Image Classification in the Dark using Quanta Image Sensors

Abhiram Gnanasambandam and Stanley H. Chan

Purdue University, West Lafayette, IN 47907, USA
{agnanasa,stanchan}@purdue.edu

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 can 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.

Keywords: Quanta Image Sensors, 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) [18,19]. 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. [48, 49], with a number of theoretical and algorithmic contributions by researchers at EPFL [6, 64], Harvard [4], and Purdue [11, 16, 17, 27, 28]. Today, the latest QIS prototype can perform color imaging with a read noise of $0.25e^-$/pix (compared to at least several electrons in CIS [22]) and dark current of $0.068e^-$/pix/s at room temperature (compared to $> 1e^-$/pix/s in CIS) [27, 49].

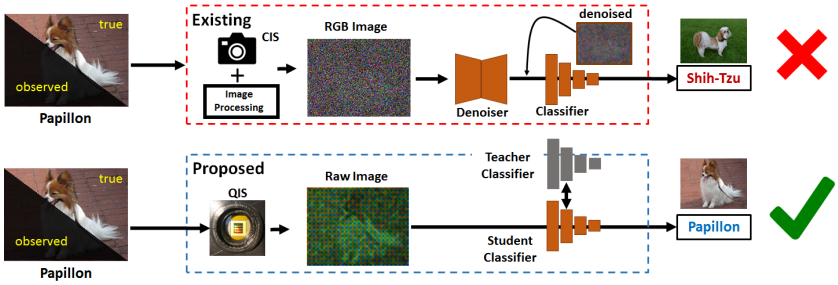


Fig. 1. [Top] Traditional image classification methods are based on CMOS image sensors (CIS), followed by a denoiser-classifier pipeline. [Bottom] The proposed classification method comprises a novel image sensor QIS and a novel student-teacher learning protocol. QIS generates significantly stronger signals, and student-teacher learning improves the robustness against noise.

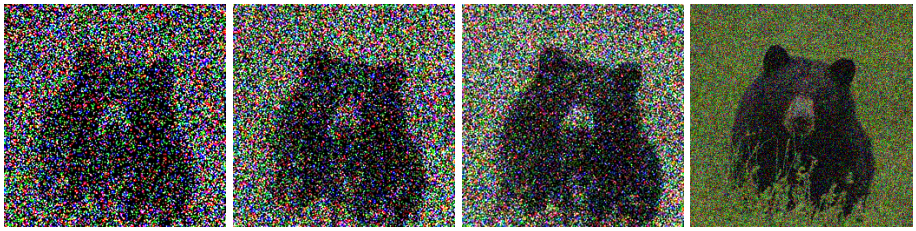
While prior works have demonstrated the effectiveness of using QIS for low-light image formation, there is no systematic study of how QIS can be utilized to perform better image classification in the dark. The goal of this paper is to fill the gap by proposing the first QIS image classification solution. 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 candidates for the third generation image sensors after CCD and CMOS. Fabricated using the commercial



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.

3D stacking technology, the current sensor has a pixel pitch of 1.1μm, with even smaller sensors being developed. The advantage of QIS over the conventional CMOS image sensors is that at 1.1μm, the read noise of QIS is as low as $0.25e^-$ whereas a typical 1μm 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 [49] so that pixels can be packed together without causing strong stray capacitance. The non-avalanche design also differentiates QIS from single photon avalanche diodes (SPAD). SPADs are typically bigger $> 5\mu$m, have lower fill factor $< 70\%$, have lower quantum efficiency $< 50\%$, and have significantly higher dark count $> 10e^-$. See [27] 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 [2, 24, 42, 55], although new results in HDR imaging has been reported [31, 50]. 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 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 $2e^-$ [59]. 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 progress has been made over the past 15 years. Readers interested in the sensor development can consult recent keynote reports, e.g., [21]. On the algorithmic side, several theoretical signal processing results and reconstruction algorithms have been proposed [3,5,27], including some very recent methods based on deep learning [6, 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 [32].

Low-light Classification. The 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 [7], who observed that low-light classification is achievable by using a few photons. In the same year, Diamond et al. [14] 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 [35], training a classifier to help denoising [61] or using an ensemble method [15], or training a denoiser that are better suited for pre-trained classifiers [44, 45].

