Predicting Important Photons for Energy-Efficient Single-Photon Videography

Shantanu Gupta*, Varun Sundar*, Lucas J. Koerner,
Claudio Bruschini, Edoardo Charbon, and Mohit Gupta

wisionlab.com/project/predicting-important-photons/

Abstract—Single-photon avalanche diodes (SPAD) detect individual photons with fine temporal resolutions, enabling capabilities like imaging in near-total darkness, extreme dynamic range, and rapid motion. Due to these capabilities, and coupled with the recent emergence of high-resolution (> 1MP) arrays, SPADs have the potential to become workhorses for computer vision systems of the future that need to operate in a wide range of challenging conditions. However, SPADs' sensitivity comes at a high energy cost due to the underlying avalanche process, which consumes substantial energy per detected photon, limiting the scalability and practicality of high-resolution SPAD arrays. To address this, we propose approaches to predict and sample only the most salient photons for a given vision task. To this end, we design computationally lightweight photon-sampling strategies that allocate energy resources for detecting photons only in areas with significant motion and spatial variation, while continually adapting to changing signals. We demonstrate the effectiveness of the proposed methods in recovering comparable video to a fully-sampled SPAD capture using only a small fraction of the photons (up to $10\times$ fewer), across diverse real-world scenes with motion, high dynamic range, and varying light conditions.

Index Terms—Computational Photography, Single-Photon Avalanche Diodes, Practical Single-Photon Imaging

1 Introduction

Single-photon avalanche diodes (SPADs) are an emerging imaging modality capable of capturing individual photons [5], [6] at ultra-high speeds. SPADs are rapidly becoming the sensor-of-choice in LiDAR time-of-flight imaging, and are even being integrated into high-resolution passive cameras [5], recently reaching the Megapixel threshold [7], [8]. This breakthrough, for the first time, has provided sufficient resolution for SPADs to be deployed as general-purpose cameras across a range of computer vision tasks, such as object detection, simultaneous localization and mapping (SLAM), and semantic segmentation, under extreme conditions involving rapid motion, high dynamic range, and low light [9], [10].

SPADs can detect individual photons, which enables their impressive capabilities. However, this comes at a price: each photon detection costs energy, which is a *unique challenge* faced by SPADs. Unlike in CMOS cameras, photon detection energy in SPAD sensors is considerable, and increases with scene brightness [11] (Fig. 1(a)). Another challenge is that SPADs capture data at high frame rates, reaching up to 100,000 FPS [12], demanding large bandwidth. For example, a 10 megapixel SPAD camera operating at 100 kHz could consume around 60 Watts of power (as much as a commodity laptop!), and produce 1 terabit/sec of data, requiring over a hundred USB-3.0 cables connected in parallel to read out the sensor data! These resource costs are prohibitive for many practical applications and would severely limit the scope of SPAD sensors. Therefore, to enable widespread adoption of SPADs, it is imperative to develop capture and algorithm strategies that *considerably lower the energy consumption and bandwidth requirements* of SPAD arrays. To this end, we raise the following questions: do SPADs need to detect *every* photon, or can some be skipped? Can we *predict* which photons are likely to be the most informative before capture, allowing us to focus our often-scarce sensing and compute resources towards the salient ones, thereby optimizing energy and bandwidth use?

We observe that a photon-counting camera needs high photon sampling rates only in salient regions with *significant motion and spatial variation*; motion blur is imperceptible for a patch of constant intensity. Based on this observation, we predict which photon measurements are more pertinent through two computationally efficient operations: a spatial gradient and a temporal change-point detector. This salience measure adapts pixel-wise photon-sampling rates (Fig. 1(b,c)), allocating resources like photon detection energy to the most informative photons. Neighborhoods with strong gradients *and* motion, such as the moving vehicles in Fig. 1(c), are sampled at the maximal temporal resolution, while others are recorded at a slower rate. In addition to the salience-region predictor and a non-uniform approach to measuring photons, we devise post-capture algorithms that can produce low-noise yet blur-free videos from the modified photon stream. (Fig. 1d shows an example reconstruction of a traffic scene.) Our approach uses a small fraction of photon measurements as the full single-photon data on a variety of scenes containing both slow and fast motion, as well as significant illumination variations; Fig. 8 presents a grid of real-world results. In scenarios with high flux and small amounts of motion, the measurement fraction may be even smaller, resulting in higher energy savings.

The proposed techniques act as a resource allocator at the

* *denotes equal contribution.*
S. Gupta, V. Sundar, and M. Gupta are with the University of Wisconsin-Madison, USA 53703. Contact: sgupta@cs.wisc.edu.
L. J. Koerner is with the University of St. Thomas, St. Paul, MN, USA 55105.
C. Bruschini and E. Charbon are with the École polytechnique fédérale de Lausanne, CH-2002 Neuchâtel 2, Switzerland.

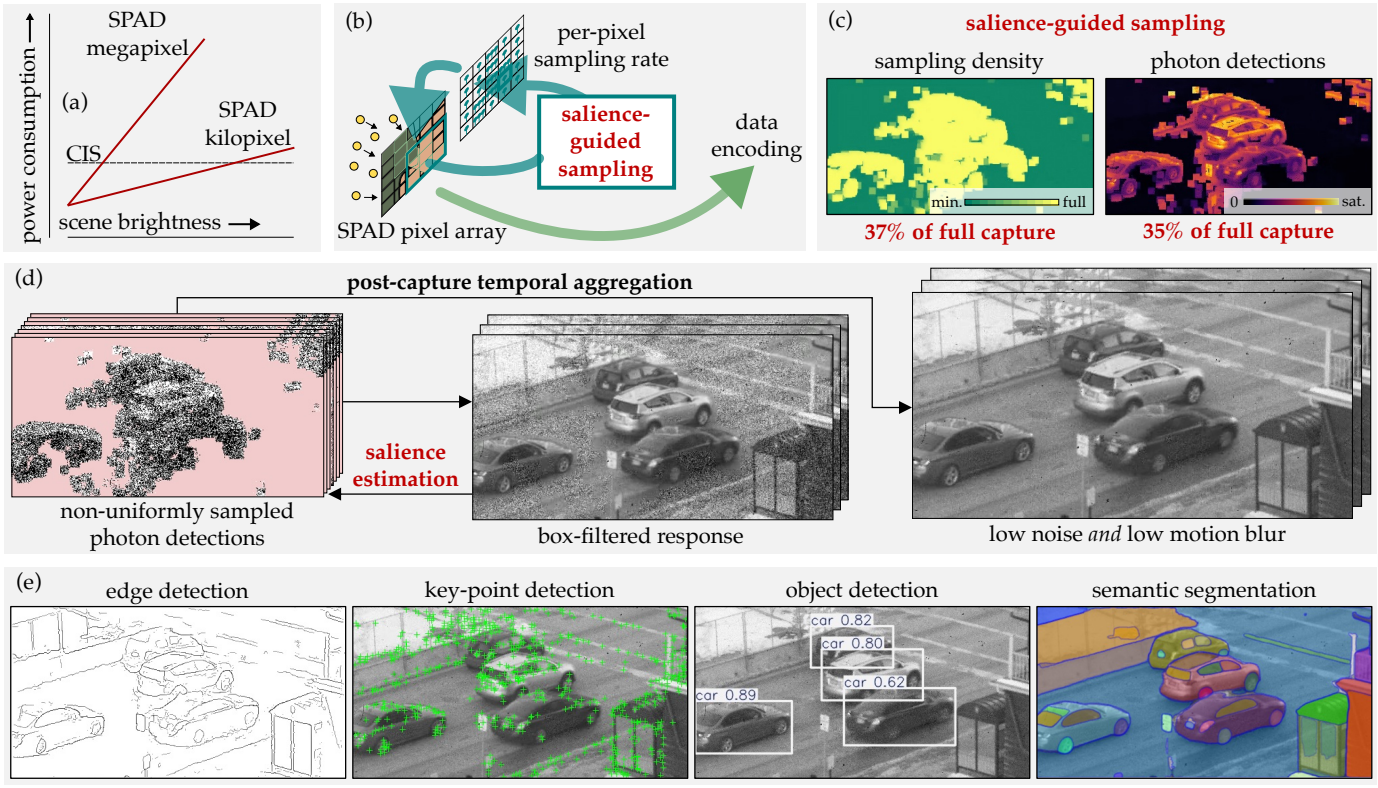


Fig. 1. **Salience-guided sampling for energy-efficient SPAD imaging.** (a) Unlike CMOS image sensors (CIS), the power consumption of SPAD sensors increases with scene brightness due to avalanches, which can limit the adoption of high-resolution SPAD cameras. (b) We propose a non-uniform sampling method for videography, where we use *salient* regions, which are regions with considerable spatial *and* temporal variation, to predict informative photons—moving cars in this scene. (c) Our salience-guided sampling reduces photon detections to $\sim$35% of full (dense) capture to lower the energy cost of avalanches. (d) Computing a box-filter response (Eq. (3)) of the salience-guided measurements reveals their non-uniform allocation with more measurements in salient regions. We propose post-capture temporal aggregators (Sec. 4.1) that combine non-uniformly sampled measurements to produce a low-noise *and* low-blur video. The resulting video enables plug-and-play inference of many downstream algorithms, such as the Structured Edge detector [1], SIFT feature detector [2], YOLOv11 object detector [3], and the Segment Anything Model [4].

sensor level, selectively blocking non-salient pixel locations to create a unique *pre-capture* computational imaging layer. Despite modifying the raw captured data, its nature and compatibility with downstream tasks are largely preserved. We demonstrate this with computer-vision tasks at multiple levels, including low-level (temporal aggregation), mid-level (edge and keypoint detection), and high-level (object detection and semantic segmentation) tasks, all of which successfully operate on the captured photon-stream data (shown in Fig. 1e). Our salience-guided photon-sampling approach can apply to a variety of (passive) SPAD pixel architectures; as case studies, we demonstrate its applicability to SPADs that capture photon measurements as binary-valued detections and exposure-bracket imaging models that are suited for capturing high-dynamic range scenes. Further, our capture policy can be coupled with eventful data encoding, providing bandwidth-efficient readout on top of capture-time power reduction. The proposed approach could become a modular component of scalable SPAD array designs in the future, paving the way for higher-resolution arrays and/or enabling deployment on power-constrained devices like cellphones, mobile robots, and AR/VR wearables.

The layout for the remaining text is as follows. Sec. 2 describes related work in the imaging literature, and Sec. 3 the mathematical model of SPAD sensing we assume throughout the paper. Sec. 4 describes how a sensor may implement a spatially varying measurement allocation, and how the resulting non-uniform measurements may be aggregated over time to recover a visually recognizable video. Sec. 5 describes a particular salience measure to drive this allocation, computed from image gradients and a temporal change-point detector. We present results with our overall design applied to real-world SPAD data in Sec. 6. Limits on the performance of this salience detector are identified via simulation of a synthetic probing signal in Sec. 7. We estimate the potential feasibility of the proposed method in the context of recent developments in SPAD image sensing in Sec. 8, and a more general discussion about other potential design choices follows in Sec. 9.

Scope and limitations. Our non-uniform allocation model assumes that SPAD pixels can be individually turned on and off at fast rates. We demonstrate results *emulated* from full dense measurements captured using a prototype SPAD array. Hardware on-chip implementation of our allocation scheme is an important next step to assess feasibility—especially of the energy overheads of salience guidance and resource allocation. Further, by taking fewer photon measurements, we fundamentally trade off imaging and vision performance for energy savings; our goal is to design practical techniques that strike a balance between the two. While our work does not present a formal analysis

of this tradeoff, a quantitative evaluation of imaging and edge-detection performance is included in Sec. 7.3, and in Appendix F within the supplement. Limitations to our approach include challenging scenarios with low light and high-speed motion, which pose challenges to our salience-identification and guidance steps—we provide methods to gracefully handle these conditions.

2 Related Work

Motion-aware videography addresses the temporal resolution limits of traditional cameras that lead to motion blur. For instance, coded-exposure techniques for photography [13] have been extended to increase the effective frame rate of videos via per-frame exposure codes [14] or multi-bucket pixels [15], [16]. Post-capture processing of optical flow balances temporal and spatial resolution at the pixel level to minimize blur and preserve detail [17]. Recently, sensors with closed-loop adaptation of pixel-wise exposures have been developed for high-dynamic range video [18], [19]. Similar to our method, Naghara et al. [20] calculate a motion mask to select one of two pixel-wise exposure patterns for minimal blur or image quality. Our work extends these approaches to the domain of single-photon videography, which is constrained by energy consumption rather than sensor noise.

