





Dynamic Neural Radiance Field From Defocused Monocular Video

Xianrui Luo¹, Huiqiang Sun¹, Juewen Peng², and Zhiguo Cao¹

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>

Abstract. Dynamic Neural Radiance Field (NeRF) from monocular videos has recently been explored for space-time novel view synthesis and achieved excellent results. However, defocus blur caused by depth variation often occurs in video capture, compromising the quality of dynamic reconstruction because the lack of sharp details interferes with modeling temporal consistency between input views. To tackle this issue, we propose D^2RF , the first dynamic NeRF method designed to restore sharp novel views from defocused monocular videos. We introduce layered Depth-of-Field (DoF) volume rendering to model the defocus blur and reconstruct a sharp NeRF supervised by defocused views. The blur model is inspired by the connection between DoF rendering and volume rendering. The opacity in volume rendering aligns with the layer visibility in DoF rendering. To execute the blurring, we modify the layered blur kernel to the ray-based kernel and employ an optimized sparse kernel to gather the input rays efficiently and render the optimized rays with our layered DoF volume rendering. We synthesize a dataset with defocused dynamic scenes for our task, and extensive experiments on our dataset show that our method outperforms existing approaches in synthesizing all-in-focus novel views from defocus blur while maintaining spatial-temporal consistency in the scene.

Keywords: Dynamic Novel View Synthesis · Neural Radiance Field · Depth-of-Field

1 Introduction

Capturing videos from cameras or smartphones has become a new norm for daily life. However, videos are typically recorded from a monocular camera, restricting the captured scene to fixed viewpoints. To depict the dynamic scene from flexible viewpoints, dynamic view synthesis [23, 24, 38] is proposed to facilitate the generation of photorealistic novel views from arbitrary camera angles and perspectives, thus allowing for free viewpoints. The ability to capture a 3D dynamic scene in

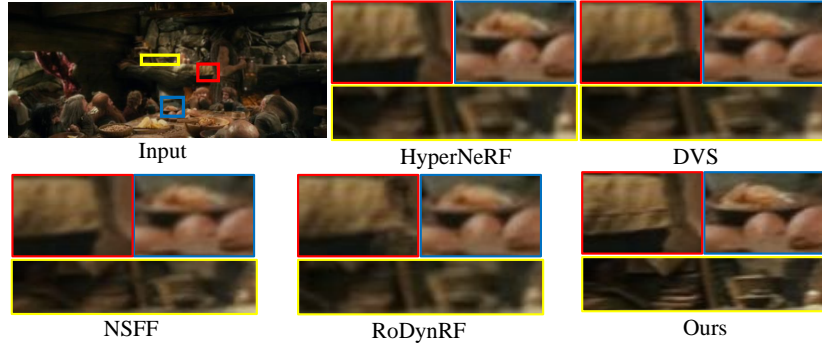


Fig. 1: Given a monocular video captured with defocus blur, existing dynamic NeRF approaches fail to recover high-quality details and tend to produce blurry views, and our method D^2RF synthesizes sharp novel views.

a monocular setting enhances the importance of dynamic novel view synthesis in video-related tasks such as video stabilization, augmented reality, and view interpolation. To interpolate new views from existing sparse input views, recent works use neural networks to train a neural radiance field (NeRF) [31]. The network is a multilayer perceptron (MLP) that learns to map spatial coordinates and view directions to volume densities and view-dependent RGB colors.

Recent dynamic NeRF methods [23, 24, 35] have made progress by establishing spatial-temporal consistency to achieve space-time novel view synthesis. However, these methods are designed under the assumption that the model is trained on all-in-focus image sequences, where all objects within the scene remain focused and sharp during the entire filming. Defocus blur has not been addressed by current dynamic NeRF models, and the blur can result in performance degradation for the reconstructed dynamic NeRF, as it poses challenges for modeling dynamic object motions due to the lack of sharp details and the inability to model temporal consistency. As shown in Fig. 1, current dynamic NeRF methods perform poorly when applied to monocular videos affected by defocus blur, resulting in an inability to restore sharp contents. Maintaining all objects in the scene in focus may offer a solution, however, maintaining every frame in focus is challenging when capturing monocular videos. Defocus blur occurs when the camera’s aperture is wide, and the objects are not positioned at the focal distance. Therefore, defocus blur tends to manifest in videos due to the scene comprising multiple objects with significant depth variations coupled with the common equipment of large aperture for more natural light and enhanced imaging quality. Furthermore, the focal distance may constantly fluctuate in frames if the photographer lacks expertise. Considering that defocus blur is almost unavoidable during video capture, recovering a sharp dynamic NeRF is a valuable problem yet to be explored.

Several works have been proposed to address defocus blur from static scenes. Physical scene prior and kernel estimation [19, 28] are introduced to mitigate

blurring, and an explicit scatter-gather technique [56] is proposed to better model the defocus blur. However, these methods are specifically tailored for static multi-view inputs and are not capable of representing defocused dynamic scenes. Previous methods apply blurring after the last step of NeRF: volume rendering. The blurring is before optimization but after the typical rendering.

To tackle defocus blur in monocular dynamic videos, we propose D^2RF , a dynamic framework designed to incorporate the defocus blur model into dynamic NeRF training, enabling the recovery of sharp novel views from defocused monocular videos. Our approach involves modeling defocus blur by estimating blur kernels and introducing layered Depth-of-Field (DoF) volume rendering to the input rays. We connect the classical volume rendering with DoF rendering, seamlessly integrating the blurring process into the NeRF pipeline. We modify the layer-based kernel in DoF rendering to the ray-based kernel to fit into volume rendering in NeRF. To decrease computational costs, we transition from using a traditional disk blur kernel to employing a sparse one [28] and optimize the rays originating from these sparse points. We use an MLP to predict both the 3D positions of the kernel points and their corresponding weights. Subsequently, the optimized rays and the estimated kernels serve as inputs for the aforementioned layered DoF volume rendering process. It is worth noting that our D^2RF enables the dynamic NeRF model to learn a sharp scene representation from defocused input views, leading to sharp novel views.

To improve the quality of reconstructed static regions, we utilize the blending technique [23] to independently predict the static and dynamic scenes. The blended color is then rendered with our layered DoF volume rendering pipeline. Furthermore, we adopt a cross-time rendering approach to represent temporal consistency in dynamic scenes. This approach models scene correlation across multiple input views and utilizes deformed rays for layered DoF volume rendering. Integrating scene information from other frames into the target frame enhances the representation of dynamic scenes.

To address the absence of existing dynamic NeRF defocus deblurring datasets, we collect and synthesize defocused video sequences from a stereo dataset [55] and conduct an extensive analysis to demonstrate D^2RF can effectively address defocus blur inputs and render high-quality sharp novel views. We compare with current dynamic NeRF methods as well as the combination of single-image deblurring and dynamic NeRF. Results show that D^2RF outperforms existing methods. Additionally, we conduct ablation studies to validate the effectiveness of each component. In summary, our main contributions are as follows.

