NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-aware Diffusion

Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, Ravi Ramamoorthi
ICML 2023 · Apple Inc., UC San Diego, MPI, UPenn

Introduction and Motivation

TLDR: We introduce NerfDiff for single-image view synthesis, which combines NeRF with a 3D-aware conditional diffusion model (CDM)

Input
VisionNeRF
Ours
Nearest GT

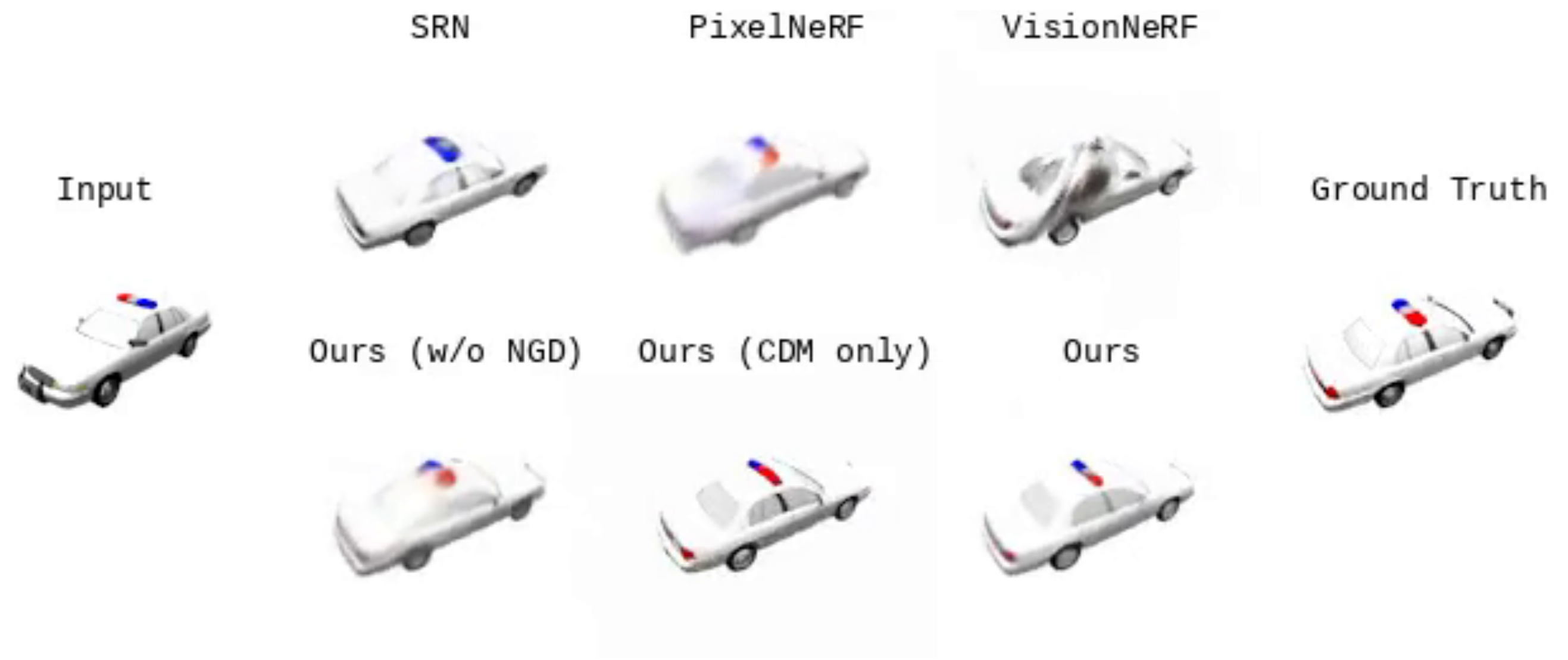
Task: Single-image View Synthesis

Given a single unposed image, our goal is to create a 3D representation which is 1). *consistent with the input image and 2). plausibly synthesizes sharp details behind occlusions.* These two are often at odds with another, making the task extremely challenging.

