Panoramas from Photons

Sacha Jungerman[†] sjungerman@wisc.edu

Atul Ingle[§] ingle2@pdx.edu

Mohit Gupta[†] mohitg@cs.wisc.edu

[†]University of Wisconsin-Madison [§]Portland State University

Abstract

Scene reconstruction in the presence of high-speed motion and low illumination is important in many applications such as augmented and virtual reality, drone navigation, and autonomous robotics. Traditional motion estimation techniques fail in such conditions, suffering from too much blur in the presence of high-speed motion and strong noise in low-light conditions. Single-photon cameras have recently emerged as a promising technology capable of capturing hundreds of thousands of photon frames per second thanks to their high speed and extreme sensitivity. Unfortunately, traditional computer vision techniques are not well suited for dealing with the binary-valued photon data captured by these cameras because these are corrupted by extreme Poisson noise. Here we present a method capable of estimating extreme scene motion under challenging conditions, such as low light or high dynamic range, from a sequence of high-speed image frames such as those captured by a single-photon camera. Our method relies on iteratively improving a motion estimate by grouping and aggregating frames after-the-fact, in a stratified manner. We demonstrate the creation of high-quality panoramas under fast motion and extremely low light, and super-resolution results using a custom single-photon camera prototype. For code and supplemental material see our project webpage.

1. Introduction

Accurate recovery of motion from a sequence of images is one of the most fundamental tasks in computer vision, with numerous applications in robotics, augmented reality, user interfaces, and autonomous navigation. When successfully estimated, motion information can be used to locate and track the camera or different objects in the scene [7], perform motion-aware video compression [19] or stabilization [22], relate multiple sensors, merge information across different viewpoints, and even reconstruct city-scale 3D models using only images from the web [1, 34, 28].

Image sequences can be used to estimate different kinds of motion ranging in complexity and degrees of freedom, from global motion models, such as simple translations, projective warps, or 3D (6-DoF) camera pose, to non-rigid, local motion models such as optical flow. However, regardless of the motion model, traditional methods cannot recover motion that is simply too fast for the camera to capture. This is especially challenging when capturing scenes in low-light conditions—the camera will compensate by increasing the exposure, thereby introducing motion blur, as seen in Fig. 1(a), or increasing the gain (ISO), thereby introducing noise [14]. Fundamentally, the image degradation associated with faster motion or a darker scene causes traditional motion estimation methods to fail.

One way to handle fast motion is by using specialized high-speed cameras. However, such cameras are not only bulky and costly but also suffer from extremely low signal-to-noise ratio due to both low signal values and high readout noise, at least an order of magnitude higher than conventional CMOS cameras[1]. This requires the scenes to be well illuminated, often in a controlled setting, further limiting their scope and widespread adoption.

Fortunately, there is an emerging class of sensors called single-photon cameras, which are capable of high-speed imaging in low-light conditions. Single-photon cameras based on single-photon avalanche diode (SPAD) technology [6] provide extreme sensitivity, are cheap to manufacture, and are increasingly becoming commonplace, recently getting deployed in consumer devices such as iPhones. The key benefit of SPADs is that they do not suffer from read-noise, enabling captures at hundreds of thousands of frames per second even in extremely low flux, while being limited only by the fundamental photon noise.

Although single-photon cameras can capture scene information at unprecedented sensitivity and speed, each individual captured frame is binary valued: a pixel is "on" if at least one photon is detected during the exposure time and "off" otherwise. This binary imaging model presents unique challenges. Traditional image registration techniques rely on feature-based matching, or direct optimization using differences between pixel intensities, both of which rely on image gradients to converge to a solution. Individual binary

[‡] This research was supported in part by an NSF CAREER award 1943149, NSF award CNS-2107060 and NSF ECCS-2138471.

[1] For example, the Phantom v2640 has read noise up to $58e^-$.

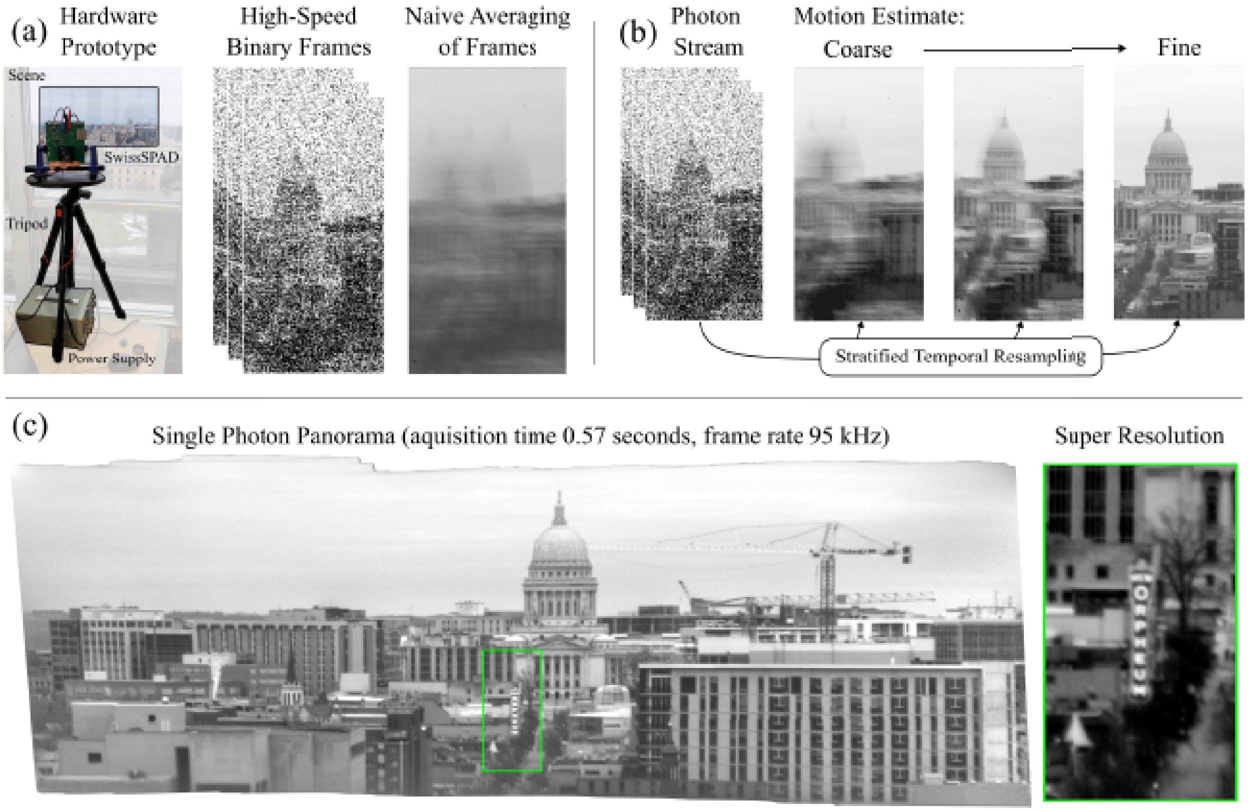

Figure 1. **Single photon panoramas:** (a) Our single-photon camera prototype can capture binary frames at , frames per second. Conventional processing techniques that average the raw image frames struggle due to extreme motion blur. (b) Our proposed method recovers a high-quality scene reconstruction by iteratively refining a motion estimate and re-aggregating the raw photon data. (c) An example high-speed panorama reconstructed from single-photon frames in merely half a second total capture time.