- We present the first dynamic NeRF model D^2RF designed to recover sharp novel views from a defocused monocular video.
- We introduce layered DoF volume rendering in dynamic NeRF to model the defocus blur and restore a sharp representation.
- We transform the layered blur kernel to the ray-based kernel to fit into NeRF training and incorporate the optimized sparse kernel with the layered DoF volume rendering, enabling an efficient blurring process.

2 Related Work

Depth-of-Field Rendering. Depth-of-field (DoF) rendering, also known as bokeh rendering, generates realistic defocus blur in out-of-focus areas to enhance the visual prominence of the focused subject within an image. Physically based methods [1, 21] rely on 3D scene information and they are time-consuming. Recent methods use neural networks [12, 13] for end-to-end training. DeepFocus [58] entails an exact depth map to render realistic bokeh. However, the exact depth map is hard to acquire. Therefore, some methods [50, 53] instead predict depth for DoF rendering. To automate the rendering process, bokeh is directly rendered [11, 27] from an all-in-focus input without depth maps and controlling parameters. Synthesized data [36] from ray tracing is introduced for training to combine classical and neural rendering. To generate partial occlusion, some methods [37, 42] decompose the scene into multiple layers and separately blur each layer before compositing them, inspired by layered rendering [4, 61]. In this work, we show the connection between DoF rendering and NeRF rendering, integrating the layered blurring kernels with volume rendering.

Image Defocus Deblurring. To recover a sharp image from defocus blur, traditional methods conduct a two-stage pipeline: (1) estimate a defocus map [33, 43] from image priors, (2) the defocus map is used as guidance to deblur the image from non-blind deconvolution [6, 22]. With the current advances in neural networks, recent methods train end-to-end on datasets [2]. Modifications on convolutions [20, 46] are proposed to improve the deblurring performance. The Transformer architecture and a novel multi-scale feature extraction module [59] are introduced for high-resolution restoration. Additional inputs, such as dual pixel [2], light field [40], video sequences [3], and depth images [32] are utilized to guide defocus deblurring. Current methods do not take multi-view 3D geometry and dynamic scenes into account. We aim to explore defocus deblurring in dynamic NeRF synthesis, and our method integrates the DoF rendering into dynamic NeRF to synthesize defocus blur, which in turn facilitates deblurring.

Neural Radiance Field. Neural radiance field (NeRF) is an implicit representation [26, 30, 44, 45] for photo-realistic novel view synthesis, which aims to generate new camera perspectives given a specified set of input viewpoints. Different from explicit representations [7, 14], NeRF [31] learns an implicit continuous function to model the complex geometry and appearance of a scene. Subsequent studies aim to synthesize high-quality novel views from abnormal inputs. To recover a sharp NeRF from defocus-blurred inputs, Deblur-NeRF [28] proposes a deformable sparse kernel to model defocus blur, and DP-NeRF [19] adopts physical scene priors to mitigate blurring. DoF-NeRF [56] introduces the Concentrate-and-Scatter technique to explicitly enable controllable DoF effects. However, all these methods render DoF in a post-process manner and do not consider the volume sampling process. LensNeRF [17] restores a sharp NeRF from defocused static scene. The technique is valid for static scenes but does not model the temporal photometric consistency between video frames. Current methods are only effective in static scenes, where the objects are not moving. Dynamic NeRF synthesis with defocus blur remains unexplored.

Dynamic Neural Radiance Field. Recent works extend NeRF from static scenes to dynamic scenes [24, 57]. Given a monocular video, some approaches apply a deformation field on a canonical representation [34, 35, 38, 48, 52] for space-time novel view synthesis. Scene flow also works as a constraint to establish temporal consistency and model object motions in dynamic scenes [8, 23, 25, 51]. These methods require sharp inputs and fail when the defocus blur is in the input monocular video. Compared with existing methods, we synthesize novel views given the defocus blur. Furthermore, we propose to tackle the blur by introducing the layered DoF volume rendering in dynamic NeRF rendering.

3 Method

We propose D^2RF , a dynamic radiance field pipeline to recover sharp novel views from a defocused monocular video. In the following, we first briefly revisit the concept of NeRF and the principle of DoF rendering. In section 3.1, we analyze the connection and the distinction between DoF rendering and volume rendering in NeRF training, proposing layered DoF volume rendering with optimized ray-based blur kernels. Then we show how to represent dynamic scenes in a space-time manner, blend dynamic and static scenes with layered DoF volume rendering, and ensure temporal consistency within the scene in Section 3.2. Finally, we provide the training details in Section 3.3.

Preliminary: Neural Radiance Field Our method follows the principle of Neural Radiance Field (NeRF), which represents a static 3D scene as an implicit continuous function F_Θ . This function is parameterized by a multilayer perceptron (MLP) and it takes a spatial point location $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{R}^3$ as inputs. F_Θ maps the inputs to color \mathbf{c} and volume density σ :

$$(\mathbf{c}, \sigma) = F_\Theta(\mathbf{x}, \mathbf{d}). \quad (1)$$

To acquire the RGB value of a pixel coordinate p , NeRF first specifies a ray $\mathbf{r}_p(t) = \mathbf{o} + t\mathbf{d}_p$ emitted from camera projection center \mathbf{o} along the viewing direction \mathbf{d}_p . Then followed by the classical volume rendering technique [15], the color of $\mathbf{r}_p(t)$ can be represented as an integral of all colors along the ray, computed as:

$$\hat{C}(\mathbf{r}_p) = \sum_{i=1}^k T_i (1 - \exp(-\sigma \delta_i)) \mathbf{c}(\mathbf{r}_p(t_i), \mathbf{d}_p), \quad (2)$$

where

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma \delta_j\right) \quad (3)$$

is the accumulated transmittance along ray \mathbf{r}_p , and $\delta_i = t_{i+1} - t_i$ is the distance between the two sampled points. The vanilla NeRF assumes a pinhole camera model, where the color of a pixel $\hat{C}(\mathbf{r}_p)$ is calculated by points on a single ray $\mathbf{r}_p(t)$, so it fails to model the defocus blur.

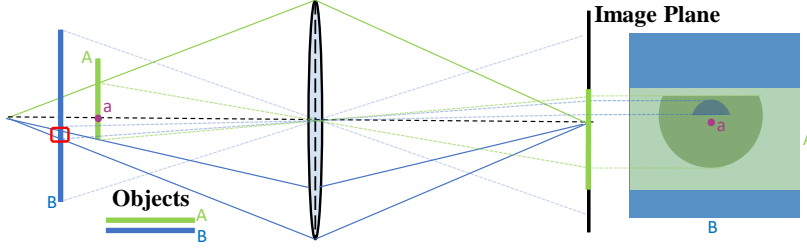


Fig. 2: The defocus blur formation. We model the defocus blur of the center purple pixel. The blur and green objects are from different layers. We define the light path as the left-right view and the image plane shows the front-back view of the defocus blur. The CoC of the purple point gathers its neighboring pixels on the green foreground and those on the blue background. This defocus modeling indicates that the object originally occluded (red square from the blue object) by the green object under the pinhole camera view can contribute to the rendered purple point by the green and blue semi-circles. In Eq.5-7 we model the defocus blur from layer visibility W_i , then use the link between visibility and opacity to integrate DoF and volume rendering in Eq.8. The visibility is learned from the layered DoF volume rendering pipeline.