Input
VisionNeRF
Ours (w/o NGD)
Ours
Nearest GT

Problems with Existing Methods

Existing methods usually rely on one of two mechanisms. In the first, points are projected to the image plane where local image features can be gathered to condition the NeRF (see pixelNeRF, VisionNeRF, GRF). Under severe occlusion, these features have no informative about occluded scene content. In the second mechanism, a global latent code is optimized based on the input image (see AutoRF, SRNs, CodeNeRF). The global bottleneck often hinders rendering sharp details even near the input.

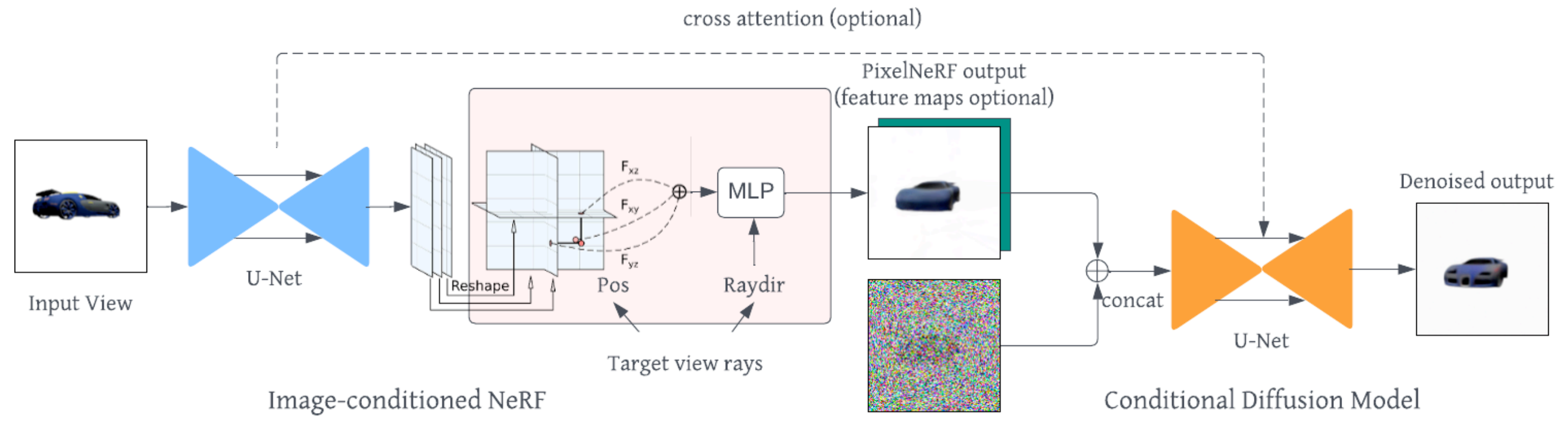


Overview

We simultaneously train a single-image NeRF model which predicts a triplane from an input image and a CDM which is conditioned on these renderings. At test-time, given a single image, we utilize the sharp CDM outputs to fine-tune the NeRF at multiple virtual cameras, effectively inferring behind occlusions.

