Dynamic Low-light Imaging with Quanta Image Sensors

Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H. Chan

Abstract—Imaging in low light is difficult because the number of photons arriving at the sensor is low. Imaging dynamic scenes in low-light environments is even more difficult because as the scene moves, pixels in adjacent frames need to be aligned before they can be denoised. Conventional CMOS image sensors (CIS) are at a particular disadvantage in dynamic low-light settings because the exposure cannot be too short lest the read noise overwhelms the signal. We propose a solution using Quanta Image Sensors (QIS) and present a new image reconstruction algorithm. QIS are single-photon image sensors with photon counting capabilities. Studies over the past decade have confirmed the effectiveness of QIS for low-light imaging but reconstruction algorithms for dynamic scenes in low light remain an open problem. We fill the gap by proposing a student-teacher training protocol that transfers knowledge from a motion teacher and a denoising teacher to a student network. We show that dynamic scenes can be reconstructed from a burst of frames at a photon level of 1 photon per pixel per frame. Experimental results confirm the advantages of the proposed method compared to existing methods.

Index Terms—Quanta image sensors, single-photon imaging, low light, burst photography

1 Introduction

Imaging in photon-starved situations is one of the biggest technological challenges for applications such as security, robotics, autonomous cars, and health care. However, the growing demand for higher resolution, smaller pixels, and smaller form factors have limited the photon sensing area of the sensors. This, in turn, puts a fundamental limit on the signal-to-noise ratio that the sensors can achieve. Over the past few years, there is an increasing amount of effort in developing alternative sensors that have photon-counting ability. Quanta Image Sensors (QIS) are one of these new types of image sensors that can count individual photons at a very high frame rate and have a high spatial resolution [1], [2]. Various prototype QIS have been reported, and numerous studies have confirmed their capability for high speed imaging [3], high dynamic range imaging [4], [5], color imaging [6], [7], and tracking [8].

Despite the increasing literature on QIS sensor development [1], [2], [9] and signal processing algorithms [10], [11], one of the most difficult problems in QIS is image reconstruction for *dynamic* scenes. Image reconstruction for dynamic scenes is important for broad adoption of QIS: solving the problem can open the door to a wide range of low-light applications such as videography, moving object detection, non-stationary facial recognition, etc. However, motion in low light is difficult because it must deal with two types of distortions: low light causes shot noise which is random and affects the entire image, whereas motion causes geometric warping which is often local. In this paper, we address this problem with a new algorithm.

Figure 1 summarizes our objective. Figure 1(a) shows real data captured by a conventional CMOS image sensor (CIS). The photon level is 0.5 photons per pixel (ppp). Figure 1(b) shows the data captured by a QIS at the same photon level. To illustrate the effect of motion, we show the average of 8 consecutive frames. Figure 1(c) shows the result of the proposed image reconstruction algorithm applied to the 8 QIS frames. Although the scene is in motion, the presented approach recovers most of the image details. This brings out the two contributions of this paper:

(i) We demonstrate low-light image reconstruction of dynamic scenes at a photon level of 1 photon per pixel (ppp) per frame. This is lower than most of the results reported in the computational photography literature.

(ii) We propose a student-teacher framework and show that this training method is effective in handling noise and motion simultaneously.

2 Background

2.1 Quanta Image Sensors

Quanta Image Sensors (QIS) were originally proposed in 2005 as a candidate solution for the shrinking pixel problem [12], [13]. The idea is to partition a CIS pixel into many tiny cells called "jots" where each jot is a single-photon detector. By oversampling the scene in space and time, the underlying image can be recovered using a carefully designed image reconstruction algorithm. Numerous studies have analyzed the theoretical properties of these sensors, including their performance limit [14], photon statistics [9], threshold analysis [5], dynamic range [4], and color filter array [7]. On the hardware side, a number of prototypes have become available [1], [15], [16]. The prototype QIS we use in this paper is based on [2].

Y. Chi, A. Gnanasambandam and S. H. Chan are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA. Email: {agnanasa, chi14, stanchan}@purdue.edu. V. Koltun is with Intel Labs, Santa Clara, CA 95054, USA.

This work is supported, in part, by the National Science Foundation under grant CCF-1718007.

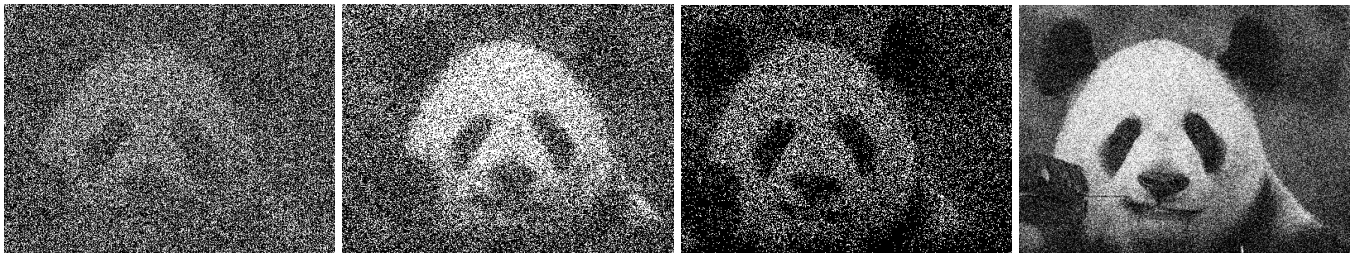
This paper is presented in the 16-th European Conference on Computer Vision (ECCV), Glasgow, United Kingdom, August 2020.

(a) Real image by CIS
avg of 8 frames, 0.5 ppp

(b) Real image by QIS
avg of 8 frames, 0.5 ppp

(c) Our reconstruction
using 8 QIS frames

Fig. 1. **Goal of this paper**. The images above are the *real* captures by a CMOS Image Sensor (CIS) and a QIS prototype [6] at the same photon level of 0.5 photons per pixel (ppp) per frame. The strong shot noise and read noise of CIS makes signal acquisition difficult, whereas the QIS can obtain a better image. Using the proposed method, we are able to reconstruct images with dynamic content.



(a) CIS (real)
0.25 ppp

(b) QIS (real)
0.25 ppp

(c) Ideal sensor
0.25 ppp

(d) i.i.d. Gaussian
$\sigma = 50/255$

Fig. 2. **Photon level and sensor limitations.** (a) and (b) show a pair of real images captured by CIS and QIS at 0.25 ppp. (c) shows a simulated image acquired by an "ideal sensor" which is free of read noise and dark current. The random shot noise in this ideal image suggests that although QIS has higher sensitivity than CIS, image reconstruction algorithms still play a critical role because there is a fundamental limit due to the Poisson statistics. (d) shows an image distorted by i.i.d. Gaussian noise of a strength $\sigma = 50/255$, somewhat high in the denoising literature.

As photon counting devices, QIS share many similarities with single-photon avalanche diodes (SPAD) [15]. However, SPAD amplify signals using avalanche multiplication. This requires a high electrical voltage (typically higher than 20V) to accelerate the photoelectron. Because avalanche multiplication requires space for electrons to multiply, SPAD have high dark current ($> 10e^-/\text{pix/s}$), large pitch ($> 5\mu$m), low fill-factor ($< 70\%$), and low quantum efficiency ($< 50\%$). In contrast, QIS do not require avalanche multiplication. They have significantly better fill-factor, quantum efficiency, dark current, and read noise. SPAD are excellent candidates for resolving time-stamps, e.g., time-of-flight applications [17]–[21], although new studies have shown other applications [22]. QIS have higher resolution which makes them suitable for low-light photography. Recent literature provides a more detailed comparison [6].