Scene-adaptive imaging. Foveation, an example of scene-adaptive resolution allocation, can be realized in imaging systems by redistributing angular resolution with micro-electromechanical (MEMs) mirrors [21] or electronically [22]. LiDAR systems can improve efficiency in a scene-adaptive manner by steering the active illumination [23] or gating laser pulses based on motion [24]. Other examples of scene adaptivity include optimal exposure-time sequences [25], pixel-wise transmission masks [26], and light modulators [27] that allocate resources across input intensities for efficient HDR imaging. Bio-inspired event cameras reduce bandwidth by producing scene-adaptive, parsimonious readout in response to brightness changes [28], [29]. Event encoding has been applied to SPAD cameras to reduce bandwidth, but this application does not mitigate avalanche power [10]. Recent work proposed on-sensor inhibition to improve the photon efficiency of *static* SPAD imaging [30]. In contrast, we target single-photon videos and adaptively allocate energy constraints across a range of input motions.

Passive SPAD imaging adoption and capabilities are growing. Recently, SPAD cameras have achieved [8] and exceeded 1 Mpixel resolution [7], used timing information to expand the dynamic range [31], [32], and extracted intensity fluctuations >GHz [33]. Advances in pixel architectures have partially addressed light-dependent power for intensity imaging [34], [35], but do not incorporate higher-level salience when optimizing a pixel's power consumption. SPADs capture photon detections as digital quantities, and so are inherently compatible with on-sensor computation; a near-sensor imager with a reconfigurable processor [36] has been applied to emulate motion-compensating cameras [37] and event-based readout [10]. Other single-photon cameras, such as jots [38], [39] and superconducting nanowire single-photon detectors (SNSPDs [40]) either lack the temporal resolution of SPADs or do not enable on-sensor processing.

3 Background: SPAD Imaging Model

SPAD sensors can capture photon detections as a sequence of binary-valued video frames, at speeds as high as 100 kHz. Concretely, let the average number of photons that arrive per exposure at pixel location $\mathbf{i}$ and frame index n be $H[\mathbf{i}, n]$. By assuming photon arrivals follow a Poisson process [41], we can model the SPAD sensor's output as a Bernoulli random variable

$$\mathcal{Z}[\mathbf{i}, n] \sim \text{Bernoulli}(1 - \exp(-H[\mathbf{i}, n])). \tag{1}$$

We do not model quantum efficiency or dark counts explicitly and instead fold them into the definitions of H. Further, our exposition focuses on SPADs operated using a clocked-recharge mechanism [11] and with a fixed gate duration [12]. However, our proposed techniques operate at a higher level of abstraction than SPAD recharge policies and can be extended with appropriate modifications to other recharge mechanisms, e.g., event-driven recharge [42].

4 Salience-Guided Photon Sampling

Our goal is to improve the resource efficiency of SPAD arrays by predicting the *salient* photons to sample, and entirely skipping the measurement of non-salient photons. Such photon-sampling policies can lower resource costs by reducing the number of measurements, which can lower readout energy, and the number of photon detections, which directly reduces avalanche energy[1]. We design our photon-sampling model to operate temporally in a block-wise manner. Within each block, we reduce measurements, or *inhibit* them [30], using uniform sub-sampling that is determined on a per-pixel basis. Specifically, assuming an integer block size M (128 or 256 in most of our experiments), these sub-sampled measurements are represented by

$$\begin{gathered}\mathcal{Z}_{\text{block}}[\mathbf{i}, l, p] \coloneqq \mathcal{Z}[\mathbf{i}, lM + p \cdot \mathcal{S}[\mathbf{i}, l]] \\ l \in \{0, 1, 2, \ldots\}, p \in \{0, 1, 2, \ldots, \lfloor M/\mathcal{S}[\mathbf{i},l] \rfloor - 1\}.\end{gathered} \tag{2}$$

Here, $\mathcal{S}[\mathbf{i}, l] \in \{1, 2, 3, \ldots, M\}$ is the sub-sampling factor which we update over time for each pixel location. We initially set $\mathcal{S}[\mathbf{i}, 0] = 1$ for all $\mathbf{i}$, representing a densely-sampled set of measurements in the first block. When $\mathcal{S}[\mathbf{i}, l] > 1$, we take fewer measurements compared to dense sampling. Further, l denotes the block index and p denotes the index of a 1-bit measurement within a block. Finally, while our sampling model is temporally uniform within a block, since we vary the sub-sampling factor $\mathcal{S}$ across blocks, our overall approach is non-uniform across space and time.

To fully describe our photon-sampling model, we first specify how to temporally aggregate the non-uniformly sampled values $\mathcal{Z}_{\text{block}}$ to recover scene intensities and for downstream processing. Next, we specify how the allocation rule, determined by $\mathcal{S}$, can be driven by a salience measure—a rule to predict which spatiotemporal photon detections are important. Sec. 5 provides a prototypical salience measure based on image gradients.

1. The relative importance of detections and measurements to the overall resource cost depends on the hardware implementation (e.g., avalanche quenching mechanisms). Thus, we individually report the number of measurements and photon detections in our results.

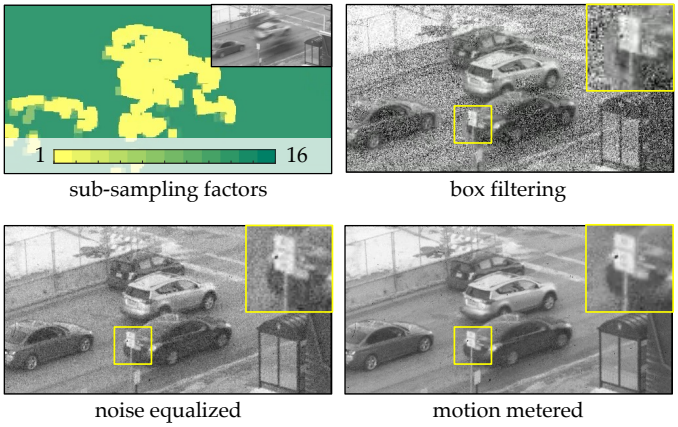


Fig. 2. **Temporally integrating non-uniformly sampled measurements**. Top: measurement allocations in this scene prioritize regions with motion and spatial gradient. Inset shows a long exposure depicting the range of motion in this scene. Considerable noise persists with box filtering (Eq. (3)) in pixels with fewer allocations (e.g., near the bus-stop sign and road here). Bottom: our noise-equalized aggregator (Eq. (5)) reduces noise in these regions by integrating over multiple allocation blocks. Our motion-metered aggregator (Eq. (8)) further decreases noise by accounting for the motion observed at each pixel.

4.1 Aggregating Sub-sampled Measurements

Measurements $\mathcal{Z}_{\text{block}}$ are non-uniformly sub-sampled SPAD outputs, which are binary-valued and, owing to photon noise, stochastic quantities that cannot be readily interpreted or consumed by downstream perception. For this reason, we temporally aggregate $\mathcal{Z}_{\text{block}}$—a simple approach is to sum measurements across time, like box filtering:

$$\mathcal{Z}_{\text{box}}[\mathbf{i}, l] \coloneqq \frac{1}{\lfloor M/\mathcal{S}[\mathbf{i},l] \rfloor} \sum_{p} \mathcal{Z}_{\text{block}}[\mathbf{i}, l, p]. \tag{3}$$

Fig. 2 *(top row)* shows an example of the box-filter response along with its per-pixel sub-sampling factors (which also corresponds to the scene shown in Fig. 1). Since the sub-sampling factor prioritizes moving regions with strong gradients, we observe that slow-moving and textureless regions, which are allocated fewer measurements, feature considerable photon noise. Fortunately, it is possible to reduce noise in these regions by temporally smoothing values, e.g., by computing an exponential moving average (EMA) across multiple blocks:

$$\begin{aligned}\mathcal{Z}_{\text{EMA}}[\mathbf{i}, l] \coloneqq \mathcal{A}[\mathbf{i}, l]\, \mathcal{Z}_{\text{box}}[\mathbf{i}, l] \\ + (1 - \mathcal{A}[\mathbf{i}, l])\, \mathcal{Z}_{\text{EMA}}[\mathbf{i}, l-1].\end{aligned} \tag{4}$$

We set $\mathcal{Z}_{\text{EMA}}[\mathbf{i}, 0] = \mathcal{Z}_{\text{box}}[\mathbf{i}, 0]$ for all $\mathbf{i}$. The EMA decay coefficient, $0 \leq \mathcal{A}[\mathbf{i}, l] \leq 1$, controls the extent of temporal smoothing and can be set based on the sub-sampling rate. For instance, using

$$\mathcal{A}_{\text{noise-equalized}}[\mathbf{i}, l] \coloneqq \frac{2}{1 + \mathcal{S}[\mathbf{i}, l]} \tag{5}$$

returns the original box-filter response if there is no sub-sampling ($\mathcal{S}[\mathbf{i}, l] = 1$), and increases the extent of smoothing as $\mathcal{S}$ increases. Eq. (5) is derived so that for a motionless scene, the variance reduction is independent of the sub-sampling rate, making noise more spatially uniform (see Appendix A in the supplement for details). We shall refer to the exponentially-smoothed output that uses Eq. (5) as the *noise-equalized* response. Fig. 2 *(bottom left)* shows an example noise-equalized response: we notice near-uniform photon noise in the image, invariant to the measurement allocation. Our noise-equalized response is compatible with burst-photography approaches [43] that are commonly used in state-of-the-art SPAD intensity-restoration techniques [44], [45], [46]; we exploit this compatibility in Sec. 6.1. However, burst photography can be computationally expensive. We now introduce a lightweight approach for producing low-noise and low-blur video outputs.

Adaptive Temporal Integration via Motion Metering

Our noise-equalized aggregator is agnostic to the extent of motion in a scene. Consequently, its temporal smoothing can be too conservative in slow-moving regions, while simultaneously producing motion blur when there is fast motion. We now aggregate non-uniformly sampled measurements in a *motion metered* manner. To do this, we set the EMA decay coefficient $\mathcal{A}$ according to the level of motion observed at a pixel—allowing us to strongly reduce photon noise in slow-moving and textureless regions. We track the temporal distribution of a pixel's measurements and model the time to the last abrupt change, or the *run-length*. We use the run-length to drive the EMA decay, denoted as $\mathcal{A}_{\text{motion}}$.

We adopt a Bayesian change detection algorithm, BOCPD [47], [48], and maintain K forecasters (typically 20), $\{\nu_k\}_{k=1}^{K}$ at each spatial location $\mathbf{i}$. These forecasters are initialized at frame numbers $\{n_k\}_{k=1}^{K}$, and denote the probability of the run-length at frame index n being $\{n - n_k\}_{k=1}^{K}$. The expected run length is then given by

$$\mathcal{R}[\mathbf{i}, n] = \sum_{k=1}^{K} \nu_k (n - n_k). \tag{6}$$

To reduce the motion blur of the aggregator, we apply minimum filtering to runlengths $\mathcal{R}$ in local windows (e.g., 9×9 pixel patches), producing $\mathcal{R}_{\text{min-filt.}}$. We then set

$$\mathcal{A}_{\text{motion}}[\mathbf{i}, l] = 1 - \exp(-\mathcal{S}[\mathbf{i}, l] / \mathcal{R}_{\text{min-filt.}}[\mathbf{i}, lM]). \tag{7}$$

This particular form of Eq. (7), notably the exponentiation, arises from using the run-length estimate to drive the time constant of the exponential moving average. We provide more details on how the forecasters are initialized and updated in Appendix B in the supplement. Since BOCPD operates on Bernoulli random variables, our motion-metered aggregator directly operates on 1-bit SPAD data as

$$\begin{aligned}\mathcal{Z}_{\text{motion}}[\mathbf{i}, n] \coloneqq \mathcal{A}_{\text{motion}}[\mathbf{i}, \lfloor n/M \rfloor]\, \mathcal{Z}[\mathbf{i}, n] \\ + (1 - \mathcal{A}_{\text{motion}}[\mathbf{i}, \lfloor n/M \rfloor])\, \mathcal{Z}_{\text{motion}}[\mathbf{i}, n-1].\end{aligned} \tag{8}$$

Fig. 2 *(bottom right)* shows the improved noise reduction of our motion-metered aggregator by contrasting its response to the box filter and the noise-equalized response.

Our run-length modeling assumes the underlying flux to be a piece-wise constant function. However, the temporal aggregator is capable of capturing more general flux changes. As an example, Fig. 3 shows the motion-metered response to a moving Gaussian blob. We further analyze the performance of motion metering in response to a moving edge in Sec. 7.2. We shall use our motion-metered aggregator as our default choice for producing intensity outputs from non-uniformly sampled measurements.

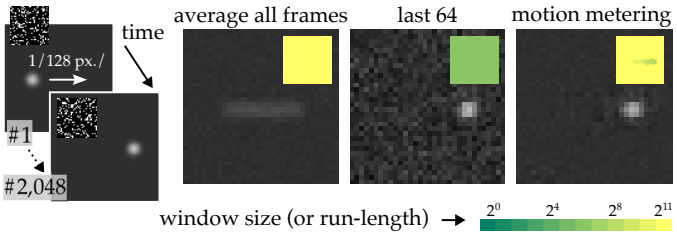


Fig. 3. **Motion-metered response to a moving Gaussian blob**. *(left to right)* We simulate 2048 SPAD frames from a moving point-like Gaussian signal (at a speed of $1/128$ pixels per frame). We illustrate the noise-blur tradeoff with a spatially uniform temporal integration, with the maximal and a shorter window size, respectively. Motion metering (Eq. (8)) results in a spatially-varying integration window (inset), with stationary points benefiting from reduced noise, but without blur in dynamic regions.

