Burst Vision Using Single-Photon Cameras – Supplementary Technical Report

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

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

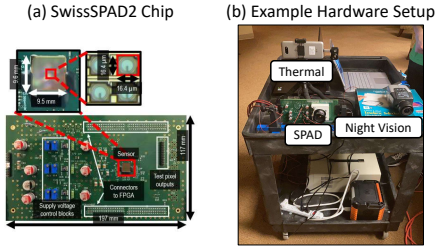

Figure 1. **Hardware setup. (a)** We use SwissSPAD2 [5] to capture binary sequences. Image reproduced from [5] with permission. **(b)** Example hardware setup with a SPAD camera, a thermal camera and a night vision camera on the same cart. This enables capturing synchronized videos where the cameras share the same trajectories.

1. Hardware Setup

We capture binary sequences with a SwissSPAD2 [5] (Fig. 1(a)) to demonstrate the capability of quanta vision. We read off the binary pixel values for half of the SPAD array, which gives a spatial resolution of 512×256. The camera can capture up to 130,000 binary frames at a maximum frame rate of 96.8kfps which are stored on-chip for offline processing. To capture long sequences for video tasks such as SLAM, we also developed an FPGA design that continuously streams the binary frames to a PC with an SSD via USB 3.0 while capturing, which supports arbitrarily long binary sequences at up to 17.3kfps. We also capture synchronized videos from commodity cameras such as DSLR, cellphone, night vision cameras and thermal cameras. Fig. 1(b) shows an example setup where we place the SPAD camera, a thermal camera and a night vision camera on a cart. The cart can then be moved during capturing such that the three cameras share the same trajectory.

2. Additional Results

This section discusses quanta vision sequences that are not included in the main paper due to space constraints.

Resolving blur-noise trade-off using burst vision. Fig. 2 shows a more detailed visualization on how the blur-noise trade-off is resolved by quanta vision (Fig. 2 in the main paper). We capture a binary sequence in a garage during nighttime, with all the lights turned off. This results in an extremely low light level, as shown by the single binary frames. We run YOLOv3 [4] on naive average and burst reconstruction respectively. Naive average results suffer from the blur-noise trade-off. The images are either too noisy or too blurry, and object detection fails on all of them. Explicit burst vision is able to solve the trade-off by compensating for the motion, generating images with enough signal when a sufficiently long integration window is used.

Recovering visual features for camera tracking. Many vision tasks involve robust recovery and tracking of structural features, which is sensitive to noise and blur. Fig. 3 shows an indoor, handheld sequence for simultaneous localization and mapping (SLAM). We run ORB-SLAM3 [1] on reconstructed images offline. We switch a lamp on (7.5lx) and off (0.02lx) during the sequence as shown in the synchronized images from a static DSLR (Canon EOS Rebel T5i). To show how challenging this lighting condition is, we also run an ArCore demo app[1] on a cellphone and ORB-SLAM3 on a night vision camera side-by-side. At first the lamp is on, and all three SLAM systems work properly. When the lamp is turned off, the room becomes extremely dark such that the DSLR gives an almost completely black image with a few noise spots. Cellphone and night vision camera lost tracking due to low image quality. SPAD still recovers sufficient features for estimating camera motion.

Recovering spatial details under strobing light. Fig. 4 shows an example of scanning a fast moving QR code under a rapidly flashing light (about 2Hz). The DSLR images are noisy and blurry, especially when the flash is off. The night vision camera captures noisy and blurry images when the flash is off, and gets overexposed images when the flash is on as it fails to adapt to high lighting level immediately. The thermal camera cannot capture the QR code which is

[1]ARCore Elements: https://play.google.com/store/apps/details?id=com.google.ar.unity.ddelements

