Image Classification in the Dark using Quanta Image Sensors

Abhiram Gnanasambandam, and Stanley H. Chan, *Senior Member, IEEE*

Abstract—State-of-the-art image classifiers are trained and tested using well-illuminated images. These images are typically captured by CMOS image sensors with at least tens of photons per pixel. However, in dark environments when the photon flux is low, image classification becomes difficult because the measured signal is suppressed by noise. In this paper, we present a new low-light image classification solution using Quanta Image Sensors (QIS). QIS are a new type of image sensors that possess photon counting ability without compromising on pixel size and spatial resolution. Numerous studies over the past decade have demonstrated the feasibility of QIS for low-light imaging, but their usage for image classification has not been studied. This paper fills the gap by presenting a student-teacher learning scheme which allows us to classify the noisy QIS raw data. We show that with student-teacher learning, we are able to achieve image classification at a photon level of one photon per pixel or lower. Experimental results verify the effectiveness of the proposed method compared to existing solutions.

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

◆

1 INTRODUCTION

Quanta Image Sensors (QIS) are a type of single-photon image sensors originally proposed by E. Fossum as a candidate solution for the shrinking full-well capacity problem of the CMOS image sensors (CIS) [1], [2]. Compared to the CIS which accumulate photons to generate signals, QIS have a different design principle which partitions a pixel into many tiny cells called the jots with each jot being a single-photon detector. By oversampling the space and time, and by using a carefully designed image reconstruction algorithm, QIS can capture very low-light images with signal-to-noise ratio much higher than existing CMOS image sensors of the same pixel pitch [3]. Over the past few years, prototype QIS have been built by researchers at Dartmouth and Gigajot Technology Inc. [4], [5], with a number of theoretical and algorithmic contributions by researchers at EPFL [6], [7], Harvard [8], and Purdue [9]–[13]. Today, the latest QIS prototype can perform color imaging with a read noise of $0.25e^-/\text{pix}$ (compared to at least several electrons in CIS [14]) and dark current of $0.068e^-/\text{pix/s}$ at room temperature (compared to $> 1e^-/\text{pix/s}$ in CIS) [5], [9].

While prior work has demonstrated the effectiveness of using QIS for low-light image formation, there is no systematic study of how QIS can be utilized to perform better image classification in the dark. The goal of this paper is to fill the gap by proposing the first QIS image classification solution. Our proposed method is summarized in Figure 1. Compared to the traditional CIS-based low-light image classification framework, our solution leverages the unique single-photon sensing capability of QIS to acquire very low-light photon count images. We do not use any image processing, and directly feed the raw Bayer QIS data into our classifier. Our classifier is trained using a novel student-teacher learning protocol, which allows us to transfer knowledge from a teacher classifier to a student classifier. We show that the student-teacher protocol can effectively alleviate the need for a deep image denoiser as in the traditional frameworks. Our experiments demonstrate that the proposed method performs better than the existing solutions. The overall system – QIS combined with student-teacher learning – can achieve image classification on real data at 1 photon per pixel or lower. To summarize, the two contributions of this paper are:

(i) The introduction of student-teaching learning for low-light image classification problems. The experiments show that the proposed method outperforms existing approaches.

(ii) The first demonstration of image classification at a photon level of 1 photon per pixel or lower, on real images. This is a very low photon level compared to other results reported in the image classification literature.

2 BACKGROUND

2.1 Quanta Image Sensor

Quanta Image Sensors are considered as one of the most promising candidates for the third generation image sensors after CCD and CMOS. Fabricated using the commercial 3D stacking technology, the current sensor has a pixel pitch of $1.1\mu m$, with even smaller sensors being developed. The advantage of QIS over the conventional CMOS image sensors is that at $1.1\mu m$, the read noise of QIS is as low as $0.25e^-$ whereas a typical $1\mu m$ CMOS image sensor is at least several electrons. This low read noise (and also the low dark current) is made possible by the unique non-avalanche design [5] so that pixels can be packed together without
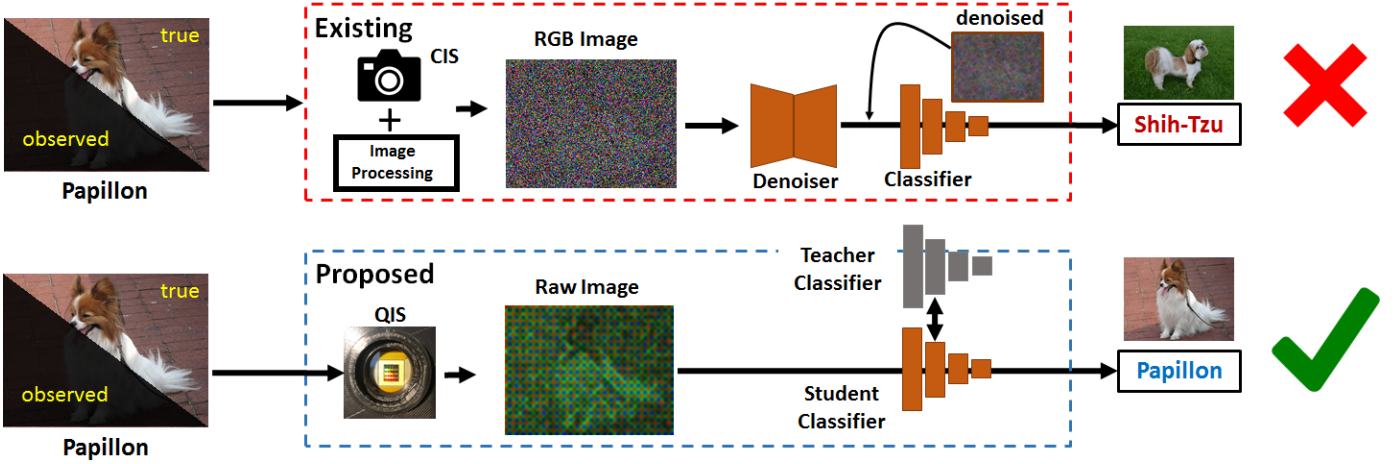
Fig. 1. [Top] Traditional image classification methods are based on CMOS image sensors (CIS), followed by a denoiser-classifier pipeline. [Bottom] The proposed classification method comprises a novel image sensor QIS and a novel student-teacher learning protocol. QIS generates significantly stronger signals, whereas student-teacher learning improves the robustness against noise.

