

# Supplementary material for “A Little Bit Too Much? High Speed Imaging from Sparse Photon Counts” - PID 29

In this document, we summarize the list of videos and include additional results and comparisons.

## Video Description

The videos corresponding to different scenes are placed in separate folders. Kindly note that we have not included the full duration of videos in order to limit the file size. The following set of videos are present:

- **1bit\_2bit4bit\_oscilloscope** - Contains three videos corresponding to 1-bit, 2-bit and 4-bit inputs. Comparisons with other methods are included in the 4-bit video.
- **1bit\_2bit\_3bit\_glass** - Contains two videos corresponding to two different time instances. Each video contains three pairs of images. As we move from left to right, the results of 1-bit, 2-bit and 3-bit sequences along with their inputs are seen. Note that in this experiment, all the three sequences have the same frame rate.
- **1bit\_2bit\_Synthetic** - Sample videos from our synthetic test set for 1-bit and 2-bit scenarios. Input, result of our method and ground-truth sequences are placed from left to right.
- **1bit\_4bit\_Tool** - Videos of real experiments showing results at 1-bit resolution and 4-bit resolution.
- **2bit\_4bit\_Tool** and **2bit\_4bit\_Water** - Videos of real experiments showing results at 2-bit resolution and 4-bit resolution.
- **3bit\_Synthetic** - Examples from the synthetic test set.
- **4bit\_balloon\_comparisons** - A real sequence showing performance of different methods.

## 3DCNN1B and 3DCNN2B evaluation under different scales

As mentioned in the main paper, we had downsampled (by a factor of 7) the video sequence while generating the training and test sets for 3DCNN1B and 3DCNN2B. To test the robustness, we reduced the scaling factor in the test-set by different amounts. The performance on the reduced downsampling sequences is shown in Table 1. As the downsampling rate is reduced, the variations across the frames increases thereby reducing the temporal coherence. Table 1 indicates that the performance gradually degrades as the downsampling factor reduces.

Table 1: Evaluation of 3DCNN1B and 3DCNN2B for different levels of downsampling. The downsampling factor is indicated in brackets.

Measure	3DCNN1B (6)	3DCNN1B (5)	3DCNN1B (4)	3DCNN2B (6)	3DCNN2B (5)	3DCNN2B (4)
PSNR	25.21	25.18	24.99.64	26.57	26.53	26.33
SSIM	0.739	0.736	0.719	0.796	0.794	0.780

## Comparison on oscilloscope sequence

While we have included actual videos along with this supplementary material, in Fig. 1, for comparison, we show a frame from the results of different video denoising techniques when the input was a 4-bit sequence. Fig. 1 (f) shows that our scheme is able to recover the fine structural information quite well.

