiance Fields Revisited

Neural Radiance Fields (NeRF) represents a scene as a continuous volumetric field, where the density $\sigma \in \mathbb{R}$ and radiance $\mathbf{c} \in \mathbb{R}^3$ at any 3D position $\mathbf{x} \in \mathbb{R}^3$ under viewing direction $\mathbf{d} \in \mathbb{R}^2$ are modeled by a multi-layer perceptron (MLP) $f_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, with θ as learnable parameters. To render a pixel, the MLP first evaluates points sampled from the camera ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ to get their densities and radiance, and then the color $\mathbf{C}(\mathbf{r})$ is estimated by volume rendering equation approximated using quadrature [11]:

$$\hat{\mathbf{C}}(\mathbf{r}; \sigma, \mathbf{c}) = \sum_k T_i(\sigma)(1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \quad (2)$$

where $\delta_i = t_{i+1} - t_i$ and $T_i(\sigma) = \exp(-\sum_{j < i} \sigma_j \delta_j)$. $\hat{\mathbf{C}}$ is conditioned on σ, \mathbf{c} , and T is conditioned on σ to sim-

plify follow-up descriptions. We denote the contribution of a point to the cumulative color as its weight ω_i :

$$\omega_i = T_i(\sigma)(1 - \exp(-\sigma_i \delta_i)) \quad (3)$$

NeRF is optimized by minimizing the photometric loss:

$$\mathcal{L}_{pm} = \|\hat{\mathbf{C}} - \mathbf{C}\|_2 \quad (4)$$

The depth value t^* along the ray can be estimated by computing the expected termination depth [35]:

$$t^*(\mathbf{r}; \sigma) = \sum_k \omega_i t_i = \sum_k T_i(\sigma)(1 - \exp(-\sigma_i \delta_i)) t_i \quad (5)$$

NeRF models view-dependent effects by taking viewing direction \mathbf{d} as the network input. But as will be demonstrated in Tab. 1, the incorporation of \mathbf{d} only enables NeRF to represent low-frequency view-dependent effects. Directly applying NeRF on scenes with severe reflections leads to a mixed geometry containing both the transmitted part and the reflected part of the scene, where the former is modeled as *translucent* to get correct view reconstructions. As illustrated in the first row of Fig. 1, the books in the bookcase appear to be “foggy”, and the mirror surface is completely transparent.

The drawbacks are two-folds. First, depth estimation obtained by Eq. (5) can be highly inaccurate, staying somewhere between the real depth of the reflective surface and the virtual depth of the reflected image due to the existence of translucent geometries. In the gray box of Fig. 1(a) we visualize the weights of points along the ray corresponding to the yellow cross. The weight distribution of NeRF appears to be bi-modal, with the left peak contributed by the real surface points and right peak by the virtual surface points. The expected termination depth along this ray is biased towards the virtual depth, visualized as a darker color in the depth map on the bottom left. Second, it is hard to get correct view synthesis results when the multi-view consistency is violated. We illustrate this in Fig. 1(b), where some of the reflected objects can only be seen when front-facing the mirror. But as NeRF models the reflected objects as real geometries, we can still observe foggy contents seeing from the side of the mirror as shown in the magnified area.

3.2. NeRFReN

To better deal with scenes containing reflections, we propose to decompose a scene into a transmitted NeRF and a reflected NeRF. The transmitted field has density σ^t and radiance \mathbf{c}^t , and the reflected field has density σ^f and radiance \mathbf{c}^f . A reflection fraction value α is learned for each 3D position to measure the reflection property of objects in different materials. To render a pixel along ray \mathbf{r} , we first render the two fields respectively to get $\hat{\mathbf{C}}(\mathbf{r}; \sigma^t, \mathbf{c}^t)$ and $\hat{\mathbf{C}}(\mathbf{r}; \sigma^r, \mathbf{c}^r)$,

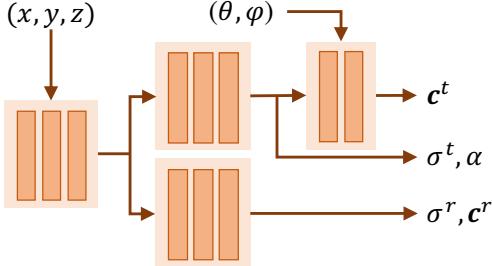


Figure 2. Network architecture of NeRFReN. The transmitted and reflected properties are predicted by separate branches of the network. Note that the reflection fraction value α is predicted from the transmitted branch.

where $\hat{\mathbf{C}}$ is defined in Eq. (2). The reflection fraction β corresponding to the pixel is accumulated via volume rendering based on the geometry of the transmitted part:

$$\beta(\mathbf{r}; \sigma, \alpha) = \sum_k T_i(\sigma^t)(1 - \exp(-\sigma_i^t \delta_i)) \alpha_i \quad (6)$$

The reflected color is attenuated by β and composed with the transmitted color additively to get the final pixel color:

$$\hat{\mathbf{C}} = \hat{\mathbf{C}}(\mathbf{r}; \sigma^t, \mathbf{c}^t) + \beta(\mathbf{r}; \sigma^t, \alpha) \hat{\mathbf{C}}(\mathbf{r}; \sigma^r, \mathbf{c}^r) \quad (7)$$

Fig. 2 shows the network architecture of NeRFReN. Detailed network configurations can be found in the supplementary material. Note that the reflection fraction value α is predicted along with transmitted density in the transmitted branch as it is a property of the reflective surface, uncorrelated with the reflected component. Also, the transmitted color is conditioned on viewing direction in our design, so that the low-frequency view-dependent effects (e.g. highlights) are modeled by the transmitted field, and the stable virtual image is modeled by the reflected field. The reflected color and the reflection fraction value are not conditioned on viewing direction. These assumptions approximate most scenes reasonably well in our experiments due to their validity¹ and greatly reduce network complexity, making the ill-posed decomposition problem much easier to learn.

3.3. Geometric Priors

Decomposing the scene into transmitted and reflected components is an under-constrained problem. There are infinite number of solutions and bad local minima that may produce visually pleasing rendering results on the training images but fail to separate apart the reflected radiance field from the transmitted radiance field. Humans identify the reflected virtual image correctly because we are aware of the real-world geometry. Inspired by this, we propose to utilize two geometric priors, namely a depth smoothness prior and a bidirectional depth consistency (BDC) prior, to guide the decomposition of the scene.

