Supplementary Material for Dynamic Neural Radiance Field From Defocused Monocular Video

Xianrui Luo[1], Huiqiang Sun[1], Juewen Peng[2], and Zhiguo Cao[1]

[1] Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of AIA, Huazhong University of Science and Technology, China
{xianruiluo,shq1031,zgcao}@hust.edu.cn
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
juewen.peng@ntu.edu.sg
https://github.com/xianrui-luo/D2RF

This document includes the following contents:

1. Discussion of more DoF-aware works.
2. Layered composition for DoF rendering.
3. Details of the dataset
4. Failure case.
5. More experiment results, including the results on the real defocused scene, quantitative results on individual scenes, comparison with state-of-the-art methods, and ablation study.

1 Discussion of More DoF-aware works

Here we discuss the distinctions between our pipeline and several more DoF-aware works. The inputs of HDR-NeRF [8] and NeRFocus [15] are all-in-focus sharp static images, and HDR-NeRF synthesizes HDR novel views with DoF effects, NeRFocus also generates defocused novel views. Our work, on the other hand, generates sharp novel views from defocused dynamic videos.

[5] is an unsupervised generative model, the training images are **single** images. [5] focuses on the unsupervised generative task and the generated images do not correspond with inputs. Ours is a supervised novel view synthesis task with defocused videos.

[4] does not explore NeRF task, it only uses an implicit model to generate DoF, and the inputs are static image stack.

2 DoF Rendering

We further describe the principles and technical details of DoF rendering. The layered composition in DoF rendering is similar to the multiplane image [14] (MPI). The RGB image C with its visibility weight W is divided into layers by depth discretization. For current layered rendering [1, 17], the composition

weight W is predefined by a certain fixed algorithm. On the other hand, in our method, we define the visibility weight W from its physical principle, utilizing NeRF volume rendering to learn the compositing formation.

3 The Dataset

We conduct our experiments on dynamic scenes from a stereo dataset VDW [16]. The dataset is collected from four data sources: movies, animations, documentaries, and web videos. The image sequences are over 1080p and are cropped at a resolution of 1880×720 or 1880×800. Since the dataset provides RGB sequences with their corresponding aligned disparity sequences, we generate defocus blur from the state-of-the-art DoF rendering method [10]. This method is proven to have better performance than current classical rendering and neural rendering methods. To be as realistic as real-world video-capturing scenarios, we adjust the focal distance along the scene disparity like the typical focusing.

Although the VDW dataset is a large-scale dataset, not all scenes are suitable for our task due to two reasons: (1) many scenes lack moving objects and are static; (2) a large number of scenes exhibit minimal camera motion, resulting in insufficient parallax to obtain camera parameters. Therefore, we carefully select 8 dynamic scenes from the dataset that are eligible for dynamic NeRF methods. These scenes consist of diverse object movements such as moving cars, walking, and opening a gate. The 8 scenes are named Camp, Shop, Car, Mountain, Dining1, Dining2, Dock, Gate. We use the defocused sequences as inputs for DPT [11] to obtain depth maps, and we utilize RAFT [13] to generate optical flows, we also employ an instance segmentation network (Mask R-CNN [3]) to obtain motion masks for moving objects.

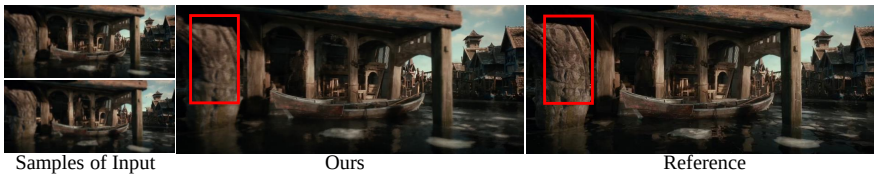

Fig. 1: The failure case. When the input views have severe defocus blur, the novel views may be unable to recover sharp details.