images suffer from severe noise and quantization (only having 1-bit worth of information per pixel), and are inherently non-differentiable, making it challenging, if not impossible, to apply conventional image registration and motion estimation techniques directly on binary frames. Aggregating sequences of binary frames over time increases signal (Fig. 1(a)) but comes at the cost of potentially severe motion blur, creating a fundamental noise-vs-blur tradeoff.

We present a technique capable of estimating rapid motion from a sequence of high-speed binary frames captured using a single-photon camera. Our key insight is that these binary frames can be aggregated in post-processing in a *motion-aware manner* so that more signal and bit-depth are collected, while simultaneously minimizing motion blur. As seen in Fig. 1(b), our method iteratively improves the initial motion estimate, ultimately enabling scene reconstruction under rapid motion and low light and conditions.

Scope and Capabilities: We demonstrate the recovery of global projective motion (homography), enabling the capture of high-speed panoramas with super-resolution and high dynamic range capabilities. As shown in Fig. 1(c), our algorithm can reconstruct a high-quality panorama, captured in less than a second over a wide field-of-view, while simultaneously super-resolving details such as text from a long distance (1300 m). The ideas presented in this paper could also be used to enhance recent one-shot local motion compensation work [27, 35, 20] as they are complementary.

Limitations: Although single-photon camera technology is rapidly evolving, today's SPAD arrays suffer from limitations such as low fill-factors, low spatial resolution, and lack of high-quality color filters. This limits the visual quality of the experimental results shown here. Fortunately, given the trend towards higher resolution SPAD arrays [31] and the increasing commercial availability of this technology [32], these are not fundamental limitations.

2. Related Work

Image Stitching: Merging multiple images together to create a large cohesive image, referred to as a panorama or mosaic, is a classical problem in computer vision. It consists of two main steps, namely image registration, and merging. To register images, most approaches either rely on computing image features, such as SIFT features [25], or direct optimization of the warps, such as the Lucas-Kanade algorithm [26] or variants thereof [3]. Once features are extracted, it is possible to match them between images to compute the warps that relate one image to another [5]. More recent techniques use learning-based methods to extract features [24] allowing them to match cross-domain images such as satellite and map images [41]. Unfortunately, the presence of extreme Poisson noise causes traditional stitching approaches to fail for high-speed binary frames.

Structure from Motion (SfM): SfM techniques estimate both the 3D geometry of the environment and the location of the camera simultaneously. Most SfM pipelines (e.g., COLMAP [34]) use feature-based approaches to match frames. Others have extended the optimization-based approaches of Lucas-Kanade to 3D pose estimation [4, 9]. While these methods can be robust to some types of noise, such as Gaussian, salt and pepper, and speckle [33], they all rely on computing image gradient either as part of the optimization process or as part of the feature extraction process and thus are not well suited for noisy and quantized high-speed images such as binary images.

Burst Photography and Denoising: Techniques that remove image noise can be used as a pre-processing step to aid the feature matching or direct optimization process. For instance, blind denoisers [8] or state-of-the-art deep video denoising networks [38] could be used. The burst processing of binary frames is also possible [27] but relies on motion compensation to denoise images which can be expensive as it requires computing optical flow on each frame. For many scenes, computing optical flow is unnecessary as motion might be primarily dominated by ego-motion. We propose an iterative approach that refines a global motion estimate and enables high-quality scene recovery while being computationally less burdensome than optical flow approaches that use patch-based processing.

Event-based Processing: Event cameras (dynamic vision sensors) superficially resemble high-speed binary frames: they produce high frequency, low-bit depth observation of a scene and have been used for fast tracking [21] and odometry [16]. However, event cameras suffer from high sensor noise and event clutter caused by camera motion [13] leading many works to use a fusion approach, combining events with conventional camera frames. Our method focuses on intensity frames, whether captured from a single photon camera or other sensing modality.

3. Image Formation via Virtual Exposures

Consider a series of images captured by a camera as the scene and/or camera undergoes motion. Suppose our goal is to register a pair of consecutive images, under a given motion model (local or global). In ideal imaging conditions (sufficient light, relatively small motion), conventional motion estimation and registration techniques perform robustly. This is demonstrated in Fig. 2 (a), where a SIFT-based feature matching technique is able to find reliable matches across images. However, in settings involving low-light and rapid motion, the number of successful feature matches drop, resulting in erroneous motion estimation. This is due to the fundamental noise-vs-blur trade-off—the captured images either have strong noise or large motion blur, depending on the exposure length used, both of which prevent reliable feature detection and image registration (Fig. 2 (b-d)). More generally, this trade-off limits the performance of several computer vision and imaging techniques that require motion estimation across a sequence of images [15, 23]. Is there a way to overcome this trade-off?

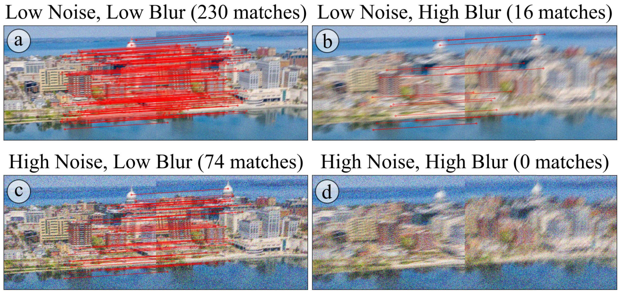
Low Noise, Low Blur (230 matches)
Low Noise, High Blur (16 matches)
High Noise, Low Blur (74 matches)
High Noise, High Blur (0 matches)
a
b
c
d

Figure 2. **Registration Accuracy vs. Noise and Blur:** Conventional feature matching techniques, such as exhaustively matched SIFT features shown here, work well with low noise, and low motion blur frames. The number of successful feature matches drops off at higher noise and blur levels.

