BeNeRF: Neural Radiance Fields from a Single Blurry Image and Event Stream

Wenpu Li[1,5*], Pian Wan[1,2*], Peng Wang[1,3*], Jinghang Li[4], Yi Zhou[4], and Peidong Liu[1†]

[1]Westlake University, [2]EPFL, [3]Zhejiang University, [4]Hunan University
[5]Guangdong University of Technology

https://akawincent.github.io/BeNeRF

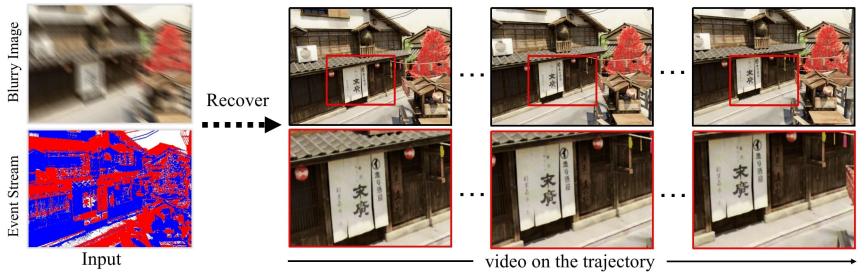

Fig. 1: Given a single blurry image and its corresponding event stream, BeNeRF can synthesize high-quality novel images along the camera trajectory, recovering a sharp and coherent video from the single blurry image.

Abstract. Implicit scene representation has attracted a lot of attention in recent research of computer vision and graphics. Most prior methods focus on how to reconstruct 3D scene representation from a set of images. In this work, we demonstrate the possibility to recover the neural radiance fields (NeRF) from a single blurry image and its corresponding event stream. To eliminate motion blur, we introduce event stream to regularize the learning process of NeRF by accumulating it into an image. We model the camera motion with a cubic B-Spline in SE(3) space. Both the blurry image and the brightness change within a time interval, can then be synthesized from the NeRF given the 6-DoF poses interpolated from the cubic B-Spline. Our method can jointly learn both the implicit scene representation and the camera motion by minimizing the differences between the synthesized data and the real measurements without any prior knowledge of camera poses. We evaluate the proposed method with both synthetic and real datasets. The experimental results demonstrate that we are able to render view-consistent latent sharp images from the learned NeRF and bring a blurry image alive in high quality.

Keywords: Neural Radiance Fields · Event Stream · Pose Estimation · Deblurring · Novel View Synthesis · 3D from a Single Image

⋆ Equal contribution: {liwenpu, wangpeng}@westlake.edu.cn, pianwan@outlook.com
† Corresponding author: Peidong Liu(liupeidong@westlake.edu.cn).

1 Introduction

Neural Radiance Fields (NeRF) [28] has drawn much attention due to its extraordinary ability in representing 3D scenes and synthesizing novel views. Given multi-view sharp RGB images and calibrated camera poses from COLMAP [40], NeRF takes corresponding 3D spatial location and 2D view direction as input, and optimizes a multi-layer perception (MLP) to represent the 3D scene. More recent advanced methods also exploit explicit octree [52], multi-resolution hash encoding [30] etc., to represent the 3D scene to improve both the training and rendering efficiency.

Prior methods usually rely on multi-view images to learn the 3D representation. Several pioneering works recently attempt to exploit a single image to learn the underlying neural radiance fields [3, 36, 38, 53]. They usually rely on a large dataset to pre-train the networks to learn priors to address the ill-posed problem. A blurry image further aggravates the problem due to the image quality degradation. Although motion blur is usually not preferred by most vision algorithms, they actually encode additional camera motion trajectory and more structural information compared to a sharp image. In this paper, we explore the possibility of recovering the neural radiance fields and camera motion trajectory from a single blurry image. Instead of learning priors from a large dataset as in previous works, we exploit the usage of an additional event stream to better constrain the problem.

Event stream can be acquired by an event camera [23] which captures pixel intensity changes caused by the relative motion between the static scene and camera. Unlike standard frame-based cameras, event camera captures asynchronous events with very low latency, leading to extremely high temporal resolution [8]. This characteristic compensates with the image formation process of a blurry image (i.e. integral of photon measurements across time). Several prior works thus take advantage of both modalities for high quality single image deblurring [33,44,47]. However, these methods are unable to recover the camera motion trajectory and extract structural details from a single blurry image, thereby limiting their applicability in 3D computer vision tasks. Some NeRF-based methods that incorporate event stream [10, 16, 25, 34, 39] demonstrate the capability to achieve image deblurring and accurate reconstruction of neural radiance fields. Nonetheless, these methods necessitate input images from multiple viewpoints alongside event data. In contrast, we explore the usage of only a single blurry image for the NeRF recovery with unknown camera motions in this work.

We represent the continuous camera motion with a cubic B-Spline in SE(3) space and define it as the trajectory of both frame-based camera and event camera. Given the neural 3D representation and interpolated poses from the cubic B-Spline, we can synthesize both the blurry image and the brightness change within a time interval via the physical image formation process. The NeRF and motion trajectory can then be jointly optimized by minimizing the difference between the synthesized data and the real measurements. To evaluate the performance of our method, we conduct experiments with both synthetic and real datasets. The experimental results demonstrate that our method is able to

recover the neural radiance fields from a blurry image and its corresponding event stream without knowing prior knowledge of poses. We are thus able to render view-consistent latent sharp images encoded in a single blurred image from learned NeRF, effectively enhancing the quality of the blurry image. The experimental results further demonstrate that our method is even able to reach same performance as E^{2}NeRF [34], which targets for the same problem, but with multi-view images and longer event data. In summary, our key contributions are as follows:

- We propose a NeRF-based method that can recover the neural 3D representation from a single blurry image and its corresponding event stream, without knowing any ground truth poses;
- Our method is able to estimate the complex continuous trajectory of camera motion during the imaging process from a single blurry image, providing multi-view geometric information;
- We experimentally validate that our approach is able to recover high quality latent sharp images and high frame-rate video from a single blurry image, without any generalization issue. Furthermore, we are able to reach same level of performance as E^{2}NeRF [34], which requires multi-view images and longer event data.