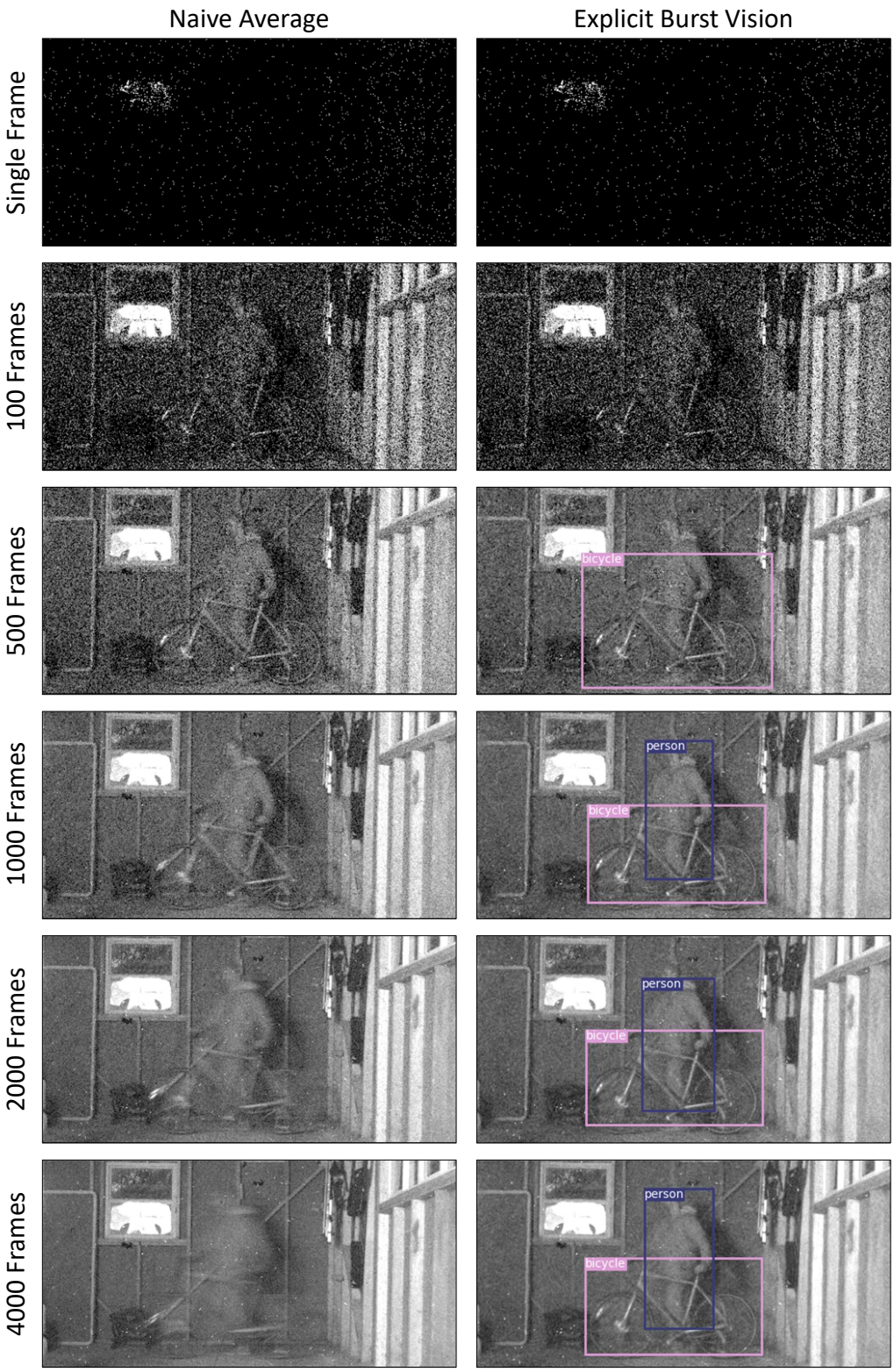

Figure 2. **Resolving blur-noise trade-off using burst vision.** We capture a binary sequence in a garage during nighttime, with all the lights off. **(Left)** Naive average images suffer from the blur-noise trade-off and are either too noisy or too blurred. Object detection fails on all the images. **(Right)** Burst vision resolves the blur-noise trade-off and gives high-quality images for successful object detection when a sufficiently long integration window is used ($\geq$ 1000 frames, 10ms software-defined exposure time).

