# Mirror-NeRF: Learning Neural Radiance Fields for Mirrors with Whitted-Style Ray Tracing Supplementary Material

Junyi Zeng[*]
Zhejiang University
Hangzhou, China
zengjunyi@zju.edu.cn

Chong Bao[*]
Zhejiang University
Hangzhou, China
chongbao@zju.edu.cn

Rui Chen
Zhejiang University
Hangzhou, China
22221111@zju.edu.cn

Zilong Dong
Alibaba Group
Hangzhou, China
list.dzl@alibaba-inc.com

Guofeng Zhang
Zhejiang University
Hangzhou, China
zhangguofeng@zju.edu.cn

Hujun Bao
Zhejiang University
Hangzhou, China
bao@cad.zju.edu.cn

Zhaopeng Cui[†]
Zhejiang University
Hangzhou, China
zhpcui@zju.edu.cn

This supplementary material provides additional details of our proposed method, including model details in Section A, and experiment setups in Section B. Moreover, we present additional experimental results and analysis in Section C. Dynamic qualitative results can be viewed in our supplementary video.

## A MODEL DETAILS

We elaborate on the detailed architecture of our unified neural field as introduced in Sec. 3.1. The unified neural field is a fully-connected neural network and consists of five parts, $\mathcal{F}_{geo}, \mathcal{F}_\sigma, \mathcal{F}_c, \mathcal{F}_n, \mathcal{F}_m$. $\mathcal{F}_{geo}$ has 8 layers with 256 hidden dimensions and ReLU activation. $\mathcal{F}_{geo}$ uses positional encoding [8] with 10 frequencies on spatial location input $\mathbf{x}$, and outputs geometry feature $f_{geo}$ with a dimension of 256. $\mathcal{F}_\sigma$ is a volume density head MLP with 1 layer which takes $f_{geo}$ as input and outputs volume density $\hat{\sigma}$. $\mathcal{F}_c$ is a radiance head MLP with 2 layers and 128 hidden dimensions which takes $f_{geo}$ and positional encoded view direction $\mathbf{d}$ (with 4 frequencies) as input and outputs the RGB radiance $\hat{c}$. $\mathcal{F}_n$ is a normal head MLP with 2 layers and 128 hidden dimensions which takes $f_{geo}$ as input and outputs the smooth surface normal $\hat{n}$. $\mathcal{F}_m$ is a reflection probability head MLP with 2 layers and 128 hidden dimensions, which takes $f_{geo}$ as input and outputs the reflection probability $\hat{m}$ normalized by the Sigmoid activation. During rendering, we take the hierarchical sampling with a coarse neural field and a fine neural field. We first uniformly sample 64 coarse points and then sample 128 fine points by importance sampling along each ray.

## B EXPERIMENT SETUPS

### B.1 Data Preparation

We evaluate Mirror-NeRF on synthetic and real datasets. Synthetic datasets contain five synthetic indoor rooms downloaded from BlenderSwap [4]. Images are captured 360 degrees around the scene with a horizontal circular camera trajectory located at the center of the room and looking towards the center, which is established by utilizing the Bezier curve. We use the Cycle engine in Blender to render high-fidelity images with a resolution of 400x400. For real datasets, we use 3D Scanner App [1] in IPad Pro 5th generation to capture images and camera poses in the clothing store, lounge, market and discussion room. The images are taken along a horizontal round trajectory in front of the mirror, with 480x360 pixels per image.

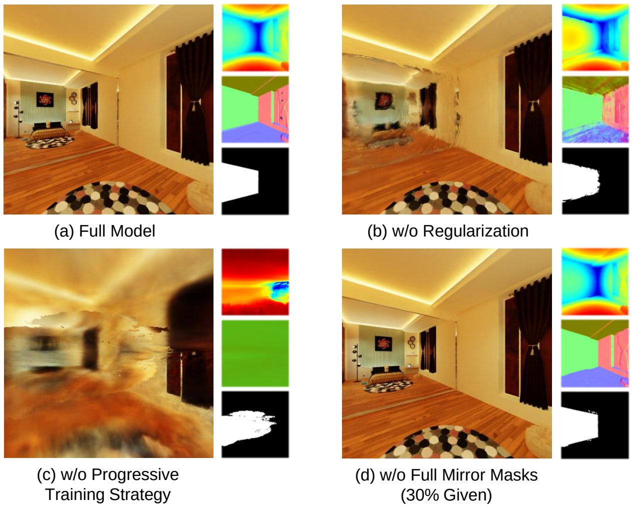
(a) Full Model
(b) w/o Regularization
(c) w/o Progressive Training Strategy
(d) w/o Full Mirror Masks (30% Given)