2 Related Work

We roughly categorize our related works into three main areas: neural implicit scene representation, single image deblurring and event-enhanced image deblurring.

Neural Implicit Scene Representation. NeRF has attracted lots of attention due to its powerful ability of implicit 3D scene representation and novel view synthesis [28]. Many following works have been proposed to improve NeRF's performance or extend NeRF to other fields. [2, 24, 49] jointly trained NeRF with inaccurate camera poses. [6, 22, 26, 27, 32, 53] improved the performance of NeRF with degraded images, including noisy or few images etc. Recently, several event based NeRF [10, 17, 34, 39] have also been proposed.

We will mainly focus on the methods that are most related to ours. BAD-NeRF [48] aims to recover true underlying 3D scene representation from multi-view blurry inputs and inaccurate camera poses estimated from COLMAP [40]. However, BAD-NeRF and its variants [20,58] struggle to address situations where the input is limited to a single image, primarily due to the severe illness of the problem. Event-Enhanced NeRF (E^{2}NeRF) [34] aims to recover the 3D scene representation from multiple blurry images and event streams. To train NeRF without optimizing the camera poses, E^{2}NeRF manually segments the event stream using preset parameters, aiming to recover sharp images individually through EDI [33]. However, in cases of serve camera motion, the recovered images from event stream may still be blurry, leading to inaccurate pose estimation in COLMAP [40]. This two-stage approach introduces errors (i.e. either from

EDI [33] or COLMAP [40]) and fails to accurately model the continuous camera trajectory. In contrast, we are able to train both the NeRF and the camera motion trajectory jointly from a single blurry image and its corresponding event stream, which is harder than prior methods.

Single Image Deblurring. A blurred image can be formulated as the convolution result of a sharp image and a kernel. Therefore, classical approaches [11,14,21,42,51] generally regard the deblurring problem as a joint optimization of the blur kernel and the latent virtual sharp images. With the development of deep learning, many learning based end-to-end debluring method have been proposed [13,18,19,31,43,46,54]. These methods usually demonstrate better qualitative and quantitative deblurring results. However, learning based deblurring methods are usually trained on a large dataset which contains paired blurry-sharp images. They would thus usually have limited generalization performance to domain-shifted images. Since NeRF is a test-time optimization approach, our method does not have the generalization performance issue.

Event Enhanced Image Deblurring. Since event camera is able to capture high dynamic temporal information [8], prior methods usually exploit event measurements to enhance the image deblurring performance [12,33,44,45,47,50,57]. EDI [33] is a simple and effective model, which is able to generate a sharp video under various types of blur by solving a single variable non-convex optimization problem. Different from EDI [33], [44,45,47] design end-to-end neural networks for event enhanced image deblurring/frame interpolation. The critical distinction lies in our method's capacity to extract potential camera motion trajectories from the event stream, thereby enhancing subsequent 3D vision tasks with additional geometric insights.

3 Methodology

Given a single blurry image and its corresponding event stream, our method recovers the underlying 3D scene representation and the camera motion trajectory jointly. The details of our method are shown in Fig. 2. In particular, we represent the 3D scene with neural radiance fields and the camera motion trajectory with a cubic B-Spline in SE(3) space. Both the blurry image and accumulated events within a time interval can thus be synthesized from the 3D scene representation providing the camera poses. The camera trajectory, NeRF, are then optimized by minimizing the difference between the synthesized data and the real measurements. The details are as follows.

3.1 Neural Implicit Representation

We adopt Multi-layer Perceptron (MLP) to represent the 3D scene as the original NeRF [28]. More advanced representations, such as multi-resolution hash encoding [30], can also be exploited to further improve its performance. In particular, the scene model is represented by a learnable mapping function $\mathbf{F}_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, which requires a Cartesian coordinates $\mathbf{x} \in \mathbb{R}^3$ and a

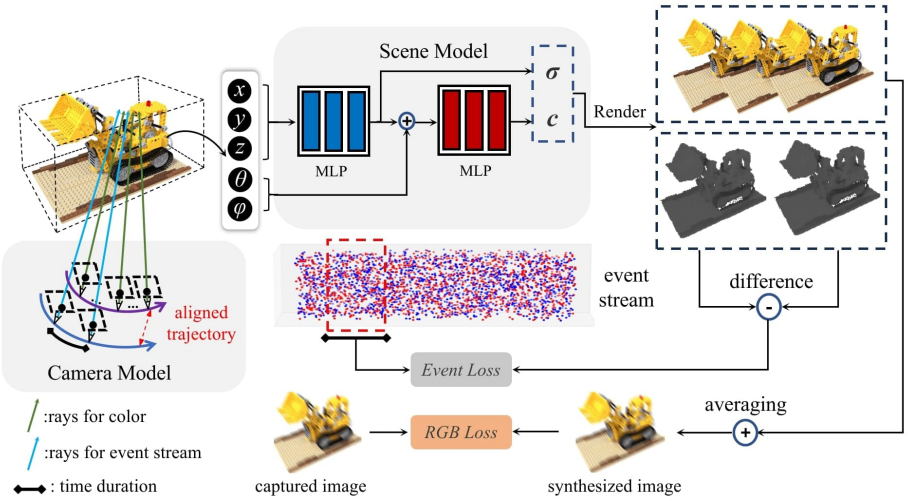

Fig. 2: The pipeline of our method. Given a motion blurry image and its corresponding event stream, we aim to recover both the implicit sharp scene representation and its camera motion trajectory within exposure time. We exploit a continuous time representation for the motion trajectory, and maximize the coherence between both the real measurements and synthesized data for the recovery.

