Dark3R: Learning Structure from Motion in the Dark

Supplementary Material

We provide supplemental implementation details and results here, and ***video results and comparisons to baselines are provided in the project webpage.***

S1. Supplemental Implementation Details

S1.1. Datasets

Handheld-captured dataset. We fine-tune the model on a dataset comprising 21,688 well-exposed raw images captured across 92 distinct scenes, with approximately 100–400 images per scene (or about 200 on average—see Table S1). All images are captured using a Sony Alpha I camera in handheld burst mode at 10 frames per second while moving around each scene to ensure significant viewpoint overlap between adjacent frames. The exposure time is fixed to $1/200$ s to avoid motion blur, while the ISO is adjusted to achieve an exposure value (EV) of approximately -0.3 relative to the metered exposure, ensuring that no regions are over-exposed. The aperture is set to an f-number of at least $f/11$ to maximize depth of field, and decreased when necessary to maintain the target exposure time. After downsampling, our captures are clean and exhibit minimal noise.

We use a calibrated Poisson–Gaussian noise model for the camera to synthesize noisy raw images from the clean captures. For each clean image, we sample a random scaling factor to determine the mean pixel intensity, then compute the corresponding noise variance according to the noise model. Poisson-Gaussian noise is then added to produce images spanning a target range of signal-to-noise ratios (SNRs) between -1 and -7 dB. This noise synthesis is performed on the fly during fine-tuning, enabling continuous variation in brightness and noise conditions. All images, together with full capture metadata—including lens model, ISO, aperture, exposure time, and noise parameters—will be publicly released.

The noise model is calibrated following the procedure of Plötz and Roth [46]. For each RGB channel of the Sony Alpha 1's color filter array, we capture GretagMacbeth ColorChecker images at three exposure values and compute the mean–variance pairs over all homogeneous patches. Fitting a linear model to these measurements yields the noise parameters for each channel.

Tripod-captured (exposure-bracketed) dataset. We further evaluate our approach on a captured low-light dataset consisting of 12 scenes (see Fig. S2), each containing 300–500 viewpoints recorded using the same Sony Alpha I camera (see Table S2). For each scene, we acquire nine exposures per viewpoint in a bracketed sequence, spaced by 0.7 EV per step, yielding a dynamic range that spans from well-exposed to extremely low-SNR conditions. The lowest two exposures reach mean SNRs below 0 dB (as low as -5 dB on average and -10 dB at the pixel level). All captures are performed on a tripod using an aperture of $f/22$ for maximum depth of field and an ISO of 102,400, with exposure times adjusted according to the measured scene illuminance (typically 1–15 lux) to maintain consistent brightness across scenes. The longest exposure in each bracketed sequence serves as a reference for pose estimation: we compute ground-truth camera poses using COLMAP on these well-exposed images. For radiance field reconstruction, we train on 90% of the frames and hold out 10% for evaluation as described in the main text.

S1.2. Structure From Motion in the Dark

Input preprocessing. The inputs to the model are 14-bit raw sensor measurements, either simulated or captured using a Sony $\alpha 1$ mirrorless camera. Each captured raw image is center-cropped from the native camera resolution of 8640×5760 pixels to 8196×5632, then demosaiced by subsampling pixels from each color channel in the Bayer mosaic and averaging the two green channels. The demosaiced image is subsequently downsampled by a factor of 8 using OpenCV's `INTER_AREA` interpolation, resulting in a final input resolution of 512×352 pixels—matching the maximum resolution supported by MASt3R. Finally, the images are normalized to the range $[0, 1]$ in `float32` using a fixed scale factor of $(2^{14} - 1)^{-1}$ based on the sensor's bit depth.

Data loading. At each iteration, the dataloader provides a batch consisting of four images: two high-SNR (clean) and two low-SNR (noisy). A clean image is paired with a nearby-captured frame that observes an overlapping region of the scene to form a clean image pair $(\mathbf{I}^{(1)}_{\text{clean}}, \mathbf{I}^{(2)}_{\text{clean}})$. The corresponding noisy pair $(\mathbf{I}^{(1)}_{\text{noisy}}, \mathbf{I}^{(2)}_{\text{noisy}})$ is obtained either using simulation or by capturing images from the same viewpoints using a short exposure. This configuration uses roughly 48 GB of GPU memory.

S1.3. View Synthesis in the Dark

We build Dark3R-NeRF on Nerfacto [57]. Following RawNeRF [42], we supervise and render in linear raw space. To visualize the NeRF renders, we apply an ISP to recover sRGB images. The ISP performs black level subtraction, clipping, scaling, white balance, and gamma cor-

Table S1. **Dataset Statistics.** Summary of the 92 scenes from our handheld-captured dataset used for simulated low-light finetuning. We break down the number of images along with a description per scene.