**Figure A: More ablation studies. In each subplot, the image on the left is the novel view and the images on the right are depth, surface normal, and reflection mask from top to bottom respectively.**

### B.2 Comparison Details

All experiments are performed on NVIDIA RTX 3090 GPU (24GB). The scene coordinates are scaled to be within [-1,1]. We train Ref-NeRF [9] with the default setting, except for tuning the far bound of ray sampling for different scenes to avoid the result collapsing and reducing the batch size to 2048 to avoid the "out of memory" error in a single GPU with decreasing the learning rate by the scale factor that batch size is reduced by. NeRFReN [2] uses two radiance fields to learn the reflected part and transmitted part of the scene separately. In cases where geometric constraints are not dominant, NeRFReN is easy to get stuck in a local optimum by exclusively

[*] Junyi Zeng and Chong Bao contributed equally to this research. The authors from Zhejiang University are also affiliated with the State Key Lab of CAD&CG.
[†] Corresponding author: Zhaopeng Cui.

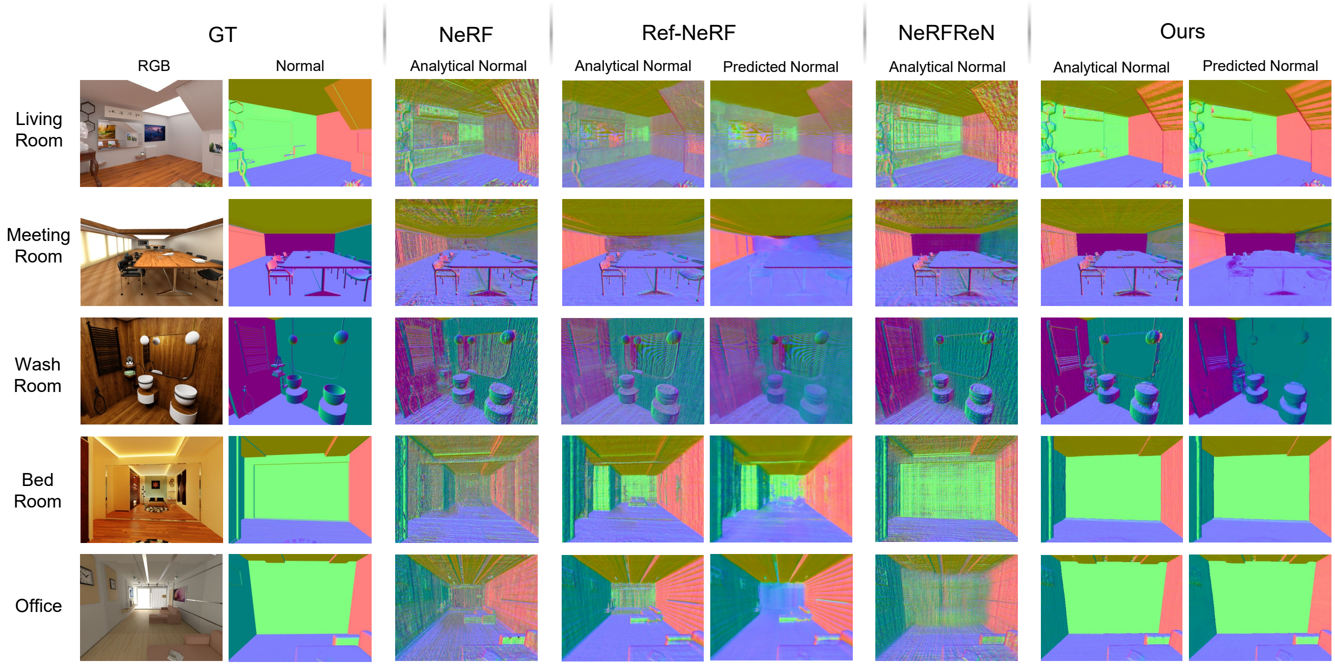

**Figure B: Qualitative comparison of estimated surface normal maps on synthetic scenes with mirrors.**

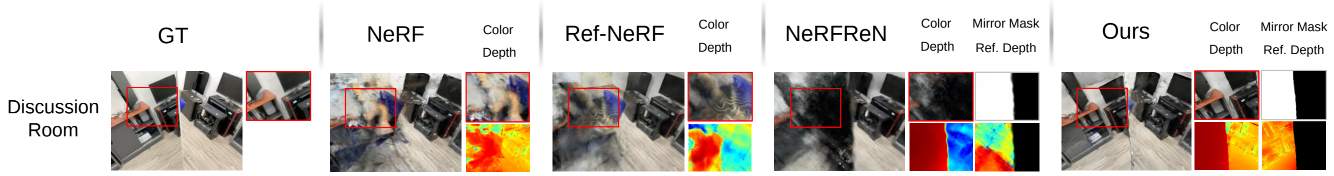

**Figure C: Qualitative comparison of novel view synthesis from challenging novel viewpoints out of the training set distribution on the real-world discussion room.**

utilizing the radiance field of a single part to learn the entire scene while leaving the other part degenerate. Thus, we manually tune its weighting coefficients of mask loss and depth smoothness loss in each dataset which ranges in [1.0, 5.0].

Instead, our method takes a progressive training strategy in Sec. 3.4 to stabilize the geometry optimization of the mirror. Specifically, In the initial stage of training, we enable $\lambda_c$ with $\mathcal{L}_{cm}$ and disable the other weight factors. After 2 epochs, we enable geometry constraints $\lambda_n, \lambda_{n_{reg}}, \lambda_m, \lambda_{pc}$. In the 5th epoch, we replace $\mathcal{L}_{cm}$ with $\mathcal{L}_c$ until the end of training.