2.2 How Dark is One Photon Per Pixel?

All photon levels in this paper are measured in terms of photons per pixel (ppp). "Photons per pixel" is the average number of photons a pixel detects during the exposure period. We use photons per pixel as the metric because the amount of photons detected by a sensor depends on the exposure time and sensor size. A large sensor can collect more photons, and longer exposure time would allow more photons to arrive at the sensor. Therefore, even for the same scene with the same illuminance (measured in lux), the number of photons per pixel seen by two sensors can be different. To give readers an idea of the amount of noise we are dealing with in this paper, Figure 2(a,b) shows a pair of real images captured by CIS and QIS at 0.25 ppp. Note that the signal at this photon level is significantly worse than what is commonly considered "heavy noise" in the denoising literature, illustrated in Figure 2(d). We should also highlight that while QIS is a better sensor, at low light the signal-to-noise ratio is upper bounded by the fundamental limit of the Poisson process. As shown in Figure 2(c), an ideal sensor with zero read noise and zero dark current will still produce an image contaminated by shot noise. Therefore, reconstruction algorithms are needed to recover the images even though QIS have higher photon sensitivity than CIS.

2.3 Related Work

QIS Image Reconstruction. Image reconstruction for QIS is challenging because of the unique Poisson-Gaussian statistics of the sensor. Early reconstruction techniques are based on solving maximum-likelihoods using gradient descent [10], dynamic programming [23], and convex optimization techniques [24], [25]. The first non-iterative algorithm for QIS image reconstruction was proposed by Chan et al. [26]. It was shown that if one assumes spatial independence, then the truncated Poisson likelihood can be simplified to Binomial. Consequently, the Anscombe binomial transform can

be used to stabilize the variance, and off-the-shelf denoising (e.g., BM3D [27]) can be used to denoise the image. Choi et al. [28] followed the idea by replacing the denoiser with a deep neural network. Alternative solutions using end-to-end deep neural networks have also been proposed for QIS [29] and SPAD [11]. To the best of our knowledge, ours is the first dynamic scene reconstruction for QIS.

Low-light Denoising. The majority of existing denoising algorithms are designed for CIS. Single-frame image denoising methods are abundant, e.g., non-local means [32], BM3D [27], Poisson denoising [33], and many others [34]–[37]. On the deep neural network side, there are numerous networks dedicated to single-image denoising [38]–[41]. However, recent benchmark experiments found that BM3D is often better than deep learning methods for real sensor data [42], [43]. Specific to low-light imaging, Chen et al. [44], [45] observed that by modeling the entire image and signal processing pipeline using an end-to-end network, better reconstruction results can be obtained from the raw sensor data. However, since the images are still captured by CIS, the photon levels are much higher than what we study in this paper.

For dynamic scenes, extensions of the static methods to videos are available, e.g., based on non-local means [46]–[48], optical flow [49]–[51], and sparse representation [52], [53]. The most relevant approach for this paper is the burst photography technique [54], which can be traced back to earlier methods based on optical flow [49], [51], [55]. Recent reports on burst photography have focused on using deep neural networks [56]–[59]. Among these, the kernel prediction network (KPN) by Mildenhall et al. [30] is the most relevant work for us. However, as we will demonstrate later in the paper, the performance of KPN is not as satisfactory in the extreme noise conditions we deal with.

3 METHOD

The proposed method consists of the QIS and a new image reconstruction algorithm. Before we discuss the algorithm, we first discuss how images are formed on QIS, as well as the challenges of imaging dynamic scenes in low-light. After that, we discuss the proposed solution using student-teacher learning, and the intuitions behind the method.

3.1 QIS Imaging Model

We now present the image formation model. Our model is based on the prototype QIS reported in [2] and is more detailed than the models used in existing literature such as [14], [26].

As light travels from the scene to the sensor, the main mathematical model is the Poisson process which describes how photons arrive. However, due to various sources of distortions, the measured QIS signal, $\boldsymbol{x}_{\text{QIS}}$, is given by

$$\underbrace{\boldsymbol{x}_{\text{QIS}}}_{\text{observed}} = \text{ADC}\Big\{ \underbrace{\text{Poisson}}_{\text{photon arrival}} \Big(\underbrace{\alpha}_{\text{sensor gain}} \cdot \big(\underbrace{\boldsymbol{x}_{\text{true}}}_{\text{scene}} + \cdots + \underbrace{\boldsymbol{\eta}_{\text{dc}}}_{\text{dark current}} \big) \Big) + \underbrace{\boldsymbol{\eta}_{\text{r}}}_{\text{read noise}} \Big\}. \quad (1)$$

Here we assume that the sensor is monochromatic because the real data reported in this paper are based on a monochromatic prototype QIS. To simulate color data we need to include a sub-sampling step to model the color filter array. $\boldsymbol{\eta}_{\text{dc}}$ denotes the dark current and $\boldsymbol{\eta}_{\text{r}}$ denotes the read noise arising from the read-out circuit. The analog-to-digital converter (ADC) describes the sensor output. In single-bit QIS, the output is a binary signal obtained by thresholding the Poisson count [5]. In multi-bit QIS, the output is the Poisson count clipped to the maximum number of bits. To image a dynamic scene, we use QIS to collect a stack of short-exposure frames. Akin to previous work [14], [26], we assume that noise is independent over time.

For the prototype sensor we use in this paper, the dark current $\boldsymbol{\eta}_{\text{dc}}$ in Equation (1) has an average value of $0.0068e^-/\text{pix/s}$ and the read noise $\boldsymbol{\eta}_{\text{r}}$ takes the value of $0.25e^-/\text{pix}$ [2]. The sensor gain α controls the exposure time and the dynamic range, which changes from scene to scene. For all experiments we conduct in this paper, the analog-to-digital conversion is 3-bit. The spatial resolution of the sensor is 1024×1024, although we typically crop regions of the image for analysis.

3.2 The Dilemma of Noise and Motion

At the heart of dynamic image reconstruction is the coexistence of noise and motion. The dilemma here is that they are intertwined. To remove noise in a dynamic scene, we often need to either align the frames or construct a steerable kernel over the space-time volume. The alignment step is roughly equivalent to estimating optical flow [60], whereas constructing the steerable kernel is equivalent to non-local means [47], [48] or kernel prediction [30]. However, if the images are contaminated by noise, then both optical flow and kernel prediction will fail. When this step fails, denoising will be difficult because we will not be able to easily find neighboring patches for filtering.

Existing algorithms in the denoising literature can usually only handle one of the two situations. For example, the kernel prediction network (KPN) [30] can extract motion information from a dynamic scene but its performance drops when noise becomes heavy. Similarly, the residual encoder-decoder networks REDNet [31] and DnCNN [39] are designed for static scenes. In Figure 3, we show the results of a synthetic experiment. The results illustrate the limitations of the motion-based KPN [30] and the single-frame REDNet (sRED) [31]. Our goal is to leverage the strengths of both.

3.3 Student-Teacher Learning

If a kernel prediction network can handle clean image sequences well and a denoising network can handle static image sequences well, is there a way we can leverage their strengths to address the dynamic low-light setting? Our solution is to develop a training scheme using the concept of student-teacher learning.

Figure 4 describes our method. There are three players in this training protocol: a teacher for motion (based on kernel prediction), a teacher for denoising (based on image denoiser networks), and a student which is the network we are going to use eventually. The two teachers are individually

(a) QIS raw data, 8-frame avg | (b) KPN [30], 23.09 dB | (c) sRED [31], 17.74 dB | (d) Ours, 26.74 dB

Fig. 3. **The dilemma of noise and motion**. (a) A simulated QIS sequence at 2 ppp, averaged over 8 frames. (b) Result of Kernel Prediction Network (KPN) [30], a burst photography method that handles motion. (c) Result of a single-frame image denoiser sRED [31] applied to the 8-frame avg. (d) Result of our proposed method.

