Quanta Video Restoration

Prateek Chennuri[1], Yiheng Chi[1], Enze Jiang[1], G.M. Dilshan Godaliyadda[2],
Abhiram Gnanasambandam[2], Hamid R. Sheikh[2],
Istvan Gyongy[3], and Stanley H. Chan[1]

[1] Purdue University
[2] Samsung Research America
[3] University of Edinburgh
{pchennur,chi14,jiang708,stanchan}@purdue.edu
{dilshan.g,abhiram.g,hr.sheikh}@samsung.com
{igyongy2}@exseed.ed.ac.uk

(a) Ground Truth 16-bit   (b) CMOS 16-bit, 60 fps at 1 lux   (c) CMOS 16-bit, 240 fps at 1 lux   (d) CMOS 16-bit, 2000 fps at 1 lux

(e) Quanta 3-bit, 2000 fps at 1 lux   (f) Quanta 3-bit, reconstructed from 11-frame avg. using Traditional Methods [47]   (g) Quanta 3-bit, reconstructed from 66-frame avg. using Traditional Methods [47]   (h) Quanta 3-bit, reconstructed from 11-frame avg using **QUIVER**
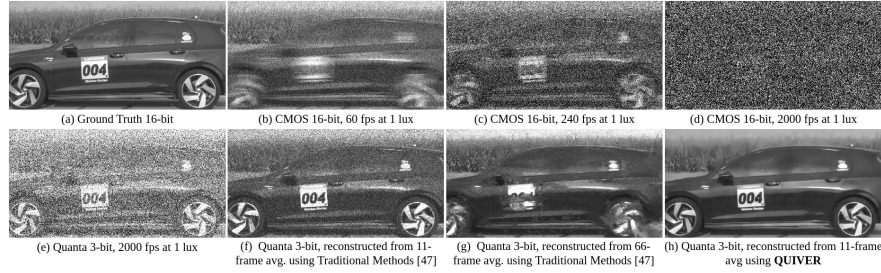
Fig. 1: Goal of this paper. (a) Blur-free video frame of a moving car. (b)-(d) CMOS image sensor simulations using realistic sensor parameters. The strong shot noise and read noise ($5.1 \, e^-$/pix) of CMOS sensor make the signal acquisition difficult. (e) With low read noise ($0.2 \, e^-$/pix), low-bit single-photon detectors capture valuable information. (f)-(g) Existing state-of-the-art algorithm, QBP [47] cannot handle strong motion and noise. (h) The proposed algorithm, QUIVER, produces high quality results.

Abstract. The proliferation of single-photon image sensors has opened the door to a plethora of high-speed and low-light imaging applications. However, data collected by these sensors are often 1-bit or few-bit, and corrupted by noise and strong motion. Conventional video restoration methods are not designed to handle this situation, while specialized quanta burst algorithms have limited performance when the number of input frames is low. In this paper, we introduce Quanta Video Restoration (QUIVER), an end-to-end trainable network built on the core ideas of classical quanta restoration methods, i.e., pre-filtering, flow estimation, fusion, and refinement. We also collect and publish I2-2000FPS, a high-speed video dataset with the highest temporal resolution of 2000 frames-per-second, for training and testing. On simulated and real data, QUIVER outperforms existing quanta restoration methods by a significant margin. Code and dataset available at `https://github.com/chennuriprateek/Quanta_Video_Restoration-QUIVER-`

Keywords: Single Photon Detectors · Video Restoration · High-Speed Dataset

1   Introduction

Over the past decade, the astonishing growth of single-photon detectors has fundamentally changed the landscape of computational imaging. With the invention and proliferation of quanta image sensors (QIS) [21] and single-photon avalanche diodes (SPAD) [52, 62], there is an unprecedented volume of new applications in low-light imaging [6, 10, 66], computer vision [27, 31, 40], high-speed videography [47, 48], time-of-flight sensing [32, 63], and 3D imaging [30, 44]. In most of these use cases, the main core question that lies is how to recover the image from the photon counts measured in the scene. Specifically, given a video stream of 1-bit or few-bit data captured from a scene involving moving objects, how do we reconstruct a gray-scale image/video while eliminating the noise without incurring motion blur?

To give the reader a visual perspective of the problem scope, Fig. 1 depicts a blur-free video of a moving car. We simulate the captured images at 1 lux assuming 60 fps, 240 fps, and 2000 fps CMOS image sensors with realistic sensor specifications. As illustrated in the figure, the resulting CMOS outputs are either severely blurred due to strong motion or completely distorted by noise due to sparse photons. In the same figure, we demonstrate a simulated single-photon camera output (a 3-bit QIS in this case) where the content the largely preserved despite heavy noise. Upon utilizing state-of-the-art Quanta Burst Photography (QBP) [47] for reconstructing the frames, provided the motion is slow, a decent output can be obtained. However, as the temporal window narrows down, as shown in Fig. 1 (f), the noise remains. In this paper, we address this problem with a new algorithm, designed to remove the noise while avoiding distortions in the presence of fast motion while utilizing only a few frames.

The core innovation of this paper is QUanta VIdeo REstoration (QUIVER), a deep-learning based video restoration algorithm for quanta image data. QUIVER is specialized for few-bit data (3-bit) captured at thousands of frames-per-second (2000 fps) with an average motion range of 1 to 7 pixels per frame. The main contributions of this paper can be summarized as follows.

- We propose QUIVER, an end-to-end trainable quanta (video) restoration method built by embracing the core ideas from traditional quanta restoration algorithms. On a comprehensive evaluation dataset containing both simulated and real data, QUIVER outperforms all methods we compared in this paper by a significant margin.

- We introduce I2-2000FPS, the first high-speed video dataset with a temporal resolution of 2000 frames-per-second for training and testing image and video reconstruction neural networks. We captured a total of 280 high-speed videos covering 114 distinct scenes with ground truth and simulated 3-bit videos.
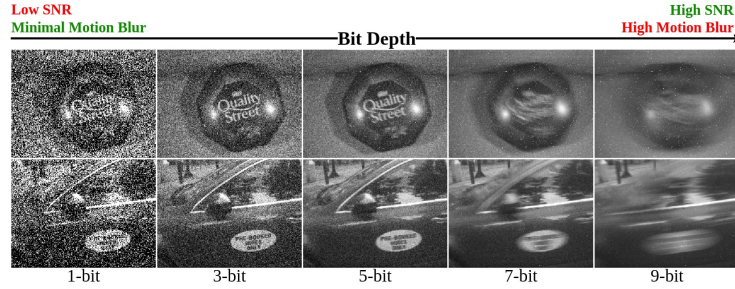
Fig. 2: Motion Blur and SNR Trade-off. The effects of bit depth on SNR and motion blur are illustrated using *real* captures by a single-photon sensor. For the motion range we target, 3-bit single-photon detectors provide the best trade-off between blur and SNR. The images are captured using a 1-bit SPAD [17] at 10k fps at an average photon level of 0.51 and 0.40 photons-per-pixel (PPP) per frame, respectively. Higher bit-depth outputs are generated through temporal frame averaging.

Table 1: Frame-rate, motion, read-noise, and data-rate statistics for various bit-depths at the same exposure level.

