Supplementary for
NeRF-DS: Neural Radiance Fields for Dynamic Specular Objects

Zhiwen Yan Chen Li Gim Hee Lee
Department of Computer Science, National University of Singapore
{yan.zhiwen, lichen}@u.nus.edu gimhee.lee@nus.edu.sg

1. Qualitative Result Videos

We include a few videos rendered by our model and the baseline models in the supplementary zip file as a better demonstration of the qualitative performance comparison.

2. Implementation Details

The details of the mask prediction module (Fig. 1), deformation prediction module (Fig. 2), hyper coordinate prediction module (Fig. 3) and canonical NeRF (Fig. 4) module are illustrated in the respective figure. Positional encoding is performed on spatial coordinates $\mathbf{x}$, $\mathbf{x}'$, viewing direction $\omega_\mathbf{o}$ and surface normal $\mathbf{n}$. Different encoding widths and annealing widths are used for different input as shown in Tab. 1. The Gaussian applied to the weights w' for mask volumetric rendering has an exponentially decreasing standard deviation β from 1 to 0.1 during the first 30k iterations. Then it stays constant at 0.1 for the rest of the training.

2.1. Details of Ref-NeRF Experiments

We use the official integrated Ref-NeRF [2] code from Multi-NeRF [1]. To accommodate our dynamic specular dataset, we slightly adjust the scene offset and scaling logic to ensure the scene is well centered and bounded.

2.2. Parameter and Training Time

The full model contains 1.45M parameters, compared to the 1.30M parameters of the baseline model. The experiment with 480x270 resolution videos takes 6 hours to train on 4 RTX A5000 GPUs, compared to the 5 hours training time of the baseline model.

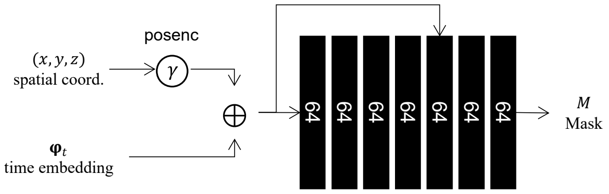


Figure 1. Architecture of the mask prediction module.

	width	anneal	delay iter.	inc. iter.	inc. func.
$\mathbf{x}$ to mask	4	Yes	0k	50k	linear
$\mathbf{x}$ to deformation	4	Yes	0k	50k	linear
$\mathbf{x}$ to hyper coord.	6	No	N/A	N/A	N/A
$\mathbf{x}$ to color branch	4	Yes	50k	50k	linear
$\mathbf{x}'$ to NeRF	8	No	N/A	N/A	N/A
$\mathbf{w}$ to NeRF	1	No	N/A	N/A	N/A
$\omega_\mathbf{o}$ to color branch	4	No	N/A	N/A	N/A
$\mathbf{n}$ to color branch	4	Yes	10k	2k	linear

Table 1. Details of the positional encoding and annealing of each input. “Width” indicates the highest k in $sin(2^k\pi\mathbf{x})$ sequence. “Anneal” indicates whether annealing coefficient $z_j(\tau)$ for positional encoding is used. If annealing is used, “delay iter.” is the number of iterations where τ stays 0 at the start of the training. “inc. iter.” and “inc. func.” are the number of increasing iterations and function after the delay.

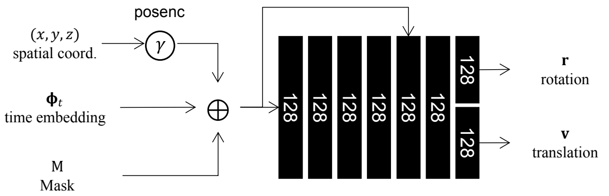


Figure 2. Architecture of the deformation field prediction module.

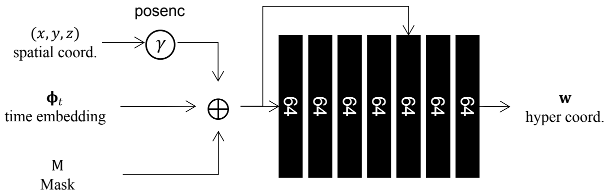


Figure 3. Architecture of the hyper coordinates prediction module.