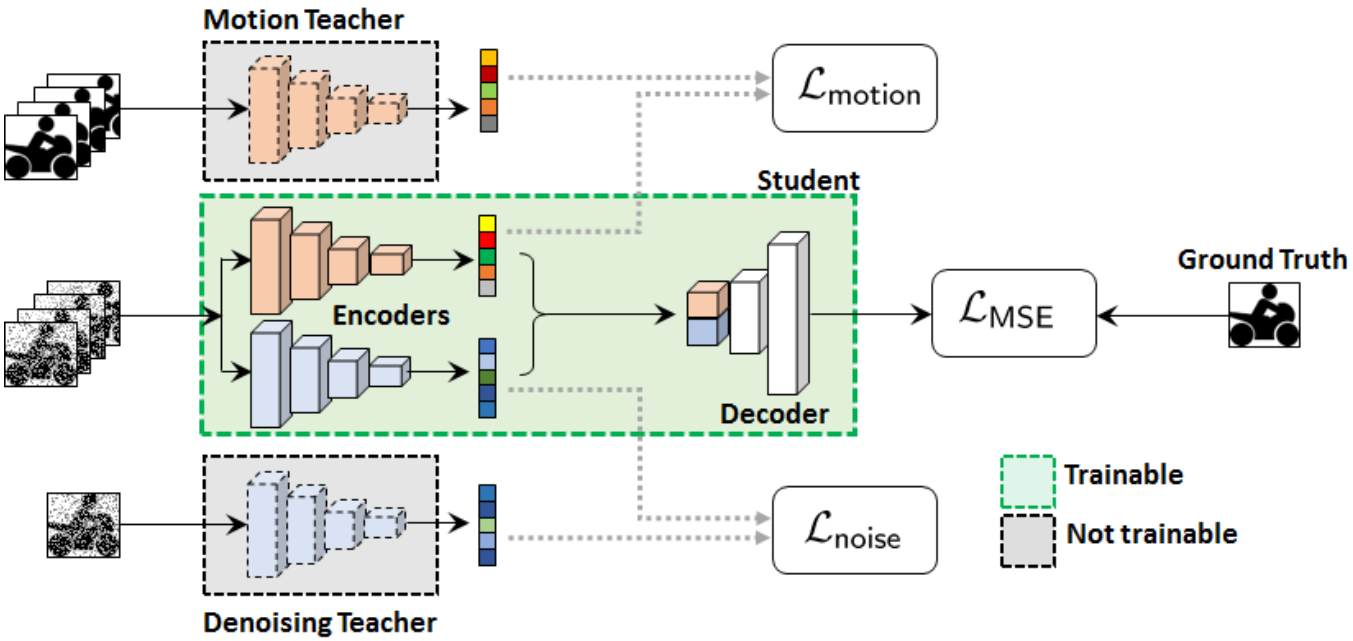


Fig. 4. **Overview of the proposed method**. The proposed student-teacher setup consists of two teachers and a student. The motion teacher shares motion features, whereas the denoising teacher shares denoising features. To compare the respective feature differences, perceptual losses $\mathcal{L}_{\text{noise}}$ and $\mathcal{L}_{\text{motion}}$ are defined. The student network has two encoders and one decoder. The final estimates are compared with the ground truth using the MSE loss $\mathcal{L}_{\text{MSE}}$.

pretrained using their respective imaging conditions. For example, the motion teacher is trained using sequences of clean and dynamic contents, whereas the denoising teacher is trained using sequences of noisy but static contents. During the training step, the teachers will transfer their knowledge to the student. During testing, only the student is used.

To transfer knowledge from the two teachers to the student, the student is first designed to have two branches, one branch duplicating the architecture of the motion teacher and another branch duplicating the architecture of the denoising teacher. When training the student, we generate three versions of the training samples. The motion teacher sees training samples that are clean and only contain motion, $\boldsymbol{x}_{\text{motion}}$. The denoising teacher sees a training sample containing no motion but corrupted by noise, $\boldsymbol{x}_{\text{noise}}$. The student sees the noisy dynamic sequence $\boldsymbol{x}_{\text{QIS}}$.

Because the student has identical branches to the teachers, we can compare the features extracted by the teachers and the student. Specifically, if we denote $\phi(\cdot)$ as the feature extraction performed by the motion teacher, $\widehat{\phi}(\cdot)$ the student motion branch, $\varphi(\cdot)$ the denoising teacher, and $\widehat{\varphi}(\cdot)$ the student denoising branch, then we can define a pair of *perceptual similarities*: the motion similarity

$$\mathcal{L}_{\text{motion}} = \| \underbrace{\widehat{\phi}(\boldsymbol{x}_{\text{QIS}})}_{\text{motion student}} - \underbrace{\phi(\boldsymbol{x}_{\text{motion}})}_{\text{motion teacher}} \|^2 \tag{2}$$

and the denoising similarity

$$\mathcal{L}_{\text{noise}} = \| \underbrace{\widehat{\varphi}(\boldsymbol{x}_{\text{QIS}})}_{\text{denoising student}} - \underbrace{\varphi(\boldsymbol{x}_{\text{noise}})}_{\text{denoising teacher}} \|^2. \tag{3}$$

Intuitively, what this pair of equations does is ensure that the features extracted by the student branches are similar to those extracted by the respective teachers, which are features that can be extracted in good conditions. If this

can be achieved, then we will have a good representation of the noisy dynamic sample and hence we can do a better reconstruction.

The two student branches can be considered as two autoencoders which convert the input images to codewords. As shown on the right side of Figure 4, we have a "decoder" which translates the concatenated codewords back to an image. The loss function of the decoder is given by the standard mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \|f(\boldsymbol{x}_{\text{QIS}}) - \boldsymbol{x}_{\text{true}}\|^2, \tag{4}$$

where f is the student network and so $f(\boldsymbol{x}_{\text{QIS}})$ denotes the estimated image. The overall loss function is the sum of these losses:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{motion}} + \lambda_2 \mathcal{L}_{\text{noise}}, \tag{5}$$

where λ_1 and λ_2 are tunable parameters. Training the network is equivalent to finding the encoders $\widehat{\phi}$ and $\widehat{\varphi}$, and the decoder f.

3.4 Choice of Teacher and Student Networks

The proposed student-teacher framework is quite general. Specific to this paper, the two teachers and the student are chosen as follows.

The motion teacher is the kernel prediction network (KPN) [30]. We modify it by removing the skip connections to maintain the information kept by the encoder. In addition, we remove the pooling layers and the bilinear upsampling layers to maximize the amount of information being fed to the feature layer. With these changes, the KPN becomes a fully convolutional-deconvolutional network.

The denoising teacher we use is a modified version of REDNet [31], which is also used in another QIS reconstruction method [28]. To differentiate this single-frame REDNet and another modified version (to be discussed in the experiment section), we refer to this single-frame REDNet denoising teacher as sRED. Like the motion teacher, we remove the residual connections since they have a negative impact on the feature transfer in student-teacher learning.

The student network has two encoders and a decoder. The encoders have exactly the same architectures as the teachers. The decoder is a stack of 15 layers where each layer is a 128-channel up-convolution. The entrance layer is used to concatenate the motion and denoising features.

4 Experiments

4.1 Experiment Settings

Training Data. The training data consists of two parts. The first part is for *global motion*. We use the Pascal VOC 2008 dataset [61] which contains 2000 training images. The second part is for *local motion*. We use the Stanford Background Dataset [62] which contains 715 images with segmentation. For both datasets, we randomly crop patches of size 64×64 from the images to serve as ground truth. An additional 500 images are used for validation. To create global motion, we shift the patches according to a random continuous camera motion where the number of pixels traveled by the camera range from 7 to 35 across 8 consecutive frames. This is approximately 1 m/s. For local motion, we fix the background and shift the foreground using translations and rotations. The implementation of the translation is the same as that of the global motion but applied to foreground objects. The rotation is implemented by rotating the object with an angle ranging from 0 to 15 degrees.

