

DyBluRF: Dynamic Neural Radiance Fields from Blurry Monocular Video

Huiqiang Sun¹ Xingyi Li¹ Liao Shen¹ Xinyi Ye¹ Ke Xian² Zhiguo Cao^{1*}

¹School of AIA, Huazhong University of Science and Technology

²School of EIC, Huazhong University of Science and Technology

{shq1031,xingyi.li,leoshen,xinyiye,kxian,zgcao}@hust.edu.cn

<https://huiqiang-sun.github.io/dyblurf>

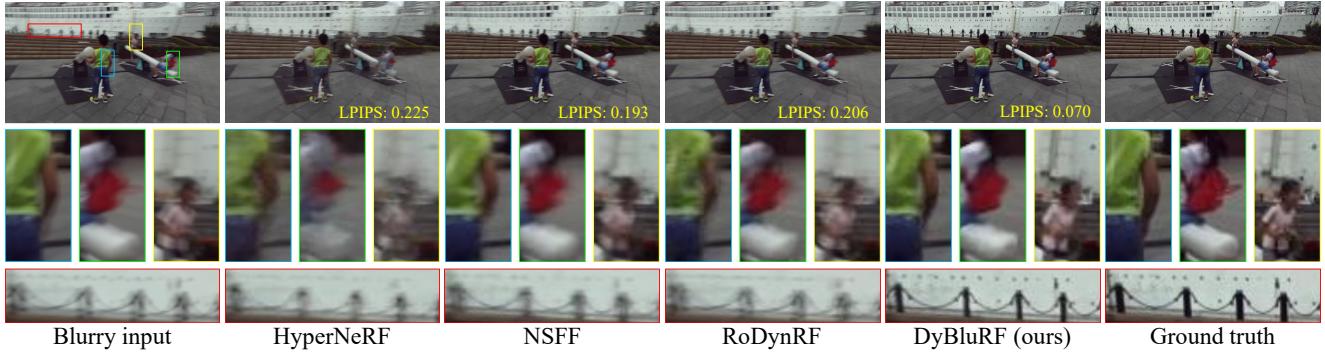


Figure 1. Given a monocular video capturing a dynamic scene with motion blur, our proposed method, DyBluRF, effectively synthesizes high-quality and sharp novel views compared to previous dynamic NeRF approaches that often yield low-quality and blurry results.

Abstract

Recent advancements in dynamic neural radiance field methods have yielded remarkable outcomes. However, these approaches rely on the assumption of sharp input images. When faced with motion blur, existing dynamic NeRF methods often struggle to generate high-quality novel views. In this paper, we propose DyBluRF, a dynamic radiance field approach that synthesizes sharp novel views from a monocular video affected by motion blur. To account for motion blur in input images, we simultaneously capture the camera trajectory and object Discrete Cosine Transform (DCT) trajectories within the scene. Additionally, we employ a global cross-time rendering approach to ensure consistent temporal coherence across the entire scene. We curate a dataset comprising diverse dynamic scenes that are specifically tailored for our task. Experimental results on our dataset demonstrate that our method outperforms existing approaches in generating sharp novel views from motion-blurred inputs while maintaining spatial-temporal consistency of the scene.

1. Introduction

The research on novel view synthesis for dynamic scenes has gained substantial importance in the community, with practical applications in a wide range of AR/VR contexts. Recent breakthroughs can be attributed to Neural Radiance Fields (NeRFs) [28]. NeRFs have made a profound impact on novel view synthesis, and recent years have witnessed rapid advancements in NeRF, including several works dedicated to dynamic scene representation [10, 22, 25, 34, 36]. Despite producing high-quality novel views, these dynamic NeRF methods may exhibit significant performance degradation when presented with inputs containing motion blur, as shown in Fig. 1. Recently, some NeRF-based methods have emerged to tackle the motion deblurring of static scenes [19, 20, 26, 48]. However, these methods cannot address blurring caused by object motion and lack the capability to effectively represent dynamic scenes.

To the best of our knowledge, there is no NeRF-based method dedicated to dynamic scene deblurring. Yet, in real-life video captures of dynamic scenarios, extended camera exposure time often introduces motion blur. Consequently, obtaining a sharp dynamic radiance field from inputs affected by motion blur remains a pressing problem. Existing dynamic NeRF methods heavily rely on sharp inputs

*Corresponding author.

to accurately represent dynamic scenes. When motion blur is present, the modulation of inter-frame relationships becomes unreliable, leading to difficulties in representing object motion and maintaining temporal consistency in dynamic scenes. Moreover, the presence of blur in input data degrades certain data priors like depth and optical flow, leading to inaccurate constraining of scene geometry. These challenges significantly hinder the construction of sharp dynamic NeRF representations from blurry inputs.

In this work, we propose DyBluRF, a model specifically designed to take monocular videos of dynamic scenes with motion blur as input and generate sharp novel views. Motion blur in dynamic scenes typically results from both camera and object movements. To model camera motion, drawing inspiration from BAD-NeRF [48], we evenly discretize the exposure time into multiple timestamps and learn camera poses at these timestamps. For object motion representation, establishing inter-frame relationships, such as directly predicting 3D scene flow, is unreliable under conditions of blurry input, thus failing to accurately present dynamic scenes. Instead, we employ an MLP to predict the global DCT trajectories of scene objects [45], thereby simulating object motion blur by expressing the DCT trajectories within an exposure time.

To enhance the temporal consistency representation in dynamic scenes, we adopt a cross-time rendering approach based on the input blurry images. This method uses the predicted DCT trajectory to model scene correlation across multiple input views and employs deformed rays for rendering by integrating scene information from other frames into the target frame. To ensure a globally consistent temporal expression, cross-time rendering is not limited to adjacent frames but instead performed across the entire temporal range. At the initial stages of training, frames close to the target frame are selected for cross-time rendering. As the training progresses, frames further away from the target frame are included until encompass the entire temporal range. This cross-time rendering approach further strengthens the representation of temporal consistency.

Furthermore, it is not reasonable to directly use depth and optical flow maps generated from blurry images to constrain the model, as the predictions of depth and optical flow can become inaccurate due to blurriness. Therefore, we introduce data-driven priors to mitigate the impact of inaccuracies in depth and optical flow from blurry images.

Given the absence of existing dynamic scene blurring datasets based on NeRF, we curate a collection of dynamic scenes sourced from the Stereo Blur Dataset [60]. We use these scenes for comparative analysis of our approach against existing dynamic NeRF methods, demonstrating its effective handling of input images with motion blur and generation of high-quality novel views. We also preprocess the input blurry images using state-of-the-art single-image

deblurring and dynamic scene video deblurring methods before feeding them into existing dynamic NeRF methods. Experimental results demonstrate that our method still outperforms existing dynamic NeRF methods. Additionally, we conduct ablation studies to validate the effectiveness of each component of our approach. In summary, our key contributions include:

- We introduce DyBluRF, the first dynamic NeRF model specifically designed to effectively tackle motion-blurred monocular video.
- We propose a dynamic scene representation method based on DCT, simulating the motion blur induced by object movement along the DCT trajectory.
- We present a cross-time rendering method that ensures temporal consistency by modeling scene relationships across different exposure times.
- We propose an effective approach that incorporates depth maps and optical flow derived from blurry images to impose constraints on NeRF.

2. Related Work

Image Deblurring. The main goal of image deblurring is to restore a sharp image from a blurry one. From the perspective of input data types, current motion deblurring methods can be categorized into single-image deblurring and multi-image/video deblurring. In the case of single-image deblurring, Some traditional methods employ unified probabilistic model [39], normalized gradient sparsity [16], or model rotational velocity of the camera during exposure time [50] to mitigate motion blur. Recently, numerous deep learning-based image deblurring methods have emerged [17, 18, 30, 42, 55], accompanied by the introduction of several datasets dedicated to image deblurring [40]. These methods often achieve superior results compared to traditional techniques.

In contrast to single-image deblurring, video deblurring requires establishing temporal consistency among different frames to ensure temporal smoothness. Some approaches employ optical flow information to establish inter-frame relationships [13, 32]. However, the optical flow in blurry videos often lacks accuracy. With the advance of deep learning, many techniques based on CNNs [41], RNNs [58], or attention mechanisms [56] have been proposed to address the problem of video deblurring. While these methods yield excellent results, they predominantly operate in image space and cannot capture scene geometry information in 3D space. Our approach leverages NeRF to represent 3D information of dynamic scenes implicitly and establishes spatial-temporal consistency to achieve deblurring.

