A Bit Too Much? High Speed Imaging from Sparse Photon Counts

Paramanand Chandramouli[1], Samuel Burri[2], Claudio Bruschini[2], Edoardo Charbon[2], and Andreas Kolb[1]
[1]University of Siegen, Germany
[2]Swiss Federal Institute of Technology (EPFL), Neuchâtel, Switzerland

Recent advances in photographic sensing technologies have made it possible to achieve light detection in terms of a single photon. Photon counting sensors are being increasingly used in many diverse applications. We address the problem of jointly recovering spatial and temporal scene radiance from very few photon counts. Our ConvNet-based scheme effectively combines spatial and temporal information present in measurements to reduce noise. We demonstrate that using our method one can acquire videos at a high frame rate and still achieve good quality signal-to-noise ratio. Experiments show that the proposed scheme performs quite well in different challenging scenarios while the existing approaches are unable to handle them.

Index Terms—**Photon Counting Sensors, SPAD Imaging, Convolutional Neural Network, High Speed Imaging**

I. Introduction

We address the problem of imaging fast dynamic scenes through single photon counting sensors [1]. Single photon avalanche diode (SPAD) detectors are endowed with the ability of photon counting and time-stamping. They are getting increasingly popular for a variety of imaging applications [2]–[6]. We attempt to the enhance the fast motion capture capability of SPAD sensors by developing a robust video recovery algorithm. Consider the example shown in Fig. 1. An analog oscilloscope with a traversing sinusoid was imaged by a SwissSPAD camera [7]. Fig. 1 (a), shows a frame from the captured video obtained at 156k frames per second (fps). The value of each pixel in this frame is a binary number wherein positive values indicate the detection of a photon. Note that in a typical DSLR camera, the number of photons captured is of the order of thousands of photons per pixel [3], [8] while the frame in Fig. 1 (a) shows the detection of *only one photon*. Due to the extremely low photon count, one can hardly infer any structure present in the scene. One could average consecutive frames to reduce the effect of noise while sacrificing the temporal resolution (Fig. 1 (c)). To preserve temporal resolution as well as recover accurate scene reflectance, we develop a convolutional neural network (CNN) that takes the low-photon count sequence as input and generates a high-photon count estimate at the *same frame rate*. Our scheme effectively combines the spatio-temporal information present in the input sequence for video recovery. The resultant frame from our method is shown in Fig. 1 (b). Note that in Figs 1 (a) and (b), one can observe the localization of the sinusoidal wave while in Fig. 1 (c), the temporal information is lost. In Fig. 1 (b), we also observe that the details of the static regions have been recovered quite well.

Previously, in the context of time-correlated SPAD imaging, regularization-based approaches have been developed to reconstruct scene reflectivity [3], [9]. For oversampled binary observations, image reconstruction algorithms such as [10] are applicable. In this paper, our objective is to jointly recover spatial and temporal variations of radiance in dynamic scenes without spatial oversampling. Existing video denoising schemes devise methods for combining local and non-local structures present across space and time [11]–[13]. In extremely noisy scenarios, explicitly determining such information does not work well. Instead of "hand-crafted" approaches to combine structural information, our method uses convolutional neural networks consisting of 3D filtering across the spatio-temporal volume [14]. By accumulating a set of binary frames, one can obtain video sequences with lesser noise and reduced frame rate. In this paper, we address different scenarios in which different number of the binary frames are combined. Although we show the application of our method on SwissSPAD cameras, our scheme can be applied to any other photon counting or binary imaging sensors [15], [16]. High speed consumer cameras typically require significantly bright illumination [17]. In contrast, we do not use any high intensity illumination and operate in normal lighting conditions.

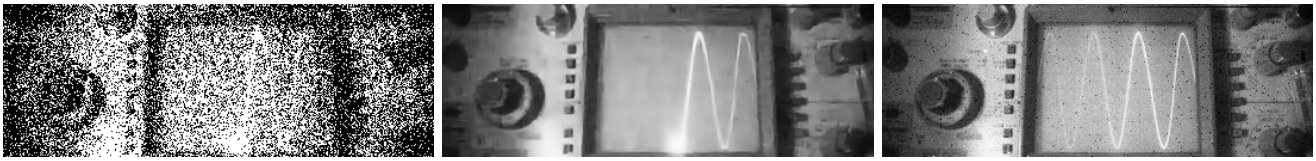
A. Related work

Since our work is related to SPAD imaging, photon counting sensors and video denoising, we briefly discuss relevant prior works in these topics.

SPAD-based imaging SPAD sensors are photodetectors in which photon radiation can be detected from the resulting large avalanche currents. SPAD sensor arrays are capable of photon counting at a high speed with high timing resolution and are useful in a variety of applications such as fluorescence lifetime imaging microscopy (FLIM), positron emission tomography (PET), time-of-flight imaging etc. [2], [3], [18]. Recently, Burri et al. developed a SPAD array known as SwissSPAD [7]. The SwissSPAD is fabricated in a high-voltage CMOS process and features a large 512×128 array with global gating. In this paper, we use data from different SwissSPAD sensor arrays for demonstrating our high speed video recovery scheme.

Recent works in range imaging through SPAD sensors include [3], [19], [20]. SPAD cameras have been used to perform challenging tasks such as transient imaging [4], [5], [21]–[23],

978-1-7281-3263-1/$31.00 ©2019 IEEE



(a) (b) (c)

Fig. 1. Imaging wave propagation in an oscilloscope: (a) A 1-bit frame from the input to our algorithm captured at 156000 fps. (b) Corresponding resultant frame of the proposed scheme. (c) Average of 255 frames with the frame shown in (a) as the center (the dark pixels are due to sensor defects). Kindly refer to the supplementary material for the complete video.

non-line-of sight imaging [6], [24]–[26] and imaging through fog [27].