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 [52] and contrast enhancement [23,30,37,53]. Deep neural network methods have recently become the main driving force [47, 56, 57, 62, 67, 68], including the recent series on "seeing in the dark" by Chen et al. [8, 9]. Burst photography [12, 33, 41, 54] (with some older work in [39,43,46]) 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 ($\boldsymbol{x}_{\text{rgb}}$), 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

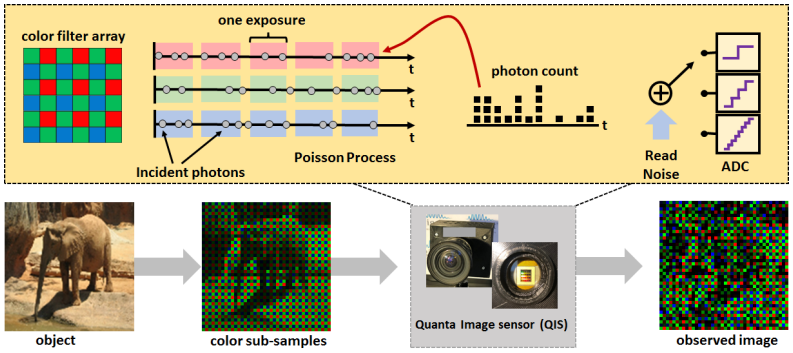


Fig. 3. QIS Image Formation Model. The basic image formation of QIS consists of a color filter array, a Poisson process, read noise, and an analog-to-digital converter (ADC). Additional factors are summarized in (1).

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 $\{0, 1, \dots L\}$, where L represents the maximum signal level. The image formation process can be summarized using the following equation

$$\underbrace{\boldsymbol{x}_{\text{QIS}}}_{\mathbb{R}^{M\times N}} = \text{ADC}_{[0,L]}\Big\{ \underbrace{\text{Poisson}}_{\text{photon arrival}}\Big(\underbrace{\alpha}_{\text{sensor gain}} \cdot \text{CFA}\big(\underbrace{\boldsymbol{x}_{\text{rgb}}}_{\mathbb{R}^{M\times N\times 3}} \big)\Big) + \underbrace{\boldsymbol{\eta}}_{\text{read noise}} \Big\}, \quad (1)$$

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 $\alpha \cdot \text{CFA}(\boldsymbol{x}_{\text{rgb}})$. The typical dark current of the QIS is $0.068e^-$/pix/s. Second, we model the pixel response non-uniformity (PRNU). PRNU is a pixel-wise multiplication applied to $\boldsymbol{x}_{\text{rgb}}$, and is unique for every sensor. Readers interested in details on the image formation model and statistics can consult previous works such as [3, 16, 20, 65].

3.2 Student-Teacher Learning

Inspecting (1), we notice that even if the read noise $\boldsymbol{\eta}$ is zero, the random Poisson process will still create a fundamental limit due to the shot noise in $\boldsymbol{x}_{\text{QIS}}$. Therefore, when applying a classification method to the raw QIS data, some capability of removing the shot noise becomes necessary. The traditional solution to this problem (in the context of CIS) is to denoise 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.

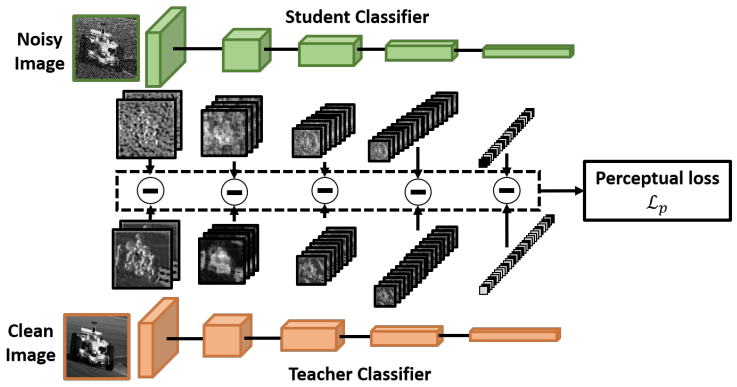


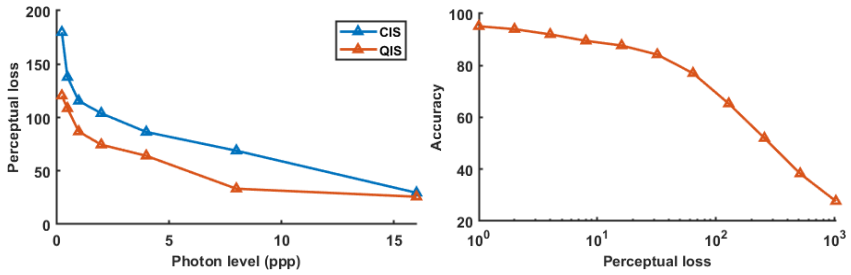
Fig. 4. Student-Teacher Learning. Student-teacher learning comprises two networks: A teacher network and a student network. The teacher network is pre-trained using clean samples whereas the student is trained using noisy samples. To transfer knowledge from the teacher to the student, we compare the 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.

