Burst Vision Using Single-Photon Cameras

Sizhuo Ma[1] sizhuoma@cs.wisc.edu
Paul Mos [2] paul.mos@epfl.ch
Edoardo Charbon [2] edoardo.charbon@epfl.ch
Mohit Gupta [1] mohitg@cs.wisc.edu

[1]University of Wisconsin-Madison, USA
[2]École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

Single-photon avalanche diodes (SPADs) are novel image sensors that record the arrival of individual photons at extremely high temporal resolution. In the past, they were only available as single pixels or small-format arrays, for various active imaging applications such as LiDAR and microscopy. Recently, high-resolution SPAD arrays up to 3.2 megapixel have been realized, which for the first time may be able to capture sufficient spatial details for general computer vision tasks, purely as a passive sensor. However, existing vision algorithms are not directly applicable on the binary data captured by SPADs. In this paper, we propose developing quanta vision algorithms based on burst processing for extracting scene information from SPAD photon streams. With extensive real-world data, we demonstrate that current SPAD arrays, along with burst processing as an example plug-and-play algorithm, are capable of a wide range of downstream vision tasks in extremely challenging imaging conditions including fast motion, low light (< 5 lux) and high dynamic range. To our knowledge, this is the first attempt to demonstrate the capabilities of SPAD sensors for a wide gamut of real-world computer vision tasks including object detection, pose estimation, SLAM, and text recognition. We hope this work will inspire future research into developing computer vision algorithms in extreme scenarios using single-photon cameras.

1. Quanta Computer Vision

Single-photon avalanche diodes (SPADs) are a novel sensor technology that promises high sensitivity and the ability to time-tag photons with picosecond precision. Despite (or perhaps due to) these capabilities, most current SPAD-based imaging systems are limited to being *active* where the sensor needs to be precisely synchronized with an active light source, *e.g*. laser. These active imaging systems are driving many new applications, including non-line-of-sight (NLOS) imaging [4, 44, 5], microscopy [3, 24], and LiDAR [22, 48, 26, 45, 34] – the last being especially pertinent to computer vision. Indeed, SPADs are fast emerging as the sensor of choice for automotive LiDAR due to their high time-resolution [29] and lower costs.

Are SPADs ready for passive computer vision? Till a few years ago, SPAD sensors were available as single-pixel or small arrays such as 32×32. While they showed promising imaging performance in lab settings, these sensors were largely restricted to mechanical scanning [27] and relatively simple scenes [1]. Fortunately, the resolution of SPAD arrays and other capabilities (signal-to-noise-ratio, dead-time) have improved rapidly [49, 41] (Fig. 1(a)), reaching up to 3.2 MPixel [42] last year, which for the first time allows capturing high-frequency spatial details needed for general computer vision applications *in-the-wild*, including object detection, text recognition, SLAM, *etc*. Furthermore, due to their compatibility with mainstream CMOS fabrication, the cost continues to decrease. In light of these recent advancements, this paper explores the following questions: Can we expand the scope of SPADs as a purely passive general-purpose sensor with a broader range of computer vision applications beyond LiDAR (Fig. 1(c))? If so, what are their benefits over conventional cameras?

SPAD arrays are capable of capturing binary frames at high frame rates reaching 100kfps. Each pixel receives a series of 0s and 1s, or a *photon stream*. To fully exploit the imaging capabilities of single-photon image sensors, we must design novel *quanta vision* algorithms (Fig. 1(b)) that are optimized for data captured by SPADs as well as other single-photon image sensors such as jots [16].

Quanta burst vision: It is challenging to extract meaningful scene information from an individual binary frame (only 1-bit dynamic range). A quanta vision algorithm must combine information from multiple frames. One such approach is to simply average multiple binary frames to emulate a multi-bit camera, which can be achieved by on-chip counters on existing SPAD [42] and jots sensors [36]. This approach, while easy to implement, runs into the fundamental imaging trade-off between blur and noise, as shown in Fig. 2 (Top). For dynamic scenes in low-light conditions, the naively averaged images are either too noisy or too blurred, thereby preventing accurate object detection.

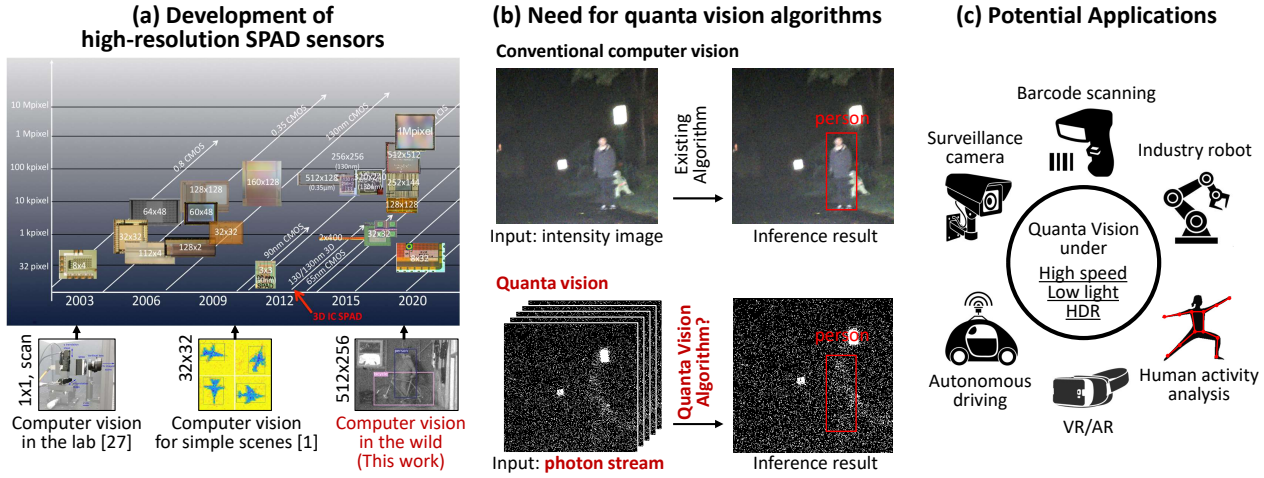

Figure 1. **Quanta vision. (a)** High-resolution SPAD sensors have been developed in recent years, which paves the way for computer vision in-the-wild. Figure reproduced from [10] with permission. **(b)** However, conventional vision algorithms do not directly work on photon stream captured by SPADs. We propose developing quanta vision algorithms based on burst processing of SPAD photon streams. **(c)** Quanta vision enables sensing and perception in high-speed, low-light or HDR environments, with a wide range of potential applications.