Preliminary: Defocus Blur and DoF Rendering To model the defocus blur and render DoF, first we introduce bokeh rendering. Bokeh is the aesthetic quality of the blur in the out-of-focus region, caused by Circle of Confusion (CoC). Previous works [36, 50] have shown the CoC formation process. The radius γ of each pixel is

$$\gamma = af \left| \frac{1}{z} - \frac{1}{z_f} \right|, \quad (4)$$

where a is the camera aperture radius, f is the focal length. z is the depth of the pixel and z_f is the focal distance. Now that we have the blur radius of each pixel, we adopt layered bokeh rendering [37]. To blur the input image with kernels, classical bokeh rendering can be categorized into *gather* and *scatter* operations. We chose *gather* because it fits the layered bokeh rendering pipeline and is compatible with the volume rendering in NeRF described in section 3.1. We define the layer as a set of fronto-parallel planes at fixed depths by discretization. Now we have the blur kernel $K(\gamma_i)$ for layer i , we apply an alpha composition for DoF rendering. The rendered DoF image B is formulated as:

$$B = \frac{\psi(C_i)}{\psi(\mathbb{I})}, \quad (5)$$

where

$$\psi(X) = \sum_{i=1}^N \prod_{j=i+1}^N (1 - W_j * K(\gamma_j))(XW_i * K(\gamma_i)). \quad (6)$$

$\psi(X)$ is the compositing blurring function with the image X as input, $*$ is the convolution operation, N is the number of layers, \mathbb{I} is an image with all values

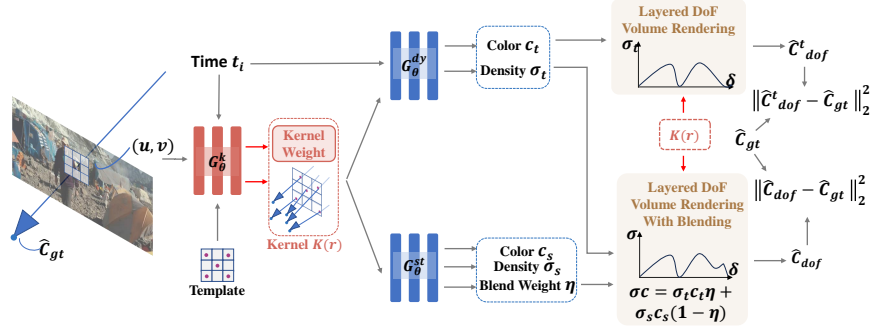


Fig. 3: Pipeline of our framework. The framework takes a set of plane coordinates (u, v) , the time embedding t_i , and defines a kernel template as inputs, the outputs are the blur kernels $K(r_i)$ consisting of the sparse optimized rays with their corresponding weights. The rays are then fed to the two MLPs G_{θ}^{st} and G_{θ}^{dy} to independently represent static and dynamic scenes. The final color is rendered by layered DoF volume rendering (Section 3.1), $\hat{C}_{dof}(\mathbf{r})$ is from Eq. 8 and $\hat{C}_{dof}^t(\mathbf{r})$ is from Eq. 13. The rendered defocused results (dynamic and blended) are supervised by the input defocused views. For testing we directly render the rays without layered DoF volume rendering and the kernel.

as 1. Each layer i encodes an RGB image C_i , with their corresponding visibility weight W_i , ensuring the correct blending of different layers. We show the defocus blur formation and how rays gather between layers in Fig. 2, demonstrating why we should apply the visibility term in DoF rendering. This layered visibility-aware rendering gathers the rays with inter-layer weights for each layer, then blends the layers with the weights from back to front.

3.1 Layered DoF Volume Rendering

Now that we have a basic understanding of NeRF and DoF rendering, we introduce our layered DoF volume rendering module. As shown in Fig. 3, we introduce layered DoF volume rendering, incorporating DoF rendering into volume rendering in dynamic NeRF. To seamlessly fuse Eq. 5 and Eq. 6 in NeRF training, we utilize the connections between the blur formation and volume rendering in Eq. 2. The volume rendering assumes that, for each ray $\mathbf{r}(t_i)$ from a scene, the sampling points in the volume all have an alpha value $\alpha_i = 1 - \exp(-\sigma\delta_i)$, which indicates the opacity of the point. A larger alpha value shows greater absorption of light, making it more likely for the light to pass through that point. In contrast, a smaller alpha value means less absorption of light at that point, and the object at that point is more transparent. In Eq. 6, the term W_i represents the visibility between the current layer i and all the layers behind it. This visibility W has the same physical meaning as the opacity value α in volume rendering. Furthermore, in NeRF rendering, the volume of the scene is sampled in a discretized manner, which typically involves uniformly sampling discrete points along the ray path. This layer discretization is compatible with the one in DoF rendering.

Therefore we first transform Eq. 5 from the format of layered images into the format of individual layered pixels:

$$B(\mathbf{r}) = \frac{\sum_{i=1}^k \left(\prod_{j=i+1}^k (1 - w_j * K(\mathbf{r})) w_i \mathbf{c}_i * K(\mathbf{r}) \right)}{\sum_{i=1}^k \left(\prod_{j=i+1}^k (1 - w_j * K(\mathbf{r})) w_i * K(\mathbf{r}) \right)}, \quad (7)$$

where \mathbf{c}_i and w_i are the colors and layer visibilities of the pixels surrounding \mathbf{r} within the blur kernel $K(\mathbf{r})$. Then we propose to integrate layered DoF volume rendering into dynamic NeRF training, the rendering can be formulated as:

$$\hat{C}_{dof}(\mathbf{r}_p) = \frac{\sum_{i=1}^k \left((T_i * K(\mathbf{r})) (1 - \exp(-\sigma \delta_i)) \mathbf{c}(\mathbf{r}(t_i), \mathbf{d}) * K(\mathbf{r}) \right)}{\sum_{i=1}^k \left((T_i * K(\mathbf{r})) (1 - \exp(-\sigma \delta_i)) * K(\mathbf{r}) \right)}, \quad (8)$$

where the variable name is consistent as Eq. 2 and Eq. 7, *e.g.* k is the number of sampling layers both in DoF rendering and volume rendering, \mathbf{c}_i is the emitted color of layer i blurred by $K(\mathbf{r})$.