Scene	Num. Images	Scene	Num. Images	Scene	Num. Images	Scene	Num. Images
Bookstore 1	249	Furniture Store 9	248	Kitchen 1	172	Old College 5	492
Bookstore 2	196	Furniture Store 10	336	Lab 1	191	Old College 6	212
Bookstore 3	231	Furniture Store 11	225	Lecture Hall 1	270	Old College 7	251
Catering 1	184	Furniture Store 12	173	Library 1	409	Parking Lot 1	216
Classroom 1	242	Furniture Store 13	154	Library 2	230	Parking Lot 2	245
Classroom 2	302	Furniture Store 14	284	Library 3	235	Parking Lot 3	309
Conference Room 1	156	Furniture Store 15	55	Library 4	159	Parking Lot 4	192
Department Store 1	207	Furniture Store 16	278	Library 5	151	Parking Lot 5	511
Department Store 2	166	Furniture Store 17	232	Library 6	201	Parking Lot 6	208
Department Store 3	185	Furniture Store 18	188	Library 7	292	Parking Lot 7	257
Department Store 4	162	Furniture Store 19	223	Library 8	191	Parking Lot 8	227
Department Store 5	307	Furniture Store 20	186	Library 9	258	Storage Room 1	270
Department Store 6	222	Furniture Store 21	196	Lobby 1	175	Storage Room 2	178
Department Store 7	201	Furniture Store 22	205	Mini Golf 1	212	Storage Room 3	175
Department Store 8	153	Furniture Store 23	242	Museum 1	192	Subway Station 1	318
Furniture Store 1	236	Furniture Store 24	137	Museum 2	166	Subway Station 2	105
Furniture Store 2	205	Grocery Store 1	340	Office 1	195	Utilities 1	242
Furniture Store 3	208	Grocery Store 2	195	Office 2	242	Workshop 1	271
Furniture Store 4	337	Grocery Store 3	417	Office 3	311	Workshop 2	295
Furniture Store 5	243	Grocery Store 4	206	Old College 1	191	Workshop 3	333
Furniture Store 6	263	Grocery Store 5	260	Old College 2	222	Workshop 4	337
Furniture Store 7	241	Grocery Store 6	155	Old College 3	229	Workshop 5	152
Furniture Store 8	206	Hardware Space 1	178	Old College 4	400	Workshop 6	383

