RnGCam: High-speed video from rolling & global shutter measurements

Supplementary Material - RnGCam

This appendix material is organized as follows. In Appendix A, we implement several processing steps to properly apply the proposed method to the hardware data, including image size alignment, white balance correction, and memory limitations during reconstruction. In Appendix B, we provide details on the prototype for creating an RnG Cam. We present an affordable method for making a random diffuser, along with specifications on the optical system. This includes information on achieving shift invariance in the optical system and calibrating the point spread function (PSF). Additionally, we will discuss the proper setup for the system time settings and the neural space-time model. We also present additional results, including ablations, and full model results based on experimental data in Appendix C.

A. Handling different sensor sensitivities and resolution

A.1. Aligning rolling and global shutter images

To align measurements from the two sensors, we simultaneously capture a static calibration image on both camera. We deconvolve the diffuser-coded image on the rolling shutter to obtain the scene from the RS viewpoint. The calibration measurement from the GS sensor is then aligned with the deconvolved RS scene by aligning two features in the scene with a scale and rotate transformation.

A.2. White balance correction between global and rolling shutter measurements

The GS and RS sensors have different sensitivities per channel. Additionally, they have different optical elements in front of the sensors. The GS sensor has a lens, while the RS has a random optical diffuser.

To calibrate the two sensor white balance and energy levels, we predict the per-channel color correction coefficients using the static INR as 3 extra outputs $\lambda_r, \lambda_g, \lambda_b$ which we represent as a matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_r & 0 & 0 \\ 0 & \lambda_g & 0 \\ 0 & 0 & \lambda_b \end{bmatrix}. \tag{1}$$

We slightly modify the optimization objective in eq. 14 in the main text to incorporate the correction factor $\mathbf{\Lambda}$ as follows:

$$\widehat{\theta} = \arg\min_{\theta} \|\mathbf{A}_G \mathbf{v}_\theta - \mathbf{b}_G \Lambda\|_2^2 + \psi \|\mathbf{A}_R \mathbf{v}_\theta - \mathbf{b}_R\|_2^2 + \tau T V_S(\mathbf{U}). \tag{2}$$

A.3. Memory limit, evaluating subset of INR

In our current implementation, there is a field stop, which makes the image 0 outside the region defined by the field stop. For memory efficiency, and reducing the extent of downsampling, we evaluate the model only inside the field stop region, containing nonzero intensities.

A.4. Modeling details

Equation (3) describes the forward model of a general camera with static psf $h(x, y)$ and shutter function $S(x, y, t)$. Substituting the discrete implementation of linear convolution, (2), into (3) yields

$$b(x, y) = \sum_T S(x, y, t) \mathbf{C} \left[\mathbf{P}\left(h(x, y)\right) \overset{(x,y)}{\circledast} \mathbf{P}\left(v(x, y, t)\right) \right]$$

We represent this compactly as a matrix-vector multiply in (4). The system matrix can be conceptualized as

$$\mathbf{A} = \mathbf{\Sigma} \text{diag}(\mathbf{S}) \mathbf{C} \mathbf{F}^{-1} \text{diag}(\mathbf{F} \mathbf{P} \mathbf{h}) \mathbf{F}.$$

Here, $\mathbf{\Sigma}$ is the matrix for summation over time. Pointwise multiplication by the shutter function, S, is described by $\text{diag}(\mathbf{S})$, which is a diagonal matrix comprised of the column-stacked shutter indicator, denoted $\mathbf{S}$. $\mathbf{F}$ is the 2D Discrete Fourier Transform matrix, $\mathbf{h}$ is the column-stacked point spread function (PSF), $\mathbf{P}$ and $\mathbf{C}$ are the matrix forms of zero-padding and cropping, respectively. Note that, in practice, we implement the camera model using operators; matrices are used only for compact notation here.

B. RnG Cam Prototype Detail

The overview of the RnG Cam is illustrated in Fig. 1. In the following subsections, we will discuss the importance of a well-designed diffuser and relay lens in achieving system shift invariance and extending the bandwidth limit of the measurement.