Although DoF rendering and NeRF volume rendering share similar traits, there is a difference in their implementations, as the DoF rendering is originally used on different layers of the entire image, and NeRF training is optimized by rays sampled across the near and far planes. To implement the DoF rendering in Eq. 8, we switch the layer-based kernels $K(\gamma_i)$ to ray-based kernels $K(\mathbf{r})$, ensuring that the blur kernels are still diverse based on different positions. $K(\gamma_i)$ is defined by the divided layers, and $K(\mathbf{r})$ is defined by the input rays. Furthermore, to alleviate the high computational cost, we apply a sparse kernel following [28], which replaces the dense kernel with a smaller number of sparse points. The convolution process with $K(\mathbf{r}_p)$ can be formulated as:

$$\mathbf{b}_p = \sum_{j \in S(p)} \mathbf{c}_j g_j, \quad (9)$$

where \mathbf{b}_p is the result of point p after convolution, $S(p)$ represents the sparse points surrounding point p , g_j and \mathbf{c}_j are the corresponding weights and colors from the sparse points. We set $\sum_{j \in S(p)} g_j = 1$ to ensure the brightness consistency. Though we fix the number of sparse points, we additionally apply ray deformation as [28] to optimize the rays as the spatially varying real-world blur kernels. As shown in Fig. 3, we jointly optimize the translation of the center ray origin of the blur kernel. For a plane coordinate at the center of kernel (u, v) , we use an MLP G_θ^k to predict the deformation of the kernel points as well as the kernel weights:

$$(\Delta \mathbf{j}, g_j) = G_\theta^k((u, v), \mathbf{j}, t_l), \quad (10)$$

where t_l is the embedding of the input frame, \mathbf{j} are the original rays from the kernel template, g_j are the corresponding kernel weights, and the final optimized ray inputs are calculated as $\mathbf{r}_j = \mathbf{j} + \Delta \mathbf{j}$.

3.2 Dynamic Radiance Field

With the layered DoF volume rendering module in place, we turn to the dynamic scene representation. Given a monocular video of a dynamic scene, we take the video frames and camera parameters as inputs. Following previous dynamic NeRF methods, we represent the scene with an MLP G_θ^{dy} , which maps the input spatial point location \mathbf{x} , viewing direction \mathbf{d} , and time t , to the corresponding color, volume density, scene flow, and disocclusion weight:

$$(\mathbf{c}_t, \sigma_t, f_t, \mathcal{W}_t) = G_\theta^{\text{dy}}(\mathbf{x}, \mathbf{d}, t), \quad (11)$$

where G_θ^{dy} is parameterized by θ . The scene flow f_t and the disocclusion weight \mathcal{W}_t are used for the following cross-time rendering. Based on the representation, we next describe our twist on the dynamic radiance field.

Blending Dynamic and Static Scene. We use two MLPs to model static and dynamic scenes as shown in Fig. 3. One MLP is designed for modeling moving objects by various loss terms (cross-time *etc.*), and the other learns the static scene, allowing for more stable rendered still objects. To improve the quality of time-invariant static regions, we apply a blending technique as previous dynamic methods [23, 25], using an additional MLP to represent the static scene:

$$(\mathbf{c}, \sigma, \eta) = G_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}), \quad (12)$$

where η denotes the blending weight used to blend static and dynamic models. We denote the static color as \mathbf{c}_s , the dynamic color as \mathbf{c}_t , the static volume density and the dynamic volume density as σ_s and σ_t . As shown in Fig. 3, the layered DoF volume rendering with the blending weight is defined as:

$$\hat{C}_{\text{dof}}(\mathbf{r}) = \frac{\sum_{i=1}^k T_i^b \text{volume}_i^\epsilon}{\sum_{i=1}^k T_i^b \text{volume}_i^\epsilon}, \quad (13)$$

where $T_i^b = \exp(-\sum_{j=1}^{i-1} \sigma_s \sigma_t \delta_j) * K(\mathbf{r})$. volume_i^ϵ and volume_i^ϵ are rendered with $\eta(t)$ as weights:

$$\begin{aligned} \text{volume}_i^\epsilon &= \eta(t)(1 - \exp(-\sigma_s \delta_i)) \mathbf{c}_s(\mathbf{r}(t), \mathbf{d}) * K(\mathbf{r}) \\ &+ (1 - \eta(t))(1 - \exp(-\sigma_t \delta)) \mathbf{c}_t(\mathbf{r}(t), \mathbf{d}) * K(\mathbf{r}), \end{aligned} \quad (14)$$

and

$$\begin{aligned} \text{volume}_i^\epsilon &= \eta(t)(1 - \exp(-\sigma_s \delta_i)) * K(\mathbf{r}) \\ &+ (1 - \eta(t))(1 - \exp(-\sigma_t \delta)) * K(\mathbf{r}). \end{aligned} \quad (15)$$

Finally, we apply the rendering loss for the blending results:

$$\mathcal{L}_{\text{color}}^b = \|\hat{C}_{\text{dof}}(\mathbf{r}) - \hat{C}_{\text{gt}}(\mathbf{r})\|_2^2. \quad (16)$$

Apart from blending the static and dynamic scenes, we also constrain the rendered result of dynamic scene $\hat{C}_{\text{dof}}^t(\mathbf{r})$ by $\mathcal{L}_{\text{color}}^t = \|\hat{C}_{\text{dof}}^t(\mathbf{r}) - \hat{C}_{\text{gt}}(\mathbf{r})\|_2^2$.

Cross-time rendering. Suppose we only apply photometric consistency on individual time stamps like static NeRF. In that case, the model might fail to build temporal consistency in dynamic scenes, resulting in overfitting input views and failing to synthesize correct novel views. Therefore, temporal consistency between neighboring frames [8, 24] is crucial for training dynamic radiance fields. To establish temporal consistency between corresponding frames, we apply cross-time rendering under layered DoF volume rendering. For a timestamp t_m from the m -th frame, we assign the three-dimensional scene flow of a spatial point \mathbf{x} from t_m to timestamp t_n as $f_{\mathbf{x}}^m(n)$, where n denotes the neighbor frames of m . The corresponding point of \mathbf{x} at t_n can be computed as $\mathbf{x}^{m \rightarrow n} = \mathbf{x} + f_{\mathbf{x}}^m(n)$. This correspondence works on all the sampled points of a ray, therefore, we use the layered DoF volume rendering to acquire the color of the warped ray:

$$\hat{C}_{dof}^{n \rightarrow m}(\mathbf{r}) = \frac{\sum_{i=1}^k \left((T_i^n * K(r))(1 - \exp(-\sigma_n \delta_i)) \mathbf{c}_t(\mathbf{r}(t_i)^{m \rightarrow n}, \mathbf{d}) * K(r) \right)}{\sum_{i=1}^k \left((T_i^n * K(r))(1 - \exp(-\sigma_n \delta_i)) * K(r) \right)}, \quad (17)$$

where $T_i^n = \exp(-\sum_{j=1}^{i-1} \sigma_n \delta_j)$, and $\mathbf{r}(t_i)^{m \rightarrow n} = \mathbf{r}(t_i)^m + f_{\mathbf{p}}^m(n)$. Then we compare the rendering loss between $\hat{C}_{dof}^{n \rightarrow m}(\mathbf{r})$ and $\hat{C}_{gt}^m(\mathbf{r})$:

$$\mathcal{L}_{\text{cross}} = \sum_{n \in \mathcal{N}(m)} \mathbf{W}^{n \rightarrow m}(\mathbf{r}) \|\hat{C}_{dof}^{n \rightarrow m}(\mathbf{r}) - \hat{C}_{gt}^m(\mathbf{r})\|_2^2, \quad (18)$$

where $n \in \mathcal{N}(m)$ are the neighboring frames of the m -th frame, $\mathbf{W}^{n \rightarrow m}(\mathbf{r})$ is calculated from volume rendering the weight \mathcal{W}_t in Eq. 11. The principle of disocclusion weight comes from the ambiguity of scene flow. Although scene flow is effective in establishing pixel-level alignment, it performs poorly when the pixels in one image are occluded. Therefore, we need a confidence map to demonstrate whether the pixel is occluded and the flow is correct. A possible solution is to use forward-backward consistency check [5] to alleviate the problem. Here we let the network learn the weight \mathbf{W} in an unsupervised manner [23].

3.3 Optimization

Monocular reconstruction of complex dynamic scenes is highly ill-posed and prone to local minima during optimization [8, 23]. Therefore, we introduce data-driven priors $\mathcal{L}_{\text{data}}$ to supervise the scene geometry apart from the previous photometric constraint. Similar to previous works [23], $\mathcal{L}_{\text{data}}$ consists of (1) scale-shift invariant monocular depth loss and (2) scene flow consistency and L1 regularization. We use RAFT [47] and DPT [39] to obtain ground truth optical flow and depth for the input images.

Implementations. We set the max kernel size as 10, which denotes the sparse kernel points are inquired under a radius of 10. The number of kernel points is 5. We employ the Adam optimizer [18] to jointly optimize the static and

dynamic MLPs and the blur kernels. We use COLMAP [41] to estimate the camera intrinsics and extrinsics. The learning rate is 5×10^{-4} . We train each scene for 250k iterations, which takes around two days on a single NVIDIA RTX 3090 GPU. The rendering takes roughly 13 seconds for each 940×360 frame.

4 Experiments

There are no available datasets for defocused inputs in dynamic NeRF, so we curated 8 defocused dynamic scenes from an existing dataset VDW [55] for evaluation. VDW consists of sharp stereo image sequences and their corresponding disparity sequences. We then use the bokeh rendering pipeline in BokehMe [36] for defocus blur generation. Given the disparity sequences, the synthesized bokeh quality surpasses current classical rendering and neural rendering methods. We collected 8 dynamic scenes of 1880×720 suitable for our task from this dataset, the image is resized to 940×360 for training. To closely resemble actual capturing scenarios, unlike static blur dataset [28] with a random focusing scheme, we gradually adjust the focal distance along the scene disparity, mimicking the typical focusing process observed in videos. The aperture is fixed for each scene. We utilize the left-view defocused image sequences for training and the corresponding right-view sharp image sequences for evaluation. The 8 scenes are named Camp, Shop, Car, Mountain, Dining1, Dining2, Dock, Gate.

4.1 Baseline Comparisons

The baselines include two types: 1) state-of-the-art Dynamic NeRF, *e.g.* scene flow methods NSFF [23] and DVS [8], deformation-based method HyperNeRF [35], robust dynamic method RoDynRF [25] and 2) an image-space baseline that uses single-view image deblurring [46] then trains the dynamic NeRF with the deblurred inputs. As all the current dynamic methods are not designed for defocus blur, we choose the robust RoDynRF and the flow-based DVS to train with [46] for baseline comparison. We use the pre-trained deblurring model for fairness.

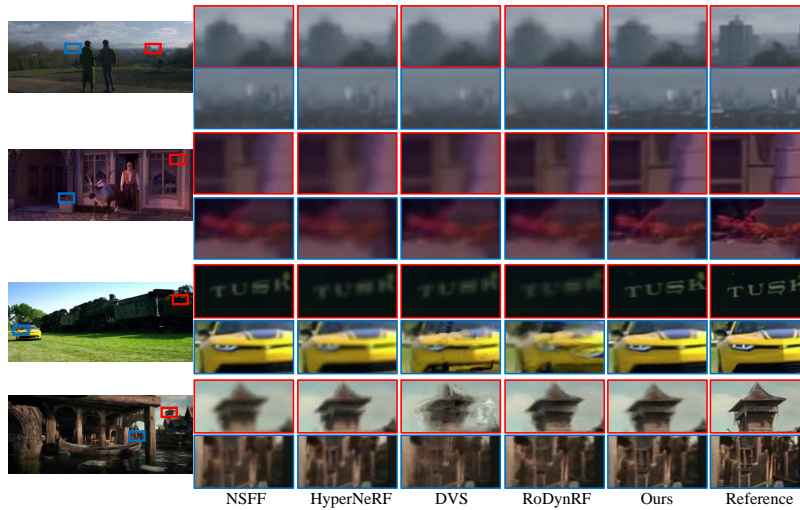
4.2 Qualitative Results

We show qualitative comparisons in Fig. 4 and Fig. 5. In Fig. 4, we compare our method with existing dynamic NeRF methods, and in Fig. 5, we compare our approach with the chosen dynamic NeRF models and their counterparts with 2D image-deblurring for preprocessing. We present the results on all scenes in these two figures to show our method’s effectiveness, 4 scenes for each.

One can observe: 1) current dynamic NeRF methods can not recover sharp details from the defocus blur; 2) the single-view 2D deblurring preprocessing [46] can somewhat recover sharp details in some regions, however, each view with the deblurred content lacks consistency across views, and the performance is still less stable and reliable than our method; 3) our method D^2RF outperforms these baselines in handling defocus blur and generating sharp novel views. please refer to the supplementary material for more results.

Table 1: Quantitative comparisons against all baselines. The best performance is **boldfaced**, and the second is underlined.

Methods	PSNR↑	SSIM↑	LPIPS↓	Methods	PSNR↑	SSIM↑	LPIPS↓
HyperNeRF [35]	26.96	0.780	<u>0.208</u>	RoDynRF [25]	<u>26.18</u>	0.770	0.227
RoDynRF [25]	26.18	0.770	0.227	DVS [8]	25.43	0.764	0.242
NSFF [23]	<u>27.01</u>	<u>0.803</u>	0.209	[46] + [25]	25.79	<u>0.776</u>	<u>0.196</u>
DVS [8]	25.43	0.764	0.242	[46] + [8]	24.52	0.757	0.208
D^2RF (Ours)	27.30	0.816	0.130	D^2RF (Ours)	27.30	0.816	0.130

**Fig. 4: 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 Mountain, Shop, Car, Dock.

4.3 Quantitative Results

We compare D^2RF with the 6 baseline methods quantitatively on our synthesized dataset. We use PSNR, SSIM, and LPIPS [60] as metrics for evaluation.