Mitigating Blur-Noise Trade-off via Virtual Exposures: Conventional cameras integrate the scene's radiance during an exposure, and produce a single image. Under the challenging conditions described above, this image may be too noisy or blurred, depending on the exposure duration. Suppose instead, we were able to record the arrival time of each photon during an exposure, creating a *3D photon cube* (and spatial dimensions, and an extra photon arrival time dimension) [10, 11]. This information is richer than what a conventional camera image can afford us, but what can we do with it? While we can reconstruct a conventional camera image by simply summing over the time slices of the photon cube, we can combine this photon data in multiple ways post hoc. We could apply arbitrary transformations to each time slice before collapsing it into one or more final images. We refer to this idea of aggregating photon information after-the-fact as *virtual exposures*. In contrast, once a conventional image has been captured, undoing the effect of motion artifacts is severely ill-posed.

Stratified Temporal Re-Sampling: Our key insight, enabled by the concept of virtual exposures, is that we can compensate for motion at the level of individual photon arrivals to create high-fidelity aggregate frames, which in turn can be used to further refine the motion estimates. Virtual exposures are created by re-sampling the photon-cube post-capture, allowing arbitrary, fluid, and even overlapping exposures, enabling us to resolve higher speed motion.

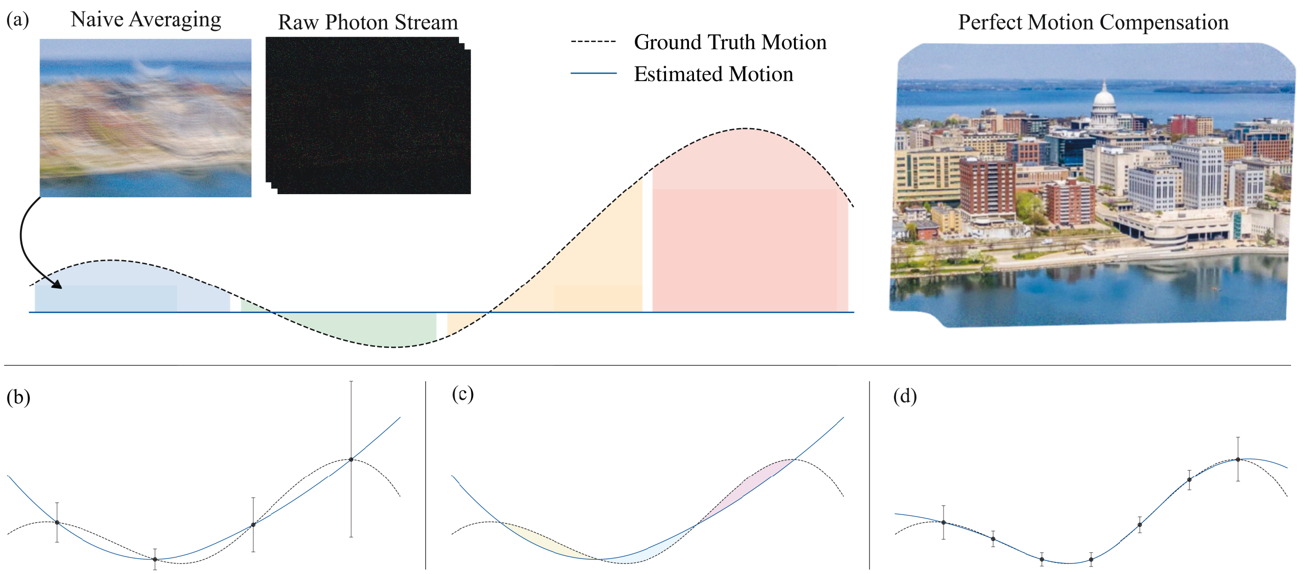

Figure 3. **Motion Estimation using Stratified Temporal Re-Sampling:** Perfect scene reconstruction can be achieved if the scene motion is precisely known. (a) We cannot recover a motion estimate from the raw photon data as it is binary-valued, extremely noisy, and non-differentiable. An initial motion estimate (blue line) is obtained using locally averaged groups of frames (shaded regions). (b) This blur causes the registration algorithm to produce noisy motion estimates (black error bars) from which we can update our estimated motion trajectory. (c) With this new trajectory, the apparent motion is smaller (shaded area), leading to higher-quality virtual exposures. We can also sample new virtual exposures as needed, here we show new frames centered around the midpoints of previous frames. (d) These lead to improved motion estimates. The estimated motion trajectory converges to the true motion over several iterations.

An abstract example of this concept is illustrated in Fig. 3. We start with an initial set of virtual exposures, which are simply aggregate frames with no motion compensation akin to a sequence of short exposures from a conventional camera. From these, we estimate a coarse motion trajectory using an off-the-shelf motion model. Although conventional techniques will output potentially erroneous estimates, we propose an iterative approach, where these motion estimates are used to spatiotemporally warp the underlying photon data and re-combine it into less blurry images. This is repeated to create additional virtual exposures until convergence, resulting in improved motion estimates.

How to Capture Photon Cubes? Thus far, we assumed having access to a continuous-time stream of photons that contains precise timing and location information. In practice, we approximate this photon stream with a camera capable of high-speed temporal sampling. While several high-speed sensing technologies exist today, we focus on single-photon avalanche diode (SPAD) sensors. In the following, we describe their unique image formation model that enables high-speed photon-level sensing, which can emulate virtual exposures whose signal-to-noise ratio (SNR) is limited only by the fundamental limits of photon noise.

SPAD Image Formation Model: For a static scene with a radiant flux (photons/second) of ϕ, during an exposure time τ, the probability of observing k incident photons on a SPAD camera pixel follows a Poisson distribution:

$$P(k) = \frac{(\phi\tau)^k e^{-\phi\tau}}{k!}. \tag{1}$$

After each photon detection, the SPAD pixel enters a dead time during which the pixel circuitry resets. During this dead time, the SPAD cannot detect additional photons. The SPAD pixel output during this exposure τ is binary-valued and follows a Bernoulli distribution given by[2]:

$$P(k=0) = e^{-\phi\tau}, \qquad P(k=1) = 1 - e^{-\phi\tau}. \tag{2}$$

Emulating Virtual Exposures: Given n binary observations B_i of a scene, we can capture a virtual exposure using the following maximum likelihood estimator [40]:

$$\widehat{\phi} = -\frac{1}{\tau}\ln\left(1 - \frac{1}{n}\sum_{i=1}^{n} B_i\right). \tag{3}$$

Different virtual exposures can be emulated by varying the starting index i and the number n of binary frames. The granularity and flexibility of these virtual exposures is limited only by the frame rate of the SPAD array, which reaches up to $\sim 100kfps$, enabling robust motion estimation at extremely fine time scales. Furthermore, SPAD arrays have negligible read noise and quantization noise, leading to significantly higher SNR as compared to conventional images captured over the same exposures.