Quanta Imaging Closely related to the topic of single photon counting is that of Binary/Quanta image sensors [1], [28]. These sensors were developed with the objective of shrinking the pixel pitch. In Quanta imaging, the densely defined sensors oversample the scene radiance to generate binary measurements [28], [29]. Image reconstruction schemes have been proposed for these sensors [10], [30]–[32]. The main difference between these reconstruction methods and our scheme is that these algorithms consider the availability of a higher number of samples to estimate the image intensity at a pixel location. These methods have mainly focused on recovering a static image. In contrast, our focus is on recovering videos.

Video denoising Video denoising has been widely studied for many years and different kinds of algorithms have been proposed. Because of the vastness of the topic, we restrict our discussion to some of the popular and recent works. When compared with applying image denoising on each frame, video denoising has an advantage because the high temporal coherence can be leveraged to make a better prediction. Consequently, denoising algorithms adopt different strategies such as non-local means [12], [33], motion estimation [34], [35], 3D dictionary representations [36] etc. Maggioni et al. [11] propose a denoising method popularly known as VBM4D and is based on the collaborative filtering scheme (VBM3D) [37]. They apply this filtering scheme to a stack of 3-D spatio-temporal volumes that are obtained through non-local grouping. Sutour et al., [12] develop a variational denoising scheme by adaptively combining nonlocal means with total variation on spatio-temporal volumes. Their method adapts to different noise levels. The authors in [38] propose a denoising method for the scenario of mixed noise (like Gaussian noise mixed with impulsive noise). They formulate the denoising problem as low-rank matrix completion problem. This work has been extended in [39] to avoid pre-detection of outliers.

Recently, techniques based on deep learning have been proposed for video denoising. Chen et al. [40] propose a deep recurrent neural network for video denoising. Their method does not assume a specific noise model and performs close to state-of-the art VBM4D scheme [11]. Xue et al. [41] propose task oriented flow to achieve a specific video processing objective such as denoising. Instead of trying to achieve a precise flow estimation, they train a model whose objective is to predict a motion field tailored for a specific task. In our experiments, we observe that video denoising methods fail in the scenario of SPAD imaging. This could be due to the fact that grouping of similar structures across the spatio-temporal volume fails when the number of photons in the observation is sparse.

Contributions The contributions of this paper can be summarized as follows: i) We address the problem of joint spatio-temporal radiance estimation from single photon counting sensors. ii) We develop a CNN-based video recovery scheme that can handle very high noise levels and still maintain the high frame rate iii) We devise methods to obtain appropriate real and synthetic datasets for training and evaluation. This data will be made available subsequently.

II. Image formation

In this section, we describe the image formation process for SwissSPAD cameras which have a globally gated sensor array [7]. The imaging model can also be applied to any other gated SPAD arrays [15], [16].

Each pixel in the SwissSPAD array is composed of a SPAD p-n junction suitably biased for enabling photon triggered avalanche. A one-bit counter is present at every pixel. The SPAD array has a global shutter in which all the pixels can be kept active for a duration as low as 3.8 ns. The pixel counter content is transferred from the sensor pixels via a fast digital readout which takes about 6.4 μs for transferring the contents of the whole array of 128×512 pixels. In effect, a 1-bit frame of size 128×512 indicating the detection of *one* photon will be read out in 6.4 μs resulting in a frame rate of 156kfps [7].

Due to the dark counts generated in a SPAD by thermal events, the number of photon counts per unit time follows a Poisson distribution [42]. The probability of k photon counts in unit time is given by

$$p_c\left(k\right) = \frac{\chi^k e^{-\chi}}{k!} \tag{1}$$

where χ is the expected value of counts and is related to the impinging count rate, dark count rate and photon detection efficiency [42]. Within a particular time frame, the SwissSPAD sensor array can only report whether one or more photons were detected. Consequently, the probability of recording a detection in one readout time is $P\left(\text{count} > 0\right) = 1 - e^{-\chi}$. Since the number of photons impinging at a pixel depends on the scene radiance corresponding to that pixel, one can conclude that the scene radiance is sensed non-linearly (according to $1 - e^{-\chi}$). This non-linear mapping has been experimentally verified in [42].

Since the SwissSPAD camera records only a 'binary pixel' in each frame, at a time instant t, the intensity $I_t(i,j)$ observed at a pixel (i,j) is a Bernoulli sample whose probability depends on the intensity of the corresponding scene point. Fig. 2 (a) shows a single 1-bit image of a resolution chart obtained by the SwissSPAD camera. When 3 such frames were captured and averaged, the resultant image is shown in Fig. 2 (b). This is a 2-bit image corresponding to 2^2-1 frames. Similarly Figs. 2 (c), (d) and (e) show 4-bit, 8-bit and 14-bit images. As the number of samples are increased, the noise in the observation reduces. The isolated bright pixels present in all the observations of Fig. 2 correspond to the hot pixels.

A. Objective

When a binary image sequence I_t is averaged upto b bit levels, one would get another sequence denoted by u^b_τ with frame rate reduced by the factor $N_b = 2^b - 1$. The sequence with bit resolution b is given by

$$u^b_\tau = \frac{1}{N_b} \sum_{t=0}^{N_b-1} I_{t+\tau N_b} \tag{2}$$

While imaging a static scene, one can afford to collect as many frames as possible and average them for obtaining an accurate estimate of the scene reflectance map. However, as seen in Fig. 1, while imaging fast dynamic scenes, if one were to average many frames, the temporal information would be lost. Our objective is to overcome this trade-off between frame rate and intensity resolution. For a particular scene, consider that one requires to have a specific frame rate and thereby the bit resolution b gets fixed to a certain level. Let u^b_τ denote the corresponding sequence. Our aim is to estimate a high bit intensity sequence $u^{\tilde{b}}_\tau$ where $\tilde{b}$ is much greater than b and also at the same time preserve the frame rate of the original sequence.

In this paper, we consider the scenarios where N_b takes the values of either 1, 3, 7, and 15. i.e., either 1-bit, 2-bit, 3-bit or 4-bit input sequences, respectively. From image sequences of such low photon counts, we attempt to recover sequences corresponding to a very high bit resolution $\tilde{b} > 12$ bits. Note that, beyond a certain value of number of bits, any further increase would not be adding new information [43]. We observed that when $\tilde{b} > 12$, the noise becomes imperceptible.