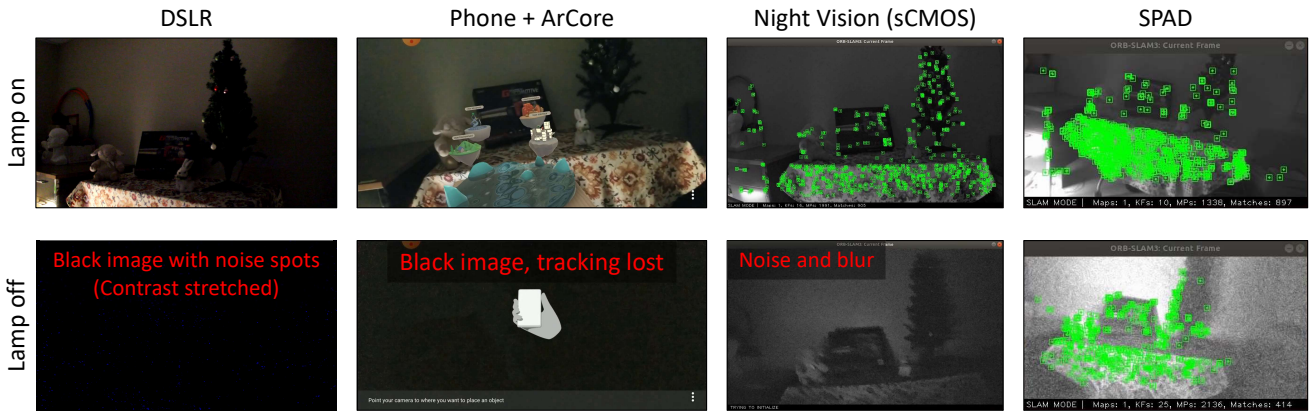

Figure 3. **Recovering visual features for camera tracking.** Handheld camera tracking for an indoor scene. The scene is illuminated by a lamp which switches between on (7.5lx) and off (0.02lx). We run ORB-SLAM3 on the reconstructed SPAD images offline. To show how challenging the lighting is, we show DSLR images from a static side view. We also run a demo ArCore app on a Galaxy S8 and ORB-SLAM3 on a night vision cameras, which are rigidly fixed to the SPAD. **(Top)** When the lamp is on, all cameras see clear images and all the SLAM systems work. **(Bottom)** When the lamp is off, the DSLR image looks almost black with only noise spots. The cellphone and night vision camera fail to track due to low image quality. The SPAD images contain sufficient features for reliable tracking.