3. Ablation Qualitative Results

We present the qualitative comparison between the full and ablation versions of our models. The comparison between our NeRF-DS model and the ablation version without surface-aware dynamic NeRF is shown in Fig. 5. The comparison between our NeRF-DS model and the ablation

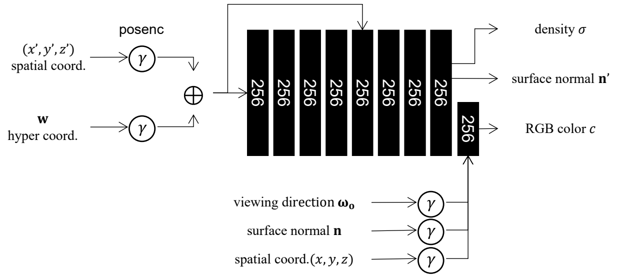


Figure 4. Architecture of the canonical NeRF module.

Positional Encoding Annealing	MS-SSIM↑	PSNR↑	LPIPS↓
constant 4	0.911	25.4	0.118
increase to 4 for 50k iter.	0.915	25.7	0.119
delay 10k, increase to 4 for 50k iter.	0.914	25.6	0.121
delay 50k, increase to 4 for 50k iter.	0.918	25.7	0.115
delay 100k, increase to 4 for 50k iter.	0.917	25.7	0.117
without x	0.913	25.5	0.120

Table 2. Quantitative results on different annealing strategy for adding observation space coordinate $\mathbf{x}$ to the color branch of the canonical NeRF. Experiments are performed on the "Sheet" scene. The best and second best results are color coded.

version without mask guided deformation field is shown in Fig. 6.

4. Additional Experiment Results

We use delayed positional encoding for the spatial location $\mathbf{x}$ and sharp volumetric weights w_i' for the mask rendering. In this section, we present additional ablation experiments to determine the best hyper-parameters for these two techniques.

We evaluate the performance of the NeRF-DS model on the "Sheet" scene in the dynamic specular dataset, under different annealing strategy of the positional encoding for the observation space spatial coordinate $\mathbf{x}$ before it is fed to the NeRF color branch. Specifically, we evaluate the reconstruction performance with different schedules for the annealing coefficient τ of the jth term in the position encoding as shown in:

$$z_j(\tau) = \frac{1 - cos(\pi \cdot clamp(\tau - j, 0, 1))}{2}. \quad (1)$$

We present the quantitative results in Tab. 2. Supported by the quantitative results, we choose to delay the use of $\mathbf{x}$ in the NeRF color branch for the first 50k iterations, and slowly increase the bandwidth to a maximum of 4 during the next 50k iterations.

We also evaluate the performance of the NeRF-DS model on the "Sheet" scene in the dynamic specular dataset, with different sharp weights w_i' for mask rendering. Particularly, we evaluate the reconstruction performance with different schedules for decreasing standard deviation β in the Gaussian filter $\mathcal{N}(k_{\text{max}}, \beta)$ applied to weight w_i based on its ray distance k_i:

$$w_i^* = w_i \cdot P(k_i|\mathcal{N}(k_{\text{max}}, \beta)), w_i' = w_i^* / (\sum_j w_j^*). \quad (2)$$

We present the quantitative results in Tab. 3. Supported by the quantitative results, we choose to gradually decrease standard deviation for sharp mask weights from 1 to 0.1 during the first 30k iterations of the training.

Additionally, we evaluate the performance of the NeRF-DS model on the "Sheet" scene in the dynamic specular dataset, with surface normal $\mathbf{n}$ calculated from different spaces. The surface normal in the observation space used in our main results are warped from the surface normal in the canonical space to ensure cross frame consistency, *i.e.* $\mathbf{n} = \mathrm{R}^\top \mathbf{n}'$. We compare the results with the model using surface normal calculated in the canonical space and the surface normal directly calculated in observation space as shown in Tab. 4. The canonical space normal means $\mathbf{n} = \mathbf{n}'$. The observation space normal means the normal is supervised by the gradient of density with respect to the spatial coordinate in observation space:

$$\hat{\mathbf{n}} = -\frac{\nabla\sigma(\mathbf{x})}{\|\nabla\sigma(\mathbf{x})\|}, \quad (3a)$$

$$\mathcal{L}_{norm} = \textstyle\sum_i w_i \|\mathbf{n} - \hat{\mathbf{n}}\|^2. \quad (3b)$$

Supported by the quantitative comparison, we choose to use the surface normal warped from the canonical space for the better consistency over time.