III. Proposed method

We propose to learn a mapping function f for generating a sequence with a high bit resolution $\hat{u} = f\left(u^b;\theta\right)$, such that it is close to the noise-free sequence $u^{\tilde{b}}$. Note that we have dropped the time index τ to simplify the notation. The term θ denotes the network parameters.

A. Network architecture

Our network architecture is composed of 3D convolutional layers and residual blocks. The network design is motivated by the fact that many image restoration methods employ architectures with a similar structure (but with 2D convolutions) [44]–[46]. Following such an approach also helps to avoid gradient exploding/vanishing problems [47]. Since 2D CNNs in ResNet style have achieved significant success, we intend to use 3D convolutional layers for processing videos. Fig. 3 shows one residual block of our network. Totally, our network consists of $K = 3$ such residual blocks. Each residual block, has units that are composed of 3D convolutional filters of size $3\times3\times3$. The input convolutional unit consists of one input and 60 output channels. While the three intermediate 'conv' units consist of 60 input and 60 output channels, the output unit has 1 output channel. Except the output unit, all the others are followed by a 'Leaky' ReLU to model non-linearities. For comparison, we also train with filter size $5\times5\times5$.

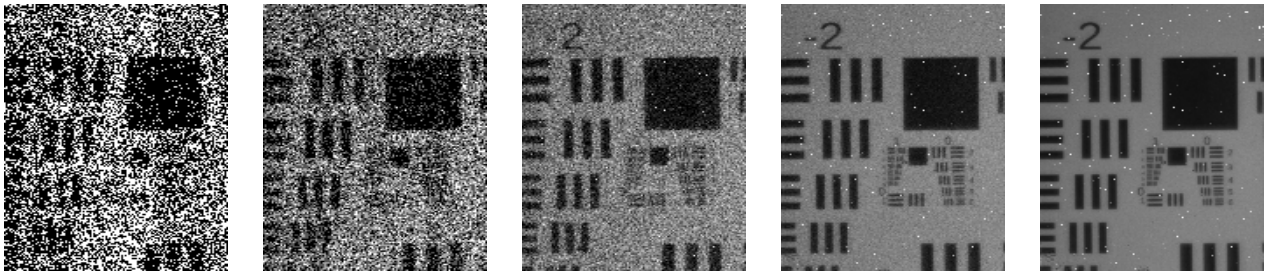
The input to the network is a spatio-temporal patch. At the end of each residual block indexed by k, the input is added to the resultant of the 'output conv' layer to arrive at $\hat{u}_k$, an estimate of the spatio-temporal patch corresponding to the clean high-bit sequence. For training the network, we minimize the loss function which is composed of the loss functions of each residual block. Since it is observed in [44] that the Charbonnier penalty function leads to good performance as against the standard ℓ_2 penalty, we also use the Charbonnier penalty to define our loss function. i.e., the final loss function is given by

$$E\left(\hat{u}_k, u^{\tilde{b}}, \theta\right) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \rho\left(\hat{u}_k - u^{\tilde{b}}\right) \tag{3}$$

where N is the number of training samples in a batch and the penalty term is given by $\rho(x) = \sqrt{x^2+\eta^2}$. The value of η is chosen to be 10^{-3} following [44].

B. Training data

For each bit-level, we train our model using simulated data. For the case of 4-bit sequences, we also train with real pairs of low-bit and corresponding high-bit sequences obtained from SPAD camera. For the 1-bit and 2-bit scenarios, the frame rates that can be achieved by SwissSPAD are 156k fps and 78k fps, respectively. At such high frame rates, one can expect that the temporal variations are quite low. Hence, for these two scenarios, we generate a video dataset with high temporal coherence. The raw videos used in the training of video-deblurring scheme of [48], consists of image sequences at 240fps. The spatial extent of these sequences is 1280×720. We spatially downsample these sequences by a factor of 7 to obtain sequences with reduced variations across time. We randomly crop these downsampled sequences at different temporal locations and arrive at 2500 sequences of dimension $100\times100\times64$. These sequences are directly considered to be the high intensity resolution sequences ($u^{\tilde{b}}$). To generate the low-bit sequences, for every frame, we average N_b Bernoulli sampled (binary) instances. In each Bernoulli instance, the probability of getting a one at a pixel is equal to the corresponding normalized true intensity value at that pixel. i.e., brighter pixels are more likely to generate a 1. To generate such a sample, we generate a uniform random number at every pixel. At a particular pixel, if the randomly generated number is less than the true normalized image intensity value, then that pixel is assigned as one, and zero otherwise. We



(a) (b) (c) (d) (e)

Fig. 2. Resolution chart imaged by the SwissSPAD camera at different bits: (a) 1-bit frame. (b) 2-bit frame. (c) 4-bit frame. (d) 8-bit frame. (e) 14-bit frame.

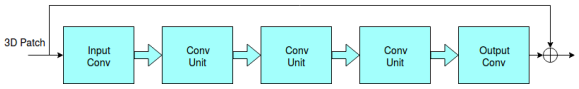


Fig. 3. One residual unit from the proposed ConvNet architecture. Overall, the network consists of K such cascaded residual blocks.

also randomly add salt-and-pepper-noise at a few points to simulate hot-pixels. For 3-bit and 4-bit scenarios, we found that training with videos from UCF 101 dataset [49] that have a normal frame-rate was sufficient. We explain the procedure of obtaining real training data from the SPAD camera in the supplementary material.

Model training We trained our network models on a NVIDIA GeForce GTX 1080 Ti using stochastic gradient descent without batch normalization. The models named 3DCNN1B, 3DCNN2B, 3DCNN3B and 3DCNN4B denote the CNNs trained with 1-bit, 2-bit, 3-bit and 4-bit sequences, respectively. The network trained using 4-bit real data is referred to as 3DCNNR. From the 2500 sequences, we use 2400 for training and the rest for testing. The input in each batch of training is obtained by randomly cropping 3D patches of size $60\times60\times38$ from any of the sequences of the training dataset. We randomly resize, flip, and rotate by multiples of 90° for data augmentation.