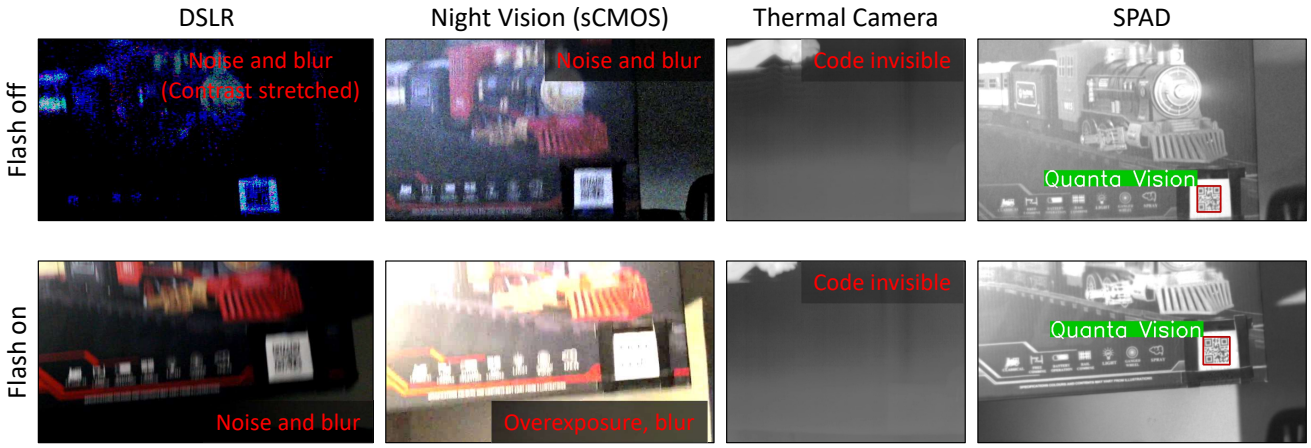

Figure 4. **Recovering spatial details under strobing light.** A QR code moves fast under a rapidly flashing light (about 2Hz). DSLR images are noisy and blurry. The night vision images are noisy and blurry when the flash is off, and overexposed when the flash is on. The thermal camera cannot capture the QR code in the visible light range. The SPAD captures high-quality images in both lighting conditions for robust QR code decoding.

only discernible in visible light range. The SPAD camera captures high-SNR, low-blur images in both lighting conditions, and the QR code is correctly decoded in both cases.

Background subtraction in low light. In addition to object detection, tracking, QR decoding, scene text detection and SLAM, we also evaluate the quanta vision for other tasks which are not included in the main paper due to page constraints. Fig. 5 shows a person walking and running in a dark room (same sequence as Fig. 7 in the main paper). We run a traditional PCA-based background subtraction algorithm [7] on the naive average and burst-reconstructed images. To help understand what the scene looks like, we first show the burst reconstructions in the first row, and then show the background-subtracted images for both naive average and burst reconstruction. The foreground mask from naive average suffers from incomplete body parts, while the result from burst reconstruction is more accurate.

Fast-moving human pose estimation. Fig. 6 shows a person doing jumping jacks in a dark room. This is a challenging sequence because of the fast motion and cluttered background. We run a learning-based human pose estimator [3] on the images. Naive average cannot estimate the human pose correctly. Burst reconstruction recovers most of the body parts, with the exception of arms in some of the images because arms are thin and move extremely fast.