viewing direction $\mathbf{d} \in \mathbb{S}^2$ as input, and outputs the corresponding volume density $\sigma \in \mathbb{R}$ and color $\mathbf{c} \in \mathbb{R}^3$. Both the 3D point $\mathbf{x}$ and viewing direction $\mathbf{d}$ are defined in the world coordinate frame. They are a function of the pixel coordinate, camera pose and the corresponding intrinsic parameters. To query the intensity of a pixel, we can apply volume rendering by sampling 3D points along the ray, which originates from the camera center and passes through the pixel. The volume rendering can be formally defined as follows:

$$\mathbf{I}(\mathbf{x}) = \sum_{i=1}^{n} T_i (1 - \exp(-\sigma_i \delta_i)) \boldsymbol{c}_i, \tag{1}$$

where n is the number of sampled points along the ray, $\boldsymbol{c}_i$ and σ_i are the predicted color and volume density of the i^{th} sampled 3D point via $\mathbf{F}_\theta$, δ_i is the distance between the i^{th} and $(i+1)^{th}$ sampled point, T_i is the transmittance which represents the probability that the ray does not hit any particle until the i^{th} sampled point. T_i can be computed via:

$$T_i = \exp(-\sum_{k=1}^{i-1} \sigma_k \delta_k). \tag{2}$$

3.2 Camera Motion Trajectory Modeling

We use a differentiable cubic B-Spline in SE(3) space to better model the camera motion trajectory. The spline is represented by a set of learnable control knots

$\boldsymbol{T}_{c_i}^{w} \in \mathbb{SE}(3)$ for $i = 0, 1, ..., n$. $\boldsymbol{T}_{c_i}^{w}$ represents the i^{th} control knot, which is defined as a transformation matrix from the camera coordinate frame to world frame. For brevity, we denote $\boldsymbol{T}_{c_i}^{w}$ with $\boldsymbol{T}_i$ for subsequent derivations. We assume the control knots are sampled with a uniform time interval Δt and the trajectory starts from t_0. Spline with a smaller Δt can represent a smoother motion, with an expense of more control knots to optimize. Since four consecutive control knots determine the value of the spline curve at a particular timestamp, we can thus compute the starting index of the four control knots for time t by:

$$k = \lfloor \frac{t - t_0}{\Delta t} \rfloor, \tag{3}$$

where $\lfloor * \rfloor$ is the floor operator. Then we can obtain the four control knots responsible for time t as $\boldsymbol{T}_k$, $\boldsymbol{T}_{k+1}$, $\boldsymbol{T}_{k+2}$ and $\boldsymbol{T}_{k+3}$. We can further define $u = \frac{t-t_0}{\Delta t} - k$, where $u \in [0, 1)$ to transform t into a uniform time representation. Using this time representation and based on the matrix representation for the De Boor-Cox formula [35], we can write the matrix representation of a cumulative basis $\mathcal{B}$(u) as

$$\mathcal{B}(u) = \mathcal{M} \begin{bmatrix} 1 \\ u \\ u^2 \\ u^3 \end{bmatrix}, \quad \mathcal{M} = \frac{1}{6} \begin{bmatrix} 6 & 0 & 0 & 0 \\ 5 & 3 & -3 & 1 \\ 1 & 3 & 3 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

The pose at time t can be computed as:

$$\boldsymbol{T}(u) = \boldsymbol{T}_k \cdot \prod_{j=0}^{2} \exp(\mathcal{B}(u)_{j+1} \cdot \boldsymbol{\Omega}_{k+j}), \tag{5}$$

where $\mathcal{B}(u)_{j+1}$ denotes the $(j+1)^{th}$ element of the vector $\mathcal{B}(u)$, $\boldsymbol{\Omega}_{k+j} = \log(\boldsymbol{T}_{k+j}^{-1} \cdot \boldsymbol{T}_{k+j+1}))$.

Since we only consider the continuous camera motion corresponding to a single blurry image, the time interval is thus usually short. We found that four control knots would be sufficient to deliver satisfying results. Therefore, we use the minimal configuration for the following experiments and they are initialized randomly around the identity pose.

3.3 Blurry Image Formation Model

A motion blurred image $\mathbf{B}(\mathbf{x}) \in \mathbb{R}^{\mathrm{W} \times \mathrm{H} \times 3}$ is physically formed by collecting photons during the exposure time and it can be mathematically modeled as:

$$\mathbf{B}(\mathbf{x}) \approx \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{I}_{\mathrm{i}}(\mathbf{x}), \tag{6}$$

where both W and H are the width and height of the image respectively, n is the number of sampled images, $\mathbf{x} \in \mathbb{R}^2$ represents the pixel location, $\mathbf{I}_{\mathrm{i}}(\mathbf{x}) \in \mathbb{R}^{\mathrm{W} \times \mathrm{H} \times 3}$ is the i^{th} virtual sharp image sampled within the exposure time. The virtual sharp images can be rendered from the neural 3D scene representation along the previous defined camera trajectory. It can be seen that $\mathbf{B}(\mathbf{x})$ is differentiable with respect to the parameters of NeRF and the motion trajectory.

3.4 Event Data Formation Model

An event camera records changes of the brightness as a stream of events asynchronously. Every time a pixel brightness change reaches a contrast threshold (i.e. $|L(\mathbf{x}, t_i+\delta t)-L(\mathbf{x}, t_i)| \geq C$), the camera will trigger an event $e_i = (\mathbf{x}, t_i, p_i)$, where $p_i \in (-1, +1)$ is the polarity of the event, $L(\mathbf{x}, t_i) = \log(\mathbf{I}(\mathbf{x}, t_i))$ is the brightness logarithm of pixel $\mathbf{x}$ at timestamp t_i, C is the contrast threshold.