Training the Teachers. The motion teacher is trained using a set of noise-free and dynamic sequences. The loss function is the mean squared error (MSE) loss suggested by [30]. The network is trained for 200 epochs using the dataset described above. The denoising teacher is trained using a set of noisy but static images. Therefore, for every ground-truth sequence we generate a triplet of sequences: A noise-free dynamic sequence for the motion teacher, a noisy static image for the denoising teacher, and a noisy dynamic sequence for the student. We remark that such a data synthesis approach works for our problem because the simulated QIS data matches the statistics of real measurements.

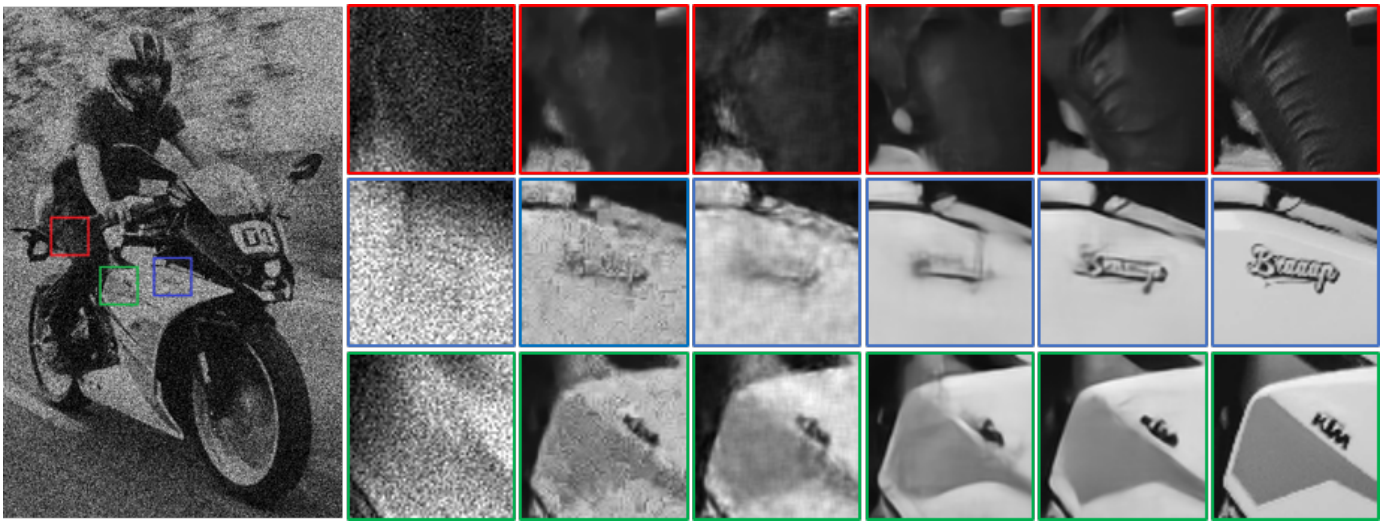
Baselines. We compare the proposed methods with three existing dynamic scene reconstruction methods: (i) BM4D [63], (ii) Kernel Prediction Network (KPN) [30], and (iii) a modified version of REDNet [31]. Our modification generalizes REDNet to multi-frame inputs, by introducing a 3D convolution at the input layer to pool the features. We refer to the modified version as multi-frame RED (mRED). Note that mRED has residual connections while sRED (denoising teacher) does not. We consider mRED a more fair baseline since it takes an input of 8 consecutive frames rather than a single frame. For KPN, the original method [30] suggested using a fixed kernel size of $K = 5$; we modify the setting by defining K as the maximum number of pixels traveled by the motion.

Implementation. All networks are implemented using Keras [64] and TensorFlow [65]. The student-teacher training is done using a semi-annealing process. Specifically, the regularization parameters λ_1 and λ_2 are updated once every 25 epochs such that λ_1 and λ_2 decay exponentially for the first 100 epochs. For the next 100 epochs, λ_1 and λ_2 are set to 0 and the overall loss function becomes $\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{MSE}}$.

4.2 Synthetic Experiments

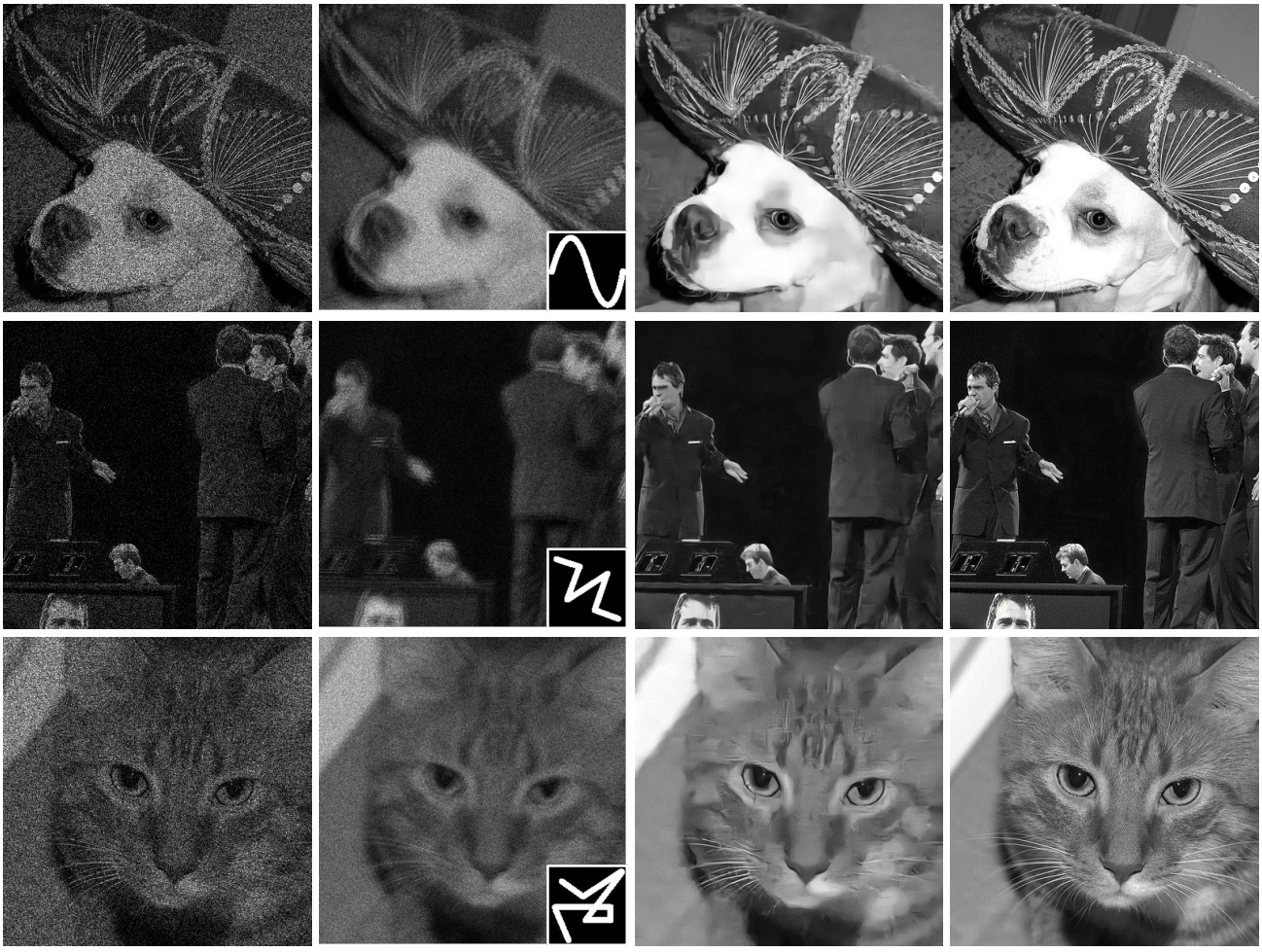
We begin by conducting synthetic experiments. We first visually compare the reconstructed images of the proposed method and the competing methods. Figure 5 shows some results using global translation. The motion magnitude is 28 pixels across 8 frames, at 2 ppp. Figure 6 shows some results using arbitrary global motion, at 4 ppp. The motion trajectory is shown in the inset in the figure. Figure 7 shows some results of local motion. We simulate QIS data with a real motion video of 30 fps. The photon level is 1.5 ppp. The average inference time of KPN on a 512×512 patch is 0.0886 seconds using an NVIDIA GeForce RTX 2080 Ti graphics card. For the same testing setting, mRED takes 0.0653 seconds, and the proposed method takes 0.1943 seconds. The average time for BM4D (MATLAB version) is 23.6985 seconds.