¹These assumptions may not hold if there exist reflectors in the reflected objects or due to severe Fresnel Effect.

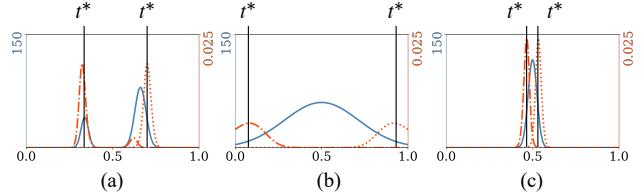


Figure 3. Illustrations for bidirectional depth consistency. Here we list three common density distributions (solid blue curve) along a ray in a neural radiance field. For each graph, the horizontal axis represents point locations along a ray. The dashdot and dotted orange curve are point weights (Eqs. (3) and (9)) at each location for calculating the forward and backward depth (Eqs. (5) and (9)) respectively. The computed depth values are marked with solid black lines. Only in case (c) we have a low BDC value, which corresponds to the expected shell-like geometry.

Depth smoothness. We exploit a general prior that the depth map of the *transmitted component* should be locally smooth. Specifically, we apply the following regularization:

$$\mathcal{L}_d = \sum_p \sum_{q \in \mathcal{N}(p)} \omega(p, q) \|t^*(p) - t^*(q)\|_1, \quad (8)$$

$$\omega(p, q) = \exp(-\gamma \|\mathbf{C}(p) - \mathbf{C}(q)\|_1)$$

where $t^*(p)$ is the approximated depth defined in Eq. (5), p denotes each pixel in an image, $\mathcal{N}(p)$ is the set of its 8-connected neighbors, \mathbf{C} is the image color and γ is a hyper-parameter. $\omega(p, q)$ is a decay factor to re-weight (relax) the constraint based on the color gradient since depth discontinuities are often accompanied with abrupt color changes. In this way, \mathcal{L}_d is edge-preserving and works only on small areas, avoiding over-smoothing in most cases.

This prior is reasonable under our setting since most of the reflectors in the real world are planar surfaces, which exhibit smooth depth changes in the reflected areas. Local depth smoothness priors are widely adopted in traditional stereo matching [21, 22, 33], but rarely explored in the training of neural representations. To apply this regularization during training, we sample patches instead of pixels.

Bidirectional depth consistency. The depth smoothness prior encourages correct geometry for the transmitted part. We now put constraints on the reflected part by assuming it has a simple geometry, where each ray only hits a single opaque surface. This assumption can help since we can only see the reflected objects from limited viewing angles so that they are, to some extend, similar to a textured *solid shell*. To describe this kind of simplicity in mathematical form, we propose a novel bidirectional depth consistency prior to encourage the aforementioned characteristic. We first define a new *backward depth* \overleftarrow{t}^* along ray \mathbf{r} as the expected termination point seeing the volume from the opposite di-

rection of \mathbf{r} :

$$\overleftarrow{t}^*(\mathbf{r}; \sigma) = \sum_k \overleftarrow{\omega}_i t_i = \sum_k \overleftarrow{T}_i(\sigma)(1 - \exp(-\sigma_i \delta_i)) t_i \quad (9)$$

where $\overleftarrow{T}_i(\sigma) = \exp(-\sum_{j>i} \sigma_j \delta_j)$. Denote \overrightarrow{t}^* as previous t^* , then our proposed bidirectional depth consistency (BDC) is defined as:

$$\mathcal{L}_{bdc} = \sum_{\mathbf{r}} \|\overrightarrow{t}^*(\mathbf{r}; \sigma^r) - \overleftarrow{t}^*(\mathbf{r}; \sigma^r)\|_1 \quad (10)$$

This regularization poses restriction on the density distribution along a ray, forcing it to be uni-modal and have small variance. The idea is demonstrated in Fig. 3. Here we consider **three representative types of density distribution** along the ray in a neural radiance field. If the ray hits multiple surfaces as in (a) or marches in a “foggy” area without a clear surface as in (b), the forward and backward depth is not consistent as marked in the figure. The two depth values become close only if the density distribution is uni-modal and has small variance as in (c). Intuitively, this requires the reflected component to have a “shell”-like geometry. Note that $\mathcal{L}_{bdc} = 0$ could also correspond to an empty ray. We claim that this would not happen because the transmitted component alone cannot faithfully reproduce the input images due to the existence of the depth smoothness prior. Some concurrent works [15, 18] also address the density distribution problem and propose regularizations to concentrate the densities at some **known depth**, while BDC achieves this with **unknown depth**.

3.4. Warm-up Training

The overall loss used to optimize NeRFReN is the weighted combination of photometric loss and the proposed geometric priors:

$$\mathcal{L} = \mathcal{L}_{pm} + \lambda_d \mathcal{L}_d + \lambda_{bdc} \mathcal{L}_{bdc} \quad (11)$$

The magnitude of the geometric priors can greatly influence the decomposition results, requiring careful tuning of the weighting factors which can hardly be shared across scenes. It is difficult to get a balance between the photometric loss and the geometric constraints especially at the early stage of training. If the geometric constraints dominate, the model will stuck at a bad local minima by using only one of the components to explain the whole scene while leaving the other empty. On the other hand, if the photometric loss dominates, insufficient geometric regularizations will lead to sub-optimal decomposition results.

Based on the above observations, we propose a universal training strategy to avoid the need for tuning hyper-parameters for each scene which effectively stabilizes the training process. We take inspirations from learning rate

warm-up and propose to “warm-up” the geometric constraints. λ_d and λ_{bdc} are initialized with small values, and are first increased then decreased as the training progresses. We also mask out the input viewing direction at early stage of training to avoid “leakage” of the reflected component into the transmitted one. We refer to the supplementary material for more details.

3.5. Interactive Setting

The proposed geometric priors and training strategy work well when the transmitted geometry can be correctly estimated from the images. However, there exist more challenging cases, such as texture-less reflectors like mirrors, where the unsupervised decomposition often fails. Thanks to the additive formulation in Eqs. (1) and (7), we can utilize extra information such as manually labeled reflection fraction maps.