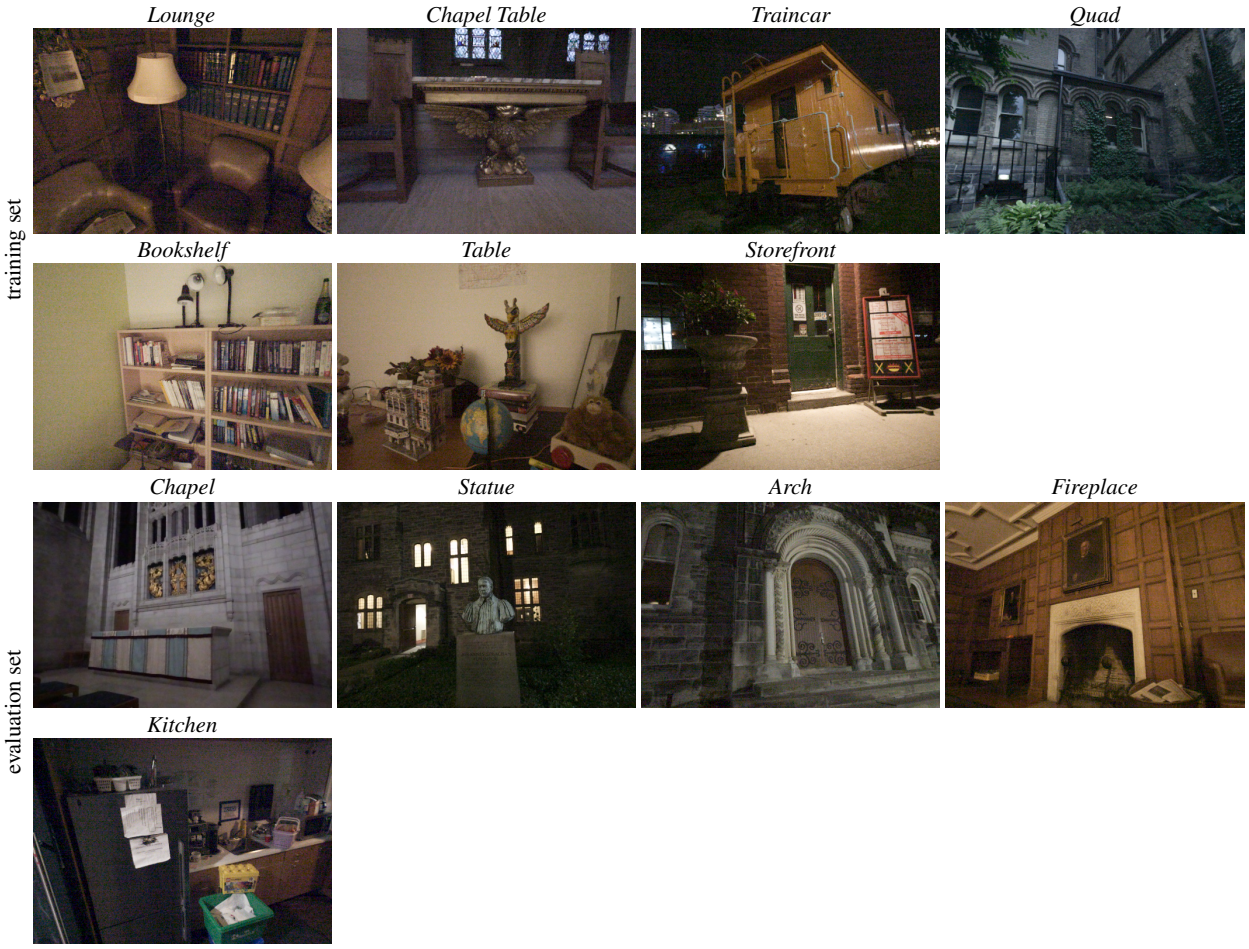

Figure S1. Scenes included in the tripod-captured dataset. We capture 12 different scenes and use seven for training and five for evaluation.

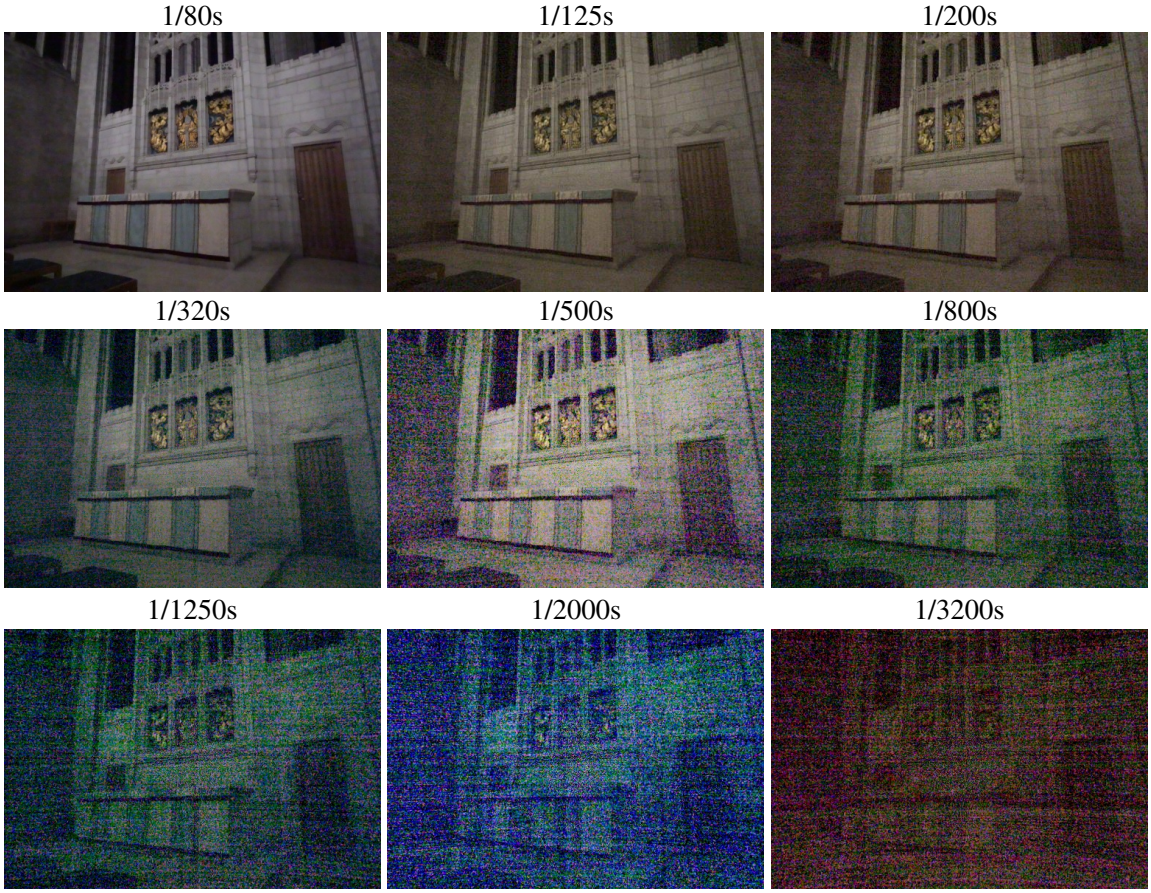

Figure S2. We visualize the different exposures included per scene in our dataset with shutter speeds indicated above each image. We capture nine different exposures for each scene across a wide range of SNRs. Images are contrast-stretched to facilitate visualization.

Table S2. **Dataset scenes.** A summary of the scenes used in our dataset, detailing the number of images, environment type, and train/test splits. Each scene contains nine different exposures with ground truth camera poses.