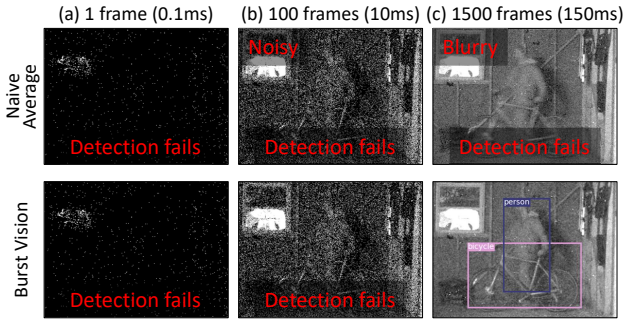

Figure 2. **Resolving blur-noise trade-off using burst vision.** A binary sequence captured in a dark garage (night, lights off). Naive averaging and burst reconstruction [37] are used to reconstruct an intensity image, which is then passed to pre-trained YOLOv3 for object detection. **(Top)** Naive average images are either too noisy or too blurred. Consequently, object detection fails for all integration window lengths. **(Bottom)** Burst vision is able to generate clear images that provide sufficient signal for detection of the person and the bike with sufficiently large integration windows.

It is possible to mitigate this blur/noise trade-off by compensating for the motion in a burst of binary images, potentially at the single-photon level, before extracting scene information for downstream tasks. We call this approach *quanta burst vision*, inspired by burst photography for conventional cameras. We show that, with existing burst image reconstruction techniques [37], a high-SNR and low-blur intensity image can be reconstructed from a binary sequence, which provides sufficient information for various computer vision tasks (Fig. 2 (Bottom)). Moreover, by allowing *software-defined* overlapping integration windows, it is possible to achieve high-SNR reconstruction at video frame rates for video-based tasks, even in extremely dark environments. The same reconstruction technique can be used as a plug-and-play module for a wide range of tasks, without needing to redesign, re-implement or retrain the downstream algorithms.

Quanta vision in the real world. To demonstrate that recently-developed SPADs can be deployed in real-world vision applications, we capture binary sequences for various ($\sim$10) downstream tasks in different lighting conditions, scene/camera motions at various speeds, including both indoor and outdoor environments (Fig. 3). The dataset consists of over 50 million binary frames, captured at frame rates ranging from 10kfps to 96.8kfps. We also capture synchronized sequences using commercial cameras including DSLR, cellphone, night vision camera and thermal camera. By applying quanta vision with state-of-the-art traditional and learning-based computer vision algorithms, we qualitatively show that it is possible to perform a wide gamut of computer vision tasks in challenging conditions, including extremely low light and rapid motion. We also provide an evaluation protocol based on *temporally-sampled annotation*: Human annotators label the burst reconstructed images at discrete timestamps, which provide a manageable way to quantitatively evaluate quanta vision algorithms. We will release this dataset and annotations, which can serve as a test set for burst reconstruction and facilitate the development of future quanta vision algorithms.

Scope and Contributions: The paper aims to draw attention to the potential of SPADs for general computer vision and the need for developing quanta vision algorithms. We propose burst vision as the canonical processing paradigm for quanta vision. We use an existing burst image reconstruction algorithm as an example plug-and-play module for several downstream image or video inference tasks. We capture the first-of-its-kind large-scale quanta vision

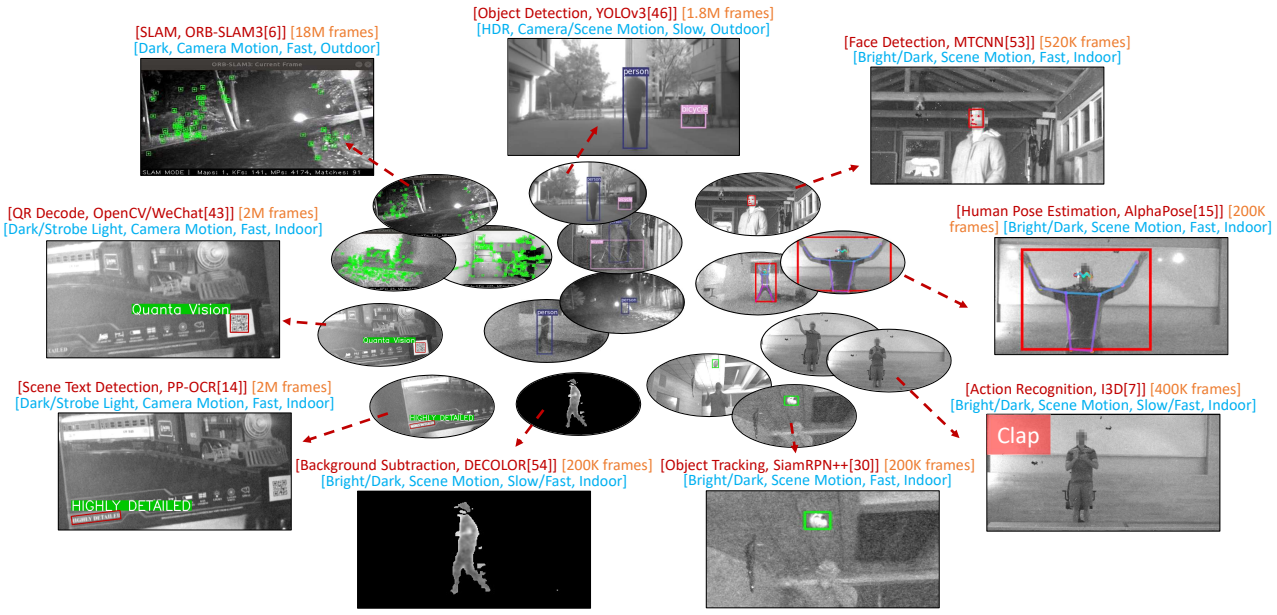

Figure 3. **Quanta vision tasks and sequences.** We capture binary sequences for a wide range of tasks, consisting of over 50 million binary images in total. Tested tasks and algorithms including: QR decoding [43], scene text detection [14], object detection [46], SLAM [6], face detection [53], human pose estimation [15], action recognition [7], background subtraction [54], object tracking [30]. For each task, one representative sequence is highlighted with details provided. Faces are blurred for anonymity.

dataset, and use it to demonstrate the performance of burst vision in-the-wild for a wide range of downstream tasks and challenging imaging conditions (Fig. 3) where conventional cameras are often inadequate.

