Inter-Photon-Limited Videography: Supplemental Material

Andrew Xie[1, 2] Dongyu Du[1, 2] Sotiris Nousias[3] David B. Lindell[1, 2]
Kiriakos N. Kutulakos[1, 2]

[1]University of Toronto [2]Vector Institute [3]Purdue University

Contents

A Related Work

A.1 Image Restoration

Classical single-image restoration methods rely on well-established image priors such as smoothness or patch self-similarity [2, 3]. More recently, deep neural networks have been leveraged for their expressive capacity to incorporate and learn inductive biases from data. Self-supervised approaches eliminate the requirement for clean targets by exploiting pixel independence [1] or assuming structure by masking pixels to predict signal from surrounding context [14]. Ulyanov *et al*. further demonstrated a strong image prior can arise using solely the structure of a generator network [29]. While these methods provide valuable insights into effective image denoising, they fail in extremely low-SNR regimes such as those encountered in the inter-photon-limited regime. Poisson-specific denoising methods [12, 23], attempt to address this regime by enforcing stronger self-consistency assumptions, but typically achieve only limited success, recovering simple structures at best, because 1-bit images contain little recoverable information. Moreover, directly applying any image-based denoising methods to video often introduces temporal artifacts such as flickering [16], since they do not account for spatiotemporal correspondences across frames.

A.2 Video Restoration

Self-supervised video denoising extends image-based ideas by exploiting temporal redundancy. Frame-to-frame fine-tuning uses motion-compensated neighboring frames as pseudo-targets, enabling learning when, temporally, noise is independent while signal is structured [7, 11]. Multi-frame variants add flow and occlusion handling to prevent target leakage [9] and other fully unsupervised objectives based on video priors [25] or known per-frame noise models [33]. Deep Video Prior [16] dispenses with explicit flow, using the structure of a generative CNN to capture spatiotemporal self-similarity while maintaining temporal consistency. However, these methods break down for single-photon datasets because they are extremely photon-sparse, so direct application produces severe temporal artifacts. Beyond estimating scene flow, these methods also treat time implicitly, making it challenging to capture long-range temporal structure beyond the model's perceptive window. This is particularly problematic for single-photon datasets, where individual frames are information-poor, but considerable redundancy exists across an acquisition interval. In addition, these methods are computationally expensive when applied to single-photon data and do not address the bandwidth and storage challenges posed. This gap motivates our approach, which massively aggregates large-scale temporal redundancy with a time and frequency-informed model and designs optimization objectives consistent with the Poissonian statistics of photon arrivals.

A.3 Quanta Image and Video Restoration

Quanta restoration techniques aim to fill the shortcomings of the general image and video denoisers by designing methods for 1-bit capture. Liu *et al*. [18] split binary volumes of a few binary frames according to the splittable nature of Poisson process statistics into noisy input-target pairs for self-supervised training. This method works because a shared 3D CNN learns to complete the noisy inputs, capturing self-similarities across space and time within the binary cubes. Similarly, Quanta Burst Photography [19] adapts burst photography techniques to align binary frames to compensate for inter-frame motion, thereby using spatiotemporal patch recurrence in natural videos to denoise a central reference frame. These methods fail when global signals extend beyond the temporal receptive field of the CNN, such as strong periodic components that require longer-range modeling. Wei *et al*. [31] instead employ an implicit periodicity prior by estimating Fourier coefficients directly from probing measurements. This enables flux recovery even more than a period per photon per pixel. However, because processing is performed pixel-by-pixel, spatial features are neglected, and thus, this approach cannot handle settings where not enough photons arrive at each pixel to estimate flux reliably.

A.4 Neural Video Representations

Our approach builds on neural video representations, which parameterize videos with implicit neural networks mapping spatiotemporal coordinates to pixel intensities [5, 6, 15, 17, 20, 32]. Prior work has primarily explored these representations for compact video compression, with denoising effects appearing only incidentally through the networks' low-rank bias. We show that these representations can be adapted to the inter-photon-limited regime—well beyond the conditions considered in earlier work. To make this effective for photon arrival sequences, we introduce frequency-based inductive biases in

the embedding layers and replace conventional reconstruction losses with likelihood objectives grounded in photon arrival statistics. These modifications enable our method to operate in regimes where photon interarrival rates fall well below the timescale of temporal flux variations.

B Inter-Photon Spectrum Estimation

In this section, we describe how we estimate the inter-photon spectra for the datasets analysed in the main paper.