B.1. Diffuser Design and Manufacturing

A diffuser must fulfill three essential requirements: First, its PSF should create a random pattern to avoid periodicity. Second, it should cover as much of the sensor area as possible to enhance the bandwidth of a snapshot. Third, the feature size needs to be small enough to be sensitive to motion. To achieve this, we create randomly positioned unifocal lenslets using 9/16-inch ball bearings, resulting in a focal length of approximately 28 mm. With this focal length, we can place the diffuser against the camera housing to generate a sharp point on the sensor when the incident light is collimated.

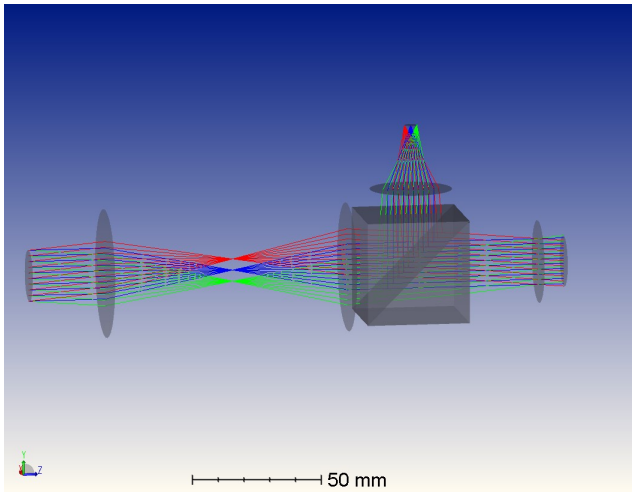

Figure 1. 3D Zemax design

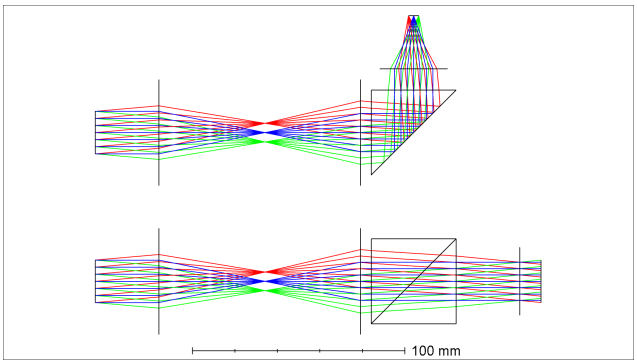

Figure 2. 2D Zemax diagram

The diffuser we used for the experiment is both low-cost and easy to make. First, randomly dent the polished aluminum block using a 9/16-inch stainless steel ball. Then, apply optical epoxy to cover the dented area. Next, place a clear cover slide over the epoxy and cure it using the appropriate wavelength of UV light. Finally, the diffuser is created by peeling off the cover slide from the aluminum block.

B.2. Shift Invariant Imaging System

In the inverse problem of the diffuser cam, it is essential to position the diffuser at the aperture stop of the optical system to ensure that the problem remains shift invariant to simplify the deconvolution and system calibration. Our setup includes two sensors that utilize the same camera lens as the primary lens, but the aperture stop is inaccessible because it is located within the camera lens. Thus, the follow-up collimated lens serves two purposes: one is to collimate the light for the beam splitter to reduce aberrations, and the other is to reimage the stop plane to a physically accessible location as the bottom diagram in Fig.2 shows.

B.3. Optical System Details

We present the prototype of the system in Fig. 6. The system uses global- and rolling-shutter cameras to capture the same scene simultaneously through a primary lens followed by a relay optical system with a beam splitter at the conjugate plane. The primary lens is a Sigma 50mm EX DG HSM Lens with f-number 1.4, and a field stop is set at the focal plane to control the field of view. Then, two 2-inch 98 mm focal length doublets with effective focal lengths around 45 mm collimate the light and image the aperture of the primary lens to a physically accessible plane where the diffuser will be located to avoid vignetting. The distance between the two doublets' last surface and the pupil's image is designed to be sufficient to fit in a 1 inch visible light beam splitter, which has 90/10 uneven energy distribution. The RS arm requires more energy because of spatial multiplexing. Therefore, the higher intensity arm has a 1" format RS camera (Basler ace acA5472-17uc, IMX183) with a diffuser containing random microlenses with a 9/16-inch radius and an effective focal length of around 28mm. The weaker intensity arm of the beam splitter has one 1/1.2" format global shutter camera (Basler ace acA1920-155uc, IMX174) with a Fujinon 25mm 1.4 f-number machine vision lens to form the static reference image.