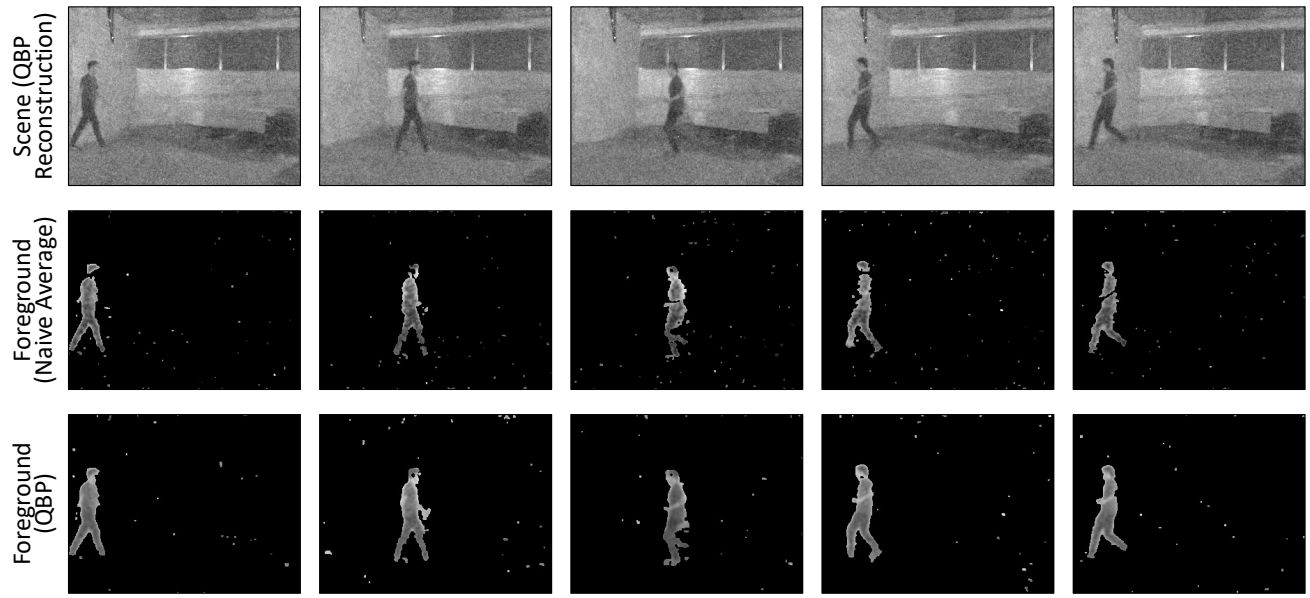

Figure 5. **Background subtraction in low light.** A person walks and runs (foreground) in a dark room with lights off (background). **(Top)** We show burst reconstructions as visualizations of the scene. **(Middle)** Foreground mask from naive average images does not completely cover the body parts. **(Bottom)** Foreground mask from burst reconstruction better covers the entire body.

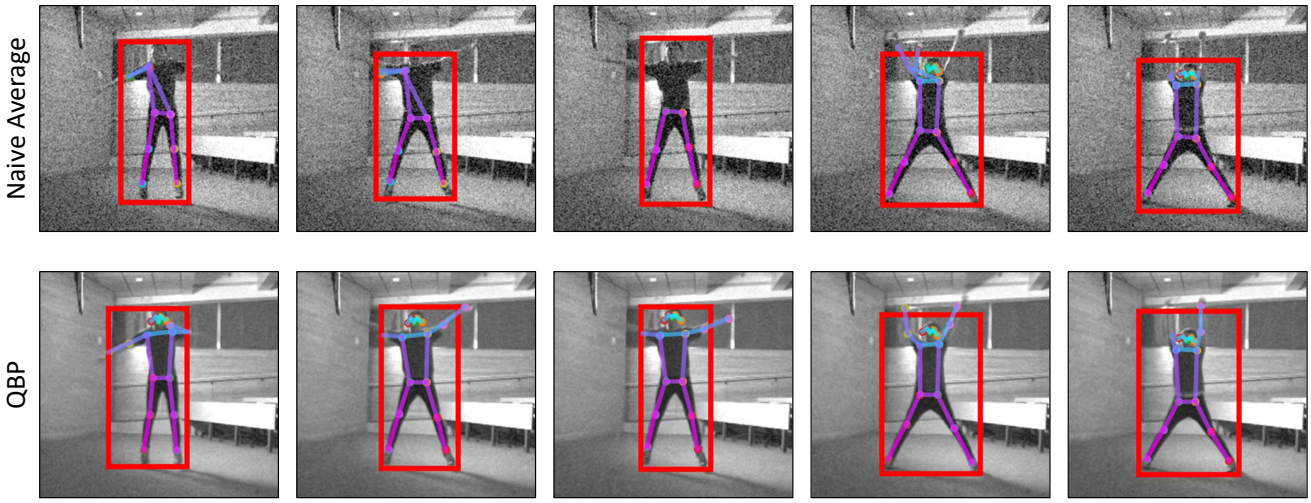

Figure 6. **Fast-moving human pose estimation.** A person doing jumping jacks in a dark room. **(Top)** Naive average fails to estimate all the body parts correctly. **(Bottom)** Burst reconstruction is able to recover poses with high accuracy, despite recovery of the arms being challenging because they are thin and move extremely fast.