Given the flux spectrum $\Phi(\mathbf{x}, f)$ at a pixel $\mathbf{x}$ and its mean inter-photon interval $\tau(\mathbf{x})$, we estimate the inter-photon spectrum $\Psi(\mathbf{x}, f_p)$ using the following equation:

$$\Psi(\mathbf{x}, f\tau(\mathbf{x})) \;=\; \Phi(\mathbf{x}, f)(1/\|\Phi(\mathbf{x}, 0)\|)(t_{\text{acq}}/\tau(\mathbf{x})), \tag{B.1}$$

where $f_p = f\tau(\mathbf{x})$.

Admittely, for real datasert, the $\Phi(\mathbf{x}, f)$ is not known and can only be estimated after the video sequence has been produced through various processing steps. For example, our analysis of the inter-photon spectrum does not account for the read noise of CMOS sensors. Because read noise is high-frequency [28] while natural videos have predominantly low-frequency content [10], we expect that read noise will shift the inter-photon spectra toward higher inter-photon frequencies.

B.1 Inter-Photon Spectra for Datasets in Figure 2

We describe how we estimate f_p for each different dataset in Fig. 2 of the main paper. For each dataset, we follow three steps: (1) estimate the flux spectrum and (2) the number of photon arrivals at each pixel, and (3) compute the inter-photon spectrum using Equation (B.1). We provide details on these steps for each dataset below.

Dancing Under the Stars [22]. We estimate the inter-photon spectrum per pixel as follows:

1. We estimate the photon count per pixel per frame as:

$$\hat{p}(\mathbf{x}, t) = \frac{I(\mathbf{x}, t)}{G} W, \tag{B.2}$$

   where $I(\mathbf{x}, t)$ is the intensity of pixel $\mathbf{x}$ at time t and W=70000, G=65536 are the full-well capacity, and number of gray levels respectively obtained from the datasheet for the Canon LI3030SAI Sensor [4]. We use the full well capacity of the green channel.

2. We compute the flux at each pixel as the ratio between the photon counts and the frame's exposure, *i.e.*, $\phi(\mathbf{x}, t) = \frac{\hat{p}(\mathbf{x}, t)}{t_{\text{acq}}}$. We compute the mean-inter photon interval as the ratio between total exposure time and the detected photons at a pixel, $\tau(\mathbf{x}) = \frac{t_{\text{acq}}}{N(\mathbf{x})}$ where $N(\mathbf{x})$ is the total number of photons detected at pixel $\mathbf{x}$ over t_{acq}.

3. We take the Discrete Fourier Transform (DFT) $\phi(\mathbf{x}, t)$ to get the spectrum $\Phi(\mathbf{x}, f)$.

4. We compute $\Psi(\mathbf{x}, f_p)$ using Eq. (B.1) and $\Phi(\mathbf{x}, f)$, $\tau(\mathbf{x})$ estimated from Steps 3, and 2 respectively. The inter-photon frequency is $f_p = f\tau(\mathbf{x})$.

High-speed Phantom Camera video dataset [24]. We follow the same procedure as for the Dancing Under the Stars dataset, substituting full-well capacity, frame rate, and exposure time with values from the Phantom camera datasheet [30].

Flying with Photons [21]. The input data is a histogram of photon counts at each pixel $h(\mathbf{x}, t)$, where time t is quantized in 4000 bins of $b = 4$ ps width each. Let N_h denote the total number of photons detected in the histogram. We estimate the inter-photon spectrum per pixel as follows:

1. We compute the flux at each pixel as the ratio of photon counts to bin width, *i.e.*, $\phi(\mathbf{x}, t) = \frac{h(\mathbf{x}, t)}{b}$.

2. We take the DFT of $\phi(\mathbf{x}, t)$ to obtain $\Phi(\mathbf{x}, f)$. We calculate the mean inter-photon interval as the histogram period divided by the total number of detected photons, $\tau = \frac{4000b}{N_h}$.

3. We compute the inter-photon spectrum as:

$$\Psi(\mathbf{x}, f\,\tau(\mathbf{x})) = 1/f_{\text{sync}}\Phi(\mathbf{x}, f_{\text{sync}} f_p), \tag{B.3}$$

where $f_{sync} = 10$ MHz is the repetition rate of the laser [21]. We apply this scaling to account for the stroboscopic imaging effect induced by the synchronization operation described in Figure 3 of the main paper.