To relate NeRF representation with the event stream, we accumulate the real measured events within a time interval Δt to an image $\mathbf{E}(\mathbf{x})$. The accumulation is defined as:

$$\mathbf{E}(\mathbf{x}) = C\{e_i(\mathbf{x}, t_i, p_i)\}_{t_k<t_i<t_k+\Delta t}, \tag{7}$$

where $e(\mathbf{x}, t_i, p_i)$ is the i^{th} event within the defined time interval corresponding to pixel $\mathbf{x}$. For real event cameras, the contrast threshold C changes over time and varies pixel by pixel. We therefore normalize the accumulated event as in [9, 17] to eliminate the effect of unknown C:

$$\mathbf{E}_n(\mathbf{x}) = \frac{\mathbf{E}(\mathbf{x})}{\|\mathbf{E}(\mathbf{x})\|_2}, \tag{8}$$

Given the interpolated start pose and end pose corresponding to the time interval Δt from the spline, we are able to render two gray-scale images (i.e. $\mathbf{I}_{start}$ and $\mathbf{I}_{end}$) from NeRF. The synthesized accumulated event image $\hat{\mathbf{E}}$ can then be computed as:

$$\hat{\mathbf{E}}(\mathbf{x}) = \log(\mathbf{I}_{end}(\mathbf{x})) - \log(\mathbf{I}_{start}(\mathbf{x})), \tag{9}$$

where $\hat{\mathbf{E}}(\mathbf{x})$ depends on the parameters of both the cubic spline and NeRF, and is differentiable with respect to them. We can also normalize $\hat{\mathbf{E}}(\mathbf{x})$ to $\hat{\mathbf{E}}_n(\mathbf{x})$ similarly as in Eq. (8) for loss computation.

3.5 Loss Functions

We minimize the sum of a photo-metric loss $\mathcal{L}_p$ and an event loss $\mathcal{L}_e$:

$$\mathcal{L}_{total} = \mathcal{L}_p + \beta\mathcal{L}_e, \tag{10}$$

where $\mathcal{L}_p$ represents the loss for the frame-based camera, $\mathcal{L}_e$ represents the loss for the accumulated events within a randomly sampled time interval, and β is a hyper-parameter. Both losses are defined as follows:

$$\mathcal{L}_p = \left\|\mathbf{B}(\mathbf{x}) - \hat{\mathbf{B}}(\mathbf{x})\right\|^2, \tag{11}$$

$$\mathcal{L}_e = \left\|\mathbf{E}_n(\mathbf{x}) - \hat{\mathbf{E}}_n(\mathbf{x})\right\|^2, \tag{12}$$

where $\hat{\mathbf{B}}(\mathbf{x})$ is the real captured blurry image.

4 Experiments

4.1 Experimental Setup

Synthetic datasets. We generate synthetic datasets for both quantitative and qualitative evaluations via Unreal Engine [1] and Blender [7]. To have a more realistic synthesis, we interpolate the real camera motion trajectories from ETH3D [41] to render high frame-rate images. In total, we generate three sequences (i.e. livingroom, whiteroom and pinkcastle) via Unreal Engine and two sequences (i.e. tanabata and outdoorpool) via Blender. For thorough evaluations, we synthesized twenty blurry images and corresponding event streams for each sequence. The event streams are generated via ESIM [37] from high frame-rate video. Furthermore, we additionally employed the synthetic dataset proposed by E^{2}NeRF to compare our method with NeRF-based image-deblur methods which require multi-view training data. This dataset includes synthesized blurry images paired with their respective event streams, which expands upon the six scenes (i.e. chair, ficus, hotdog, lego, materials and mic).

Real-world datasets. We utilized the real-world datasets proposed by E^{2}NeRF, captured using the DAVIS346 color event camera in real-world scenarios. The dataset encompass five challenging scenes (i.e. letter, lego, camera, plant and toys). The exposure time for RGB frames was set to 100ms, resulting in occurrences of complex camera motion and severe motion blur within the time interval.

Baseline methods and evaluation metrics. To evaluate the performance of our method in terms of image deblurring, we compare it against state-of-the-art deep learning-based single image deblur methods, i.e. SRN-Deblur [46], HINet [5], DeblurGANv2 [19], MPRNet [55], NAFNet [4] and Restormer [54], as well as event-enhanced single image-based deblur methods, i.e. EDI [33], eSLNet [47]. We also compared our method with NeRF-based image deblur method using multi-view information. The quality of the rendered image is evaluated with the commonly used PSNR, SSIM and LPIPS [56] metrics. Since the lack of sharp reference images in real-world datasets, we conducted quantitative analysis experiments on five real scenes using the no-reference image quality assessment metrics BRISQUE [29].

Implementation details. We implement our method with PyTorch. The implicit representation of the scene from MLP (i.e. F_θ) is built from NeRF [28] without any modification. We randomly initialize trajectory control knots for cameras within a range of $(0, 0.01)$. We use two separate Adam optimizers [15] for the scene model (i.e. F_θ) and camera motion (i.e. $\mathbf{P}_i$). The learning rate for the scene model and poses decay from 5×10^{-4} with a rate of 0.1 exponentially. In each training step, 1024 pixels for brightness and 1024 pixels for color are sampled. The weight of the event loss β is selected to be 0.1 for synthetic datasets and 2 for real-world datasets, respectively. We train our model for 80K iterations for each image and its corresponding event stream.

Table 1: Ablation studies on the number of virtual sharp images. The results demonstrate that the image quality gradually saturates as the number of virtual sharp images increases. By compromising the image quality and computational efficiency, we select $n = 19$ for all our experiments.

	Livingroom			Tanabata		
n	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
7	36.16	.9291	.0711	29.28	.8533	.0712
11	36.84	.9353	.0659	30.77	.8788	.0622
15	37.03	.9368	.0635	31.90	.8979	.0529
19	37.11	.9370	.0632	32.14	.9015	.0515
23	37.11	.9375	.0629	32.35	.9042	.0506

Table 2: Ablation studies on event accumulation time lengths. The experimental results demonstrate that the image quality can be affected by the time length. For all experiments, we select $\alpha = 0.1$ for event data accumulation.