The paper does not aim to develop novel burst reconstruction algorithms. Instead, the paper merely uses the proposed quanta burst vision approach, with an existing burst reconstruction algorithm, to bootstrap the exploration of quanta vision, which will hopefully spur further research in this young but promising sub-field of computer vision.

2. Related Work

Single-photon image sensors. Single-photon avalanche diodes (SPAD) achieve sensitivity to single photons via avalanche multiplication. Recently, large-format SPAD arrays with extremely high frame rates and negligible read noise have been realized [49, 41], which makes them ideal for low-light high-speed computer vision. Quanta image sensors (QIS) based on jots are another type of single-photon image sensors, which currently achieve smaller pixel pitch and higher quantum efficiency but have a lower temporal resolution than SPADs [16, 36]. Although we focus on SPADs in this paper, the principle of burst vision benefits computer vision tasks with QIS as well.

Computer vision with passive single-photon image sensors. Previous work focuses on image classification on naive average of binary frames with few detected photons by condensing knowledge from clean images [19, 21]. Such photon-limited images exist mostly because a longer exposure cannot be used due to scene/camera motion. Their performance can therefore be further improved if burst vision techniques are applied. [11] proposes to apply computer vision algorithms directly on a stream of photons before a clean intensity image is formed. [31] uses a non-local neural network to combine features from a burst of photon-limited images. Both methods can be seen as *implicit* approaches to burst vision (see Sec. 6).

Computer vision with active single-photon image sensors. SPADs have been broadly used for LiDAR, which measures scene depth and has applications in various computer vision tasks. [1] proposes an event-based neuromorphic system that directly performs object recognition from detected photons. [40] directly trains a neural network for object detection from pixelwise photon timing histograms. In this paper, we focus on using SPADs as passive sensors.

3. Quanta Burst Vision

3.1. Why Do We Need Burst Vision?

Most existing SPADs for active sensing are event-driven, which can accurately record the timestamp of photon arrival events but are limited to low resolution due to the complexity of time-to-digital converter (TDC). In this paper, we focus on the clock-driven design proposed in [49], which enables higher resolution and can capture not only passive images but also active data with time gating. In such design, a single SPAD pixel records a binary value B which is 1 if one or more photons hit the sensor during a fixed exposure

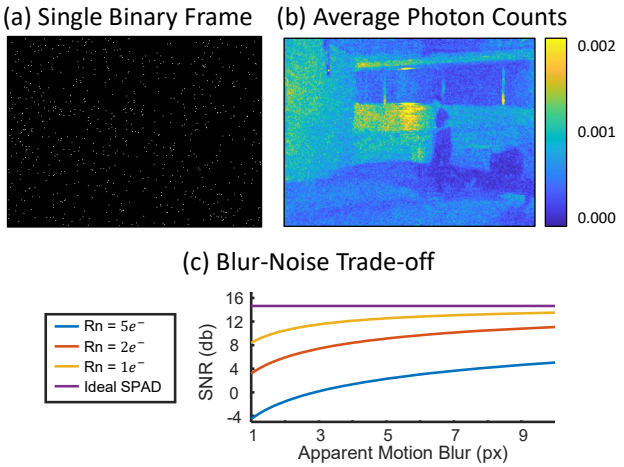

Figure 4. **Why do we need burst vision? (a)** In low light, a single binary frame contains a sparse set of detected photons. **(b)** Average photon counts are down to 0.002 per pixel. Therefore, it is necessary to apply burst vision to combine information from a large number of frames. **(c)** Conventional burst photography introduces read noise at each frame, resulting in a blur-SNR trade-off. A lower read noise gives a higher blur-SNR curve. SPAD sensors have negligible read noise, which results in an ideal flat curve for low-light burst vision. (Assuming 300 photons/second·pixel, 0.1s exposure, 100 pixel apparent motion, and SPAD having the same quantum efficiency as conventional CMOS and zero read noise.)

time, or 0 otherwise. Mathematically,

$$P\{B=0\} = e^{-\phi}, \qquad P\{B=1\} \quad = 1 - e^{-\phi}, \quad (1)$$

where ϕ is the photon flux (photons/exposure) incident on the pixel[1]. A 2D array of SPADs captures a 2D binary image $B(x, y)$. Due to the randomness of photons, a binary image is extremely noisy, making it challenging for conventional computer vision algorithms to extract meaningful information. The problem is exacerbated in extremely dark environments, as shown in Fig. 4(a,b): Only sparse photons are detected, and it seems impossible to recognize the scene by looking at the image. Statistics over a sequence of binary frames show that the photon flux goes down to 0.002 photons per pixel. For such extreme conditions, it is necessary to combine information from multiple frames for downstream computer vision tasks.

Existing SPAD [42] or jots [36] cameras simply add (or average) binary frames captured over time into multi-bit images, by using on-chip counters. In such naive averaging mode of operation, single-photon cameras can be considered similar to conventional CMOS cameras, but with lower noise. Although this naive averaging approach can capture high-quality images in the dark for static scenes, as discussed in Fig. 2, it is subject to the blur-noise trade-off, which limits the performance in dark, dynamic scenes.

[1]Here we omit the discussion of other parameters including exposure time, dark count, *etc*. A complete formula can be found in [2, 37].

Our key observation, inspired by burst photography for conventional cameras [35, 25, 20, 39, 32], is that it is possible to align and merge the binary frames into a single multi-bit image with reduced motion blur, which is then used as input to downstream tasks. By using burst reconstruction as a plug-and-play module, existing algorithms for a wide range of downstream applications can be directly applied without retraining or redesigning new algorithms.

Why are SPADs uniquely suited for burst vision in low light? It is important to notice that conventional cameras do not always benefit from burst photography, as each frame suffers from a fixed amount of read noise. Fig. 4(c) shows the effects of read noise on low-light burst photography. During a fixed budget of total exposure time, a larger number of frames results in shorter exposure per frame and therefore less blur in the merged image (assuming the frames are perfectly aligned by the algorithm). However, since each additional frame introduces a fixed amount of read noise, the SNR of the final image is also reduced. We can analytically compute the SNR of a single white patch as photon flux / (photon noise + read noise) as in [37, 27]. Intensity error due to blur is not considered so that noise and blur is disentangled. By connecting points corresponding to different number of frames, a blur-SNR curve can be plotted. Notice that different curves are plotted for different amount of read noise: Sensors with lower read noise are less affected by the trade-off, and thus have a more flattened curve. Although modern CMOS sensors have low read noise ($\sim 1e^-$ or input-referred noise of one electron for recent smartphones [13]), in extreme low-light, read noise becomes significant as each frame only has a very small number of photons. Furthermore, read noise for conventional high-speed cameras is often considerably higher.[2]

In contrast, although SPADs have a small number of spurious photon detections *per unit time* known as dark count, SPADs do not require analog to digital conversion (ADC), and thus have no/minimal *per-frame read noise*, even at high capture speeds. This absence of read noise makes SPADs the ideal sensor for low-light burst vision, with a completely flat blur-SNR curve. A quantitative comparison on downstream tasks performance can be found in Sec. 5.