Quanta video sequences. Since we need flux to compute the inter-photon spectra but our quanta sensors output binary frames, we estimate flux using different reconstruction methods. We estimate the ground truth flux for each quanta dataset using its own corresponding reconstruction method. For QBP datasets [19] and our own SPAD512 captures, we apply our method since QBP is computationally infeasible to scale to 100k frames, while for bit2bit datasets [18] we use their reconstruction output. We treat the flux output as ground truth flux $\phi(\mathbf{x},t)$ for computing the inter-photon spectrum at each pixel. We then do the following:

1. We take the DFT of $\phi(\mathbf{x},t)$ to obtain the flux spectrum $\Phi(\mathbf{x},f)$. We set $\tau(\mathbf{x})$ to $\frac{t_{\text{acq}}}{N(\mathbf{x})}$, where $N(\mathbf{x})$ is the total number of photons detected at pixel $\mathbf{x}$.
2. We compute $\Psi(\mathbf{x},f_p)$ using Equation (B.1) and $\Phi(\mathbf{x},f)$, $\tau(\mathbf{x})$ estimated from Step 1. The inter-photon frequency is $f_p = f\tau(\mathbf{x})$.

UWB—single-pixel SPAD timestamp streams [31]. We use the photon timestamps at each pixel and compute the inter-photon frequency spectrum as follows:

1. We perform frequency probing [31] to estimate the flux spectrum $\Phi(\mathbf{x},f)$ at every pixel. For computational efficiency, we probe only the harmonics of the fan and laser, similar to [31]. We set the mean inter-photon interval $\tau(\mathbf{x}) = \frac{t_{\text{acq}}}{N(\mathbf{x})}$, where $N(\mathbf{x})$ is the total number of photons detected at pixel $\mathbf{x}$.
2. We compute $\Psi(\mathbf{x},f_p)$ using Equation (B.1) and $\Phi(\mathbf{x},f)$, $\tau(\mathbf{x})$ estimated from Step 1. The inter-photon frequency is $f_p = f\tau(\mathbf{x})$.

B.2 Estimating the f_p Summary Statistic

We compute the summary statistic f_p by aggregating inter-photon spectral information across all camera pixels. For each pixel, we define

$$\text{summary statistic } \log(f_p(\mathbf{x})) = \frac{\displaystyle\int_{\varepsilon}^{\infty} \log(f_p)\,\log\left|\Psi(\mathbf{x},f_p)\right| df_p}{\displaystyle\int_{\varepsilon}^{\infty} \log\left|\Psi(\mathbf{x},f_p)\right| df_p}. \tag{B.4}$$

This quantity corresponds to the log-log centroid of the inter-photon spectrum. The lower limit ε is intended to exclude the DC component. In practice, we approximate both integrals using Riemann sums (left-hand rule) evaluated on a uniformly sampled interval between ε and the maximum frequency of incident flux. To obtain a single video-level summary, we report the 75th percentile of per-pixel f_p values across the entire frame.

Layer	Modules	Upscale Factor	Output Size $(C \times H \times W)$
0	Positional Encoding	–	$16 \times 1 \times 1$
1	MLP & Reshape	–	$128 \times 16 \times 16$
2	Conv-Upsample block	$2\times$	$128 \times 32 \times 32$
3	Conv-Upsample block	$2\times$	$64 \times 64 \times 64$
4	Conv-Upsample block	$2\times$	$32 \times 128 \times 128$
5	Conv-Upsample block	$2\times$	$16 \times 256 \times 256$
6	Conv-Upsample block	$2\times$	$8 \times 512 \times 512$
7	Head layer	–	$1 \times 512 \times 512$

Table 1: Model architecture for 512×512 videos. The model takes a 1D timestamp, applies positional encoding and an MLP, then upsamples through 5 convolutional upsampling blocks.

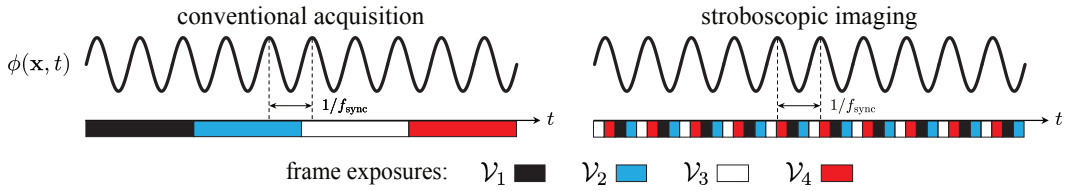

Figure C.1: Stroboscopic imaging is used for periodic signals when the exposure required to achieve sufficient SNR exceeds the signal's period. Rather than using active illumination, we achieve this by integrating non-contiguously across repeated cycles, accumulating photons over intervals consistently relative to the fundamental frequency of the signal (f_{sync}), which we synchronize to.