To demonstrate that our model has comparable performance to the baselines on non-specular dynamic scenes, we also present the experiment results of our model in the released scenes in the HyperNeRF dataset in Tab. 5. The results shown for Nerfies [2] and HyperNeRF [3] are taken from the original paper, while the performance of our model is reproduced on the same data. Please note that due to our limited hardware (compared to the 4 TPU used in the original paper), our model trained on this HyperNeRF [3] dataset is using 1/10 of the batch size and 10 times the number of iterations. The performance comparison in this way is slightly in our disadvantages, as our reproduced HyperNeRF [3] models under this setting perform worse than the reported models.

5. Additional Qualitative Analysis

To further analyse the influence of the surface normal input on the rendering, we present a qualitative case study. Taking the early stage result of NeRF-DS (w/o mask) as an example (Fig. 8), the norms predicted for the middle part of the plate in two frames are different. With this input, our NeRF-DS model can render different reflected colors of the same surface. However, HyperNeRF fails to recognize the surfaces in the two frames to be the same and renders severe

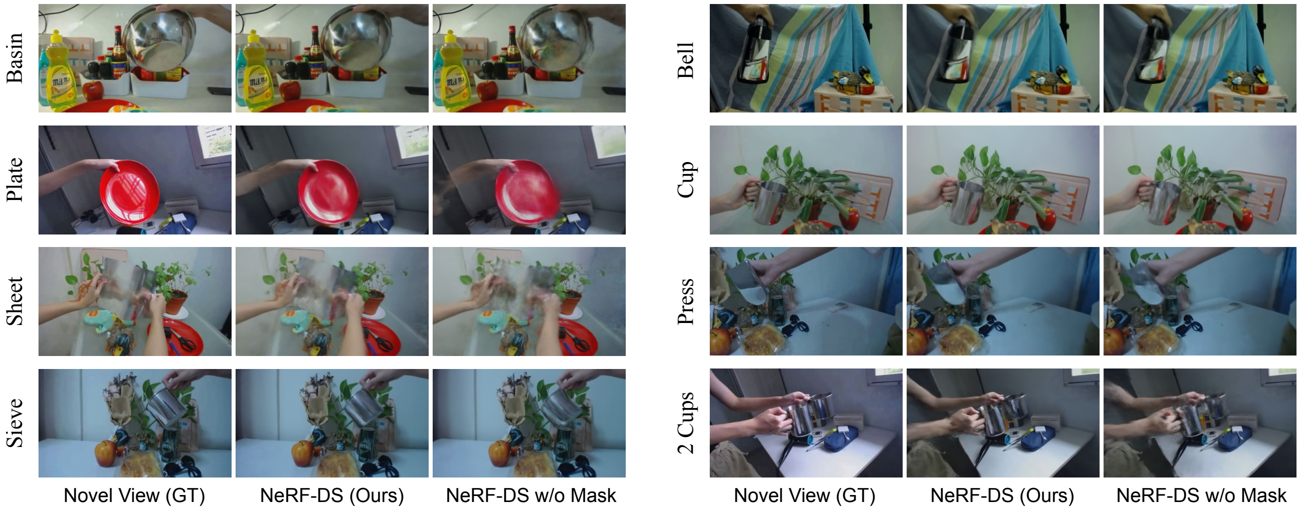


Figure 5. Qualitative comparison between our full model (NeRF-DS) and the ablation version without the surface-aware dynamic NeRF (NeRF-DS w/o Mask).