We exploit an interactive setting where the user provides binary masks of a small number of training images, with 1 and 0 denoting reflective and non-reflective regions respectively. L1 loss is used to encourage consistency between the predicted reflection fraction map β and the user-provided masks:

$$\mathcal{L}_\beta = \sum_p \|\widehat{\beta}(p) - \beta(p)\|_1 \quad (12)$$

where $\widehat{\beta}(p) = \beta(\mathbf{r}(p); \sigma^t, \alpha)$ is the estimation of the reflection fraction value, and $\beta(p)$ is the value of the user-provided binary mask at pixel p . With the help of this extra supervision, we are able to successfully isolate the reflected component in several challenging scenarios as demonstrated in Fig. 4. These scenarios are hard to deal with using existing novel view synthesis techniques.

4. Experiments

In this section, we first show the scene decomposition results in Sec. 4.1. Then, we compare with NeRF [13] and its naïve variant to show that NeRFReN achieves comparable visual quality with significantly better depth estimation results in Sec. 4.2. We validate some of our design choices in Sec. 4.3, and finally provide applications for reflection removal and scene editing in Sec. 4.4.

Baselines. We compare NeRFReN with NeRF and a variant of NeRF which we call NeRF-D. NeRF-D applies our proposed depth smoothness constraint on the original NeRF, which is expected to model the reflected image solely by the surface point itself rather than all the points behind it along this viewing direction.

Datasets. Due to the lack of real world datasets with strong reflections for the view synthesis task, we introduce RFFR (Real Forward-Facing with Reflections) dataset, containing 6 captured forward-facing scenes with strong reflection effects caused by glass and mirrors. We split the images into

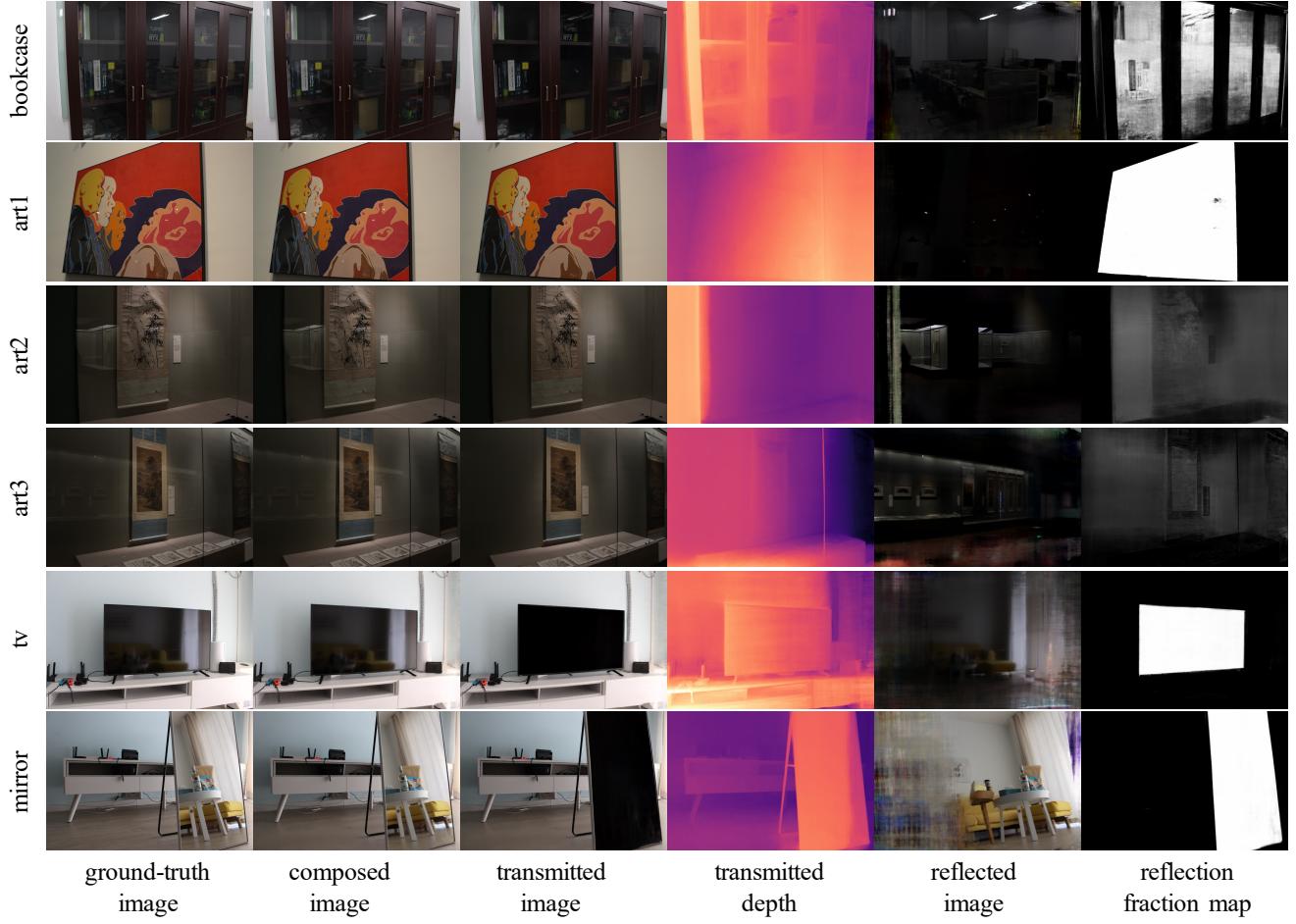


Figure 4. Decomposition results of NeRFReN on RFFR scenes. The decomposed images and depth estimations are highly consistent with human perception.

training and test set, and report the qualitative and quantitative results on test images.

4.1. Decomposition

In Fig. 4 we show how NeRFReN decomposes the scene and performs view synthesis on our RFFR scenes. The transmitted image, reflected image, and reflection fraction map are rendered from the predicted transmitted and reflected neural radiance fields. The composed image is generated by combining the transmitted image and the reflected image weighted by the reflection fraction map according to Eq. (1). The depth values are computed by Eq. (5). For *tv* and *mirror*, 1 and 4 manually-annotated reflection masks are utilized respectively as extra information, as described in Sec. 3.5. For rest of the scenes, the decomposition is inferred from the training images alone.