	Livingroom			Tanabata		
α	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
0.05	36.66	.9325	.0724	31.86	.8977	.0538
0.10	37.11	.9370	.0632	32.14	.9015	.0515
0.15	37.21	.9376	.0601	32.19	.9019	.0509
0.20	37.16	.9369	.0596	32.20	.9016	.0499
0.25	37.11	.9358	.0589	32.21	.9015	.0496

4.2 Ablation Study

We evaluate the performance of our method under various configurations. To quantify the differences, we exploit two synthetic datasets (i.e. Livingroom and Tanabata) for the experiments.

Effect of the number of virtual sharp images. We evaluate the effect of different numbers of the interpolated virtual images as mentioned in Eq. (6). The experimental results are presented in Table 1. It demonstrates that more virtual images deliver better image quality, while requiring more computational resources. By compromising the image quality and computation requirement, we choose $n = 19$ for our experiments.

Effect of time lengths for event accumulation. We study the effect of different time lengths Δt for event accumulation as in Eq. (7). The timestamps of the event stream are normalized to a range of $(0, 1)$ by its total time length. We choose different values from 0.05 to 0.25 for the ablation study. The experimental results are presented in Table 2, showing that the performance on Tanabata dataset gradually saturates as α increases until to 0.25, whereas on Livingroom dataset, performance initially improves with increasing α but subsequently declines. This may depend on the noise level in event stream of different time lengths. In all experiments, we select $\alpha = 0.1$.

Effect of trajectory representations. We explore the difference between linear interpolation and cubic B-Spline to represent the motion trajectory. The

Table 3: Ablation studies on trajectory representations. The results demonstrate that cubic B-spline can deliver better performance than linear interpolation.

	Livingroom			Tanabata		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
linear interpolation	34.16	.9035	.1133	27.42	.8164	.1117
cubic B-Spline	37.11	.9370	.0632	32.14	.9015	.0515

Table 4: Quantitative comparisons on single image deblurring with synthetic datasets. The results demonstrate that our method performs better than prior learning-based methods in terms of image quality. For HINet and NAFNet, we tests pre-trained weights from both GoPro and REDS datasets(*). Due to page limits, the results of the SSIM metric can be found in the supplementary materials.

	PSNR ↑					
	Livingroom	Whiteroom	Pinkcastle	Tanabata	Outdoorpool	*Average*
DeblurGANv2 [19]	29.26	27.64	23.16	20.09	26.89	25.41
SRN-deblur [46]	30.86	27.59	23.12	19.89	27.79	25.85
MPRNet [55]	28.57	26.49	21.60	18.20	27.02	24.38
HINet [5]	28.56	26.27	21.91	18.59	26.70	24.41
HINet* [5]	27.55	22.89	20.25	18.15	27.14	23.20
NAFNet [4]	29.92	28.16	22.41	18.96	26.75	25.24
NAFNet* [4]	28.18	23.67	20.85	18.38	27.52	23.72
Restormer [54]	29.48	27.39	22.22	18.82	27.35	25.05
BeNeRF	37.11	32.95	29.68	32.14	36.38	33.65
	LPIPS ↓					
	Livingroom	Whiteroom	Pinkcastle	Tanabata	Outdoorpool	*Average*
DeblurGANv2 [19]	.2087	.1989	.2608	.3934	.3100	.2744
SRN-deblur [46]	.2529	.2503	.3245	.4260	.3594	.3226
MPRNet [55]	.2621	.2564	.3586	.4173	.3679	.3325
HINet [5]	.2468	.2620	.3500	.4024	.3355	.3193
HINet* [5]	.3327	.3602	.3789	.5265	.4397	.4076
NAFNet [4]	.2268	.1991	.3058	.3908	.3280	.2901
NAFNet* [4]	.3182	.3566	.3943	.5271	.4257	.4044
Restormer [54]	.2391	.2493	.3373	.4248	.3664	.3234
BeNeRF	.0632	.0788	.0761	.0515	.0677	.0675

experimental results shown in Table 3 demonstrate that cubic B-Spline deliver better performance on complex motions than that of the linear interpolation. We exploit cubic B-Spline for the experiments.

4.3 Quantitative evaluations

To evaluate the performance of our method, we compare it against single-image deblurring methods, event-enhanced single-image deblurring methods, and NeRF-based image deblurring methods requiring multi-view information on both synthetic datasets and real datasets. The experimental results are presented in Table 4, Table 5, Table 6 and Table 7.

In particular, we compare against SRN-Deblur [46], DeblurGANv2 [19], MPRNet [55], HINet [5], NAFNet [4], Restormer [54] in terms of single image deblur-

Table 5: Quantitative comparisons on event-enhanced single image deblurring with synthetic datasets. The results demonstrate that our method performs better than both EDI and eSLNet. Due to page limits, the results of the SSIM metric can be found in the supplementary materials.

	PSNR ↑					
	Livingroom	Whiteroom	Pinkcastle	Tanabata	Outdoorpool	*Average*
eSLNet [47]	14.22	10.81	10.49	8.86	11.80	11.24
EDI [33]	32.61	30.33	27.24	24.87	31.64	29.34
BeNeRF	37.11	32.95	29.68	32.14	36.38	33.65
	LPIPS ↓					
	Livingroom	Whiteroom	Pinkcastle	Tanabata	Outdoorpool	*Average*
eSLNet [47]	.3981	.4236	.4902	.5067	.4676	.4572
EDI [33]	.0904	.1020	.0779	.1039	.1409	.1030
BeNeRF	.0632	.0788	.0761	.0515	.0677	.0675

ring. The experimental results shown in Table 4 demonstrates that our method significantly outperforms prior state-of-the-art methods. It shows that prior learning-based methods have limited generalization performance on domain-shifted images, especially with large motion blurs.