As shown in Table 1, D^2RF outperforms other state-of-the-art baselines. As mentioned above, to demonstrate the superiority of our method in modeling defocus blur in 3D space, we choose two dynamic NeRF baselines RoDynRF and DVS, and train these dynamic NeRF models while preprocessing the input blurry images with 2D defocus deblurring for comparison. This comparison highlights that, although the 2D deblurring method alleviates the defocus blur on each input view, this method struggles to maintain the consistency of scene information and tends to be unstable across all scenes. When dynamic NeRFs are trained on images deblurred with 2D methods, it results in undesirable artifacts



Fig. 5: The qualitative results with dynamic NeRF and their corresponding 2D image deblurring baselines. Although 2D image deblurring helps to alleviate the blur for dynamic NeRF in novel views (red box), our method is more stable and generates more reliable sharp details. The scenes are Camp, Dining1, Dining2, Gate.

in scene geometry and unsatisfying performance overall. In contrast, our method models defocus blur in 3D space across all input views, ensuring spatial-temporal consistency of scene geometry. We show the average results on all scenes, please refer to the supplementary material for separate results on all scenes.

4.4 Ablation Study

The main idea of D^2RF is to model the defocus blur by optimizing sparse blur kernels and incorporating the blur model into NeRF training. Therefore we conduct ablation studies on the layered DoF volume rendering module and the optimized kernels. Furthermore, we evaluate the effects of the static-dynamic blending scheme. As shown in Fig. 6 and Table 2, our framework works best when all the components are applied. We remove (1) the layered volume rendering module, and we use the kernel weight to directly gather the final rendered color as post-processing; (2) the optimized kernel, and we apply the full regular kernel and learn bokeh parameters; (3) static representation. We predict motion masks [23] to calculate the results on dynamic regions. We show that (1) the layered DoF volume rendering helps to model a more accurate defocus blur by integrating the blur into volume rendering, facilitating the recovery of sharp novel views; (2) the optimized kernel along with the layered DoF volume rendering provides an efficient and effective blurring pipeline for training; (3) the reconstruction of the static scene improves the performance because some

Table 2: Ablation study. The best performance is **boldfaced**, and the second is underlined. The left part calculates the results in dynamic regions, and the right part shows the results on the whole image.

Method (Dynamic)	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Method (All)	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
w/o cross-time	19.46	0.607	0.351	w/o cross-time	22.61	0.725	0.232
w/o layered volume	24.42	<u>0.738</u>	<u>0.170</u>	w/o layered volume	27.11	<u>0.811</u>	0.211
w/o optimized kernel	<u>24.54</u>	0.735	0.246	w/o optimized kernel	<u>27.25</u>	0.795	0.216
w/o static	23.93	0.719	0.179	w/o static	26.20	0.769	<u>0.177</u>
Full (Ours)	24.60	0.747	0.162	Full (Ours)	27.30	0.816	0.130



Fig. 6: The visualizations of the ablation study. The corresponding error map is visualized at the bottom, where darker regions indicate smaller errors.

regions are static and an extra scene representation helps to stabilize training. We choose two scenes for visualization and show average quantitative results on all scenes, please refer to the supplementary material for more results.

5 Conclusion

We present D^2RF , a novel framework for all-in-focus dynamic novel view synthesis from defocused monocular videos. To tackle defocus blur from video capture, we integrate the blur model into NeRF training. Particularly, we connect DoF rendering with volume rendering, proposing layered DoF volume rendering for dynamic NeRF. We modify the layered kernel to the ray-based kernel and apply an optimized sparse kernel to gather the rays. We conduct experiments to show our method can recover from defocus blur and produce satisfying results. Although this framework works well overall, some limitations remain, such as the inability to handle extreme defocus blur. We plan to address these issues in future work. We declare to release the source code upon acceptance of the paper.

References

1. Abadie, G., McAuley, S., Golubev, E., Hill, S., Lagarde, S.: Advances in real-time rendering in games. In: ACM SIGGRAPH 2018 Courses, pp. 1–1 (2018) [4](#)
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Proceedings of the European conference on computer vision (ECCV). pp. 111–126. Springer (2020) [4](#)
3. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2289–2298 (2021) [4](#)
4. 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) [4](#), [19](#)
5. Chen, Q., Koltun, V.: Full flow: Optical flow estimation by global optimization over regular grids. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4706–4714 (2016) [10](#)
6. Dong, J., Roth, S., Schiele, B.: Dwdn: deep wiener deconvolution network for non-blind image deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(12), 9960–9976 (2021) [4](#)
7. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 605–613 (2017) [4](#)
8. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV. pp. 5712–5721 (2021) [5](#), [10](#), [11](#), [12](#), [22](#)
9. 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) [20](#)
10. 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) [19](#)
11. Ignatov, A., Patel, J., Timofte, R.: Rendering natural camera bokeh effect with deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 418–419 (2020) [4](#)
12. Ignatov, A., Patel, J., Timofte, R., Zheng, B., Ye, X., Huang, L., Tian, X., Dutta, S., Purohit, K., Kandula, P., et al.: Aim 2019 challenge on bokeh effect synthesis: Methods and results. In: Proc. IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 3591–3598. IEEE (2019) [4](#)
13. Ignatov, A., Timofte, R., Qian, M., Qiao, C., Lin, J., Guo, Z., Li, C., Leng, C., Cheng, J., Peng, J., et al.: Aim 2020 challenge on rendering realistic bokeh. In: Proc. European Conference on Computer Vision Workshops (ECCVW). pp. 213–228. Springer (2020) [4](#)
14. Jimenez Rezende, D., Eslami, S., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3d structure from images. *Advances in neural information processing systems* **29**, 4996–5004 (2016) [4](#)
15. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. *ACM SIGGRAPH computer graphics* **18**(3), 165–174 (1984) [5](#)
16. 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) [19](#)

17. Kim, M.J., Gu, G., Choo, J.: Lensnerf: Rethinking volume rendering based on thin-lens camera model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3182–3191 (2024) [4](#)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [10](#)
19. Lee, D., Lee, M., Shin, C., Lee, S.: Dp-nerf: Deblurred neural radiance field with physical scene priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12386–12396 (June 2023) [2](#), [4](#)
20. Lee, J., Son, H., Rim, J., Cho, S., Lee, S.: Iterative filter adaptive network for single image defocus deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2034–2042 (2021) [4](#)
21. Lee, S., Eisemann, E., Seidel, H.P.: Real-time lens blur effects and focus control. ACM Transactions on Graphics (TOG) **29**(4), 1–7 (2010) [4](#)
22. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM transactions on graphics (TOG) **26**(3), 70-es (2007) [4](#)
23. 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) [1](#), [2](#), [3](#), [5](#), [9](#), [10](#), [11](#), [12](#), [13](#), [22](#)
24. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: CVPR. pp. 4273–4284 (2023) [1](#), [2](#), [5](#), [10](#)
25. 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) [5](#), [9](#), [11](#), [12](#), [22](#)
26. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Transactions on Graphics **38**(4CD), 65.1–65.14 (2019) [4](#)
27. Luo, X., Peng, J., Xian, K., Wu, Z., Cao, Z.: Defocus to focus: Photo-realistic bokeh rendering by fusing defocus and radiance priors. Information Fusion **89**, 320–335 (2023) [4](#)
28. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblurnerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022) [2](#), [3](#), [4](#), [8](#), [11](#)
29. 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) [19](#)
30. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019) [4](#)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 405–421. Springer (2020) [2](#), [4](#)
32. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2021) [4](#)