To quantitatively analyze the performance, we use the linear global motion to plot two sets of curves as shown in Figure 8. In the first plot, we show PSNR as a function of the motion magnitude. The magnitude of the motion is defined as the number of pixels traveled along the dominant direction, over 8 consecutive frames. As shown in Figure 8(a), the



(a) QIS raw data, 1 frame | (b) Avg of 8 frames | (c) BM4D 23.04 dB | (d) KPN 25.45 dB | (e) mRED 26.42 dB | (f) Ours 29.39 dB | (g) Ground Truth

Fig. 5. **Simulated QIS data with linear global motion**. (a) The raw QIS image is simulated at 2 ppp, with a global motion of 28 pixels uniformly spaced across 8 frames. (b) An average 8 QIS raw frames. (c) BM4D [63] (d) KPN [30]. (e) mRED, a modification of REDNet [31]. (f) Proposed method. (g) Ground truth.

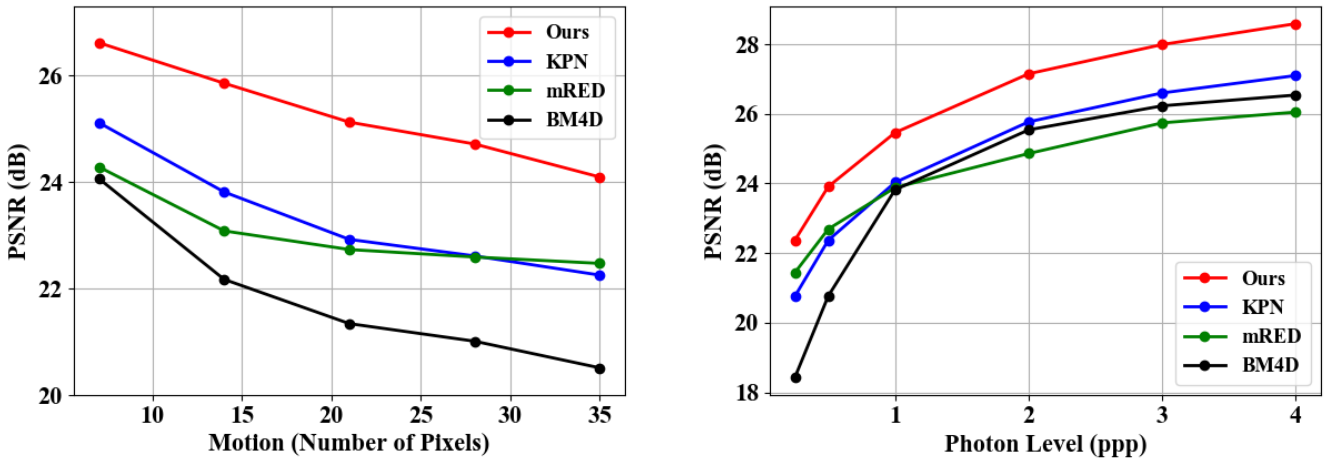


(a) QIS raw | (b) avg 8 frames | (c) Ours | (d) Ground truth

Fig. 6. **Simulated QIS data with arbitrary global motion**. (a) QIS raw data simulated at 4 ppp. The motion trajectory is shown in the insect. (b) Average of 8 frames. (c) Proposed method. (d) Ground truth.

(a) QIS raw (b) avg 8 frames (c) Ours (d) Ground truth

Fig. 7. **Simulated QIS data with local motion**. In this example, only the car moves. The background is static. (a) Raw QIS frame assuming 1.5 ppp. (b) The average of 8 QIS frames. (c) Proposed algorithm. (d) Ground truth.



(a) PSNR vs. Motion at photon level of 2 ppp

(b) PSNR vs. Photon Level at motion magnitude of 4 pixels

Fig. 8. **Quantitative analysis using synthetic data**. (a) PSNR as a function of the motion magnitude, at a photon level of 2 ppp. The magnitude of the motion is defined as the number of pixels traveled along the dominant direction, over 8 consecutive frames. (b) PSNR as a function of photon level. The motion magnitude is fixed at 4 pixels, but the photon level changes. Notice the consistent performance improvement of our method compared to BM4D [63], KPN [30] and mRED (a modified version of [31]).

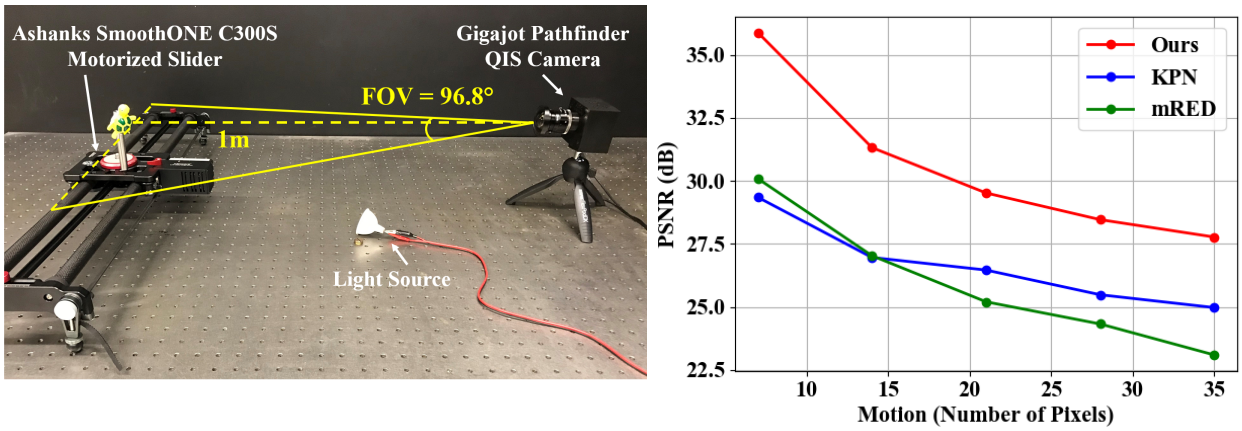
proposed method has a consistently higher PSNR compared to the three competing methods, ranging from 1.5 dB to 3 dB. This suggests that the presence of both teachers has provided a positive impact on solving the motion and noise dilemma, which is difficult for both KPN and mRED. The second set of curves is shown in Figure 8(b) and reports PSNR as a function of the photon level. The curves in Figure 8(b) suggest that for the photon levels we have tested, the performance gap between the proposed method and the competing methods is consistent. This provides additional evidence of the effectiveness of the proposed method.