Neural Radiance Field. NeRF [28] has demonstrated impressive results in novel view synthesis, and subsequent works have aimed to enhance it from various perspectives, including fundamental enhancements [2, 3], fast training

and rendering [6, 7, 9, 29], pose estimation [4, 24, 49], or sparse viewpoint input [52, 54]. Since vanilla NeRF imposes high requirements on input image quality, several follow-up methods have explored generating high-quality novel views in degraded input conditions, such as in-the-wild images [27], input images with burst noise [35], or low dynamic range (LDR) inputs with varying exposures [12]. Recently, several NeRF-based approaches have been devised to handle blurry inputs. Deblur-NeRF [26] and DP-NeRF [19] restore sharp radiance fields by modeling a blur kernel from various blur types. BAD-NeRF [48], which is most relevant to our work, models motion blur by estimating camera trajectory. ExBluRF [20] leverages a voxel-based radiance field to handle extreme motion blur. While these methods can remove blur from input images and generate high-quality novel views, they cannot handle dynamic scenes. Our approach can effectively reduce motion blur caused by both camera and object motion, generating novel views of dynamic scenes.

Dynamic Neural Radiance Field. Many recent efforts have extended NeRF to represent dynamic scenes. These methods input multi-view videos [21] or a monocular video [23, 51] of dynamic scenes to achieve space-time novel view synthesis. To ensure temporal consistency, some approaches represent dynamic scenes by deforming the canonical space points [33, 34, 36, 44, 47]. Other methods leverage optical flow constraints to establish scene coherence between adjacent frames [10, 22, 25, 46]. In particular, NSFF [22] employs neural scene flow fields to represent complex object motions in dynamic scenes; RoDynRF [25] jointly learns camera parameters to enhance model robustness. Additionally, many methods are currently dedicated to fast training and rendering of dynamic NeRF [1, 5, 8]. However, these dynamic NeRF methods require sharp input images and often struggle to synthesize reliable novel views in the presence of motion blur in the input images.

3. Method

We propose DyBluRF, a dynamic radiance field approach synthesizing sharp novel views from a monocular video affected by motion blur. The overall pipeline is illustrated in Fig. 2. In this section, we first provide background information on neural radiance fields and volume rendering in Section 3.1. We then elaborate on modeling genuine physical image formation process of motion blur in Section 3.2. Next, Section 3.3 outlines our approach for representing dynamic scenes and ensuring temporal consistency within the scene. Finally, we provide details of the training and optimization process in Section 3.4.

3.1. Preliminary: Neural Radiance Field

We begin by describing Neural Radiance Fields (NeRFs) that our method builds upon. NeRF represents a scene as an

implicit function that inputs a spatial point location $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{R}^3$, and outputs corresponding volume density σ and color \mathbf{c} :

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

To enable learning scene details, point location \mathbf{x} and viewing direction \mathbf{d} are transformed into a high-dimensional space through positional embedding $\gamma(\cdot)$ before sending to NeRF. We also employ this trick in our methods but omit to mention the position encoding in our writing for clarity.

NeRF adopts classical volume rendering [14] to render RGB images from any given camera pose. For a ray \mathbf{r} emitted from camera center through a given pixel on the image plane, its color is computed as:

$$\mathbf{C}(\mathbf{r}) = \int_{s_n}^{s_f} T(s) \sigma(\mathbf{r}(s)) \mathbf{c}(\mathbf{r}(s), \mathbf{d}) ds, \quad (2)$$

where

$$T(s) = \exp \left(- \int_{s_n}^s \sigma(\mathbf{r}(m)) dm \right) \quad (3)$$

is accumulated transmittance along ray \mathbf{r} which denotes the probability that the ray will not hit any object from s_n to s .

3.2. Motion Blur Formulation

The physical process of motion blur can be considered as the result of camera or object motion within exposure time τ . The mathematical modeling of the motion blur generation process can be written as:

$$\mathbf{B}(\mathbf{x}) = g \int_0^\tau \mathbf{C}_t(\mathbf{x}) dt, \quad (4)$$

where $\mathbf{C}_t(\mathbf{x})$ denotes the sharp image captured at time t , $\mathbf{B}(\mathbf{x})$ is the blurry image, g is a normalization factor. To approximate the integral operation, we discretize the exposure time into n timestamps uniformly and consider the captured blurry image as the average of n sharp images during the exposure time:

$$\mathbf{B}(\mathbf{x}) \approx \frac{1}{n} \sum_{l=1}^n \mathbf{C}_l(\mathbf{x}). \quad (5)$$

In theory, as n increases, the simulated blur generation process becomes more close to a genuine one.

According to Eq. 5, generating a blurry image demands obtaining sharp images at each timestamp within the exposure time. Therefore, it becomes crucial to model both camera and object motion during the exposure time. For the object motion representation, we will provide a detailed explanation in Section 3.3. For the camera motion, we adopt a strategy that jointly learns camera parameters and NeRF. Specifically, given N input blurry images, we initialize the

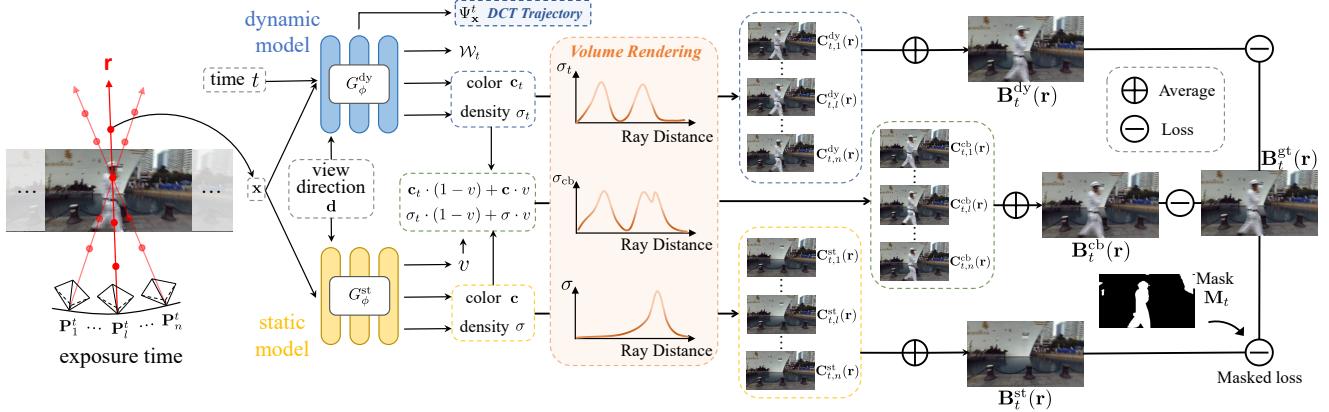


Figure 2. Overall pipeline. To model the motion blur of input images, we discretize the exposure time into n timestamps. Subsequently, we perform ray sampling for the same pixel based on the camera poses for each timestamp. We employ two MLPs to represent dynamic scenes. The dynamic model takes spatial point coordinates \mathbf{x} , viewing direction \mathbf{d} , and time t as inputs and predicts color \mathbf{c}_t , volume density σ_t , DCT coefficients $\Psi_{\mathbf{x}}^t$, and disocclusion weights \mathcal{W}_t . The static model only takes \mathbf{x} and \mathbf{d} as inputs and predicts color \mathbf{c} , volume density σ , and a blending weight v for blending static and dynamic results. After obtaining colors and volume densities for static, dynamic, and blended scenes, we use volume rendering to compute pixel RGB values $\mathbf{C}(\mathbf{r})$. We average the RGB values for the same pixel within the exposure time to obtain the predicted blurry image, and calculate losses with the input blurry image. For dynamic and blended images, losses are directly computed against the input image. For the static loss, we use a mask image to calculate the static regions only.

camera poses \mathbf{P}_1^i and \mathbf{P}_n^i at the beginning and end timestamps of exposure time i , where $i \in \{1, \dots, N\}$. Leveraging the short exposure time, we utilize linear interpolation to obtain camera poses for intermediate timestamps. Following the ideas of BAD-NeRF [48], we employ Lie-algebra of $SE(3)$ to interpolate the intermediate camera poses \mathbf{P}_l^i , $l \in \{2, \dots, n-1\}$.

3.3. Temporal Consistency Modeling

Dynamic radiance field. Similar to existing dynamic NeRF methods, we employ an MLP to represent dynamic scenes. This MLP takes the input spatial point location \mathbf{x} , viewing direction \mathbf{d} , and time t , and outputs the corresponding color and volume density: $G_{\phi}^{\text{dy}} : (\mathbf{c}_t, \sigma_t) = G_{\phi}^{\text{dy}}(\mathbf{x}, \mathbf{d}, t)$, where G_{ϕ}^{dy} indicates the MLP parameterized by ϕ . Handling a monocular video where scene information at different times is limited to individual frames poses an ill-posed problem. Establishing temporal consistency in dynamic scenes becomes essential to resolve this issue. Although employing scene flow information between adjacent frames is a direct approach [22, 25], accurately determining scene motion within the exposure time using adjacent blurry images remains challenging. To enforce a more reliable temporal consistency, we establish relationships across global time intervals rather than relying solely on adjacent frames. Leveraging the efficiency of the Discrete Cosine Transform (DCT) in representing complex motion trajectories in a smooth and concise manner [45, 46], we choose to model the object DCT trajectory to express dynamic scenes.