Action recognition in low light. Fig. 7 shows a person clapping and waving hands in a dark room. We run a state-of-the-art action recognition algorithm [2] on a sliding window of 32 frames across the reconstructed sequence. Naive average images are noisy and blurry (*e.g.* waving hand). As a result, the action is not correct recognized. Burst reconstruction generates better quality images, where the clapping and waving action is correctly recognized.

Face-moving face detection in low light. Fig. 8 demonstrates running face detection [6] on a jumping person, which is challenging due to the dark environment and the fast vertical movement. With burst vision, it is still able to detect the faces from the reconstructed images.

References

[1] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and

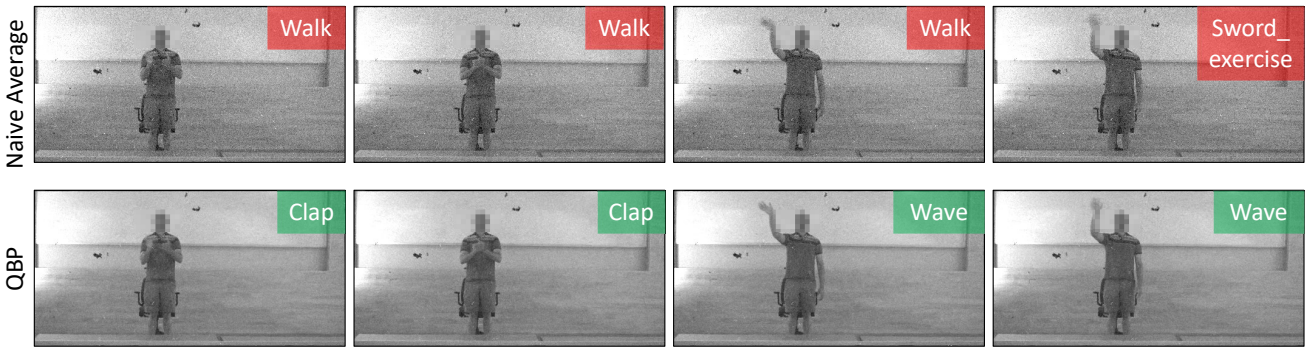

Figure 7. **Action recognition in low light.** A person clapping and waving hands in a dark room. **(Top)** Naive average fails to recognize the action due to heavy noise and blur. **(Top)** Burst reconstruction generates higher quality images which enable correct recognition of the actions. Faces are blurred for anonymity.

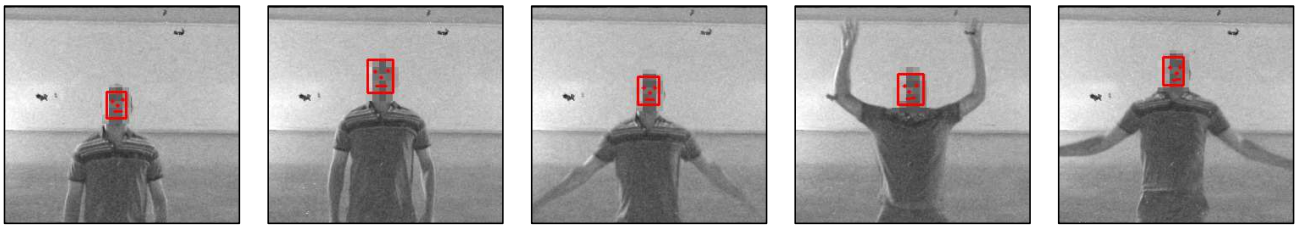

Figure 8. **Fast-moving face detection in low light.** We run face detection on a jumping person. This is a challenging task due to the low light level and fast movement of the person. With burst vision, the faces are successfully detected. Faces are blurred for anonymity.

Multi-Map SLAM. *IEEE Transactions on Robotics*, pages 1–17, 2021.

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

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

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

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

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

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