4.3 Real Experiments

We verify the results using real QIS data. The real data is collected using a prototype Gigajot PathFinder camera [2]. The camera has a spatial resolution of 1024×1024. The integration time of each frame is 75 μs. Each reconstruction is based on 8 consecutive QIS frames. At the time this experiment is conducted, the readout circuit of this camera is still a prototype that is not optimized for speed. Thus, instead of demonstrating a real high-speed video, we capture a slowly moving real dynamic scene where the motion is continuous but slow. We make the exposure period short so that it is equivalent to a high-speed video. We expect that the problem will be solved in the next generation of QIS.

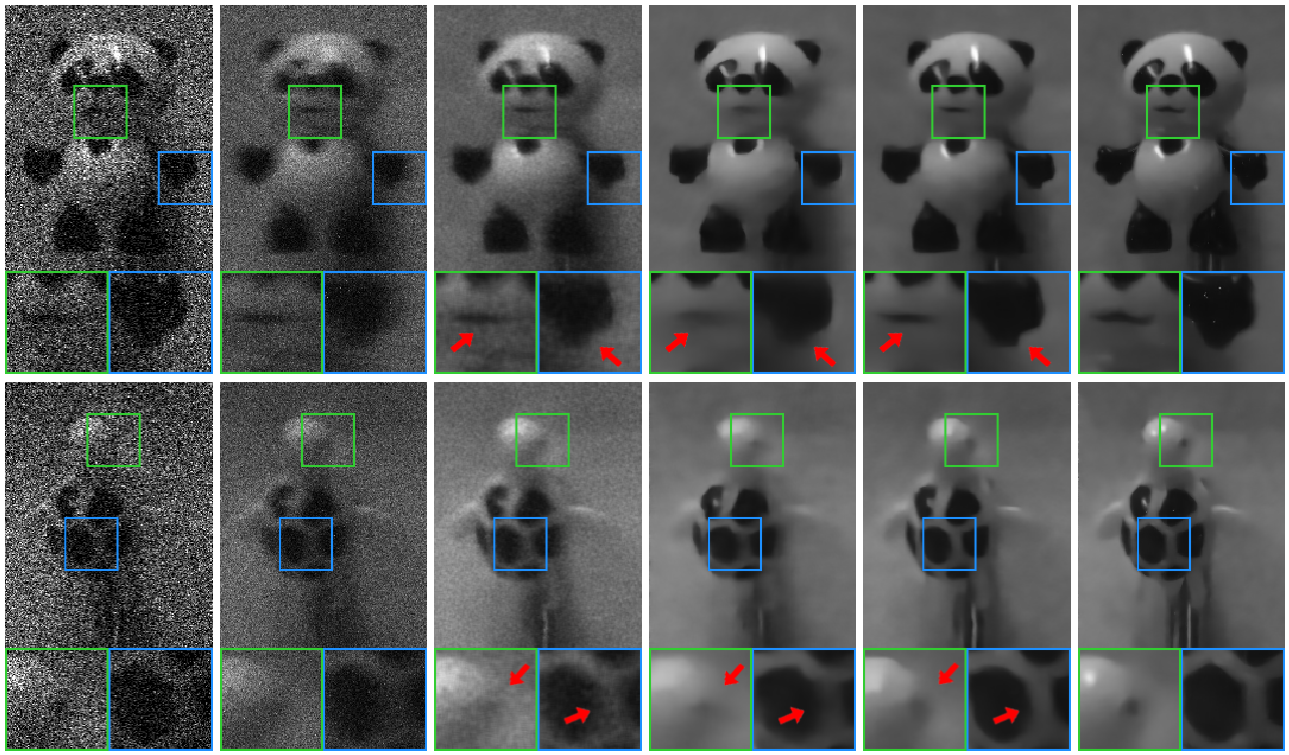
The physical setup of the experiment is shown in Figure 9(a). We put the camera approximately 1 meter away from the objects. The photon level is controlled by a light source. To create motion, the objects are mounted on an Ashanks SmoothONE C300S motorized camera slider, which allows us to control the location of the objects remotely. The "ground truth" (reference images) in this experiment is obtained by capturing a static scene via 8 consecutive QIS frames. Since these static images are noisy (due to photon shot noise), we apply mRED to denoise the images before using them as the references.

A visual comparison for this experiment is shown in Figure 10. The quantitative analysis is shown in Figure 9(b), where we plot the PSNR curves as functions of the number of pixels traveled by the object. As we can see, the performance of the proposed method and the competing methods are similar to those reported in the synthetic experiments.



(a) Experimental Setup (b) PSNR vs. motion (pixels)

Fig. 9. **(a) Setup of QIS data collection**. The QIS camera is placed 1 meter from the object which is attached to a motorized slider. The horizontal field of view (FOV) of the lens is $96.8°$. The motion is continuous but slow. **(b) Quantitative analysis on real data**. The plot shows the PSNR values as a function of the motion magnitude, under a photon level of 0.5 ppp. The "reference" in this experiment is determined by reconstructing an image using a stack of static frames of the same scene. The reconstruction method is based on [28].



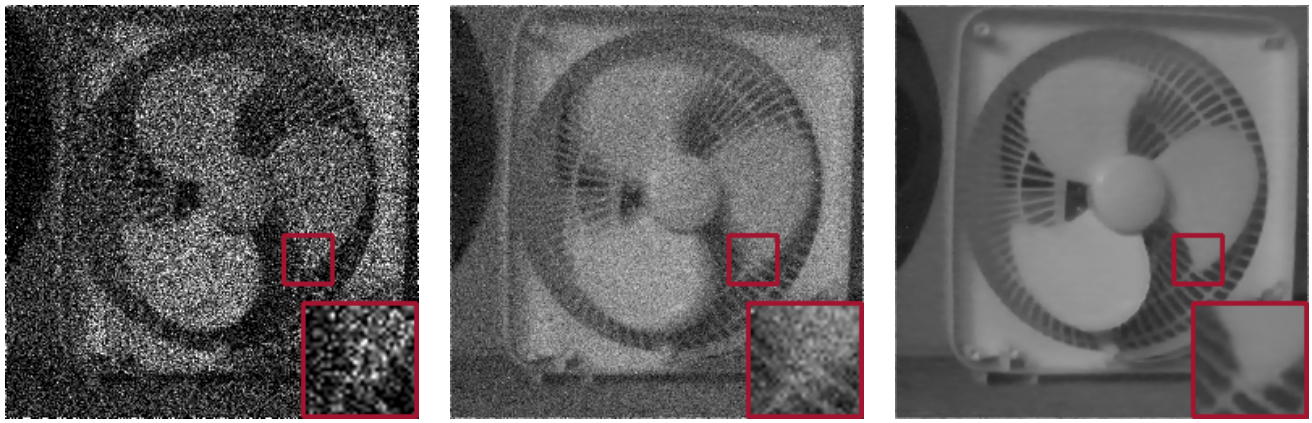
(a) QIS raw 1 frame (b) Average 8 frames (c) KPN 25.08 dB (d) mRED 25.33 dB (e) Ours 30.97 dB (f) Reference

Fig. 10. **Real QIS data**. (a) A snapshot of a real QIS frame captured at 2 ppp per frame. The number of pixels traveled by the object over the 8 frames is 28 pixels. (b) The average of 8 QIS frames. Notice the blur in the image. (c) Reconstruction result of KPN [30]. (d) Reconstruction result of mRED, a modification of [31]. (e) Our proposed method. (f) Reference image is a static scene denoised using mRED.

The gap appears to be consistent with the synthetic experiments. An additional real data experiment is shown in Figure 11, where we use QIS to capture a rotating fan scene.