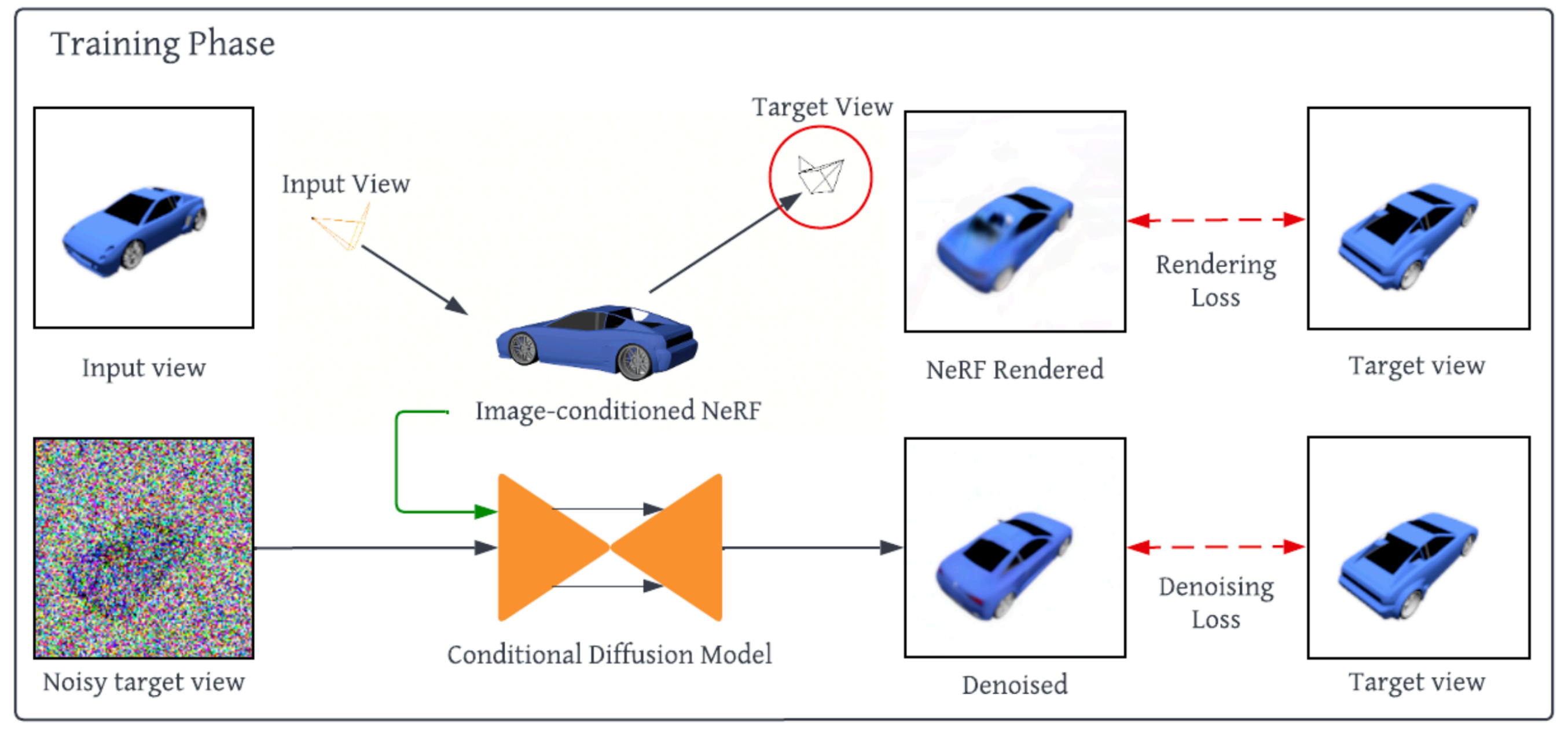
Architecture

Using a UNet, we first map an input image to a camera-aligned triplane-based NeRF representation. This triplane efficiently conditions volume rendering from a target view, resulting in an initial rendering. This rendering conditions the diffusion process so the CDM can consistently denoise at that target pose.