Table 6: Quantitative comparisons on NeRF-based image deblurring with synthetic datasets from E^{2}NeRF. The results indicate that our method outperforms both NeRF and Deblur-NeRF, and exhibits performance comparable to E^{2}NeRF in terms of the PSNR metric. Moreover, our method even surpasses E^{2}NeRF with the LPIPS metric. Due to page limits, the results of the SSIM metric can be found in the supplementary materials.

	PSNR ↑						
	Chair	Ficus	Hotdog	Lego	Materials	Mic	*Average*
NeRF [28]	24.29	22.98	27.75	21.95	19.99	20.50	22.91
Deblur-NeRF [26]	25.87	22.86	24.62	24.47	20.54	11.92	21.71
E^{2}NeRF [34]	31.28	30.00	34.34	28.11	27.27	27.60	29.77
BeNeRF	31.17	30.81	34.31	28.09	27.44	26.13	29.66
	LPIPS ↓						
	Chair	Ficus	Hotdog	Lego	Materials	Mic	*Average*
NeRF [28]	.1254	.1037	.1158	.2103	.1512	.1579	.1441
Deblur-NeRF [26]	.2185	.1541	.2138	.2053	.2562	.3706	.2364
E^{2}NeRF [34]	.0608	.0362	.0660	.1078	.0919	.0724	.0725
BeNeRF	.0500	.0299	.0539	.0745	.0708	.0738	.0588

We also compare against prior event-enhanced single image deblurring methods, such as EDI [33] and eSLNet [47]. The results in Table 5 demonstrates that our method has superior performance when compared to them. eSLNet demonstrates poor generalization performance, since we are unable to fine-tune it on our evaluation datasets. It demonstrate the benefit on incorporating event streams to enhance single image deblurring task under the framework of NeRF.

Furthermore, we conducted detailed comparisons with NeRF-based image deblurring methods that require multi-view information on the synthetic dataset proposed by E^{2}NeRF. We compared against NeRF [28], Deblur-NeRF [26] and E^{2}NeRF [34]. The experimental results in Table 6 demonstrate that despite utilizing only a single blurred image and event stream of a small time interval, our method achieves performance comparable to E^{2}NeRF [34], which utilizes multi-view images and a longer event stream, in terms of PSNR metric. Moreover, our method even surpasses E^{2}NeRF [34] in terms of the LPIPS metric.

Finally, we select the best-performing algorithms on the synthetic dataset, excluding our method, from single-image deblurring methods, event-enhanced single-image deblurring methods, and NeRF-based image deblurring methods, which are SRN-deblur [46], EDI [33], and E^{2}NeRF [34], respectively. We compare against with these methods on the real dataset and provide the BRISQUE [29] metric. The results in Table 7 indicate a significant improvement of our method over the aforementioned methods. This is attributed to our method's incorporation of a physical model for the imaging process of blurry images, enabling better performance on real-world datasets.

Table 7: Quantitative comparisons on real-world datasets from E^{2}NeRF. We exploit the used real-world dataset proposed by E2NeRF [34] for the evaluations, which is collected via a DAVIS event camera. The results indicates that our method outperforms EDI, SRN-Deblur and even E^{2}NeRF on the real-world datasets. Note that E^{2}NeRF [34] requires multi-view images while ours only need a single image. Since E^{2}NeRF does not provide the trained model and the code for the metric computation, we re-trained E^{2}NeRF for this experiment and compute the metric with the MATLAB implementation of the BRISQUE metric for fair comparisons.

	BRISQUE ↓					
	Camera	Lego	Letter	Plant	Toys	*Average*
EDI [33]	29.74	29.35	28.74	31.09	37.09	31.20
SRN-Deblur [46]	32.20	34.91	40.82	37.45	46.10	38.30
E^{2}NeRF [34]	33.40	33.85	37.41	32.02	43.00	35.94
BeNeRF	19.47	25.86	27.37	21.46	25.20	23.87

4.4 Qualitative evaluations

The qualitative evaluation results are shown in Fig. 3 and Fig. 4 for both synthetic and real datasets respectively. The experimental results demonstrate that our method deliver better performance than prior methods even when the image is severely blurred. In particular, Fig. 3 shows that prior learning-based methods struggle to generalize to domain-shifted images. Notably, EDI [33] performs well on synthetic datasets due to the high quality of event data. Fig. 4 shows that our method outperforms all prior methods (even trained with multi-view images) on real noisy dataset, which demonstrates the advantage of our method