We can see from Fig. 4 that NeRFReN can perform transmission-reflection decomposition that is consistent with human perception while producing high-quality view synthesis results. The learned reflection fraction map roughly indicates the surface material, with large values

(white) for reflectors and small values (black) for non-reflectors. Note that we only intend to model the stable virtual image by the reflected field, and leave the low-frequency highlights to the view-dependency in the transmitted field. For example, there are highlights on the frames of the bookcase, where the reflection fraction is almost zero.

Accurately estimating the location of transparent surfaces like glass is difficult for lack of traceable image features. In such cases like *art2* and *art3*, NeRFReN treats the opaque surface behind as the reflective surface, which is not strictly correct but does not affect view synthesis quality due to the image-space composition.

4.2. Comparisons

We compare NeRFReN with NeRF and NeRF+D on the test set of our RFFR dataset under three metrics: PSNR, SSIM [28] and LPIPS [34] (see Tab. 1). We also provide qualitative comparisons in Fig. 5.

Quantitatively, NeRFReN significantly outperforms NeRF on *art1* and *mirror* in terms of PSNR. While performing slightly better on rest of the scenes, NeRF produces

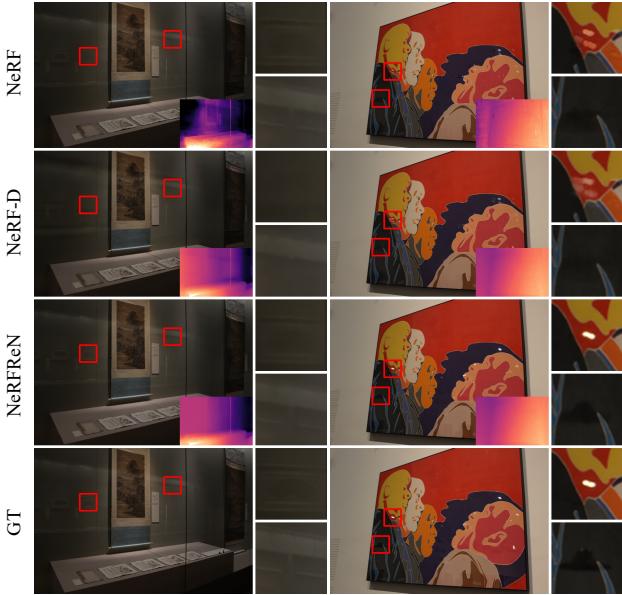


Figure 5. Visual comparisons between NeRF, NeRF-D and NeRFReN. NeRFReN achieves comparable or better novel view synthesis quality with physically correct depth estimation. Zoom in for more details.

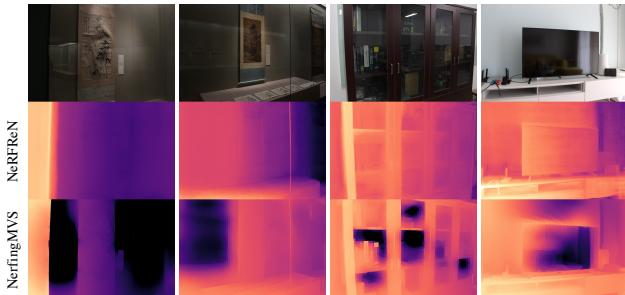


Figure 6. Qualitative comparisons with NerfingMVS [29] on depth estimation. NerfingMVS fails in such scenes with severe reflections, behaving similar to NeRF.

inaccurate depth estimation due to the mixed geometry of the transmitted and reflected components as demonstrated in Sec. 3.1, which is shown on the bottom left corner of the NeRF rendering results in Fig. 5. NeRFReN models the scene with separate neural radiance fields, and provides physically correct depth map by taking the depth estimation of the transmitted field. Also, NeRF generates sub-optimal renderings when the multi-view consistency is violated while NeRFReN does not suffer from such restrictions, as shown in the *mirror* case in Fig. 1(b). NeRF-D produces more correct depth estimation than NeRF by adopting the depth smoothness regularization, and condition the surface color on viewing direction. However, it synthesizes blurry results due to the insufficient representation power of NeRF’s view-dependency in modeling complex reflections. The quantitative results conform with this claim that NeRF-D consistently underperforms NeRF and NeRFReN.

To demonstrate the depth estimation quality of NeR-

Method	PSNR↑					
	art1	art2	art3	bookcase	tv	mirror
NeRF	36.48	42.63	41.18	29.50	33.47	32.22
NeRF-D	36.49	41.58	39.80	29.78	33.02	32.33
NeRFReN	39.49	42.33	41.03	30.18	33.24	33.30
Method	SSIM↑					
	art1	art2	art3	bookcase	tv	mirror
NeRF	0.975	0.974	0.973	0.883	0.960	0.944
NeRF-D	0.971	0.968	0.965	0.876	0.954	0.939
NeRFReN	0.978	0.972	0.970	0.890	0.955	0.947
Method	LPIPS↓					
	art1	art2	art3	bookcase	tv	mirror
NeRF	0.029	0.059	0.053	0.078	0.054	0.066
NeRF-D	0.027	0.070	0.117	0.136	0.063	0.072
NeRFReN	0.029	0.067	0.072	0.100	0.061	0.071

Table 1. View synthesis results of NeRF, NeRF-D and NeRFReN on RFFR scenes in PSNR, SSIM [28] and LPIPS [34]. NeRFReN performs comparable with NeRF and better than NeRF-D.

FReN, we also qualitatively compare with NerfingMVS [29] on RFFR scenes. Results are shown in Fig. 6. NerfingMVS utilizes monocular depth estimation priors and post-filtering to achieve promising results on ScanNet scenes, but struggles on scenes with complex reflections. This is because NerfingMVS initializes the depth maps based on Multi-View-Stereo output, which cannot give faithful depth estimations in such cases.

To conclude, NeRFReN achieves visually pleasing view synthesis results and physically correct depth estimation results at the same time. The formulation of NeRFReN naturally relax the multi-view consistency of NeRF by combining the transmitted and reflected part in the image domain, which greatly facilitate novel view synthesis in challenging scenes like those with mirrors.

4.3. Ablation Studies

We discuss the necessity of several key design choices, and demonstrate their importance to getting faithful decompositions and correct view synthesis results.