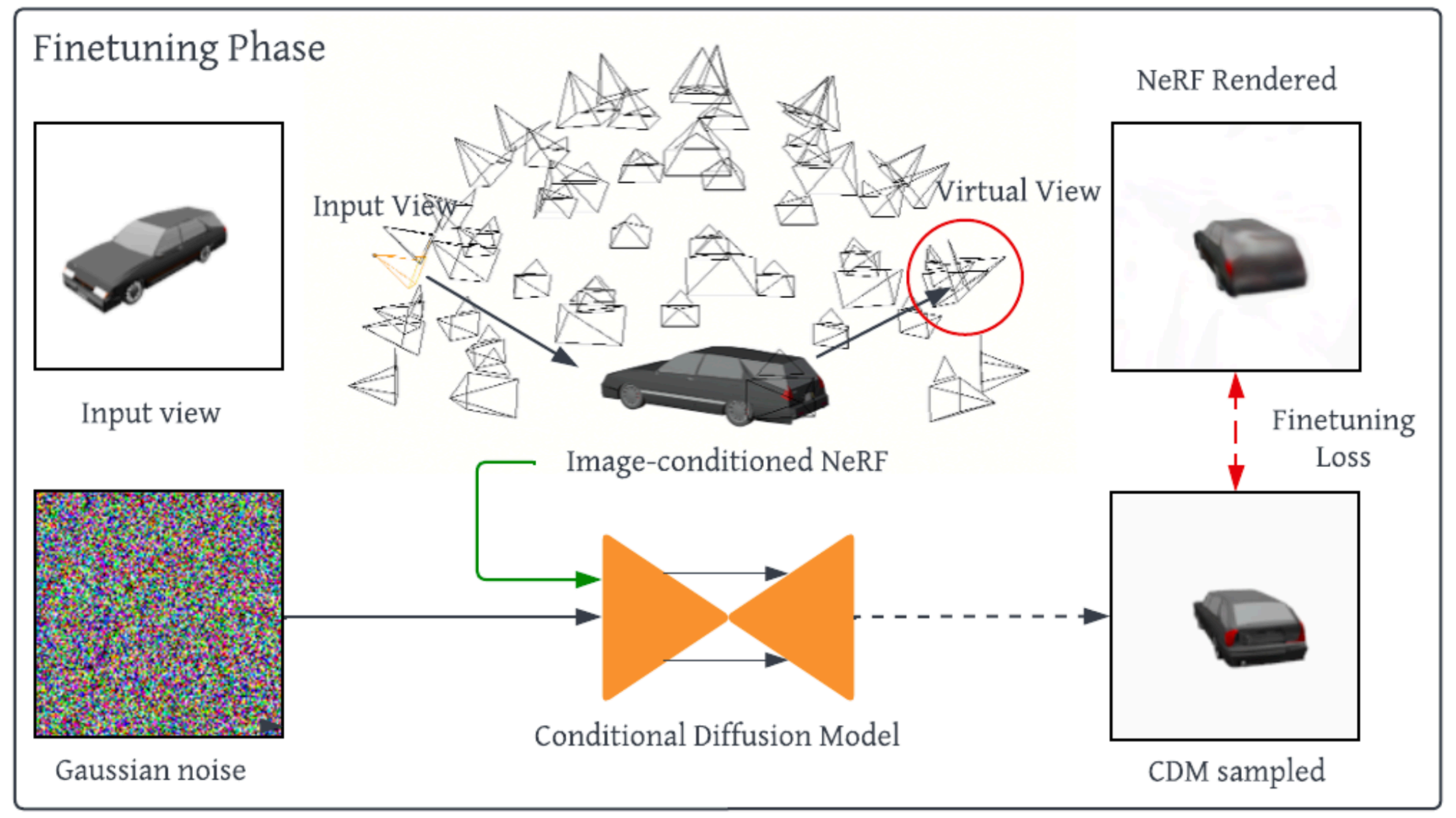
Training stage:

We first learn the single-image NeRF and 2D CDM which is conditioned on the single-image NeRF renderings.



Fine-tuning stage:

At test time, we use the learned network parameters to predict an initial NeRF representation for fine-tuning. The NeRF-guided denoised images from the frozen CDM then supervise the NeRF in-turn (right).



NeRF-guided distillation

The core of our algorithm distills the knowledge of a 3D-aware CDM into the single-image NeRF from multiple virtual views for generating high-quality images. In the meantime, the multi-view diffusion process is guided by the NeRF representation to preserve 3D consistency of the diffusion. The details of the algorithm is shown below:

Algorithm 1 Finetuning with NeRF-guided distillation.