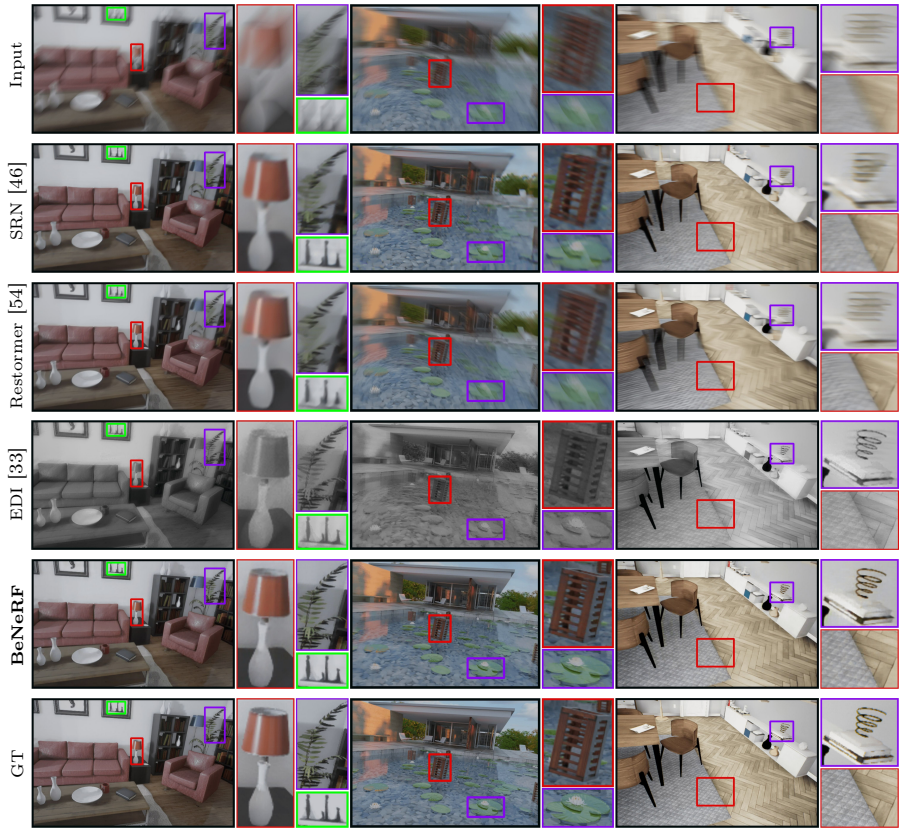

Fig. 3: Qualitative results of different methods with synthetic datasets. It demonstrates that our method delivers better performance compared to prior approaches. The learning based methods fail to generalize on severely blurry images.

and the necessity to jointly optimize the camera motion and the implicit 3D representation.

5 Conclusion

In conclusion, we present a novel method to jointly recover the underlying 3D scene representation and camera motion trajectory from a single blurry image and its corresponding event stream. Extensive experimental evaluations with both synthetic and real datasets demonstrate the superior performance of our method over prior works, even for those requiring multi-view images and longer event streams.

Acknowledgements. This work was supported in part by NSFC under Grant 62202389, in part by a grant from the Westlake University-Muyuan Joint Research Institute, and in part by the Westlake Education Foundation.

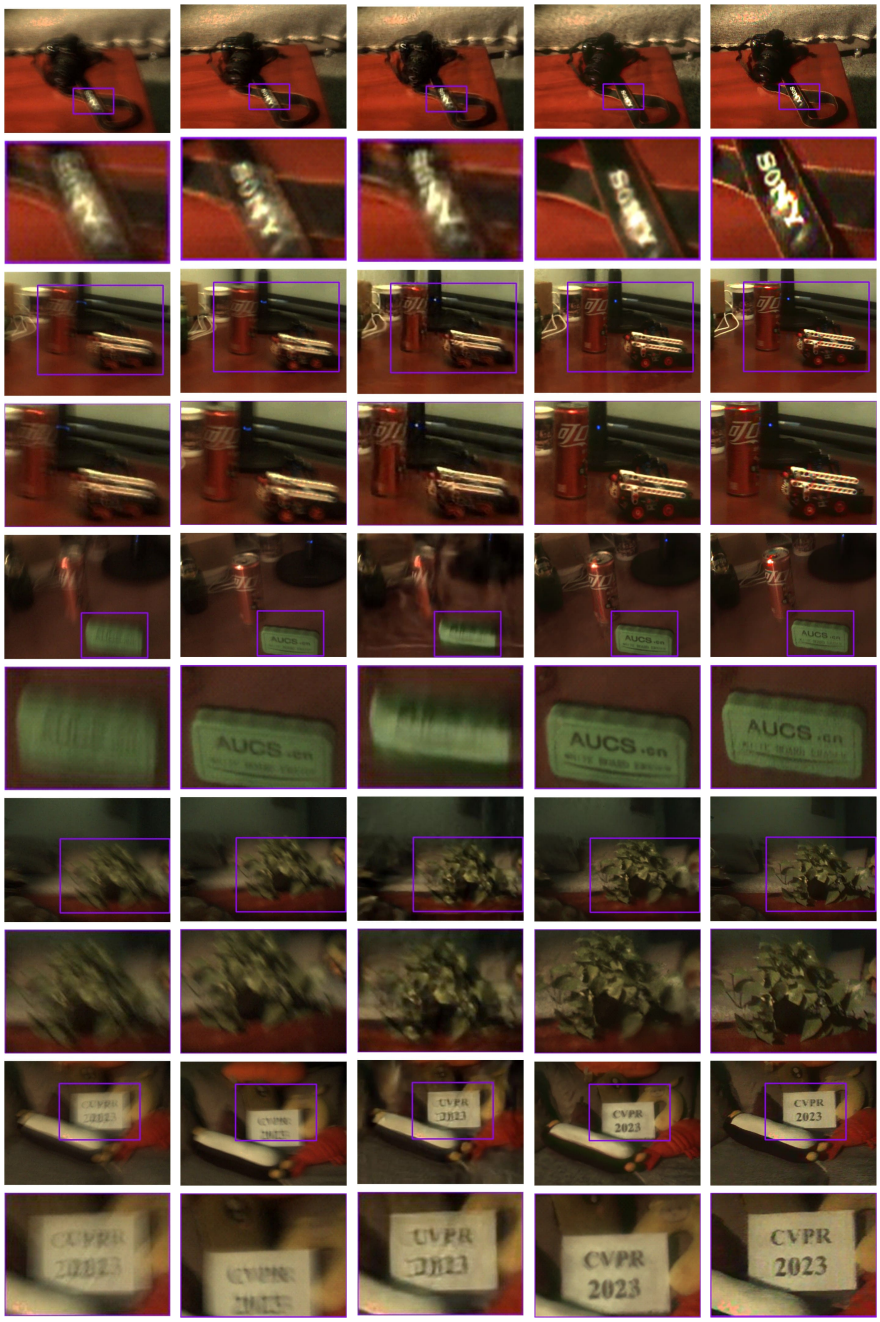

Fig. 4: Qualitative results of different methods on the real datasets. The experimental results demonstrate that our method delivers superior performance on the real DAVIS datasets from E^{2}NeRF. We are even achieve better performance than prior methods requiring multi-view images and longer event stream.