Geometric priors. We train NeRFReN without any geometric priors and with only one of the priors, to see the impact of these priors on the decomposition quality. The results are shown in Fig. 7. Without any priors (row 1), we find the optimization process to be highly under-controlled, largely depending on the network initialization. In this case, the decomposition results are not reasonable since the transmitted and reflected part are mixed together. Without the depth smoothness prior (row 2), there is not enough constraint on the transmitted geometry, letting it contain reflected objects and producing wrong depth. Without the bidirectional depth consistency prior (row 3), there will be redundant contents left in the reflected component (the bookcase frame). The BDC loss encourages uni-modal peak distribution for densities along a ray, which effectively

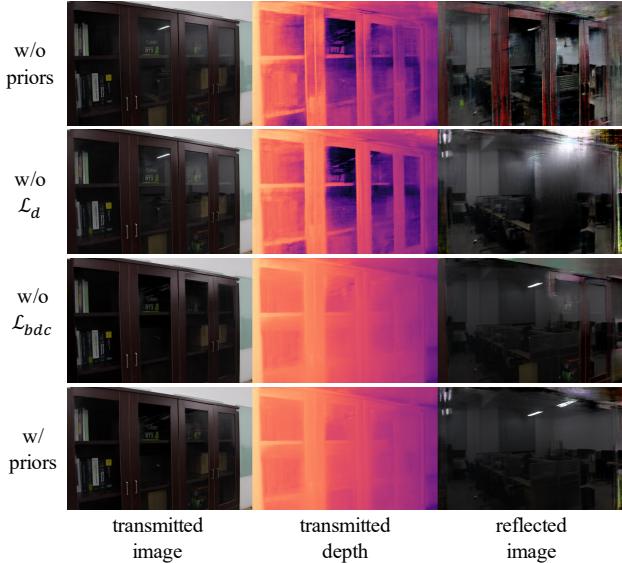


Figure 7. Effectiveness of the proposed geometric priors. Both priors are essential to get physically sound decompositions.

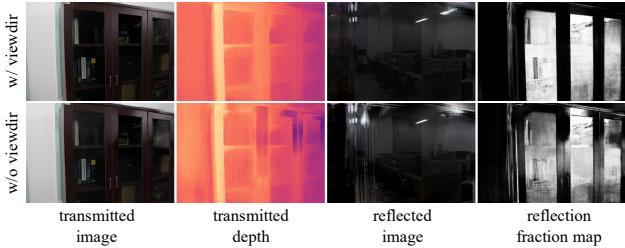


Figure 8. The effectiveness of conditioning transmitted radiance on viewing direction. Without this dependency, low-frequency reflections like highlights are modeled with foggy contents in the reflected component, leading to inaccurate depth estimation.

removes occlusions to get a “clean” reflected image. We claim that both priors are essential to getting reasonable solutions for this under-constrained problem (row 4).

View-dependency. NeRFReN conditions the color of the transmitted component on viewing direction to model low-frequency view-dependent effects. We remove this dependency on the *bookcase* scene and show the results in Fig. 8. Without the view-dependency, highlights on the bookcase frame are modeled with the combined effect of the reflected image and the reflection fraction map. This results in inaccurate depth estimation (dark regions on the predicted depth map), and redundancies in the reflected component which do not belong to the virtual geometry.

4.4. Applications

Enabling user editing for neural scene representations has been attracting attention in the community [2, 6, 9, 26, 32, 35]. Existing techniques either achieve material editing or relighting by reflectance decomposition [2, 26, 35] or



Figure 9. Examples of replacing the reflected image (with the room scene in [13]). This could be utilized to synthesize self-reflections of the user for higher realism.

model the scene at object level for manipulation [6, 9, 32]. Benefit from the formulation in Eq. (1), NeRFReN enables scene editing in the image domain while maintaining view consistency, which is more intuitive and makes it possible to work with other scene representations.

Reflection removal. Our formulation naturally supports the application for reflection removal by taking only the transmitted image. See column 3 of Fig. 4.

Reflection substitution. We achieve reflection substitution by replacing the reflection image I_r by images coming from other neural radiance fields, or even from other scene representations like mesh. Fig. 9 shows two examples of replacing the reflections with images rendered from another NeRF model trained on the *room* scene [13]. This could be further promoted to synthesize self-reflections of the user to provide even more immersive experiences. We hope the image-based formulation of NeRFReN can inspire future research about combining multiple scene representations.

5. Limitations and Conclusion

Some major limitations of our method include modeling curved reflective surfaces and multiple non-coplanar reflective surfaces. Also, we do not consider the view-dependent effects of reflected objects, and ignore the view-dependency of reflection fraction values caused by Fresnel effect.

To conclude, we propose NeRFReN to enhance NeRF in scenes with reflections. We decompose the scene into transmitted and reflected components, model them with separate neural radiance fields, and synthesize novel views by weighted combination. We exploit geometric priors and special training strategies to encourage proper decomposition, where user-provided reflection masks can be utilized in challenging cases like mirrors. Our method enables high-quality view synthesis along with physically sound decomposition results. Finally, potential applications including depth estimation, reflection removal and reflection substitution are investigated to encourage further research about scene understanding and neural editing.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Project Number 62132012), Research Grant of Beijing Higher Institution Engineering Research Center, and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

- [1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. [2](#)
- [2] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. [3, 8](#)
- [3] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. [2](#)
- [4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. [2](#)
- [5] John Flynn, Michael Broxton, Paul Debevec, Matthew Du-Vall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. [2](#)
- [6] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. [8](#)
- [7] Peter Hedman, Tobias Ritschel, George Drettakis, and Gabriel Brostow. Scalable inside-out image-based rendering. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. [2](#)
- [8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [11](#)
- [9] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. *arXiv preprint arXiv:2105.06466*, 2021. [8](#)
- [10] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. [2](#)
- [11] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. [3](#)
- [12] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. [2](#)
- [13] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. [1, 2, 3, 5, 8, 11](#)
- [14] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deep sdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. [2](#)
- [15] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. *arXiv preprint arXiv:2111.14643*, 2021. [5](#)
- [16] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020. [2](#)
- [17] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021. [2](#)
- [18] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *arXiv preprint arXiv:2112.03288*, 2021. [5](#)
- [19] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [2](#)
- [20] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. [2](#)
- [21] Sudipta N Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. [3, 4](#)
- [22] Sudipta N Sinha, Drew Steedly, and Richard Szeliski. Piecewise planar stereo for image-based rendering. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1881–1888. IEEE, 2009. [4](#)
- [23] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. [2](#)
- [24] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019. [2](#)
- [25] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. [2](#)
- [26] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. [8](#)