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 $\phi^j(\boldsymbol{x}_{\text{QIS}})$, where $\phi^j(\cdot)$ maps $\boldsymbol{x}_{\text{QIS}}$ to a feature vector, and we define $\widehat{\phi}^j(\boldsymbol{x}_{\text{rgb}})$ as the feature vector extracted by the teacher network. The perceptual loss is

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

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 $\boldsymbol{x}_{\text{QIS}}$ before predicting the label.

We conduct a simple experiment to demonstrate the impact of input noise on 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 in-



(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 $\mathcal{L}_p(\boldsymbol{x}_{\text{QIS}}, \boldsymbol{x}_{\text{rgb}})$. 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.

creases. This is illustrated in Figure 5(a). Then in Figure 5(b), we evaluate the classification accuracy 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. [34] which proposed an effective way to compress networks. Several follow up ideas have been proposed, e.g., [1, 29, 63, 69], including the MobileNet [36]. 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. [38] and Gatys et al. [25, 26], among many others [10, 44, 48, 51, 58, 60, 66]. The method we propose here is different because we are not compressing the network. 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 $\mathcal{L}_{\text{c}}$, measures the difference between true label y and the predicted label $f_{\boldsymbol{\Theta}}(\boldsymbol{x}_{\text{QIS}})$ generated by the student network, where $f_{\boldsymbol{\Theta}}$ is the student network. The overall loss is mathematically described as

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \Big\{ \mathcal{L}_{\text{c}}\big(y^n, f_{\boldsymbol{\Theta}}(\boldsymbol{x}^n_{\text{QIS}})\big) + \lambda \mathcal{L}_{\text{p}}\big(\boldsymbol{x}^n_{\text{rgb}}, \boldsymbol{x}^n_{\text{QIS}}\big) \Big\}, \tag{3}$$

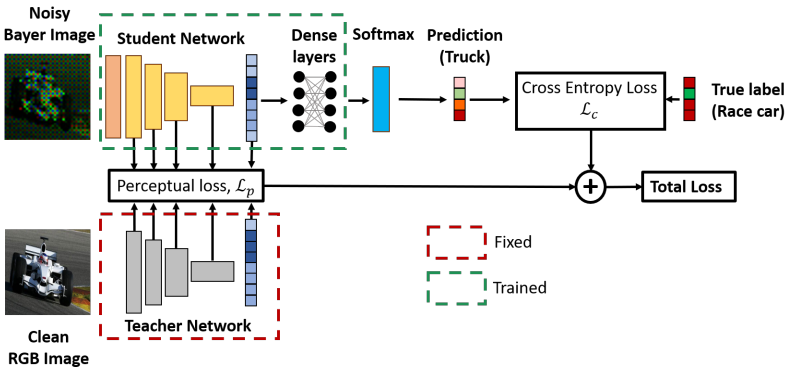


Fig. 6. Proposed Method. The proposed method trains a classification network with two training losses: (1) cross-entropy loss to measure the prediction quality, and (2) perceptual loss to transfer knowledge from teacher to student. During testing, only the student is used. We introduce a 2-layer entrance (colored in orange) for the student network so that the classifier can handle the Bayer image.

where $\boldsymbol{x}^n$ 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

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{\Theta}). \tag{4}$$

During testing, we feed a testing sample $\boldsymbol{x}_{\text{QIS}}$ to the student network and evaluate the output:

$$\widehat{y} = f_{\widehat{\boldsymbol{\Theta}}}(\boldsymbol{x}_{\text{QIS}}). \tag{5}$$

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 the 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 to differentiate the performance regime of the proposed method and its benefits over other state-of-the-art networks.