C Implementation Details

C.1 Network Architecture Details

We provide the architecture details of our Neural Flux Field which has a decoder CNN architecture in Table 1. On a 512×512 video, given the timestamp index t, we first apply a 1D positional encoding with 32 log-spaced frequencies to obtain a temporal embedding. This embedding is passed through a 4-layer MLP (hidden dimensions 128, 128, and 256) and reshaped into a $128 \times 16 \times 16$ feature grid. We then stack 5 convolutional blocks with 3 convolutional layers per block and pixelshuffle upsampling factors $(2,2,2,2,2)$. The corresponding channel widths are $(128,64,32,16,8)$. The final head layer maps the last feature map to a single output frame with one channel at full spatial resolution and applies a softplus activation.

C.2 Computational Stroboscopy

We describe how to implement periodic integration domains for stroboscopic imaging. We use computational stroboscopy for the scene with both a spinning fan and illumination from a pulsed laser (See Figure 6) to address periodic flux from two vastly different timescales.

Multiple fundamental frequencies. Consider the case where a flux function contains multiple, non-harmonically related frequencies $\{f^{(1)}, f^{(2)}, \dots\}$, each with period $T^{(i)}$. As in the single frequency case, we can identify multiple frequencies using harmonic probing on the raw photon timestamps [31]. Then, we define a set of periodic exposures for each frequency:

$$\mathcal{V}_{k_i}^{(i)} = \bigcup_{m=1}^{M_i} \left[t_{\mathcal{V}_{k_i}^{(i)}} + mT^{(i)},\, t_{\mathcal{V}_{k_i}^{(i)}} + mT^{(i)} + \Delta t^{(i)} \right], \tag{C.5}$$

where the start time $t_{\mathcal{V}_{k_i}^{(i)}}$ and duration $\Delta t^{(i)}$ may vary across frequencies. To selectively integrate flux corresponding to specific timing offsets across all frequencies, we define a combined exposure $\mathcal{V}_k$, with index $k = (k_1, k_2, \ldots)$, as the intersection of the per-frequency exposures:

$$\mathcal{V}_k = \bigcap_i \mathcal{V}_{k_i}^{(i)}. \tag{C.6}$$

Thus, the exposure $\mathcal{V}_k$ comprises the collection of time intervals that are simultaneously active for all frequencies.

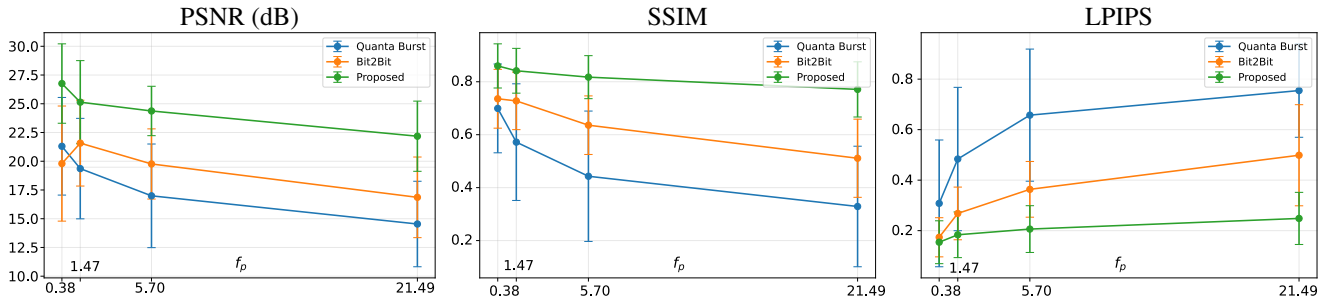

Figure D.2: f_p is an intrinsic measure of the difficulty of reconstructing a signal. It considers both the photon inter-arrival times and how rapidly flux varies over time. We reconstruct nine simulated quanta videos derived from high-speed footage and plot reconstruction metrics for different methods as a function of f_p. We find that across scenes, lower f_p (i.e., faster photon arrival rates, slower scene motion) correlates with improved reconstruction quality for all methods. We also demonstrate that the proposed method outperforms competing approaches across all evaluated perceptual metrics.