Scene	Num Images	Environment	Split
Lounge	430	Indoors	Train
Chapel Table	420	Indoors	Train
Traincar	400	Outdoors	Train
Quad	400	Outdoors	Train
Bookshelf	361	Indoors	Train
Table	332	Indoors	Train
Storefront	330	Outdoors	Train
Chapel	500	Indoors	Test
Statue	400	Outdoors	Test
Arch	370	Outdoors	Test
Fireplace	360	Indoors	Test
Kitchen	360	Indoors	Test

rection. The ground truth used to compute quantitative metrics is the clean raw images captured at the longest exposure for the held out viewpoint.

Table S3. **Camera pose estimation using a 2D denoiser.** We first process raw noisy images using the 2D denoiser LED [28] before using those images as input for MASt3R-SfM [16]. While using a 2D denoiser improves performance slightly compared to the baseline of providing raw input images to MASt3R-SfM, it performs much worse than Dark3R, which is trained to operate directly on noisy raw images. These results were computed on the chapel scene using 120 images.

Method	Input	ATE ↓	RPE T ↓	RPE R ↓
MASt3R-SfM [16]	raw	0.080	0.061	0.289
MASt3R-SfM [16]	denoised raw	0.048	0.048	0.238
Dark3R	raw	0.038	0.028	0.146

S1.4. Baseline Implementation Details

For COLMAP and VGGT, the sRGB images are processed through a standard in-camera image signal processor (ISP) including demosaicing via subsampling, black-level sub-

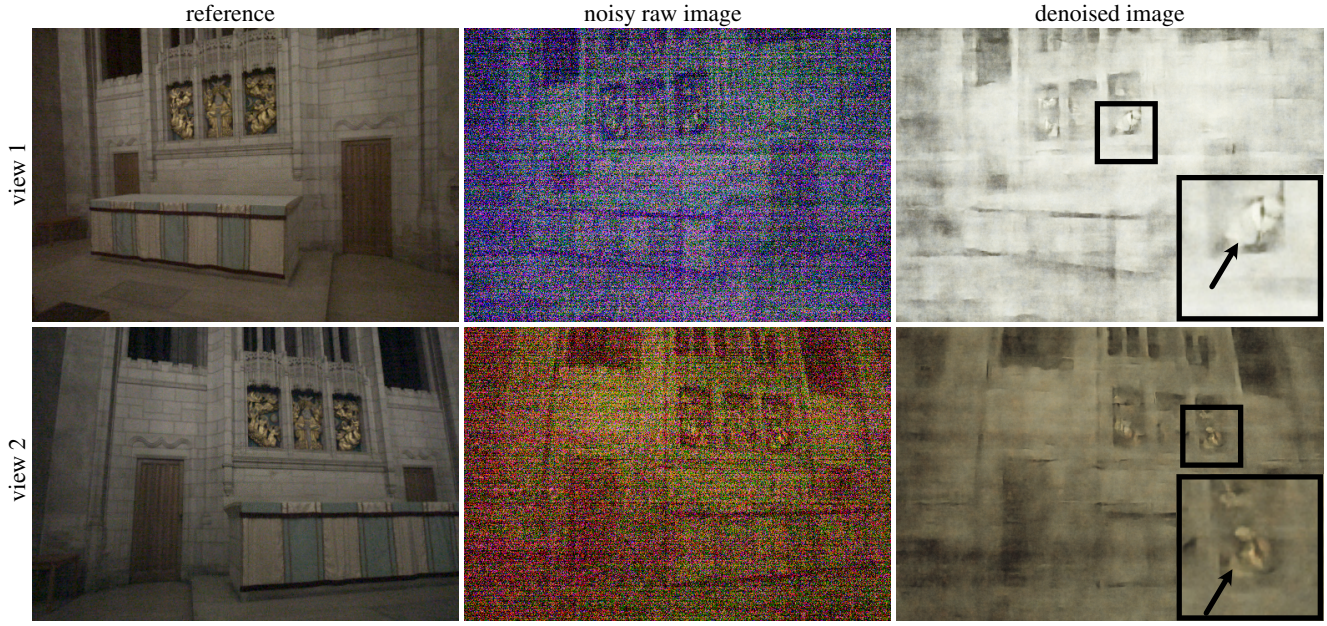

Figure S3. We illustrate the result of applying the 2D denoiser LED [28] to captured images in our dataset. Inspecting the output of the denoiser shows that the appearance of image features is not multi-view consistent (see insets, arrows), which complicates feature matching used by conventional structure-from-motion pipelines.

traction and clipping, white balancing, and gamma correction. The raw images are demosaiced by subsampling.

For radiance field reconstruction, we compare against two low-light extensions of common methods. We choose to test LE3D [29] as one of our baseline methods. LE3D is a state-of-the-art 3D Gaussian-Splatting [32] method designed for noisy raw input images. We run their codebase directly. Initialization is performed using the refined point cloud produced by Dark3R during its bundle-adjustment step.