TABLE I
QUANTITATIVE EVALUATION ON SIMULATED DATA.

Measure	3DCNN1B	3DCNN1B (5)	3DCNN2B	3DCNN2B (5)
PSNR	25.44	25.37	27.40	26.86
SSIM	0.744	0.740	0.815	0.803

IV. EXPERIMENTS

To run our algorithm on a GPU for videos of realistic size, we divide an input sequence into overlapping 3D spatio-temporal patches and subsequently merge the outputs. For sequences of a particular bit-level, we use its corresponding CNN. If the bit-level of an input is not known, it can be determined easily by checking the number of unique levels in the pixel intensities.

1-bit and 2-bit synthetic experiments We used the test set of simulated data to evaluate the performance of our video reconstruction method. Table I shows the peak signal to noise ratio (PSNR) and structural similar index measure (SSIM) averaged over all the 100 test sequences. Note that 3DCNN1B and 3DCNN2B had different inputs, but the ground-truth sequences were the same as seen in the representative examples shown in Fig. 4. In Table I, the value 5 within parentheses indicates that the filter size used in the CNN was $5\times5\times5$ instead of $3\times3\times3$. To check if any other denoising method works for this scenario, we applied the algorithms proposed in [11], [12], [41], and [50] on five of these image sequences. None of these methods were able to restore videos for these sequences. We varied the parameters of the algorithms and searched for the optimal values. Out of these other algorithms, the best performance was obtained from [12] (when applied with Poisson noise statistics). However even this is not quite satisfactory. The SSIM values of the outputs from [12] ranged between 0.5 and 0.64. For visual comparison, we show one example output of [12] in Fig. 6. The synthetic experiments clearly demonstrate that existing video denoising methods cannot be used for observations with sparse photon counts. However, we subsequently notice improvements in their performance when the bit-level improves.

We also checked if a single network can be used to recover both 1-bit and 2-bit sequences. For this purpose, we trained another CNN wherein the inputs for training consisted of both 1-bit and 2-bit sequences (with equal probability). We observe a slight reduction in the performance of this jointly-trained CNN compared to that of the specific networks seen in Table I. For the case of 1-bit test sequences, the resultant mean PSNR and SSIM from the jointly-trained CNN are 24.97 and 0.726, respectively. For the 2-bit test sequences, the resultant mean PSNR and SSIM are 26.9 and 0.80, respectively. We have also tested the proposed networks on video sequences with downsampling factors less than 7. In these test sequences, the spatio-temporal variations will be increased when compared to the training data. These results are reported in the supplementary material.

Real experiments We initially show an example of 1-bit real sequence. Fig. 7 shows a scene wherein the SwissSPAD camera was placed close to a rotating tool. In this particular setup, because of limitations in the data-transfer rate, 1-bit frames were captured at the rate of 42 kfps. A binary image from the input sequence is shown in Fig. 7 (a). Its corresponding output is shown in Fig. 7 (b). For comparison, we averaged 15