Given a sequence of N blurry images, the exposure time

for each image can be divided into n timestamps, resulting in a total of $n \times N$ timestamps. The trajectory with DCT coefficients $\Psi \in \mathbb{R}^{3K}$ can be expressed as:

$$\mathcal{T}(t) = \sqrt{\frac{2}{nN}} \sum_{k=1}^K \Psi(k) \cos\left(\frac{\pi}{2nN}(2t+1)k\right), \quad (6)$$

where $\mathcal{T} \in \mathbb{R}^3$ represents the time-dependent trajectory, $\Psi(k) \in \mathbb{R}^3$ denotes the k -th coefficient in 3D space, K is a hyperparameter. According to Eq. 6, for a 3D point \mathbf{x} at time t , its global time motion trajectory $\mathcal{T}_{\mathbf{x}}^t$ can be modeled using the corresponding DCT coefficient $\Psi_{\mathbf{x}}^t$, and $\mathcal{T}_{\mathbf{x}}^t(t')$ represents its 3D coordinates at time t' . We use the MLP to predict the DCT coefficients. Thus, the dynamic radiance field can be rewritten as $G_{\phi}^{\text{dy}} : (\mathbf{c}_t, \sigma_t, \Psi_{\mathbf{x}}^t) = G_{\phi}^{\text{dy}}(\mathbf{x}, \mathbf{d}, t)$.

To handle occlusion in dynamic scenes, we adopt a similar approach to [22], where the MLP additionally outputs a disocclusion weight $\mathcal{W}_t = (w_{\text{fw}}, w_{\text{bw}})$. However, unlike [22], our method predicts disocclusion weights to represent the occlusion possibility to the corresponding timestamps within the subsequent or preceding exposure time. In summary, our dynamic radiance fields G_{ϕ}^{dy} can be expressed as:

$$(\mathbf{c}_t, \sigma_t, \Psi_{\mathbf{x}}^t, \mathcal{W}_t) = G_{\phi}^{\text{dy}}(\mathbf{x}, \mathbf{d}, t). \quad (7)$$

To supervise the model using input images, we need to obtain the blurry image at each frame through the model. For the i -th blurry image $\mathbf{B}_i(\mathbf{r})$, we calculate the timestamps t_l^i , $l \in \{1, \dots, n\}$ within the i -th camera exposure time. As per Section 3.2, the camera pose for each timestamp is denoted as \mathbf{P}_l^i . Using the MLP depicted in Eq. 7

and the volume rendering from Eq. 2, we derive predicted sharp images $\mathbf{C}_{i,l}^{\text{dy}}(\mathbf{r})$ of timestamps l within the i -th exposure time. Then, we average these sharp images to obtain the predicted blurry image $\mathbf{B}_i^{\text{dy}}(\mathbf{r})$:

$$\mathbf{B}_i^{\text{dy}}(\mathbf{r}) = \frac{1}{n} \sum_{l=1}^n \mathbf{C}_{i,l}^{\text{dy}}(\mathbf{r}), \quad i \in \{1, \dots, N\}, \quad (8)$$

and calculate rendering loss with input blurry image $\mathbf{B}_i^{\text{gt}}(\mathbf{r})$:

$$\mathcal{L}_{\text{RGB}}^{\text{dy}} = \|\mathbf{B}_i^{\text{dy}}(\mathbf{r}) - \mathbf{B}_i^{\text{gt}}(\mathbf{r})\|_2^2. \quad (9)$$

Static model integration. Similar to many previous dynamic NeRF methods [22, 25], we also employ an additional MLP to enhance the representation of static regions:

$$(\mathbf{c}, \sigma, v) = G_{\phi}^{\text{st}}(\mathbf{x}, \mathbf{d}), \quad (10)$$

where v denotes the blending weight used to combine the results from static and dynamic models:

$$\begin{aligned} \mathbf{c}_{\text{combine}} &= \mathbf{c}_t \cdot (1 - v) + \mathbf{c} \cdot v, \\ \sigma_{\text{combine}} &= \sigma_t \cdot (1 - v) + \sigma \cdot v. \end{aligned} \quad (11)$$

We utilize Eq. 2 to obtain static sharp images $\mathbf{C}_{i,l}^{\text{st}}(\mathbf{r})$ and combined sharp images $\mathbf{C}_{i,l}^{\text{cb}}(\mathbf{r})$, and we average these images in an exposure time to obtain the predicted static blurry image $\mathbf{B}_i^{\text{st}}(\mathbf{r})$ and combined blurry image $\mathbf{B}_i^{\text{cb}}(\mathbf{r})$. Subsequently, we apply rendering loss from both static and combined results to constrain the MLPs:

$$\begin{aligned} \mathcal{L}_{\text{RGB}}^{\text{cb}} &= \|\mathbf{B}_i^{\text{cb}}(\mathbf{r}) - \mathbf{B}_i^{\text{gt}}(\mathbf{r})\|_2^2, \\ \mathcal{L}_{\text{RGB}}^{\text{st}} &= (\mathbf{1} - \mathbf{M}_i) \cdot \|\mathbf{B}_i^{\text{st}}(\mathbf{r}) - \mathbf{B}_i^{\text{gt}}(\mathbf{r})\|_2^2. \end{aligned} \quad (12)$$

where \mathbf{M}_i denotes the motion mask of input frame i . The inclusion of static model enhances the ability to represent static scenes.

Cross-time rendering. If we only rely on Eq. 9 for reconstruction, the model may inadequately capture the temporal consistency in dynamic scenes, potentially leading to overfitting to input images and failing to produce high-quality novel views. To mitigate this issue, we utilize DCT trajectories to establish consistency between the i -th frame and other frames, as shown in Fig. 3. Establishing temporal consistency between neighboring frames plays a critical role in the dynamic radiance field due to their strong correlation [10, 23]. However, blurry images relate to multiple images within the exposure time rather than a single image. Therefore, we express temporal consistency by establishing relationships between corresponding timestamps across different exposure times. Specifically, for a timestamp t_l^i in an exposure time from frame i , the scene flow of a spatial point \mathbf{x} from t_l^i to timestamp t_l^j can be represented as:

$$f_{\mathbf{x}}^{i,l}(j) = \mathcal{T}_{\mathbf{x}}^{t_l^i}(t_l^j) - \mathcal{T}_{\mathbf{x}}^{t_l^i}(t_l^i), \quad (13)$$

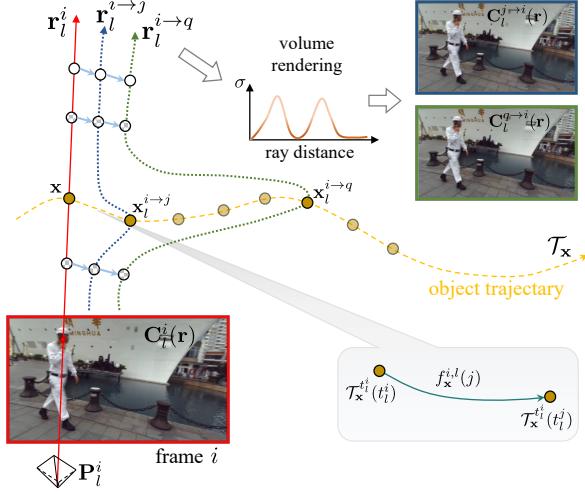


Figure 3. Cross-time rendering. We use cross-time rendering to ensure temporal consistency in dynamic scene representation. We render the target timestamp image $\mathbf{C}_l^j(\mathbf{r})$ in frame i utilizing scene information predicted from the corresponding timestamp in other frames. In addition to selecting adjacent frames $j \in \mathcal{N}(i)$, we also choose an extra frame q from the global time to ensure the global consistency of the DCT trajectory.

where $j \in \mathcal{N}(i)$ denotes the neighbor frames of i . The corresponding 3D point of \mathbf{x} at t_l^j can be computed as $\mathbf{x}_l^{i-j} = \mathbf{x} + f_{\mathbf{x}}^{i,l}(j)$. Establishing such relationships for every point on a ray \mathbf{r} can obtain a warped ray \mathbf{r}_l^{i-j} from t_l^i to t_l^j . Subsequently, we utilize volume rendering to obtain the color value of the warped ray $\mathbf{C}_l^{j-i}(\mathbf{r})$:

$$\mathbf{C}_l^{j-i}(\mathbf{r}) = \int_{s_n}^{s_f} T_{t_l^i}^j(s) \sigma_{t_l^i}(\mathbf{r}_l^{i-j}(s)) \mathbf{c}_{t_l^i}(\mathbf{r}_l^{i-j}(s), \mathbf{d}_{t_l^i}) ds. \quad (14)$$