Bit-Depth	fps	Motion (pixels/frame)	σ_{read} (/pixel/sec)	Data-rate (Mb/sec)
1	10k	$0-1$	$2000\ \mathrm{e}^-$	96
3	1428	$2-3$	$285.6\ \mathrm{e}^-$	41.13
5	323	$6-12$	$64.6\ \mathrm{e}^-$	15.5
7	78	$25-30$	$15.6\ \mathrm{e}^-$	5.24
9	20	$70-80$	$4\ \mathrm{e}^-$	1.73

2   Background

2.1   Few-bit Single-Photon Detectors

What is few-bit photon counting? Single-photon detectors (QIS and SPAD) differ from conventional CMOS pixels by their extraordinary photon counting capability. QIS uses a two-stage pump-gate technique and correlated double sampling to suppress the read noise, while SPAD uses avalanche multiplication to amplify the photocharge. In both cases, the sensors are capable of resolving photons up to a single-photon sensitivity. We refer readers interested in the sensor development of QIS and SPAD to consult, for example, [5, 9, 18, 45].

Along with the single-photon detectors' unique capability to count individual photons, these devices can generate data at a bit-depth as low as 1-bit to as high as 16-bit or even more. However, higher bit-depth is accompanied by longer integration time. If the scene contains motion, longer integration time will eventually result in strong motion blurs as shown in Fig. 2. On the other hand, 1-bit sensing with high frame rates will result in motion-blur-free but extremely noisy images. Therefore, from a pure data acquisition perspective, there exists an optimal bit-depth with respect to the motion that will give us mini-

mal/no motion-blur data with a minimum per-frame signal-to-noise ratio (SNR) required for good quality reconstruction.

How about 1-bit and reconstruct afterward? Readers familiar with single-photon counting may wonder whether we can collect as many 1-bit frames as possible and then process the data afterward. The problem is power consumption and data rate. Fixing the same level of exposure, as described in Tab. 1, a 1-bit video at 10k fps would require 96 Mb/sec whereas a 9-bit video at 20 fps would only need 1.73 Mb/sec. Another problem is read noise accumulation. For sensors with non-zero read noise (such as QIS), every frame contributes to a finite amount of read noise. The more frames we read, the more read noise we accumulate. Therefore, recording 1-bit is not always the best option.

2.2   Related Work

Image and Video Denoising. Classical state-of-the-art methods utilize a non-local strategy to identify similar patches across an image/video [3, 38, 50]. Deep neural networks have been proven to be successful in producing high quality denoised outputs [13, 69–71]. Among these architectures, Vision Transformers [41, 42] have been rated the state-of-the-art in recent times. However, all these solutions make simplistic assumptions on noise statistics, thus failing to perform on real noisy images or videos [55].

Coming to low light, Burst denoising [34, 43], where images are aligned, merged and denoised, is one of the most popular methods. However, these methods fail without robust alignment. To overcome this, a number of alternative solutions with learnable alignment modules have been proposed [29, 54, 73]. Recent solutions have focused on practical noise models that replicate real camera sensor noise, to produce visually appealing results [51, 74]. Nevertheless, the existing solutions utilize images captured using CMOS image sensors, resulting in a notably higher photon level compared to the one utilized in our study.

SPADs, Event, and Spike Cameras. Gariepy *et al.* [24, 25] firstly introduced the utilization of Single Photon Avalanche Diodes (SPADs) at pico-second temporal resolution to capture light in motion. Gyongy *et al.* [33] demonstrated 2D motion tracking of rigid planar objects using SPADs at 10k fps. Recently, Ma *et al.* [47] and Seets *et al.* [64] utilized SPADs in a passive imaging setting to capture motion in low illumination. However, all these methods utilize extremely high-temporal resolutions hindering the deployment of these sensors into consumer devices where bandwidth is the bottleneck. Event [53, 60] and Spike Cameras [84, 86] also have demonstrated their effectiveness in capturing high-speed motion. However, these cameras focus on luminance/brightness variation and record a spike only when variation is above a threshold (changes based on factors like temperature, event rate, etc.) [15, 60]. Therefore, unlike single-photon detectors (QIS and SPADs), these cameras are NOT designed for single-photon counting and cannot operate in extreme low-light conditions.

QIS Reconstruction. Reconstructing quanta images is a challenging task due to the underlying Poisson-Gaussian statistics. Initial solutions to this problem included utilizing standard gradient descent [78], greedy algorithms [79],
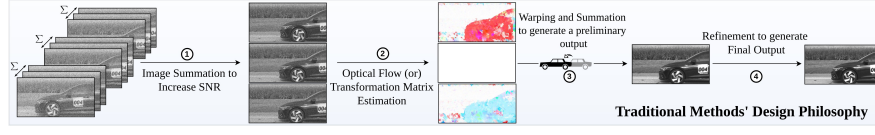
Fig. 3: Traditional Methods' Design. Depiction of existing classical quanta restoration algorithms' design philosophy. *Best viewed in zoom.*

(ADMM) [7, 8], among others [19, 20, 23, 26, 61, 77]. Chan *et al.* [6] were the first to propose a non-iterative approach using Anscombe transform for reconstructing quanta images. Choi *et al.* [12] proposed the first end-to-end trainable deep neural network (DNN) for QIS reconstruction. Alternative DNN-based solutions include utilizing vision transformers [75], Dual Prior Integrated networks [80], among others [81]. Nonetheless, all these methods fail to produce good results when the scene is in motion. Chi *et al.* [10] is the only method which focuses on capturing dynamic scenes using QIS but only targets at slow motion (1 pixel/frame).

3 QUanta VIdeo REstoration (QUIVER)

3.1 Design Philosophy

In this section we present the design of our proposed algorithm. We start by briefly reviewing the design of classical methods [33, 47] which, to an extent, have been successful in restoring quanta images. As shown in Fig. 3, classical methods' algorithm design can be divided into four stages: (1) computing sum images to increase SNR, (2) optical flow (or) transformation matrix estimation for aligning the input frames, (3) warping and linear combination for generating preliminary restored output, and (4) refinement for producing the final output. While the steps seem intuitive and straightforward, existing methods are heavily vulnerable to extreme noise and strong motion in the input frames primarily due to two reasons. (1) none of the stages are designed to handle extreme noise and strong motion simultaneously (will be discussed further). (2) Since all the stages are sequential yet independent of each other, it is difficult to obtain an optimal result for a wide range of noise and motion. Our proposed algorithm QUIVER, leverages the design philosophy of existing classical methods while designing each stage to simultaneously handle both noise and motion. Moreover, QUIVER is an end-to-end trainable model making all the stages inter-dependent, thus leading to good restoration outputs.

3.2 Design of QUIVER

QUIVER is a deep-learning-based video restoration method for quanta imaging. The design philosophy of QUIVER is to adopt the intuitive thoughts behind
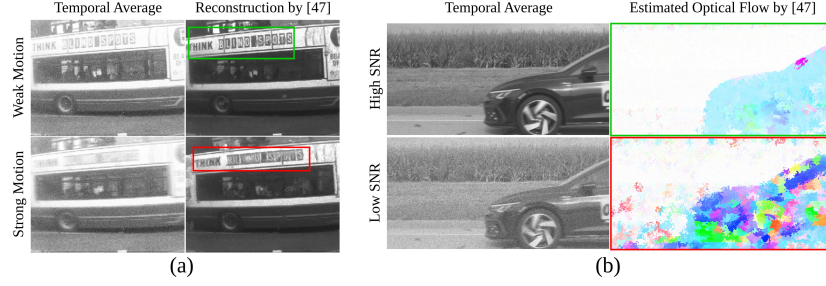
Fig. 4: Traditional Methods' Limitations. (a) Traditional methods' [33, 47] predenoising/temporal-averaging fails in strong motion. It is visible in the restored images that an input with strong motion between the frames results in several artifacts in the output even though SNR levels are similar. (b) Traditional methods [47] utilize a patch-based pre-trained optical flow module similar to [34]. This optical flow module fails to compensate for motion in the presence of noise.
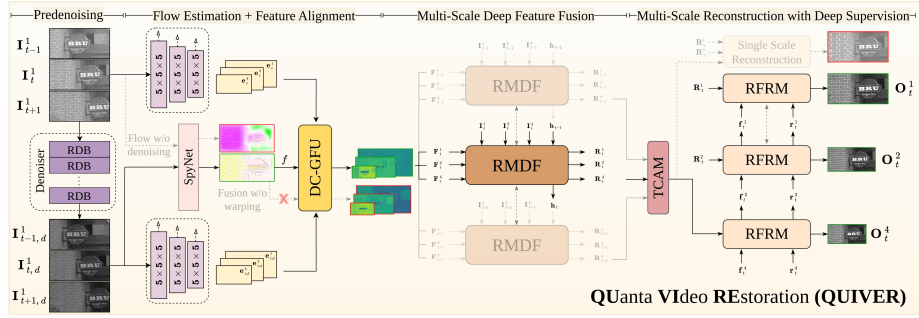