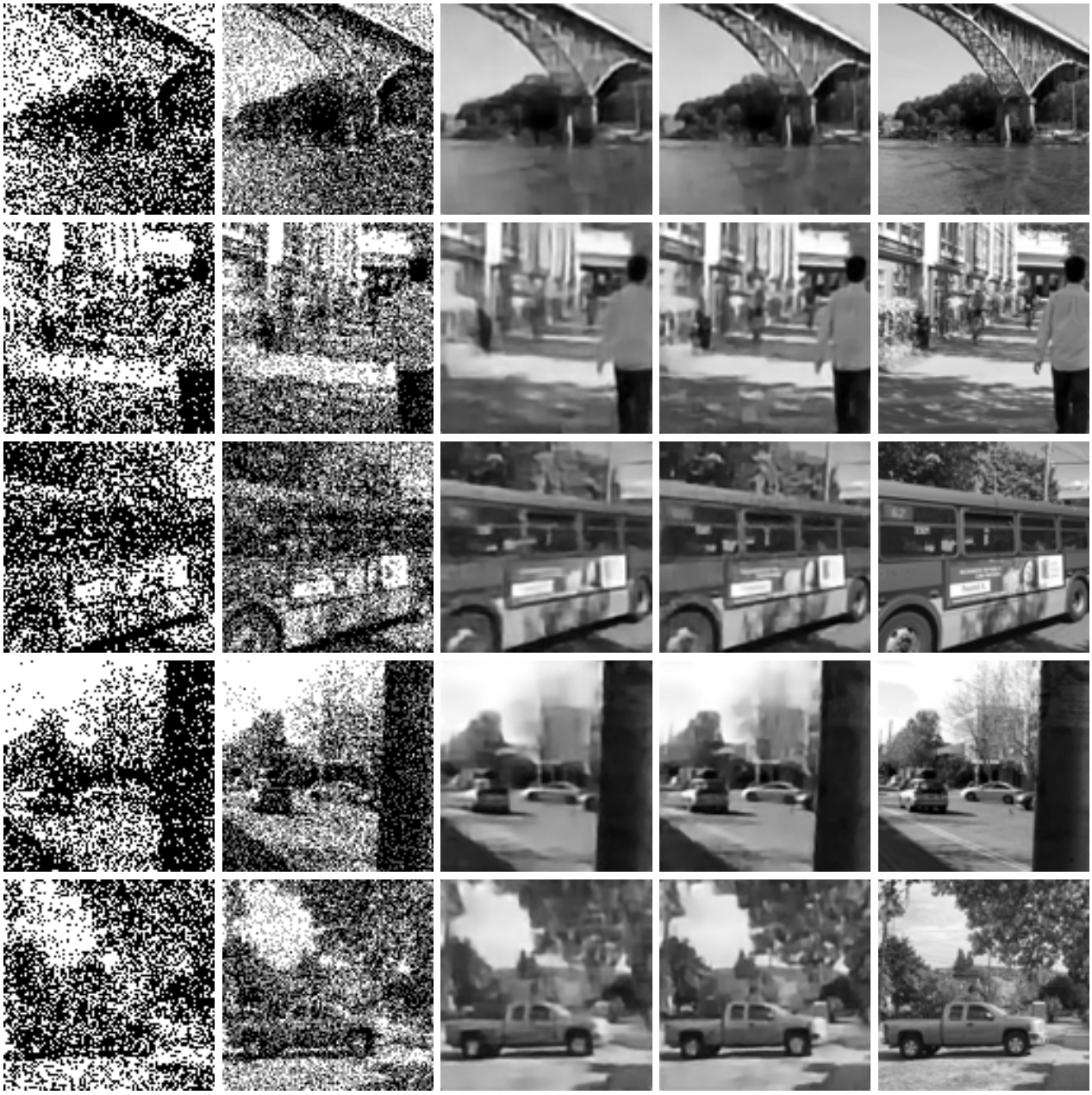


Fig. 4. Representative examples from our test set (Videos are in supplementary material)

TABLE II
QUANTITATIVE RESTORATION PERFORMANCE COMPARISON ON THE REAL SPAD TEST DATASET.

Measure	[41]	VBM4D [11]	[12]	3DCNNR
PSNR	28.53	30.51	31.41	35.54
SSIM	0.7328	0.7816	0.8218	0.909

binary frames to generate a 4-bit sequence. This 4-bit sequence was fed as input to our 3DCNN4B network. In the videos included as supplementary material, one can clearly see the loss of temporal resolution in the output of 3DCNN4B when the averaged sequence was input.

Fig. 8 shows another rotating tool present in a static background. The input in this scenario was a 2-bit sequence captured at about 25 kfps. While Fig. 8 (a) shows an input frame, Fig. 8 (b) shows the corresponding result from 3DCNN2B. By averaging five frames of this sequence, one can obtain a 4-bit sequence. Figs. 8 (c) and (d) show a 4-bit input and a resultant (3DCNN4B) frame, respectively. We observe that in the static regions, the output of 3DCNN2B does come close to that of 3DCNN4B. The supplementary material contains a comparison of our output with a high-intensity resolution reference image captured when the scene was still.

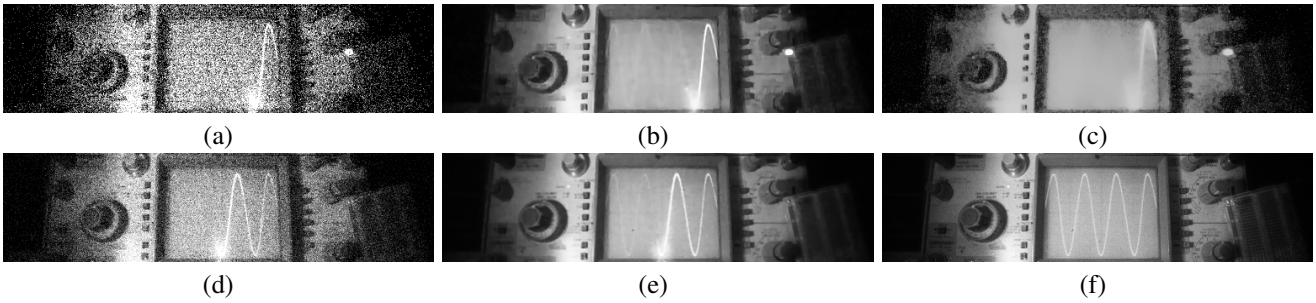


Fig. 5. (a) A 2-bit frame from the input to our algorithm captured at 78000 fps, and (b) corresponding resultant frame (3DCNN2B). (c) Output of [12]. (d) A 4-bit frame, and (e) corresponding output from 3DCNNR. (f) Average of 120 4-bit frames.

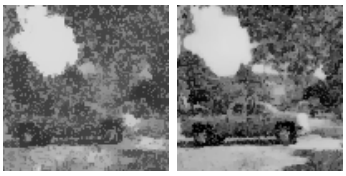


Fig. 6. Result of the algorithm in [12] for the input shown in the last row of Fig. 4: Output for (left) 1-bit sequence and (right) two-bit sequence.

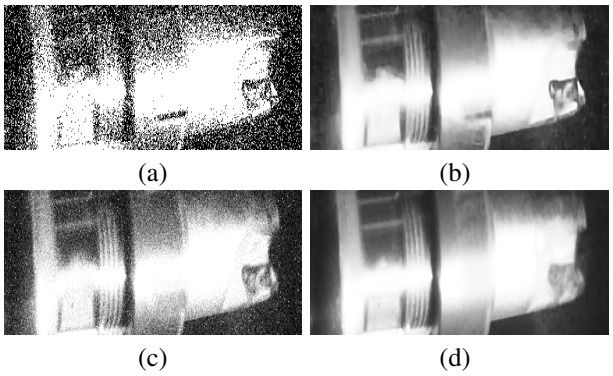


Fig. 7. High speed rotating tool (1-bit): (a) A frame from the input sequence and (b) its corresponding output from 3DCNN1B. (c) A 4-bit input frame generated by averaging. (d) Output of 3DCNN4B on the 4-bit input.

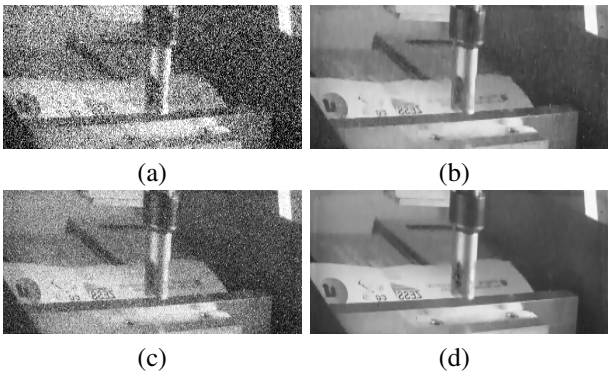


Fig. 8. High speed rotating tool (2-bit): (a) A frame from the input sequence and (b) its corresponding output from 3DCNN2B. (c) A 4-bit input frame generated by averaging. (d) Output of 3DCNN4B on the 4-bit input.