Then we average the images within the exposure time to yield the blurred image $\mathbf{B}^{j-i}(\mathbf{r})$, and we calculate the cross-time rendering loss between $\mathbf{B}^{j-i}(\mathbf{r})$ and $\mathbf{B}_i^{\text{gt}}(\mathbf{r})$:

$$\mathbf{B}^{j-i}(\mathbf{r}) = \frac{1}{n} \sum_{l=1}^n \mathbf{C}_l^{j-i}(\mathbf{r}), \quad (15)$$

$$\mathcal{L}_{\text{cross}} = \sum_{j \in \mathcal{N}(i)} \mathbf{W}^{j-i}(\mathbf{r}) \|\mathbf{B}^{j-i}(\mathbf{r}) - \mathbf{B}_i^{\text{gt}}(\mathbf{r})\|_2^2, \quad (16)$$

where $\mathbf{W}^{j-i}(\mathbf{r})$ represents a motion disocclusion weight, which is the average outcome of n disocclusion weights $\mathbf{W}_l^{j-i}(\mathbf{r})$ computed by \mathcal{W}_t in Eq. 7. For the specifics of disocclusion weight settings, we offer detailed information in the supplementary material. Through the aforementioned cross-time rendering, the DCT coefficients Ψ predicted for the same point at different times remain consistent, which preserves the temporal consistency of the scene.

In addition to using adjacent frames $j \in \mathcal{N}(i)$ for cross-time rendering, we include an additional frame q for ensuring global time consistency. During the early training stages, q is selected to be close to i . As training progresses, q is gradually chosen further away from i until it encompasses the entire global time range. We employ a similar rendering process to obtain the blurry image $\mathbf{B}^{q \rightarrow i}(\mathbf{r})$ and compute cross-time rendering loss with input image. In summary, the cross-time rendering loss can be written as:

$$\mathcal{L}_{\text{cross}} = \sum_{j \in \{\mathcal{N}(i), q\}} \mathbf{W}^{j \rightarrow i}(\mathbf{r}) \|\mathbf{B}^{j \rightarrow i}(\mathbf{r}) - \mathbf{B}_i^{\text{gt}}(\mathbf{r})\|_2^2. \quad (17)$$

3.4. Training & Optimization

Data-driven priors. We introduce data-driven priors $\mathcal{L}_{\text{data}}$ to constrain the scene geometry. We utilize RAFT [43] and MiDaS [37] to obtain ground truth optical flow and depth for the input blurry images. Intuitively, we only need to calculate the depth maps for each timestamp, and optical flow maps from each timestamp to the next exposure time, subsequently averaging them within a single exposure time and constraining them with the ground truth. However, we encountered issues where the depth and optical flow do not have a translucent effect in the blurry edges, as opposed to observed in RGB data (we show a depth map as an example in Fig. 4(a)). Therefore, using simple averaging is unsuitable for constraining depth and optical flow. To address this issue, we implement Extreme Value Constraints (EVC) for both depth and optical flow maps. Specifically, after calculating the depth maps $\{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n\}$ and optical flow maps $\{\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_n\}$ projected from 3D scene flow within an exposure time, we derive the predicted blurry result by taking the minimum depth value and maximum optical flow value, and calculate the loss with ground truth:

$$\begin{aligned} \mathbf{D}(\mathbf{r}) &= \min\{\mathbf{D}_1(\mathbf{r}), \mathbf{D}_2(\mathbf{r}), \dots, \mathbf{D}_n(\mathbf{r})\}, \\ \mathbf{F}(\mathbf{r}) &= \max\{\mathbf{F}_1(\mathbf{r}), \mathbf{F}_2(\mathbf{r}), \dots, \mathbf{F}_n(\mathbf{r})\}, \end{aligned} \quad (18)$$

$$\mathcal{L}_{\text{data}} = \|\mathbf{D}(\mathbf{r}) - \mathbf{D}_{\text{gt}}(\mathbf{r})\|_1 + \|\mathbf{F}(\mathbf{r}) - \mathbf{F}_{\text{gt}}(\mathbf{r})\|_1. \quad (19)$$

After applying EVC for data-driven constraints, a sharper result can be generated for each timestamp (as shown in Fig. 4(b)). Please refer to supplementary material for principles and technical details of EVC.

Scene flow modeling. For a more accurate representation of scene motion, we use the predicted DCT coefficients Ψ to calculate the scene flow and impose constraints using \mathcal{L}_{sf} . Similar to NSFF [22], \mathcal{L}_{sf} consists of three components: cycle consistency of the scene flow, spatial-temporal smoothness constraints, and L1 regularization to the scene flow.

Learning DCT basis. We model the motion trajectories using Eq. 6; however, we discovered that employing a learnable DCT basis produced superior outcomes. To achieve

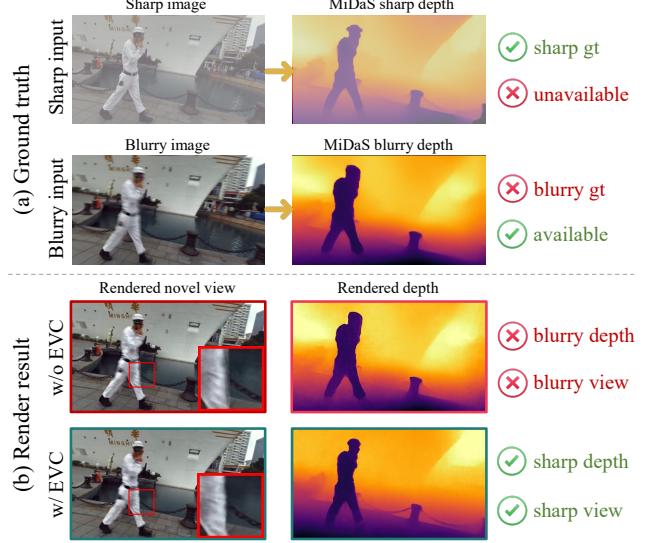


Figure 4. (a) We utilize depth map from MiDaS [37] as the supervising signal for optimizing the model. Compared with sharp inputs, blurry images are available but lead to inaccurate depth ground truth. In the example, MiDaS interprets the blurry edges of the person as foreground, causing the person on the depth map to appear ‘fatter’ than in the ideal sharp one. (b) Using the MiDaS blurry depth to optimize the model directly may predict inaccurate depth maps and distorted novel views. To mitigate this issue, we employ EVC to enhance our model for predicting sharp depth maps and novel views from blurry images.

this, we initialize the DCT basis using the expression $\cos\left(\frac{\pi}{2* n N}(2t+1)k\right)$, and treat them as learnable embeddings, facilitating joint learning with the models throughout the training process.

Implementation details. In experiments, we set $n = 7$ and $K = 6$. We employ the Adam optimizer [15] to jointly optimize the static and dynamic MLPs, camera parameters, and DCT basis. The learning rates are set to 5×10^{-4} for MLPs, 1×10^{-3} for camera parameters, and 1.25×10^{-4} for DCT basis. We train each scene for $300k$ iterations, which takes approximately two days on a single NVIDIA RTX A6000 GPU.