Fig. 5: The proposed QUIVER network. The corresponding stages of QUIVER, built by embracing the intuitive thoughts behind existing classical methods. *Best viewed in zoom.*

classical quanta restoration methods and develop an end-to-end trainable, robust to noise and motion deep learning based framework, as shown in Fig. 5. Specifically, QUIVER can be divided into four main stages:

Pre-Denoising to improve SNR: Since the input quanta frames possess extreme noise, classical methods adopt naive averaging to increase the SNR and thereby predict better optical flows or transformation matrices. However, as shown in Fig. 4(a), the simple averaging is vulnerable to motion and will negatively impact subsequent processing, ultimately leading to distorted outputs. Simply eliminating this stage is not the solution, because it leads to poor optical flow estimation, resulting in over-smoothed outputs with lack of low-level intricate details, as shown in Fig. 5 and Fig. 11. Therefore, a preliminary denoising step robust to noise and motion is crucial. In QUIVER, we leverage a computational undemanding single image denoiser built using Residual Dense Blocks [83] (RDBs) to provide minimal pre-preprocessing of the input quanta
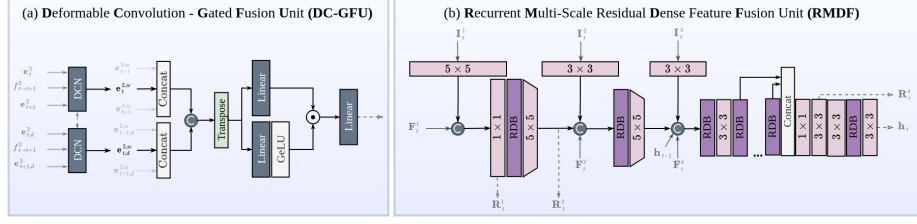
Fig. 6: Design of the proposed modules DC-GFU and RMDF.

data. A multi-frame denoiser is not an option due to it computationally demanding nature. We use RDB based network due to its history in handling noise while preserving details with its simple yet effective design.

Optical Flow Estimation + Feature Alignment using Deformable Convolution - Gated Fusion Unit: Classical methods utilize an off-the-shelf pre-trained optical flow estimation module or predict a transformation matrix to compensate for motion between the frames. The basic assumption behind such approaches is that the motion between the frames is limited and the SNR is high enough. However, when such assumption is not met, the motion compensation is sub-optimal, as shown in Fig. 4(b). As most state-of-the-art pre-trained optical flow estimators are optimized on the CMOS RGB sensor images, it leads to sub-optimal performance when applied on quanta frames. Eliminating the flow estimation step is not recommended since experiments reveal the critical role it plays in motion compensation, as shown in Fig. 5 and Fig. 11. For QUIVER, we deploy a learnable optical flow estimation module and utilize SpyNet [57] owing to its computational efficiency while using a multi-scale approach.

We deploy 3D convolution blocks to extract multi-scale spatio-temporal features from both the noisy and denoised quanta frames. We reuse the noisy frames to compensate for any information lost in the pre-denoising stage.

The estimated multi-scale robust-to-noise optical flows are utilised for feature-level alignment of the extracted multi-scale spatio-temporal features. We utilize the deformable convolution with residual offsets proposed by [4] to warp the features. Inspired by the superior performance of Gated Linear Units (GLUs) in Transformers [65] we design and add a GLU based multi-layer-perceptron layer with GeLU activation for efficiently fusing the aligned features extracted from both the noisy and denoised frames. As shown in Fig. 6(a), we name this deformable convolution-GLU combination as the DC-GFU module. At this fusion-stage each frame is processed separately and the fusion is performed only along the channel dimension. Recurrence is applied for the alignment stage across all the multi-scale features of each frame. For the fusion module, we do not employ recurrence, as different scale features capture distinct long-range dependencies owing to their varying receptive fields.

Deep Feature Fusion using Recurrent Multi-scale Residual Dense Feature Fusion Unit: Post warping we want to perform a robust-to-noise dense feature fusion while taking advantage of the temporal correlations among
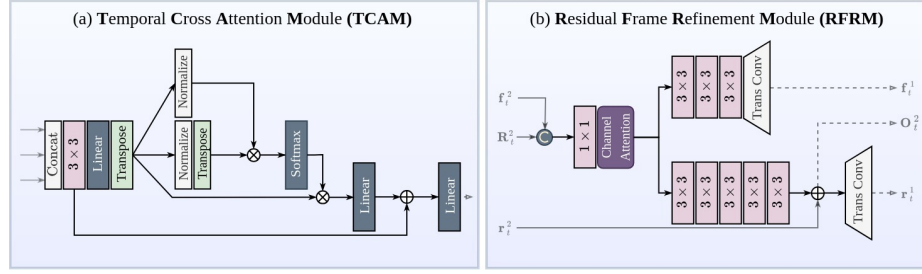
Fig. 7: Design of the proposed modules TCA and RFRM.

the features of all the input frames and also the spatial correlations between the multi-scale features within the same frame. For this task, we design and propose a Recurrent Multi-scale Residual Dense Feature Fusion Unit (RMDF) as shown in Fig. 6(b). The recurrence comes from the fact that the same RMDF module is applied progressively to all the frames' features. For any frame t, the RMDF takes in the corresponding frame's multi-scale features $\{\mathbf{F}_t^1, \mathbf{F}_t^2, \mathbf{F}_t^4\}$, bi-linearly interpolated noisy frames $\{\mathbf{I}_t^1, \mathbf{I}_t^2, \mathbf{I}_t^4\}$ and a hidden state $\mathbf{h}_{t-1}$ as inputs. The multi-scale features are progressively fused in a feed-forward fashion, with Residual Dense Block (RDB) as the basic block, to effectively extract both the short and long range dependencies required for good reconstruction. As shown in Fig. 6(b), multi-scale features are extracted from the noisy frames and fused with the other corresponding input features to minimize any errors accumulated through the previous stages. While these features are utilized to exploit the spatial correlations within the frame, the hidden state $\mathbf{h}$ captures the temporal correlations between all the input frames. Thus, the design of RMDF enables it to extract densely fused multi-scale spatio-temporal features required for enhanced quality outputs.