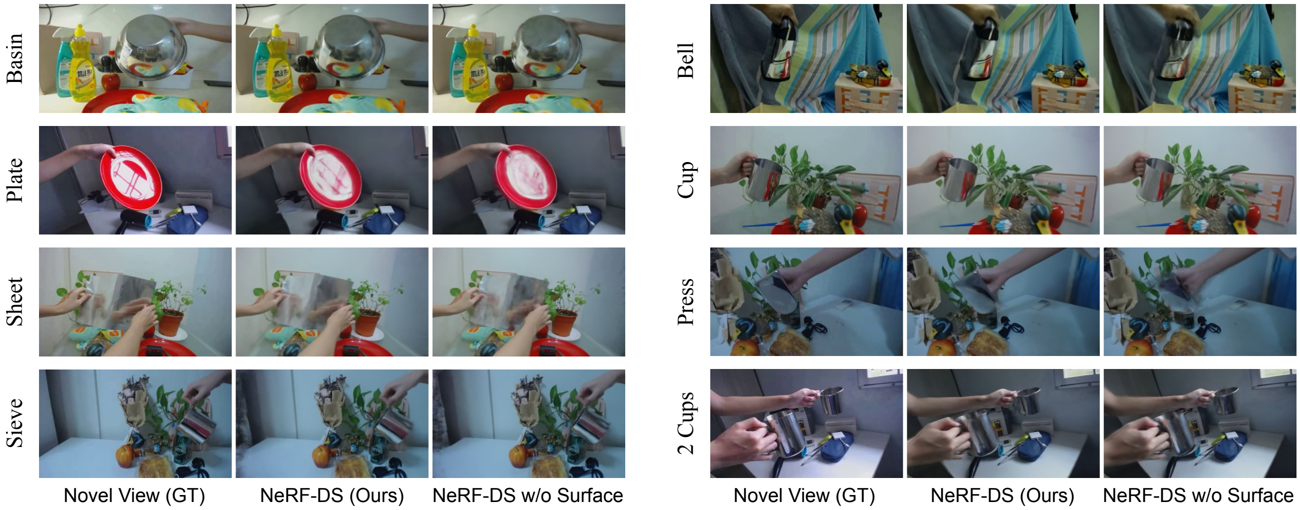


Figure 6. Qualitative comparison between our full model (NeRF-DS) and the ablation version without the mask guided deformation field (NeRF-DS w/o Surface).

Standard Deviation Schedule	MS-SSIM↑	PSNR↑	LPIPS↓
1 to 0.01 for 30k iter.	0.916	25.8	0.122
1 to 0.03 for 30k iter.	0.905	25.3	0.125
1 to 0.1 for 30k iter.	0.918	25.7	0.115
1 to 0.3 for 30k iter.	0.917	25.7	0.120
without sharping	0.909	25.6	0.126

Table 3. Quantitative results on different schedule for decreasing the standard deviation β for the Gaussian filter to sharp the mask weights. Experiments are performed on the “Sheet” scene. The best and second best results are color coded.

Surface Normal	MS-SSIM↑	PSNR↑	LPIPS↓
Warped from canonical space	0.918	25.7	0.115
Canonical space normal	0.913	25.5	0.119
Observation space normal	0.913	25.6	0.117

Table 4. Quantitative results on types of surface normal $\mathbf{n}$ used. Experiments are performed on the “Sheet” scene. The best and second best results are color coded.

geometric artifacts. Additional masks can further suppress the geometric artifacts, but our ablation study suggests that the surface normal alone also contributes significantly to the performance (20.3% LPIPS improvement from baseline).