4.2 Updating the Measurement Allocation

We update the sub-sampling factor $\mathcal{S}$ over time as

$$\mathcal{S}[\mathbf{i}, l+1] = \min\left(S_{\max.}, \texttt{round}\left(1/\mathcal{W}[\mathbf{i},l]\right)\right), \tag{9}$$

where round is towards the nearest integer, and $\mathcal{W}[\mathbf{i}, l] \in [0, 1]$ denotes the salience measure used. In the next section, we describe a prototypical measure based on estimated image gradients. We also set an upper limit on the extent of sub-sampling: using $S_{\max.} \leq M$ to ensure that at least one measurement is made over the block. The optimal setting of $S_{\max.}$ is scene- and motion-dependent, and we presently treat it as a user-provided hyperparameter. At a high level, aggressive sub-sampling can introduce excessive noise in subsequent blocks and make it difficult to accurately identify salient pixels. An analog of Amdahl's law for the design of parallel computer systems also applies here, in that extreme sub-sampling yields diminishing returns on energy reduction, since the measurements and photon detections become more heavily determined by the most salient pixels—ideally, these salient pixels would not be sub-sampled. Concretely, in a scene where 10% of the pixel locations are determined to be salient and are not sub-sampled ($\mathcal{S} = 1$), increasing $S_{\max.}$ from 10 to 20 (a 100% increase) reduces the total measurements by just 23%.

5 Gradient-based Salience Estimation

Sec. 4 described how salience-guided measurements may be temporally aggregated for downstream consumption. In this section, we describe how salience maps may be derived from spatial and temporal gradients. Fig. 4 illustrates the overview of our proposed salience estimation. We identify scene regions with motion and strong spatial gradients, which we subsequently dilate to account for (potential) motion across temporal blocks and use to drive the allocation for the next-to-be-observed set of measurements.

We estimate the spatial gradient using the discrete differentiation kernels proposed by Farid and Simoncelli [49]. These are designed to obtain as rotationally symmetric a response as possible from a compact and separable set of filters. We use the 7-tap kernel, which is applied *directly* to box-filter response, $\mathcal{Z}_{\text{box}}$ (Eq. (3)), even when it has spatially-varying noise due to non-uniform measurement allocations. Direct application avoids errors arising from any excessive temporal smoothing in the adaptive integrators, and simplifies the estimation of noise in the gradient; the latter is described in Sec. 5.1. Our salience measure is initially based on the standard or z-score of the gradient, and is defined in Sec. 5.2. We augment it with a proxy of temporal gradients in the form of change points (abrupt changes within a measurement block), to mask out spatial gradients that are strong but from static or slow-moving regions, and therefore not necessary to sample rapidly. These details are described in Sec. 5.3.

5.1 Estimating Noise in the Gradient

We compute the spatial gradient as

$$\mathcal{G}_{\{x,y\}}[\mathbf{i}] \coloneqq \sum_{\mathbf{i}'} g_{\{x,y\}}[\mathbf{i}']\ \mathcal{Z}_{\text{box}}[\mathbf{i} - \mathbf{i}'], \tag{10}$$

where $g_{\{x,y\}}$ denotes the gradient kernel along one of the two spatial axes; we drop the temporal index l for brevity. Let $\mathcal{Z}_{\text{sm.}}$ denote a spatial smoothing of the box-filter response (we use a Gaussian blur kernel of standard deviation 3 pixels). The variance of $\mathcal{G}[\mathbf{i}]$ may be approximated as [50]:

$$\text{Var}(\mathcal{G}_{\{x,y\}}[\mathbf{i}]) \approx \frac{\mathcal{Z}_{\text{sm.}}[\mathbf{i}](1 - \mathcal{Z}_{\text{sm.}}[\mathbf{i}])}{\lfloor M/\mathcal{S}[\mathbf{i}] \rfloor} \sum_{\mathbf{i}'} |g_{\{x,y\}}[\mathbf{i}']|^2. \tag{11}$$

The sum expression in Eq. (11) can be pre-computed: for the gradient kernels used here, this evaluates identically for both x and y. The above assumes the entire neighborhood is acquired using the same number of measurements (see Appendix C in the supplement for details), which is inexact for general non-uniform sampling but is effectively correct in our design *at the pixel locations predicted to have strong gradients*, due to the dilation of salience measures which we discuss shortly.

5.2 Salience from z-scores

We convert the gradient response to a z-score as

$$\mathcal{Z}_{\mathcal{G}}[\mathbf{i}] \coloneqq \frac{\|\mathcal{G}_x[\mathbf{i}], \mathcal{G}_y[\mathbf{i}]\|_2}{\sqrt{\text{Var}(\mathcal{G}_{\{x,y\}}[\mathbf{i}])}}. \tag{12}$$

The above expression ignores the correlation between the x- and y-responses, and thus, implicitly assumes a Rayleigh distribution for the 2D gradient magnitude. This assumption underestimates the variance, somewhat offsetting the over-estimation in Eq. (11) from using a local mean flux estimate.

Next, we map the z-score to the $[0, 1]$ range. We define a scalar parameter z_0 representing an estimate of the noise floor, shared by all pixels and set ahead of time. In our experiments, we set $z_0 \in [2, 6]$, typically 3. We then compute the salience measure pixel-wise as (indexing omitted for brevity):

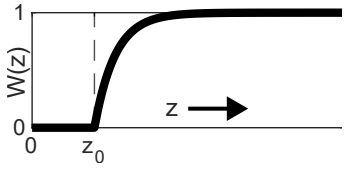


Fig. 5. z-score to salience mapping.

$$\mathcal{W}_{\mathcal{G}}(\mathcal{Z}_{\mathcal{G}}) \coloneqq 1 - \exp\left(-\max\left(0, \mathcal{Z}_{\mathcal{G}} - z_0\right)\right). \tag{13}$$

This z-score to salience mapping is sketched in Fig. 5.

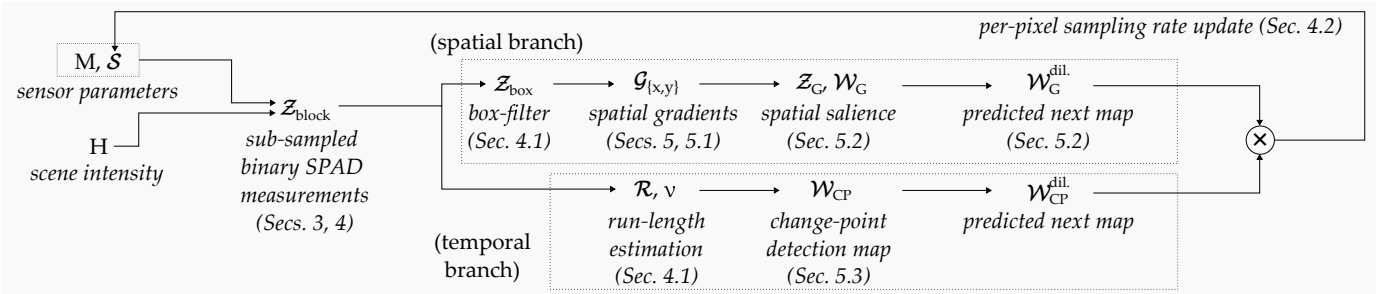


Fig. 4. **Outline of gradient-based salience estimation,** with references to relevant method sections. We select regions with strong spatial gradients and motion. The spatial and temporal selection masks are dilated to account for inter-block motion and subsequently drive the allocation for the next set of measurements. The spatial branch computes gradients on the box-filter response (Eq. (3)), while the temporal branch uses the run-length estimates constructed for motion metering (Eq. (8)).

Predicting the Salient Pixels in the Next Block

A crucial step in our allocation scheme is to use the current set of salient pixels to predict the salient pixels in the next block. Because of the high temporal resolution of SPADs, when the block length M is not too large, the motion between adjacent blocks is small, at most a few pixels for the sensor and scenes tested here. We determine the salient pixels in the next block by dilating the (current) set of salient pixels—which we implement using a local maximum filter with a window size greater than the gradient kernel size plus twice the maximal motion between adjacent blocks. Dilation may be interpreted as enforcing a uniformly *and maximally* sampled image at the salient pixels. The final result, denoted as $\mathcal{W}_{\mathcal{G}}^{\text{dil.}}$, serves as the salience score in Eq. (9) and drives the sub-sampling rate for the next block.

5.3 Temporal Change Detection to Identify Static Boundaries

Regions with strong spatial gradients but which are (nearly) static do not require maximal allocation in each block, since the motion metering of Sec. 4.1 automatically extracts a longer integration window, so infrequent measurements are sufficient for a low-noise estimate. For this reason, we pick points with strong spatial *and* temporal changes detected locally, which we implement as point-wise binary masking with the (non-dilated) salience measure $\mathcal{W}_G$. We note that masking is different from using a spatio-temporal gradient, which would be functionally closer to an OR gate.

We detect changes using the per-pixel forecasters mentioned in Sec. 4.1 by keeping track of the forecaster associated with the binary-frame index that corresponds to the last abrupt change. Denoting this forecaster by ν_{k^*}, we detect a change point whenever $\nu_{k^*} < \max_k \nu_k$. Once a change is detected, we set $k^* = \arg\max_k \nu_k$. A similar dilation as for the spatial salience is performed here as well.

Fig. 6 shows an outline of the measurement allocation process for the traffic scene from Fig. 1. Spatial gradients result in allocations towards static regions such as road signs and markings, and the bus stop on the lower right. Change-points have many false positives outside the moving objects in the scene. However, when combined, the resulting measurement allocation pattern focuses on regions in the proximity of strong spatial gradients and temporal changes, and hence reduces the resource consumption.

Rapidly changing illumination: the combination of spatial and temporal cues plays a significant role in improving energy-efficiency in scenes with artificial lighting, which contains a rapidly varying flicker component (Fig. 7). In these conditions, change-points detected pixel-wise may not be localized to textured regions, and if we were to rely solely on temporal changes to drive the salience measure (and the subsequent measurement allocation), the gains in energy-efficiency would be limited (54% photon detections in the example of Fig. 7). By combining temporal changes with spatial gradients, the allocation scheme is more selective (21% detections), while still yielding a comparable image.

5.4 Fallback to Uniform Sampling on Detector Failure

In very low light or when the sensor is saturated (flux $>>1$ photons per pixel per binary frame, or 1 ppp), the salience measure, $\mathcal{W}_G$, may be zero at nearly all pixels. In this case, Eq. (9) allocates very few measurements, creating inaccurate gradients, and potentially leaving the sensor unprepared for any eventual changes in illumination levels. To prevent this, we fall back to uniform sampling when the number of pixels with non-zero $\mathcal{W}_G$ is under a threshold (around 2.5% of pixels). With a fixed threshold, we occasionally observe oscillations in the sampling rates (some examples may be seen in the videos included in the supplementary material). These may be prevented with hysteresis—by keeping track of past allocations—in a future version. We note that variations in sampling rates do not generate video artifacts because the motion-metered response accounts for irregular time gaps between samples [51].

5.5 Extension to Exposure Brackets for HDR Imaging

Exposure brackets, i.e., measurements sequences with different exposure times (like brackets in conventional CMOS imaging), are often used for efficiently imaging high-dynamic range scenes—since individual binary frames can be quite wasteful for measuring flux levels when photon-detection probabilities are very close to 0 or 1. However, exposure-bracketed measurements cannot be modeled as identically Bernoulli- or binomial-distributed for a given true flux level, invalidating the temporal integration approaches described in Sec. 4.1. Further, the original noise model for the spatial gradient given in Sec. 5.1 also cannot be applied directly to exposure-bracketed measurements.

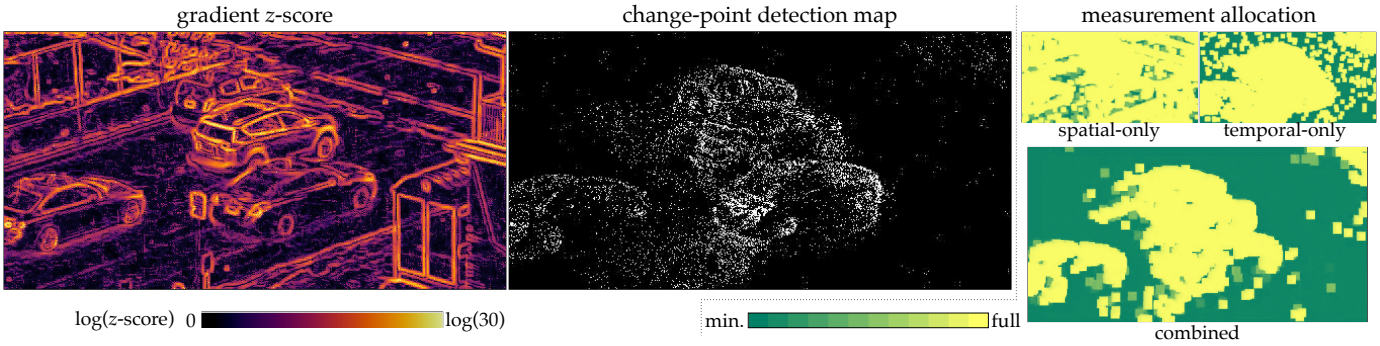


Fig. 6. **Combining of spatial and temporal cues for measurement allocation**. Left & middle: gradient z-scores and a per-pixel binary map of changes detected, for the scene shown in Fig. 1. Right: measurement allocations based on using spatial gradients alone (total measurements: 85% of the full data), change-points alone (68%), and their combination (38%).