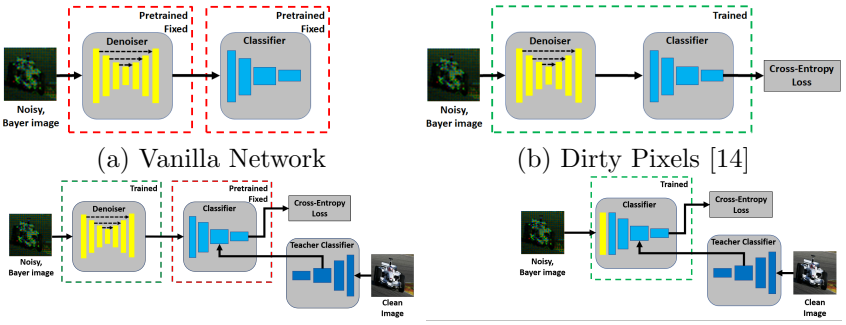
(a) Animal Dataset (Easier) (b) Dog Dataset (Harder)

Fig. 7. The two datasets for our experiments.

The construction of the two datasets is as follows. For the Animal dataset, we randomly select 10 classes of animals from ImageNet [13], as shown in Figure 7(a). Each class contains 1300 images, giving a total of 13K images. Among these, 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 [40], 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

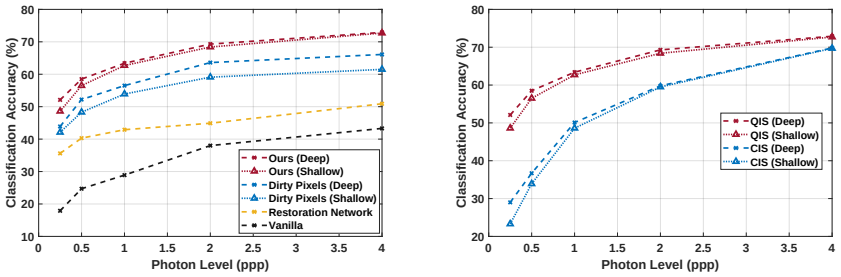
We compare our method with three existing low-light classification methods as shown in Figure 8. The three competing methods are (a) Vanilla denoiser + classifier, an "off-the-shelf" solution using pre-trained models. The denoiser is pre-trained on the QIS data and the classifier is pre-trained on clean images. (b) Dirty Pixels [14], same as Vanilla denoiser + classifier, but trained end-to-end using the QIS data. (c) Restoration Network [44,45], which trains a denoiser but uses a classifier pre-trained on clean images. This can be viewed as a middle-ground solution between Vanilla and Dirty Pixels.



(a) Vanilla Network (b) Dirty Pixels [14]

(c) Restoration Network [44] (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.



(a) Different Classifiers (b) Different Sensors

Fig. 9. Synthetic Data on Dog Dataset. (a) Comparing different classification methods with QIS images. (b) Comparing QIS and CIS using proposed classifier.

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 will certainly affect 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 [49]. 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 $\sigma = 2.0e^-$ [59].

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.

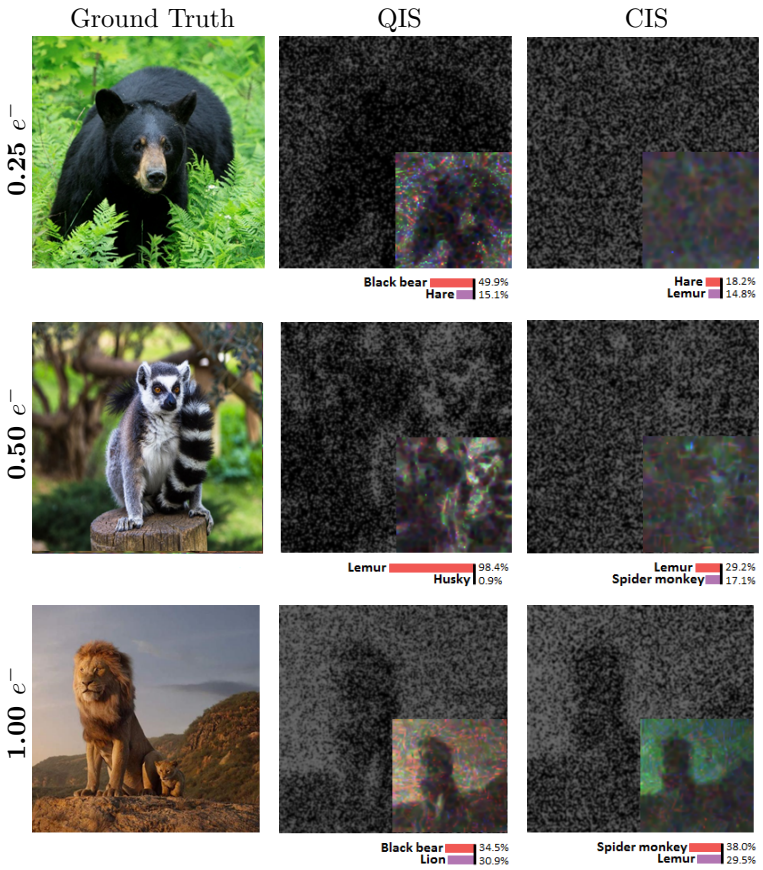
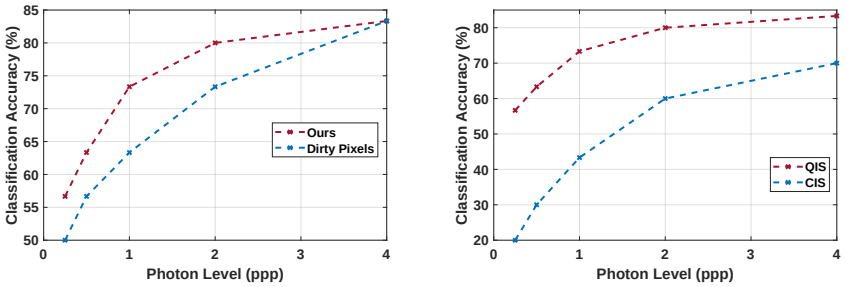