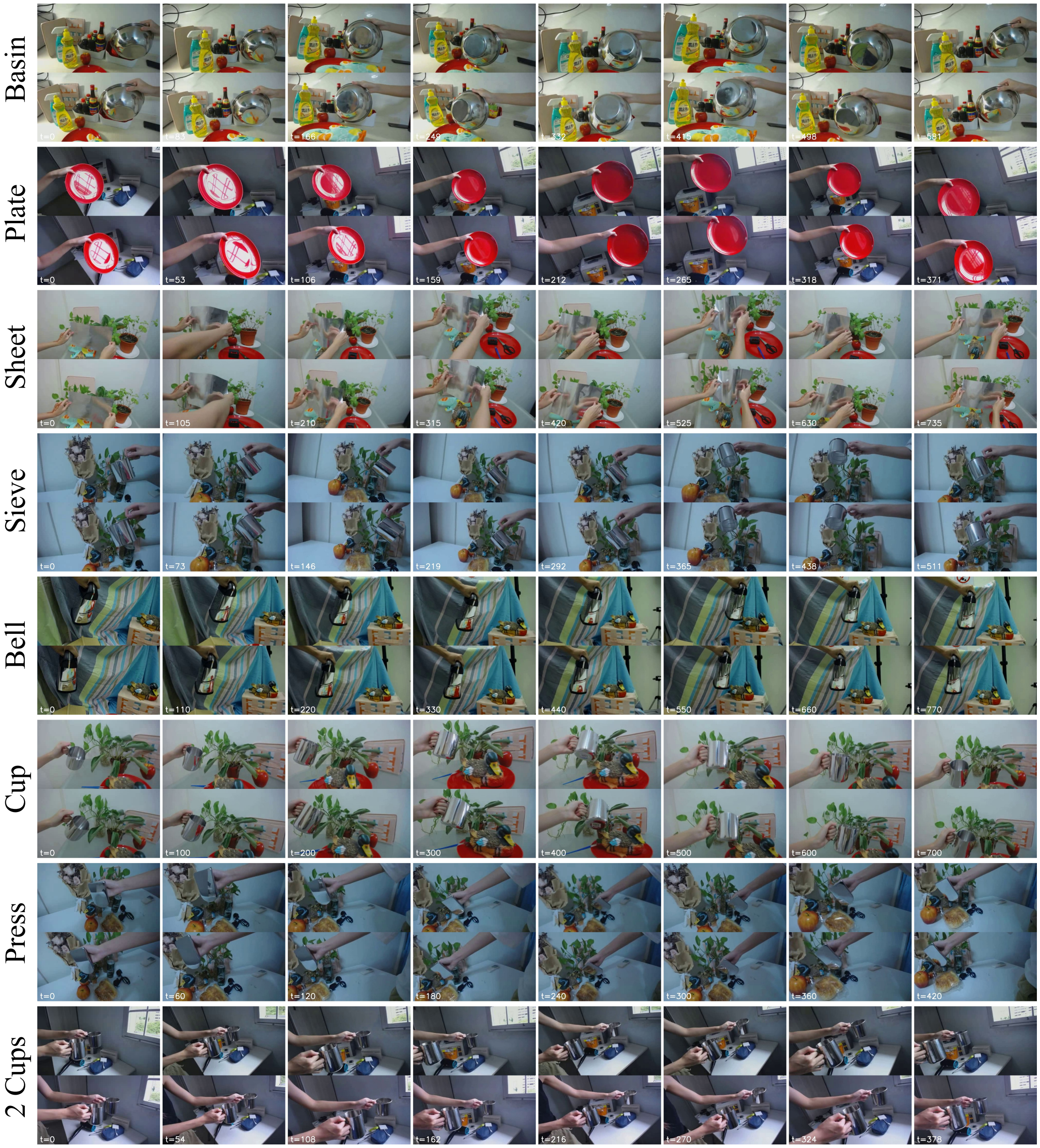


Figure 7. A snippet of the dynamic specular dataset for both cameras in 8 scenes. The training camera video is shown on the top and the test camera video is shown on the bottom.

	Printer	Broom	Chicken	Banana	Mean
	PSNR↑	PSNR↑	PSNR↑	PSNR↑	PSNR↑
Nerfies [2]	20.0	19.3	26.9	23.3	22.4
HyperNeRF [3]	20.0	**20.6**	27.6	**24.3**	**23.1**
NeRF-DS (Ours)	**21.0**	19.6	**27.9**	22.8	22.8

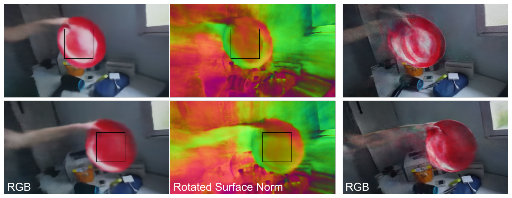
Table 5. Performance on non-specular HyperNeRF [3] dataset.

6. Dynamic Specular Dataset Details

The dataset consists of 8 scenes of various dynamic specular objects in everyday environments. Two rigidly connected cameras are used to capture the scenes for 480x270 resolution. Different types of objects and surfaces are used as shown in Tab. 6. A snippet of the dataset is shown in Fig. 7.

References

[1] Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ricardo Martin-Brualla, and Jonathan T. Barron. Multi-



NeRF-DS(w/o Mask) at 5k iter HyperNeRF at 250k iter

Figure 8. A case study on how different surface norms can guide rendering different reflected colors.

Scene Name	# frames	Object Attribute
Basin	668	Curved+Flat, Metallic
Plate	424	Curved+Flat, Plastic, Colored
Sheet	846	Soft, Metallic, Non-Rigid Deformation
Sieve	584	Curved, Metallic, Porous Bottom
Bell	881	Slightly Curved, Metallic
Cup	807	Curved+Flat, Metallic
Press	487	Flat, Metallic
2 Cups	437	Curved+Flat, Metallic, 2 Objects

Table 6. Details of each scene in the dynamic specular dataset.

NeRF: A Code Release for Mip-NeRF 360, Ref-NeRF, and RawNeRF, 2022. 1

[2] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1, 2, 4

[3] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2, 4