3.2. Plug-and-Play Burst Reconstruction

What is the right burst reconstruction algorithm for binary frames captured by SPADs? Fig. 5 shows example sequences with naive average as a baseline. Results are too noisy when averaging over a short sequence, and blurred when averaging over a long sequence.

Conventional burst denoising. Traditional burst denoising approaches are inadequate for quanta images because each

[2]For example, Phantom v1840 has a read noise of $7.2e^-$: `https://www.phantomhighspeed.com/products/cameras/ultrahighspeed/v1840`

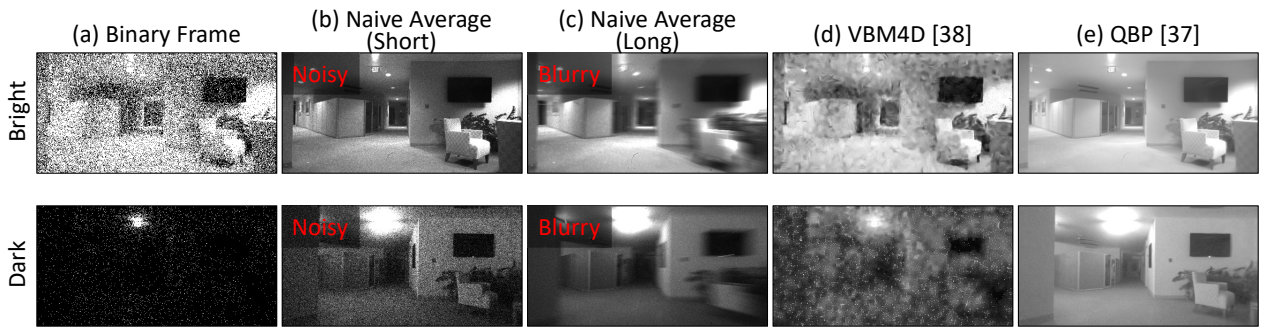

Figure 5. **Methods for burst reconstruction. (a)** Binary images of an extremely dark scene (0.03 photons per pixel on average). Images look bright due to 200X contrast stretching; the bright spots are very dim exit signs. **(b)** Naive average over a short sequence gives a noisy image. **(c)** Naive average over a long sequence gives a blurred image. **(d)** Conventional burst denoising method (VBM4D) creates blocky artifacts, especially for dark images. **(e)** Burst reconstruction for quanta images (QBP) is able to generate a high-SNR low-blur image.

quanta image is extremely noisy and does not contain sufficient information to compute the motion and merge them robustly. Fig. 5(d) shows the result of applying a state-of-the-art video denoising algorithm (VBM4D [38]) on a binary sequence, which contains heavy blocky artifacts.

Burst reconstruction of binary images. Early reconstruction algorithms for single-photon images focus on statistical formulation [51, 17] with smoothness priors [9, 8] on a static scene. Several approaches have then been proposed to reconstruct an intensity images from a burst of quanta images. [23, 28, 47] estimates the motion from the temporal statistics of binary images, assuming the scene consists of rigid objects. [12] trains a neural network to reconstruct an image from a small number of photon-limited multi-bit QIS images. In this paper, we consider quanta burst photography (QBP) [37] as a plug-and-play module for quanta vision because QBP makes a less restrictive motion assumption (patch-wise translation) and works for binary sequence of arbitrary length. Fig. 5(e) shows that QBP-reconstructed images have a higher visual quality than naive averaging and burst denoising. In Sec. 5 we demonstrate that such higher visual quality leads to better downstream inference performance than naive averaging and burst denoising.

3.3. Video Inference using Burst Vision

For video inference tasks, it is possible to shift the integration window in time to reconstruct frames at different time instants. Suppose the binary frames are captured at a frame rate f. The *integration window* t_w is the number of frames used to create a single intensity frame with burst reconstruction. Ideally, t_w should be chosen to be large enough so that a high-SNR intensity frame can be obtained. In practice, an extremely long integration window will not further improve the result as the field-of-view of the later frames may not overlap with the earlier frames. We also define *inference period* t_p, which is the amount of shift between neighboring integration windows. Inference period is application-dependent, and is determined from how

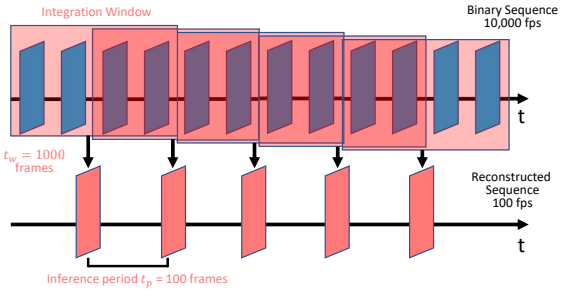

Figure 6. **Burst vision for video inference.** The integration window size and the reconstruction period are independent. Here an intensity image is reconstructed for every 100 binary frames while combining information from 1000 neighboring frames.

frequently the inference results are needed for the task and how much computation is affordable. In practice, an effective minimum inference period exists since motion between even shorter periods becomes too small (a small fraction of a pixel) to be precisely estimated.

One interesting observation is that the choice of inference period t_p and integration window size t_w are independent. This decoupling enables the notion of *software-defined exposure times*: The effective exposure time is given by the integration window, which is completely defined by the post-capture software processing. Furthermore, it is possible to choose an exposure time much longer than the inference period (effective frame rate), which allows *overlapping exposure*. Fig. 6 shows an example where $t_p = 100$ (0.01s at 10kfps capture rate) and $\delta_U + \delta_L = 1000$ (0.1s). This enables reconstruction of high-quality images at high frame rates for low light vision, which is not possible with naive averaging. A quantitative analysis on the effect of exposure time is shown in Sec. 5.