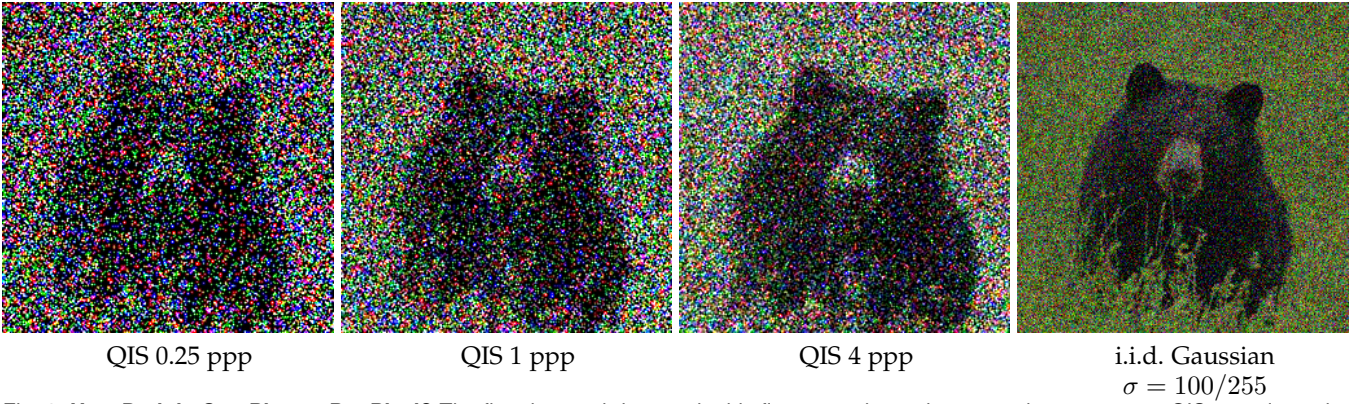| QIS 0.25 ppp | QIS 1 ppp | QIS 4 ppp | i.i.d. Gaussian $\sigma = 100/255$ |

Fig. 2. **How Dark is One Photon Per Pixel?** The first three sub-images in this figure are the real captures by a prototype QIS at various photon levels. The last sub-image is a simulation using additive i.i.d. Gaussian noise of a level of $\sigma = 100/255$, which is often considered as heavy degradation in the denoising literature. Additional examples can be found in Figure 10.

causing strong stray capacitance. The non-avalanche design also differentiates QIS from single photon avalanche diodes (SPAD). SPADs are typically bigger in size $> 5\mu$m, have lower fill factor $< 70\%$, have lower quantum efficiency $< 50\%$, and have significantly higher dark count $> 10e^-$. See [9] for a detailed comparison between CIS, SPAD and QIS. In general, SPADs are useful for applications such as time-of-flight imaging because of their speed [15]–[18], although new results in HDR imaging has been reported [19], [20]. QIS have better resolution and works well for passive imaging.

2.2 How Dark is One Photon Per Pixel?

When we say low-light imaging, it is important to clarify the photon level. The photon level is usually measured in the terms of lux. However, a more precise definition is the unit of photons per pixel (ppp). "Photons per pixel" is the average number of photons a pixel sees 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 inherently detects more photons, so does long exposure. For example, under the same low-light condition, images formed by the Keck telescope (aperture diameter = 10m) certainly

has better signal-to-noise than an iPhone camera (aperture diameter = 4.5mm). A high-end 3.5μm camera today has a read noise greater than $2e^-$ [21]. Thus, our benchmark choice of 1 ppp is approximately half of the read noise of a high-end sensor today. To give readers an idea of the amount of noise we should expect to see at 1 ppp, we show a set of real QIS images in Figure 2. Signals at 1 ppp is significantly worse than the so called "heavy noise" images we observe in the denoising literature and the low-light classification literature.

2.3 Prior Work

Quanta Image Sensors. QIS were proposed in 2005, and since then significant progresses have been made over the past 15 years. Readers interested in the sensor development can consult recent keynote reports, e.g., [22]. On the algorithmic side, a number of theoretical signal processing results and reconstruction algorithms have been proposed [3], [9], [23], including some very recent methods based on deep learning [7], [11]. However, since the sensor is relatively new, computer vision applications of the sensor are not yet common. To the best of our knowledge, the only available method for tracking applications is [24].
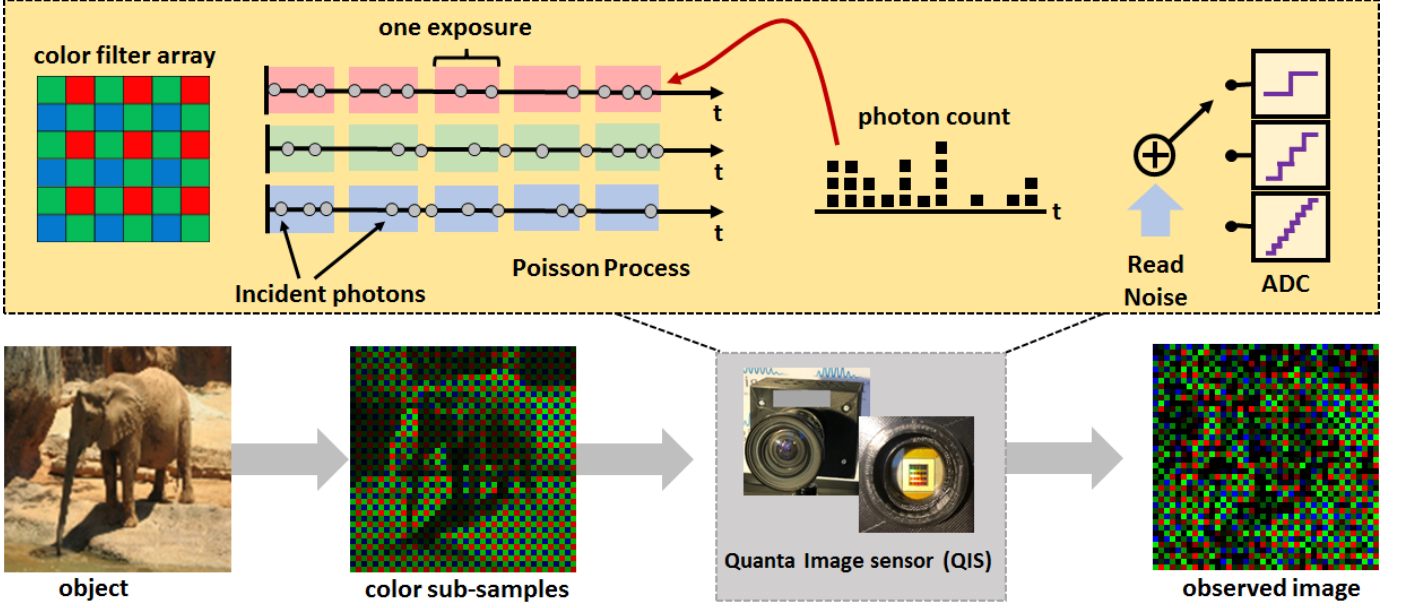
Fig. 3. **QIS Image Formation Model**. The basic image formation of QIS consists of a color filter array, a Poisson process, read noise, and an analog-to-digital converter (ADC). Additional factors are summarized in (1).

Low-light Classification. Majority of the existing work in classification is based on well-illuminated CMOS images. The first systematic study of the feasibility of low-light classification was presented by Chen and Perona [25], who observed that low-light classification is achievable by using a few photons. In the same year, Diamond et al. [26] proposed the "Dirty Pixels" method by training a denoiser and a classifier simultaneously. They observed that less aggressive denoisers are better for classification because the features are preserved. Other methods adopt similar strategies, e.g., using discrete cosine transform [27], training a classifier to help denoising [28] or using an ensemble method [29], or training a denoiser that are better suited for pre-trained classifiers [30], [31].