[2] Source of noise such as dark counts and non-ideal quantum efficiency can be absorbed into the value of ϕ.

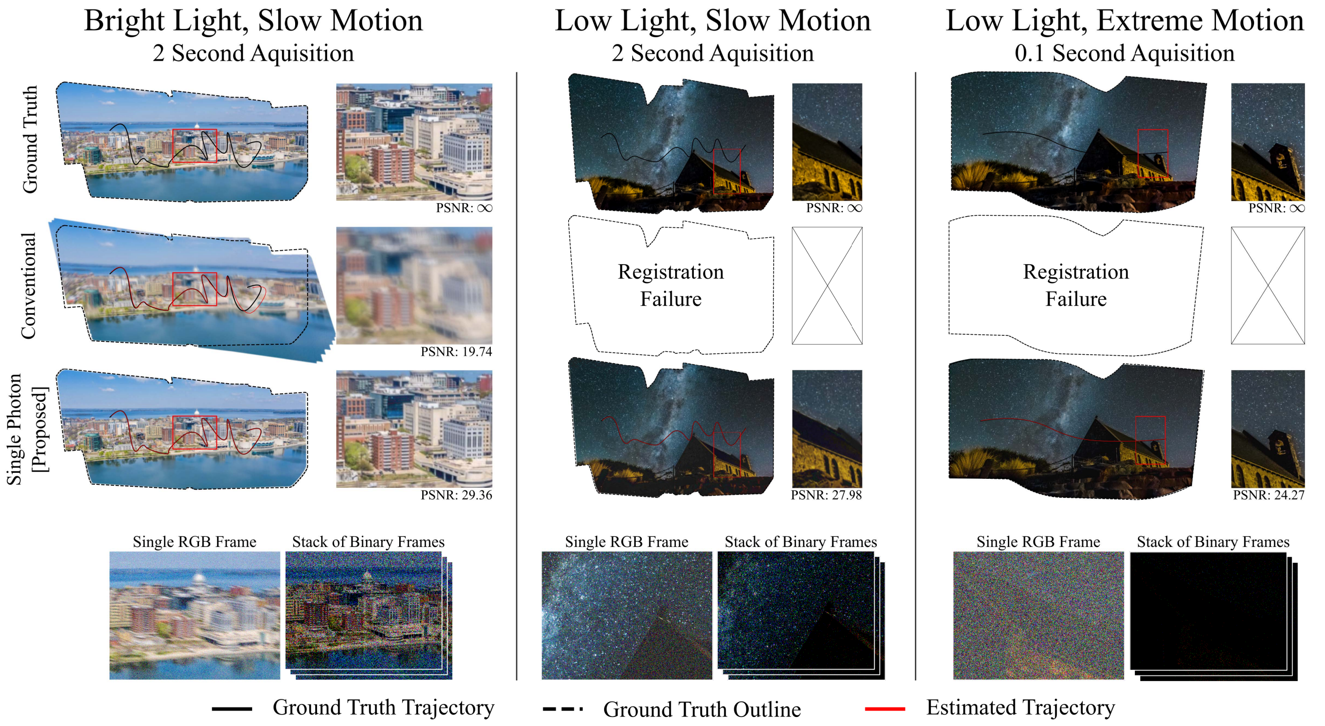

Figure 4. **Simulated Panoramas:** (left) We show a ground truth panorama, and a zoomed-in section of it, as well as one created with realistic blurry RGB frames and another created under the same conditions using binary frames. Reliable reconstruction using traditional RGB frames is impossible due to large motion blur, however, we can compensate for fine-grain motion using binary frames resulting in perfect reconstruction. (middle) The baseline fails in low light, this scene is darker than the left one. (right) Our method works even with less light. The baseline reconstruction fails as each conventional image is dominated by read noise.

4. Stratified Estimation of High-Speed Motion

While the ideas presented in Section 3 are applicable to a wide range of motion models, we focus on image homographies, a global motion model. We propose a modular technique for homography estimation from photon cube data, which is capable of localizing high-speed motion even in ultra-low light settings. As an example application, we demonstrate panoramic reconstruction from photon cubes by using the estimated homographies to warp binary frames onto a common reference frame. Given a temporal sequence of binary frames , we compute and iteratively refine image homographies and the resulting reconstruction through the following steps:

Re-sample: Sample binary frames across the photon cube which will be merged together.

Merge: Merge the sampled frames using the current per-frame homography estimate.

Locate: Apply an off-the-shelf motion estimation algorithm to the merged frames.

Interpolate: Interpolate the estimated homographies to the granularity of individual binary frames.

With successive iterations of the above method, the homography estimates are refined. Once convergence is reached, the per-frame estimated warps are used to assemble the final panorama.

Re-sample: The entire sequence of binary frames is resampled and grouped into subsets that are later aligned and merged. We use midpoint sampling as the grouping strategy. Given a group size of , during the first iteration, we split the binary frames into non-overlapping groups. A single frame within each group is chosen to be the reference frame whose warp is later estimated in the "Locate" step. Initially, we choose the center frame of each group to be the reference frame. In subsequent iterations, the binary frame sequence is *re-sampled* to create new groups consisting of frames that are chosen such that they are centered between the previous iteration's groups. This introduces overlapping groups and ensures a progressively denser sampling of the motion trajectory. Fig. 6 illustrates what happens if we omit this crucial step—regions where the motion trajectory is more complex exhibit blur and ghosting artifacts.

This stratified re-sampling approach is crucial for dealing with the motion blur and noise tradeoff. The number of frames ( ) per group plays an important role in dealing with this tradeoff: a larger value of helps counteract Poisson noise but also causes motion blur. In practice, if SPAD