The second method we test is RawNeRF [42], which we re-implemented within the Nerfacto framework [57] to ensure a consistent backbone with our method. It employs the exponential loss proposed in the original RawNeRF paper, along with black-level subtraction and clipping. Our method differs by introducing coarse-to-fine optimization, depth supervision, and operating directly on unnormalized raw data. All methods are trained for the same number of iterations (90,000) for fair comparison.

S2. Supplemental Results

2D denoising baseline. We show an example of applying a 2D denoiser (LED [28]) to a noisy captured scene in Fig. S3. Since the denoiser does not preserve image features in a multi-view consistent fashion, it is still challenging to perform pose estimation on the denoised images. Specifically, in Table S3 we find that while MASt3R-SfM performs slightly better in terms of pose accuracy on the denoised images vs. using raw images, Dark3R yields significantly better performance when given the raw images as input.

Additional pose results. In Figure S4 we show additional plots of pose and depth accuracy metrics versus image SNR, broken out by scene, on all images from the captured evaluation dataset. We find that Dark3R outperforms MASt3R-SfM in most cases, especially as the SNR decreases. We see similar trends in the corresponding results shown in Table S4 across the test scenes. These results are calculated for a single exposure-bracket setting in each scene, and the average image SNR per scene ranges from -4.76 dB to -2.99 dB.

Additional view synthesis results. We show the results on view synthesis for each scene in the evaluation dataset in Table S5, and we use the same single exposure-bracket setting for each scene as in Table S4. We see the same trends as in the main paper—on average using Dark3R for pose estimation and Dark3R-NeRF for view synthesis outperforms the baselines. We note that in some cases the performance of an oracle (where poses are obtained from the clean images) performs worse than other approaches; however, we attribute this to the fact that the oracle is still trained on noisy images and so there are still artifacts in the reconstruction that can degrade image quality.

Averaging and motion blur in the captured dataset. To illustrate that our dataset has significant disparity between views, we show the result of naive averaging of sequentially