A function generator syncs cameras with a hardware trigger to start simultaneously, triggering the subsequent three global shutter frames between a RS frame.

B.4. PSF calibration

The point spread function of the system, h, shown in Figure 6 (a) is experimentally obtained by shining a point light source to the main lens of the system shown in Figure 6 (b).

B.5. System Timing

The downsampling factor of 16 in space and time was chosen because of a combination of memory limitations and the very high resolution of the RS sensor ($T, W, H \approx 3648 \times 5472 \times 3648$). Because our effective frame rate is limited by the downsampling we have to fit the coordinate network to our machine (48GB NVIDIA A40), a lower resolution RS sensor with a similar line time could get us to around $77,000$ fps.

B.6. Space-time fusion model implementation details

In our implementation, we instantiate the time-varying SIREN F_θ^D with 3 hidden layers, and 128 hidden features,

the static SIREN F_θ^S and motion SIREN F_θ^M with 2 hidden layers and 32 hidden features. We apply a non-negativity constraint on the outputs of F_θ^S and F_θ^D and apply a sigmoid activation to ensure that $\alpha \in [0, 1]$.

To increase the ability of the scene representation to represent higher frequency features [43], we apply positional encoding $\gamma(\cdot)$ to the 3D coordinate inputs $(x, y, t) \in \mathbf{X}$, where

$$\gamma(x) = (x, \cos(2^i \pi x), \sin(2^i \pi x), \dots), \text{for } i = 0, \dots, L-1. \tag{3}$$

$L \in \mathbb{Z}^+$ is a tunable parameter. Larger values of L increase the ability of the network to represent high-frequency information. However, as seen in [2, 34] large L values may introduce high-frequency distortions and overfitting in the final reconstructions. For all the subnetworks, $F_\theta^M, F_\theta^D, F_\theta^S$, we consider L as a tunable parameter.

F_θ^M and β is an additional weight the temporal total variation sparsity. (We found that β values from 10 - 10000 yielded the best results, depending on the scene).

The estimated measurements are used to compute the mean squared error with b_G and b_R and we minimize this loss with respect to the parameters of the coordinate networks. We minimize this using the Adam optimizer with learning rate 0.5e-5. We ran all experiments for a fixed number of iterations, with average total runtimes of approximately three hours.

C. Miscellaneous Results

C.1. Ablations on global shutter and motion regularization

In this section, we perform ablations to examine the contributions of different components of both the Neural Spacetime model and our design including both RS and GS measurements. The results are summarized in Tab. 1 and Fig. 3.

Without motion regularization we observe, for example, distortion in the appearance of the captions at the bottom of the frames. Without RS measurement, we do not correctly recover the initial appearance of smoke emerging from the barrel. Without the GS measurement, we do not correctly recover the full puff of smoke at the final frames. We demonstrate that all the components in our design together contribute to recovery complex motion with a dense background.

C.2. Full model results on experimental data

We visualize the full reconstruction of our network, with all of its intermediate components in Fig. 4. We present the magnitude of the motion encoding in b), the time-varying alpha mask which enables blending of the dynamic motion d) with the static background e) resulting in the reconstructed scene f).

label	GS	RS	Motion Reg	Static network	PSNR
Our method	✓	✓	✓	✓	**31.99 dB**
no static	✓	✓	✓	×	29.42 dB
no motionreg	✓	✓	×	✓	21.52 dB
no RS	✓	×	✓	✓	22.72 dB
no GS	×	✓	✓	✓	22.39

Table 1. Ablation test on bullet scene. See Fig. 7. We test the effects of removing individual parts of our model, including the static network, regularization on the motion field, rolling shutter measurements, and global shutter measurements. We show that a combination of every component yields the best reconstruction.