Low-light Reconstruction. A closely related area of low-light classification is low-light reconstruction, e.g., denoising. Classical low-light reconstruction usually follows the line of Poisson-based inverse problems [32] and contrast enhancement [33]–[36]. Deep neural network methods have recently become the main driving force [37]–[42], including the recent series on "seeing in the dark" by Chen et al. [43], [44]. Burst photography [45]–[48] (with some older work in [49]–[51]) is related but not directly applicable to us since the methods are developed for multi-frame problems.

3 METHOD

The proposed method comprises QIS and a novel student-teacher learning scheme. In this section, we first discuss how images are formed by QIS. We will then present the proposed student-teacher learning scheme which allows us to overcome the noise in QIS measurements.

3.1 QIS Image Formation Model

The image formation model is shown in Figure 3. Given an object in the scene (x_{rgb}), we use a color filter array (CFA) to bandpass the light to subsample the color. Depending on the exposure time and the size of the jots, a sensor gain α is applied to scale the sub-sampled color pixels. The photon arrival is simulated using a Poisson model. Gaussian noise is added to simulate the read noise arising from the circuit. Finally, an analog-to-digital converter (ADC) is used to truncate the real numbers to integers depending on the number of bits allocated by the sensor. For example, a single-bit QIS will output two levels, whereas multi-bit QIS will output several levels. In either case, the signal is clipped to take value in $\{0, 1, \ldots L\}$, where L represents the maximum signal level. The image formation process can be summarized using the following equation

$$\underbrace{x_{\text{QIS}}}_{\mathbb{R}^{M \times N}} = \text{ADC}_{[0,L]}\left\{ \underbrace{\text{Poisson}}_{\text{photon arrival}} \left(\underbrace{\alpha}_{\text{sensor gain}} \right. \right. \tag{1}$$
$$\left. \left. \times \text{CFA}\left(\underbrace{x_{\text{rgb}}}_{\mathbb{R}^{M \times N \times 3}} \right) \right) + \underbrace{\eta}_{\text{read noise}} \right\},$$

In addition to the basic image formation model described in (1), two other components are included in the simulations. First, we include the dark current which is an additive noise term to $\alpha \cdot \text{CFA}(x_{\text{rgb}})$. The typical dark current of the QIS is $0.068e^-/\text{pix/s}$. Second, we model the pixel response non-uniformity (PRNU). PRNU is a pixel-wise multiplication applied to x_{rgb}, and is unique for every sensor. Readers interested in details on the image formation model and statistics can consult previous works such as [3], [13], [52], [53].

3.2 Student-Teacher Learning

Inspecting (1), we notice that even if the read noise η is zero, the random Poisson process will still create a fundamental limit due to the shot noise in x_{QIS}. Therefore, when applying a classification method to the raw QIS data, some capability of removing the shot noise becomes necessary. Traditional solution to this problem (in the context of CIS) is to denoise
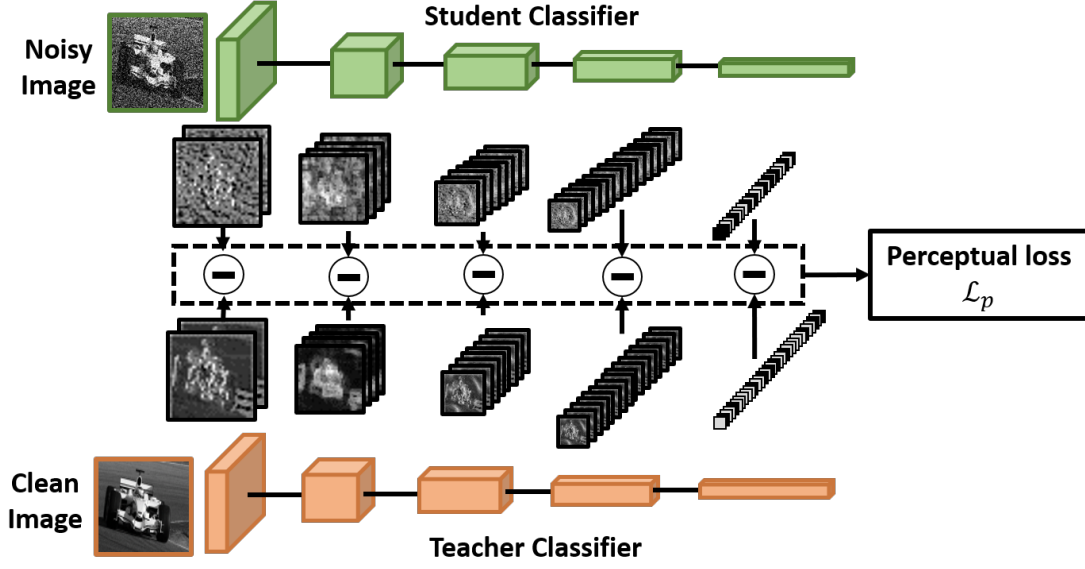
Fig. 4. **Student-Teacher Learning**. Student-teacher learning comprises two networks: A teacher network and a student network. The teacher network is pre-trained using clean samples whereas the student is trained using noisy samples. To transfer knowledge from the teacher to the student, we compare the features extracted by the teacher and the student at different stages of the network. The difference between the features is measured as the perceptual loss.
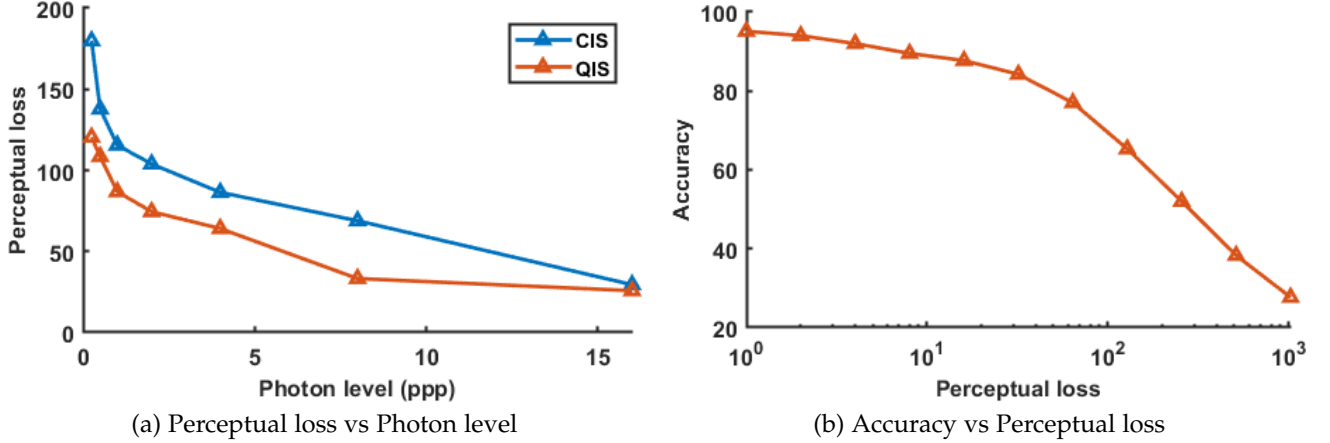

(a) Perceptual loss vs Photon level

(b) Accuracy vs Perceptual loss

Fig. 5. **Effectiveness of Student-Teacher Learning**. (a) Perceptual loss as a function of photon level. (b) Classification accuracy as a function of the perceptual loss $\mathcal{L}_p(\boldsymbol{x}_{\mathrm{QIS}}, \boldsymbol{x}_{\mathrm{rgb}})$. The accuracy is measured by repeating the synthetic experiment described in the Experiment Section. The negative correlation suggests that perceptual loss is indeed an influential factor.

the images as shown in the top of Figure 1. The objective of this section is to introduce an alternative approach using the concept of student-teacher learning.