Multi-Scale Reconstruction using Residual Frame Refinement Module: Considering the heavy noise in the input quanta frames, this ill-posed problem's restored image subspace can be quite large. To output a restored image close to the ground truth we prefer deep supervision that lets the model preserve critical details of the scene. We opt for a multi-scale reconstruction approach where the image at each scale is reconstructed in a progressive fashion. Experiments reveal the efficacy of this approach (Fig. 5, Fig. 11). The stage starts with the lowest scale features extracted from all the frames using RMDF being concatenated and sent into a newly designed Temporal Cross Attention (TCA) module. As shown in Fig. 7(a), the TCA module is similar to the multi-head attention in vision transformers [16] in terms of generating queries, keys, and values. However, we maintain the number of heads to be one and apply attention only on the channel dimension. The cross attention is from the fact that we input features extracted from all the input frames. The extracted cross-attention features are then fed into a newly designed Residual Frame Refinement Module (RFRM). As shown in Fig. 7(b), RFRM takes in a residual frame $\mathbf{r}_t^\alpha$, a

hidden state $\mathbf{f}_t^\alpha$, and the features as input, concatenates the hidden state with the features to input into the channel attention block [82] to emphasize critical spatio-temporal information. Further, we divide the module into 2 branches. While the former is designed to extract multi-scale spatial correlation information and output a modified hidden state $\mathbf{f}_t^{\alpha/2}$, the latter focuses on refining the residual frame to output a corresponding scale reconstructed image $\mathbf{O}_t^\alpha$ and a residual frame $\mathbf{r}_t^{\alpha/2}$. The main purpose of this setup is to initially restore the high-level features through estimating $\mathbf{O}_t^4$, followed by focusing on the low-level intricate details while refining the residual frames for scales 2 and 1.

Loss Function. We train QUIVER with a multi-scale loss. The overall loss function can be represented as

$$\mathcal{L}_{\mathrm{Q}} = \; \lambda_1 \cdot \mathcal{L}(\mathbf{I}^{1,\mathrm{GT}}, \mathbf{I}_d^1) + \lambda_2 \cdot \mathcal{L}(\mathbf{I}_t^{1,\mathrm{GT}}, \mathbf{O}_t^1) + \lambda_3 \cdot \mathcal{L}(\mathbf{I}_t^{2,\mathrm{GT}}, \mathbf{O}_t^2) + \cdots$$
$$\lambda_4 \cdot \mathcal{L}(\mathbf{I}_t^{4,\mathrm{GT}}, \mathbf{O}_t^4), \quad (1)$$

where $\mathbf{I}_t^{\alpha,\mathrm{GT}}$ is the captured t^{th} ground truth frame bicubically downsampled by α, and $\mathcal{L}(\mathbf{I}_a, \mathbf{I}_b) = ||\mathbf{I}_a - \mathbf{I}_b||_1 + ||\nabla_x \mathbf{I}_a - \nabla_x \mathbf{I}_b||_1 + ||\nabla_y \mathbf{I}_a - \nabla_y \mathbf{I}_b||_1$. Here, ∇_x and ∇_y represent the operations of computing horizontal and vertical gradients.

4   Proposed I2-2000FPS dataset

While several high-frame-rate datasets have been open-sourced in recent times [36, 49, 60, 67, 68, 72], these datasets mainly feature videos tainted by severe motion blur, making them unsuitable in our problem setting. Moreover, features such as high motion speed and sufficient number of videos are also not always guaranteed. A visual representation of existing datasets' comparison is shown in Fig. 8(a). To address the gaps, we introduce the I2-2000FPS dataset, a high frame rate video collection meticulously designed to capture high-speed motion with precision.

The I2-2000FPS dataset has a temporal resolution of 2000 FPS and a spatial resolution of 512×1024 pixels, comprising 280 unique videos spanning 114 diverse scenes. The videos are captured using the Chronos 1.4 high-speed CMOS sensor-based camera from Kron Technologies. Notably, I2-2000FPS incorporates dark current calibration, leveraging the camera's capabilities to mitigate dark current effects. Throughout the data collection process, analog and digital gain were consistently maintained at 0 dB to avoid amplification of noise. To minimize noise, the videos are exclusively captured outdoor with ambient lighting conditions. *More details on I2-2000FPS can be found in the supplementary.*

5   Experiments

5.1   Image Formation Model

For experiments involving synthetic data, we use a single-photon detector simulator based on an underlying image formation model discussed below. We build
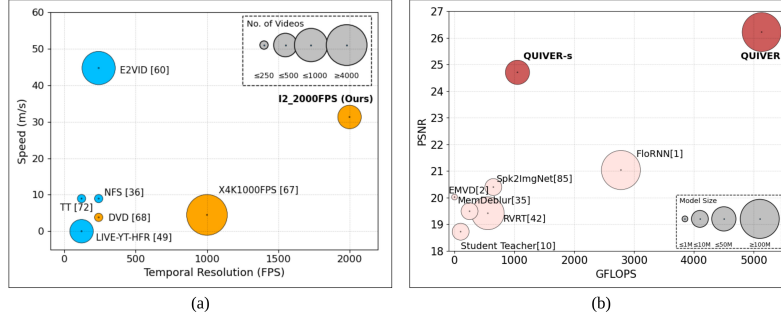
(a)   (b)

Fig. 8: (a) Benchmarking high-speed video datasets. Horizontal axis represents the temporal resolution and the vertical axis indicates the maximum speed captured by the dataset, assuming a fixed camera-object distance. The circles in blue and orange indicate blur and blur-free videos, respectively. **(b) Benchmarking Quanta Video Restoration Models** on the I2-2000FPS dataset. Horizontal axis represents the computational complexity in terms of GFLOPs and the vertical axis indicates the PSNR acquired at 3.25PPP.

upon the prototype initially suggested in [46] adopted in prior works [6, 10, 11, 20, 26, 28, 56].

Given the quanta exposure [22], $\mathbf{I}^{\mathrm{GT}}$, dependent on the photon flux and exposure time, the observed signal by the sensor can be represented as a Poisson-Gaussian random variable, where the Poisson represents the photon arrival process and the Gaussian models the read noise. The readout process involves various sources of distortions and an Analog-to-Digital Converter (ADC) to convert the real numbers into integers $\{0, 1, 2, ..., \mathrm{L}\}$, where $L = 2^{\mathrm{Nbits}} - 1$ depending on the bit-depth (Nbits) allocated to the sensor. The final sensor readout, $\mathbf{Y}$, can be represented using the following equation,

$$\mathbf{Y} \sim \mathrm{ADC}_{[0,\mathrm{L}]}\{\mathrm{Poisson}(\mathrm{QE} \times \mathbf{I}^{\mathrm{GT}} + \theta_{\mathrm{dark}}) + \underbrace{\mathrm{Gauss}(0, \sigma_{\mathrm{read}}^2 \mathbf{1})}_{\text{read noise}}\}. \qquad (2)$$

Akin to previous works [6, 10, 20, 26, 28], we assume our sensor to be monochromatic as we utilize monochromatic real data in our experiments. For our sensor prototype, we utilize a Quantum Efficiency (QE) of 0.80. The dark current (θ_{dark}) and read noise (σ_{read}) are set to $1.6\,\mathrm{e^-/pix/sec}$ and $0.2\,\mathrm{e^-/pix}$, respectively.