We next show results on additional oscilloscope sequences of [7] (Fig. 5). The 2-bit sequence was fed as input to 3DCNN2B, and 4-bit sequence was input to 3DCNN4B. One can observe that the quality of our 2-bit output does come close to that of 4-bit and even with the high-bit observation (Fig. 5 (f)) at static regions. This shows that our algorithm is capable of producing high-quality outputs from only 2-bit frames. The output from [12] on the 2-bit sequence (Fig. 5 (c)) looks quite inferior to our output. The supplementary material contains complete videos and comparisons. We have also shown the result from [10] and compared it with our method.

Subsequently, we compare the performances of 3DCNN1B, 3DCNN2B and 3DCNN3B on the same scene. The image sequence corresponds to breaking of glass with a resolution chart in the background. The scene was captured at 156kfps and at 1-bit resolution. We divided the sequence into groups of seven frames. For 1-bit observations, we keep the central frame in each group and drop six other frames. For 2-bit observations, we average the middle three frames and drop the rest. For 3-bit observations, we average all the seven frames in the group. Essentially, we arrive at 1-bit, 2-bit and 3-bit sequences with similar temporal variations and corresponding to about 22kfps. With these inputs, we obtain the outputs from our restoration scheme. Fig. 9 shows inputs and outputs at two different instants of time. In the second instant, we see that the particles have been scattered after the breaking of glass. Despite the inputs being highly noisy, we see that the structural information has been recovered quite well. This shows that even with just one-bit measurements, our network model is able to reconstruct scene information quite robustly.

4-bit sequences We quantitatively evaluate different video denoising methods using our real SPAD dataset that has both 4-bit noisy sequence and the corresponding high-bit sequence. We evaluate the performance on ten randomly selected test-image sequences. The performance of different schemes are presented in Table II. The algorithms of [11] and [12], do not handle hot pixels. For these methods, to evaluate score, we replace the intensity of a hot-pixel by the value equal to the median of the 3×3 neighborhood of that hot pixel (excluding the damaged pixels while calculating median). The other algorithms can handle outliers and this step is not necessary. The table shows that our scheme of 3DCNNR clearly outperforms other methods. A representative example from the SwissSPAD dataset is shown in Fig. 11. On close observation of different regions, we can see that the reconstruction is more faithful in

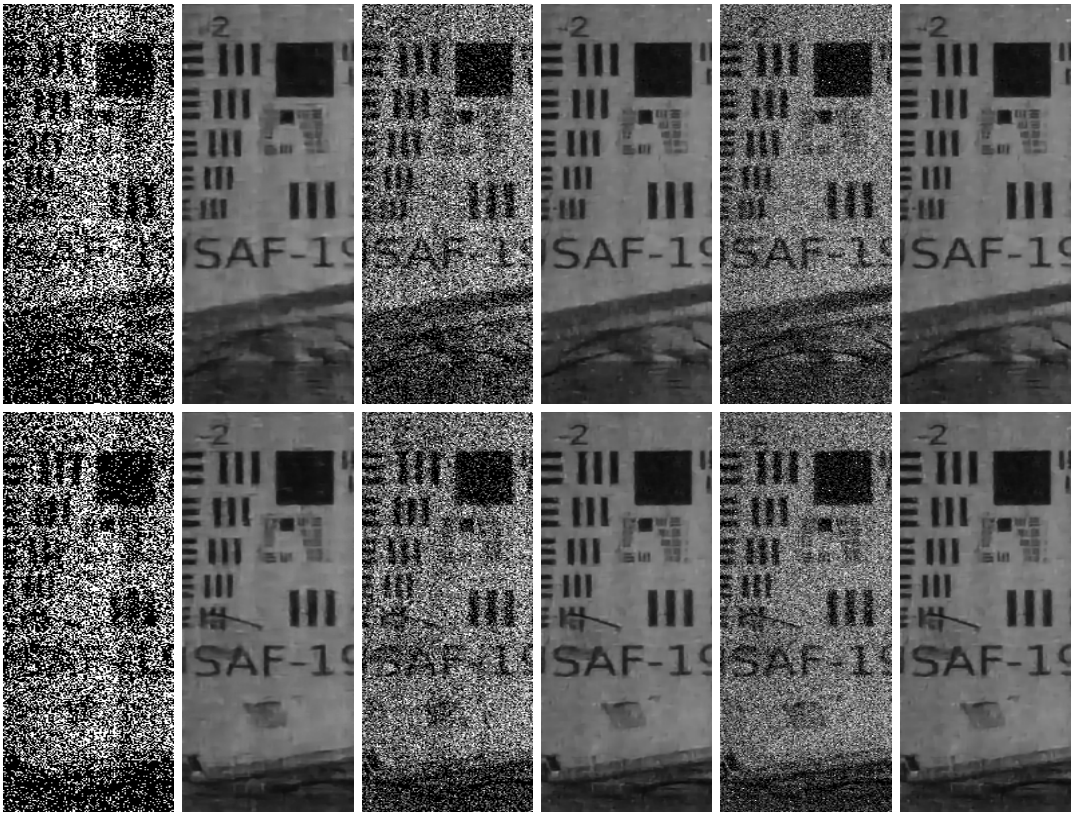


Fig. 9. Each row indicates a different time instant. (Left Pair) 1-bit observation and result. (Central) 2-bit observation and result. (Right) 3-bit observation and result.

