Supplementary Material for NeRFReN: Neural Radiance Fields with Reflections

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

1. Network and Training

1.1. Network Configurations

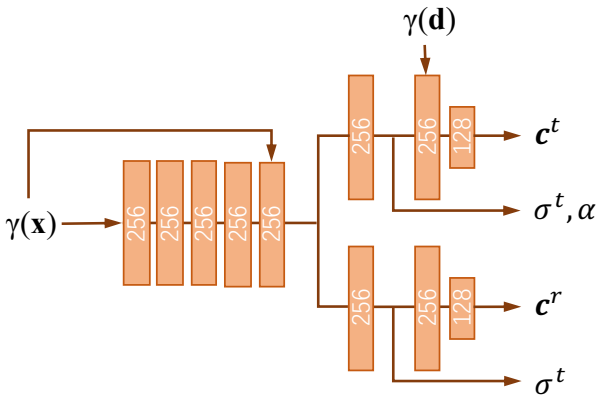

Figure 1. Detailed network architecture of NeRFReN.

The detailed network architecture of NeRFReN is shown in Fig. 1. 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 [2] network.

1.2. Warm-Up Training

An illustration of how λ_d and λ_{bdc} change with the training process is shown in Fig. 2. 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. 4.1.

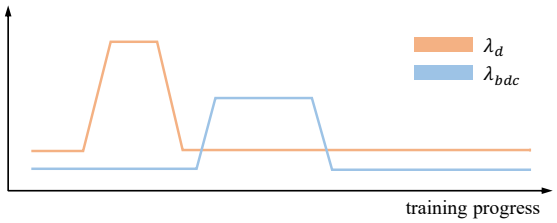

Figure 2. Illustration for the proposed warm-up training. Weightings for geometric priors are first increased then decreased.

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. 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 [1] 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:

$$\mathcal{L}_{init} = ||\widehat{\mathbf{C}}(\mathbf{r}; \sigma^t, \mathbf{c}^t) - \widehat{C}||_2 \tag{1}$$

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

$$\mathcal{L}_{\beta} = \sum_{p} \sum_{q \in \mathcal{N}(p)} ||\beta(p) - \beta(q)||_1, \tag{2}$$

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

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 1. View synthesis results of NeRF and NeRFReN on LLFF dataset.

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.

2. Limitations and Failure Case

Figure 3. 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 formulation. An example can be seen in Fig. 3. 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.

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. 1, 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. 4. In Fig. 5 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.

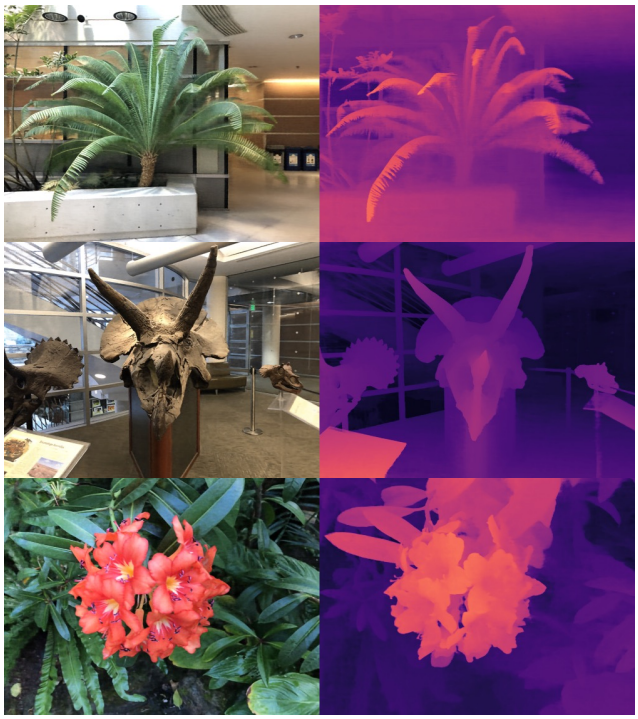

Figure 4. Novel view synthesis and depth estimation results of NeRFReN on some of the LLFF scenes.

4. Ablation Studies

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. 6. 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 de-

Figure 5. 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.