When comparing reflection in the mirror from the novel viewpoints out of the training set distribution in Tab. 2, we capture scenes from a group of challenging test camera poses. These viewpoints are closer to the mirror than training viewpoints and can see the reflection in the mirror that is unobserved from the training viewpoints. For synthetic scenes, we use the Cycle engine to render the images and the mirror reflection masks from the test viewpoints in Blender. To assess the PSNR of the reflection synthesized by each method, we use the ground-truth mirror reflection mask to filter the pixels outside the mirror and only evaluate PSNR for pixels inside the mirror. Since evaluating SSIM [10] and LPIPS [11] requires a complete 2D image, we use ground-truth mirror reflection mask to zero out the pixels outside the mirror for both rendered images and ground-truth images before evaluating their SSIM and LPIPS.

# C MORE EXPERIMENTS

## C.1 Ablation Study on Regularization

To ablate the naïve training without regularization, we turn off all regularization terms $\lambda_{pc}, \lambda_{n_{reg}}$, and joint optimization described in Sec. 3.3. As demonstrated in Tab. A and Fig. A(b), the geometry of the mirror is broken due to the underconstrained density field. The "foggy" geometry of the mirror prevents us from synthesizing the precise color and reflection probability of the mirror. Instead, with the proposed regularization, we can obtain the smooth depth and normal of the mirror to improve the rendering quality of the reflection.

## C.2 Ablation Study on Progressive Training Strategy

When ablating the progressive training strategy described in Sec. 3.4, we train the whole model by enabling all weight factors in Eq. (16)

| Settings | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| w/o Regularization | 25.588 | 0.812 | 0.182 |
| w/o Progressive Training Strategy | 18.523 | 0.569 | 0.585 |
| w/o Full Mirror Masks (30% Given) | 32.305 | 0.927 | 0.052 |
| Full Model | 32.422 | 0.933 | 0.047 |

**Table A: More ablation studies on the synthetic bedroom.**

| Methods | MAE° ↓ |
|---|---|
| NeRF | 38.555 |
| Ref-NeRF | 31.930 |
| Ref-NeRF (predict) | 31.778 |
| NeRFReN | 22.717 |
| Ours | 8.387 |
| Ours (predict) | 7.422 |