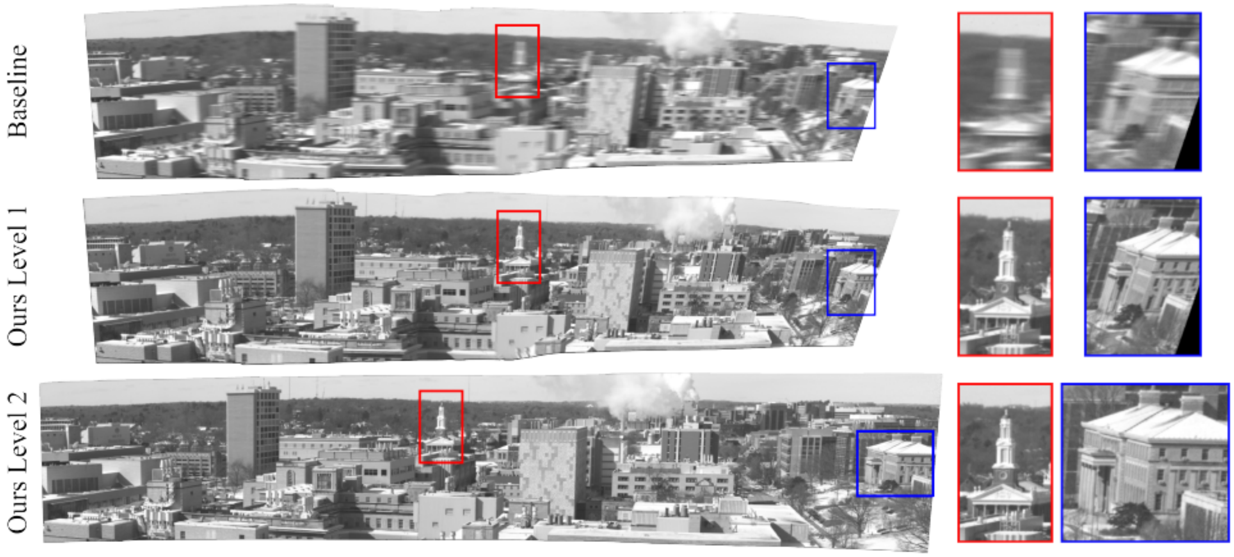

Figure 5. **Multi-level refinement of panorama:** We show a panorama created by naively averaging adjacent frames, in groups of 1000 (baseline), and two iterations of our method also using a group size of $m = 1000$ (ours, 1 and 2). Blurrier regions such as the tower (red inset) become sharper and the building (blue inset) is reconstructed without distortion after only two iterations.

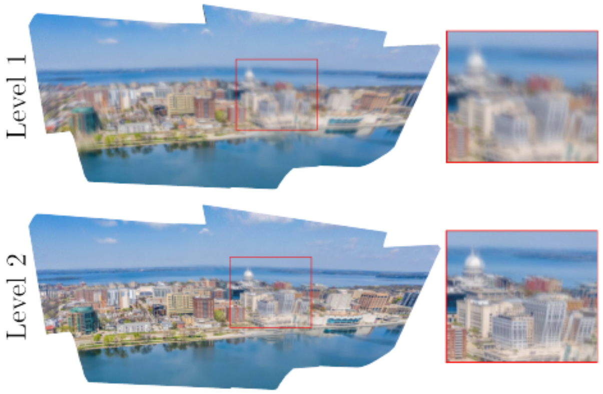

Figure 6. **Constant number of groups without resampling:** Without adding new virtual exposures at each level, the motion trajectory estimates are unreliable, resulting in blur and ghosting artifacts.

binary frames are available at $\sim$ 100 kHz, setting $m \approx$250-750 achieves high-quality results across different motion speeds and light levels, with higher values better suited for extremely low light, and lower values for fast motion. See supplementary material for details on the asymptotic behavior of this grouping policy, and the impact of the choice of the reference frame for each group.

Merge: The frames within each group are warped and merged. The warp operation is applied locally within each group. Applying these warps locally with respect to the group's center frame (instead of a global reference frame) is critical; it ensures that the frames within each group only need to be warped by small amounts. The warped frames are then merged using Eq. (3) and tone-mapped to sRGB.

Locate: The pairwise warps between merged frames are estimated using an off-the-shelf method. Any drift introduced in this step is corrected during subsequent iterations.

Interpolate: The "Locate" step estimates homographies across groups of ***merged*** binary frames. In this step, we interpolate these estimated homography matrices across time to get the fine-scale warps later used to warp ***individual binary frames***. A natural way to interpolate homographies is using a geodesic interpolation [12]. In practice, an extended Lucas-Kanade formulation [3] is more robust since it avoids computing matrix inverses and is numerically more stable. We perform cubic interpolation on the eight free parameters, p_i, of the 3×3 homography matrix:

$$H = \begin{bmatrix} 1+p_1 & p_3 & p_5 \\ p_2 & 1+p_4 & p_6 \\ p_7 & p_8 & 1 \end{bmatrix}. \qquad (4)$$

The resulting interpolated homographies are able to resolve extremely high-speed motion at the granularity of individual binary frames ($\sim$ 100 kHz), thus significantly mitigating the noise-blur tradeoff.

Computational Considerations: Instead of the proposed re-sampling method, one could create virtual exposures centered around each time instance, in a sliding window manner, and register those. This would not only be computationally expensive, as one would need to construct (and extract features from) hundreds of thousands of aggregate frames, but would also produce blurry results as the registration process will be sensitive to the blur introduced in each aggregate frame. The iterative nature of the proposed method allows for progressively better localization as at each iteration, the merged frames are progressively less blurry. This is shown in Fig. 3(d) as the error bars on existing points get smaller in the second iteration. Thus, the naive sliding-window approach would also need to be iter-

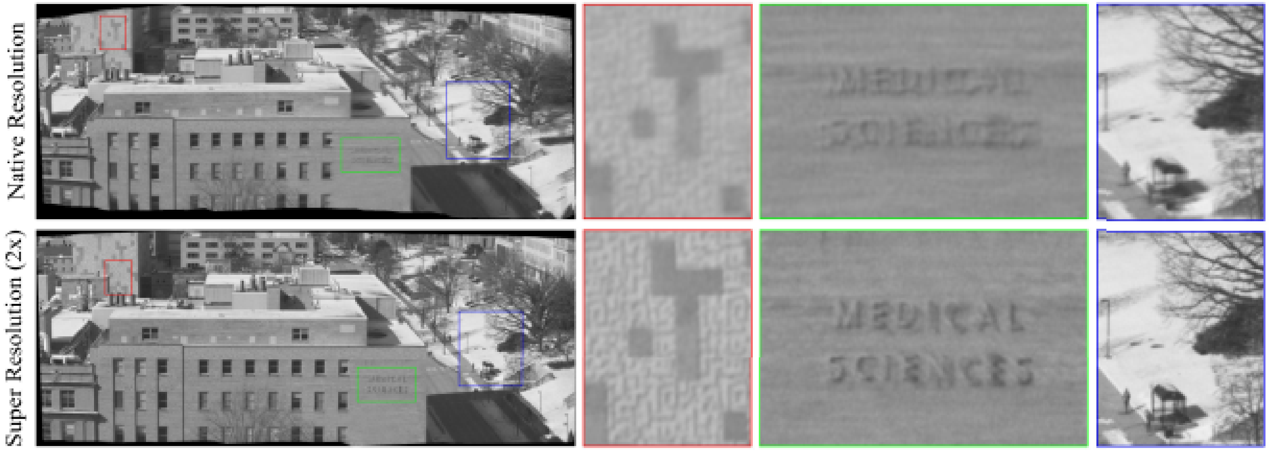

Figure 7. **Super-resolution on experimental data:** By interpolating homographies over additional virtual exposures, our method can super-resolve the sensor's native resolution ( ) by . Details such as text on the building (red, green insets) and finer structures such as tree branches (blue inset) are super-resolved.

ated upon, yielding a complexity of $\mathcal{O}($), for processing frames over windows of size (assuming the number of iterations stays constant). In contrast, our iterative method has an asymptotic runtime of $\mathcal{O}($), providing significant speedup over the sliding-window approach.