4.4 Ablation Study

We conduct an ablation study to evaluate the significance of the proposed student-teacher training protocol. Figure 12 summarizes the 5 configurations we study. Config A is a vanilla baseline where the denoising and motion teachers are pretrained. Config B uses a single encoder instead of two encoders. Ours-I uses a student-teacher setup to train the denoising encoder. Ours-II is similar to Ours-I, but we use the motion teacher in lieu of the denoising teacher. Ours-full uses both teachers. All networks are trained using the same set of noisy and dynamic sequences. The experiments are conducted using synthetic data, at a photon level of 1

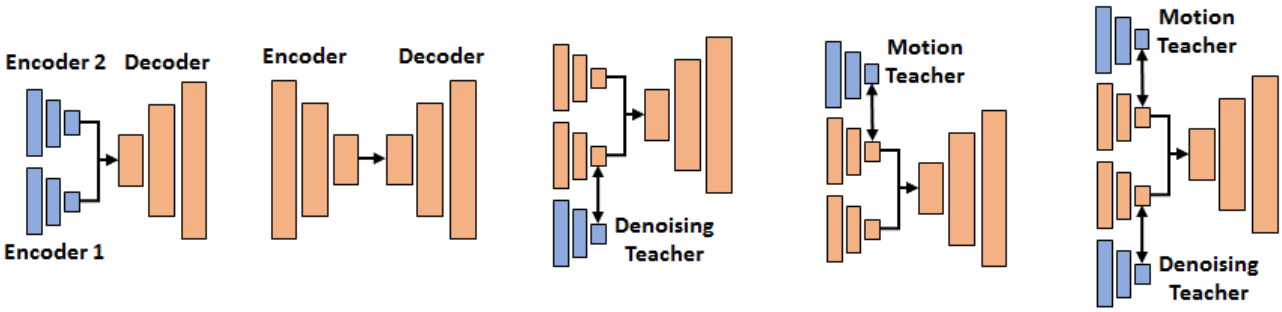


(a) Real image by QIS
1 frame, 1.5 ppp

(b) Real image by QIS
avg of 8 frames, 1.5 ppp

(c) Our reconstruction
using 8 QIS frames

Fig. 11. **Real QIS data with rotational motion**. The image is captured at 1.5 ppp. Notice the rotation blur in the 8-frame average, and the reconstructed result.



(a) Config A (b) Config B (c) Ours-I (d) Ours-II (e) Ours-full

Fig. 12. **Configurations for ablation study**. (a) Config-A: Uses pre-trained teachers. (b) Config-B: Uses a single encoder instead of two smaller encoders. (c) Ours-I: Uses denoising teacher only. (d) Ours-II: Uses motion teacher only. (e) Our-full: The complete model. In this figure, blue colored layers are pre-traiend and fixed. Orange layers are trainable.

ppp and motion of 28 pixels across 8 frames. The results are summarized in Table 1.

Is student-teacher training necessary? Configurations A and B do not use any teacher. Comparing with Ours-full, the PSNR values of Config A and Config B are worse by more than 1dB. Even if we compare with a single teacher, e.g., Ours-I, it is still 0.8dB ahead of Config B. Therefore, the student-teacher training protocol has a positive impact on performance.

Do teacher encoders extract meaningful information? Config A uses two pretrained encoders and a trainable decoder. The network achieves 21.51dB, which means that some features are useful for reconstruction. However, when comparing with Ours-full, it is substantially worse (23.87dB compared to 21.51dB). Since the network architectures are identical, the performance gap is likely caused by the training protocol. This indicates that the student-teacher setup is a better way to transfer knowledge from teachers to a student network.

Which teacher to use? Ours-I and Ours-II both use one teacher. The results suggest that if we only use one teacher, the motion teacher has a small gain (0.1dB) over the denoising teacher. However, if we use both teachers as in the proposed method, we observe another 0.2dB improvement. Thus, the presence of both teachers is helpful.

5 Conclusion

Dynamic low-light imaging is an important capability in application such as autonomous driving, security, and health care. CMOS image sensors (CIS) have fundamental limitations due to their inability to count photons. This paper considers Quanta Image Sensors (QIS) as an alternative solution. By developing a deep neural network using a new student-teacher training protocol, we demonstrated the effectiveness of transferring knowledge from a motion teacher and a denoising teacher to the student network. Experimental results indicate that the proposed method outperforms existing solutions trained under the same conditions. The proposed student-teacher protocol can also be applied to CIS problems. However, at a photon level of 1 photon per pixel or lower, QIS are necessary. Future work will focus on generalizing the reconstruction to more complex motions.

References

[1] Jiaju Ma and Eric Fossum, "A pump-gate jot device with high conversion gain for a Quanta Image Sensor," *IEEE Journal of Electron Devices Soc.*, vol. 3, no. 2, pp. 73–77, January 2015.

[2] Jiaju Ma, Saleh Masoodian, Dakota Starkey, and Eric R Fossum, "Photon-number-resolving megapixel image sensor at room temperature without avalanche gain," *Optica*, vol. 4, no. 12, pp. 1474–1481, December 2017.

Configuration	# of Encoders	Which Teacher?	Test PSNR
A	2	None	21.51 dB
B	1	None	22.74 dB
Ours-I	2	Denoising	23.53 dB
Ours-II	2	Motion	23.65 dB
Ours-full	2	Both	23.87 dB

TABLE 1
Ablation Study Results. This table summarizes the influence of different teachers on the proposed method. The experiments are conducted using synthetic data, at a photon level of 1 ppp and a motion of 28 pixels along the dominant direction.

[3] Samuel Burri, Yuki Maruyama, Xavier Michalet, Francesco Regazzoni, Claudio Bruschini, and Edoardo Charbon, "Architecture and applications of a high resolution gated SPAD image sensor," *Optics Express*, vol. 22, no. 14, pp. 17573–17589, 2014.

[4] Abhiram Gnanasambandam, Jiaju Ma, and Stanley H. Chan, "High Dynamic Range imaging using Quanta Image Sensors," in *Intl. Image Sensors Workshop*, 2019.

[5] Omar A. Elgendy and Stanley H. Chan, "Optimal threshold design for Quanta Image Sensor," *IEEE Trans. Computational Imaging*, vol. 4, no. 1, pp. 99–111, December 2017.

[6] Abhiram Gnanasambandam, Omar Elgendy, Jiaju Ma, and Stanley H. Chan, "Megapixel photon-counting color imaging using Quanta Image Sensor," *Optics Express*, vol. 27, no. 12, pp. 17298–17310, June 2019.

[7] Omar A. Elgendy and Stanley H. Chan, "Color Filter Arrays for Quanta Image Sensors," *arXiv preprint arXiv:1903.09823*, 2019.

[8] Istvan Gyongy, Neale Dutton, and Robert Henderson, "Single-photon tracking for high-speed vision," *Sensors*, vol. 18, no. 2, pp. 323, January 2018.

[9] Eric R Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors," *IEEE Journal of the Electron Devices Society*, vol. 1, no. 9, pp. 166–174, 2013.

[10] Feng Yang, Yue M. Lu, Luciano Sbaiz, and Martin Vetterli, "An optimal algorithm for reconstructing images from binary measurements," in *Proc. SPIE*, 2010, vol. 7533.

[11] Paramanand Chandramouli, Samuel Burri, Claudio Bruschini, Edoardo Charbon, and Andreas Kolb, "A bit too much? High speed imaging from sparse photon counts," in *ICCP*, 2019.

[12] Eric R. Fossum, "Some thoughts on future digital still cameras," in *Image sensors and signal processing for digital still cameras*, 2006.