D Quantitative Reconstruction Error

D.1 Simulating Photon Arrivals from High-Speed Videos

We simulate photon arrivals from high-speed videos captured by conventional cameras using a method similar to [27]:

1. We collect high-speed videos from a Phantom camera [24], originally recorded at 1 kfps. We download the video at 1920×1080, downsample the spatial resolution to 640×360, and apply a center crop to obtain 512×256 pixels per frame.
2. We map the gamma-corrected videos to linear intensity by applying the inverse gamma transformation. We use an exponent of 2.2.
3. We interpolate videos by 16 times using RIFE [13], which produces a video with a nominal frame rate of 16 kfps.
4. We linearly interpolate 7 frames between each pair of consecutive frames at 16 kfps, to produce a final nominal frame rate of 128 kfps (8× temporal upsampling).
5. We generate photon timestamps at each pixel by thinning an inhomogeneous Poisson process whose rate is given by the intensity estimated from the previous step.
6. Finally, we quantize the photon timestamps to 10 μs resolution. If multiple photons have the same quantized timestamp, we keep only the first arrival.

D.2 Simulated Evaluation

In Table 2 we evaluate our video restoration approach across nine diverse simulated scenes, each simulated for four different photon-arrival regimes to capture a broad range of flux conditions. We compare against QBP [19] and bit2bit [18], the current state-of-the-art in photon-limited video reconstruction. While QBP is computationally prohibive [1] for practical frame-by-frame video synthesis, we include it because of its consistently strong reconstruction quality and widespread use as a reference method, even in comparisons against supervised techniques [8].

We also show plots of reconstruction quality versus inter-photon frequency in Figure D.2, and we find that the inter-photon frequency is inversely correlated with image quality. As expected, as the photon arrival rate decreases or the flux variations become more rapid, then the difficulty of the reconstruction problem increases and the photometric quality is reduced.

[1] Using the implementation of [19], processing the raw $130k \times 512 \times 512$ binary data and rendering 300 output frames requires 2–3 days on a CPU.

Table 2: Quantitative comparisons across nine simulated scenes at four thinning levels. Higher PSNR, SSIM, and lower LPIPS indicate better image reconstruction quality. The average interphoton frequency f_p is calculated across the nine scenes. The best result among the three methods for each scene and thinning factor (fp setting) is indicated in red. Note that as f_p increases, the reconstruction problem becomes increasingly more challenging. Our method outperforms all others for almost all datasets and inter-photon frequencies above one period per photon.