4. Quanta Vision in the Real World

To demonstrate the capability of quanta vision in real-world scenarios, we create a diverse SPAD image dataset

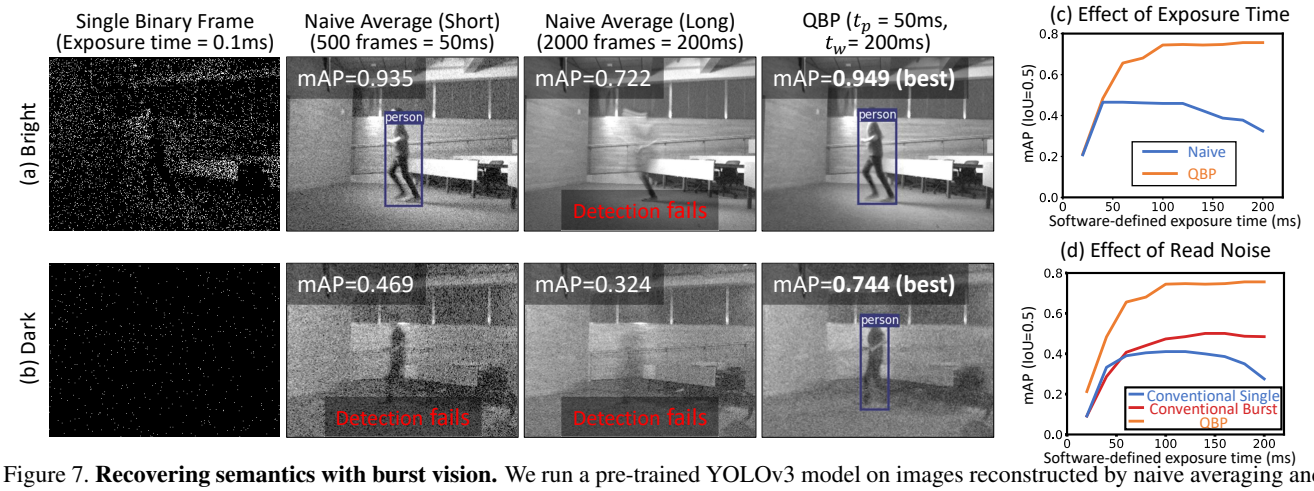

Figure 7. **Recovering semantics with burst vision.** We run a pre-trained YOLOv3 model on images reconstructed by naive averaging and QBP. **(a)** In a bright scene, both naive average (short) and QBP generate quality images for person detection. **35X contrast stretched. (b)** In a dark scene, naive average suffers from the blur-noise trade-off. QBP is able to reconstruct a clear image for detection. **370X contrast stretched**. **(c)** Naive averaging always performs worse than QBP and has a best operating point due to the blur-noise trade-off. QBP keeps improving as software-defined exposure time increases. **(d)** We simulate images that would have been captured by a conventional camera with read noise of 1 photon. Conventional single shot shows similar performance as naive average due to the blur-noise trade-off. Conventional burst photography has improved accuracy when a longer exposure time is used, but is significantly lower due to read noise.

(Fig. 3) consisting of 47 scenes, many of which are captured under different lighting conditions, resulting in a total of 137 binary sequences. Sequence length ranges from 20 seconds to 15 minutes. The sequences include both indoor and outdoor scenes, with both scene motion and camera motion, ranging from slow handheld motion to fast car driving. Ambient illumination varies from 0.02 lux in a dark room to 50,000 lux directly under sunlight. All the sequences are captured using SwissSPAD2 [49], which captures binary frames at a resolution of 512×256, with frame rates between 10kfps and 96.8kfps. This comprehensive dataset demonstrates the potential of using SPADs for real-world applications, and can be used as a test-bed for future burst reconstruction algorithms.

Evaluating quanta vision algorithms. For quantitative evaluation, it is important to provide ground truth for the collected dataset. However, it is extremely challenging to annotate a quanta image due to its low SNR. We propose evaluating quanta vision algorithms using *temporally-sampled annotations*. We use QBP to reconstruct intensity images at fixed inference period (*e.g.* every 200 binary frames), which gives high-quality images for human annotators to work on. An algorithm has the freedom to return results at any time instants, and the error is computed by comparing the most recent result with the ground truth available at fixed timestamps. We evaluate the algorithms by the mean errors across the timestamps because an algorithm that works well at *sampled* timestamps also is more likely to perform well at other time instants, which enables manageable quanta vision evaluation.

5. Experimental Results

Recovering semantics with burst vision. Fig. 7 demonstrates how scene semantics can be recovered with burst vision. We capture sequences with the same scene motion (running person) under two different light levels. We reconstruct intensity images using naive averaging and QBP, and then run YOLOv3 [46] pre-trained on COCO [33] to detect the person. Both methods succeed in the bright setting. In dim lighting, detection fails for both naive averaging using a short window (50ms) and a long window (200ms) due to the blur-noise trade-off. QBP achieves an inference period of 50ms while using a integration window of 200ms via overlapping exposure, and successfully detects the person. We annotate the bounding box using the proposed temporally-sampled annotation scheme and quantitatively evaluate if the methods can detect the person stably across the sequence. QBP achieves best mean average precision (mAP, at IoU=0.5) among the three methods, as shown in Fig. 7(a, b). **For better visualization of the results in this paper, please refer to the supplementary video.**

Performance analysis. We further study the performance of SPAD as a function of software-defined exposure time. Fig. 7(c) plots the mAP for both methods on the dark sequence. Naive averaging always performs worse than QBP and has a best operating point due to the blur-noise trade-off, while QBP keeps improving and flattens at long exposure times. Therefore, it is always possible to use a long exposure time for QBP for better detection precision. Notice that the computation complexity also increases with the exposure time. In practice, the shortest software-defined ex-

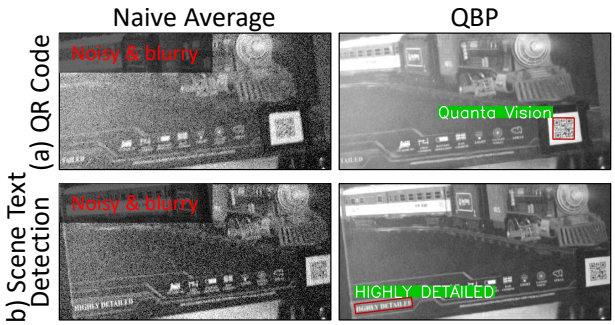

Figure 8. **Recovering high-frequency spatial details.** A package box moves fast in the dark. Naive average images are noisy and blurry. Burst results ($t_p = 50ms, t_w = 200ms$) are clear and sharp and the text and QR code are correctly detected, which would be challenging if a low-resolution SPAD was used.