The idea of student-teacher learning can be understood from Figure 4. There are two networks in this figure: A teacher network and a student network. The teacher network is trained using *clean* samples, and is pre-trained, i.e., its network parameters are fixed during training of the student network. The student network is trained using *noisy* samples with the assistance from the teacher. Because the teacher is trained using clean samples, the features extracted are in principle "good", in contrast to the features of the student which are likely to be "corrupted". Therefore, in order to transfer knowledge from the teacher to the student, we propose minimizing a *perceptual loss* as defined below. We define the j-th layer's feature of the student network as $\phi^j(\boldsymbol{x}_{\mathrm{QIS}})$, where $\phi^j(\cdot)$ maps $\boldsymbol{x}_{\mathrm{QIS}}$ to a feature vector, and we define $\widehat{\phi}^j(\boldsymbol{x}_{\mathrm{rgb}})$ as the feature vector extracted by the teacher

network. The perceptual loss is

$$\mathcal{L}_{\mathrm{p}}(\boldsymbol{x}_{\mathrm{QIS}}, \boldsymbol{x}_{\mathrm{rgb}}) = \sum_{j=1}^{J} \underbrace{\frac{1}{N_j} \left\| \widehat{\phi}^j(\boldsymbol{x}_{\mathrm{rgb}}) - \phi^j(\boldsymbol{x}_{\mathrm{QIS}}) \right\|^2}_{j\text{-th layer's perceptual loss}}, \quad (2)$$

where N_j is the dimension of the j-th feature vector. Since the perceptual loss measures the distance between the student and the teacher, minimizing the perceptual loss forces them to be close. This, in turn, forces the network to "denoise" the shot noise and read noise in $\boldsymbol{x}_{\mathrm{QIS}}$ before predicting the label.

We conduct a simple experiment to demonstrate the impact of input noise to perceptual loss and classification accuracy. We first consider a pre-trained teacher network by sending QIS data at different photon levels. As photon level drops, the quality of the features also drops and hence the perceptual loss increases. This is illustrated in Figure 5(a). Then in Figure 5(b), we evaluate the classification accu-
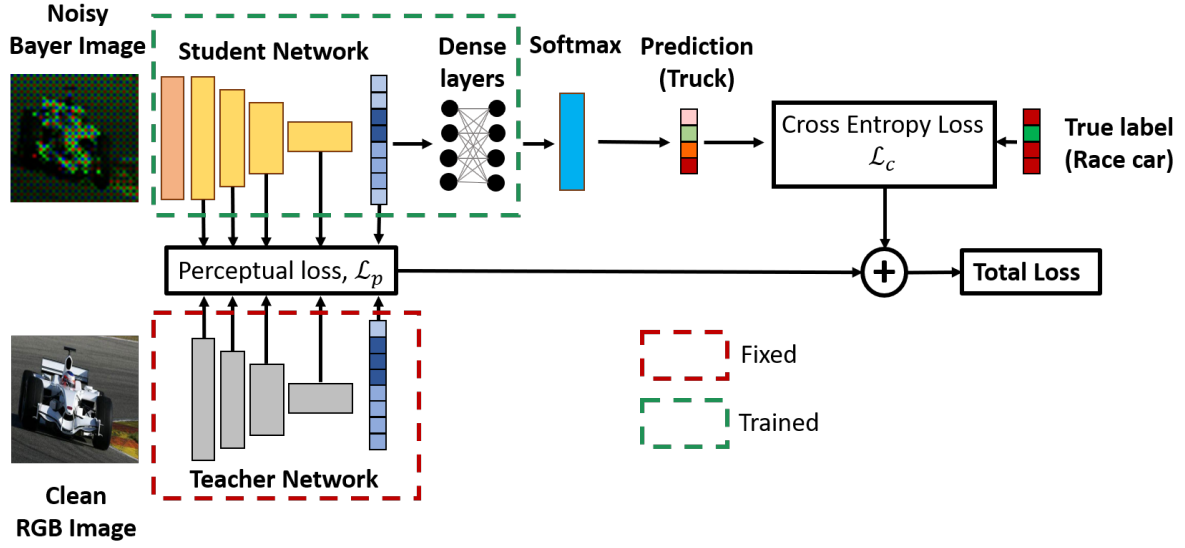
Fig. 6. **Proposed Method**. The proposed method trains a classification network with two training losses: (1) cross-entropy loss to measure the prediction quality, and (2) perceptual loss to transfer knowledge from teacher to student. During testing, only the student is used. We introduce a 2-layer entrance (colored in orange) for the student network so that the classifier can handle the Bayer image.

racy by using the synthetic testing data outlined in the Experiment Section. As the perceptual loss increases, the classification accuracy drops. This result suggests that if we minimize the perceptual loss then the classification accuracy can be improved.

Our proposed student-teacher learning is inspired by the knowledge distillation work of Hinton et al. [54] which proposed an effective way to compress networks. A number of follow up ideas have been proposed, e.g., [55]–[58], including the MobileNet [59]. The concept of perceptual loss has been used in various computer vision applications such as the texture-synthesis and style-transfer by Johnson et al. [60] and Gatys et al. [61], [62], among many others [4], [30], [63]–[67]. The method we propose here is different in the sense that we are not compressing the network. Also, we are not asking the student to mimic the teacher because the teacher and the student are performing two different tasks: The teacher classifies clean data, whereas the student classifies noisy data. In the context of low-light classification, student-teacher learning has not been applied.

3.3 Overall Method

The overall loss function comprises the perceptual loss and the conventional prediction loss using cross-entropy. The cross-entropy loss $\mathcal{L}_c$, measures the difference between true label y and the predicted label $f_\Theta(x_{\text{QIS}})$ generated by the student network, where f_Θ is the student network. The overall loss is mathematically described as

$$\mathcal{L}(\Theta) = \sum_{n=1}^{N} \left\{ \mathcal{L}_c\big(y^n, f_\Theta(x_{\text{QIS}}^n)\big) + \lambda \mathcal{L}_p\big(x_{\text{rgb}}^n, x_{\text{QIS}}^n\big) \right\}, \quad (3)$$

where x^n denotes the n-th training sample with the ground truth label y^n. During the training, we optimize the weights of the student network by solving

$$\widehat{\Theta} = \underset{\Theta}{\arg\min}\ \mathcal{L}(\Theta). \quad (4)$$

During testing, we feed a testing sample x_{QIS} to the student network and evaluate the output:

$$\widehat{y} = f_{\widehat{\Theta}}(x_{\text{QIS}}). \quad (5)$$

Figure 6 illustrates the overall network architecture. In this figure, we emphasize that training is done on the student only. The teacher is fixed and is not trainable. In this particular example, we introduce a very shallow network consisting of 2 convolution layers with 32 and 3 filters respectively. This shallow network is used to perform the necessary demosaicking by converting raw Bayer pattern to the full RGB before feeding into a standard classification network.