			Scene								
Metric	Method	fp	falling	pillow fight	wok	bee	surfacing	wine	hair-flip	cereal	coffee
PSNR ↑	Quanta Burst	0.38	21.87	21.27	13.98	18.38	16.99	27.23	22.29	23.89	25.80
		1.47	20.74	20.71	14.02	15.34	15.35	27.41	16.89	20.39	23.40
		5.70	17.80	16.94	14.10	12.86	12.61	24.27	13.64	16.45	24.26
		21.49	14.08	13.74	13.63	11.67	10.75	20.31	11.69	13.78	21.20
	bit2bit	0.38	18.08	17.12	13.91	18.86	17.52	21.72	20.19	18.93	31.84
		1.47	20.17	19.77	13.89	20.96	21.38	27.43	24.32	22.76	23.49
		5.70	18.39	17.87	13.79	21.84	20.06	23.94	18.44	20.90	22.66
		21.49	14.95	15.26	14.11	10.67	18.58	22.12	20.29	18.81	16.96
	Proposed	0.38	32.50	22.51	27.88	27.31	23.65	29.21	26.38	29.21	22.17
		1.47	25.85	21.60	20.23	23.12	21.97	27.82	27.30	31.27	27.10
		5.70	23.71	21.93	26.15	21.15	24.04	27.49	26.70	23.24	24.94
		21.49	17.67	20.05	25.00	22.26	18.73	23.91	24.88	20.72	26.36
LPIPS ↓	Quanta Burst	0.38	0.210	0.076	0.120	0.774	0.437	0.075	0.480	0.510	0.092
		1.47	0.351	0.168	0.275	0.987	0.641	0.214	0.703	0.710	0.302
		5.70	0.510	0.391	0.493	1.057	0.827	0.344	0.881	0.914	0.498
		21.49	0.664	0.593	0.618	0.986	0.887	0.510	0.897	0.995	0.643
	bit2bit	0.38	0.172	0.136	0.106	0.132	0.316	0.122	0.271	0.215	0.094
		1.47	0.223	0.264	0.208	0.261	0.441	0.133	0.416	0.299	0.167
		5.70	0.293	0.430	0.273	0.392	0.511	0.210	0.517	0.382	0.263
		21.49	0.335	0.596	0.323	0.928	0.589	0.295	0.580	0.447	0.396
	Proposed	0.38	0.129	0.187	0.113	0.103	0.351	0.054	0.192	0.127	0.131
		1.47	0.148	0.217	0.138	0.142	0.386	0.069	0.242	0.154	0.156
		5.70	0.154	0.307	0.146	0.169	0.377	0.075	0.253	0.213	0.162
		21.49	0.191	0.348	0.168	0.235	0.417	0.085	0.309	0.298	0.185
SSIM ↑	Quanta Burst	0.38	0.814	0.815	0.549	0.472	0.523	0.948	0.605	0.718	0.849
		1.47	0.710	0.748	0.548	0.236	0.344	0.903	0.410	0.498	0.750
		5.70	0.606	0.522	0.540	0.104	0.181	0.836	0.256	0.285	0.656
		21.49	0.534	0.304	0.506	0.051	0.101	0.716	0.181	0.150	0.416
	bit2bit	0.38	0.813	0.636	0.551	0.764	0.642	0.787	0.752	0.751	0.925
		1.47	0.803	0.673	0.562	0.440	0.766	0.589	0.686	0.794	0.778
		5.70	0.770	0.533	0.550	0.675	0.500	0.797	0.530	0.697	0.672
		21.49	0.684	0.353	0.535	0.360	0.395	0.767	0.493	0.597	0.415
	Proposed	0.38	0.928	0.797	0.879	0.939	0.713	0.713	0.930	0.859	0.764
		1.47	0.908	0.761	0.818	0.893	0.674	0.934	0.826	0.921	0.837
		5.70	0.890	0.715	0.843	0.847	0.675	0.931	0.809	0.856	0.791
		21.49	0.823	0.652	0.820	0.827	0.580	0.936	0.749	0.767	0.786

E Experiments

We use a SPAD512 single-photon camera from Pi Imaging to capture the following scenes. Unless otherwise stated, the timestamp resolution is 10 microseconds.

E.1 Additional Details for Figure 1 Experiment

We provide additional details for the experiment in Figure 1 of the main paper.

Scene. The elevator sequence is designed to contain motion at a broad range of inter-photon frequencies within a single video. In the same scene, we observe slow motion from the opening and closing elevator doors, medium-speed motion from a person repeatedly jumping, and high-frequency flicker from the elevator lights. These different dynamics correspond to different effective f_p values, making the reconstruction problem heterogeneous in difficulty even within one sequence. To further challenge the methods, we simulate increasingly photon-starved conditions by applying binomial thinning up to a factor of 1024, which reduces the detected photon counts and exacerbates the difficulty of recovering both the slow and fast components of the motion. It is established that Bernoulli thinning of events produced by a Poisson point process yields an inhomogeneous Poisson process with rate scaled by the thinning factor α [26].

Acquisition. We recorded binary photon detections over 1.3 seconds. The timestamp resolution was 10 μs. The average detection rate was 2117 photons per pixel per second (PPS), varying from over 50000 PPS in bright regions (ceiling lights) to an average of less than 100 PPS in the dark areas of the hallway.

Simulating inter-photon frequencies. We simulate different inter-photon frequencies by applying thinning to the acquired timestamps. We use thinning factors of 1, 1/8, 1/32, 1/64, 1/256, 1/512, and 1/1024, resulting in inter-photon frequencies f_p of 2.37, 18.2, 68.5, 130, 376, 853, and 5284, respectively.

E.2 Additional Details for Figure 6 Experiments

Figure 6 of the main paper contains experiments from several different acquisition settings, including quanta-sensor videos, simulated photon streams derived from conventional high-speed videos, and picosecond-resolution transient measurements. For the quanta camera captures, we reduce the amount of light reaching the sensor by stopping down the aperture, which physically darkens the scene and lowers the detected photon rate. When needed, we further synthetically thin the recorded photon streams relative to their original inter-photon-frequency levels to evaluate performance deeper in the inter-photon-limited regime. In the thinning factors listed below, a factor of 1 denotes the original photon stream, and small factors indicate more aggressive thinning relative to that original level.