References

1. Unreal Engine: The most powerful real-time 3D creation tool. https://www.unrealengine.com/en-US/
2. Bian, W., Wang, Z., Li, K., Bian, J., Prisacariu, V.A.: NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
3. Cai, S., Obukhov, A., Dai, D., Van Gool, L.: Pix2NeRF: Unsupervised Conditional pi-GAN for Single Image to Neural Radiance Fields Translation. In: CVPR (2022)
4. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple Baselines for Image Restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 17–33. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_2
5. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: HINet: Half Instance Normalization Network for Image Restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021)
6. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
7. Foundation, B.: Blender.org - Home of the Blender project - Free and Open 3D Creation Software
8. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-Based Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 154–180 (Jan 2022). https://doi.org/10.1109/TPAMI.2020.3008413
9. Hidalgo-Carrió, J., Gallego, G., Scaramuzza, D.: Event-aided Direct Sparse Odometry (Apr 2022). https://doi.org/10.48550/arXiv.2204.07640
10. Hwang, I., Kim, J., Kim, Y.M.: Ev-nerf: Event based neural radiance field. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 837–847 (2023)
11. Jia, J.: Single image motion deblurring using transparency. In: 2007 IEEE Conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
12. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning Event-Based Motion Deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2020)
13. Jin, M., Meishvili, G., Favaro, P.: Learning to extract a video sequence from a single motion-blurred image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6334–6342 (2018)
14. Joshi, N., Zitnick, C.L., Szeliski, R., Kriegman, D.J.: Image deblurring and denoising using color priors. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1550–1557. IEEE (2009)
15. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (Jan 2017). https://doi.org/10.48550/arXiv.1412.6980
16. Klenk, S., Koestler, L., Scaramuzza, D., Cremers, D.: E-nerf: Neural radiance fields from a moving event camera. IEEE Robotics and Automation Letters (2023)
17. Klenk, S., Koestler, L., Scaramuzza, D., Cremers, D.: E-NeRF: Neural Radiance Fields From a Moving Event Camera. IEEE Robotics and Automation Letters **8**(3), 1587–1594 (Mar 2023). https://doi.org/10.1109/LRA.2023.3240646

18. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8183–8192 (2018)
19. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8878–8887 (2019)
20. Lee, D., Oh, J., Rim, J., Cho, S., Lee, K.M.: Exblurf: Efficient radiance fields for extreme motion blurred images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
21. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithms. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1964–1971. IEEE (2009)
22. Li, M., Wang, P., Zhao, L., Liao, B., Liu, P.: USB-NeRF: Unrolling Shutter Bundle Adjusted Neural Radiance Fields. In: International Conference on Learning Representations (ICLR) (2024)
23. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128\times 128 120 dB 15 Ms Latency Asynchronous Temporal Contrast Vision Sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (Feb 2008). https://doi.org/10.1109/JSSC.2007.914337
24. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: BARF: Bundle-Adjusting Neural Radiance Fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)
25. Low, W.F., Lee, G.H.: Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
26. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblur-NeRF: Neural Radiance Fields From Blurry Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)
27. 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. CVPR (2022)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis (Aug 2020). https://doi.org/10.48550/arXiv.2003.08934
29. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
30. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics **41**(4), 1–15 (Jul 2022). https://doi.org/10.1145/3528223.3530127
31. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
32. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
33. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6820–6829 (2019)

34. Qi, Y., Zhu, L., Zhang, Y., Li, J.: E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13254–13264 (2023)
35. Qin, K.: General matrix representations for B-splines. In: Proceedings Pacific Graphics '98. Sixth Pacific Conference on Computer Graphics and Applications (Cat. No.98EX208). pp. 37–43 (Oct 1998). https://doi.org/10.1109/PCCGA.1998.731996
36. Rebain, D., Matthews, M., Yi, K.M., Lagun, D., Tagliasacchi, A.: LOLNeRF: Learn from One Look. In: CVPR (2022)
37. Rebecq, H., Gehrig, D., Scaramuzza, D.: ESIM: An Open Event Camera Simulator. In: Proceedings of The 2nd Conference on Robot Learning. pp. 969–982. PMLR (Oct 2018)
38. Rematas, K., Martin-Brualla, R., Ferrari, V.: ShaRF: Shape-conditioned Radiance Fields from a Single View. In: ICML (2021)
39. Rudnev, V., Elgharib, M., Theobalt, C., Golyanik, V.: EventNeRF: Neural Radiance Fields From a Single Colour Event Camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4992–5002 (2023)
40. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Schops, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 134–144 (2019)
42. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. Acm transactions on graphics (tog) **27**(3), 1–10 (2008)
43. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 769–777 (2015)
44. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 412–428. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19797-0_24
45. Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18043–18052 (2023)
46. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-Recurrent Network for Deep Image Deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8174–8182 (2018)
47. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event Enhanced High-Quality Image Recovery. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 155–171. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_10
48. Wang, P., Zhao, L., Ma, R., Liu, P.: BAD-NeRF: Bundle Adjusted Deblur Neural Radiance Fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4170–4179 (2023)
49. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF–: Neural Radiance Fields Without Known Camera Parameters (Apr 2022). https://doi.org/10.48550/arXiv.2102.07064

50. Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.S., Jia, X., Qiao, Z., Liu, J.: Motion Deblurring With Real Events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2583–2592 (2021)
51. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. pp. 157–170. Springer (2010)
52. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
53. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
54. 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)
55. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-Stage Progressive Image Restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14821–14831 (2021)
56. 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 Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
57. Zhang, X., Yu, L., Yang, W., Liu, J., Xia, G.S.: Generalizing Event-Based Motion Deblurring in Real-World Scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10734–10744 (2023)
58. Zhao, L., Wang, P., Liu, P.: BAD-Gaussians: Bundle Adjusted Deblur Gaussian Splatting. In: European Conference on Computer Vision (ECCV). Springer (2024)