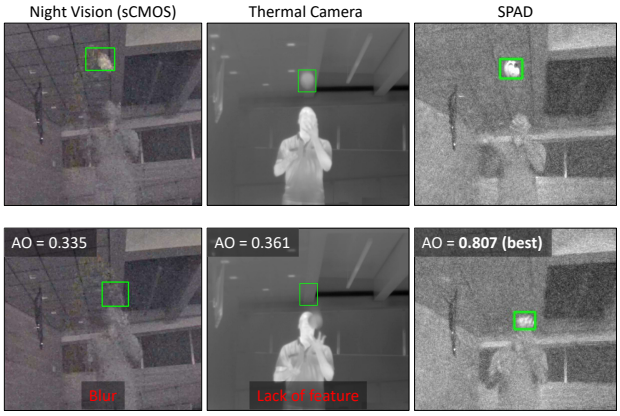

Figure 9. **Fast object tracking in extremely low light.** (t_0) We manually mark the bounding box of the ball at the beginning of the sequence, where the ball starts falling at a low speed. (t_1) In a later frame of the sequence, the ball drops at a higher speed. The night vision camera records a blurred image. The thermal camera does not capture the visual features of the ball. The SPAD camera ($t_p = 10ms, t_w = 200ms$) gets a clear, sharp burst-reconstructed image for tracking the ball and achieves the best AO among the three cameras. **150X contrast stretched**.

posure time needed to achieve a given precision at current light level should be used. We analyze the benefit of SPAD as a function of light level in Fig. 11.

Quanta burst vision vs. conventional burst photography. To evaluate the effect of read noise on burst photography, we simulate conventional images by adding read noise (mean=0, std=1e^-) to linear images reconstructed from SPAD data (Fig. 7(d)).[3] A burst of 10 frames with 20ms exposure each are generated, with the same total exposure time as the other two cases. Same burst reconstruction algorithm [37] is applied to conventional and quanta data. Similar to naive averaging of binary frames, the mAP of conventional single image increases at first and then decreases due to the blur-noise trade-off. The mAP of conventional burst photography keeps increasing as a longer software-defined exposure time is used. However, due to the per-frame read noise, the overall precision is much lower than QBP.

Recovering high-frequency spatial details. In addition to high-level semantics, many vision tasks also involve recovery of high-frequency spatial details, which is impossible with a low-resolution SPAD array and is very sensitive to noise and blur. Fig. 8 shows a sequence where the goal is to detect text and a qr code on a fast-moving box in a dark room. Naive averaging gives a noisy and blurry image, while burst reconstruction results in a clear, sharp image, with the text and the code correctly detected by existing QR decoder [43] and scene text detection algorithms [14].

How does quanta vision compare to other low-light imaging sensors? Fig. 9 shows a ball juggling scene in extremely low light where we use SiamRPN++ [30] to track the ball. We show both the initial frame (t_0) and a later frame (t_1) where the ball drops at a high speed. Synchronized images are shown from a night vision camera (Bosch DINION IP starlight 7000 HD), which is optimized for low light and widely used as surveillance cameras. However, the image contains heavy motion blur when the ball moves fast. A thermal camera (HT 301) detects IR and gets a clear image even in this dark environment, but it does not capture features in visible light. The SPAD reconstructed images are sharp and preserve the visual features on the ball, which is tracked successfully. The three sequences are annotated separately and the average overlap (AO) is reported.

Notice that this is not meant to be a direct comparison between camera technologies as they have different sensor sizes, exposure times and optics. Instead, through this example we show that there are extremely challenging imaging conditions where current imaging technologies still struggle at, while SPADs may provide a potential solution.

Quanta vision in the wild. Fig. 10 demonstrates quanta vision capabilities in challenging outdoor scenes. We fix the SPAD camera and the night vision camera side-by-side above the dashboard of a car. In the object detection task, the SPAD can detect a pedestrian from a long distance beyond the reach of the headlights. In the SLAM task, we run ORB-SLAM3 [6] on reconstructed images offline. The SPAD can track the camera motion and build a map despite fast driving in the night. The night vision camera fails in both cases. Please see the supplementary video for the video sequence, including a SLAM reconstruction.

Quanta vision across wide dynamic range. In addition to dark environments, Fig. 11 shows that SPAD cameras can also perform computer vision tasks across a wide dynamic range. The sequence starts from a dark room down to 1 lux, and gets directly exposed to sunlight at the end (50000

[3] Assuming SPAD and conventional CMOS pixels have comparable fill factor and quantum efficiency, which has been achieved recently [42].

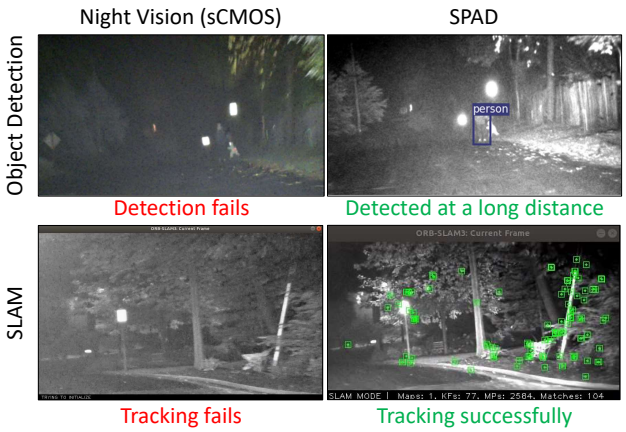

Figure 10. **Quanta vision in the wild.** We placed a SPAD camera and a night vision camera side-by-side on the dashboard of a car. **(a)** SPAD ($t_p = 50ms, t_w = 500ms$) gives a clearer image so the person can be detected from a long distance. The bright spots are traffic signs illuminated by the headlight (this street has no streetlights). The scene is very dark ($\sim$ 1 lux, 0.05 photons per pixel on average). Images are contrast stretched (**15X**) for visualization. **(b)** SPAD ($t_p = 50ms, t_w = 500ms$) is able to reconstruct a high-quality image to enable SLAM tracking for nighttime driving while the night vision camera fails to detect features.