- [27] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. [2](#)
- [28] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. [6](#), [7](#)
- [29] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. [7](#)
- [30] Suttisak Wizadwongsu, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. [2](#)
- [31] Jiamin Xu, Xiuchao Wu, Zihan Zhu, Qixing Huang, Yin Yang, Hujun Bao, and Weiwei Xu. Scalable image-based indoor scene rendering with reflections. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. [3](#)
- [32] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. [8](#)
- [33] Chi Zhang, Zhiwei Li, Rui Cai, Hongyang Chao, and Yong Rui. As-rigid-as-possible stereo under second order smoothness priors. In *European Conference on Computer Vision*, pages 112–126. Springer, 2014. [4](#)
- [34] 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. [6](#), [7](#)
- [35] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970*, 2021. [3](#), [8](#)

Supplementary Material

In this supplementary material, we provide additional details for the network and training procedure (Sec. A.1), further discussions on limitations along with an example of failure case (Sec. A.2), results and comparisons on LLFF Dataset [13] (Sec. A.3), and more ablation studies with qualitative demonstrations (Sec. A.4). We also provide video results on the [project webpage](#).

A.1. Network and Training

A.1.1. Network Configurations

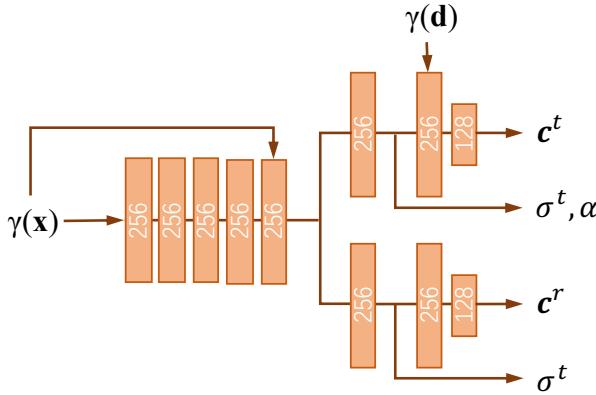


Figure A1. Detailed network architecture of NeRFReN.

The detailed network architecture of NeRFReN is shown in Fig. A1. Orange blocks are fully-connected layers. Each layer is followed by ReLU activation except for the output layers. $\gamma(\mathbf{x})$ and $\gamma(\mathbf{d})$ are positional encodings of the input coordinate \mathbf{x} and viewing direction \mathbf{d} . The network is designed to have approximately same amount of parameters with the original NeRF [13] network.

A.1.2. Warm-Up Training

An illustration of how λ_d and λ_{bdc} change with the training process is shown in Fig. A2. Note that we first increase λ_d to ensure correct geometry for the transmitted component and then increase λ_{bdc} to remove redundancies in the reflected component. We find this stabilize training compared to optimizing both at the same time. Then we gradually decrease the weights to concern more on photometric loss to get more accurate renderings. Effects of warm-up training are demonstrated in Sec. A.4.1.

A.1.3. Training Details

We sample 4×4 patches instead of individual pixels to compute the depth smoothness constraint, which requires depth value of pixels in a local area. λ_d and λ_{bdc} are set to 0.01 and 1e-4 at the beginning of training. λ_d is increased to 0.1 at iteration 1k, and decreased to 0.01 at iteration 5k.

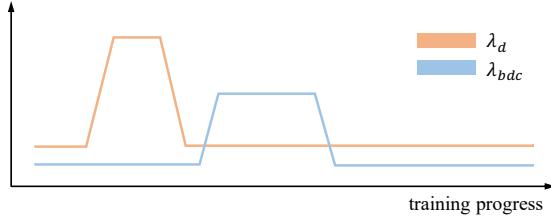


Figure A2. Illustration for the proposed warm-up training. Weightings for geometric priors are first increased then decreased.

λ_{bdc} is increased to 0.05 at epoch 20, decreased to 1e-4 at epoch 12 and further decreased to 0 at epoch 15. We mask out the input viewing direction before epoch 10. All models are trained for 40 epochs using Adam [8] optimizer with an initial learning rate of 5e-4. The learning rate is exponentially decayed to 5e-6 in the last 20 epochs.

To further stabilize training, we force the transmitted image to be close to the input image for the first 1k iterations:

$$\mathcal{L}_{init} = \|\hat{\mathbf{C}}(\mathbf{r}; \sigma^t, \mathbf{c}^t) - \hat{\mathbf{C}}\|_2 \quad (\text{A.1})$$

We also apply a smoothness constraint on the reflection fraction map:

$$\mathcal{L}_\beta = \sum_p \sum_{q \in \mathcal{N}(p)} \|\beta(p) - \beta(q)\|_1, \quad (\text{A.2})$$

λ_{init} and λ_β are set to 0.01 and 1e-4 respectively.

Same as the original NeRF, we simultaneously optimize a coarse and a fine network. When training the coarse network, the transmitted and reflected field share the same set of samples. For the fine network, we get two different sets of fine samples respectively for the transmitted and reflected field based on the weights of their coarse samples, since the transmitted and reflected component have independent geometries. A side-effect would be that we have to evaluate the fine network twice for a query point. This computation overhead is carefully taken care of in the comparisons with the original NeRF: 64 fine samples are evaluated for NeRFReN and 128 fine samples for NeRF.

A.2. Limitations and Failure Case

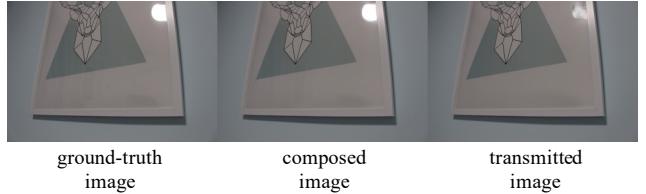


Figure A3. Illustration for a failure case where the virtual image is not stable due to the curved reflective surface. Zoom in for more details.