**Table B: Quantitative comparison of mean angular error (MAE) on five synthetic scenes with mirrors. Here "predict" denotes that the predicted normal from MLP is used in the comparison, otherwise we use analytical surface normal from volume density to compute mean angle error. The best is marked in red and the second best is marked in orange.**

from scratch without using the masked photometric loss $\mathcal{L}_{cm}$. As shown in Tab. A and Fig. A(c), the color, depth, normal and reflection probability degenerate due to the conflict between the color supervision of the mirror and the geometry regularization in the early stage of training. When the geometry of the mirror has not converged yet, the color supervision and the geometry regularization confuse the optimization direction of training, *i.e.*, to learn a virtual scene in the mirror or to learn a planar surface at the mirror. With the progressive training strategy, we can easily avoid the degeneration situation by first learning the geometry of the mirror and then modeling the reflection in the mirror.

## C.3 Ablation Study on Mirror Reflection Masks

We support training with some mirror reflection masks not given, while the quality has no significant degradation. To analyze the impact of training with the partial mirror reflection masks, we segment out the mirror in 30% of the images with the off-the-shelf segmentation tool [3] to obtain the mirror reflection masks. In the first two stages of training, the model is supervised by the images with the reflection masks to learn the accurate geometry and reflection probability of the mirror. In the last stage, all images are used to supervise the joint optimization of the color inside and outside the mirror. As demonstrated in Tab. A and Fig. A(d), we can still synthesize high-fidelity reflection in the mirror with partial reflection masks, and the rendering quality is comparable with the results from full(100%) reflection masks. This demonstrates the robustness of our method.

## C.4 Comparison of the Surface Normal

We compare the estimated surface normal of our method with NeRF [5], Ref-NeRF [9], NeRFReN [2] on five synthetic datasets, as shown in Fig.B. The ground-truth surface normal maps of novel views are rendered by Blender. We use the mean angular error (MAE) to quantitatively compare the volume-rendered analytical surface normal $\mathbf{N}$ with the ground truth.

$$\mathrm{N}(\boldsymbol{r}) = \sum_{i=1}^{M} T_i \alpha_i \mathbf{n}_i, \tag{1}$$

where $\mathbf{n}$ is derived from Eq. (3). Since Ref-NeRF and our method also predict the smooth surface normal $\hat{\mathbf{n}}$ which is parameterized by an MLP, we also compare the volume-rendered predicted surface normal $\hat{\mathrm{N}}$ in Eq. (6) with the ground truth. The quantitative results are shown in Tab. B. Both the predicted surface normal and the analytical surface normal of our method outperform all the compared methods, which reveals the superiority of our method in learning the geometry of the scene with the mirror. Here MAE of our predicted surface normal is smaller than that of the analytical surface normal, which demonstrates the effectiveness of our smooth surface normal parameterization. The qualitative comparisons are shown in Fig. B. We can produce a smoother surface normal than all the compared methods, which is closer to the ground truth.

## C.5 Analysis on Training and Rendering Efficiency

To evaluate the training and rendering efficiency, we conduct quantitative comparisons between NeRF and our method with the same setting on the "office" dataset. The batch size is 1024 and the chunk size is 8192.

The training time for NeRF with 20 epochs is 8.012 hours, while our method takes 28.539 hours. Our additional computations compared to NeRF during training are mainly caused by training neural networks of normal and reflection probability fields, computing regularization loss terms, ray tracing, and some calculations like computing the gradient of volume density.

As for rendering efficiency, NeRF takes 5.878s and our method takes 8.386s on average to render a 400x400 image of the "office" dataset. Our additional computations compared to NeRF are mainly caused by tracing the reflected rays from the mirror-like surface and extra networks to predict the normal and reflection probability of sampled points. Benefiting from the efficient formulation of Whitted Ray Tracing, only rays hitting mirror-like surfaces are traced with one reflected ray for each, and most rays terminate at diffuse surfaces quickly. The latest acceleration approaches [6, 7] can be also exploited to accelerate both the training and rendering efficiency of our method, which is considered as future work.