4 Failure Case

As shown in Fig. 1, our method may be unable to recover from extreme defocus blur. Extreme blur occurs when the focal distance is too close or too far from the camera, increasing the blur amount of the out-of-focus areas. The extreme defocus blur poses challenges for our method in two aspects: (1) the excessive

information loss of the defocused regions may hinder the ability of the depth and flow prediction networks, perturbing the representation of dynamic scenes; (2) the model may get stuck in local minima reconstructing the scene when supervised by extreme blur. Furthermore, our model may fail to reconstruct sharp NeRF when all the frames of the sequences have consistent defocus blur, which means a certain object keeps focused for the whole sequence and other objects are consistently out-of-focus. However, capturing a dynamic scene often results in a shift in focal distance, so consistent blur is not common. We require a larger memory cost than existing methods due to DoF modeling, but the additional cost is acceptable (Tab. 1). For future work, we aim to explore the integration of explicit representations to better utilize the in-focus regions from input frames and address the extreme blur issue.

Table 1: Calculation cost.

	NSFF	HyperNeRF	DVS	RoDynRF	Ours
Memory (G)	11.9	14.9	12.4	11.9	20.1

5 Experiments

We show the quantitative results with state-of-the-art methods on all scenes in Tab. 2. We also collect one real dynamic defocused video for evaluation. As shown in Fig. 2, our method is more stable and generates more reliable sharp details. The ablation results are in Tab. 3 and Tab. 4 We show qualitative comparisons with state-of-the-art methods in Figs. 3 and 4. We choose different scenes from the main article to show results on all scenes. We visualize the ablation results in Fig. 5.

Fig. 2: Results on real defocused video.

Table 2: Quantitative results on the synthesized dataset. The best performance is in **boldface**, and the second best is underlined. Although the baselines achieve better results on a certain indicator in some scenes, our method achieves better visualization quality and restores sharper details than other baselines as shown in the qualitative results and the supplementary video.

Method	Camp			Shop			Car			Mountain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DVS [2]	17.18	0.525	0.399	27.05	0.842	0.216	24.57	0.795	0.240	**33.54**	0.896	0.174
NSFF [6]	**21.17**	0.643	0.310	26.54	0.838	0.207	25.57	0.799	0.223	32.11	0.891	0.158
HyperNeRF [9]	21.08	0.627	0.294	27.21	0.834	0.216	25.44	0.769	0.216	33.02	0.886	0.160
RoDynRF [7]	20.99	0.597	0.312	**28.53**	0.844	0.213	22.62	0.726	0.264	28.71	0.858	0.194
[12] + DVS	17.07	0.531	0.352	26.47	0.850	0.155	23.72	0.786	0.171	31.87	**0.900**	0.128
[12] + RoDynRF	20.53	0.564	0.345	28.15	**0.856**	0.159	22.65	0.756	0.197	27.46	0.863	0.149
D^2RF (Ours)	20.73	**0.644**	**0.207**	27.01	0.854	**0.117**	**26.79**	**0.852**	**0.123**	32.44	**0.900**	**0.079**
Method	Dining1			Dining2			Dock			Gate		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DVS [2]	23.07	0.712	0.287	27.85	0.822	0.165	27.54	0.807	0.220	22.60	0.713	0.231
NSFF [6]	30.54	0.865	0.175	27.54	0.812	0.152	**28.42**	0.822	0.216	24.17	0.754	0.228
HyperNeRF [9]	30.48	0.854	0.198	27.55	0.818	0.165	27.94	0.806	0.212	22.99	0.647	0.202
RoDynRF [7]	29.21	0.803	0.243	**28.25**	0.816	0.171	27.07	0.789	0.214	24.06	0.726	0.201
[12] + DVS	22.91	0.724	0.252	27.78	**0.826**	0.153	26.94	0.790	0.189	19.36	0.647	0.265
[12] + RoDynRF	28.96	0.812	0.205	27.92	0.815	0.156	26.83	0.797	0.178	23.85	0.741	**0.182**
D^2RF (Ours)	**31.05**	**0.877**	**0.105**	27.84	0.817	**0.094**	28.12	**0.823**	**0.131**	**24.41**	**0.757**	0.185

Fig. 3: The qualitative results with all dynamic NeRF baselines. Compared with existing dynamic NeRF methods, our method generates sharper novel views that are more faithful and have more details. The scenes are Camp, Dining1, Dining2, Gate.