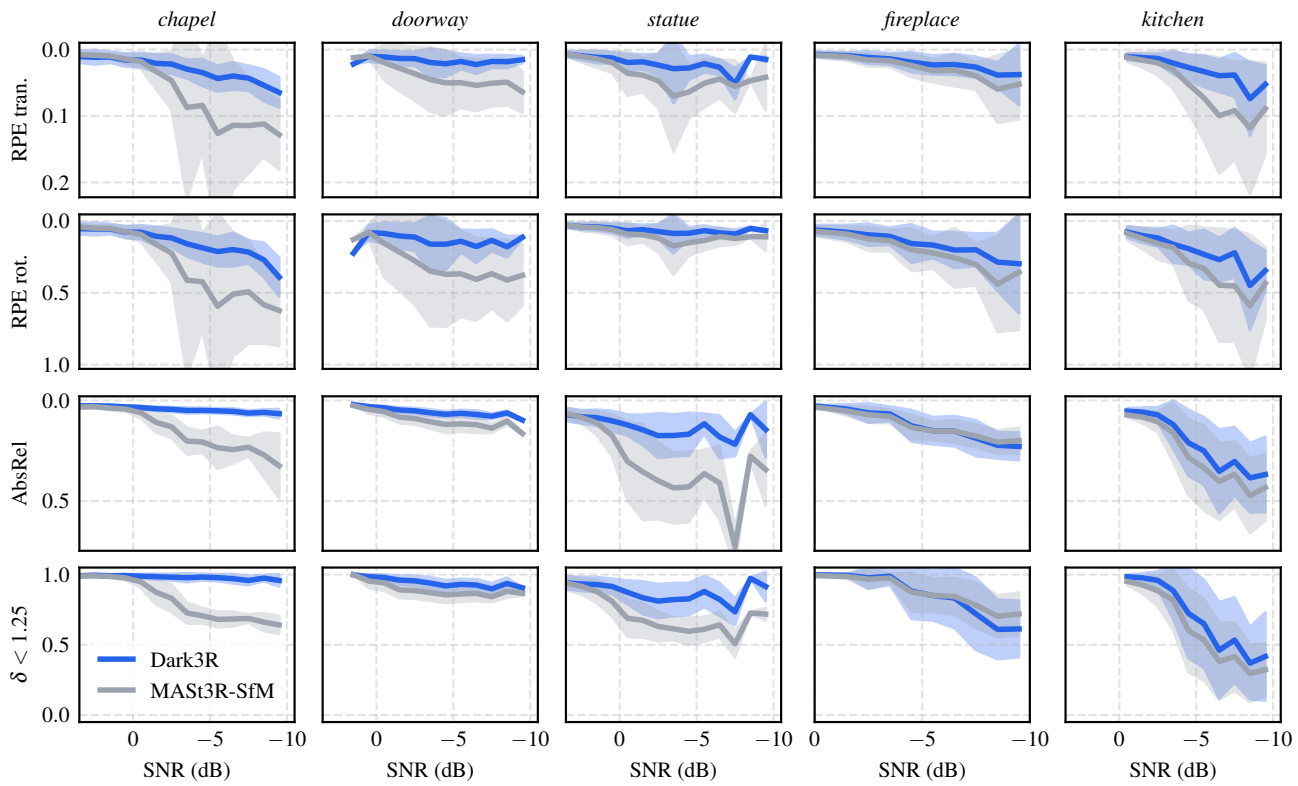

Figure S4. Pose prediction assessment for all evaluated scenes. We compare Dark3R (blue) against MASt3R-SfM [16] (gray) across five pose and depth metrics: relative pose error in translation (RPE tran.), relative pose error in rotation (RPE rot.), absolute relative depth error (AbsRel), and the accuracy threshold $\delta<1.25$. Each curve shows mean performance with shaded regions indicating standard deviation across scenes. As image SNR decreases, Dark3R generally maintains lower pose and depth errors and higher reconstruction accuracy compared to MASt3R-SfM.

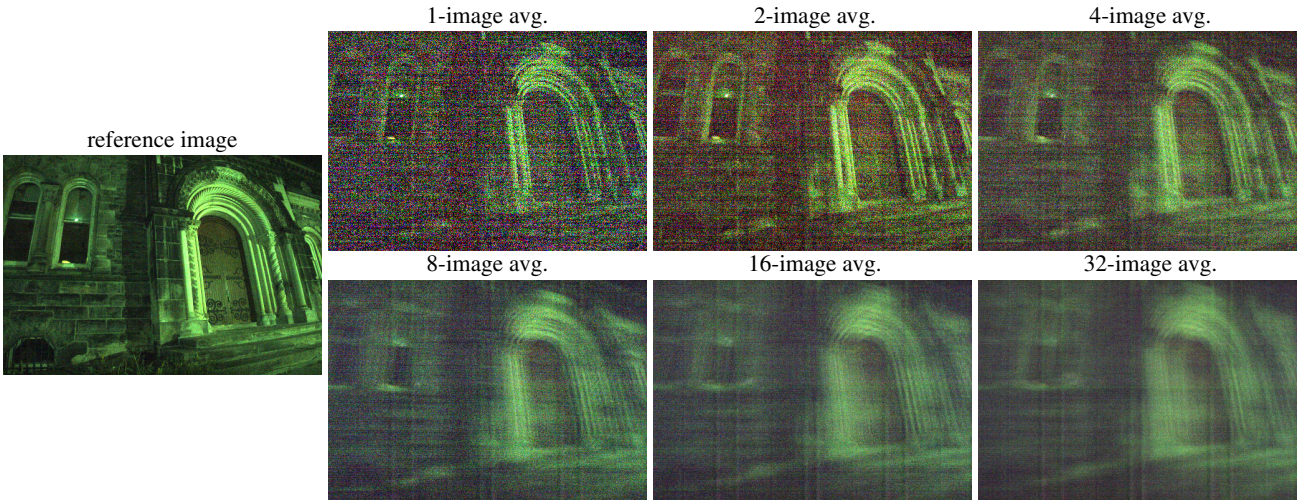

Figure S5. Visualization of naive averaging on a captured scene. We average an increasing number of sequentially captured images together to perform denoising, but this comes at the cost of blurring the image. Hence, the dataset contains a non-trivial amount of disparity between the captured images, and the amount of noise makes it challenging to apply conventional burst denoising techniques that require image alignment.