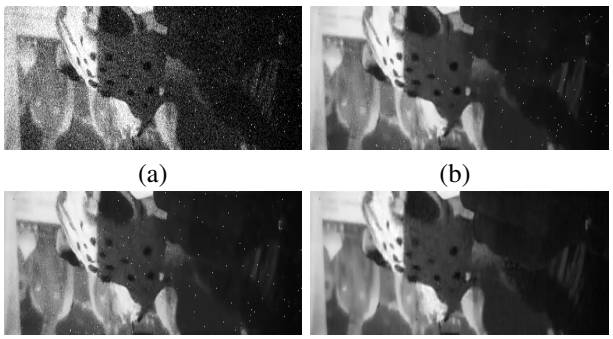


Fig. 10. Real example of a balloon bursting sequence. Frame from (a) input sequence, result of (b) [11], (c) [12], (d) 3DCNN4B.

the proposed CNN model.

We next show a result on a 4-bit sequence captured at 12 kfps. In Fig. 10, show sample frames from the resultant sequences. From the videos, one can clearly observe that the proposed method performs quite well. There has been an improvement in the performance of video denoising schemes when compared to the 1-bit and 2-bit scenarios. However, in the results of [11], [12] and [41] (supplementary material), we do observe artifacts clearly in regions where there is no motion. In the supplementary material, we show additional results and also include a quantitative evaluation of 3DCNN3B on a synthetic dataset.

V. Conclusions

We proposed a video recovery scheme for single-photon counting cameras with sparse photon counts. The performance of our model is quite good in real scenarios despite the training on simulated data. The level of performance achieved by our method in 1-bit and 2-bit scenarios is not possible with any existing approach. Even for 3-bit and 4-bit scenarios, our method outperforms existing video denoising schemes.

In other applications of SPAD cameras such as range imaging and time-of-flight imaging, the number photon counts is of much higher magnitude than the numbers seen in this paper. Our work could serve as a template for developing more photon-efficient techniques to perform these tasks. We would explore this direction in future.

References

[1] E. R. Fossum, N. Teranishi, A. Theuwissen, D. Stoppa, and E. Charbon, *Photon-Counting Image Sensors*. MDPI, 2018.

[2] E. Charbon, "Spad based image sensors," in *IEEE International Electron Devices Meeting (IEDM)*, 2014.

[3] D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappa, V. K. Goyal, F. N. Wong, and J. H. Shapiro, "Photon-efficient imaging with a single-photon camera," *Nature Communications*, vol. 7, 2016.

[4] G. Gariepy, N. Krstajić, R. Henderson, C. Li, R. R. Thomson, G. S. Buller, B. Heshmat, R. Raskar, J. Leach, and D. Faccio, "Single-photon sensitive light-in-flight imaging," *Nature Communications*, vol. 6, 2015.

[5] M. O'Toole, F. Heide, D. B. Lindell, K. Zang, S. Diamond, and G. Wetzstein, "Reconstructing transient images from single-photon sensors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1539–1547.

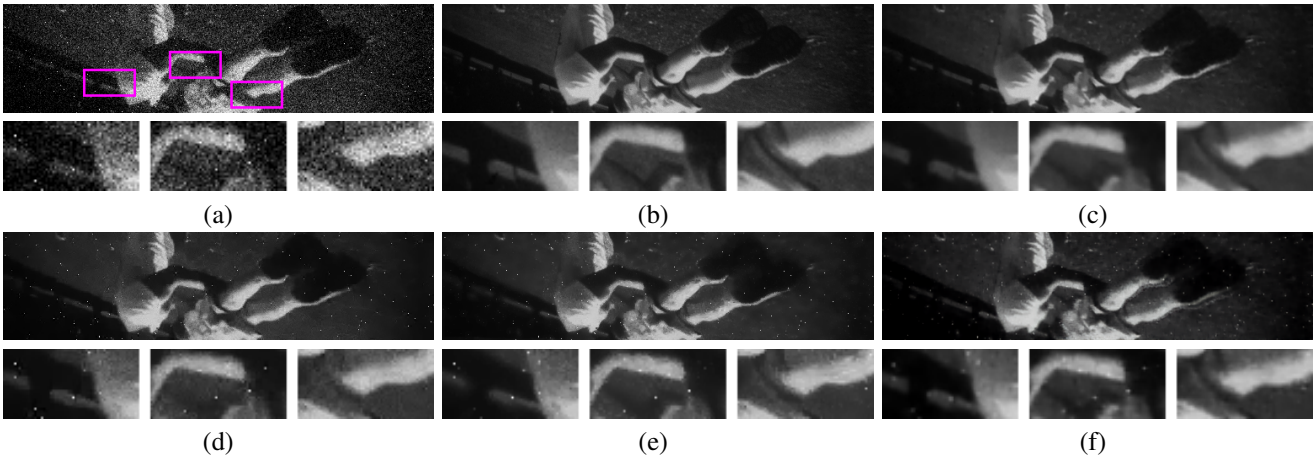


Fig. 11. A representative example from our real test set. (a) A 4-bit frame from the input sequence. (b) Corresponding high-bit resolution image. Frame from recovered video using (c) 3DCNNR, (d) [11], (e) [12] and (f) [41].

[6] F. Heide, M. O'Toole, K. Zhang, D. Lindell, S. Diamond, and G. Wetzstein, "Robust non-line-of-sight imaging with single photon detectors," *arXiv preprint arXiv:1711.07134*, 2017.

[7] S. Burri, Y. Maruyama, X. Michalet, F. Regazzoni, C. Bruschini, and E. Charbon, "Architecture and applications of a high resolution gated spad image sensor," *Optics Express*, vol. 22, no. 14, pp. 17 573–17 589, 2014.

[8] J. Nakamura, *Image sensors and signal processing for digital still cameras*. CRC press, 2017.

[9] K. Yan, L. Lifei, D. Xuejie, Z. Tongyi, L. Dongjian, and Z. Wei, "Photon-limited depth and reflectivity imaging with sparsity regularization," *Optics Communications*, vol. 392, pp. 25–30, 2017.

[10] S. H. Chan, O. A. Elgendy, and X. Wang, "Images from bits: Non-iterative image reconstruction for quanta image sensors," *Sensors*, vol. 16, 2016.