Q-fan. We use the SPAD512 to capture of a rotating fan in a dark room. During capture, we additionally reduce the incident flux using aperture control. We evaluate thinning factors of 1 and $1/16$, resulting in inter-photon frequencies f_p of 44.7 and 5,188, respectively.

Q-nerf. We use the SPAD512 to capture shooting a foam-bullet from a nerf gun. As with the other real captures, we reduce the detected flux through aperture control and then further evaluate multiple thinning levels relative to the original photon stream. We use thinning factors of 1, $1/4$, $1/16$, and $1/256$, resulting in inter-photon frequencies f_p of 0.447, 1.41, 6.76, and 309, respectively.

Phantom-cereal. This scene is generated from a conventional high-speed camera video and converted into simulated photon streams. We use thinning factors of approximately $1/7.25$, $1/29.0$, and $1/116$, resulting in inter-photon frequencies f_p of 0.468, 1.82, and 7.08, respectively.

T-UWB-fan. This scene comes from the ultra-wideband single-photon imaging dataset [31], acquired using a picosecond-resolution single-pixel SPAD scanning setup. We use a thinning factor of $1/10$, resulting in an inter-photon frequency f_p of 1,412,537, and use the same measurements to reconstruct videos at two timescales, 100 kHz and 6 GHz.

Q-QBP-guitar. This scene uses the guitar dataset from Quanta Burst Photography [19]. We use thinning factors of $1/32$ and $1/256$, resulting in inter-photon frequencies f_p of 4.07 and 21.4, respectively.

Q-bit2bit-drill. This scene uses the drill dataset provided by the bit2bit authors [18]. We evaluate thinning factors of $1/8$ and $1/128$, resulting in inter-photon frequencies f_p of 1.995 and 32,359, respectively.

Figure F.3: Performance of our method under varying SNR levels. **Top:** Varying inter-photon frequency, constant SNR. **Bottom:** Varying inter-photon frequency with decreasing SNR (higher inter-photon frequency corresponds to lower SNR).

F Thinning with Fixed Dark Count Rate

In this section, we evaluate the performance of our approach under various signal-to-noise-ratio (SNR) levels in the scene captured in Figure 1. Note that uniformly thinning all timestamps by a constant factor does not change the SNR, since both signal photons and dark counts are reduced proportionally. To vary the SNR, we thin the original set of captured timestamps while maintaining the same level of dark counts.

We estimate the dark count rate of the SPAD512 single-photon camera by capturing binary frames with the lens cap on. After removing hot pixels following [18], we compute the average dark count rate as the total number of detected photons divided by the total acquisition time and number of pixels, yielding 25 counts per second (cps) for our camera. To simulate different SNR levels, we thin the original timestamps by a desired factor α and add timestamps from a homogeneous Poisson process with rate $(1-\alpha) \times 25$ cps. This reduces the signal by approximately factor α while maintaining a constant dark count rate of 25 cps, effectively lowering the SNR.

F.1 Results

We observe no significant degradation in our reconstructions when $f_p < 130$, corresponding to a thinning factor of 1/64, as shown in Figure F.3. However, at higher inter-photon frequencies, artifacts begin to appear, as indicated by the red ellipses in Figure F.3. This is expected since $f_p = 853$ corresponds to a thinning factor of 1/512, at which point the number of signal photons per frame becomes comparable to the dark count level.

References

[1] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *Int. Conf. Mach. Learn.*, pages 524–533, 2019. 2

[2] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18(11):2419–2434, 2009. 2

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 60–65, 2005. 2

[4] Canon Inc. Li3030sai cmos image sensor datasheet. https://mm.digikey.com/Volume0/opasdata/d220001/medias/docus/6222/LI3030SA.pdf. 4

[5] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[6] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[7] Zikang Chen, Tao Jiang, Xiaowan Hu, Wang Zhang, Huaqiu Li, and Haoqian Wang. Spatiotemporal blind-spot network with calibrated flow alignment for self-supervised video denoising. In *AAAI Conf. Artif. Intell.*, pages 2411–2419, 2025. 2