Figure 3. **Ablations on bullet scene.** We compare our method (b) to ablations removing different components of our design: c) we remove the static INR; d) without motion regularization; e) without RS measurements; f) without the three GS captures.

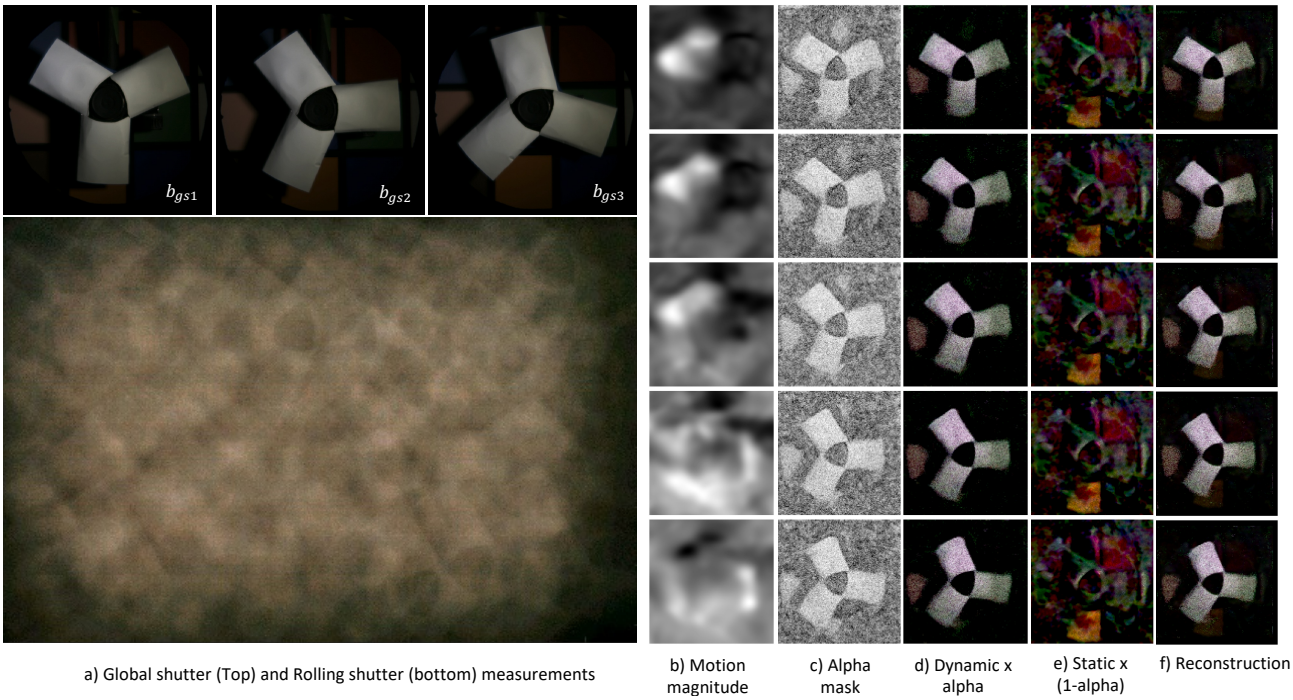

Figure 4. **Visualizing intermediate components of our method on an experimental scene. 1) Reconstructions of a spinning propeller (a)** Spatiotemporally encoded RS measurements (bottom), and the 3 global shutter measurements acquired over the same period (top). **(b)** Magnitude of the motion encoding from the motion network. **(c)** Time-varying alpha mask used to blend estimated static (e) and dynamic (d) scenes. **(d)** Dynamic estimate multiplied by alpha mask. **(e)** Static scene (contrast stretched for visualization) multiplied by (1-alpha mask). **(f)** Full scene reconstruction. **2) Reconstructions of a tennis ball leaving hand.** We demonstrate that our system is able to simulatenously resolve both the dense background, and the dynamics of the tennis ball.