Curved reflective surfaces that do not produce stable virtual images cannot be modeled by our image-based formu-

Method	PSNR↑								
	fern	flower	fortress	horns	leaves	orchids	room	trex	average
NeRF	26.01	31.70	33.23	32.92	22.98	21.38	37.04	32.03	29.66
NeRFReN	25.36	29.95	33.60	31.59	22.75	21.89	37.52	30.64	29.16

Table A1. View synthesis results of NeRF and NeRFReN on LLFF dataset.

lation. An example can be seen in Fig. A3. Note the distorted light and curtain in the reflections of the ground-truth image. In this case, the reflected component does not have a consistent geometry, leading to inaccurate decomposition and rendering results.

Another limitation lies in the modeling of multiple non-coplanar surfaces. However, we find that in real life, reflection images from different non-coplanar surfaces rarely coincide because they are often far away from each other or only observed from limited angles. This makes it possible to model them by a single reflected field as we do in the paper. Potential failure cases could be alleviated by utilizing multiple reflected MLPs. An interesting direction for addressing the above limitations would be to model reflected rays as in ray tracing, which we regard as a future work.

A.3. Results on LLFF Dataset

We experiment on LLFF dataset, where most of the scenes do not exhibit strong reflections. As is shown in Tab. A1, NeRFReN achieves a competitive average PSNR (29.16) compared to NeRF (29.66). This demonstrates that our method maintains high representational ability despite of all the specifically-designed priors in training. Some qualitative results are provided in Fig. A4. In Fig. A5 we show the decomposition results and improved depth prediction on the *room* scene, where we make use of manually-annotated masks on the TV screen for 2 of the training images. No meaningful decomposition is achieved in other LLFF scenes.

A.4. Ablation Studies

A.4.1. Warm-up Training

We exploit two alternative training strategies to demonstrate the effectiveness of warm-up training: (1) training with strong geometric constraints ($\mathcal{L}_d = 0.1$, $\mathcal{L}_{bdc} = 0.05$) directly without warm-up; (2) training with viewing directions directly from the very beginning without masking. The qualitative results are shown in Fig. A6. For the first setting, the transmitted depth can be overly smoothed, and the reflected component quickly converges to empty due to the strong constraints without proper initialization. For the second setting, the reflected image is explained by view-dependency, leading to blurry reflections lacking fine details. Training with the warm-up strategy does not exhibit

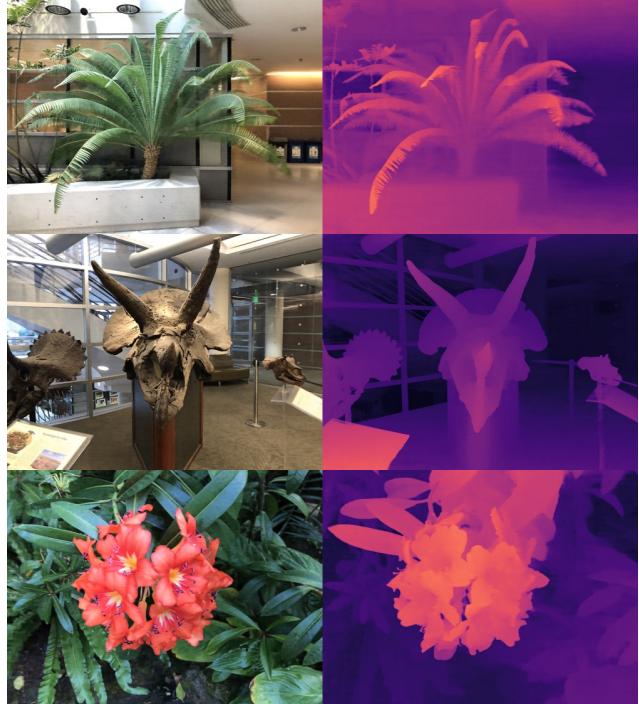


Figure A4. Novel view synthesis and depth estimation results of NeRFReN on some of the LLFF scenes.

such problems, as shown in the first row.

A.4.2. Interactive Setting

We use user-provided reflection masks to guide the decomposition for the *mirror* and *tv* scene. Fig. A7 shows the effects of different numbers of masks to the decomposition results. Without masks, the network finds it hard to distinguish between the transmitted and reflected geometry. For the *mirror* scene, fair decomposition results can be achieved by utilizing 4 masks. And only 1 mask is needed for the *tv* scene. This demonstrates that our method can deal with hard cases with only minimum user inputs.

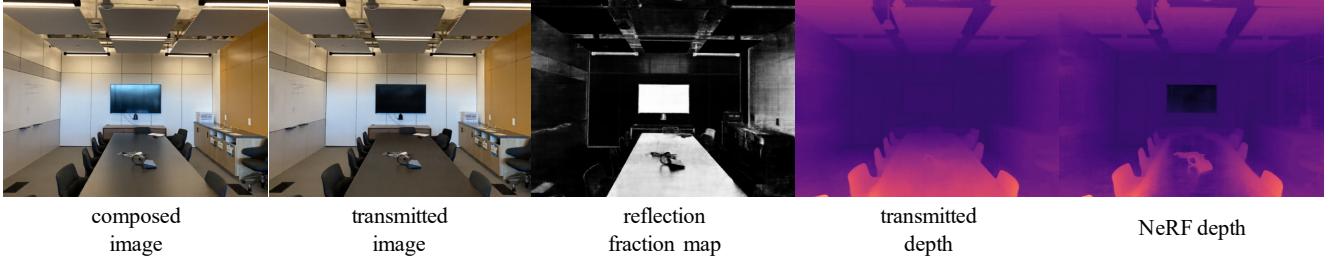


Figure A5. Decomposition results of NeRFReN on the *room* scene with 2 manually annotated reflection masks on the screen. NeRFReN provides reasonable decomposition along with high-quality depth estimation results.