[8] Prateek Chennuri, Yiheng Chi, Enze Jiang, GM Dilshan Godaliyadda, Abhiram Gnanasambandam, Hamid R Sheikh, Istvan Gyongy, and Stanley H Chan. Quanta video restoration. In *Eur. Conf. Comput. Vis.*, 2024. 8

[9] Valéry Dewil, Jérémy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. Self-supervised training for blind multi-frame video denoising. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 2724–2734, 2021. 2

[10] Dawei W Dong and Joseph J Atick. Statistics of natural time-varying images. *Network: computation in neural systems*, 6(3):345, 1995. 4

[11] Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo, and Pablo Arias. Model-blind video denoising via frame-to-frame training. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11369–11378, 2019. 2

[12] Raja Giryes and Michael Elad. Sparsity-based poisson denoising with dictionary learning. *IEEE Trans. Image Process.*, 23(12): 5057–5069, 2014. 2

[13] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Eur. Conf. Comput. Vis.*, pages 624–642, 2022. 8

[14] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2129–2137, 2019. 2

[15] Joo Chan Lee, Daniel Rho, Jong Hwan Ko, and Eunbyung Park. Ffnerv: Flow-guided frame-wise neural representations for videos. In *ACM Int. Conf. Multimedia*, pages 7859–7870, 2023. 2

[16] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):356–371, 2022. 2

[17] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. In *Eur. Conf. Comput. Vis.*, pages 267–284. Springer, 2022. 2

[18] Yehe Liu, Alexander Krull, Hector Basevi, Ales Leonardis, and Michael Jenkins. bit2bit: 1-bit quanta video reconstruction via self-supervised photon prediction. In *Adv. Neural Inform. Process. Syst.*, pages 88443–88485, 2024. 2, 5, 8, 10, 11

[19] Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Trans. Graph*, 39(4):79–1, 2020. 2, 5, 8, 10

[20] Long Mai and Feng Liu. Motion-adjustable neural implicit video representation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[21] Anagh Malik, Noah Juravsky, Ryan Po, Gordon Wetzstein, Kiriakos N Kutulakos, and David B Lindell. Flying with photons: Rendering novel views of propagating light. In *Eur. Conf. Comput. Vis.*, pages 333–351. Springer, 2024. 4, 5

[22] Kristina Monakhova, Stephan R Richter, Laura Waller, and Vladlen Koltun. Dancing under the stars: video denoising in starlight. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 16241–16251, 2022. 4

[23] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local pca. *J. Math. Imaging Vis.*, 48(2):279–294, 2014. 2

[24] Bruno Saravia Vega. The world in slow motion in 8K ultra HD. [Online Video]. Available: https://youtu.be/A2eNVDGesYY, 2021. 4, 8

[25] Dev Yashpal Sheth, Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter A Crozier, Mitesh M Khapra, Eero P Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *IEEE/CVF Int. Conf. Comput. Vis.*, 2021. 2

[26] Donald L Snyder and Michael I Miller. *Random point processes in time and space*. Springer Science & Business Media, 2012. 10

[27] Varun Sundar, Matthew Dutson, Andrei Ardelean, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Generalized event cameras. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 25007–25017, 2024. 8

[28] Hui Tian. *Noise analysis in CMOS image sensors*. Stanford university, 2000. 4

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *Int. J. Comput. Vis.*, 128(7):1867–1888, 2020. 2

[30] Vision Research. Emva 1288 data sheet. https://www.phantomhighspeed.com/-/media/project/ameteksxa/visionresearch/ametekphantomhighspeed/secure/emva-spec-updates/emvareport_explainerguide.pdf?la=en&revision=ed2f4972-f961-49cc-a6aa-7b6f28463200&hash=3B70D0D7CBC0E9B74A8C34EA833BD5C9. 4

[31] Mian Wei, Sotiris Nousias, Rahul Gulve, David B Lindell, and Kiriakos N Kutulakos. Passive ultra-wideband single-photon imaging. In *IEEE/CVF Int. Conf. Comput. Vis.*, 2023. 2, 5, 6, 10

[32] Qi Zhao, M Salman Asif, and Zhan Ma. Pnerv: Enhancing spatial consistency via pyramidal neural representation for videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[33] Huan Zheng, Tongyao Pang, and Hui Ji. Unsupervised deep video denoising with untrained network. In *AAAI Conf. Artif. Intell.*, pages 3651–3659, 2023. 2