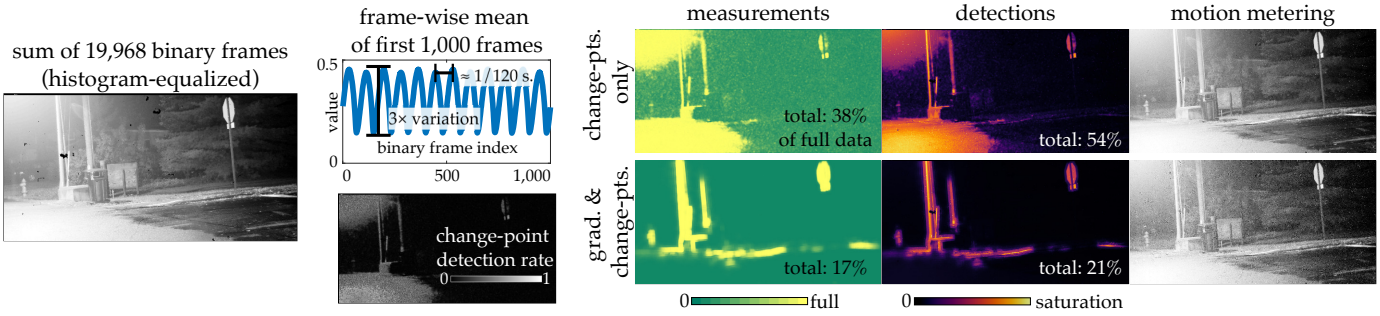


Fig. 7. **Response to rapid illumination changes**. This scene is imaged by a SPAD at 10 kHz from a stopped vehicle at night. There is no camera or scene motion, but substantial illumination variation due to the street lamp flicker. Left: an average of all frames smooths the flicker. Second from left, top: frame-wise mean values reveal a flicker frequency of 120 Hz; bottom: change-points are detected continually at nearly all strongly illuminated pixels. Right half, top row: when measurements rely solely on change-points (dilating over a 3×3 window), resources are inefficiently allocated to untextured regions. Bottom row: combining change-points with spatial gradients restricts the allocation to edge-like pixels and reduces detection energy by $2.5\times$. Right-most column: motion-metered response (Eq. (8)).

We now construct a version of the proposed salience-guided sampling method using *spatial* gradients and the *noise-equalized* response. (Because of technical difficulties in working with non-Bernoulli distributions, we do not extend motion metering and temporal change detection to this version.) We convert bracketed measurements to a (maximum-likelihood) flux estimator and substitute the binomial variance expression in Eq. (11) with the variance of this flux estimate. Unfortunately, the latter is challenging to derive analytically for the maximum-likelihood estimator. Instead, we use a proxy, in the form of the optimal *linear* combination of the bracketed measurements, which has been previously analyzed in detail [52]. The rest of the method proceeds similarly to the case of a single fixed binary-frame exposure, including compensating for non-uniform sample allocation via the noise-equalizing aggregator. A detailed description of maximum-likelihood estimation for exposure bracketing is provided in Appendix D within the supplement, including the variance expression for the linear combination proxy.

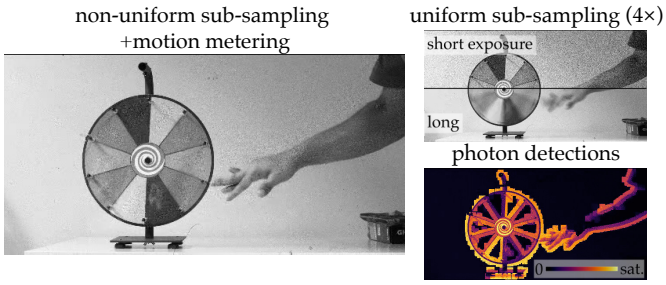
Our compatibility modifications for exposure bracketing show the requisites for extending the proposed allocation process to other SPAD imaging models [30], [32], [35], [42], [53]: 1) a per-pixel flux estimator accessible at the temporal block granularity and 2) its accompanying noise model.

6 Results on Real SPAD Data

We demonstrate salience-guided sampling on a variety of scenes captured using the SwissSPAD2 sensor [12], which has a spatial resolution of 512×256 (in half-array mode); the sensor is run at 96.8 kHz in most scenes. As a pre-processing step, we replace known "hot" pixels (i.e., very high dark count rate) with values from their nearest neighbors. The SPAD data is captured on-sensor densely, and our method is implemented in software.

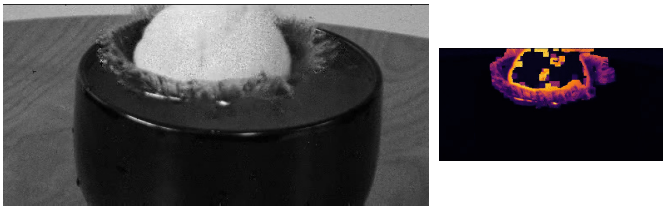
Fig. 8 shows that our salience detection largely succeeds in restricting the photon detection budget towards locations of motion. Naturally, a static camera offers more scope for efficiency gains—as seen in Fig. 8(a–d). When there is no motion (ego or subject), the allocations automatically drop, as seen in the left column of Fig. 8(e) where both the photon detections and measurements are under 10% of dense sampling. The presence of ego motion still permits significant savings due to many image regions being texture-less, especially in indoor scenes containing solid-color walls, such as the middle column of Fig. 8(e). Even in outdoor scenes, there is scope for savings (e.g., sky regions). Our proposed algorithm can run robustly without requiring manual intervention over long durations, which we show in Fig. 8(f); this is a low-light scene with significant illumination variations that leads to failures without the fallback option discussed in Sec. 5.4 (automatically resorting to dense sampling when too many salient pixels leave the field-of-view).

We show results for a high-dynamic range scene in the top rows of Fig. 9(a, b). Due to a combination of camera and subject motion, and intricate scene content, measurements reductions via *spatial* allocations are relatively small (∼30–



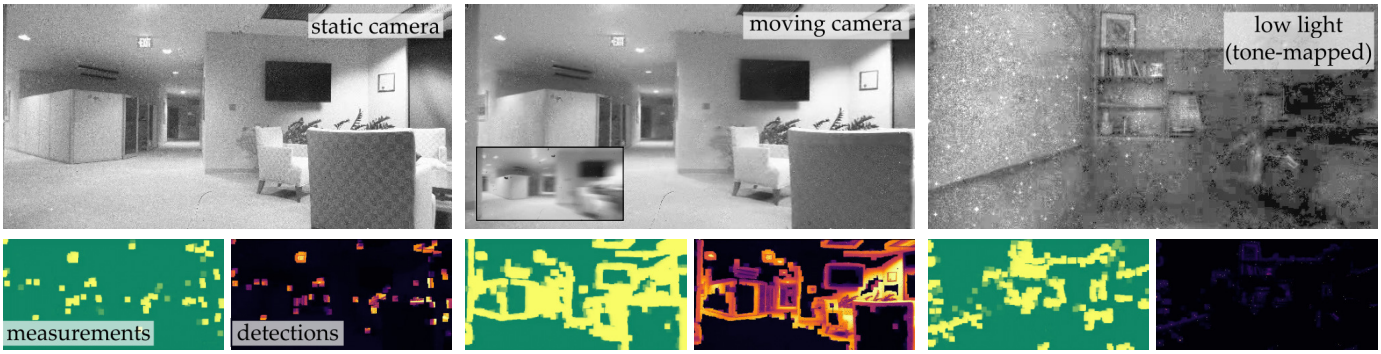
(a) *Spinning prize wheel.* Uniform sub-sampling leads to noise and blur with short and large block lengths, respectively. Salience guidance provides a higher-quality output, using $\sim 25\%$ of the dense data and photon detections. We used $M = 64$ and $S_{\text{max}} = 16$ in this scene.

(b) *Toy pops up.* Motion metering applied to salience-guided sampling captures the fast dynamics of a jack-in-the-box toy as it pops up. Inset shows the range of motion as a long exposure. Total detections and measurements are 24% of full data. We used $M = 256$, $S_{\text{max}} = 16$.

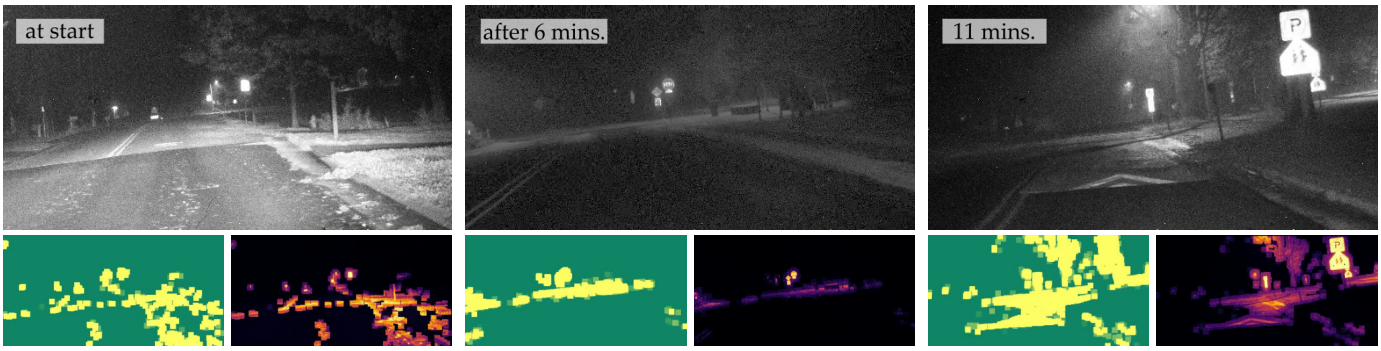


(c) *Water splash.* Our approach can also capture non-rigid dynamics, such as the water splashed by a falling tennis ball. Cumulative detections and measurements are 35% and 27%, respectively, of full data. We used $M = 128$ and $S_{\text{max}} = 16$ in this scene.

(d) *Winter stroll.* A slight camera motion results in branches and road outlines also getting sampled at a higher rate in addition to the moving pedestrians. Total detections and measurements are 57% and 53%, respectively, of full data. We used $M = 256$ and $S_{\text{max}} = 16$ here.



(e) *Hallways translation.* A SPAD camera (run at 16 kHz) is translated across a hallway in a moving cart. There are large illumination variations in this scene as the cart exits the hallway and enters a dark room (right column). Cumulative detections and measurements are both 42% of full data. We use GIMP [54] for tone-mapping the motion-metered response in low light. We used $M = 256$ and $S_{\text{max}} = 16$ in this scene.



(f) *Nighttime driving.* We run our algorithm continuously throughout this 15 min sequence (SPAD run at 10 kHz), with no manual intervention or reset. There are rapidly flickering light sources in the scene. We show the noise-equalized response instead of the motion-metered response, which is slightly too motion-blurred here. Road markings and signs are visible even after sub-sampling, with a total 30% of the detections of a dense capture (and 24% of the measurements). We used $M = 128$ and $S_{\text{max}} = 16$. The middle frame is originally very dark and is tone mapped for visualization.

Fig. 8. **Results on real SPAD data.** We show the motion-metered response, measurement allocations and photon-detection distributions on a variety of scenes. These scenes include slow and fast motion, rigid and non-rigid dynamics, and low light to daytime environments. The duration of these sequences is about 1 second in (a)–(d), 1 minute in (e), and 15 minutes in (e). We highlight measurement sparsities and photon-detection distributions, and the general lack of motion blur in the recovered frames. Please see the supplementary material for video visualizations.

40%). However, this reduction can be coupled with reduced photon detections that stem from using exposure brackets, which we show next.

6.1 Combination With Exposure Brackets for HDR

As described in Sec. 5.5, our approach is compatible with flux estimators from exposure bracketed measurements and using saliency determined from (only) spatial gradients. We demonstrate this compatibility in Fig. 9, where we use an exposure bracket with two measurements with the original SPAD binary exposure, and one measurement each with $4\times$ and $16\times$ the exposure, respectively. We use the "saturation look-ahead inhibition" policy that improves the efficiency of (sequentially captured) exposure brackets [30]; this policy inhibits photons at pixels predicted to saturate to reduce avalanche power with minimal impact on signal-to-noise.