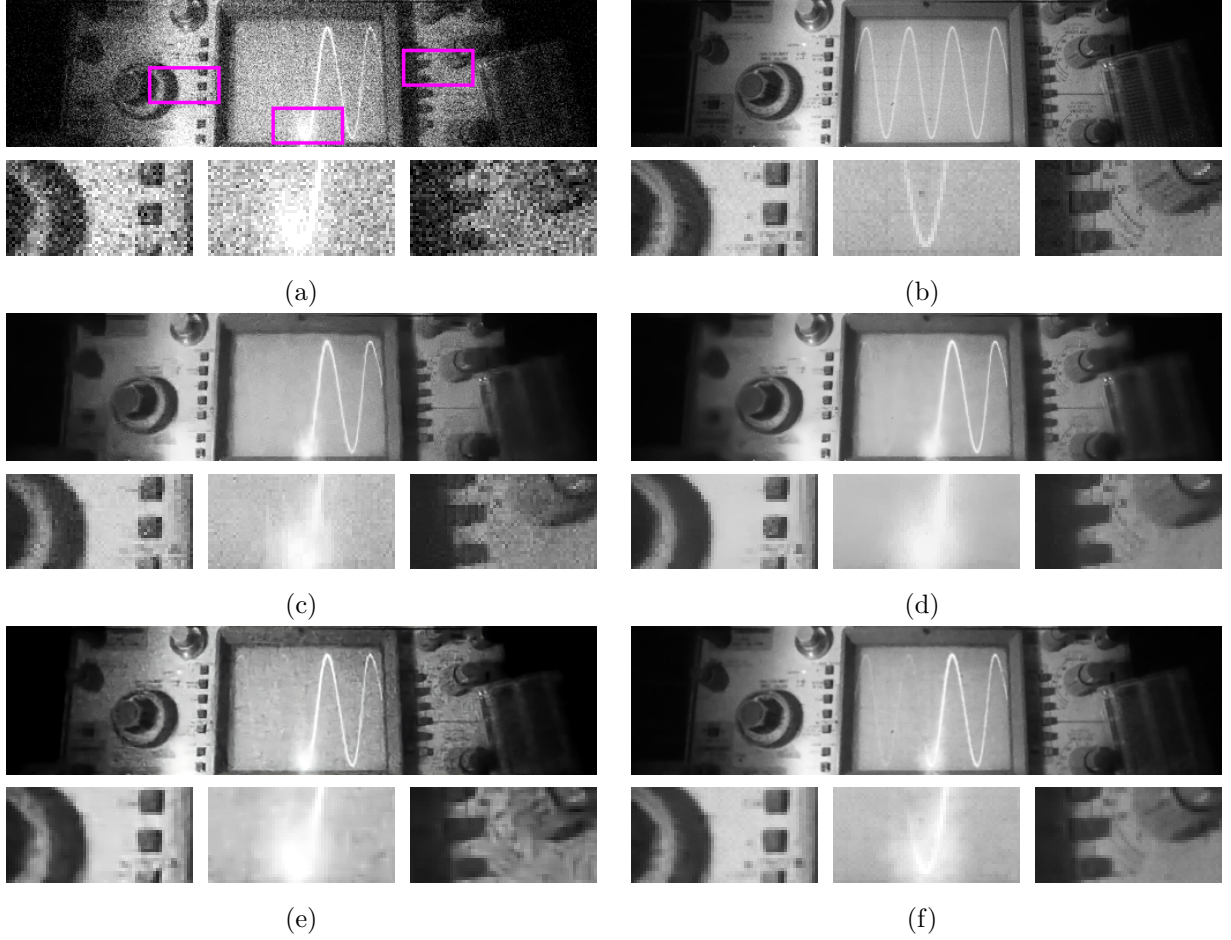


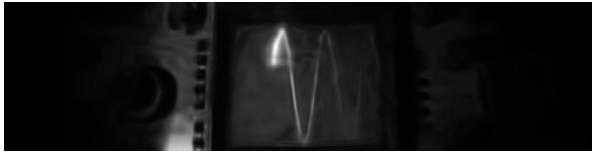
Figure 1: Imaging wave propagation in an oscilloscope: (a) A 4-bit frame from the input sequence. (b) Average of 120 frames. Frame from recovered video using (c) VBM4D [11], (d) Sutour et al. [12], (e) [41], and (f) Proposed 3DCNNR.

The image reconstruction algorithm proposed in [10] has also been applied on the oscilloscope sequence. Here, we compare the output of [10] with our method. In Fig. 2, the results from [10] are directly reproduced from that paper. In [10], a particular frame is inferred using its adjacent frames as input. If 4 frames were used by the algorithm in [10], we have compared it with the result of 3DCNN2B (2-bit input sequence). Similarly when 16 frames were used, we have compared it with our scheme corresponding to 4-bits. Fig. 2 shows that our method leads to higher quality outputs. There is a bright spot in our output corresponding to 3DCNN2B but not in that of [10] because, in our experiment, we had used a different input sequence which also had a bright spot at that location (as seen in the video).

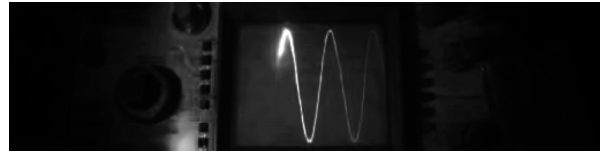
## Comparison with a high frame count reference image

In the real experiments corresponding to the high speed rotating blade (Figs. 7 and 8 in the main paper), we also captured a 16-bit image when the scene was static. For these two experiments, in Fig. 3 we compare the high-bit intensity resolution image against our output (which was either 1-bit or 2-bit).