4 Experiments

4.1 Dataset

Dataset. We consider two datasets. The first dataset (Animal) contains visually distinctive images where the class labels are far apart. The second dataset (Dog) contains visually similar images where the class labels are fine-grained. The two different datasets can help differentiating the performance regime of the proposed method, and its benefits over other state-of-the-art networks.

The construction of the two datasets are as follows. For the Animal dataset, we randomly select 10 classes of animals from ImageNet [68], as shown in Figure 7(a). Each class contains 1300 images, giving a total of 13K images. Among which, 9K are used for training, 1K for validation, and 3K for testing. For the Dog dataset, we randomly select 10 classes of dogs from the Stanford Dog dataset [69], as shown in Figure 7(b). Each class has approximately 150 images, giving a total of 1919 images. We use 1148 for training, 292 for validation, and 479 for testing.

4.2 Competing Methods and Our Network

We compare our method with three existing low-light classification methods as shown in Figure 8. The three competing
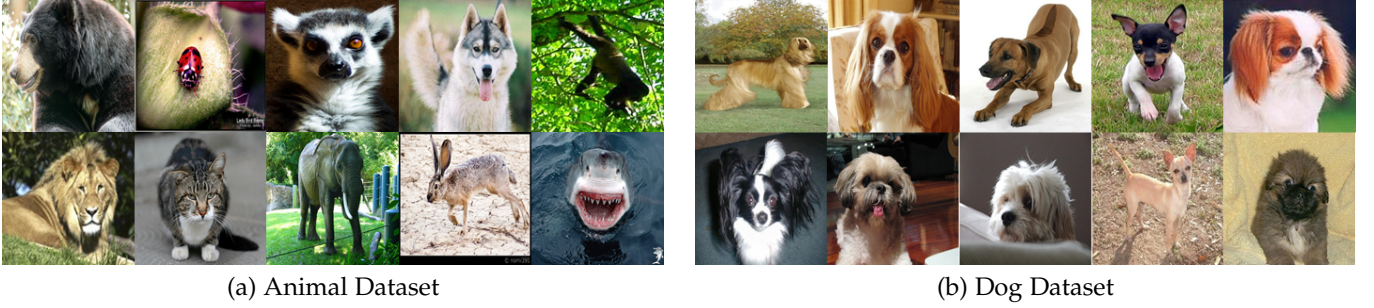
(a) Animal Dataset

(b) Dog Dataset

Fig. 7. The two datasets for our experiments.

(a) Vanilla Network

(b) Dirty Pixels [26]

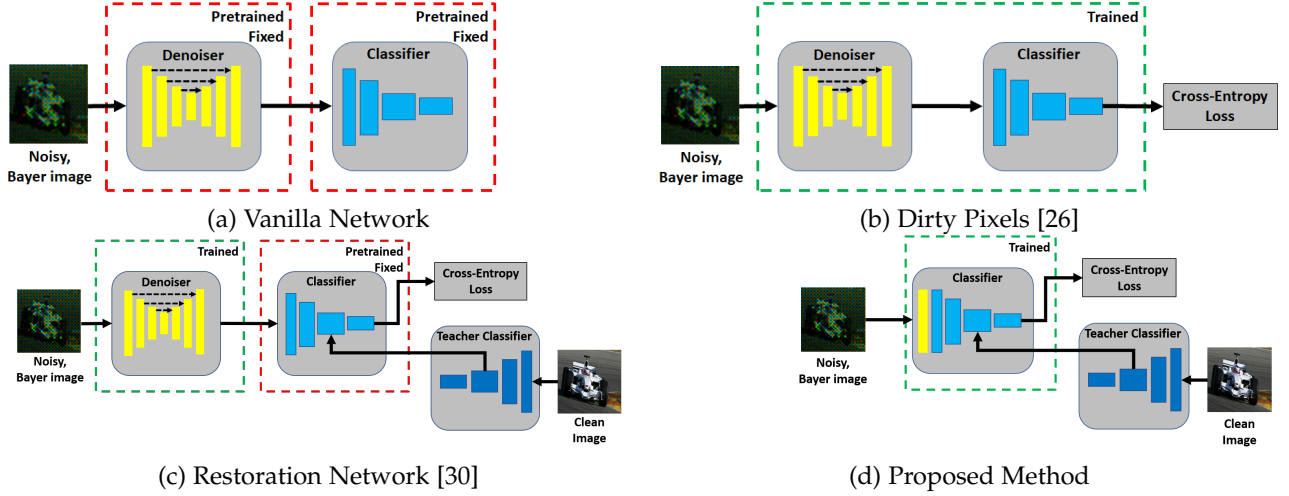(c) Restoration Network [30]

(d) Proposed Method

Fig. 8. **Competing Methods**. The major difference between the networks are the trainable modules and the loss functions. For Dirty Pixels and our proposed method, we further split it into two versions: Using a deep denoiser or using a shallow entrance network.

methods are (a) Vanilla denoiser + classifier, an "off-the-shelf" solution using pre-trained models but fine-tuned using QIS data. THe denoiser is pre-trained on the QIS data and the classifier is pre-trained on clean images. (b) Dirty Pixels [26], same as Vanilla denoiser + classifier, but trained end-to-end using the QIS noisy data. (c) Restoration Network [30], [31], which trains a denoiser but uses a pre-trained classifier. This can be viewed as a middle-ground solution between Vanilla and Dirty Pixels.

To ensure that the comparison is fair w.r.t. the training protocol and not the architecture, all classifiers in this experiment (including ours) use the same VGG-16 architecture. For methods that use a denoiser, the denoiser is fixed as a UNet. This particular combination of denoiser and classifier does have some influence to the final performance, but the effectiveness of the training protocol can still be observed. Combinations beyond the ones we report here can be found in the ablation study. For Dirty Pixels and our proposed method, we further split them into two versions: (i) Using a deep denoiser as the entrance, i.e., a 20-layer UNet, and (ii) using a shallow two-layer network as the entrance to handle the Bayer pattern, as we described in the proposed method section. We will analyze the influence of this component in the ablation study.

4.3 Synthetic Experiment

The first experiment is based on synthetic data. The training data are created by the QIS model. To simulate the QIS data, we follow Equation (1) by using the Poisson-Gaussian process. the read noise is $\sigma = 0.25e^-$ according to [5]. The analog-to-digital converter is set to 5 bits, so that the number of photons seen by the sensors is between 0 and 31. We use a similar simulation procedure for CIS with the difference being the read noise, which we set to $\sigma = 2.0e^-$ [21].