[13] Eric R. Fossum, "Gigapixel digital film Sensor (DFS) proposal," *Nanospace Manipulation of Photons and Electrons for Nanovision Systems*, 2005.

[14] Feng Yang, Yue M Lu, Luciano Sbaiz, and Martin Vetterli, "Bits from photons: Oversampled image acquisition using binary poisson statistics," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1421–1436, 2011.

[15] Neale AW Dutton, Istvan Gyongy, Luca Parmesan, Salvatore Gnecchi, Neil Calder, Bruce R Rae, Sara Pellegrini, Lindsay A Grant, and Robert K Henderson, "A SPAD-based QVGA image sensor for single-photon counting and quanta imaging," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 189–196, 2015.

[16] Neale AW Dutton, Luca Parmesan, Andrew J Holmes, Lindsay A Grant, and Robert K Henderson, "320× 240 oversampled digital single photon counting image sensor," in *2014 Symposium on VLSI Circuits Digest of Technical Papers*, 2014.

[17] Anant Gupta, Atul Ingle, and Mohit Gupta, "Asynchronous single-photon 3D imaging," in *ICCV*, October 2019.

[18] Matthew O'Toole, Felix Heide, David B Lindell, Kai Zang, Steven Diamond, and Gordon Wetzstein, "Reconstructing transient images from single-photon sensors," in *CVPR*, 2017.

[19] David B Lindell, Matthew O'Toole, and Gordon Wetzstein, "Single-photon 3D imaging with deep sensor fusion," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–12, 2018.

[20] Clara Callenberg, Ashley Lyons, Dennis den Brok, Robert Henderson, Matthias B Hullin, and Daniele Faccio, "EMCCD-SPAD camera data fusion for high spatial resolution time-of-flight imaging.," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2019.

[21] Genevieve Gariepy, Nikola Krstajić, Robert Henderson, Chunyong Li, Robert R Thomson, Gerald S Buller, Barmak Heshmat, Ramesh Raskar, Jonathan Leach, and Daniele Faccio, "Single-photon sensitive light-in-fight imaging," *Nature communications*, vol. 6, no. 1, pp. 1–7, 2015.

[22] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Brushini, Edoardo Charbon, and Mohit Gupta, "Quanta burst photography," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, Jul. 2020.

[23] Feng Yang, Luciano Sbaiz, Edoardo Charbon, Sabine Süsstrunk, and Martin Vetterli, "Image reconstruction in the gigavision camera," in *ICCV Workshops*, 2009.

[24] Stanley H. Chan and Yue M. Lu, "Efficient image reconstruction for gigapixel Quantum Image Sensors," in *IEEE Global Conf. Signal and Info. Process.*, 2014.

[25] Stanley H. Chan, Xiran Wang, and Omar A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Computational Imaging*, vol. 3, no. 1, pp. 84–98, November 2016.

[26] Stanley H. Chan, Omar A. Elgendy, and Xiran Wang, "Images from bits: Non-iterative image reconstruction for Quanta Image Sensors," *Sensors*, vol. 16, no. 11, pp. 1961, November 2016.

[27] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, July 2007.

[28] Joon Hee Choi, Omar A. Elgendy, and Stanley H. Chan, "Image reconstruction for Quanta Image Sensors using deep neural networks," in *ICASSP*, 2018.

[29] Tal Remez, Or Litany, and Alex Bronstein, "A picture is worth a billion bits: Real-time image reconstruction from dense binary threshold pixels," in *ICCP*, 2016.

[30] Ben Mildenhall, Jonathan Barron, Jiawen Chen, et al., "Burst denoising with kernel prediction networks," in *CVPR*, 2018.

[31] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[32] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, "A review of image denoising algorithms, with a new one," *SIAM Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, July 2005.

[33] Markku Makitalo and Alessandro Foi, "Optimal inversion of the Anscombe transformation in low-count Poisson image denoising," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 99–109, July 2010.

[34] Henrik Malm, Magnus Oskarsson, Eric Warrant, et al., "Adaptive enhancement and noise reduction in very low light-level video," in *ICCV*, 2007.

[35] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang, "Deblurring low-light images with light streaks," in *CVPR*, 2014.

[36] Xiaojie Guo, Yu Li, and Haibin Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, October 2016.

[37] Qingtao Fu, Cheolkon Jung, and Kaiqiang Xu, "Retinex-based perceptual contrast enhancement in images using luminance adaptation," *IEEE Access*, vol. 6, pp. 61277–61286, October 2018.

[38] Tal Remez, Or Litany, Raja Giryes, and Alex Bronstein, "Deep convolutional denoising of low-light images," *arXiv preprint arXiv:1701.01687*, 2017.

[39] Kai Zhang, Wangmeng Zuo, Yunjin Chen, et al., "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, February 2017.

[40] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, January 2017.

[41] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, May 2018.

[42] Tobias Plotz and Stefan Roth, "Benchmarking denoising algorithms with real photographs," in *CVPR*, 2017.

[43] Jun Xu, Hui Li, Zhetong Liang, et al., "Real-world noisy image denoising: A new benchmark," *arXiv preprint arXiv:1804.02603*, 2018.

[44] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun, "Learning to see in the dark," in *CVPR*, 2018.

[45] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun, "Seeing motion in the dark," in *ICCV*, 2019.

[46] Axel Davy, Thibaud Ehret, Jean-Michel Morel, et al., "A non-local CNN for video denoising," in *ICIP*, 2019.

[47] Camille Sutour, Charles-Alban Deledalle, and Jean-François Aujol, "Adaptive regularization of the NL-means: Application to image and video denoising," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3506–3521, 2014.

[48] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, "Denoising image sequences does not require motion estimation," in *IEEE Conf. Adv. Video and Signal Based Surveillance*, 2005, pp. 70–74.

[49] Ce Liu and William Freeman, "A high-quality video denoising algorithm based on reliable motion estimation," in *ECCV*, 2010.

[50] Manuel Werlberger, Thomas Pock, Markus Unger, and Horst Bischof, "Optical flow guided TV-L 1 video interpolation and restoration," in *Intl. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2011.

[51] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun, "Fast burst images denoising," *ACM Trans. Graphics*, vol. 33, no. 6, pp. 232, November 2014.

[52] Matan Protter and Michael Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–35, 2008.

[53] Hui Ji, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu, "Robust video denoising using low rank matrix completion," in *CVPR*, 2010.

[54] 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 Trans. Graphics*, vol. 35, no. 6, pp. 192, November 2016.

[55] Neel Joshi and Michael Cohen, "Seeing Mt. Rainier: Lucky imaging for multi-image denoising, sharpening, and haze removal," in *ICCP*, 2010.

[56] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti, "Basis prediction networks for effective burst denoising with large kernels," *arXiv preprint arXiv:1912.04421*, 2019.

[57] Clément Godard, Kevin Matzen, and Matt Uyttendaele, "Deep burst denoising," in *ECCV*, 2018.

[58] Filippos Kokkinos and Stamatis Lefkimmiatis, "Iterative residual CNNs for burst photography applications," in *CVPR*, 2019.

[59] Miika Aittala and Frédo Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *ECCV*, 2018.

[60] Berthold KP Horn and Brian G Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*. Intl. Society Optics and Photonics, 1981, vol. 281.

[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," *Intl. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[62] Stephen Gould, Richard Fulton, and Daphne Koller, "Decomposing a scene into geometric and semantically consistent regions," in *ICCV*. IEEE, 2009, pp. 1–8.

[63] Matteo Maggioni, Vladimir Katkovnik, Karen Egiazarian, and Alessandro Foi, "Nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 119–133, 2012.

[64] François Chollet et al., "Keras," https://keras.io, 2015.

[65] Martín Abadi, Ashish Agarwal, Paul Barham, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.