5.2   Experimental Settings

Training data. We curate a set of 249 videos from the I2-2000FPS collection and employ it as the training dataset for all the deep-learning models in our experiments. Each training sample is fetched on the fly from each clip. A training sample here is defined as a tuple containing the ground-truth/target frames and the 3-bit quanta frames simulated at 3.25 photons-per-pixel (PPP) (~ 1 lux assuming a 1.1μm pixel pitch and a $1/2000$ second exposure time) using the
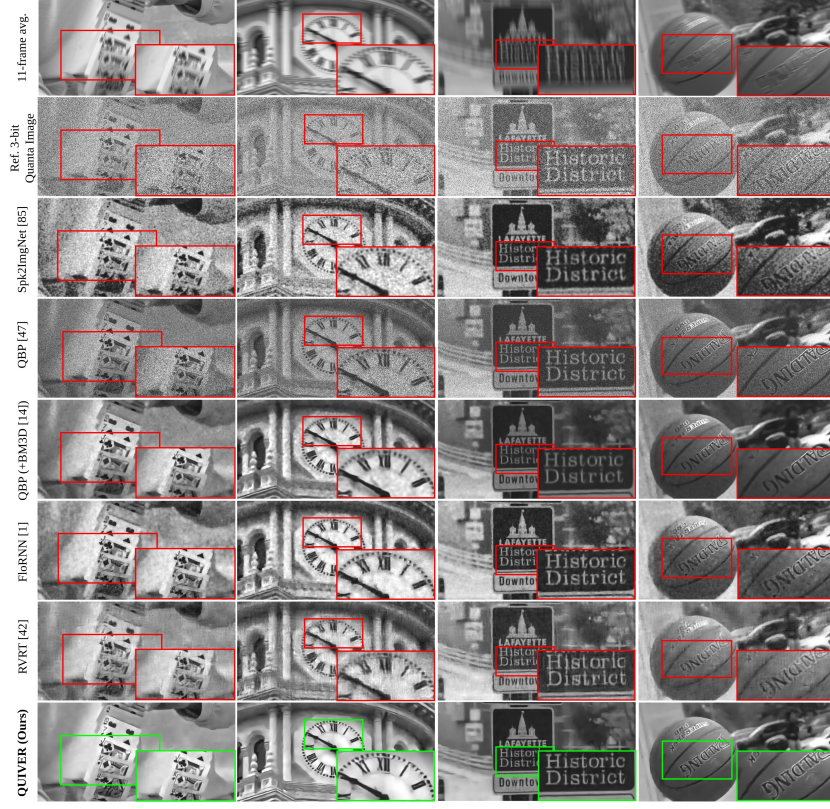
Fig. 9: Visual comparisons of the reconstructed results on test videos from the proposed I2-2000FPS dataset. For fair comparison, all methods utilize 11 3-bit quanta frames simulated at 3.25 PPP per frame ($\sim$ 1 lux) to produce a restored frame. *Best viewed in zoom.*

image formation model described in Sec. 5.1.

Testing data. To effectively analyze the performance of various methods, we carefully sample 31 videos from I2-2000FPS containing various motion types, shapes, and speeds. To test the generalizability, we also test the algorithms on X4K1000FPS [67] test dataset containing 15 videos from distinct scenes. Lastly, to measure the performance on real-world data, we collect binary frames using a SPAD sensor [17] and compare the reconstructed outputs. More details will be discussed in Sec. 5.3.

Baselines. We compare the proposed method with eight existing dynamic scene reconstruction algorithms, namely Transform Denoise [6], QBP [47], Student-Teacher [10], RVRT [42], EMVD [2], FloRNN [1], MemDeblur [35], and Spk2ImgNet [85]. We also add an off-the-shelf denoiser BM3D [14] to QBP, denoted QBP (+BM3D), as a baseline for comparison. As we will discuss in Sec. 5.3, QUIVER beats all the baselines, both quantitatively and qualitatively.
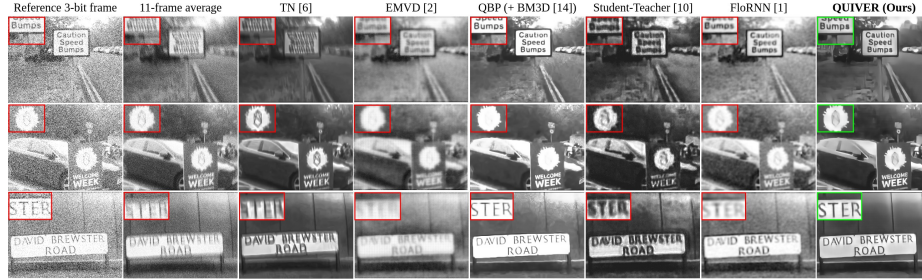
Fig. 10: Performance on Real Quanta Data. We capture real 1-bit quanta data using a SPAD [17] and generate 3-bit frames through temporal averaging. All deep learning based models are trained using a photon level of 4.9 PPP per frame. *Best viewed in zoom.*

Training QUIVER. We utilize the function mentioned in Eq. (1) as the cost function for training QUIVER with regularization parameters $\lambda_1 = 0.2$, $\lambda_2 = 0.85$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.05$. The training data is extracted with patch size 128×128 and a batch size of 4. The weights are initialized with Lecun initialization [39]. The network is trained using the Adam optimizer [37] with an initial learning rate of 2.5×10^{-5}. The low learning rate is driven by the inherent instability of recurrent networks, as it mitigates the risk of divergent behavior during training. We use a learning rate scheduler that reduces the learning rate by a factor of 2 when a plateau is reached. QUIVER takes approximately 1.5 days to train on a NVIDIA A100 Tensor Core GPU using Pytorch.

5.3   Results

Synthetic Data Experiments. We begin with the synthetic experiments where we utilize 3-bit quanta frames, simulated using the parameters mentioned in Sec. 5.1 at 3.25, 9.75, 19.5, and 26 PPP to test the algorithms' performance. Tab. 2 and Tab. 3 demonstrate the PSNR and SSIM [76] of various methods extracted by predicting 6017 I2-2000FPS frames and 345 X4K1000FPS frames. To further substantiate the efficacy of QUIVER's design, we introduced a scaled-down variant, QUIVER-s (Refer Fig. 8(b) for complexity comparison). Quantitative results indicate that both QUIVER and QUIVER-s offer substantially better performance than all the baselines across a range of light levels. Fig. 9 depicts visual results of all the methods on the I2-2000FPS dataset. It is evident that existing methods fail to handle both motion and noise simultaneously, whereas, our proposed method, QUIVER, produces blur free high SNR outputs while preserving high-frequency details to a large extent.

  Real Data Experiments. We verify the methods' performance on real data. The real data is collected as binary frames using a SPAD sensor [17] at 10000 FPS with a spatial resolution of 240×320. As SPADs possess zero read noise, the binary frames are summed up to generate 3-bit frames. The average observed

Table 2: Performance comparison on the proposed I2-2000FPS dataset across various light levels. Models are trained using the I2-2000FPS dataset. QUIVER performs significantly better than the existing methods.

Photons-Per-Pixel (PPP)	3.25		9.75		19.5		26	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Transform Denoise [6]	21.3170	0.7184	23.1521	0.7671	22.7748	0.7812	22.3096	0.7811
QBP [47]	15.9411	0.1293	19.1856	0.2654	20.4000	0.3713	20.7978	0.4114
QBP (+ BM3D [14])	21.5476	0.7033	22.2001	0.6899	22.8351	0.7696	22.8617	0.7832
Student-Teacher [10]	18.7200	0.4006	16.5195	0.2479	15.7636	0.2133	13.2889	0.0735
RVRT [42]	19.4115	0.3539	21.6714	0.4568	22.0826	0.5021	21.7528	0.4968
EMVD [2]	20.0194	0.5873	21.0559	0.6048	22.4403	0.5592	23.4053	0.5576
FloRNN [1]	21.0341	0.6785	25.6132	0.7091	27.4322	0.7395	27.8520	0.7784
MemDeblur [35]	19.4877	0.3868	14.4906	0.1112	16.1775	0.1667	16.0058	0.1712
Spk2ImgNet [85]	20.3945	0.5642	19.6665	0.6733	22.9372	0.7008	14.9769	0.6861
QUIVER-s (Ours)	24.7013	0.7565	**26.8676**	**0.7883**	27.2989	0.8432	27.8659	0.8408
QUIVER (Ours)	**26.2143**	**0.7897**	26.8058	**0.8250**	**27.7538**	**0.8563**	**27.9377**	**0.8446**

Table 3: Performance comparison on the X4K1000FPS dataset across various light levels. Models are trained using the I2-2000FPS dataset. QUIVER performs significantly better than the existing methods.