[11] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking and enhancement through separable 4-d nonlocal spatiotemporal transforms," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3952–3966, 2012.

[12] C. Sutour, C.-A. Deledalle, and J.-F. Aujol, "Adaptive regularization of the nl-means: Application to image and video denoising," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3506–3521, 2014.

[13] B. Wen, Y. Li, L. Pfister, and Y. Bresler, "Joint adaptive sparsity and low-rankness on the fly: An online tensor reconstruction scheme for video denoising," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[15] N. A. W. Dutton, L. Parmesan, A. J. Holmes, L. Grant, and R. K. Henderson, "320x240 oversampled digital single photon counting image sensor," *2014 Symposium on VLSI Circuits Digest of Technical Papers*, pp. 1–2, 2014.

[16] I. Gyongy, N. Calder, A. Davies, N. Dutton, P. Dalgarno, R. Duncan, C. Rickman, and R. Henderson, "256× 256, 100kfps, 61% fill-factor time-resolved spad image sensor for microscopy applications," in *IEEE International Electron Devices Meeting (IEDM)*, 2016.

[17] "Lighting for high speed," http://http://www.lovehighspeed.com/lighting-for-high-speed/, accessed: 2018-09-10.

[18] D.-U. Li, J. Arlt, J. Richardson, R. Walker, A. Buts, D. Stoppa, E. Charbon, and R. Henderson, "Real-time fluorescence lifetime imaging system with a 32× 32 0.13 μm cmos low dark-count single-photon avalanche diode array," *Optics Express*, vol. 18, no. 10, pp. 10 257–10 269, 2010.

[19] A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F. N. Wong, J. H. Shapiro, and V. K. Goyal, "First-photon imaging," *Science*, vol. 343, no. 6166, pp. 58–61, 2014.

[20] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3d imaging with deep sensor fusion," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, 2018.

[21] ——, "Towards transient imaging at interactive rates with single-photon detectors," in *IEEE International Conference on Computational Photography (ICCP)*, 2018.

[22] Q. Sun, X. Dun, Y. Peng, and W. Heidrich, "Depth and transient imaging with compressive spad array cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 273–282.

[23] A. K. Pediredla, A. C. Sankaranarayanan, M. Buttafava, A. Tosi, and A. Veeraraghavan, "Signal processing based pile-up compensation for gated single-photon avalanche diodes," *arXiv preprint arXiv:1806.07437*, 2018.

[24] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten, "Non-line-of-sight imaging using a time-gated single photon avalanche diode," *Optics Express*, vol. 23, no. 16, pp. 20 997–21 011, 2015.

[25] G. Gariepy, F. Tonolini, R. Henderson, J. Leach, and D. Faccio, "Detection and tracking of moving objects hidden from view," *Nature Photonics*, vol. 10, no. 1, pp. 23–26, 2016.

[26] M. OToole, D. B. Lindell, and G. Wetzstein, "Confocal non-line-of-sight imaging based on the light-cone transform," *Nature*, vol. 555, no. 7696, p. 338, 2018.

[27] G. Satat, M. Tancik, and R. Raskar, "Towards photography through realistic fog," in *IEEE International Conference on Computational Photography (ICCP)*, 2018, pp. 1–10.

[28] E. R. Fossum, "What to do with sub-diffraction-limit (sdl) pixels? a proposal for a gigapixel digital film sensor (dfs)."

[29] F. Yang, L. Sbaiz, E. Charbon, S. Susstrunk, and M. Vetterli, "Image reconstruction in the gigavision camera," in *ICCV Workshops*, 2009.

[30] J. H. Choi, O. Elgendy, and S. H. Chan, "Image reconstruction for quanta image sensors using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing.*, 2018.

[31] T. Remez, O. Litany, and A. Bronstein, "A picture is worth a billion bits: Real-time image reconstruction from dense binary threshold pixels," in *IEEE International Conference on Computational Photography (ICCP)*, 2016.

[32] R. A. Rojas, W. Luo, V. Murray, and Y. M. Lu, "Learning optimal parameters for binary sensing image reconstruction algorithms," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2791–2795.

[33] A. Buades, B. Coll, and J.-M. Morel, "Denoising image sequences does not require motion estimation," in *IEEE Conference on Advanced Video and Signal Based Surveillance.*, 2005, pp. 70–74.

[34] C. Liu and W. T. Freeman, "A high-quality video denoising algorithm based on reliable motion estimation," in *European Conference on Computer Vision*, 2010, pp. 706–719.

[35] M. Werlberger, T. Pock, M. Unger, and H. Bischof, "Optical flow guided tv-l1 video interpolation and restoration," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2011, pp. 273–286.

[36] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 27–35, 2009.

[37] K. Dabov, A. Foi, and K. O. Egiazarian, "Video denoising by sparse 3d transform-domain collaborative filtering," *2007 15th European Signal Processing Conference*, pp. 145–149, 2007.
[38] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1791–1798.
[39] H. Ji, S. Huang, Z. Shen, and Y. Xu, "Robust video restoration by joint sparse and low rank matrix approximation," *SIAM Journal on Imaging Sciences*, vol. 4, no. 4, pp. 1122–1142, 2011.
[40] X. Chen, L. Song, and X. Yang, "Deep rnns for video denoising," in *Applications of Digital Image Processing*. International Society for Optics and Photonics, 2016.
[41] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *arXiv preprint arXiv:1711.09078*, 2017.
[42] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon, "Nonuniformity analysis of a 65-kpixel cmos spad imager," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 57–64, 2016.
[43] E. R. Fossum, J. Ma, S. Masoodian, L. Anzagira, and R. Zizza, "The quanta image sensor: Every photon counts," *Sensors*, vol. 16, no. 8, p. 1260, 2016.
[44] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[45] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3855–3863.
[46] J. Jiao, W.-C. Tu, S. He, and R. W. Lau, "Formresnet: Formatted residual learning for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1034–1042.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[48] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 237–246.
[49] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[50] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.