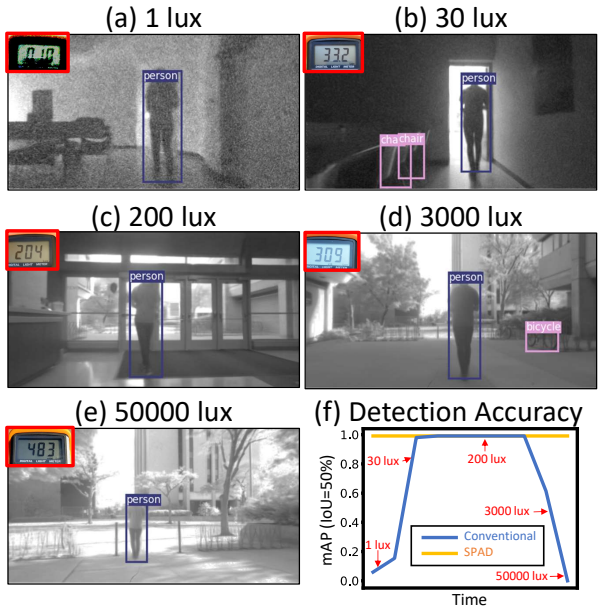

Figure 11. **Quanta vision across wide dynamic range. (a-e)** A SPAD camera captures a person walking from a dark room (1 lux) to directly under sunlight (50000 lux). Burst method ($t_p = 50ms, t_w = 2s$) reconstructs HDR images, enabling object detection across the sequence. **(f)** The person is detected for almost every frame reconstructed by SPAD. The simulated conventional camera fails to reconstruct good images for dark and bright conditions, resulting in failure of the person detection algorithm.

lux). We also simulate conventional burst photography results following the same process as in Fig. 7 assuming 10-bit ADC and $1e^-$ read noise. Conventional camera fails for extremely dark and bright scenarios, while SPAD is able to reconstruct high-quality images across the entire spectrum of lighting conditions, and the person is detected in almost every reconstructed frame. For a fair comparison, the exposure time and lens aperture is kept constant for both cameras. In the future, we expect SPAD cameras to adapt to a wider range of lighting conditions with auto-exposure and exposure bracketing [18] implemented.

Please refer to the supplementary technical report for more results that are not included due to the page limit.

6. Discussion and Future Outlook

Implicit burst vision. It is also possible to directly take multiple binary frames as input [11] and combine the information *implicitly*, which we call *implicit burst vision*. Similar ideas have been explored in multi-frame and video neural networks for conventional cameras [50, 55], but the goal is usually to combine semantic information across frames and has not focused on improving inference results with bursts of photon-limited frames until recently [31].

The challenge for implicit burst vision is that algorithms need to be redesigned and training data needs to be collected for each task, which is challenging as high-resolution SPAD cameras are not yet accessible to everyone. On the other hand, implicit burst vision can be optimized for each individual task for better performance. In this paper, we focus on explicitly applying burst reconstruction to quanta images for downstream tasks (*explicit burst vision*). We hope this will bootstrap the exploration of quanta vision and spur further research on both explicit and implicit burst vision.

On the future of SPAD image sensors for vision applications. One limitation of burst reconstruction such as QBP is that they require more computing resources and higher bandwidth for transferring the binary frames, as compared to on-chip naive averaging and only transferring intensity images for downstream vision. Therefore, it is critical to couple SPADs to powerful computational capabilities *in situ*, so as to take advantage of the natively digital nature of SPADs and the massive parallelism intrinsic to large pixel arrays. Fast *in situ* functionality has been embedded in SPAD image sensors [52], mostly to compress time-of-flight data into event-driven packets of preprocessed partial histograms. A higher level of computing sophistication including advanced data interpretation, ideally through machine learning and neuromorphic computing, is possible through 3D stacking, which enables combination of older technologies that host SPADs in the top tier and advanced technology nodes to host processing architectures.

Acknowledgement. This research was supported by NSF CAREER Award 1943149, Intel-NSF award CNS-2003129, and Swiss National Science Foundation grant 20QT21_187716.

References

[1] Saeed Afshar, Tara Julia Hamilton, Langdon Davis, Andre Van Schaik, and Dennis Delic. Event-Based Processing of Single Photon Avalanche Diode Sensors. *IEEE Sensors Journal*, 20(14):7677–7691, July 2020.

[2] Ivan Michel Antolovic, Samuel Burri, Claudio Bruschini, Ron Hoebe, and Edoardo Charbon. Nonuniformity Analysis of a 65-kpixel CMOS SPAD Imager. *IEEE Transactions on Electron Devices*, 63(1):57–64, Jan. 2016.

[3] Claudio Bruschini, Harald Homulle, Ivan Michel Antolovic, Samuel Burri, and Edoardo Charbon. Single-photon avalanche diode imagers in biophotonics: Review and outlook. *Light: Science & Applications*, 8(1):87, Dec. 2019.

[4] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics Express*, 23(16):20997, Aug. 2015.

[5] Clara Callenberg, Zheng Shi, Felix Heide, and Matthias B. Hullin. Low-cost SPAD sensing for non-line-of-sight tracking, material classification and depth imaging. *ACM Transactions on Graphics*, 40(4):1–12, Aug. 2021.

[6] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. *IEEE Transactions on Robotics*, pages 1–17, 2021.

[7] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[8] Stanley Chan, Omar Elgendy, and Xiran Wang. Images from Bits: Non-Iterative Image Reconstruction for Quanta Image Sensors. *Sensors*, 16(11):1961, Nov. 2016.

[9] Stanley H. Chan and Yue M. Lu. Efficient image reconstruction for gigapixel quantum image sensors. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 312–316, Atlanta, GA, USA, Dec. 2014. IEEE.

[10] Edoardo Charbon. Lidar and consumer photography: applications with a common denominator. Image Sensor Europe, 2021.

[11] Bo Chen and Pietro Perona. Vision without the Image. *Sensors*, 16(4):484, Apr. 2016.

[12] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H. Chan. Dynamic low-light imaging with quanta image sensors. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 122–138, Cham, 2020. Springer International Publishing.

[13] William J. Claff. Input-referred Read Noise versus ISO Setting. https://www.photonstophotos.net/Charts/RN_e.htm, 2020.

[14] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A Practical Ultra Lightweight OCR System. *arXiv:2009.09941 [cs]*, Oct. 2020.

[15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, Venice, Oct. 2017. IEEE.

[16] Eric R. Fossum. What To Do With Sub-Diffraction Limit (SDL) Pixels?—A proposal for a gigapixel digital film sensor (DFS). In *IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors*, pages 214–217, 2005.

[17] Eric R. Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The Quanta Image Sensor: Every Photon Counts. *Sensors*, 16(8):1260, Aug. 2016.

[18] Abhiram Gnanasambandam and Stanley H. Chan. HDR Imaging with Quanta Image Sensors: Theoretical Limits and Optimal Reconstruction. *IEEE Transactions on Computational Imaging*, 6:1571–1585, 2020.

[19] Abhiram Gnanasambandam and Stanley H. Chan. Image Classification in the Dark using Quanta Image Sensors. In *European Conference on Computer Vision (ECCV)*, pages 484–501. Springer, 2020.