5. Experiments and Results

5.1. Setup: Simulations and Hardware

We demonstrate our technique in simulation and through real-world experiments using a SPAD hardware prototype.

Simulation Details: We simulate a SPAD array capturing a panoramic scene by starting with high-resolution panoramic images downloaded from the internet. We create camera trajectories across the scene such that the SPAD's field of view (FOV) sees only a small portion of the panorama at a time. At each time instant of the trajectory, we simulate a binary frame from the FOV of the ground truth image by first undoing the sRGB tone mapping to obtain linear intensity estimates, and then applying Eq. (2) to simulate the binary photon stream. RGB images are simulated by averaging the ground truth linear intensities over a certain exposure and adding Gaussian noise [14].

Hardware Prototype: For real experiments, we use the SwissSPAD [39] to capture binary frames (Fig. 1(a)). The sensor has a usable resolution of 254 496 pixels. It does not have micro-lenses, or a color filter array, and the fill factor is 10 5% with 16 8 pixel pitch. Despite these limitations, it is capable of capturing binary frames at 100 kHz.

Implementation Details: We use OpenCV's registration algorithm based on SIFT and RANSAC homography fitting to match virtual exposures. Our implementation takes roughly ten minutes per iteration to process 100 frames. While factors such as resolution and window size ( ) will affect runtime, our implementation is throttled by the underlying registration algorithm which recomputes features at every level. Further optimizations and feature caching would greatly improve runtime.

5.2. Results and Capabilities

Fast Motion Recovery: Fig. 4 (left) shows an example panorama reconstruction in a challenging scenario where the camera moves along an arbitrary trajectory across the full FOV. Conventional panorama reconstruction techniques fail, even if there is sufficient light in the scene, because individual frames suffer from extreme motion blur, making it difficult to find reliable feature matches. By iteratively creating staggered virtual exposures, our method can resolve motion that would otherwise be entirely contained within a single exposure of a conventional camera image. Observe that our approach is capable of recovering a near-perfect motion trajectory, which, as seen in the zoomed-in crops, further enables high-fidelity scene reconstruction.

Low Light Robustness: Fig. 4 (center) shows the challenging scenario where the camera pans across a dark scene. Here, the conventional RGB method fails because no matches are found in the extremely noisy RGB frames. The situation gets worse in Fig. 4 (right) where low light is accompanied by extremely fast camera motion. In this extremely low flux regime, the RGB image is dominated by read noise and causes feature registration to fail. In contrast, our approach produces high-quality reconstructions.

Globally Consistent Matching: A key issue when global motion is estimated piece by piece is that of drift: any error in the pairwise registration process accumulates over time. This phenomenon is clearly visible in the RGB panorama in Fig. 4 (left)—not only does the estimated motion trajectory (red) drift away from the ground truth (black), but the panorama gets stretched as compared to the ground truth panorama outline (black dotted line). This drift gets cor-

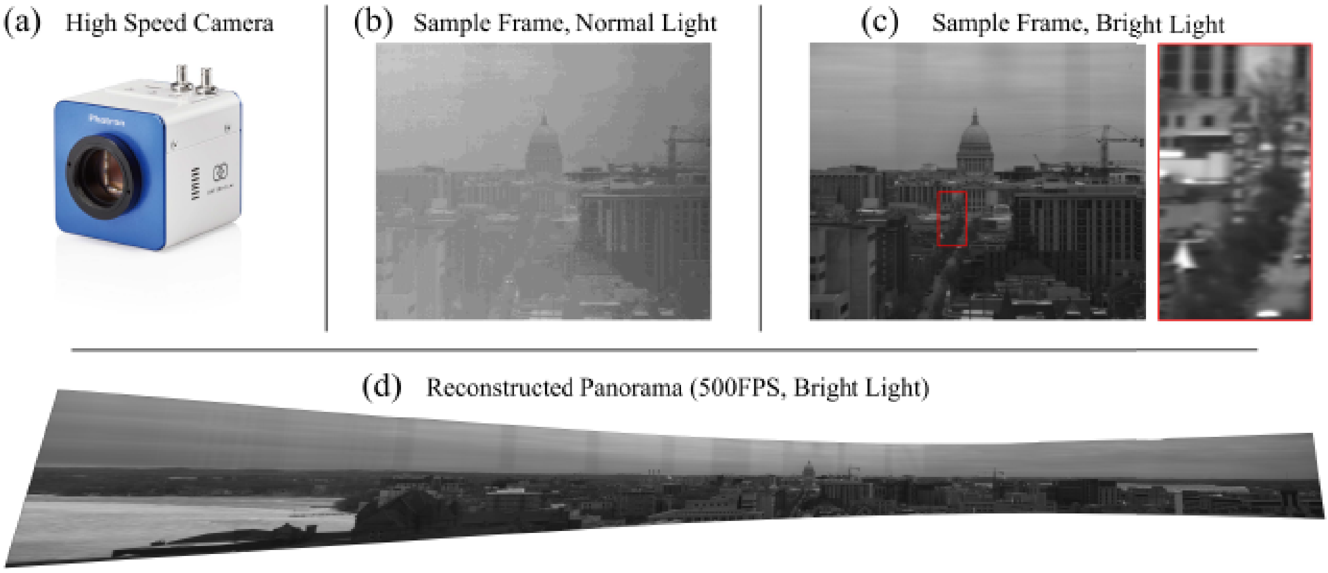
(a) High Speed Camera
(b) Sample Frame, Normal Light
(c) Sample Frame, Bright Light
(d) Reconstructed Panorama (500FPS, Bright Light)

Figure 8. **Application to conventional high-speed cameras:** (a) We use Photron's Infinicam as our high-speed camera mounted with similar optics as our SPAD camera prototype. (b) Individual image frames from the commercial high-speed camera are extremely noisy and show compression artifacts even when capturing frames at slower motion and running at fps. Panorama reconstruction using such frames fails due to a lack of reliable feature matches across frames. (c) Using a lens with a larger aperture ($f/$. instead of $f/$.) the frame quality dramatically improves. However, despite the Infinicam's high resolution of and slower motion, the text on the sign is still blurry compared to the SPAD camera result in Fig. 1(c). (d) In this higher flux regime, we can reconstruct a panorama (the vertical band artifacts are reflections from the window), albeit a blurry one as each individual frame is blurry. This demonstrates that while our method works with high-speed cameras, it is ideally suited for single-photon imaging.

rected with the proposed method due to the iterative refinement of both the motion estimate and the resulting reconstruction. Fig. 5 demonstrates such iterative refinement using real SPAD captures with our hardware prototype. As we increase the number of iterations, the global shape of the panorama gets rectified. The progressive improvement of individual aggregate frames is shown in Fig. 1(b).