The experiments are conducted for 5 different photon levels corresponding to 0.25, 0.5, 1, 2, and 4 photons per pixel (ppp). The photon level is controlled by adjusting the value of the multiplier α in Equation (1). The loss function weights λ in Equation (3) is tuned for optimal performance.

The results of the synthetic data experiment are shown in Figure 9. In Figure 9(a), we observe that our proposed classification is consistently better than competing methods the photon levels we tested. Moreover, since all methods reported in Figure 9(a) are using QIS as the sensor, the curves in Figure 9(a) reveal the effectiveness of just the classification method. In Figure 9(b), we compare the difference between using QIS and CIS. As we expect, CIS has worse performance compared to QIS.

4.4 Real experiment

We conduct an experiment using real QIS and CIS data. The real QIS data are collected by a prototype QIS camera Gigajot PathFinder [9], whereas the real CIS data are collected by using a commercially available camera. To setup the experiment, we display the images on a Dell P2314H LED screen (60Hz). The cameras are positioned 1m from the
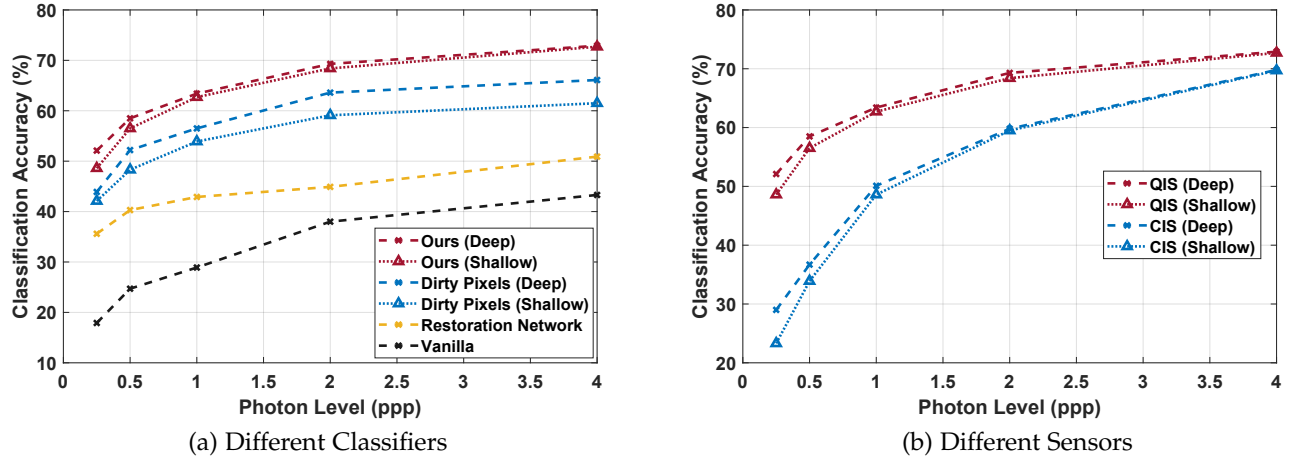
(a) Different Classifiers

(b) Different Sensors

Fig. 9. **Synthetic Data on Dog Dataset**. (a) Comparing different classification methods using QIS as the sensor. (b) Comparing QIS and CIS using our proposed classifier.
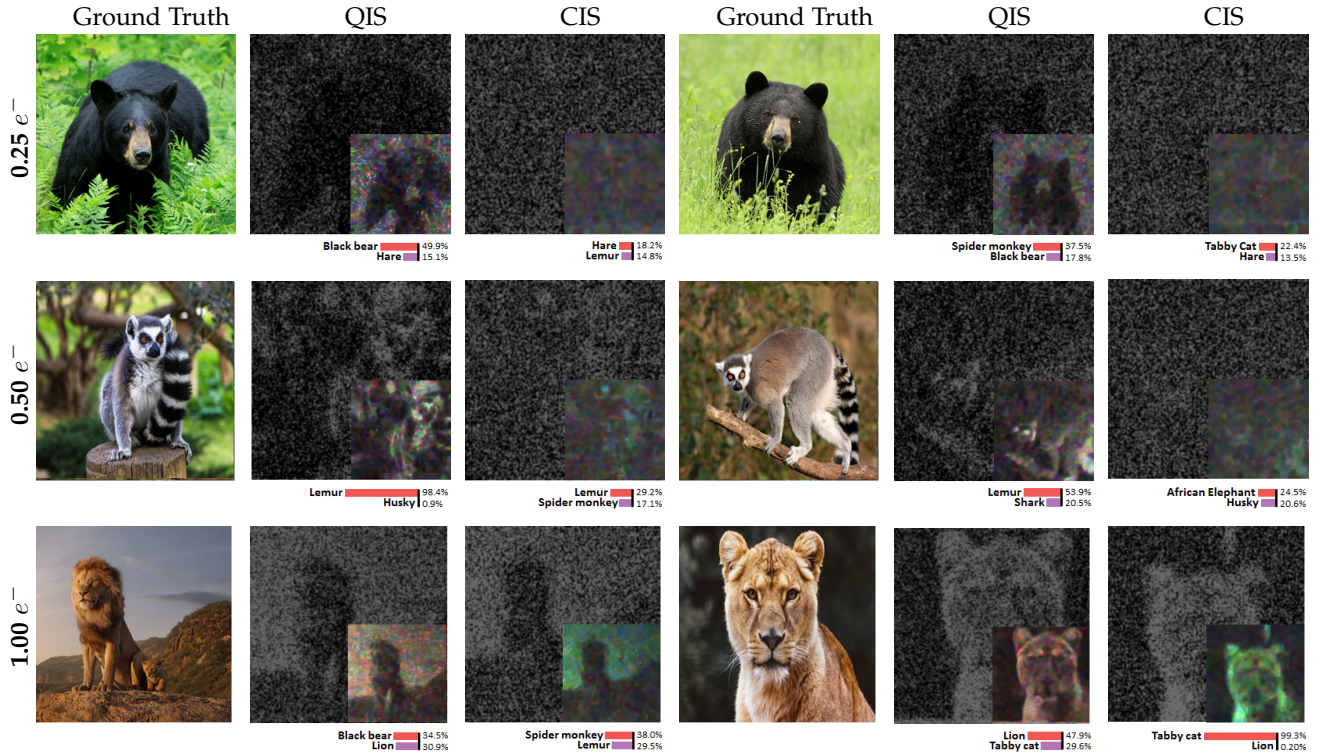

Fig. 10. **Real Image Results**. This figure shows raw Bayer data obtained from a prototype QIS and a commercially available CIS, and how they are classified using our proposed classifier. The inset images show the denoised images (by [43]) for visualization. Notice the heavy noise at 0.25 and 0.5 ppp, only QIS plus our proposed classification method can produce the correct prediction.

display so that the field of view covers 256×256 pixels of the image. The integration time of the CIS is set to $250\mu s$ and that of QIS is $75\mu s$. Since the CIS and QIS have different lenses, we control their aperture sizes and the brightness of the screen such that the average number of photons per pixel is equal for both sensors.