Photons-Per-Pixel (PPP)	3.25		9.75		19.5		26	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Transform Denoise [6]	19.6255	0.6323	22.1703	0.7044	22.9938	0.7229	22.6230	0.7204
QBP [47]	15.5634	0.2302	16.9758	0.3230	17.1798	0.3957	17.7807	0.4188
QBP (+ BM3D [14])	17.9677	0.5123	18.5308	0.5226	18.2407	0.5414	18.7917	0.5586
Student-Teacher [10]	18.8208	0.3652	16.1548	0.2608	14.9359	0.2571	13.9762	0.1186
RVRT [42]	19.9203	0.3641	21.0781	0.4472	21.4780	0.4925	20.7899	0.4919
EMVD [2]	20.5102	0.4836	21.8152	0.5595	22.9440	0.5936	22.4587	0.5860
FloRNN [1]	20.8283	0.5778	**23.5874**	0.6484	24.3214	0.6683	**25.2483**	0.7170
MemDeblur [35]	19.5534	0.3642	14.5595	0.2203	16.6749	0.3116	15.6496	0.2974
Spk2ImgNet [85]	18.9424	0.4731	19.2532	0.5722	20.3442	0.5716	16.0931	0.6106
QUIVER-s (Ours)	20.9197	0.5955	21.7990	0.6523	24.1924	0.7316	23.4411	0.7248
QUIVER (Ours)	**21.8730**	**0.6521**	23.1654	**0.7057**	**24.5956**	**0.7645**	25.0086	**0.7513**

light level after summation is 4.9 PPP. We generate results with networks trained at 4.9 PPP and demonstrate the visual results in Fig. 10. QUIVER, as opposed to existing state-of-the-art, effectively recovers high-frequency information while applying visually appealing smoothening effect to low-frequency regions of the scene. It is noteworthy that SPADs' image formation model is significantly different from that of the QIS' imaging model [58,59]. Therefore, the visual results also indicate that the proposed QUIVER can thoroughly generalize to various single-photon detectors.

6   Ablation Study

Effect of Pre-Denoising, Optical Flow and Multi-Scale. We conduct an ablation study to evaluate the effect of the pre-denoising, the learnable optical flow, and multi-scale reconstruction on performance. Upon removal of either module, we expand the features dimension of layers in the feature extraction stage and add RDB blocks to the RMDF module, thereby maintaining a similar model capacity. We train all possible combinations and display the quantitative results in Tab. 4. Results indicate that, in the absence of either one or more modules, the network's performance is substantially worse. Visual Results in Fig. 11
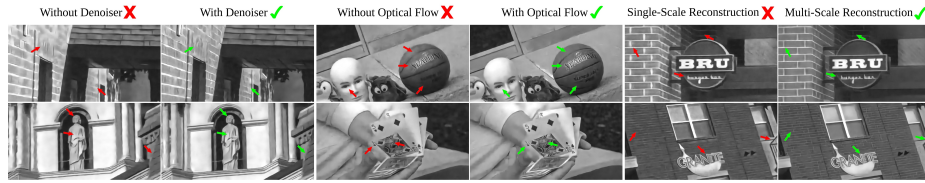
Fig. 11: Ablation Study. Visual Comparisons depicting the effectiveness of Pre-denoiser, Optical Flow, and Multi-Scale Reconstruction Modules. *Best viewed in zoom.*

indicate that these modules serve critical roles as they significantly contribute to model's performance.

Does loading pre-trained optical flow module weights help? We initialize SPyNet with its pre-trained weights and finetune the same while training QUIVER, and display its quantitative results in Tab. 4. As the pre-trained SPyNet is not robust to photon shot noise and read noise, especially in low-light conditions, initializing the module with it will result in sub-optimal performance.

Table 4: Ablation study. We conduct experiments to emphasize the role of Denoiser, Optical flow, and Multi-Scale reconstruction modules. We also show the effect of loading pretrained optical flow weights (*) on performance.

Pre-Denoising	Optical Flow	Multi-Scale	I2-2000FPS	
			PSNR↑	SSIM↑
✗	✗	✗	23.6702	0.7756
✓	✓*	✓	23.9479	0.7709
✓	✓	✗	24.3841	0.7808
✗	✗	✓	24.7445	0.7755
✗	✓	✓	24.9999	0.7753
✓	✗	✓	25.7521	0.7760
✓	✓	✓	**26.2143**	**0.7897**

7   Conclusion

In this paper, we presented a methodology to reconstruct blur-free grayscale images/videos captured using 1-bit or few-bit quanta data. While adopting the ideology of classical quanta restoration methods, we proposed an end-to-end deep learning framework, QUIVER, that utilizes pre-filtering, learnable optical flow module, and a novel multi-scale reconstruction approach to produce high-visual outputs. Experiments on synthetic and real data indicate QUIVER beats state-of-the-art and can generalize across single-photon sensors. We also introduce the world's first high-speed video dataset, I2-2000FPS, that captures fast moving scenes at 2000 fps, covering wide ranges of motion. We believe that I2-2000FPS will be a valuable asset for researchers in high-speed motion analysis and other computer vision tasks.

References

1. Li et al, J.: Unidirectional Video Denoising by Mimicking Backward Recurrent Modules with Look-Ahead Forward Ones. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. Lecture Notes in Computer Science (2022). https://doi.org/10.1007/978-3-031-19797-0_34

2. Maggioni et al, M.: Efficient Multi-Stage Video Denoising With Recurrent Spatio-Temporal Fusion. In: CVPR (2021)

3. Arias, P., Morel, J.M.: Video Denoising via Empirical Bayesian Estimation of Space-Time Patches. Journal of Mathematical Imaging and Vision **60**(1), 70–93 (Jan 2018). https://doi.org/10.1007/s10851-017-0742-4

4. Chan, K.C.K., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5972–5981 (2022)

5. Chan, S.H.: What Does a One-Bit Quanta Image Sensor Offer? IEEE Transactions on Computational Imaging **8**, 770–783 (2022). https://doi.org/10.1109/TCI.2022.3202012

6. Chan, S.H., Elgendy, O.A., Wang, X.: Images from Bits: Non-Iterative Image Reconstruction for Quanta Image Sensors. Sensors **16**(11), 1961 (Nov 2016). https://doi.org/10.3390/s16111961

7. Chan, S.H., Lu, Y.M.: Efficient image reconstruction for gigapixel quantum image sensors. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 312–316 (Dec 2014). https://doi.org/10.1109/GlobalSIP.2014.7032129

8. Chan, S.H., Wang, X., Elgendy, O.A.: Plug-and-Play ADMM for Image Restoration: Fixed-Point Convergence and Applications. IEEE Transactions on Computational Imaging **3**(1), 84–98 (Mar 2017). https://doi.org/10.1109/TCI.2016.2629286

9. Charbon, E., Fishburn, M., Walker, R., Henderson, R.K., Niclass, C.: SPAD-Based Sensors. In: Remondino, F., Stoppa, D. (eds.) TOF Range-Imaging Cameras, pp. 11–38. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-27523-4_2

10. Chi, Y., Gnanasambandam, A., Koltun, V., Chan, S.H.: Dynamic Low-Light Imaging with Quanta Image Sensors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020, vol. 12366, pp. 122–138. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_8

11. Chi, Y., Zhang, X., Chan, S.H.: HDR imaging with spatially varying signal-to-noise ratios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5724–5734. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00554, https://doi.org/10.1109/CVPR52729.2023.00554