33. Park, J., Tai, Y.W., Cho, D., So Kweon, I.: A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1736–1745 (2017) [4](#)
34. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV. pp. 5865–5874 (2021) [5](#)
35. 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) [2](#), [5](#), [11](#), [12](#), [22](#)
36. 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) [4](#), [6](#), [11](#), [19](#)
37. Peng, J., Zhang, J., Luo, X., Lu, H., Xian, K., Cao, Z.: Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 590–607. Springer (2022) [4](#), [6](#)
38. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR. pp. 10318–10327 (2021) [1](#), [5](#)
39. 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) [10](#), [20](#)
40. Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16304–16313 (2022) [4](#)
41. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016) [11](#)
42. Sheng, Y., Yu, Z., Ling, L., Cao, Z., Zhang, C., Lu, X., Xian, K., Lin, H., Benes, B.: Dr. bokeh: Differentiable occlusion-aware bokeh rendering. arXiv preprint arXiv:2308.08843 (2023) [4](#)
43. Shi, J., Xu, L., Jia, J.: Just noticeable defocus blur detection and estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 657–665 (2015) [4](#)
44. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2437–2446 (2019) [4](#)
45. Sitzmann, V., Zollhofer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/b5dc4e5d9b495d0196f61d45b26ef33e-Paper.pdf> [4](#)
46. 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](#), [11](#), [12](#), [22](#)
47. 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) [10](#), [20](#)

48. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV. pp. 12959–12970 (2021) 5
49. 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) 19
50. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (TOG)* **37**(4), 1–13 (2018) 4, 6
51. Wang, C., Eckart, B., Lucey, S., Gallo, O.: Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994* (2021) 5
52. Wang, C., MacDonald, L.E., Jeni, L.A., Lucey, S.: Flow supervision for deformable nerf. In: CVPR. pp. 21128–21137 (2023) 5
53. Wang, L., Shen, X., Zhang, J., Wang, O., Lin, Z., Hsieh, C.Y., Kong, S., Lu, H.: Deeplens: Shallow depth of field from a single image. *ACM Transactions on Graphics (TOG)* **37**(6), 1–11 (2018) 4
54. Wang, Y., Yang, S., Hu, Y., Zhang, J.: Nerfocus: Neural radiance field for 3d synthetic defocus. *arXiv preprint arXiv:2203.05189* (2022) 19
55. 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) 3, 11, 19
56. Wu, Z., Li, X., Peng, J., Lu, H., Cao, Z., Zhong, W.: Dof-nerf: Depth-of-field meets neural radiance fields. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1718–1729 (2022) 3, 4
57. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: CVPR. pp. 9421–9431 (2021) 5
58. Xiao, L., Kaplanyan, A., Fix, A., Chapman, M., Lanman, D.: Deepfocus: Learned image synthesis for computational displays. *ACM Transactions on Graphics (TOG)* **37**(6), 1–13 (2018) 4
59. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022) 4
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018) 12
61. 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) 4, 19

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.

6 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 [29] and NeRFocus [54] 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.

[16] is an unsupervised generative model, the training images are **single** images. [16] 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.

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

7 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 [49] (MPI). The RGB image C with its visibility weight W is divided into layers by depth discretization. For current layered rendering [4, 61], 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.

8 The Dataset

We conduct our experiments on dynamic scenes from a stereo dataset VDW [55]. 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 [36]. 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 [39] to obtain depth maps, and we utilize RAFT [47] to generate optical flows, we also employ an instance segmentation network (Mask R-CNN [9]) to obtain motion masks for moving objects.



Fig. 7: The failure case. When the input views have severe defocus blur, the novel views may be unable to recover sharp details.

9 Failure Case

As shown in Fig. 7, 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. 3). 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 3: Calculation cost.

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

10 Experiments

We show the quantitative results with state-of-the-art methods on all scenes in Tab. 4. We also collect one real dynamic defocused video for evaluation. As shown in Fig. 8, our method is more stable and generates more reliable sharp details. The ablation results are in Tab. 5 and Tab. 6. We show qualitative comparisons with state-of-the-art methods in Figs. 9 and 10. We choose different scenes from the main article to show results on all scenes. We visualize the ablation results in Fig. 11.

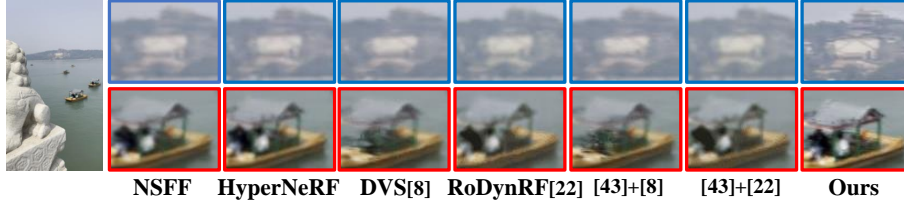
**Fig. 8:** Results on real defocused video.

Table 4: 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 \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DVS [8]	17.18	0.525	0.399	27.05	0.842	0.216	24.57	0.795	0.240	33.54	<u>0.896</u>	0.174
NSFF [23]	21.17	<u>0.643</u>	0.310	26.54	0.838	0.207	<u>25.57</u>	<u>0.799</u>	0.223	32.11	0.891	0.158
HyperNeRF [35]	<u>21.08</u>	0.627	<u>0.294</u>	27.21	0.834	0.216	25.44	0.769	0.216	<u>33.02</u>	0.886	0.160
RoDynRF [25]	20.99	0.597	0.312	28.53	0.844	0.213	22.62	0.726	0.264	28.71	0.858	0.194
[46] + DVS	17.07	0.531	0.352	26.47	0.850	<u>0.155</u>	23.72	0.786	<u>0.171</u>	31.87	0.900	<u>0.128</u>
[46] + RoDynRF	20.53	0.564	0.345	<u>28.15</u>	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	<u>0.854</u>	0.117	26.79	0.852	0.123	32.44	0.900	0.079
Method	Dining1			Dining2			Dock			Gate		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DVS [8]	23.07	0.712	0.287	27.85	<u>0.822</u>	0.165	27.54	0.807	0.220	22.60	0.713	0.231
NSFF [23]	<u>30.54</u>	<u>0.865</u>	<u>0.175</u>	27.54	0.812	<u>0.152</u>	28.42	<u>0.822</u>	0.216	<u>24.17</u>	<u>0.754</u>	0.228
HyperNeRF [35]	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 [25]	29.21	0.803	0.243	28.25	0.816	0.171	27.07	0.789	0.214	24.06	0.726	0.201
[46] + 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
[46] + RoDynRF	28.96	0.812	0.205	<u>27.92</u>	0.815	0.156	26.83	0.797	<u>0.178</u>	23.85	0.741	0.182
D^2RF (Ours)	31.05	0.877	0.105	27.84	0.817	0.094	<u>28.12</u>	0.823	0.131	24.41	0.757	<u>0.185</u>