The training of the network in this real experiment is still done using the synthetic dataset, with the image formation model parameters matched with the actual sensor parameters. However, since the real image sensors have pixel non-uniformity, during the training we multiply a random PRNU mask to each of the generated images in order to mimic the process of PRNU. For testing, we collect 30 real images at each photon level, across 5 different photon levels. This corresponds to a total of 150 real testing images.

When testing, we make two pairs of comparisons: Proposed (shallow) versus Dirty Pixels (shallow), and QIS versus CIS. The result of the first experiment is shown in Figure 11(a), where we observe that the proposed method has a consistent improvement over Dirty Pixels. The comparison between QIS and CIS is shown in Figure 11(b). It is evident that QIS has better performance compared to CIS. Figure 10 shows the visualizations. The ground truth images were displayed on the screen, and the background
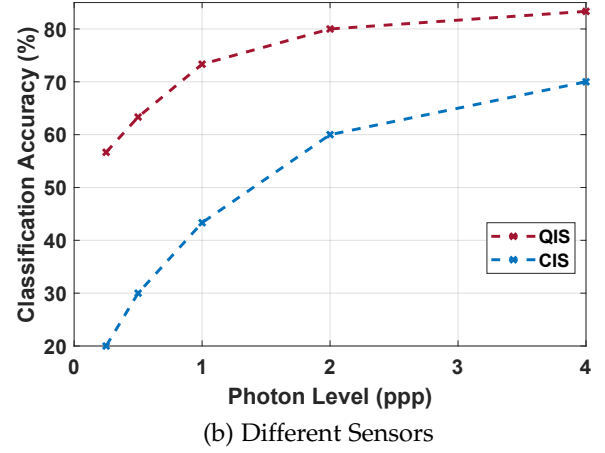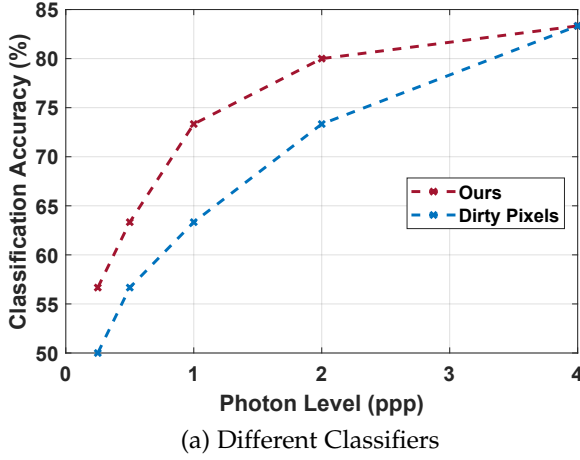
(a) Different Classifiers

(b) Different Sensors

Fig. 11. **Real data on Animal Dataset**. (a) Comparing different classification methods using QIS as the sensor. (b) Comparing QIS and CIS using our proposed classifier.

images in QIS and CIS column are actual measurements from the corresponding cameras, cropped to 256×256. The thumbnail images in the front are the denoised images for reference. They are not used during the actual classification. The color bars at the bottom report the confidence level of the predicted class. Note the significant visual difference between QIS and CIS, and the classification results.

4.5 Ablation Study

In this section we report several ablation study results, and highlight the most influencing factors to the design.

Sensor. Our first ablation study is to fix the classifier but change the sensor from QIS to CIS. This experiment will underline the impact of the sensor in the overall pipeline. The result of this ablation study can be seen from Figure 9(b). Specifically, at 4 ppp (high photon level) of the Dogs dataset, QIS + proposed has a classification accuracy of 72.9% while CIS has 69.8%. The difference is 3.1%. As the photon level drops, the gap between QIS and CIS this gap widens to 23.1% at 0.25 ppp. A similar trend is found in the Animals dataset. Thus at low light QIS has a clear advantage, although CIS can catch up when there is sufficient number of photons.

Classification Pipeline. The next ablation study is to fix the sensor but change the entire classification pipeline. This will tell us how important the classifier is, and which classifier is more effective. The results in Figure 9(a) show that the among the competing methods, Dirty Pixels is the most promising one because it is end-to-end trained. However, comparing Dirty Pixels with our proposed method, at 1 ppp Dirty Pixels (shallow) achieves an accuracy of 53.9% whereas the proposed (shallow) achieves 62.7%. The trend continues as the photon level increases. This ablation analysis shows that even if we have a good sensor (QIS), we still need a good classifier.

Student-Teacher Learning. Let us fix the sensor and the network, but change the training protocol. This will reveal the significance of the proposed student-teacher learning. To conduct this ablation study, we recognize that Dirty Pixels network structure (shallow and deep) is exactly the same as Ours (shallow and deep) since both use the same UNet

and VGG-16. The only difference is the training protocol, where ours uses student-teacher learning and Dirty Pixels is a simple end-to-end. The result of this study is summarized in Figure 9(a). It is evident that our training protocol offers advantages over Dirty Pixels.

We can further analyze the situation by plotting the training and validation error. Figure 12 shows the comparison between the proposed method (shallow) and Dirty Pixels (shallow). It is evident from the plot that without stident-teacher learning (Dirty Pixels), the network overfits. This is evident if we look at the the validation loss, which drops and then rises whereas the training loss keeps dropping. In contrast, the proposed method appears to mitigate the overfitting issue. One possible reason is that the student-teacher learning is providing some kind of regularization in an implicit form so that the validation loss is maintained at a low level.
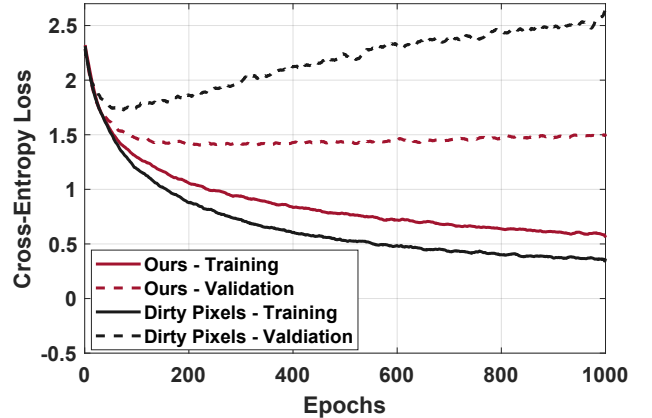

Fig. 12. Training and validation loss of Our method and Dirty Pixels. Notice that while our training loss is higher, the validation loss is significantly lower than Dirty Pixels.

Choice of Classification Network. All experiments reported in this paper use VGG-16 as the classifier. In this ablation study we replace the VGG-16 classifier by other popular classifiers, namely ResNet50 and InceptionV3. These networks are fine-tuned using the QIS data. Table 1 shows the

comparisons. Using the baseline training scheme, i.e., simple fine-tuning as in Dirty Pixels, it is observed that there is a minor gap between the different classifiers. However, by using the proposed student-teacher training protocol, we observe a substantial improvement for all the classifiers. This ablation study confirms that student-teacher learning is not limited to a particularly network architecture.

TABLE 1
Ablation study of different classifiers and different training schemes. Reported numbers are based on QIS synthetic experiments at 0.25 ppp for the Dog Dataset.