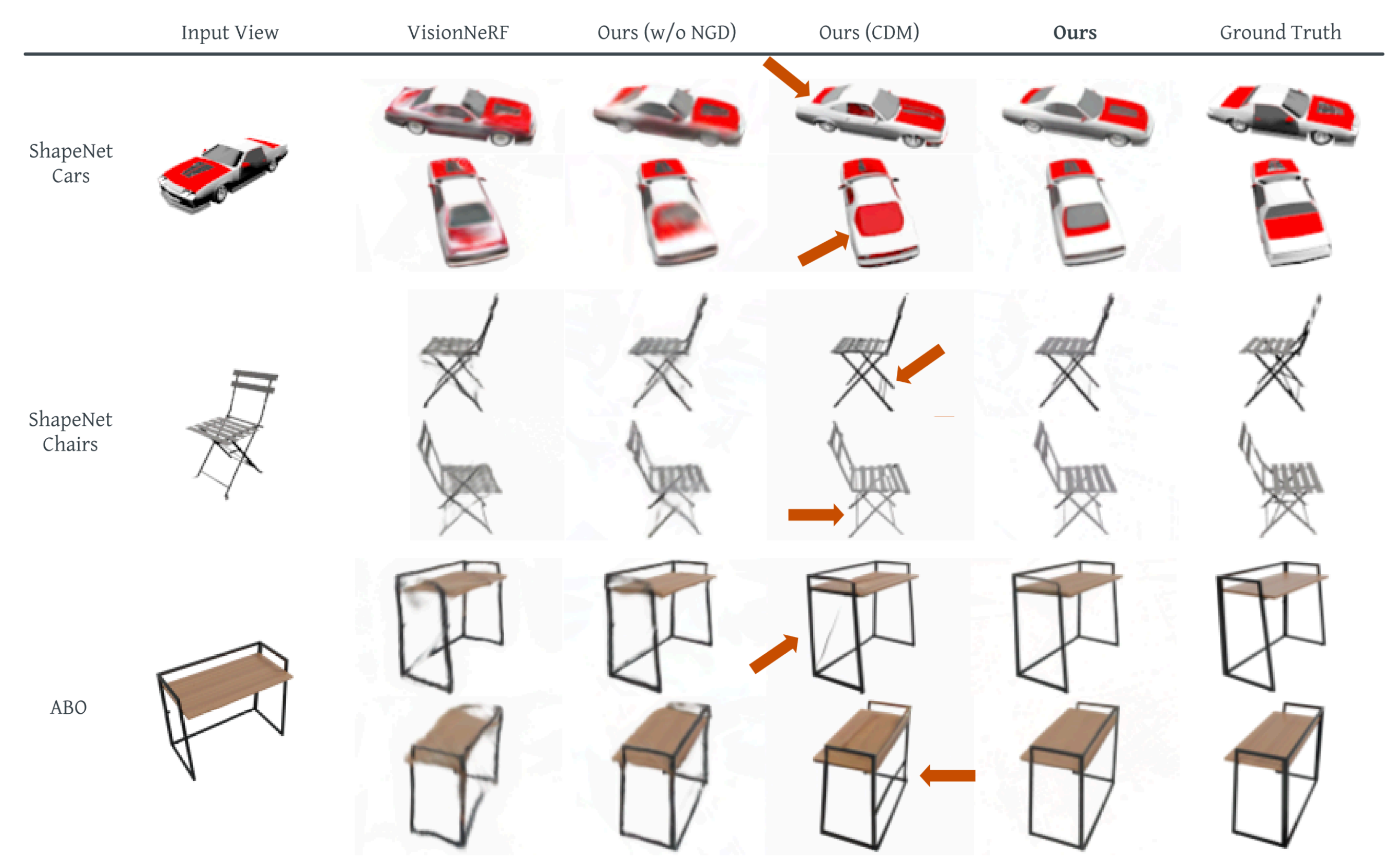
```
Input: NeRF (MLP $f_\theta$, triplanes $W$), CDM $\boldsymbol{\epsilon}_\phi$, input $\boldsymbol{I}^s, \gamma, N, B$
Initialize $I^\pi = I^\pi_{\theta,W}, \boldsymbol{\epsilon}^\pi = \boldsymbol{\epsilon}, \pi \in \Pi, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$
for $t = t_{\max} \dots t_{\min}$ do
    for $\pi \in \Pi$ do
        $\boldsymbol{Z}^\pi = \alpha_t \boldsymbol{I}^\pi + \sigma_t \boldsymbol{\epsilon}^\pi$;
        $\boldsymbol{\epsilon}^\pi = \boldsymbol{\epsilon}_\phi(\boldsymbol{Z}^\pi, \boldsymbol{I}^s) + \gamma\sigma_t/\alpha_t \cdot (\boldsymbol{I}^\pi - \boldsymbol{I}^\pi_{\theta,W})$
        $\boldsymbol{I}^\pi = (\boldsymbol{Z}^\pi - \sigma_t \boldsymbol{\epsilon}^\pi)/\alpha_t$
    for $n = 1 \dots N$ do
        for $b = 1 \dots B$ do
            Sample a view $\pi \sim \Pi$ and sample a ray $\boldsymbol{r}$ from $\pi$;
        Update $\theta, W$ with $\nabla_{\theta,W} \frac{1}{B} \sum_{\pi,\boldsymbol{r}} \|\boldsymbol{I}^\pi_{\theta,W}(\boldsymbol{r}) - \boldsymbol{I}^\pi(\boldsymbol{r})\|_2^2$
return $\theta, W$
```