4. Experiments

4.1. Datasets

Since there is a lack of datasets specifically designed for addressing motion blur in dynamic neural radiance fields, we have curated a collection of dynamic scenes with motion blur from an existing Stereo Blur Dataset [60]. This dataset was captured using a ZED stereo camera, containing sharp stereo image sequences along with their corresponding blurry image counterparts. Details about the generation of blur and other dataset information can be found

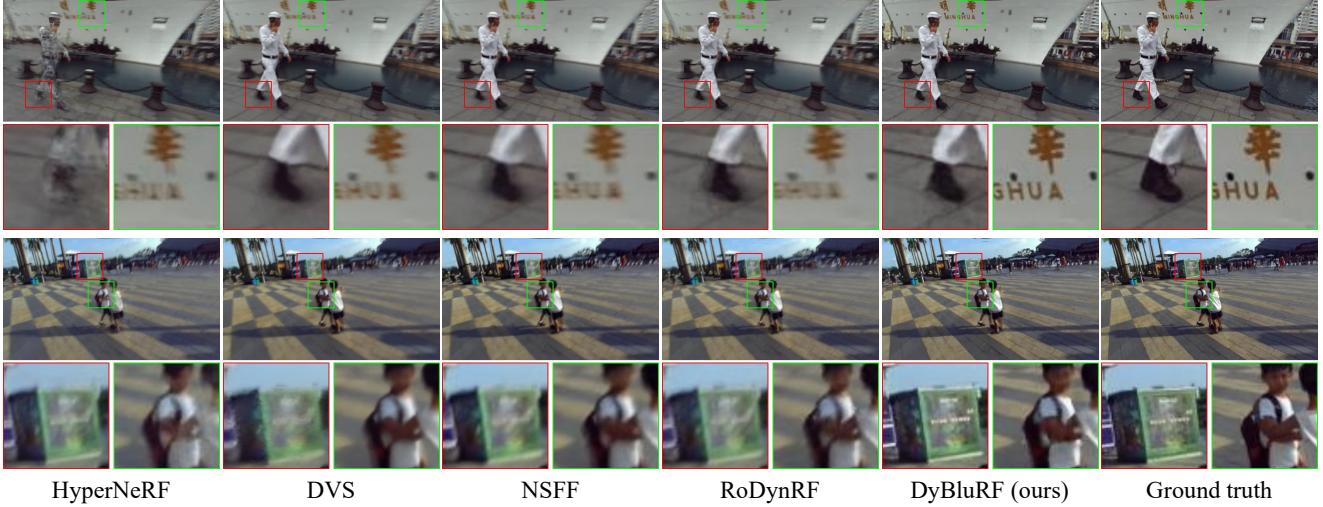


Figure 5. **Qualitative comparisons against all baselines.** Compared to existing dynamic NeRF methods, our method generates novel view images that are more faithful to the ground truth images, with less blur in both static and dynamic regions.

Methods	PSNR↑	SSIM↑	LPIPS↓	Methods	PSNR↑	SSIM↑	LPIPS↓
HyperNeRF [34]	20.15	0.744	0.221	[55] + [22]	23.17	0.839	0.143
DVS [10]	22.30	0.811	0.193	[59] + [22]	<u>23.79</u>	<u>0.854</u>	<u>0.125</u>
NSFF [22]	<u>23.33</u>	<u>0.834</u>	<u>0.181</u>	[55] + [25]	20.76	0.787	0.162
RoDynRF [25]	19.28	0.748	0.217	[59] + [25]	20.88	0.794	0.151
DyBluRF (ours)	25.66	0.895	0.079	DyBluRF	25.66	0.895	0.079

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

in [60], and we also provide additional explanations in supplementary materials. To obtain accurate camera parameters for training neural radiance fields, we require image sequences of dynamic scenes with sufficient motion parallax. We collected 6 dynamic scenes suitable for our task from this dataset. These scenes exhibit significant motion blur caused by both camera and object movements. In our experiments, we employ COLMAP [38] to acquire the camera parameters for the input images. For each scene, we extract 24 frames from the original video, utilizing the left blurry image sequences for training and the corresponding right sharp image sequences for testing.

4.2. Baselines and Error Metrics

We compare our method against several state-of-the-art monocular dynamic NeRF methods. Specifically, the comparison involves a canonical space-based method, HyperNeRF [34], two scene flow-based methods, NSFF [22] and DVS [10], and a robust dynamic NeRF method, RoDynRF [25]. Furthermore, to demonstrate the superiority of our approach over 2D image deblurring, we select the current state-of-the-art single-image deblurring method

Restormer [55] and a dynamic scene video deblurring method JCD [59] to perform deblurring preprocessing on the input blurry images. We then train NSFF and RoDynRF on the preprocessed images for comparison.

Similar to most existing dynamic NeRF methods, we use PSNR, SSIM, and LPIPS [57] to quantitatively evaluate the quality of novel views generated by different approaches.

4.3. Comparisons

We conduct a quantitative evaluation on the scenes collected from the Stereo Blur Dataset. The average results across all scenes are shown in Tab. 1 (left), clearly demonstrating the superior performance of our method compared to the existing dynamic NeRF methods. We also provide qualitative comparisons of novel view results generated by different methods, as shown in Fig. 5. Notably, existing dynamic NeRF methods cannot handle motion blur inputs, resulting in low-quality novel views in both dynamic and static regions. In contrast, our method can effectively manage inputs with motion blur and produce sharp and high-quality novel views. Furthermore, to demonstrate the superiority of our method in modeling motion blur in 3D space, we first preprocess the input blurry images with 2D deblurring and then train dynamic NeRF models, e.g., RoDynRF, for comparison. As shown in Tab. 1 (right) and Fig. 6, our method continues to outperform the alternatives. This comparison underscores that 2D deblurring methods fail to preserve the consistency of scene information. When the dynamic NeRFs are trained on these pre-processed images, it leads to undesirable artifacts in scene geometry. In contrast, our method models the blurring process in 3D space, ensuring spatial-temporal consistency of scene geometry.

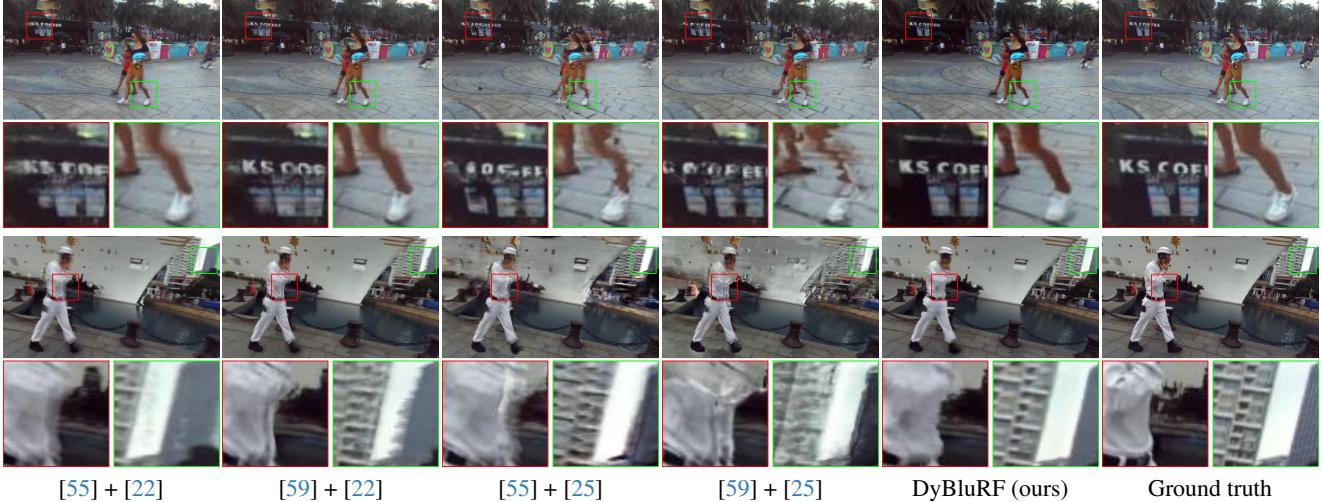


Figure 6. **Qualitative comparisons against dynamic NeRF baselines incorporated with 2D deblur method.** Even if we use preprocessed input blurry images by 2D deblur approaches to train existing dynamic NeRF methods, our method also generates more reliable novel views, with less blur in both static and dynamic regions.

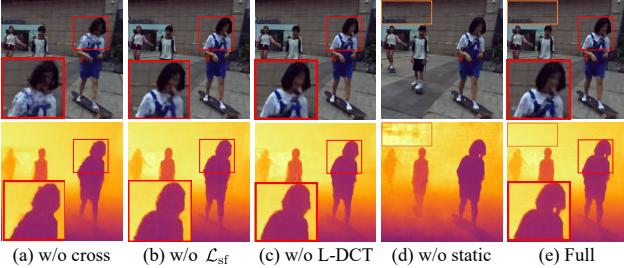


Figure 7. **Qualitative ablations.** (a) No cross-time rendering or (b) no \mathcal{L}_{sf} results in degradation of predicted RGB and depth maps, especially containing artifacts in dynamic regions. (c) Not learning DCT basis may impact the representation of scene details. (d) The absence of static model leads to poor results in static regions. (e) Our proposed full method yields high-quality results in both dynamic and static regions.