Table 3: Ablation results on the synthesized dataset. The results only calculate the dynamic regions from motion masks. The best performance is in **boldface**, and the second best is underlined. The results show that our method works best with all the modules, and missing one of them causes performance degradation.

Method	Camp			Shop			Car			Mountain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o layered volume	20.89	0.664	0.245	19.65	0.583	0.245	24.34	0.886	0.077	**26.09**	0.818	0.125
w/o optimized kernel	**20.91**	0.641	0.331	19.76	0.592	0.256	25.04	0.888	0.130	25.90	0.796	0.196
w/o static	20.68	0.658	0.252	19.42	0.567	0.221	22.96	0.867	0.085	25.31	0.780	0.125
Full (Ours)	20.70	**0.665**	**0.220**	**19.93**	**0.595**	**0.203**	**25.18**	**0.901**	**0.075**	25.95	**0.820**	**0.106**
Method	Dining1			Dining2			Dock			Gate		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o layered volume	30.99	0.866	0.099	25.29	0.724	**0.135**	27.40	**0.779**	**0.136**	20.89	0.594	0.294
w/o optimized kernel	30.61	0.851	0.144	25.66	0.733	0.221	27.47	0.767	0.219	20.97	**0.615**	0.472
w/o static	30.59	0.850	0.141	24.69	0.697	0.144	**27.65**	0.775	0.180	20.19	0.558	**0.280**
Full (Ours)	**31.01**	**0.871**	**0.085**	**25.76**	**0.745**	0.138	27.24	0.774	0.150	**21.03**	0.603	0.316

Table 4: Ablation results on the synthesized dataset. The results calculate the whole image. The best performance is in **boldface**, and the second best is underlined. The results show that our method works best with all the modules, and missing one of them causes performance degradation.

Method	Camp			Shop			Car			Mountain		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o layered volume	**20.90**	0.641	0.230	26.78	**0.854**	0.154	25.93	0.835	0.134	**32.57**	0.899	0.099
w/o optimized kernel	20.86	0.615	0.317	26.85	0.853	0.180	**26.72**	0.810	0.218	31.86	0.881	0.157
w/o static	20.77	0.633	0.238	26.33	0.825	0.180	24.69	0.798	0.149	31.02	0.862	0.117
Full (Ours)	20.73	**0.644**	**0.207**	**27.02**	**0.854**	**0.115**	26.79	**0.852**	**0.123**	32.44	**0.900**	**0.079**
Method	Dining1			Dining2			Dock			Gate		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o layered volume	30.94	0.872	0.118	27.45	0.806	0.098	28.24	0.833	**0.118**	24.26	0.749	0.189
w/o optimized kernel	30.64	0.857	0.152	**28.33**	**0.830**	0.144	**28.80**	**0.834**	**0.118**	23.93	0.681	0.438
w/o static	30.31	0.849	0.161	25.50	0.734	0.128	27.87	0.804	0.183	23.14	0.645	0.259
Full (Ours)	**31.05**	**0.877**	**0.105**	27.85	0.817	**0.094**	28.12	0.823	0.131	**24.41**	**0.757**	**0.185**