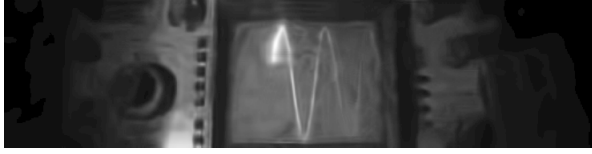
In the first pair of Fig. 3, the tool was located at different positions in the two images. In the second pair, the reference image was captured under different lighting conditions. Despite these differences, one can see that the contents in the scene seen in static reference images are reproduced quite well in our outputs.



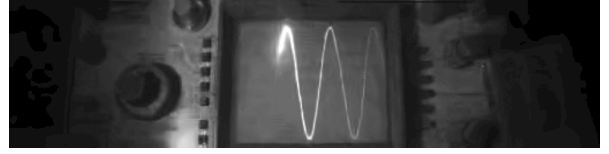
Output of [10] using 4 frames



Output of [10] using 16 frames



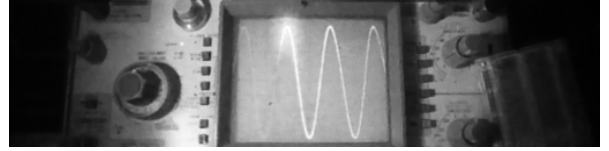
Gamma corrected result of [10] (4 frames)



Gamma corrected result of [10] (16 frames)



Output frame from 3DCNN2B



Output frame from 3DCNN4R

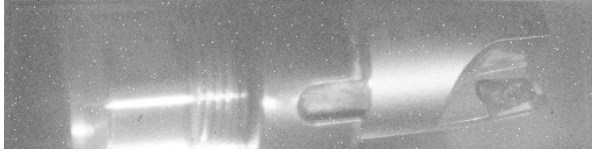
Figure 2: Comparison with the results of [10]: We have directly reproduced images from [10]. For better illustration, we also applied a Gamma correction on the output of [10].



16-bit reference image



A frame from the output of 3DCNN2B



16-bit reference image



A frame from the output of 3DCNN1B

Figure 3: Comparison with a high intensity resolution reference.

## Synthetic Experiments on 3-bit images

For testing our algorithm in 3-bit imaging, we synthetically generate image sequences using UCF 101 video dataset. We generated 40 sequences corresponding to 3-bits for evaluating the proposed models. The PSNR and SSIM measured between the estimated and true image sequence are reported in Table 2. Sample videos from this dataset can be found in the folder named `3bit_Synthetic`. Table 2 indicates that the performance of standard video denoising schemes improves significantly when compared to the scenario of 1-bit and 2-bit sequences (as discussed in the paper). However, in the results of video denoising schemes, visual artifacts are clearly seen especially at the static regions.

## Real dataset capture SPAD

For training with real data, we need to capture scenes simultaneously at high-bit resolution and low-bit resolution. We use an LCD projector for this purpose. We display videos at a very low frame rate (4 fps) and capture a 4-bit image sequence at the rate of 12 kfps from the SwissSPAD sensor array. We divide this sequence into segments of 1180 frames. Within this 1180 frames, we average 260 frames and drop the remaining 920 frames. The average of the 260 frames is considered to be a particular frame in the high-bit sequence and the central frame within the 260 frame window is considered to be the corresponding low-bit frame (4-bit). The average of 260 frames is almost equivalent to a 12-bit image and corresponds to

Table 2: Evaluation on synthetic 3-bit test set.

<b>Measure</b>	VBM4D [11]	Sutour et al. [12]	3DCNN4B	3DCNN3B
PSNR	30.79	30.21	30.64	31.26
SSIM	0.846	0.795	0.893	0.908

a high signal-to-noise ratio level. The dropping of frames and displaying of video at a low frame rate help in reducing the possibility of different scenes getting mixed within a single high-bit resolution frame due to projector lag times. From the images of dark and white scenes captured by the camera, we determine the locations of the damaged pixels. The ground-truth images are obtained by correcting the hot pixels in the high-bit count image sequence through median filtering. In the low-bit images, the hot-pixels are retained to force the network to ignore outliers. In total, we capture 1400 image sequence pairs out of which we use 1200 for training 3DCNNR.