4.4. Ablation Study

We conduct an ablation study to evaluate the contribution of each component in our model. Specifically, we evaluate the impact of (1) removing the cross-time rendering (w/o cross), (2) using average in data-driven priors instead of EVC (w/o EVC), (3) removing the static model (w/o static), (4) removing \mathcal{L}_{sf} (w/o \mathcal{L}_{sf}), and (5) directly computing DCT basis using cosine values without being set as learnable parameters (w/o L-DCT). We report quantitative results in Tab. 2. One can see that without cross-time rendering and scene flow constraints, view synthesis quality degrades significantly. Combining static model can also improve performance. Furthermore, removing EVC or not learning DCT basis can reduce performance, demonstrating the value of these training tricks for improving novel view quality. We also conduct visual comparisons in Fig. 4(b) and Fig. 7, demonstrating that each component contributes

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o cross	24.54	0.875	0.087
w/o EVC	25.42	0.890	<u>0.079</u>
w/o static	24.67	0.869	0.090
w/o \mathcal{L}_{sf}	24.97	0.879	0.085
w/o L-DCT	<u>25.49</u>	0.891	0.085
Full	25.66	0.895	0.079

Table 2. **Ablation study.** The best performance is **boldfaced**, and the second is underlined.

to the overall performance of the model. Specifically, without EVC in data-driven constraints, the incorrect depth and flow priors can misleadingly optimize the model to predict low-quality rendering results. Removing cross-time rendering or \mathcal{L}_{sf} cannot ensure the temporal consistency of dynamic scenes, resulting in obvious artifacts in dynamic regions. Static model can improve the scene representation for static elements, benefiting high-quality rendering results in static scenes. Finally, learning DCT basis can improve the representation of scene details.

5. Conclusion

We propose DyBluRF, a dynamic radiance field to synthesize sharp novel views from a monocular video with motion blur. By modeling camera trajectory and object motion trajectories within the scene, our method overcomes the impact of motion blur inputs on novel view synthesis. We also ensure the global temporal consistency of the predicted DCT trajectory through cross-time rendering. We conducted extensive experiments using scenes collected from the Stereo Blur Dataset, which demonstrate that our method outperforms existing dynamic NeRF methods in handling monocular inputs with motion blur.

References

- [1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O’Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16610–16620, 2023. 3
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 2
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 2
- [4] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4160–4169, 2023. 3
- [5] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. 3
- [6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–350, 2022. 3
- [7] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16569–16578, 2023. 3
- [8] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *ACM SIGGRAPH Asia*, pages 1–9, 2022. 3
- [9] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 3
- [10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5712–5721, 2021. 1, 3, 5, 7
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 3
- [12] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18398–18408, 2022. 3
- [13] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5426–5434, 2015. 2
- [14] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 3
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [16] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 233–240, 2011. 2
- [17] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8183–8192, 2018. 2
- [18] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8878–8887, 2019. 2
- [19] Dogyoon Lee, Minhyeok Lee, Chajin Shin, and Sangyoun Lee. Dp-nerf: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12386–12396, 2023. 1, 3
- [20] Dongwoo Lee, Jeongtaek Oh, Jaesung Rim, Sunghyun Cho, and Kyoung Mu Lee. Exblurf: Efficient radiance fields for extreme motion blurred images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17639–17648, 2023. 1, 3
- [21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5521–5531, 2022. 3
- [22] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6498–6508, 2021. 1, 3, 4, 5, 6, 7, 8
- [23] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4273–4284, 2023. 3, 5
- [24] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5741–5751, 2021. 3

- [25] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13–23, 2023. 1, 3, 4, 5, 7, 8
- [26] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12861–12870, 2022. 1, 3
- [27] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021. 3
- [28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 1, 2
- [29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 3
- [30] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3883–3891, 2017. 2
- [31] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 261–270, 2017. 2
- [32] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3043–3051, 2020. 2
- [33] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5865–5874, 2021. 3
- [34] 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 Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 1, 3, 7
- [35] Naama Pearl, Tali Treibitz, and Simon Korman. Nan: Noise-aware nerfs for burst-denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12672–12681, 2022. 3
- [36] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021. 1, 3
- [37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 44(3):1623–1637, 2020. 6, 1
- [38] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 7
- [39] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics (TOG)*, 27(3):1–10, 2008. 2
- [40] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5572–5581, 2019. 2
- [41] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1279–1288, 2017. 2
- [42] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8174–8182, 2018. 2
- [43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 6, 2
- [44] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12959–12970, 2021. 3
- [45] Jack Valmadre and Simon Lucey. General trajectory prior for non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1394–1401, 2012. 2, 4
- [46] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 3, 4
- [47] Chaoyang Wang, Lachlan Ewen MacDonald, Laszlo A Jeni, and Simon Lucey. Flow supervision for deformable nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21128–21137, 2023. 3
- [48] Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4170–4179, 2023. 1, 2, 3, 4
- [49] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance

- fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3
- [50] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 491–498, 2010. 2
- [51] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 3
- [52] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8254–8263, 2023. 3
- [53] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5336–5345, 2020. 3
- [54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 3
- [55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 2, 7, 8, 3, 5
- [56] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 581–596, 2022. 2
- [57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7
- [58] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 191–207, 2020. 2
- [59] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9219–9228, 2021. 7, 8, 3, 5
- [60] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy S Ren. Davanet: Stereo deblurring with view aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10996–11005, 2019. 2, 6, 7

DyBluRF: Dynamic Neural Radiance Fields from Blurry Monocular Video

Supplementary Material

6. Appendix

We provide the following content in this supplementary:

- Detailed description of disocclusion weights;
- Principles and technical details of EVC;
- Detailed description of scene flow constrains \mathcal{L}_{sf} ;
- Final loss;
- More information on the dataset;
- Detailed quantitative and qualitative results;
- Results on sharp inputs;
- Limitations.

A. Disocclusion weights

Recall in Eq. 7 of our main paper, the dynamic MLP additionally outputs a disocclusion weight $\mathcal{W}_t = (w_{fw}, w_{bw})$. The weights w_{fw} and $w_{bw} \in [0, 1]$ represent the probability of occlusion for a spatial point. They determine the confidence regarding occlusion occurring from a specific timestamp in the current frame to the corresponding timestamp in adjacent frames. Certainly, a value closer to 0 signifies minimal chances of the point being occluded during that interval. Conversely, a value closer to 1 indicates a higher likelihood of the point experiencing occlusion or disocclusion. Specifically, for a timestamp t_l^i from frame i , where $l \in \{1, \dots, n\}$ represents the specific timestamp in an exposure time, the disocclusion weights are denoted as $\mathcal{W}_{t_l^i} = (w_l^{i \rightarrow i+1}, w_l^{i \rightarrow i-1})$. To obtain the disocclusion weight map in 2D plane, the weight along the ray \mathbf{r}_l^i is used for volume rendering with opacity from adjacent frames:

$$\mathbf{W}_l^{j \rightarrow i}(\mathbf{r}_l^i) = \int_{s_n}^{s_f} T_{t_l^i}(s) \sigma_{t_l^i}(\mathbf{r}_l^{i \rightarrow j}(s)) (1 - w_l^{i \rightarrow j}(\mathbf{r}_l^i(s))) ds, \quad (20)$$

where $j \in \mathcal{N}(i) = \{i+1, i-1\}$ denotes the adjacent frames of frame i , $\mathbf{r}_l^{i \rightarrow j}$ represents the warped ray mentioned by Eq. 14 in the main paper.

For each timestamp within the exposure time, the disocclusion weight map $\mathbf{W}_l^{j \rightarrow i}(\mathbf{r})$ is computed using Eq. 20. Subsequently, we average these maps to yield the motion disocclusion weight $\mathbf{W}^{j \rightarrow i}(\mathbf{r})$ for cross-time rendering:

$$\mathbf{W}^{j \rightarrow i}(\mathbf{r}) = \frac{1}{n} \sum_{l=1}^n \mathbf{W}_l^{j \rightarrow i}(\mathbf{r}). \quad (21)$$

B. Extreme Value Constraints (EVC)

In this section, we will introduce the operational principles and technical details of EVC. In Fig. 4 of our main paper, we highlight an issue encountered while predicting depth and optical flow from blurry inputs. The depth prediction

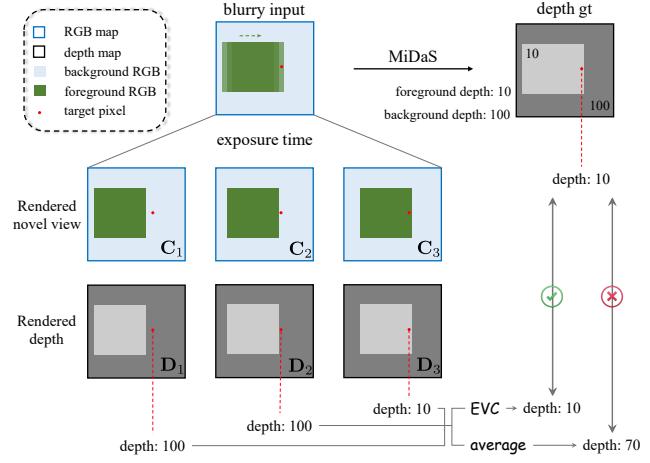


Figure 8. Principle of EVC.

network and optical flow prediction network often mistakenly interpret the blurry edges in the RGB image as foreground, resulting in inaccuracies in data priors. To enable model learning accurate scene geometry by using these inaccurate data priors, we propose the EVC data prior constraint method.