	VGG	ResNet	Inception
Baseline (fine-tuning)	42.1%	43.3%	44.3%
Proposed (student-teacher)	48.6%	49.1%	50.0%
Gap	+6.5%	+5.8%	+5.7%

Using a pre-trained classifier. This ablation study analyzes the effect of using a pre-trained classifier (trained on clean images). If we do this, then the overall system is exactly the same as the Restoration network [30] in Figure 8(c). Restoration network has three training losses: (i) MSE to measure the image quality, (ii) Perceptual loss to measure feature quality, and (iii) the cross-entropy loss. These three losses are used to just train the denoiser, and not the classifier. Since the classifier is fixed, it becomes necessary for the the denoiser to produce high-quality images or otherwise the classifier will not work. The results in Figure 9(a) suggest that when photon level is low, the denoiser fails to produce high quality images and so the classification fails. For example, at 0.25 ppp Restoration Network achieves 35.6% but our proposed method achieves 52.1%. Thus it is imperative that we re-train the classifier for low-light images.

Deep or Shallow Denoisers? This ablation study analyzes the impact of using a deep denoiser compared to a shallow entrance layer. The result of this study can be found by comparing Ours (deep) and Ours (shallow) in Figure 9(a), as well as Dirty (deep) and Dirty (shallow). In both methods, the deep version refers to using a 20-layer UNet, whereas the shallow version refers to using a 2-layer network. The result in Figure 9(a) suggests that while the deep denoiser has significant impact to Dirty Pixels, its influence is quite small to the proposed method with the QIS images. One reason is that since we are using student-teacher learning, the features are already properly denoised. The added benefit from a deep denoiser for QIS is therefore marginal. However, for CIS data at low light, the deep denoiser actually helps getting better classification performance, especially when the signal level is much lower than the read noise.

5 CONCLUSION

We proposed a new low-light image classification method by integrating Quanta Image Sensors (QIS) and a novel student-teacher training protocol. Experimental results confirmed that such combination is effective for low-light image classification, and the student-teacher protocol is a better alternative than the traditional denoise-then-classify framework. This paper also made the first demonstration of low-light image classification at a photon level of 1 photon per pixel or lower. The student-teacher training protocol is transferable to conventional CIS data, however to achieve the desired performance at low light, it is necessary for QIS to be part of the overall pipeline. Using multiple frames for image classification would be a fruitful direction for future work.

REFERENCES

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

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

[3] 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.

[4] Jiaju Ma and Eric Fossum, "A pump-gate jot device with high conversion gain for a quanta image sensor," *IEEE J. Electron Devices Soc.*, vol. 3, no. 2, pp. 73–77, January 2015.

[5] 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.

[6] Feng Yang, Yue M. Lu, Luciano Sbaiz, and Martin Vetterli, "An optimal algorithm for reconstructing images from binary measurements," in *Proc. SPIE*, January 2010, vol. 7533, pp. 158 – 169.

[7] 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.

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

[9] 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.

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

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

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

[13] 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.

[14] Boyd Fowler, Dan McGrath, and Peter Bartkovjak, "Read noise distribution modeling for CMOS Image Sensors," in *International Image Sensor Workshop*, 2013.

[15] 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.

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

[17] 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*, 2019.

[18] 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.

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

[20] 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.

[21] "FLIR Sensor Review: Mono Camera," https://www.flir.com/globalassets/iis/guidebooks/2019-machine-vision-emva1288-sensor-review.pdf.

[22] Eric R Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza, "The Quanta Image Sensor: Every photon counts," *Sensors*, vol. 16, no. 8, pp. 1260, 2016.

[23] 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.

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

[25] Bo Chen and Pietro Perona, "Seeing into darkness: Scotopic visual recognition," in *CVPR*, 2017.

[26] Steven Diamond, Vincent Sitzmann, Stephen Boyd, Gordon Wetzstein, and Felix Heide, "Dirty pixels: Optimizing image classification architectures for raw sensor data," *arXiv preprint arXiv:1701.06487*, 2017.

[27] Md Tahmid Hossain, Shyh Wei Teng, Dengsheng Zhang, Suryani Lim, and Guojun Lu, "Distortion robust image classification using deep convolutional neural network with discrete cosine transform," in *ICIP*, 2019.

[28] Jiqing Wu, Radu Timofte, Zhiwu Huang, and Luc Van Gool, "On the relation between color image denoising and classification," *arXiv preprint arXiv:1704.01372*, 2017.

[29] Samuel Dodge and Lina Karam, "Quality resilient deep neural networks," *arXiv preprint arXiv:1703.08119*, 2017.

[30] Ding Liu, Bihan Wen, Jianbo Jiao, Xianming Liu, Zhangyang Wang, and Thomas S Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3695–3706, 2020.

[31] Zhuang Liu, Tinghui Zhou, Zhiqiang Shen, Bingyi Kang, and Trevor Darrell, "Transferable recognition-aware image processing," *arXiv preprint arXiv:1910.09185*, 2019.

[32] 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.

[33] Henrik Malm, Magnus Oskarsson, Eric Warrant, Petrik Clarberg, Jon Hasselgren, and Calle Lejdfors, "Adaptive enhancement and noise reduction in very low light-level video," in *ICCV*, 2007.

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

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

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

[42] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang, "Real-world noisy image denoising: A new benchmark," *arXiv preprint arXiv:1804.02603*, 2018.

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

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

[45] 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.

[46] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll, "Burst denoising with kernel prediction networks," in *CVPR*, 2018.

[47] Axel Davy, Thibaud Ehret, Jean-Michel Morel, Pablo Arias, and Gabriele Facciolo, "A non-local CNN for video denoising," in *ICIP*, 2019.

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

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

[50] 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.

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

[52] 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.

[53] 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.

[54] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[55] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *NeurIPS*, 2014.

[56] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *CVPR*, 2018.

[57] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille, "Training deep neural networks in generations: A more tolerant teacher educates better students," in *AAAI Conf. Artificial Intelligence*, July 2019.

[58] Tianyu Guo, Chang Xu, Shiyi He, Boxin Shi, Chao Xu, and Dacheng Tao, "Robust student network learning," *IEEE Trans. Neural Networks and Learning Systems*, August 2019.

[59] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[60] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[61] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "Texture synthesis using convolutional neural networks," in *NeurIPS*, 2015.

[62] LA Gatys, AS Ecker, and M Bethge, "A neural algorithm of artistic style," *Nature Communications*, 2015.

[63] Gang Chen, Yawei Li, and Sargur N Srihari, "Joint visual denoising and classification using deep learning," in *ICIP*, 2016.

[64] Hossein Talebi and Peyman Milanfar, "Learned perceptual image enhancement," in *ICCP*, 2018.

[65] Aravindh Mahendran and Andrea Vedaldi, "Understanding deep image representations by inverting them," in *CVPR*, 2015.

[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *ICLR (Workshop Poster)*, 2014.

[67] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, "Understanding neural networks through deep visualization," *ICML Deep Learning Workshop*, 2015.

[68] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[69] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *CVPR Workshop on Fine-Grained Visual Categorization*, 2011, vol. 2.