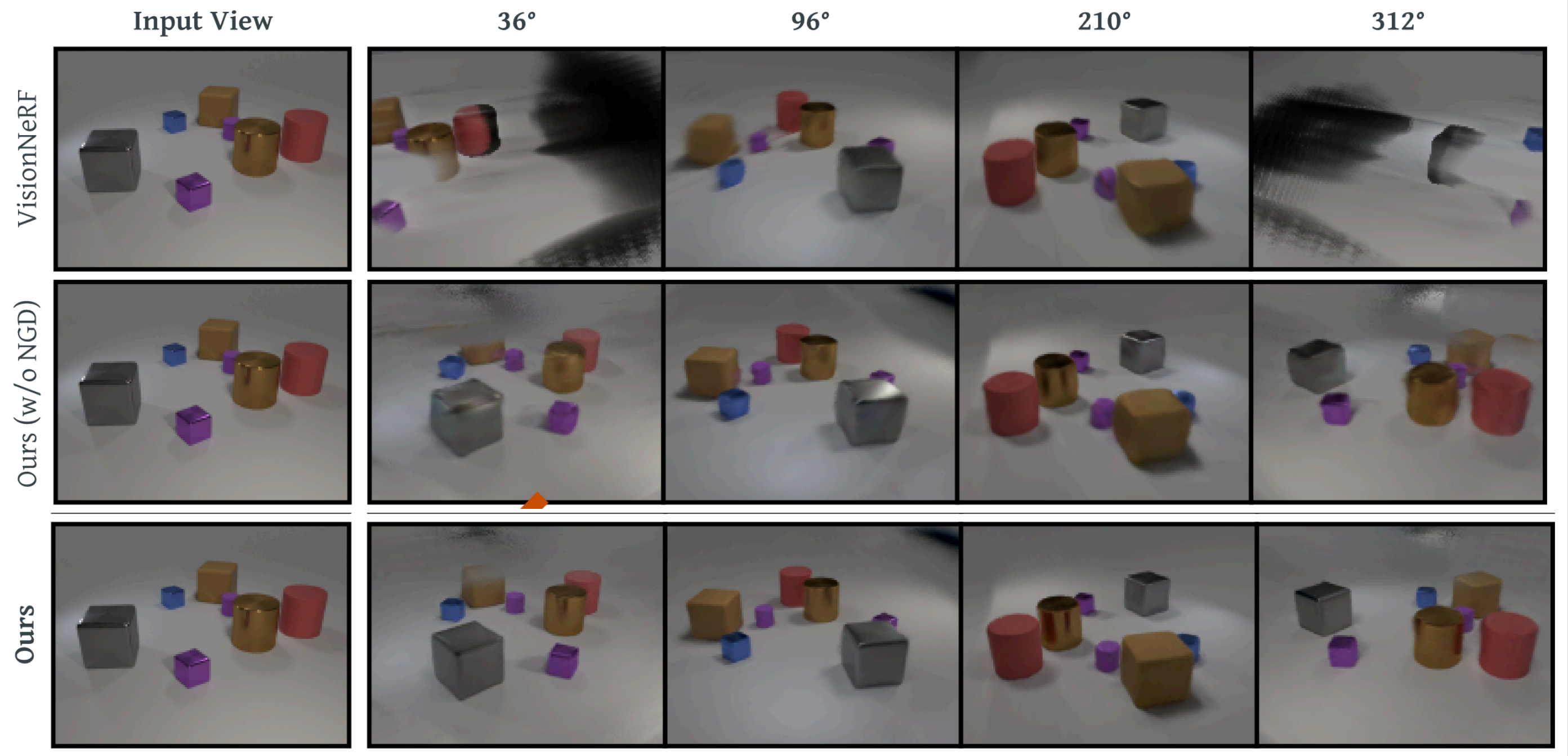
Quantitative Results

Table 1. Comparisons on ShapeNet Cars & Chairs and ABO datasets across baselines. * indicates geometry-free model. The results of the baselines except VisionNeRF (Lin et al., 2023) are copied from the official papers. – denotes the results are unavailable.

	ShapeNet Cars				ShapeNet Chairs				Amazon-Berkeley Objects			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
LFN (Sitzmann et al., 2021)*	22.42	0.89	–	–	22.26	0.90	–	–	–	–	–	–
3DiM (Watson et al., 2022)*	21.01	0.57	–	**8.99**	17.05	0.53	–	6.57	–	–	–	–
SRN (Sitzmann et al., 2019a)	22.25	0.88	0.129	41.21	22.89	0.89	0.104	26.51	–	–	–	–
PixelNeRF (Yu et al., 2021)	23.17	0.89	0.146	59.24	23.72	0.90	0.128	38.49	–	–	–	–
CodeNeRF (Jang & Agapito, 2021)	22.73	0.89	0.128	–	23.39	0.87	0.166	–	–	–	–	–
FE-NVS (Guo et al., 2022)	22.83	0.91	0.099	–	23.21	0.92	0.077	–	–	–	–	–
VisionNeRF (Lin et al., 2023)	22.88	0.90	0.084	21.31	24.48	0.92	0.077	10.05	28.61	0.93	0.095	33.38
NerfDiff-B (Ours)	23.51	**0.92**	0.082	18.09	24.79	0.94	**0.056**	5.65	32.81	0.96	0.057	7.77
w/o NGD	23.81	**0.92**	0.093	42.37	24.77	0.93	0.068	15.72	32.07	0.95	0.063	18.01
NerfDiff-L (Ours)	23.76	**0.92**	**0.076**	15.49	**24.95**	**0.94**	**0.056**	**5.34**	**32.84**	**0.97**	**0.042**	**6.31**
w/o NGD	**23.95**	**0.92**	0.092	43.26	24.80	0.93	0.070	15.50	32.00	0.96	0.061	17.73

Qualitative Results





Links

Paper: https://arxiv.org/abs/2302.10109
Project Page: https://jiataogu.me/nerfdiff/