Fig. 10. Real Image Results. This figure shows raw Bayer data obtained from a prototype QIS and a commercially available CIS, and how they are classified using our proposed classifier. The inset images show the denoised images (by [9]) for visualization. Notice the heavy noise at 0.25 and 0.5 ppp, only QIS plus our proposed classification method can produce the correct prediction.

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 [27], whereas the real CIS data are collected by using a commercially available camera. To set up the experiment, we display the images on a Dell P2314H LED screen (60Hz). The cameras are positioned 1m from the display so that the field of view covers 256×256 pixels of the image. The integration time of the CIS is set to 250μs and that of QIS is 75μs. 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.



(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.

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 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.

In Figure 11 we make two pairs of comparisons: Proposed (shallow) versus Dirty Pixels (shallow), and QIS versus CIS. 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 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 QIS in the overall pipeline. The result of this ablation study can be seen in Figure 9(b). At 4 ppp 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 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 are a sufficient number of photons.

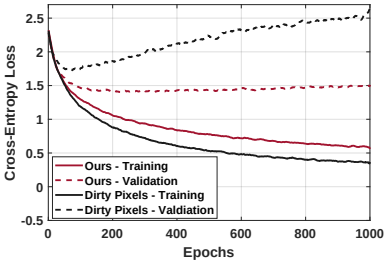
Classification Pipeline. We fix the sensor but change the entire classification pipeline to understand how important the classifier is, and which classifier is more effective. The results in Figure 9(a) show that among the competing methods, Dirty Pixels is the most promising one because it is end-to-end 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 a good sensor (QIS) does not automatically translate to better performance.

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 [Left] shows the comparison between the proposed method (shallow) and Dirty Pixels (shallow). It is evident from the plot that without student-teacher learning (Dirty Pixels), the network overfits. The validation loss 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.

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 QIS data. Figure 12 [Right] 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 particular network architecture.

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 [44] 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 denoiser to produce high-quality images or otherwise the classifier will not work. The results in Figure 9(a) suggest that when the photon level is low, the denoiser fails to produce high-quality images and so the classification fails. For example, at 0.25 ppp



	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%

Fig. 12. [Left] 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. [Right] 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.

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). The deep versions use a 20-layer UNet, whereas the shallow versions use a 2-layer network. The result in Figure 9(a) suggests that while the deep denoiser has a significant impact on Dirty Pixels, its influence is quite small to the proposed method with the QIS images. Since we are using student-teacher learning, the features are already properly handled. The benefit from a deep denoiser for QIS is therefore marginal. However, for CIS data at low light, the deep denoiser helps in getting better classification performance, especially when the signal level is much below 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, QIS must be a part of the overall pipeline. Using multiple frames for image classification would be a fruitful direction for future work.