We treat the bracketed data as a composite measurement made throughout 22 recharge periods. We define the block size for salience-guided sampling to be 6 of these composite measurements. At every block, a pixel makes between 1–6 composite bracketed measurements, directed by the gradient-driven computations described in Sec. 5. Given the original sampling rate (SPAD frame-rate) of 8 kHz, this means that the sampling budget is adjusted at a temporal resolution of approximately 60 Hz (estimated as 8 kHz/ 22 recharges / 6 composite measurements).

We show results for the bracket-based capture in the bottom rows of Fig. 9(a,b). As exposure brackets capture very few photons in a single cycle, a single temporal block does not quite yield a high-quality image, even with noise-equalized aggregation (Eq. (5)). For this reason, to produce a high-fidelity image, we aggregate information across larger spatio-temporal windows using burst reconstruction[2] [43], [44]. In strong lighting conditions (Fig. 9(a)), salience-guided bracketed measurements yield image reconstructions comparable to densely-sampled photon streams, while incurring 93% fewer detections than the full data. If we were to use brackets alone, without any spatially-selective allocation, the detection reduction is about 88%, thus highlighting the complementary and compounding effect of combining exposure-bracketing capture policies with our proposed salience-guided allocation schemes.

The high-dynamic range (HDR) setting, however, presents room for improvement. Low and intermediate flux levels are more challenging for reliably detecting salient pixels. While the pre-set threshold on the minimum number of salient pixels (5%) does detect very low light and falls back to uniform sampling, near the threshold some undesirable artifacts are possible as some salient pixels are detected but others are not, potentially leading to a spatially-varying motion blur (as seen in Fig. 9(b)). This challenge is also present with fixed-exposure data (see full video sequences in supplement for examples), but in HDR settings—where exposure brackets are typically used—the likelihood that *some* salient region is very dimly-lit is higher. Better detection of faint gradients becomes even more important in this scenario, perhaps using multi-scale processing [55].

2. Burst reconstruction is much more computationally costly than the noise-equalized or motion-metered response applied to a non-bracketed exposure, so we effectively trade off power consumption at capture versus downstream by using brackets.

6.2 Plug-and-play downstream processing

Together, our salience-guided allocation process and temporal integrators efficiently realize a high-speed video in the $\sim$ 100–1,000 Hz range. These video frames can be readily processed by many algorithms designed for single-photon imaging, such as the burst restoration techniques demonstrated in Fig. 9, and off-the-shelf computer vision algorithms (Fig. 1e). We expect this operating mode—where the output of our temporal integrators is directly consumed downstream—to be a default, (computationally) efficient option for many application scenarios.

Noise reduction at the temporal block-level: in scenarios where the temporally integrated responses are still too noisy (low light or extremely fast motion), we may apply a variety of image and/or video reconstruction algorithms to improve them first. Quanta Burst Photography, as shown in Fig. 9, is one such example, but other approaches such as QUIVER [46] or bit2bit [58] are also similarly applicable. Methods aimed at Poisson noise reduction [56] or even Gaussian denoising [57] may also be considered: while not optimal for the SPAD measurement distribution, they potentially offer other benefits like a faster implementation or being trained with larger datasets, and their performance may be acceptable in practice. Some example results are shown in Fig. 10 and compared with the original noise-equalized and motion-metered responses.

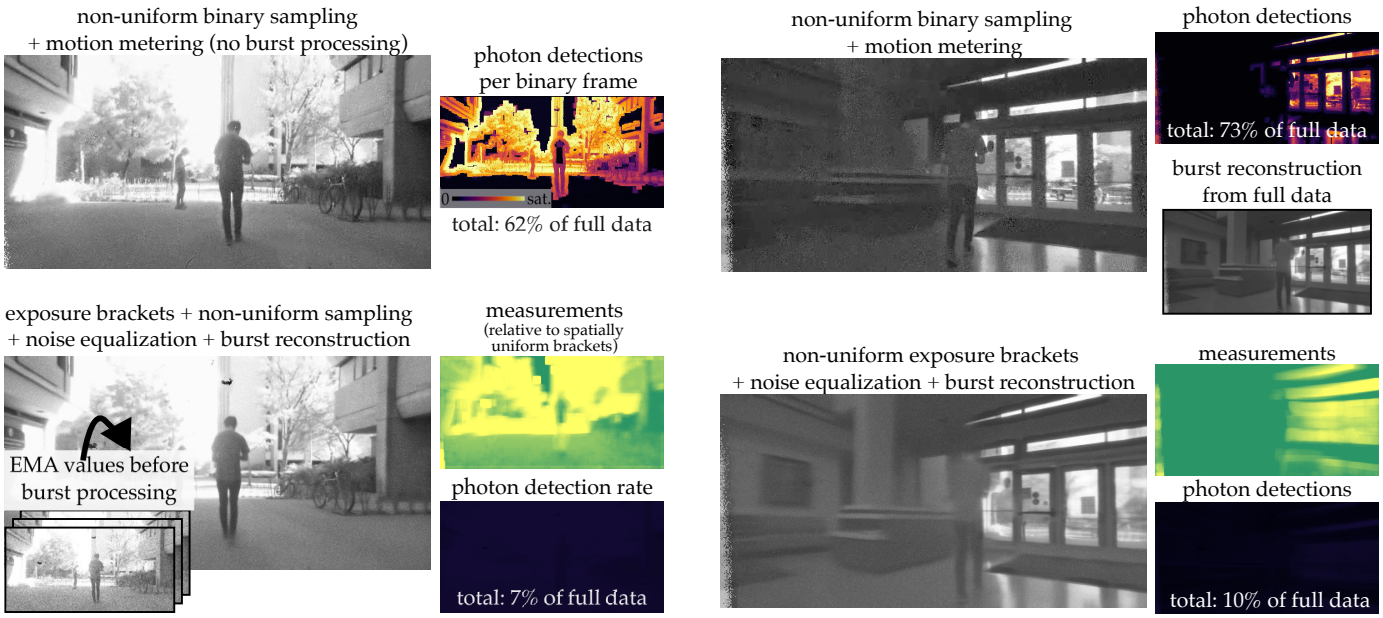
In-filling skipped binary measurements: while the proposed adaptive temporal integrators can hide the non-uniformity in sampling, reading them at block-wise granularity loses some information. At salient pixels, we would still have ultra-high-speed capture ($\sim$ 100 kHz), so an opportunity remains to fully exploit this temporal resolution, and extract and/or compensate motion even between binary frames in these locations, often in the sub-pixel range [44], [59]. However, most existing single-photon vision algorithms in these regimes expect a uniformly sampled photon cube, whether for reconstruction or higher-level tasks. For compatibility with these methods, we may *in-fill* or *hallucinate* the missing binary pixel data (light-red tinted regions in the left-most diagram in Fig. 1d), as simulated Bernoulli random variables. The next question is to set the probability of photon detection in this simulation, for which we use the motion-metered response as our estimator[3].

Fig. 11 shows burst reconstruction with such a data stream, for the scene shown in Fig. 1. Compared to reconstructing from the temporal integrator responses at block-wise granularity, we get similar results in the sub-sampled (and therefore hallucinated) regions: there is no free lunch here, as expected. However, fully-sampled patches do benefit from sub-pixel motion compensation at the binary-frame level, as demonstrated previously [44]. The sub-sampled/hallucinated regions can still be denoised further using existing approaches, similar to Fig. 10.

6.3 Bandwidth-Efficient Eventful Readout

We show further compatibility with recent event-camera inspired approaches that reduce SPAD data readout [10],

3. We may refine this estimate using another reconstruction algorithm (possibly in an iterative loop), but that direction is not pursued here.



(a) Strong-flux condition ($\sim$1 ppp). Top row results obtained similar to Fig. 8. Bottom row shows burst reconstruction applied to the noise-equalized response of non-uniformly sampled bracketed measurements; inset shows the noise-equalized response. The per-pixel distribution of the number of bracket cycles is shown in the green-tinged image, and the effective photon detection rate in the lower image. The detection rate with all bracket cycles active is 12% of the full data here, meaning a $\sim$40% reduction is obtained from non-uniform sampling (comparable to that for fixed-length exposures in the top row).

(b) Intermediate-flux condition ($\sim$0.1 ppp). Top row: similar to top row of (a) and Fig. 8, with burst reconstruction on the densely sampled data shown for reference. Bottom row shows corresponding results for the exposure brackets-based acquisition. In this scene, gradients in brighter regions (right half of image) are detected but not in the dimmer parts on the left, resulting in a spatially varying amount of motion blur. This behavior depends on the thresholds set for the gradient z-scores (here 3), and for the density of salient points required to be found in the image (here 5%), to fall back onto uniform sampling.

Fig. 9. **Results on a high-dynamic range scene.** A SPAD camera, operated at 8 kHz, captures a person walking out from a dimly-lit corridor onto the foyer and then outside in bright sunlight. Our compatibility with exposure brackets provides strong energy savings in HDR settings, particularly in very bright light. The foyer has scene regions with large intensity differences (as seen in b). For this scene, we used $M = 128$ and $S_{\text{max}} = 16$ when working with fixed-length exposures and $M = 6$ and $S_{\text{max}} = 6$ with exposure brackets. Please see the supplement for video results.

[37], [60]. We transmit the motion-metered integrator's response in an eventful manner, by transmitting its significant changes (absolute deviation > 0.05) between subsequent blocks (of size M). We encode the first motion-metered response as a dense frame and subsequent events using the compressed sparse column (CSC) encoding. We quantize the encoded integrator values using 10 bits and share frame-index information across the CSC-encoded event frame.

Fig. 12 shows eventful transmission of our motion-metered response on two scenes, one with subject motion and the other with camera motion, demonstrating a 8–20$\times$ and 50–100$\times$ readout reduction over dense-periodic transmission (every M binary measurements) and reading out raw photon detections, respectively. As the right-most column of Fig. 12 shows, intensity images can be recovered from the events by simply accumulating their values. We note that since motion-metering updates at the full 1-bit SPAD sampling rate (Eq. (8)), in principle, the frame-index information may also have the same temporal resolution, and comparable to event cameras.

7 Simulation-based Performance Analysis

We now present simulations that analyze the spatial and temporal branches of our salience-guided allocation process (overview shown in Fig. 4). We measure the accuracy and delay of salience detection for an example moving-edge signal (shown in Fig. 13a), and vary its speed of motion, the contrast in intensity across its two regions, and the mean flux level. We set the image resolution of the moving edge to be 32$\times$32, and generate 2048 SPAD frames; no subsampling is performed since the proposed algorithm targets a dense sampling near highly probable regions of saliency Our simulation aims to represent a highly localized spatio-temporal window containing a single moving edge, which motivates this choice of image resolution.

7.1 Spatial Branch: Gradient-Based Salience Detection

We use the 7-tap kernel, and sweep over the block size M used for calculating the box-filter response (Eq. (3)) that is centered on the middle frame of the sequence and used to estimate spatial gradients (Eq. (10)). We extract two sets of pixels from the ground-truth signal, representing true edge locations and known background pixels that are at least five pixels away from both the edge and the borders of the image (shown in Fig. 13a). From the thresholded z-scores of the gradient responses (set as 3), we calculate precision and recall, and in turn the F1-score.

We conduct 1024 trials and report the median F1 score across them in Fig. 13b. Salience is near-perfectly detected (F1-score close to 1) in many scenarios, especially for favorable settings where either the contrast is very strong or the motion is limited. Expectedly, while a longer block duration

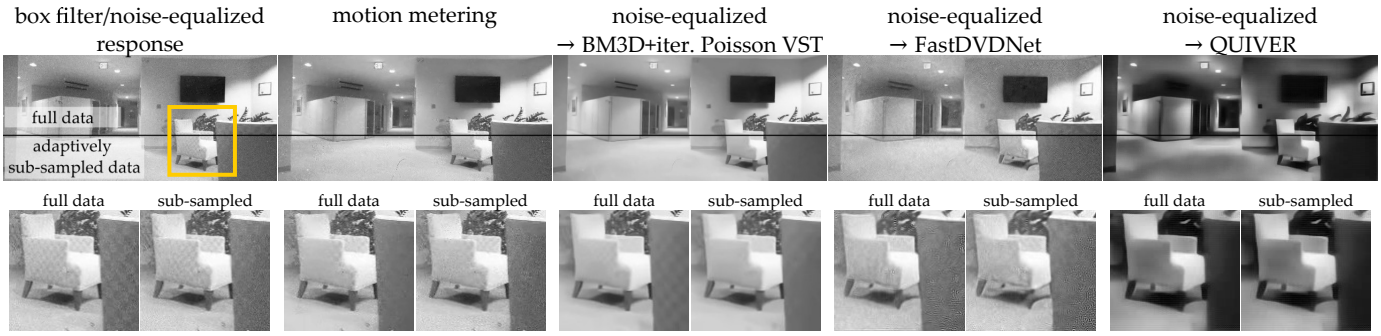


Fig. 10. **Plug-and-play reconstruction**. Comparing reconstruction results from non-uniformly sampled data with their equivalents extracted from the fully-sampled photon detection sequences, for the middle snapshot from Fig. 8e. Left-most two columns: the proposed adaptive temporal integrators; the fully-sampled equivalent of the noise-equalized response is the box filter ($M = 256$ for both). Right-most three columns: the noise-equalized response further denoised, using frame-by-frame BM3D for Poisson distributions [56], FastDVDNet for videos with Gaussian noise [57], and QUIVER for single-photon videos [46]. The QUIVER output is tone-mapped using Contrast-Limited Adaptive Histogram Equalization (CLAHE).