12. Choi, J.H., Elgendy, O.A., Chan, S.H.: Image Reconstruction for Quanta Image Sensors Using Deep Neural Networks. In: 2018 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). pp. 6543–6547 (Apr 2018). https://doi.org/10.1109/ICASSP.2018.8461685

13. Claus, M., van Gemert, J.: ViDeNN: Deep Blind Video Denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)

14. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on image processing **16**(8), 2080–2095 (2007)

15. Dong, S., Huang, T., Tian, Y.: Spike camera and its coding methods. arXiv preprint arXiv:2104.04669 (2021)

16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

17. Dutton, N.A.W., Gyongy, I., Parmesan, L., Gnecchi, S., Calder, N., Rae, B.R., Pellegrini, S., Grant, L.A., Henderson, R.K.: A SPAD-Based QVGA Image Sensor for Single-Photon Counting and Quanta Imaging. IEEE Transactions on Electron Devices **63**(1), 189–196 (Jan 2016). https://doi.org/10.1109/TED.2015.2464682

18. Dutton, N.A., Gyongy, I., Parmesan, L., Henderson, R.K.: Single photon counting performance and noise analysis of CMOS SPAD-based image sensors. Sensors **16**(7), 1122 (2016)

19. Elgendy, O.A., Chan, S.H.: Optimal Threshold Design for Quanta Image Sensor. IEEE Transactions on Computational Imaging **4**(1), 99–111 (Mar 2018). https://doi.org/10.1109/TCI.2017.2781185

20. Feng Yang, Lu, Y.M., Sbaiz, L., Vetterli, M.: Bits From Photons: Oversampled Image Acquisition Using Binary Poisson Statistics. IEEE Transactions on Image Processing **21**(4), 1421–1436 (Apr 2012). https://doi.org/10.1109/TIP.2011.2179306

21. Fossum, E., Ma, J., Masoodian, S., Anzagira, L., Zizza, R.: The Quanta Image Sensor: Every Photon Counts. Sensors **16**(8), 1260 (Aug 2016). https://doi.org/10.3390/s16081260

22. Fossum, E.R.: Modeling the performance of single-bit and multi-bit quanta image sensors. IEEE Journal of the Electron Devices Society **1**(9), 166–174 (2013)

23. Gao, J., Shang, Z., Nie, K., Luo, T.: High dynamic range image reconstruction for multi-bit quanta image sensor. Optoelectronics Letters **18**(9), 553–558 (Sep 2022). https://doi.org/10.1007/s11801-022-2014-9

24. Gariepy, G., Krstajić, N., Henderson, R., Li, C., Thomson, R.R., Buller, G.S., Heshmat, B., Raskar, R., Leach, J., Faccio, D.: Single-photon sensitive light-in-fight imaging. Nature Communications **6**(1), 6021 (Jan 2015). https://doi.org/10.1038/ncomms7021

25. Gariepy, G., Leach, J., Warburton, R., Chan, S., Henderson, R., Faccio, D.: Picosecond time-resolved imaging using SPAD cameras. In: Lewis, K.L., Hollins, R.C. (eds.) SPIE Security + Defence. p. 99920N. Edinburgh, United Kingdom (Oct 2016). https://doi.org/10.1117/12.2241184

26. Gnanasambandam, A., Chan, S.H.: HDR Imaging With Quanta Image Sensors: Theoretical Limits and Optimal Reconstruction. IEEE Transactions on Computational Imaging **6**, 1571–1585 (2020). https://doi.org/10.1109/TCI.2020.3041093

27. Gnanasambandam, A., Chan, S.H.: Image Classification in the Dark Using Quanta Image Sensors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Proceed-

ings of the European Conference on Computer Vision (ECCV). pp. 484–501. Lecture Notes in Computer Science, Cham (2020). `https://doi.org/10.1007/978-3-030-58598-3_29`

28. Gnanasambandam, A., Chan, S.H.: Exposure-Referred Signal-to-Noise Ratio for Digital Image Sensors. IEEE Transactions on Computational Imaging **8**, 561–575 (2022). `https://doi.org/10.1109/TCI.2022.3187657`

29. Godard, C., Matzen, K., Uyttendaele, M.: Deep Burst Denoising. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 538–554 (2018)

30. Gupta, A., Ingle, A., Velten, A., Gupta, M.: Photon-Flooded Single-Photon 3D Cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6770–6779 (2019)

31. Gupta, S., Gupta, M.: Eulerian Single-Photon Vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10465–10476 (2023)

32. Gutierrez-Barragan, F., Ingle, A., Seets, T., Gupta, M., Velten, A.: Compressive Single-Photon 3D Cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17854–17864 (2022)

33. Gyongy, I., Dutton, N., Henderson, R.: Single-Photon Tracking for High-Speed Vision. Sensors **18**(2), 323 (Jan 2018). `https://doi.org/10.3390/s18020323`

34. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics **35**(6), 1–12 (Nov 2016). `https://doi.org/10.1145/2980179.2980254`

35. Ji, B., Yao, A.: Multi-Scale Memory-Based Video Deblurring. In: CVPR (2022)

36. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for Speed: A Benchmark for Higher Frame Rate Object Tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1125–1134 (2017)

37. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) (2015). `https://doi.org/10.48550/arXiv.1412.6980`

38. Lebrun, M., Buades, A., Morel, J.M.: A Nonlocal Bayesian Image Denoising Algorithm. SIAM Journal on Imaging Sciences **6**(3), 1665–1688 (Jan 2013). `https://doi.org/10.1137/120874989`

39. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (Nov/1998). `https://doi.org/10.1109/5.726791`

40. Li, C., Qu, X., Gnanasambandam, A., Elgendy, O.A., Ma, J., Chan, S.H.: Photon-limited object detection using non-local feature matching and knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3976–3987 (2021)

41. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: VRT: A Video Restoration Transformer (Jun 2022). `https://doi.org/10.48550/arXiv.2201.12288`

42. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent Video Restoration Transformer with Guided Deformable Attention. Advances in Neural Information Processing Systems **35**, 378–393 (Dec 2022)

43. Liba, O., Murthy, K., Tsai, Y.T., Brooks, T., Xue, T., Karnad, N., He, Q., Barron, J.T., Sharlet, D., Geiss, R., Hasinoff, S.W., Pritch, Y., Levoy, M.: Handheld mobile photography in very low light. ACM Transactions on Graphics **38**(6), 1–16 (Dec 2019). `https://doi.org/10.1145/3355089.3356508`

44. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3D imaging with deep sensor fusion. ACM Transactions on Graphics **37**(4), 113–1 (2018)
45. Ma, J., Chan, S., Fossum, E.R.: Review of Quanta Image Sensors for Ultralow-Light Imaging. IEEE Transactions on Electron Devices **69**(6), 2824–2839 (Jun 2022). `https://doi.org/10.1109/TED.2022.3166716`
46. Ma, J., Masoodian, S., Starkey, D.A., Fossum, E.R.: Photon-number-resolving megapixel image sensor at room temperature without avalanche gain. Optica **4**(12), 1474–1481 (Dec 2017). `https://doi.org/10.1364/OPTICA.4.001474`
47. Ma, S., Gupta, S., Ulku, A.C., Bruschini, C., Charbon, E., Gupta, M.: Quanta burst photography. ACM Transactions on Graphics **39**(4) (Aug 2020). `https://doi.org/10.1145/3386569.3392470`
48. Ma, S., Mos, P., Charbon, E., Gupta, M.: Burst Vision Using Single-Photon Cameras. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5375–5385 (2023)
49. Madhusudana, P.C., Yu, X., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Subjective and Objective Quality Assessment of High Frame Rate Videos. IEEE Access **9**, 108069–108082 (2021). `https://doi.org/10.1109/ACCESS.2021.3100462`
50. Maggioni, M., Boracchi, G., Foi, A., Egiazarian, K.: Video Denoising, Deblocking, and Enhancement Through Separable 4-D Nonlocal Spatiotemporal Transforms. IEEE Transactions on Image Processing **21**(9), 3952–3966 (Sep 2012). `https://doi.org/10.1109/TIP.2012.2199324`
51. Monakhova, K., Richter, S.R., Waller, L., Koltun, V.: Dancing Under the Stars: Video Denoising in Starlight. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16241–16251 (2022)
52. Niclass, C., Rochas, A., Besse, P.A., Charbon, E.: Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. IEEE Journal of Solid-State Circuits **40**(9), 1847–1854 (Sep 2005). `https://doi.org/10.1109/JSSC.2005.848173`
53. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6820–6829 (2019)
54. Pearl, N., Treibitz, T., Korman, S.: NAN: Noise-Aware NeRFs for Burst-Denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12672–12681 (2022)
55. Plotz, T., Roth, S.: Benchmarking Denoising Algorithms With Real Photographs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1586–1595 (2017)
56. Qu, X., Chi, Y., Chan, S.H.: Spatially varying exposure with 2-by-2 multiplexing: Optimality and universality. IEEE Transactions on Computational Imaging **10**, 261–276 (2024)
57. Ranjan, A., Black, M.J.: Optical Flow Estimation Using a Spatial Pyramid Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4161–4170 (2017)
58. Rapp, J., Goyal, V.K.: A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging. IEEE Transactions on Computational Imaging **3**(3), 445–459 (Sep 2017). `https://doi.org/10.1109/TCI.2017.2706028`
59. Rapp, J., Ma, Y., Dawson, R.M.A., Goyal, V.K.: Dead Time Compensation for High-Flux Ranging. IEEE Transactions on Signal Processing **67**(13), 3471–3486 (Jul 2019). `https://doi.org/10.1109/TSP.2019.2914891`

60. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High Speed and High Dynamic Range Video with an Event Camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(6), 1964–1980 (Jun 2021). `https://doi.org/10.1109/TPAMI.2019.2963386`

61. Remez, T., Litany, O., Bronstein, A.: A picture is worth a billion bits: Real-time image reconstruction from dense binary threshold pixels. In: 2016 IEEE International Conference on Computational Photography (ICCP). pp. 1–9 (May 2016). `https://doi.org/10.1109/ICCPHOT.2016.7492874`

62. Rochas, A.: Single photon avalanche diodes in CMOS technology. Tech. rep., Citeseer (2003)

63. Ruget, A., Tyler, M., Mora Martín, G., Scholes, S., Zhu, F., Gyongy, I., Hearn, B., McLaughlin, S., Halimi, A., Leach, J.: Pixels2Pose: Super-resolution time-of-flight imaging for 3D pose estimation. Science Advances **8**(48), eade0123 (Nov 2022). `https://doi.org/10.1126/sciadv.ade0123`

64. Seets, T., Ingle, A., Laurenzis, M., Velten, A.: Motion Adaptive Deblurring With Single-Photon Cameras. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1945–1954 (2021)

65. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)

66. Shin, D., Xu, F., Venkatraman, D., Lussana, R., Villa, F., Zappa, F., Goyal, V.K., Wong, F.N.C., Shapiro, J.H.: Photon-efficient imaging with a single-photon camera. Nature Communications **7**(1), 12046 (Jun 2016). `https://doi.org/10.1038/ncomms12046`

67. Sim, H., Oh, J., Kim, M.: XVFI: eXtreme Video Frame Interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14489–14498 (2021)

68. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep Video Deblurring for Hand-Held Cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1279–1288 (2017)

69. Tassano, M., Delon, J., Veit, T.: DVDNET: A Fast Network for Deep Video Denoising. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1805–1809. IEEE, Taipei, Taiwan (Sep 2019). `https://doi.org/10.1109/ICIP.2019.8803136`

70. Tassano, M., Delon, J., Veit, T.: FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1354–1363 (2020)

71. Vaksman, G., Elad, M., Milanfar, P.: Patch Craft: Video Denoising by Deep Modeling and Patch Matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2157–2166 (2021)

72. Voeikov, R., Falaleev, N., Baikulov, R.: TTNet: Real-Time Temporal and Spatial Video Analysis of Table Tennis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 884–885 (2020)

73. Vogels, T., Rousselle, F., Mcwilliams, B., Röthlin, G., Harvill, A., Adler, D., Meyer, M., Novák, J.: Denoising with kernel prediction and asymmetric loss functions. ACM Transactions on Graphics **37**(4), 124:1–124:15 (Jul 2018). `https://doi.org/10.1145/3197517.3201388`

74. Wang, W., Chen, X., Yang, C., Li, X., Hu, X., Yue, T.: Enhancing Low Light Videos by Exploring High Sensitivity Camera Noise. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4111–4119 (2019)

75. Wang, X.: Single-Photon Cameras Image Reconstruction Using Vision Transformer. In: 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI). pp. 296–300 (May 2023). `https://doi.org/10.1109/CCAI57533.2023.10201259`

76. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (Apr 2004). `https://doi.org/10.1109/TIP.2003.819861`

77. Wong, H.T., Leung, C.S., Ho, D.: Theoretical analysis and image reconstruction for multi-bit quanta image sensors. Signal Processing **185**, 108087 (Aug 2021). `https://doi.org/10.1016/j.sigpro.2021.108087`

78. Yang, F., Lu, Y.M., Sbaiz, L., Vetterli, M.: An optimal algorithm for reconstructing images from binary measurements. In: Bouman, C.A., Pollak, I., Wolfe, P.J. (eds.) IS&T/SPIE Electronic Imaging. p. 75330K. San Jose, California (Feb 2010). `https://doi.org/10.1117/12.850887`

79. Yang, F., Sbaiz, L., Charbon, E., Süsstrunk, S., Vetterli, M.: Image reconstruction in the gigavision camera. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. pp. 2212–2219 (Sep 2009). `https://doi.org/10.1109/ICCVW.2009.5457554`

80. Zhang, D., Lian, Q., Su, Y., Ren, T.: Dual-Prior Integrated Image Reconstruction for Quanta Image Sensors Using Multi-Agent Consensus Equilibrium. IEEE/CAA Journal of Automatica Sinica **10**(6), 1407–1420 (Jun 2023). `https://doi.org/10.1109/JAS.2023.123390`

81. Zhang, D., Lian, Q., Yang, Y.: TwP: Two-stage projection framework with manifold constraint for image reconstruction. Digital Signal Processing **141**, 104186 (Sep 2023). `https://doi.org/10.1016/j.dsp.2023.104186`

82. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)

83. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual Dense Network for Image Restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(7), 2480–2495 (Jul 2021). `https://doi.org/10.1109/TPAMI.2020.2968521`

84. Zhao, J., Xiong, R., Huang, T.: High-Speed Motion Scene Reconstruction for Spike Camera via Motion Aligned Filtering. In: 2020 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5. IEEE, Seville, Spain (Oct 2020). `https://doi.org/10.1109/ISCAS45731.2020.9181055`

85. Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T.: Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11996–12005 (2021)

86. Zhao, J., Xiong, R., Xie, J., Shi, B., Yu, Z., Gao, W., Huang, T.: Reconstructing Clear Image for High-Speed Motion Scene With a Retina-Inspired Spike Camera. IEEE Transactions on Computational Imaging **8**, 12–27 (2022). `https://doi.org/10.1109/TCI.2021.3136446`