[20] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep Burst Denoising. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision (ECCV)*, pages 538–554, Cham, 2018. Springer International Publishing.

[21] Bhavya Goyal and Mohit Gupta. Photon-Starved Scene Inference using Single Photon Cameras. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2512–2521, 2021.

[22] Anant Gupta, Atul Ingle, and Mohit Gupta. Asynchronous Single-Photon 3D Imaging. In *International Conference on Computer Vision (ICCV)*, pages 7909–7918, 2019.

[23] Istvan Gyongy, Neale Dutton, and Robert Henderson. Single-Photon Tracking for High-Speed Vision. *Sensors*, 18(2):323, Jan. 2018.

[24] Istvan Gyongy, Andrew Green, Sam W. Hutchings, Amy Davies, Neale Dutton, Rory Duncan, Colin Rickman, Robert K. Henderson, and Paul Dalgarno. Fluorescence lifetime imaging of high-speed particles with single-photon image sensors. In Keisuke Goda and Kevin K. Tsia, editors, *High-Speed Biomedical Imaging and Spectroscopy IV*, page 24, San Francisco, United States, Mar. 2019. SPIE.

[25] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):1–12, Nov. 2016.

[26] Felix Heide, Steven Diamond, David B. Lindell, and Gordon Wetzstein. Sub-picosecond photon-efficient 3D imaging using single-photon sensors. *Scientific Reports*, 8(1):1–8, Dec. 2018.

[27] Atul Ingle, Andreas Velten, and Mohit Gupta. High Flux Passive Imaging with Single-Photon Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6760–6769, 2019.

[28] Kiyotaka Iwabuchi, Yusuke Kameda, and Takayuki Hamamoto. Image Quality Improvements Based on Motion-Based Deblurring for Single-Photon Imaging. *IEEE Access*, 9:30080–30094, 2021.

[29] Oichi Kumagai, Junichi Ohmachi, Masao Matsumura, Shinichiro Yagi, Kenichi Tayu, Keitaro Amagawa, Tomohiro Matsukawa, Osamu Ozawa, Daisuke Hirono, Yasuhiro Shinozuka, Ryutaro Homma, Kumiko Mahara, Toshio Ohyama, Yousuke Morita, Shohei Shimada, Takahisa Ueno, Akira Matsumoto, Yusuke Otake, Toshifumi Wakano, and Takashi Izawa. A 189×600 Back-Illuminated Stacked SPAD Direct Time-of-Flight Depth Sensor for Automotive LiDAR Systems. In *IEEE International Solid- State Circuits Conference (ISSCC)*, pages 110–112, San Francisco, CA, USA, Feb. 2021. IEEE.

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4277–4286, Long Beach, CA, USA, June 2019. IEEE.

[31] Chengxi Li, Xiangyu Qu, Abhiram Gnanasambandam, Omar A Elgendy, Jiaju Ma, and Stanley H Chan. Photon-Limited Object Detection Using Non-Local Feature Matching and Knowledge Distillation. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3976–3987, 2021.

[32] Orly Liba, Ryan Geiss, Samuel W. Hasinoff, Yael Pritch, Marc Levoy, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T. Barron, and Dillon Sharlet. Handheld mobile photography in very low light. *ACM Transactions on Graphics*, 38(6):1–16, Nov. 2019.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, Cham, 2014. Springer International Publishing.

[34] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. Single-photon 3D imaging with deep sensor fusion. *ACM Transactions on Graphics*, 37(4):1–12, Aug. 2018.

[35] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics*, 33(6):1–9, Nov. 2014.

[36] Jiaju Ma, Saleh Masoodian, Dakota A. Starkey, and Eric R. Fossum. Photon-number-resolving megapixel image sensor at room temperature without avalanche gain. *Optica*, 4(12):1474, Dec. 2017.

[37] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics*, 39(4):1–16, July 2020.

[38] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video Denoising, Deblocking, and Enhancement Through Separable 4-D Nonlocal Spatiotemporal Transforms. *IEEE Transactions on Image Processing*, 21(9):3952–3966, Sept. 2012.

[39] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst Denoising with Kernel Prediction Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2510, Salt Lake City, UT, June 2018. IEEE.

[40] Germán Mora-Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, Jonathan Leach, and Istvan Gyongy. High-speed object detection with a single-photon time-of-flight image sensor. *arXiv:2107.13407 [physics]*, July 2021.

[41] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications. *Optica*, 7(4):346–354, Apr. 2020.

[42] 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 *IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2021.

[43] OpenCV. Wechat qr code detector for detecting and parsing qr code, 2021. https://github.com/opencv/opencv_contrib/.

[44] Matthew O'Toole, David B. Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, Mar. 2018.

[45] Joshua Rapp and Vivek K. Goyal. A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging. *IEEE Transactions on Computational Imaging*, 3(3):445–459, Sept. 2017.

[46] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018.

[47] Trevor Seets, Atul Ingle, Martin Laurenzis, and Andreas Velten. Motion Adaptive Deblurring with Single-Photon Cameras. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1945–1954, 2021.

[48] Zhanghao Sun, David B. Lindell, Olav Solgaard, and Gordon Wetzstein. SPADnet: Deep RGB-SPAD sensor fusion assisted by monocular depth estimation. *Optics Express*, 28(10):14948, May 2020.

[49] Arin Can Ulku, Claudio Bruschini, Ivan Michel Antolovic, Yung Kuo, Rinat Ankri, Shimon Weiss, Xavier Michalet, and Edoardo Charbon. A 512 × 512 SPAD Image Sensor With Integrated Gating for Widefield FLIM. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–12, Jan. 2019.

[50] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully Motion-Aware Network for Video Object Detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision (ECCV)*, volume 11217, pages 557–573, Cham, 2018. Springer International Publishing.

[51] Feng Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli. Bits From Photons: Oversampled Image Acquisition Using Binary Poisson Statistics. *IEEE Transactions on Image Processing*, 21(4):1421–1436, Apr. 2012.

[52] Chao Zhang, Scott Lindner, Ivan Michel Antolovic, Juan Mata Pavia, Martin Wolf, and Edoardo Charbon. A 30-frames/s, 252×144 SPAD Flash LiDAR With 1728 Dual-

Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming. *IEEE Journal of Solid-State Circuits*, 54(4):1137–1151, 2018.

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct. 2016.

[54] Xiaowei Zhou, Can Yang, and Weichuan Yu. Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):597–610, Mar. 2013.

[55] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-Guided Feature Aggregation for Video Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 408–417, 2017.