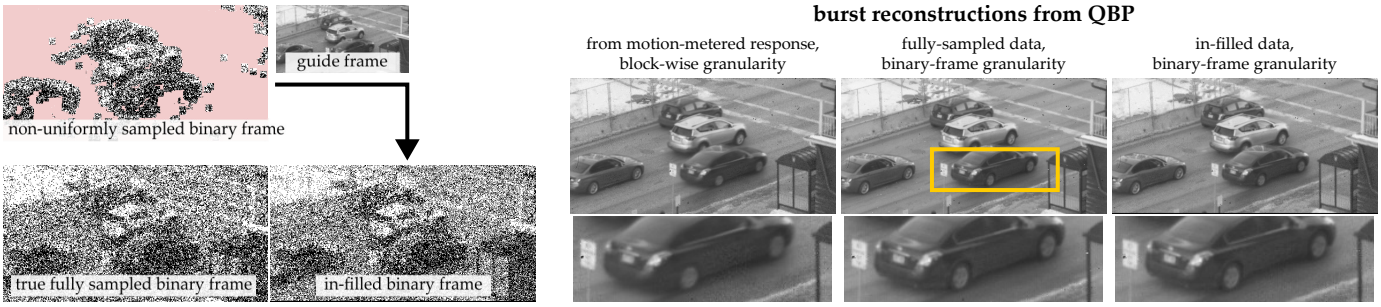


Fig. 11. **Re-generating a full binary photon cube**. Left: for the sequence of Fig. 1, binary samples are randomly generated at disabled pixels based on a guide frame (here the motion-metered response). The original fully-sampled binary SPAD data is shown for comparison. Right: quanta burst reconstruction [44] from three distinct inputs: from left to right, the motion-metered response from the non-uniformly sampled data, sub-sampled once per temporal block ($M = 256$); the original fully-sampled photon detections; and the synthetically in-filled detection sequence. The block-wise result from non-uniform sampling is close to that from the fully sampled photons at the maximal temporal resolution, except at the vehicle in the right lane which shows some motion blur. Using in-filled full-frame-rate data reduces this blur.

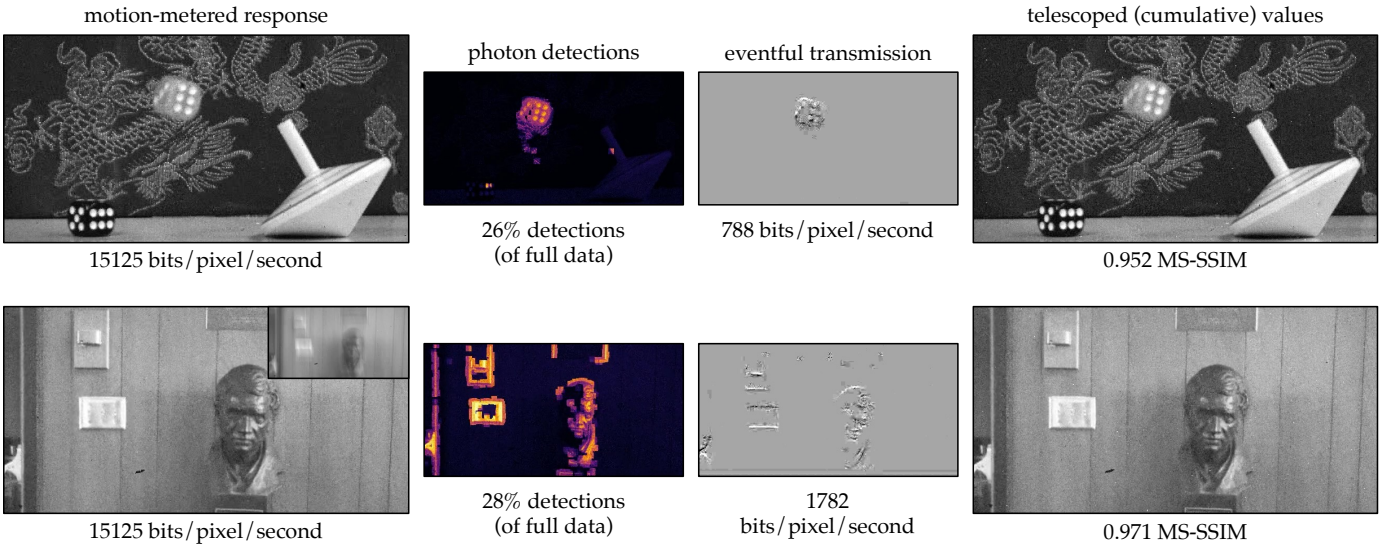


Fig. 12. **Eventful transmission for bandwidth-efficient readout.** We transmit significant changes in the motion-metered response (left-most column) across temporal blocks; the first response is encoded densely. Second from right: significant changes, or events, which predominantly encode motion information; gray denotes no change (we apply contrast stretching to more clearly show the significant changes). Eventful transmission can reduce readout by 8–$20\times$ and 50–$100\times$ compared to periodic dense readout (of motion metering) and over raw photon detections, respectively. Right most: intensity frames can be recovered from event frames by simply accumulating event values. MS-SSIM (higher is better) is computed between the original motion-metered response and accumulated event values.

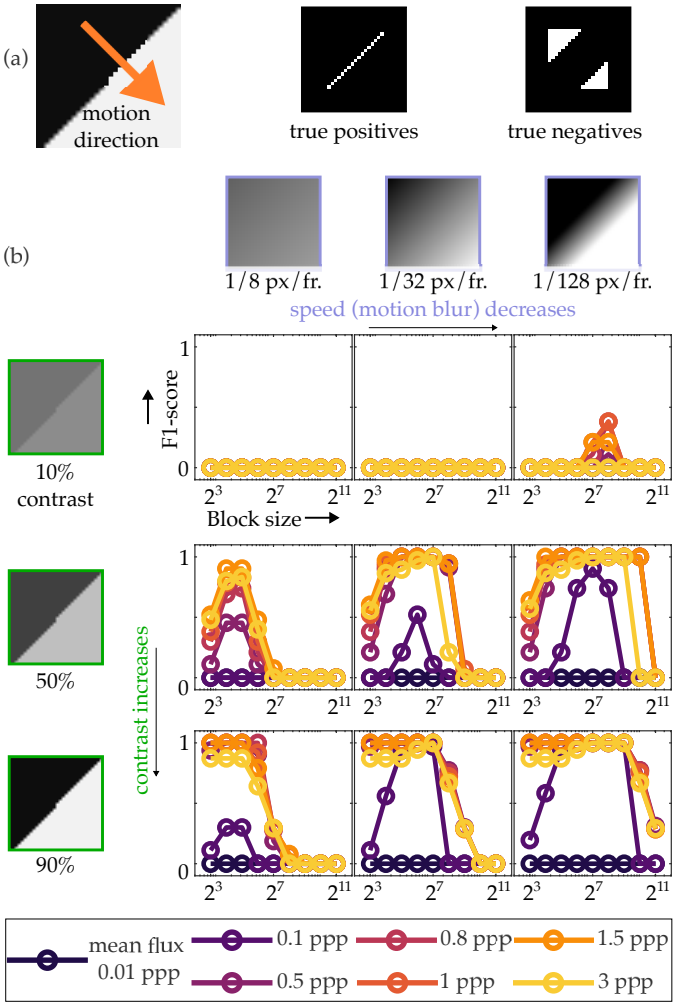


Fig. 13. **Performance of spatial-gradient estimation of a moving step edge.** (a) Visualization of reference signal and its motion direction, and the pixels considered for computing precision and recall. (b) F1-scores after thresholding z-scores under various signal and measurement parameters. Outer axes of the grid vary the motion velocity and signal contrast; within one graph the F1 score is plotted against the temporal aggregation window size M, for different mean flux levels. A score of 1 represents perfect detection: full recall and no false positives.

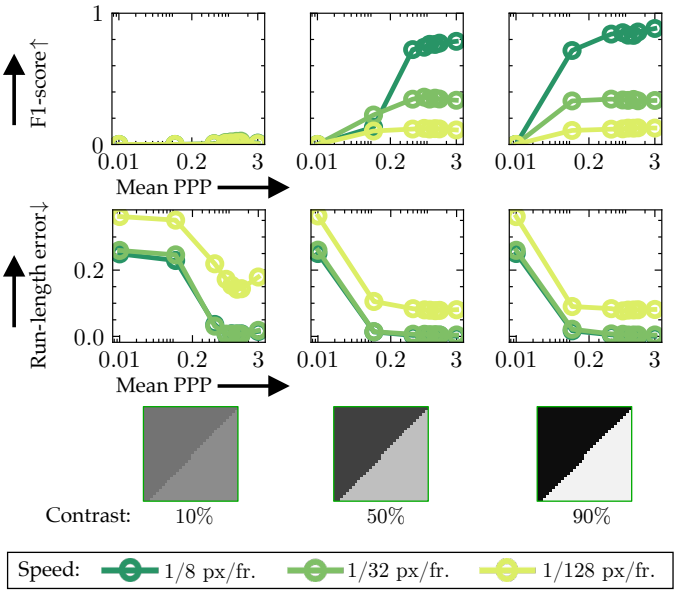


Fig. 14. **Change-point detection and run-length estimation of a moving step edge.** Top: F1-scores for change-point detection, with false positives comprising change-points flagged outside of 200 frame indices of the groundtruth change. Bottom: error in run-length estimates at the end of the photon-stream sequence (2048 binary measurements). In both rows, we see better performance for more abrupt (faster-moving edges) and large contrast changes.

reduces noise, it is also prone to motion blur and missed detections. Block size less than around $2/\text{edge\_speed}$ detects that edge velocity (and any edge that moves at a slower rate, albeit with a slightly sub-optimal efficacy).

With respect to flux levels, we see good performance from 0.1 photons per pixel per binary measurement (ppp), and up to 3 ppp. Very low light (~0.01 ppp) consistently fails. We also see failures with low edge-contrast ratios across light levels. Both these challenges are also observed with real data. We could improve the performance in these regimes by using multi-scale gradient extraction [55], or possibly by lowering the detection threshold (which would improve recall at the cost of precision).

7.2 Temporal Branch: Change Detection and Motion Metering

To analyze the temporal branch, we measure the performance of change-point detection and the error in (min filtered) run-length estimation by considering the same moving-edge input signal in Sec. 7.1. As a recap, change-points are estimated in a per-pixel manner using the estimated run lengths. Whereas, minimum-filtered run-length estimation outputs the minimum estimated run-length across overlapping patches of pixels (11×11 here). Run-length estimation error, which compares the estimated time since the last abrupt change to the actual abrupt change-point in the ground truth signal, serves as a proxy for the detection delay when processing a moving edge.

Fig. 14 shows the F1-error in change-point detection (median across 1024 trials) and run-length estimation error (mean across 128 trials). False positives for change detection comprise changes flagged outside a certain interval of the actual change: 200 frame indices here. For both change-point detection and run-length estimation, we see that faster moving edges result in better performance—which, while surprising, is a consequence of our run-length modeling that assumes incident flux to be a piece-wise constant function. Like Fig. 13, performance improves with contrast ratio. We also see good performance in similar flux ranges (~0.2–3 ppp) and failures in very low light (~0.05 ppp).

7.3 Image-quality and Edge-detection Evaluation

We compare motion metering on salience-guided sub-sampling to motion metering on uniformly sampled measurements (same sub-sampling factors for all pixels) using the Multi-Scale Structural Similarity (MS-SSIM [61], [62]) and the Feature Similarity (FSIM [63]) metrics. We provide a summary of our results here and refer the reader to Appendix F within the supplement for a detailed presentation.

Image-quality-wise, for the same number of measurements, we find that motion metering improves over box filtering, and salience-guided sampling improves over uniform sampling. However, perhaps counter-intuitively, we

find edge-detection performance to *not* be starkly different across the proposed methods (motion metering on salience-guided versus uniformly-sampled measurements). This conclusion appears to be a result of the edge detector (Structured Edges [1]) being resistant to motion blur and noise to some extent. So while the current design of the salience detector preserves neighborhoods around strong edges (which can aid image quality scores), the edges themselves are often detectable from uniformly sub-sampled data nearly as well. It would be interesting to repeat this experiment with salience detection that prioritizes faint gradients, which are indeed lost after uniform sub-sampling.