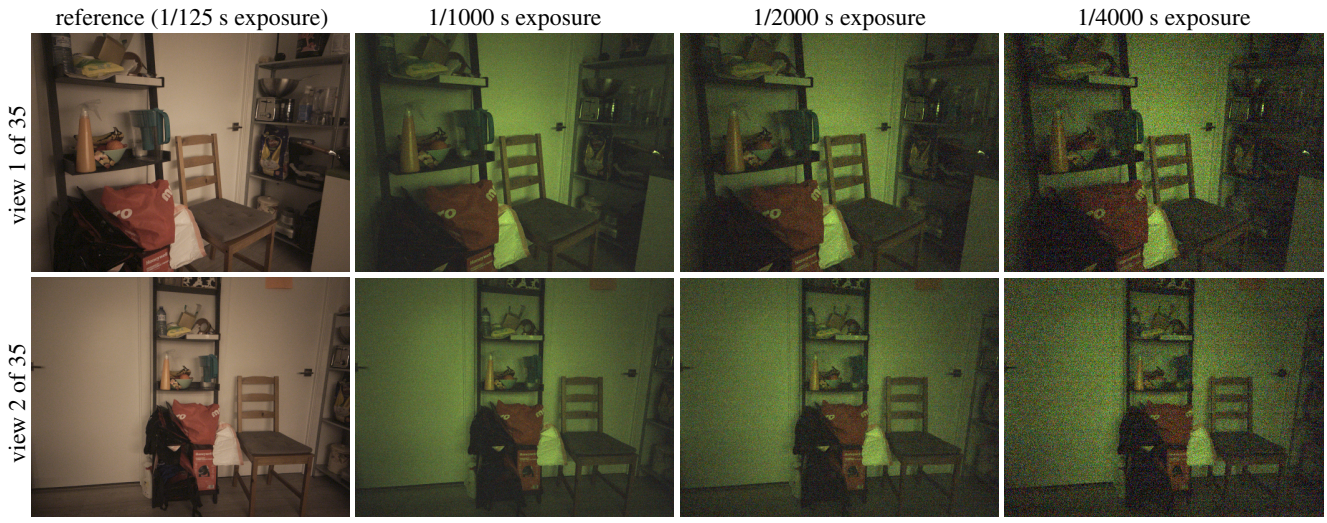

Figure S6. iPhone dataset. We test the generalization capabilities of our model by testing it on a multi-view raw image dataset captured by an iPhone 16. We capture this scene at four different exposure settings (including a reference, well-exposed capture), resulting in different signal-to-noise ratios. As shown in Table S6, the Dark3R outperforms the baseline even when applied to this dataset.

captured frames for a particular scene (Fig. S5). We observe that while the noise is reduced by averaging images together, this process also introduces significant blur.

Evaluation on the iPhone dataset. We test the generalization capabilities of Dark3R by capturing a dataset of 315 exposure-bracketed multi-view raw images for a single scene using an iPhone 16 (i.e., 9 different exposure settings with 35 images each). We compare our approach to MASt3R-SfM [16] in Table S6 and find that our approach outperforms this baseline, especially at low signal-to-noise ratios corresponding to image captures with a short exposure time. We show example images from the iPhone dataset in Fig. S6, and we will publicly release this dataset.

Evaluation on clean images. We provide an additional comparison in Table S7 showing that Dark3R has comparable pose estimation performance to MASt3R-SfM [16] on clean images, and hence generalizes to a broad SNR range.

Generalization across intrinsics. Table S8 shows that performance of running Dark3R without camera intrinsics is comparable to performance when providing intrinsics as input.

Table S4. **Camera pose estimation results by scene.** We report results for 120 sequential images (top rows) or the full set (bottom rows) from each scene, grouped by scene name. Metrics include pose estimation errors after Sim(3) alignment and 3D reconstruction quality.

	Method/Input		Camera pose error			3D reconstruction	
Scene	**Method**	**Input**	**ATE** ↓	**RPE T** ↓	**RPE R** ↓	**AbsRel** ↓	$\delta < 1.25$ ↑
Chapel	COLMAP [51]	sRGB	0.187	0.250	1.229	0.187	84.21
	MASt3R [33]	raw	0.832	0.614	2.877	0.234	60.56
	VGGT [61]	sRGB	0.405	0.451	2.197	0.212	59.24
	MASt3R-SfM [16]	raw	0.080	0.061	0.289	0.210	70.74
	Dark3R	raw	0.038	0.027	0.144	0.050	98.22
	MASt3R-SfM (Full)	raw	0.175	0.054	0.255	0.226	57.73
	Dark3R (Full)	raw	0.206	0.025	0.135	0.053	93.99
Statue	COLMAP [51]	sRGB	0.377	0.182	0.482	1.179	54.33
	MASt3R [33]	raw	0.790	0.483	2.526	0.361	26.09
	VGGT [61]	sRGB	0.325	0.216	0.622	0.266	51.80
	MASt3R-SfM [16]	raw	0.081	0.044	0.111	0.381	63.34
	Dark3R	raw	0.054	0.020	0.066	0.166	79.10
	MASt3R-SfM (Full)	raw	0.207	0.046	0.107	0.406	50.72
	Dark3R (Full)	raw	0.144	0.021	0.070	0.179	82.17
Arch	COLMAP [51]	sRGB	0.043	0.062	0.453	0.087	94.71
	MASt3R [33]	raw	0.239	0.268	1.619	0.257	64.65
	VGGT [61]	sRGB	0.080	0.096	0.488	0.101	87.39
	MASt3R-SfM [16]	raw	0.024	0.025	0.180	0.088	88.35
	Dark3R	raw	0.013	0.013	0.093	0.044	96.50
	MASt3R-SfM (Full)	raw	0.068	0.045	0.352	0.100	87.93
	Dark3R (Full)	raw	0.048	0.017	0.142	0.057	94.74
Fireplace	COLMAP [51]	sRGB	1.443	0.090	3.125	0.418	24.45
	MASt3R [33]	raw	0.430	0.268	1.830	0.245	55.98
	VGGT [61]	sRGB	0.176	0.112	0.881	0.083	97.57
	MASt3R-SfM [16]	raw	0.143	0.023	0.187	0.118	94.58
	Dark3R	raw	0.106	0.018	0.146	0.090	98.64
	MASt3R-SfM (Full)	raw	0.225	0.020	0.179	0.067	96.32
	Dark3R (Full)	raw	0.138	0.014	0.127	0.051	97.91
Kitchen	COLMAP [51]	sRGB	1.297	0.191	2.933	1.320	14.20
	MASt3R [33]	raw	1.289	0.521	4.186	0.421	7.35
	VGGT [61]	sRGB	0.272	0.203	1.048	0.500	20.39
	MASt3R-SfM [16]	raw	0.115	0.038	0.236	0.181	79.95
	Dark3R	raw	0.041	0.019	0.156	0.108	93.24
	MASt3R-SfM (Full)	raw	0.355	0.030	0.194	0.169	82.06
	Dark3R (Full)	raw	0.158	0.018	0.132	0.124	92.85