## C.6 Handling Multi-Time Reflections

As demonstrated in Eq. (11), our method can handle multi-time reflections by recursive tracing. We have verified this ability to learn multi-time reflections on a new scene with multiple mirrors. The quantitative result is 31.891 on PSNR(↑), 0.883 on SSIM(↑), 0.083 on LPIPS(↓). The qualitative result is shown in Fig. D. Due to the accumulated drift error of multi-time reflections and lack of plane consistency constraint, there is a slight imperfection in the part of the mirror with multi-time reflections. Besides, the result of rendering multi-time reflections is shown in the application of placing new mirrors in Fig. 7(a).

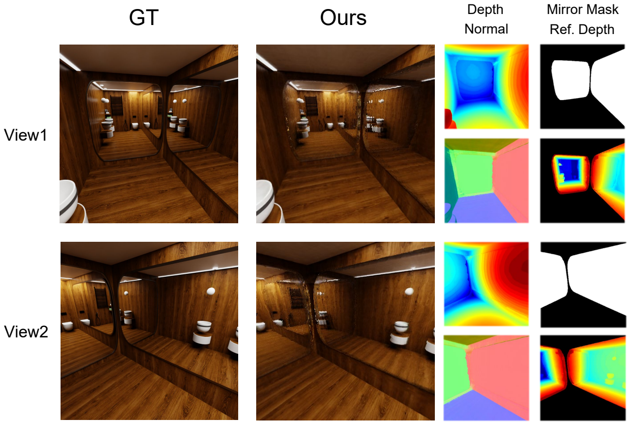

**Figure D: Our method can handle multi-time reflections by recursive tracing.**

## C.7 More Analysis on Real-World Datasets

From Tab. 1 in the main paper, we can see that our method is slightly inferior to NeRF [5] on real-world datasets. This is mainly due to the following two reasons.

First, the camera poses computed for the real-world datasets are not very precise, which makes the geometry slightly inconsistent in multiple views. NeRF can tolerate the view-dependent geometry inconsistency due to the "foggy" density field. However, our method tries to learn an accurate surface of the mirror for ray tracing. The geometry inconsistency will bias the direction of the reflected ray and incur a decrease in quality. The error of camera poses is mainly caused by the mirror since the non-Lambertian surface causes most general reconstruction algorithms to fail. The 3D Scanner App we used to capture images and camera poses also suffers from this problem. In future work, we will incorporate the joint optimization of camera poses with the neural radiance field to overcome this limitation. To be noted, for the synthetic datasets with accurate camera poses, our method achieves superior results.

Second, we find that in the test sets of real-world datasets shown in Tab. 1, the new reflection only accounts for 0.56% of the test image area on average. This means that NeRF can easily interpolate the reflection memoized from training views to reach high rendering qualities. To compare the correctness of modeling reflection, we captured some challenging test views whose viewpoints are out of the distribution of the training viewpoints. The average proportion of new reflection reaches 5.53% in a test image. On the challenging test views, our method outperforms NeRF and other works as shown in Tab. 2. For intuitive perception, we show the qualitative comparison of novel view synthesis on the challenging test view of the discussion room as an example in Fig. C. Our method synthesizes realistic reflections in the mirror while the results of NeRF and other works are corrupted on the challenging novel views.

Besides, to be noted, NeRF does not reconstruct the physically sound geometry of the mirror, while our method can recover the accurate geometry of the mirror. Moreover, our physically-inspired rendering pipeline enables synthesizing reflections unobserved in training views, and supports various applications which previous NeRF-related works cannot do.

## REFERENCES

[1] Laan Consulting Corp. 2023. 3d Scanner App. https://3dscannerapp.com/. Accessed: 2023-05-07.

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

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]

[4] John Roper Matthew Muldoon. 2022. BlenderSwap. https://www.blenderswap.com/. Accessed: 2022-11-10.

[5] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of European Conference on Computer Vision*. 405–421.

[6] Thomas Müller. 2021. *tiny-cuda-nn*. https://github.com/NVlabs/tiny-cuda-nn

[7] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[8] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems* 33 (2020), 7537–7547.

[9] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. 2021. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. *arXiv preprint arXiv:2112.03907* (2021).

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

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