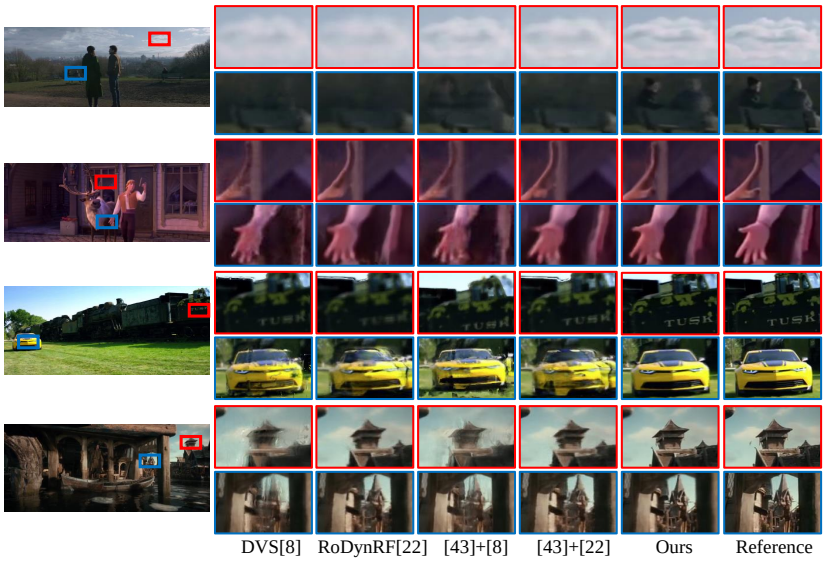

Fig. 4: The qualitative results with dynamic NeRF and their corresponding 2D image deblurring baselines. The scenes are Mountain, Shop, Car, Dock.

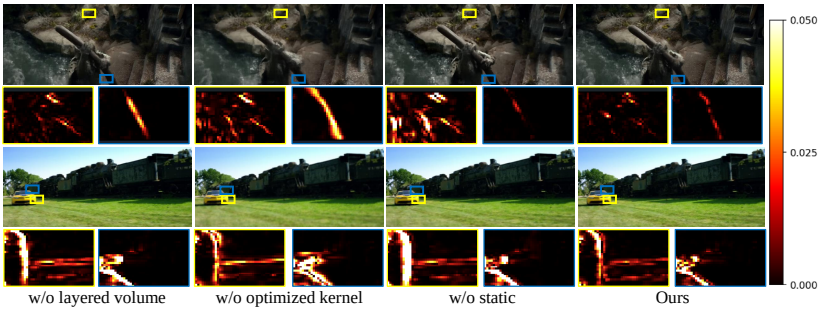

Fig. 5: The visualizations of the ablation study. The corresponding error map is visualized at the bottom, where darker regions indicate smaller errors. We define the error range from 0 to 0.05. Our full model has the smallest error overall.

References

1. Busam, B., Hog, M., McDonagh, S., Slabaugh, G.: Sterefo: Efficient image refocusing with stereo vision. In: Proc. IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 0–0 (2019) 1
2. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV. pp. 5712–5721 (2021) 4
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 2
4. Huang, X., Zhang, Q., Feng, Y., Li, H., Wang, Q.: Inverting the imaging process by learning an implicit camera model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21456–21465 (2023) 1
5. Kaneko, T.: Ar-nerf: Unsupervised learning of depth and defocus effects from natural images with aperture rendering neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18387–18397 (2022) 1
6. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: CVPR. pp. 6498–6508 (2021) 4
7. Liu, Y.L., Gao, C., Meuleman, A., Tseng, H.Y., Saraf, A., Kim, C., Chuang, Y.Y., Kopf, J., Huang, J.B.: Robust dynamic radiance fields. In: CVPR. pp. 13–23 (2023) 4
8. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16190–16199 (June 2022) 1
9. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. ACM TOG **40**(6), 1–12 (2021) 4
10. Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., Zhang, J.: Bokehme: When neural rendering meets classical rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16283–16292 (2022) 2
11. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 2
12. Son, H., Lee, J., Cho, S., Lee, S.: Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2642–2650 (2021) 4
13. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020) 2
14. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 551–560 (2020) 1
15. Wang, Y., Yang, S., Hu, Y., Zhang, J.: Nerfocus: Neural radiance field for 3d synthetic defocus. arXiv preprint arXiv:2203.05189 (2022) 1
16. Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., Xian, K., Lin, G.: Neural video depth stabilizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9466–9476 (2023) 2
17. Zhang, X., Matzen, K., Nguyen, V., Yao, D., Zhang, Y., Ng, R.: Synthetic defocus and look-ahead autofocus for casual videography. ACM Transactions on Graphics (TOG) **38**, 1–16 (2019) 1