8 Discussion: Feasibility of Implementation

The proposed approach is designed for relatively high-resolution SPAD sensors ($> 10^5$ pixels) with sub-μsec. pixel addressing and control [11]. While significant progress has been made in recent years towards building near-sensor processing architectures for SPADs [36], support for these imaging specifications is still anticipated to be a few years away. That said, we expect it to emerge eventually, analogous to developments in smartphone camera processors to support ISP and control algorithms for conventional image sensors [64]. We justify this optimism through a more detailed analysis of the complexity of the different stages of the proposed adaptive imaging system.

Pixel-wise sub-sampling

Our salience-guided sub-sampling relies on dynamically turning on or off individual SPADs. A potential hardware implementation would disable SPADs by lowering the bias voltage at selected pixels, preventing photon avalanches and suppressing downstream digital activity [8], [34], [35]. The sub-sampling control signal could be realized by a configurable in-pixel division of the global clocked-recharge signal, which operates at the frame rate of the SPAD sensor.

Adaptive temporal integration; eventful readout

Pixel-wise temporal integration relies on the availability of near-sensor arithmetic processing units; very similar operations to the proposed motion-metering have been run previously on such hardware, including (emulation-based) experiments with sensor resolutions similar to the SwissSPAD2 used here [10], [37]. The same processing units also enable on-sensor compression of the kind shown in Sec. 6.3.

Spatial filtering

While spatial convolution-like processing has been recently demonstrated on SPAD-centered image processing architectures [37], the feasibility of extracting gradients at 100 kHz frame rates at high sensor resolutions ($\sim$ 1 MP) is still unclear. Fortunately, the computations of Sec. 5 occur at a much lower cadence, due to the tolerance of the salience detector to motion blur: as shown in Sec. 7.1, motion blur of 1 to 2 pixels over a temporal block is tolerated. For the experiments presented in Sec. 6 this works out a rate of around 100 to 300 Hz, well within the capabilities of current image processing units for smartphone cameras [64], with sophisticated on-chip inter-connection networks. Our proposed operations are separable in x and y, with a constant (signal-independent) computational complexity. We may additionally trade off performance for computational cost by opting for smaller kernels, *e.g.* 3-tap vs. 7-tap, or working at a lower resolution via binning.

GPU-based empirical complexity estimate

Functioning as a stand-in parallel computing architecture, runtimes on an Nvidia RTX 3090 Ti GPU are reported in Table 1, for processing a sequence of $8{,}192$ binary frames (~ 0.1 seconds of real time). With the caveat that an eventual on-sensor implementation may yet be significantly different, we find that at least on this GPU, the complexity is largely dominated by the temporal integration and change-point detection steps, as only they execute at the SPAD's binary capture rate. Specialized SPAD-centered processing architectures are likely to enable much more efficient execution of these operations [36], [37], potentially in real-time. However, this paper does not yet prove this claim; an important next step in this direction could involve execution on a software-based emulator of these architectures.

TABLE 1
GPU-based wall-clock run-times to process $8{,}192$ binary frames (resolution 496×254) on a NVIDIA RTX 3090 Ti GPU. Parameters: block size $M = 256$, 7-tap gradient kernels, dilation window size 11×11.

Operation	Time (secs.)
temporal box-filter (Eq. (3))	1.055
noise-equalized response update (Eqs. (4) and (5))	0.153
motion-metering update (Eq. (8)) & change-point detection (Appendix B)	3.739
spatial gradient (Farid and Simoncelli [49])	0.013
salience measure computation (Secs. 5.1 to 5.4) & sampling rate update (Eq. (9))	0.327
eventful encoding (Sec. 6.3)	0.483

9 Future Outlook: Design Landscape for Predictive (Passive) SPAD Sampling

We propose a possible computational approach to improving the energy-efficiency of SPAD cameras, based on allocating photon measurements preferentially where they are needed the most – in regions with strong spatial and temporal gradients, and hence reducing avalanche energy costs in non-salient regions. We demonstrate potential measurement reductions of $3-4\times$ across a variety of scenes, which include slow to rapid motion, low light to bright outdoor conditions, and high-dynamic range settings. Measurement allocations are particularly low (up to $10\times$ fewer) for scenes with small amounts of motion and sufficient light.

Higher-level guidance. So far, we have considered salience measures and allocation policies based on the very low-level gradient and change-point operators. However, it is also possible to design salience measures based on feature detectors and trackers like ORB [65], SIFT [2], or KLT [66]. In

learned systems, we can tailor salience guidance to downstream neural networks trained for reconstruction or even other downstream perception tasks, which is an exciting future direction. An important consideration here would be computational costs and latency. While salience estimation runs at the temporal-block granularity and has greater breathing room than operations running at the SPAD frame rate, it may not be feasible or energy-efficient to run computationally involved neural networks at this granularity.

In human-operated settings, user input may be used to further modulate the sample allocation (gradient-based or otherwise) to focus on specific regions, or alter method parameters such as $S_{\max.}$ and z_0 (Eq. (9) and Sec. 5.2, respectively), similar to manually adjusting exposure, focus, *etc.*, in current cameras.

Sampling rates driven directly by motion metering. The temporal model behind the motion-metered response and change-point detection currently only contributes explicitly towards the sampling rate via a binary mask (Sec. 5.3), with the rest dictated by the spatial gradient. It is likely possible to involve the temporal model to a larger degree, encouraging sparser sampling in regions with strong gradients which are changing slowly nevertheless—that is, adaptively sampling based on temporal statistics. Adaptive sampling (and bandit) algorithms have been considered in other settings involving sequential accumulation of information, such as Monte Carlo rendering [67], [68] and gating in single-photon LiDAR acquisition [69]. A related open question is of choosing an operating point between efficient sampling techniques and sub-optimal but less complex ones, relying instead on strong priors and denoising algorithms downstream to recover quality.

Explicitly minimizing photon detections. Sub-sampling explicitly minimizes measurements, but photon detections only indirectly. However, depending on the hardware implementation, the resource costs of a photon detection—which is directly associated with avalanche energy costs—may be more prominent than whether a measurement is made. In this case, reducing detections explicitly would be desirable, where, in addition to a salience measure agnostic to scene brightness, we incorporate a prediction of the avalanche energy that would be consumed to capture a given salient region. Some resource-efficiency metrics and "inhibition policies" have been proposed recently towards this purpose, for the static imaging case [30].

Spatial gradient as an additional data stream. In addition to the modulated photon stream and the event-encoded motion-metered response, the computed spatial gradients may be read out for downstream perception (feature extraction [55], [66]; optical flow estimation [70]). An event-based encoding may also be computed from the gradient as opposed to the intensity [71], followed by Poisson solvers to reconstruct the signal [72].

How few photons can we get by with? Energy savings depend on a variety of scene factors, including the illumination and motion levels; slow-moving HDR settings offer greater scope for measurement reductions than low-light scenes involving rapid motion. The end objective of measurement inhibition—whether intensity reconstruction or a computer-vision task—can also influence the balance between energy savings and performance. This energy-distortion tradeoff is a rich space to analyze for future work, including determining the theoretical limits as to how many photons are sufficient along various axes (motion, light, and end objective).

Acknowledgments

This research was supported in part by NSF CAREER award 1943149, ONR award N00014-24-1-2155, Wisconsin Alumni Research Foundation via a Research Forward Initiative Award, and the Swiss National Science Foundation grants 200021_166289 and 20QT21_187716. We thank Paul Mos for providing us access to SwissSPAD2 acquisition software, and the anonymous reviewers for their valuable feedback.

References

[1] P. Dollar and C. L. Zitnick, "Fast Edge Detection Using Structured Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/6975234/

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 3992–4003. [Online]. Available: https://ieeexplore.ieee.org/document/10378323/

[5] E. Charbon, C. Bruschini, and M.-J. Lee, "3D-Stacked CMOS SPAD Image Sensors: Technology and Applications," in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. Bordeaux: IEEE, Dec. 2018, pp. 1–4.

[6] N. A. W. Dutton, I. Gyongy, L. Parmesan, S. Gnecchi, N. Calder, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson, "A SPAD-Based QVGA Image Sensor for Single-Photon Counting and Quanta Imaging," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 189–196, Jan. 2016.

[7] K. Morimoto, J. Iwata, M. Shinohara, H. Sekine, A. Abdelghafar, H. Tsuchiya, Y. Kuroda, K. Tojima, W. Endo, Y. Maehashi, Y. Ota, T. Sasago, S. Maekawa, S. Hikosaka, T. Kanou, A. Kato, T. Tezuka, S. Yoshizaki, T. Ogawa, K. Uehira, A. Ehara, F. Inui, Y. Matsuno, K. Sakurai, and T. Ichikawa, "3.2 Megapixel 3D-Stacked Charge Focusing SPAD for Low-Light Imaging and Depth Sensing," in *67th Annual IEEE International Electron Devices Meeting*, Dec. 2021.

[8] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," *Optica*, vol. 7, no. 4, p. 346, Apr. 2020. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=optica-7-4-346

[9] S. Ma, P. Mos, E. Charbon, and M. Gupta, "Burst Vision Using Single-Photon Cameras," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[10] V. Sundar, M. Dutson, A. Ardelean, C. Bruschini, E. Charbon, and M. Gupta, "Generalized Event Cameras," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2024, pp. 25 007–25 017.

[11] T. Takatsuka, J. Ogi, Y. Ikeda, K. Hizu, Y. Inaoka, S. Sakama, I. Watanabe, T. Ishikawa, S. Shimada, J. Suzuki, H. Maeda, K. Toshima, Y. Nonaka, A. Yamamura, H. Ozawa, F. Koga, and Y. Oike, "A 3.36- μ m-Pitch SPAD Photon-Counting Image Sensor Using a Clustered Multi-Cycle Clocked Recharging Technique With an Intermediate Most-Significant-Bit Readout," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 4, pp. 1137–1145, Apr. 2024.

[12] A. C. Ulku, C. Bruschini, I. M. Antolovic, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512 × 512 SPAD Image Sensor With Integrated Gating for Widefield FLIM," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, Jan. 2019.

[13] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: Motion deblurring using fluttered shutter," in *ACM SIGGRAPH 2006 Papers*, ser. SIGGRAPH '06. New York, NY, USA: Association for Computing Machinery, Jul. 2006, pp. 795–804.

[14] J. Holloway, A. C. Sankaranarayanan, A. Veeraraghavan, and S. Tambe, "Flutter Shutter Video Camera for compressive sensing of videos," in *2012 IEEE International Conference on Computational Photography (ICCP)*, Apr. 2012, pp. 1–9.

[15] G. Wan, X. Li, G. Agranov, M. Levoy, and M. Horowitz, "CMOS Image Sensors With Multi-Bucket Pixels for Computational Photography," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 4, pp. 1031–1042, Apr. 2012.

[16] M. Wei, N. Sarhangnejad, Z. Xia, N. Gusev, N. Katic, R. Genov, and K. N. Kutulakos, "Coded Two-Bucket Cameras for Computer Vision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 54–71.

[17] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Flexible Voxels for Motion-Aware Videography," in *Computer Vision – ECCV 2010*, vol. 6311. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 100–114. [Online]. Available: http://link.springer.com/10.1007/978-3-642-15549-9_8

[18] J. Zhang, J. P. Newman, X. Wang, C. S. Thakur, J. Rattray, R. Etienne-Cummings, and M. A. Wilson, "A Closed-Loop, All-Electronic Pixel-Wise Adaptive Imaging System for High Dynamic Range Videography," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 6, pp. 1803–1814, Jun. 2020.

[19] C. Wang, J. Zhang, M. A. Wilson, and R. Etienne-Cummings, "Pix2HDR - A Pixel-Wise Acquisition and Deep Learning-Based Synthesis Approach for High-Speed HDR Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8771–8787, Dec. 2024.

[20] H. Nagahara, T. Sonoda, D. Liu, and J. Gu, "Space-Time-Brightness Sampling Using an Adaptive Pixel-Wise Coded Exposure," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1834–1842.

[21] B. Tilmon, E. Jain, S. Ferrari, and S. Koppal, "FoveaCam: A MEMS Mirror-Enabled Foveating Camera," in *2020 IEEE International Conference on Computational Photography (ICCP)*. Saint Louis, MO, USA: IEEE, Apr. 2020, pp. 1–11. [Online]. Available: https://ieeexplore.ieee.org/document/9105183/

[22] F. Faramarzi, B. Linares-Barranco, and T. Serrano-Gotarredona, "A 128×128 Electronically Multi-Foveated Dynamic Vision Sensor With Real-Time Resolution Reconfiguration," *IEEE Access*, vol. 12, pp. 192 656–192 671, 2024.

[23] B. Tilmon, Z. Sun, S. J. Koppal, Y. Wu, G. Evangelidis, R. Zahreddine, G. Krishnan, S. Ma, and J. Wang, "Energy-Efficient Adaptive 3D Sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5054–5063.

[24] F. Mattioli Della Rocca, H. Mai, S. W. Hutchings, T. A. Abbas, K. Buckbee, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "A 128 × 128 SPAD Motion-Triggered Time-of-Flight Image Sensor With In-Pixel Histogram and Column-Parallel Vision Processor," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1762–1775, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9097202/

[25] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, Jun. 2010, pp. 553–560.