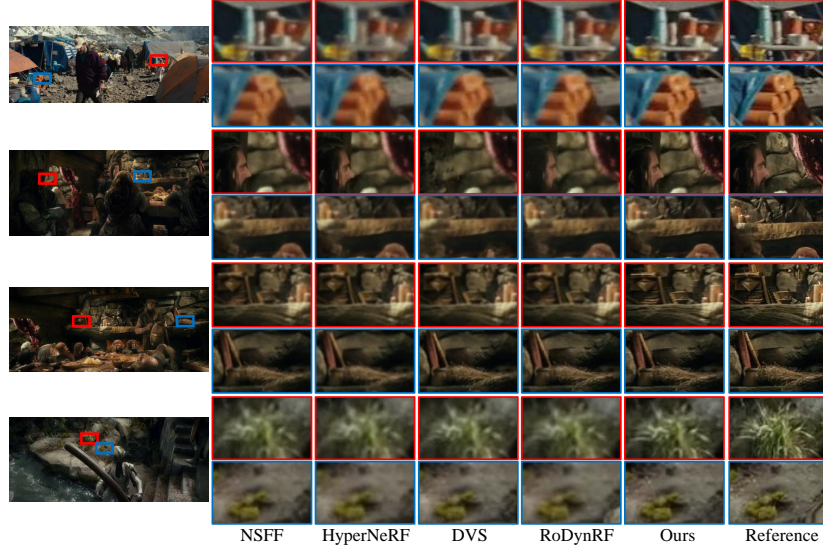


Fig. 9: The qualitative results with all dynamic NeRF baselines. Compared with existing dynamic NeRF methods, our method generates sharper novel views that are more faithful and have more details. The scenes are Camp, Dining1, Dining2, Gate.

Table 5: 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	<u>20.89</u>	<u>0.664</u>	<u>0.245</u>	19.65	0.583	0.245	24.34	0.886	<u>0.077</u>	26.09	<u>0.818</u>	<u>0.125</u>
w/o optimized kernel	20.91	0.641	0.331	<u>19.76</u>	<u>0.592</u>	0.256	<u>25.04</u>	<u>0.888</u>	0.130	25.90	0.796	0.196
w/o static	20.68	0.658	0.252	19.42	0.567	<u>0.221</u>	22.96	0.867	0.085	25.31	0.780	<u>0.125</u>
Full (Ours)	20.70	0.665	0.220	19.93	0.595	0.203	25.18	0.901	0.075	<u>25.95</u>	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	<u>30.99</u>	<u>0.866</u>	<u>0.099</u>	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	<u>25.66</u>	<u>0.733</u>	0.221	<u>27.47</u>	0.767	0.219	<u>20.97</u>	0.615	0.472
w/o static	30.59	0.850	0.141	24.69	0.697	0.144	27.65	<u>0.775</u>	0.180	20.19	0.558	0.280
Full (Ours)	31.01	0.871	0.085	25.76	0.745	<u>0.138</u>	27.24	0.774	<u>0.150</u>	21.03	<u>0.603</u>	0.316

Table 6: 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	<u>0.853</u>	0.180	26.72	0.810	0.218	31.86	0.881	0.157
w/o static	<u>20.77</u>	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	<u>26.79</u>	0.852	0.123	<u>32.44</u>	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	<u>30.94</u>	<u>0.872</u>	<u>0.118</u>	27.45	0.806	<u>0.098</u>	<u>28.24</u>	<u>0.833</u>	0.118	<u>24.26</u>	<u>0.749</u>	<u>0.189</u>
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	<u>27.85</u>	<u>0.817</u>	0.094	28.12	0.823	<u>0.131</u>	24.41	0.757	0.185

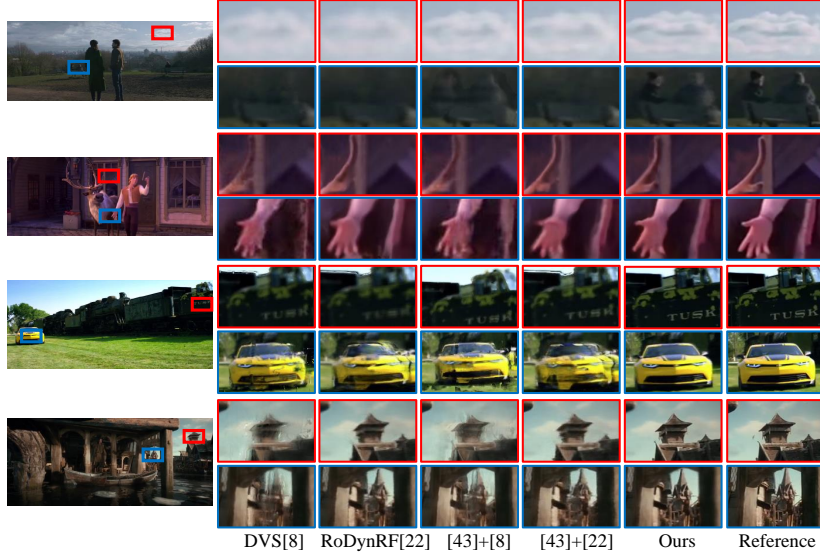


Fig. 10: The qualitative results with dynamic NeRF and their corresponding 2D image deblurring baselines. The scenes are Mountain, Shop, Car, Dock.

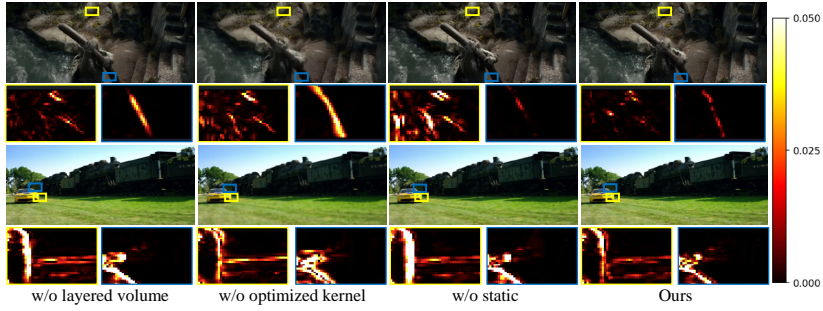


Fig. 11: The visualizations of the ablation study. The corresponding error map is visualized at the bottom, where darker regions indicate smaller errors. We define the error range from 0 to 0.05. Our full model has the smallest error overall.