Super-Resolution and Efficient Registration: Due to dense temporal sampling, and the resulting fine-grained homography estimates, the proposed method enables super-resolution in the reconstructed panoramas. This is achieved by applying a scaling transform to the estimated homographies before the merging step. This scaling transform stretches the grid of pixels into a larger grid, resulting in super-resolution. Further, to save on compute and memory costs, this scaling factor can be gradually introduced across iterations. For example, if the goal is to super-resolve by a scale of 4 , we could scale the estimated warps by a factor of two over two iterations. It is also possible to use scaling factors that are smaller than one in the initial iterations of the pipeline. This can be done to create large-scale panoramas, such as the one in Fig. 1(c), while maintaining low computational and memory footprints. An experimental result with sub-pixel registration is shown in Fig. 7.

High Dynamic Range: Single photon cameras have recently been demonstrated to have high dynamic range (HDR) capabilities [18, 17, 2]. By performing high-accuracy homography estimation and registration, the proposed method is able to merge a large number of binary measurements from a given scene point, thus achieving HDR. Fig. 9 shows a real-world example of HDR on a sequence of binary frames captured at night.

Extension to High-Speed Cameras: The stratified resampling approach can be extended to other high-speed imaging modalities that allow fast sampling. The only assumption is that individual frames contain minimal motion blur and can be combined in such a way that boosts SNR and allows for feature matching. We demonstrate this using a commercially available high-speed camera (Photron Infinicam) in Fig. 8(a). This camera captures 1000 fps at its full resolution of 1246 1024, with higher frame rates available for lower resolutions. All frames are compressed on the camera and access to raw frames is not possible. If the scene is too dark all useful information will be corrupted by compression artifacts, further impeding the creation of virtual exposures as compression and frame aggregation do not commute. This phenomenon can be seen in Fig. 8(b), it occurs when using the same optical setup as the one used with our SPAD prototype (75mm focal length, 5 6), even with a much longer exposure time (which corresponds to 500fps). To get sufficient signal to overcome these limitations we increased the aperture to allow the camera to cap-

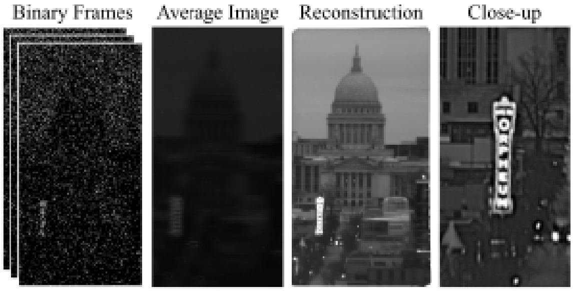

Figure 9. **High dynamic range image stabilization:** By aligning a large number of extremely dark binary frames, we can stabilize the high-frequency camera shake which causes the average image to be washed out, and faithfully reconstruct this night-time scene, recovering detail in both the dark and bright regions.

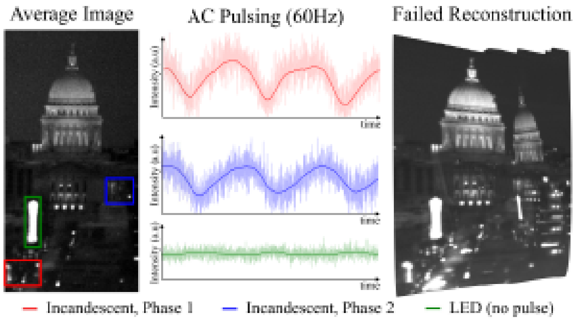

Figure 10. **Artificial flicker, a failure case:** The flickering of street lights due to the power grid's alternating current can be observed when using high frame rates. This flickering violates the brightness constancy assumption used by most registration algorithms and thus leads to a reconstruction failure.

ture 4 more light. Fig. 8(c) shows a sample frame captured at 500fps with this new setup. Despite the slower motion (2 slower than seen in Fig. 1(c)) and the much larger resolution of the Infinicam, scene details such as text appear blurred. These frames can be assembled into a larger panorama using our algorithm (Fig. 8(c)), albeit, with some residual blur.

Although high-speed sampling enables the creation of virtual exposures, the reconstruction quality deteriorates in challenging conditions due to both the high read noise and the relatively lower sampling rate of high-speed cameras, thus suggesting that the proposed techniques are ideally suited for single-photon imaging.

6. Limitations and Future Outlook

Brightness Constancy Failure: As seen in Fig. 10, our hardware prototype is fast enough to detect and even measure the flickering of artificial lighting due to the electric grid. While this can be used to categorize light sources and measure the grid's load [36], it can also cause the underlying registration algorithm to fail as the brightness constancy assumption is violated. This may be mitigated using spatially varying virtual exposures (longer exposures in regions with non-constant brightness).

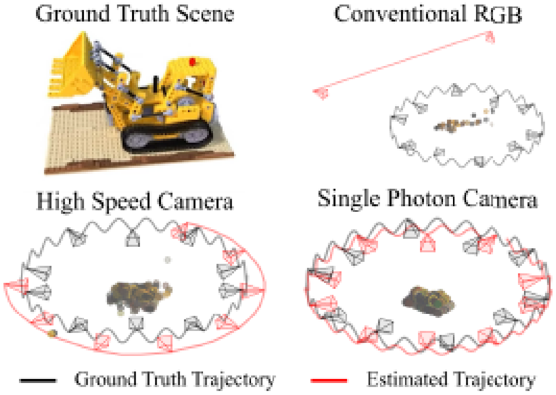

Figure 11. **High-speed 3D pose estimation using COLMAP:** Naively averaging adjacent binary frames from a single photon camera enables better 3D reconstruction and pose recovery than what is possible with a high-speed or conventional camera. A promising future research direction is to apply the stratified resampling ideas put forward in this paper to pose estimation.

Beyond the Planar Scene Assumption: The applications shown in this paper use a homography-based motion model which characterizes global camera motion under the planar scene assumption. An important next step is to extend these ideas to 3D scenes with global motion, such as 6-DoF pose estimation. Fig. 11 shows that the high temporal sampling provided by using single photon cameras improves COLMAP's [34] pose estimation and sparse reconstruction when simply averaging neighboring frames. How can this initial pose estimate be used to refine our reconstruction? We discuss two possible 3D-consistent aggregation methods below which are promising future research directions.