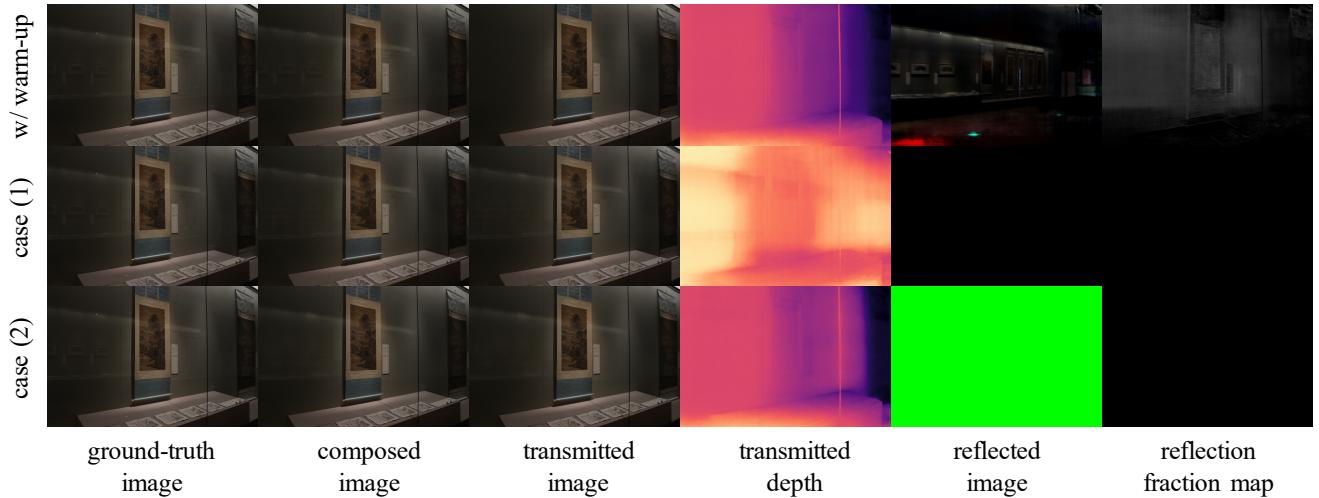


Figure A6. The effects of the proposed warm-up training strategy. The decomposition is likely to fail if strong geometric constraints (row 2) or view-dependency (row 3) are introduced from the beginning of training. We manage to get faithful decomposition by warming-up the weighting factors of the geometric constraints and masking out the viewing direction in early training stage (row 1).

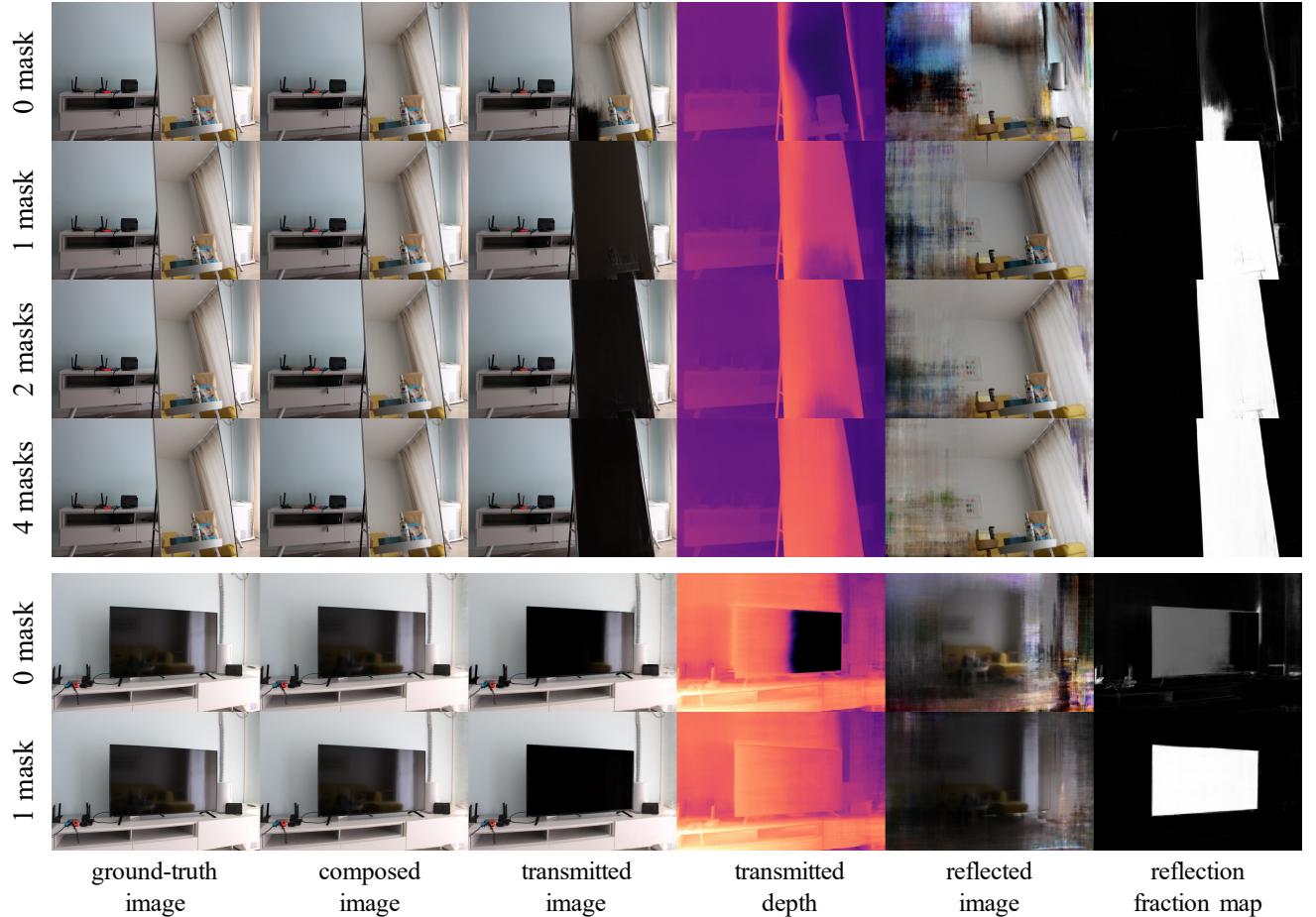


Figure A7. The effects of the user-provided reflection masks on challenging scenes. Without any masks, NeRFReN fails to distinguish between the transmitted and reflected geometry. We utilize 4 masks for the *mirror* scene and only 1 mask for the *tv* scene to get faithful decompositions.