[26] S. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 1. Hilton Head Island, SC, USA: IEEE Comput. Soc, 2000, pp. 472–479.

[27] S. Nayar and V. Branzoi, "Adaptive dynamic range imaging: Optical control of pixel exposures over space and time," in *Proceedings Ninth IEEE International Conference on Computer Vision*. Nice, France: IEEE, 2003, pp. 1168–1175 vol.2.

[28] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, May 2010, pp. 2426–2429.

[29] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-Based Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, Jan. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9138762/

[30] L. Koerner, S. Gupta, A. Ingle, and M. Gupta, "Photon Inhibition for Energy-Efficient Single-Photon Imaging," in *European Conference on Computer Vision (ECCV)*, 2024.

[31] A. Ingle, A. Velten, and M. Gupta, "High Flux Passive Imaging with Single-Photon Sensors," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6753–6762.

[32] A. Ingle, T. Seets, M. Buttafava, S. Gupta, A. Tosi, M. Gupta, and A. Velten, "Passive Inter-Photon Imaging," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 8581–8591.

[33] M. Wei, S. Nousias, R. Gulve, D. B. Lindell, and K. N. Kutulakos, "Passive Ultra-Wideband Single-Photon Imaging," in *International Conference on Computer Vision*, 2023.

[34] Y. Ota, K. Morimoto, T. Sasago, M. Shinohara, Y. Kuroda, W. Endo, Y. Maehashi, S. Maekawa, H. Tsuchiya, A. Abdelahafar, S. Hikosaka, M. Motoyama, K. Tojima, K. Uehira, J. Iwata, F. Inui, Y. Matsuno, K. Sakurai, and T. Ichikawa, "A 0.37W 143dB-Dynamic-Range 1Mpixel Backside-Illuminated Charge-Focusing SPAD Image Sensor with Pixel-Wise Exposure Control and Adaptive Clocked Recharging," in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*. San Francisco, CA, USA: IEEE, Feb. 2022, pp. 94–96.

[35] J. Ogi, T. Takatsuka, K. Hizu, Y. Inaoka, H. Zhu, Y. Tochigi, Y. Tashiro, F. Sano, Y. Murakawa, M. Nakamura, and Y. Oike, "7.5 A 250fps 124dB Dynamic-Range SPAD Image Sensor Stacked with Pixel-Parallel Photon Counter Employing Sub-Frame Extrapolating Architecture for Motion Artifact Suppression," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*. San Francisco, CA, USA: IEEE, Feb. 2021, pp. 113–115.

[36] A. Ardelean, "Computational Imaging SPAD Cameras," Ph.D. dissertation, EPFL, 2023.

[37] V. Sundar, A. Ardelean, T. Swedish, C. Bruschini, E. Charbon, and M. Gupta, "SoDaCam: Software-defined Cameras via Single-Photon Imaging," in *International Conference on Computer Vision (ICCV 2023)*, 2023.

[38] E. R. Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors," *IEEE Journal of the Electron Devices Society*, vol. 1, no. 9, pp. 166–174, 2013.

[39] J. Ma, D. Zhang, D. Robledo, L. Anzagira, and S. Masoodian, "Ultra-high-resolution quanta image sensor with reliable photon-number-resolving and high dynamic range capabilities," *Scientific Reports*, vol. 12, no. 1, p. 13869, Aug. 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-17952-z

[40] L. You, "Superconducting nanowire single-photon detectors for quantum information," *Nanophotonics*, vol. 9, no. 9, pp. 2673–2692, Jul. 2020. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/nanoph-2020-0186/html

[41] F. Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli, "Bits from Photons: Oversampled Image Acquisition Using Binary Poisson Statistics," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1421–1436, Apr. 2012. [Online]. Available: http://arxiv.org/abs/1106.0954

[42] I. M. Antolovic, C. Bruschini, and E. Charbon, "Dynamic range extension for photon counting arrays," *Optics Express*, vol. 26, no. 17, p. 22234, Aug. 2018.

[43] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics*, vol. 35, no. 6, Nov. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2980179.2980254

[44] S. Ma, S. Gupta, A. C. Ulku, C. Bruschini, E. Charbon, and M. Gupta, "Quanta burst photography," *ACM Transactions on Graphics*, vol. 39, no. 4, Jul. 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3386569.3392470

[45] T. Seets, A. Ingle, M. Laurenzis, and A. Velten, "Motion adaptive deblurring with single-photon cameras," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[46] P. Chennuri, Y. Chi, E. Jiang, G. M. D. Godaliyadda, A. Gnanasambandam, H. R. Sheikh, I. Gyongy, and S. H. Chan, "Quanta Video Restoration," in *Computer Vision – ECCV 2024*, vol. 15098. Cham: Springer Nature Switzerland, 2024, pp. 152–171.

[47] R. P. Adams and D. J. C. MacKay, "Bayesian Online Changepoint Detection," Oct. 2007. [Online]. Available: http://arxiv.org/abs/0710.3742

[48] R. Alami, O. Maillard, and R. Féraud, "Restarted Bayesian Online Change-point Detector achieves Optimal Detection Delay," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 2020, pp. 211–221. [Online]. Available: https://proceedings.mlr.press/v119/alami20a.html

[49] H. Farid and E. Simoncelli, "Differentiation of Discrete Multidimensional Signals," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 496–508, Apr. 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1284386/

[50] S. Gupta and M. Gupta, "Eulerian Single-Photon Vision," in *International Conference on Computer Vision (ICCV)*, 2023.

[51] A. Eckner, "Algorithms for unevenly spaced time series: Moving averages and other rolling operators. 2015," *URL http://eckner.com/papers/Algorithms% 20for% 20Unevenly% 20Spaced% 20Time% 20Series. pdf*, 2015.

[52] A. Gnanasambandam and S. H. Chan, "HDR Imaging with Quanta Image Sensors: Theoretical Limits and Optimal Reconstruction," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1571–1585, Nov. 2020.

[53] M. White, S. Ghajari, T. Zhang, A. Dave, A. Veeraraghavan, and A. Molnar, "A differential spad array architecture in 0.18 μm cmos for hdr imaging," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 292–296.

[54] The GIMP Development Team, "GNU image manipulation program (GIMP), version 3.0.0. community, free software (license GPLv3)," 2025, version 3.0.0, Free Software. [Online]. Available: https://gimp.org/

[55] T. Lindeberg, "Spatio-Temporal Scale Selection in Video Data," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 525–562, May 2018. [Online]. Available: http://link.springer.com/10.1007/s10851-017-0766-9

[56] L. Azzari and A. Foi, "Variance Stabilization for Noisy+Estimate Combination in Iterative Poisson Denoising," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1086–1090, Aug. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7491301/

[57] M. Tassano, J. Delon, and T. Veit, "FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 1351–1360. [Online]. Available: https://ieeexplore.ieee.org/document/9156652/

[58] Y. Liu, A. Krull, H. Basevi, A. Leonardis, and M. Jenkins, "Bit2bit: 1-bit quanta video reconstruction by self-supervised photon location prediction," in *Neural Information Processing Systems (NeurIPS)*, 2024.

[59] S. Jungerman, A. Ingle, and M. Gupta, "Panoramas from Photons," in *International Conference on Computer Vision (ICCV)*, 2023.

[60] R. Gomez-Merchan, J. A. Leñero-Bardallo, R. de la Rosa-Vidal, and Á. Rodríguez-Vázquez, "Dynamic vision with single photon detectors: A discrete dvs architecture using asynchronous sensor front-ends," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024.

[61] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Pacific Grove, CA, USA: IEEE, 2003, pp. 1398–1402.

[62] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1284395/

[63] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5705575/

[64] J. Redgrave, A. Meixner, N. Goulding-Hotta, A. Vasilyev, and O. Shacham, "Pixel Visual Core: Google's Fully Programmable Image, Vision, and AI Processor For Mobile Devices," Cupertino, California, USA, Aug. 2018.

[65] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*. Barcelona, Spain: IEEE, Nov. 2011, pp. 2564–2571. [Online]. Available: http://ieeexplore.ieee.org/document/6126544/

[66] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*. Seattle, WA, USA: IEEE Comput. Soc. Press, 1994, pp. 593–600.

[67] M. Zwicker, W. Jarosz, J. Lehtinen, B. Moon, R. Ramamoorthi, F. Rousselle, P. Sen, C. Soler, and S.-E. Yoon, "Recent Advances in Adaptive Sampling and Reconstruction for Monte Carlo Rendering," *Computer Graphics Forum*, vol. 34, no. 2, pp. 667–681, May 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/cgf.12592

[68] B. Bitterli, C. Wyman, M. Pharr, P. Shirley, A. Lefohn, and W. Jarosz, "Spatiotemporal reservoir resampling for real-time ray tracing with dynamic direct lighting," *ACM Transactions on Graphics*, vol. 39, no. 4, Aug. 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3386569.3392481

[69] R. Po, A. Pediredla, and I. Gkioulekas, "Adaptive Gating for Single-Photon 3D Imaging," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 16 333–16 342. [Online]. Available: https://ieeexplore.ieee.org/document/9878707/

[70] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, Jun. 2010, pp. 2432–2439.

[71] E. Lehtonen, T. Komulainen, A. Paasio, and M. Laiho, "Gradient events: Improved acquisition of visual information in event cameras," Jun. 2024. [Online]. Available: http://arxiv.org/abs/2409.01764

[72] J. Tumblin, A. Agrawal, and R. Raskar, "Why I Want a Gradient Camera," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. San Diego, CA, USA: IEEE, 2005, pp. 103–110. [Online]. Available: http://ieeexplore.ieee.org/document/1467255/

Shantanu Gupta (Student Member, IEEE) is a Ph.D. student at the University of Wisconsin-Madison. He received his bachelor's degree in the computer science department at the Indian Institute of Technology, Madras, in 2016. His research interests include computational photography, computer vision, and the design of resource-efficient imaging systems.

Varun Sundar is a graduate student at the University of Wisconsin–Madison, pursuing a Ph.D. in computer science. He previously received a bachelor's degree in electrical engineering from the Indian Institute of Technology, Madras, in 2020. His research interests broadly involve computer vision and computational photography, with his doctoral work focusing on efficient techniques for processing single-photon acquisition.

Lucas J. Koerner (Member, IEEE) received the B.A. (Hons.) degree in integrated science, physics, and mathematics from Northwestern University, Evanston, IL, USA, and the Ph.D. degree in physics from Cornell University, Ithaca, NY, USA. Since 2023, he has been an Associate Professor of Electrical and Computer Engineering with the University of St. Thomas, St. Paul, MN, USA. His research interests include electrical instrumentation, time-of-flight sensing, image sensors, and resource-efficient imaging systems.

Claudio Bruschini (Senior Member, IEEE) received the Laurea degree in physics from the University of Genoa, Genoa, Italy, in 1992, and the Ph.D. degree in applied sciences from Vrije Universiteit Brussel, Brussels, Belgium, in 2002. He is currently a Scientist, Lecturer and Lab Deputy with EPFL's Advanced Quantum Architecture Laboratory. His scientific interests have spanned from high energy physics and parallel computing in the early days to challenging sensor applications in humanitarian demining, concentrating since 2003, on quantum photonic devices, high-speed and time-resolved 2-D/3-D optical sensing, as well as applications thereof (biophotonics, nuclear medicine, basic sciences, security, and ranging).

Edoardo Charbon (Fellow, IEEE) received the Diploma degree from ETH Zurich, Zurich, Switzerland, in 1988, the M.S. degree from the University of California at San Diego, La Jolla, CA, USA, in 1991, and the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 1995, all in electrical engineering and computer sciences. Since 2002, he has been a member of the faculty of EPFL, where he is a Full Professor. From 2008 to 2016, he was with Delft University of Technology as the Chair of VLSI design. He has been the driving force behind the creation of deep-submicrometer CMOS SPAD technology, which is mass-produced since 2015 and is present in telemeters, proximity sensors, and medical diagnostics tools. His interests span from 3-D vision, LiDAR, FLIM, FCS, and NIROT to super-resolution microscopy, time-resolved Raman spectroscopy, and cryo-CMOS circuits and systems for quantum computing. Dr. Charbon was a recipient of the 2023 IISS Pioneering Achievement Award, a Distinguished Visiting Scholar of the W. M. Keck Institute for Space at Caltech, a Fellow of the Kavli Institute of Nanoscience Delft, and a Distinguished Lecturer of the IEEE Photonics Society.

Mohit Gupta is an Associate Professor of Computer Sciences at the University of Wisconsin-Madison. He received Ph.D. from the Robotics Institute, Carnegie Mellon University, and was a postdoctoral research scientist at Columbia University. He directs the WISION Lab with research interests broadly in computer vision and computational imaging. He has received best paper honorable mention awards at computer vision and photography conferences in 2014 and 2019 including a Marr Prize honorable mention at IEEE ICCV, a best demo award at SIGGRAPH ETech 2024, multiple Sony Faculty Innovation Awards, and an NSF CAREER award.