Fig. 8 provides an illustration of a simulation experiment on the working principle of EVC. We suppose that each exposure time is represented by three sharp images (*i.e.*, $n = 3$), with only one green square foreground moving from left to right in the scene, while the rest is a blue background. Given an input blurry frame, the model produces 3 RGB images and depth maps within the exposure time. We consider a pixel position in blurry edges of the input frame, where the pixel changes the affiliation from the background (in C_1 and C_2) to the foreground (in C_3). To ease the discussion, we set the depth of the foreground to be 10 while that of the background to be 100 respectively. An intuitive practice to simulate the depth of input blurry frames may be averaging the three depth values to a value of 70. However, due to the mistaken interpretation of blurry edges, MiDaS [37] identifies the red point as the foreground, leading to its predicted depth value of 10. Therefore the intuitive constraint cannot work for our method, manifesting as ‘foreground-fatter’ depth map predictions as shown in the first row of Fig. 4(b) in the main paper. In order to accommodate the depth prediction results, we should recognize a position as on the foreground (the green square), once it is covered by the foreground in a certain timestamp during the exposure time. This can be expressed as choosing the minimum depth value at the red-pointed position among the three sharp images (as EVC

does). This approach enables learning accurate scene geometry and rendering a sharp depth map, as shown in the second row of Fig. 4(b) in the main paper.

The process of data prior constraint for optical flow closely resembles that of depth. The model generates an optical flow map for each timestamp. This map represents the scene motion from the timestamp of current exposure time to the corresponding timestamp of adjacent exposure time. However, in contrast to depth where foreground values are typically smaller than background ones, optical flow for the foreground often exceeds that of the background. According to the principle and process of EVC, we use the maximum optical flow as the simulated blurry optical flow and compare it with the ground truth predicted by RAFT [43]. In practice, we obtain the optical flow by projecting the predicted 3D scene flow onto a 2D plane.

C. Scene flow modeling details

Recall in Sec. 3.4, \mathcal{L}_{sf} is used as a regularization loss for the scene flow calculated by Eq. 13 in the main paper. \mathcal{L}_{sf} consists of three components: scene flow cycle consistency, spatial-temporal smoothness, and minimal scene flow.

Scene flow cycle consistency enforces the forward scene flow $f_{\mathbf{x}}^{i,l}(j)$ of sampled 3D points at timestamps t_l^i consistent with the backward scene flow $f_{\mathbf{x}}^{j,l}(i)$ at the corresponding location at timestamps t_l^j :

$$\mathcal{L}_{\text{cyc}} = \sum_{j \in \{i \pm 1\}} \sum_{l=1}^n (1 - w_l^{i \rightarrow j}) \|f_{\mathbf{x}}^{i,l}(j) + f_{\mathbf{x}}^{j,l}(i)\|_1, \quad (22)$$

where $w_l^{i \rightarrow j}$ denotes the disocclusion weights used to reduce consideration of occlusion points.

The spatial-temporal smoothness is designed to maintain continuity in scene flow both spatially and temporally. To achieve spatial smoothness, we encourage consistency between the scene flows of adjacent spatial points. Specifically, we compute L1 loss between scene flows sampled at two neighboring spatial points along the ray \mathbf{r}_l^i . For temporal smoothness, we minimize the sum of forward and backward scene flow for each sampled spatial point to ensure the smoothness of predicted DCT trajectories. The spatial-temporal smoothness can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{smooth}} &= \sum_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} \sum_{j \in \{i \pm 1\}} \sum_{l=1}^n \|f_{\mathbf{x}}^{i,l}(j) - f_{\mathbf{y}}^{i,l}(j)\|_1 \\ &\quad + \frac{1}{2} \sum_{l=1}^n \|f_{\mathbf{x}}^{i,l}(i+1) + f_{\mathbf{x}}^{i,l}(i-1)\|_2^2, \end{aligned} \quad (23)$$

where the first term denotes spatial smoothness and the second term denotes temporal smoothness. $\mathcal{N}(\mathbf{x})$ represents the neighboring spatial points of point \mathbf{x} along the ray \mathbf{r}_l^i .

Finally, due to minor scene changes between adjacent frames, we additionally apply a minimal scene flow constraint to minimize the predicted 3D scene flow:

$$\mathcal{L}_{\text{min}} = \sum_{j \in \{i \pm 1\}} \sum_{l=1}^n \|f_{\mathbf{x}}^{i,l}(j)\|_1. \quad (24)$$

D. Final loss

Considering the RGB image rendering, temporal consistency of dynamic scenes, data-driven constraints, and scene flow modeling, the final training loss of our method is:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{cross}} + \lambda_{\text{data}} \mathcal{L}_{\text{data}} + \lambda_{\text{sf}} \mathcal{L}_{\text{sf}}, \quad (25)$$

where

$$\begin{aligned} \mathcal{L}_{\text{RGB}} &= \mathcal{L}_{\text{RGB}}^{\text{cb}} + \mathcal{L}_{\text{RGB}}^{\text{dy}} + \lambda_{\text{st}} \mathcal{L}_{\text{RGB}}^{\text{st}}, \\ \mathcal{L}_{\text{sf}} &= \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{min}}, \end{aligned} \quad (26)$$

λ_{data} , λ_{sf} , λ_{st} , λ_{cyc} denote the weights for $\mathcal{L}_{\text{data}}$, \mathcal{L}_{sf} , $\mathcal{L}_{\text{RGB}}^{\text{st}}$ and \mathcal{L}_{cyc} , respectively. During training, λ_{data} multiplies 0.1 every 50k iterations to prevent the model from overfitting to data priors.

E. Dataset

We conduct experiments using dynamic scenes from the Stereo Blur Dataset [60]. Similar to most datasets in image deblurring [30, 41], the Stereo Blur Dataset generates motion blur by averaging a sharp high frame rate sequence. In practice, this dataset is captured using a ZED stereo camera capable of acquiring high frame rate (60fps) image sequences of dynamic scenes at a resolution of 1280×720 . However, because the frame rate is still not high enough, direct averaging may lead to some unrealistic artifacts. Therefore, the dataset employs a video frame interpolation method [31] to increase the frame rate to 480fps. Subsequently, averaging is performed on different numbers (17, 33, 49) of consecutive frames to generate motion blur, with the sharp center frame among the consecutive frames serving as the ground truth for the blurry image.

However, not all scenes in the Stereo Blur Dataset can be used in our work. That is because of two issues: 1) Many scenes within the dataset are static and lack moving objects. 2) Several scenes have minimal camera motion, resulting in a lack of motion blur caused by camera movement and insufficient motion parallax to obtain camera parameters. Therefore, we select 6 dynamic scenes from the Stereo Blur Dataset that are tailored for NeRF-based methods. These scenes encompass both camera and object motion blur, showcasing varied size object movements like playing seesaw, walking, and skating. We employ COLMAP to acquire camera parameters from the image sequences and downsample the image resolution to 512×288 for experiments. Similar to NSFF, we process the input blurry image sequences through the MiDaS to obtain depth maps,

Methods	Sailor			Seesaw			Street			Children			Skating			Basketball		
	PSNR↑	SSIM↑	LPIPS↓															
BAD-NeRF [48]	16.96	0.631	0.333	20.58	0.795	0.220	20.27	0.670	0.190	18.10	0.650	0.386	19.08	0.691	0.345	17.99	0.731	0.271
HyperNeRF [34]	18.56	0.743	0.275	20.25	0.779	0.182	19.99	0.662	0.137	21.36	0.762	0.279	19.52	0.702	0.319	21.21	0.818	0.136
DVS [10]	<u>22.32</u>	<u>0.810</u>	<u>0.247</u>	18.14	0.775	0.158	19.06	0.669	0.176	24.00	0.823	0.307	26.18	0.907	0.124	24.10	0.880	0.143
NSFF [22]	19.06	0.726	0.290	19.92	0.807	0.178	<u>23.42</u>	<u>0.800</u>	<u>0.121</u>	<u>24.55</u>	<u>0.846</u>	<u>0.259</u>	27.96	0.923	<u>0.116</u>	<u>25.06</u>	<u>0.903</u>	<u>0.122</u>
RoDynRF [25]	12.69	0.584	0.317	<u>23.71</u>	0.894	<u>0.116</u>	19.80	0.719	0.161	15.02	0.609	0.382	21.66	0.831	0.160	22.82	0.850	0.166
DyBluRF (ours)	23.50	0.860	0.115	24.56	<u>0.882</u>	0.075	26.88	0.906	0.068	25.57	0.884	0.092	<u>27.94</u>	<u>0.917</u>	0.072	25.28	0.920	0.050

Table 3. **Quantitative comparisons for every scene against all dynamic NeRF baselines.** The best performance is **boldfaced**, and the second is underlined.