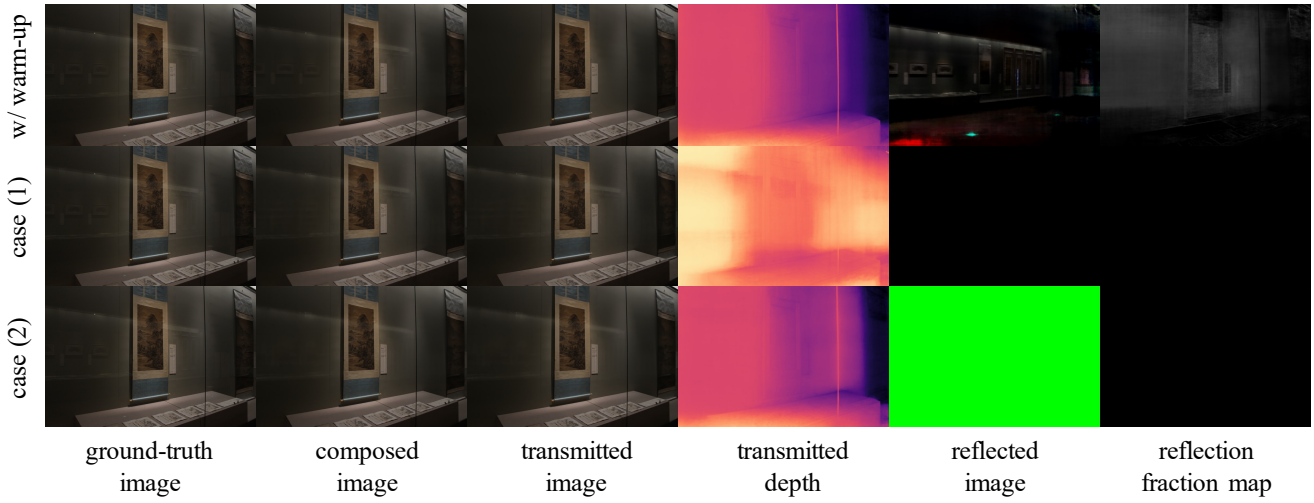

Figure 6. The effects of the proposed warm-up training strategy. The decomposition is likely to fail if strong geometric constraints (row 2) or view-dependency (row 3) are introduced from the beginning of training. We manage to get faithful decomposition by warming-up the weighting factors of the geometric constraints and masking out the viewing direction in early training stage (row 1).

tails. Training with the warm-up strategy does not exhibit such problems, as shown in the first row.

4.2. Interactive Setting

We use user-provided reflection masks to guide the decomposition for the *mirror* and *tv* scene. Fig. 7 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.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

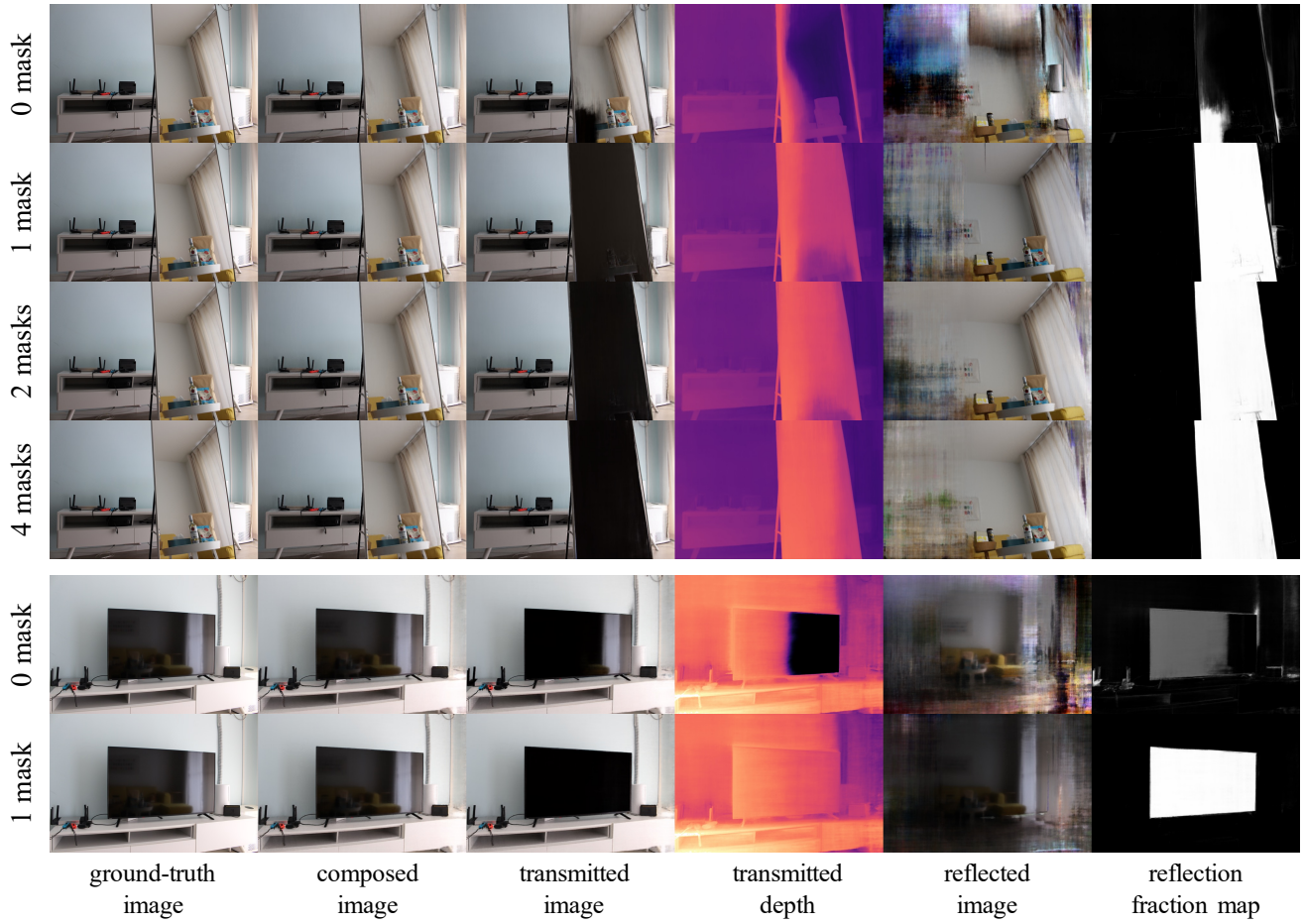

Figure 7. The effects of the user-provided reflection masks on challenging scenes. Without any masks, NeRFReN fails to distinguish between the transmitted and reflected geometry. We utilize 4 masks for the *mirror* scene and only 1 mask for the *tv* scene to get faithful decompositions.