Implicit 3D Representations: One way to perform the stratified 3D aggregation of binary frames needed to converge to a high-quality pose estimate would be to adapt the recent work done on implicit representations [29, 30, 37] to work with binary images. However numerous challenges remain such as i) how to adapt the rendering model to non-differentiable image data, ii) how to train and update this representation in an online manner, and iii) how the stochastic nature of binary frames will affect the creation and refinement of a globally consistent 3D representation.

Dense Motion Models: A more general motion model, such as optical flow, could be applied pixel-wise to allow for robust, 3D-consistent binary frame aggregation without resorting to a three-dimensional representation. A recent work [27] has applied optical flow to binary frames with the goal of reconstructing high-quality images as opposed to recovering 3D structure and motion. This method estimates per-frame optical flow in a single shot, meaning that applying our stratified algorithm in this scenario could lead to improved reconstruction and fine-grain pose estimates.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building Rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79, Sept. 2009. ISSN: 2380-7504. 1

[2] Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Dynamic range extension for photon counting arrays. *Optics Express*, 26(17):22234, Aug. 2018. 8

[3] Simon Baker and Iain Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision*, 56(3):221–255, Feb. 2004. 2, 6

[4] Simon Baker, Raju S. Patil, G. Cheung, and I. Matthews. Lucas-Kanade 20 Years On: Part 5. 2004. 3

[5] Matthew Brown and David G. Lowe. Automatic Panoramic Image Stitching using Invariant Features. *International Journal of Computer Vision*, 74(1):59–73, Aug. 2007. 3, 1

[6] 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, 2019. 1

[7] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 1

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, Aug. 2007. Conference Name: IEEE Transactions on Image Processing. 3

[9] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct Sparse Odometry, Oct. 2016. arXiv:1607.02565 [cs]. 3

[10] 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. IEEE, 2005. 3

[11] Eric R Fossum. The quanta image sensor (qis): concepts and challenges. In *Computational Optical Sensing and Imaging*, page JTuE1. Optica Publishing Group, 2011. 3

[12] Pasqualina Fragneto, Andrea Fusiello, Beatrice Rossi, Luca Magri, and Matteo Ruffini. Uncalibrated View Synthesis with Homography Interpolation. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 270–277, Zurich, Switzerland, Oct. 2012. IEEE. 6

[13] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, Jan. 2022. 3

[14] Samuel W. Hasinoff, Fredo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560, San Francisco, CA, USA, June 2010. IEEE. 1, 7

[15] 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. 3

[16] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided Direct Sparse Odometry, Apr. 2022. arXiv:2204.07640 [cs]. 3

[17] Atul Ingle, Trevor Seets, Mauro Buttafava, Shantanu Gupta, Alberto Tosi, Mohit Gupta, and Andreas Velten. Passive Inter-Photon Imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8581–8591, Nashville, TN, USA, June 2021. IEEE. 8

[18] Atul Ingle, Andreas Velten, and Mohit Gupta. High Flux Passive Imaging With Single-Photon Sensors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6753–6762, Long Beach, CA, USA, June 2019. IEEE. 8

[19] Information technology — Coding of audio-visual objects — Part 10: Advanced Video Coding. Standard, International Organization for Standardization, Mar. 2022. 1

[20] 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. Conference Name: IEEE Access. 2, 1

[21] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew Davison. Simultaneous Mosaicing and Tracking with an Event Camera. In *Proceedings of the British Machine Vision Conference 2014*, pages 26.1–26.12, Nottingham, 2014. British Machine Vision Association. 3

[22] Ken-Yi Lee, Yung-Yu Chuang, Bing-Yu Chen, and Ming Ouhyoung. Video stabilization using robust feature trajectories. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1397–1404, Sept. 2009. ISSN: 2380-7504. 1

[23] Peidong Liu, Xingxing Zuo, Viktor Larsson, and Marc Pollefeys. MBA-VO: Motion Blur Aware Visual Odometry, Mar. 2021. arXiv:2103.13684 [cs]. 3

[24] Yinyang Liu, Xiaobin Xu, and Feixiang Li. Image Feature Matching Based on Deep Learning. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1752–1756, Dec. 2018. 3

[25] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004. 2

[26] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, Aug. 1981. Morgan Kaufmann Publishers Inc. 2

[27] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics*, 39(4), Aug. 2020. 2, 3, 9, 1

[28] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections, Jan. 2021. arXiv:2008.02268 [cs]. 1

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Aug. 2020. arXiv:2003.08934 [cs]. 9

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, July 2022. 9

[31] 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, 2020. 2

[32] K Morimoto, J Iwata, M Shinohara, H Sekine, A Abdelghafar, H Tsuchiya, Y Kuroda, K Tojima, W Endo, Y Maehashi, et al. 3.2 megapixel 3d-stacked charge focusing spad for low-light imaging and depth sensing. In *2021 IEEE International Electron Devices Meeting (IEDM)*, pages 20–2. IEEE, 2021. 2

[33] Sidheswar Routray, Arun Kumar Ray, and Chandrabhanu Mishra. Analysis of various image feature extraction methods against noisy image: SIFT, SURF and HOG. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–5, Feb. 2017. 3

[34] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, Las Vegas, NV, USA, June 2016. IEEE. 1, 3, 9

[35] Trevor Seets, Atul Ingle, Martin Laurenzis, and Andreas Velten. Motion Adaptive Deblurring with Single-Photon Cameras, Dec. 2020. arXiv:2012.07931 [eess]. 2, 1

[36] Mark Sheinin, Yoav Y Schechner, and Kiriakos N Kutulakos. Computational imaging on the electric grid. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2017. 9

[37] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A Modular Framework for Neural Radiance Field Development, Feb. 2023. arXiv:2302.04264 [cs]. 9

[38] Matias Tassano, Julie Delon, and Thomas Veit. FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1351–1360, Seattle, WA, USA, June 2020. IEEE. 3

[39] Arin Can Ulku, Claudio Bruschini, Ivan Michel Antolović, 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. Conference Name: IEEE Journal of Selected Topics in Quantum Electronics. 7

[40] Feng Yang, Yue M Lu, Luciano Sbaiz, and Martin Vetterli. Bits from photons: Oversampled image acquisition using binary poisson statistics. *IEEE Transactions on image processing*, 21(4):1421–1436, 2011. 4

[41] Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep Lucas-Kanade Homography for Multimodal Image Alignment, Apr. 2021. arXiv:2104.11693 [cs]. 3