Methods	Sailor			Seesaw			Street			Children			Skating			Basketball		
	PSNR↑	SSIM↑	LPIPS↓															
[55] + [22]	18.96	0.728	0.252	20.19	0.827	0.149	<u>23.39</u>	0.790	0.125	<u>24.88</u>	0.864	<u>0.182</u>	<u>26.91</u>	0.922	0.067	24.70	<u>0.903</u>	0.084
[59] + [22]	<u>22.60</u>	<u>0.805</u>	<u>0.150</u>	21.31	0.856	0.135	23.06	<u>0.791</u>	0.114	24.27	0.853	0.195	26.88	<u>0.917</u>	0.078	24.64	<u>0.903</u>	<u>0.076</u>
[55] + [25]	17.16	0.703	0.227	23.63	<u>0.892</u>	0.090	19.75	0.727	0.149	17.41	0.665	0.282	23.22	0.869	0.110	23.38	0.868	0.114
[59] + [25]	17.28	0.772	0.225	<u>23.85</u>	0.893	<u>0.080</u>	19.69	0.700	0.153	18.39	0.683	0.226	22.54	0.849	0.126	23.51	0.867	0.098
DyBluRF (ours)	23.50	0.860	0.115	24.56	0.882	0.075	26.88	0.906	0.068	25.57	0.884	0.092	<u>27.94</u>	<u>0.917</u>	<u>0.072</u>	25.28	0.920	0.050

Table 4. **Quantitative comparisons for every scene against dynamic NeRF methods with blurry image preprocess.** The best performance is **boldfaced**, and the second is underlined.

use RAFT to generate optical flow maps, and employ an instance segmentation network (Mask r-cnn [11]) to derive motion masks for moving objects.

F. Detailed results

In this section, we present detailed quantitative and qualitative results of the comparative experiments in our main paper. In Tab. 1 of the main text, we have shown the average quantitative results of all baselines across the 6 scenes. Here, we will provide individual quantitative results for each scene, as depicted in Tab. 3 for blurry inputs and Tab. 4 for deblurring preprocess inputs. We also conduct comparative experiments with BAD-NeRF, which is a deblurring NeRF method designed for static scenes, in Tab. 3. Due to its inability to represent dynamic scenes, the performance of BAD-NeRF is much lower than ours. We also include detailed qualitative comparison results, as illustrated in Fig. 10 and Fig. 11. One can see that our method performs the best quantitative results in most scenes. Although our approach slightly trails NSFF in the Skating scene, the qualitative results of our method in the Skating scene are better than NSFF, regardless of whether blurry inputs or pre-processed sharp inputs, as shown in the second row of Fig. 10 and Fig. 11. We speculate that slightly lower metrics in the Skating scene of our method could be attributed to the background in the Skating scene having an extensive low-texture area with less motion blur, which cannot fully showcase the superiority of our method in handling motion

blur input through metrics. However, the qualitative results demonstrate that our method outperforms all baselines, especially in handling motion blur in dynamic scenes.

From Tab. 3 and Tab. 4, we can also observe that pre-processing blurry images with 2D deblurring methods may yield inferior results compared to directly using blurry inputs. This phenomenon arises from the 2D deblurring methods disrupting the temporal consistency of scene information, causing inaccurate scene representation in NeRF. This further reflects that our approach effectively ensures spatial-temporal consistency in dynamic scenes.

G. Results on sharp inputs.

Although our method is specifically tailored for dynamic scenes with motion blur, it remains capable of achieving comparable results to existing methods when dealing with sharp inputs. We evaluate our method on the Nvidia Dynamic Scene Dataset [53] without blur, configured with $n = 1$ for sharp image input. Here, we compare with NSFF, which performs second best in Tab. 1 in the main paper. As shown in Tab. 5, our method is marginally better than NSFF even in sharp inputs, which underscores our sustained efficacy even in sharp inputs. Meanwhile, our method significantly outperforms NSFF under blurry inputs in Tab. 1. That demonstrates the effectiveness of our method in handling both motion blur and representing dynamic scenes.

Methods	Balloon1			Balloon2			Jumping			Playground		
	PSNR↑	SSIM↑	LPIPS↓									
NSFF [22]	21.96	0.791	0.215	24.27	0.825	0.222	24.65	0.872	0.151	21.22	0.780	0.212
DyBluRF (ours)	21.90	0.781	0.181	24.92	0.880	0.117	26.21	0.901	0.091	23.64	0.861	0.113
Methods	Skating			Truck			Umbrella			Average		
	PSNR↑	SSIM↑	LPIPS↓									
NSFF [22]	29.29	0.936	0.129	25.96	0.863	0.167	22.97	0.769	0.295	24.33	0.834	0.199
DyBluRF (ours)	28.36	0.913	0.087	30.01	0.946	0.043	24.19	0.831	0.163	25.60	0.873	0.114

Table 5. Quantitative results on the non-blur dataset with NSFF. The better performance is **boldface**.



Figure 9. **Limitation.** It is challenging to restore a sharp image when input blur is caused by extreme object motion (*e.g.*, the leg).

H. Limitations

Although our method can handle most of the motion blur in the input images, it might not be able to synthesize high-quality novel views when the motion blur is caused by highly complex and fast motion. As shown in Fig. 9, the rapid movement of the human leg causes extreme blurriness in the input image. This challenges our method in two aspects: firstly, excessive information loss of foreground may lead to the depth and flow prediction networks being unable to identify foreground objects, thereby affecting the representation of dynamic scenes. Secondly, the model tends to incorrectly learn the scene geometry supervised by such extremely blurry images, and gets stuck in the incorrect local minimum. In the future, we aim to explore combining explicit representations to enhance the temporal coherence of dynamic objects to solve this problem.



Figure 10. **Qualitative comparisons against all baselines.** Compared to existing dynamic NeRF methods, our method generates novel view images that are more faithful to the ground truth images, with less blur in both static and dynamic regions.

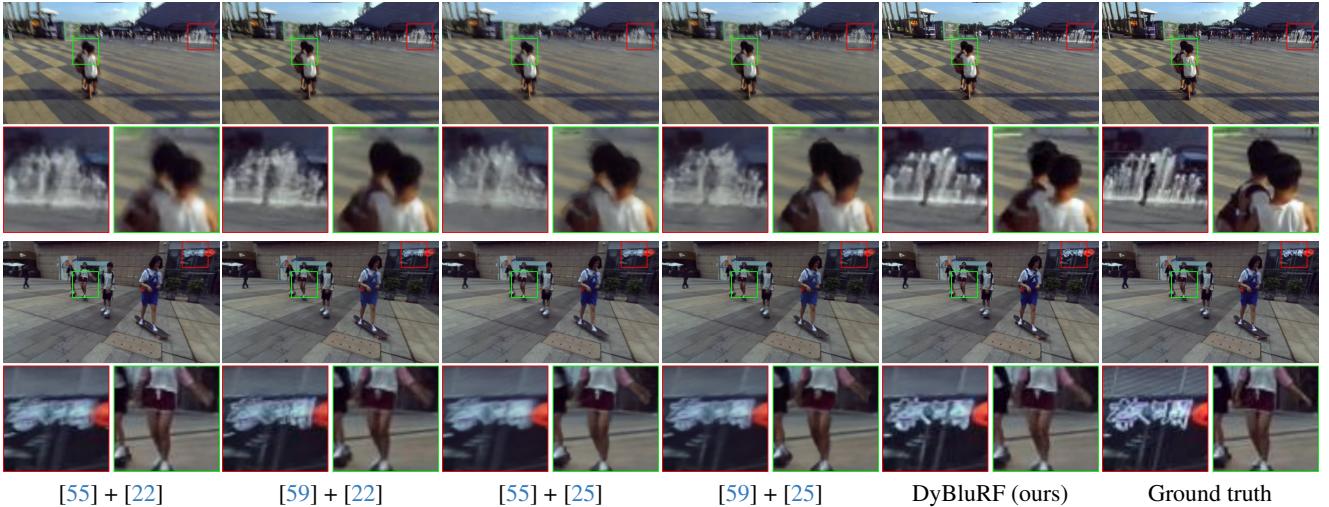


Figure 11. **Qualitative comparisons against dynamic NeRF baselines incorporated with 2D deblur method.** Even if we use preprocessed input blurry images by 2D deblur approaches to train existing dynamic NeRF methods, our method also generates more reliable novel views, with less blur in both static and dynamic regions.