Table S5. **View synthesis results by scene.** We assess photometric quality on low-SNR images. We compare to an "oracle" obtained by performing pose estimation on clean images.

Scene	Method	Camera poses	PSNR ↑	SSIM ↑	LPIPS ↓
Chapel	LE3D [29]	Dark3R	38.01	0.868	0.398
	RawNeRF [42]	Dark3R	35.59	0.871	0.347
	Dark3R-NeRF	MASt3R-SfM [16]	36.48	0.826	0.356
	Dark3R-NeRF	Dark3R	37.44	0.840	0.310
	Dark3R-NeRF	Oracle	37.87	0.873	0.270
Statue	LE3D [29]	Dark3R	25.20	0.839	0.283
	RawNeRF [42]	Dark3R	24.24	0.829	0.255
	Dark3R-NeRF	MASt3R-SfM [16]	24.47	0.814	0.277
	Dark3R-NeRF	Dark3R	25.68	0.851	0.240
	Dark3R-NeRF	Oracle	27.97	0.877	0.201
Arch	LE3D [29]	Dark3R	38.50	0.889	0.292
	RawNeRF [42]	Dark3R	37.00	0.845	0.275
	Dark3R-NeRF	MASt3R-SfM [16]	35.90	0.795	0.344
	Dark3R-NeRF	Dark3R	38.13	0.858	0.237
	Dark3R-NeRF	Oracle	38.55	0.862	0.213
Fireplace	LE3D [29]	Dark3R	39.35	0.898	0.408
	RawNeRF [42]	Dark3R	37.57	0.834	0.306
	Dark3R-NeRF	MASt3R-SfM [16]	38.65	0.869	0.302
	Dark3R-NeRF	Dark3R	41.09	0.896	0.249
	Dark3R-NeRF	Oracle	40.24	0.898	0.246
Kitchen	LE3D [29]	Dark3R	37.81	0.897	0.312
	RawNeRF [42]	Dark3R	36.81	0.860	0.272
	Dark3R-NeRF	MASt3R-SfM [16]	37.52	0.872	0.261
	Dark3R-NeRF	Dark3R	38.54	0.884	0.249
	Dark3R-NeRF	Oracle	41.17	0.900	0.210

Table S6. **Generalization to other cameras.** We assess the ability of Dark3R to generalize to other types of cameras using images captured on an iPhone 16. We captured a series of low-light multi-view images on an iPhone 16 and tested Dark3R against MASt3R-SfM [16]. Despite not being finetuned with iPhone images, Dark3R still outperforms MASt3R-SfM for camera pose estimation accuracy.

Method	Input	Exposure (s)	ATE ↓	RPE T ↓	RPE R ↓
MASt3R-SfM [16]	raw	1/1000	0.105	0.062	0.238
Dark3R	raw	1/1000	0.043	0.045	0.172
MASt3R-SfM [16]	raw	1/2000	0.124	0.127	0.380
Dark3R	raw	1/2000	0.089	0.068	0.244
MASt3R-SfM [16]	raw	1/4000	2.484	1.174	6.278
Dark3R	raw	1/4000	0.227	0.178	0.575

Table S7. **Clean image performance.** We compare the camera pose error of Dark3R to MASt3R-SfM [16]. We evaluate on all 330–500 sRGB images from each scene at the longest exposure in our captured dataset. We omit the depth metrics shown in Table 1 because they used the M-SfM output on clean images as the reference.

Method	ATE ↓	RPE T ↓	RPE R ↓
M-SfM [16]	0.088	0.006	0.031
Dark3R	0.104	0.006	0.030

Table S8. **Known vs. unknown intrinsics.** Average camera pose error and 3D reconstruction metrics with unknown and known intrinsics evaluated on the captured dataset with 120 images per scene.

		Camera pose error			3D reconstruction	
Method	Intrinsics	ATE ↓	RPE T ↓	RPE R ↓	AbsRel ↓	$\delta < 1.25$ ↑
Dark3R	unknown	0.043	0.019	0.121	0.090	93.63
Dark3R	known	0.050	0.020	0.121	0.091	93.14