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

References

1. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: NeurIPS (2014)
2. Callenberg, C., Lyons, A., den Brok, D., Henderson, R., Hullin, M.B., Faccio, D.: EMCCD-SPAD camera data fusion for high spatial resolution time-of-flight imaging. In: Computational Optical Sensing and Imaging (2019)
3. Chan, S.H., Elgendy, O.A., Wang, X.: Images from bits: Non-iterative image reconstruction for Quanta Image Sensors. Sensors **16**(11), 1961 (November 2016)
4. Chan, S.H., Lu, Y.M.: Efficient image reconstruction for gigapixel Quantum Image Sensors. In: IEEE Global Conf. Signal and Info. Process. (2014)
5. Chan, S.H., Wang, X., Elgendy, O.A.: Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. IEEE Trans. Computational Imaging **3**(1), 84–98 (November 2016)
6. Chandramouli, P., Burri, S., Bruschini, C., Charbon, E., Kolb, A.: A bit too much? High speed imaging from sparse photon counts. In: ICCP (2019)
7. Chen, B., Perona, P.: Seeing into darkness: Scotopic visual recognition. In: CVPR (2017)
8. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: ICCV (2019)
9. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: CVPR (2018)
10. Chen, G., Li, Y., Srihari, S.N.: Joint visual denoising and classification using deep learning. In: ICIP (2016)
11. Choi, J.H., Elgendy, O.A., Chan, S.H.: Image reconstruction for Quanta Image Sensors using deep neural networks. In: ICASSP (2018)
12. Davy, A., Ehret, T., Morel, J.M., Arias, P., Facciolo, G.: A non-local CNN for video denoising. In: ICIP (2019)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
14. Diamond, S., Sitzmann, V., Boyd, S., Wetzstein, G., Heide, F.: Dirty pixels: Optimizing image classification architectures for raw sensor data. arXiv preprint arXiv:1701.06487 (2017)
15. Dodge, S., Karam, L.: Quality resilient deep neural networks. arXiv preprint arXiv:1703.08119 (2017)
16. Elgendy, O.A., Chan, S.H.: Optimal threshold design for Quanta Image Sensor. IEEE Trans. Computational Imaging **4**(1), 99–111 (December 2017)
17. Elgendy, O.A., Chan, S.H.: Color Filter Arrays for Quanta Image Sensors. arXiv preprint arXiv:1903.09823 (2019)
18. Fossum, E.R.: Gigapixel digital film Sensor (DFS) proposal. Nanospace Manipulation of Photons and Electrons for Nanovision Systems (2005)
19. Fossum, E.R.: Some thoughts on future digital still cameras. In: Image sensors and signal processing for digital still cameras (2006)
20. Fossum, E.R.: Modeling the performance of single-bit and multi-bit quanta image sensors. IEEE Journal of the Electron Devices Society **1**(9), 166–174 (2013)
21. Fossum, E.R., Ma, J., Masoodian, S., Anzagira, L., Zizza, R.: The Quanta Image Sensor: Every photon counts. Sensors **16**(8), 1260 (2016)
22. Fowler, B., McGrath, D., Bartkovjak, P.: Read noise distribution modeling for CMOS Image Sensors. In: International Image Sensor Workshop (2013)
23. Fu, Q., Jung, C., Xu, K.: Retinex-based perceptual contrast enhancement in images using luminance adaptation. IEEE Access **6**, 61277–61286 (October 2018)

24. Gariepy, G., Krstajić, N., Henderson, R., Li, C., Thomson, R.R., Buller, G.S., Heshmat, B., Raskar, R., Leach, J., Faccio, D.: Single-photon sensitive light-in-fight imaging. Nature communications **6**(1), 1–7 (2015)
25. Gatys, L., Ecker, A., Bethge, M.: A neural algorithm of artistic style. Nature Communications (2015)
26. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: NeurIPS (2015)
27. Gnanasambandam, A., Elgendy, O., Ma, J., Chan, S.H.: Megapixel photon-counting color imaging using Quanta Image Sensor. Optics Express **27**(12), 17298–17310 (June 2019)
28. Gnanasambandam, A., Ma, J., Chan, S.H.: High Dynamic Range imaging using Quanta Image Sensors. In: International Image Sensors Workshop (2019)
29. Guo, T., Xu, C., He, S., Shi, B., Xu, C., Tao, D.: Robust student network learning. IEEE Trans. Neural Networks and Learning Systems (August 2019)
30. Guo, X., Li, Y., Ling, H.: LIME: Low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (October 2016)
31. Gupta, A., Ingle, A., Gupta, M.: Asynchronous single-photon 3D imaging. In: ICCV (October 2019)
32. Gyongy, I., Dutton, N., Henderson, R.: Single-photon tracking for high-speed vision. Sensors **18**(2), 323 (January 2018)
33. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Trans. Graphics **35**(6), 192 (November 2016)
34. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NeurIPS Deep Learning and Representation Learning Workshop (2015)
35. Hossain, M.T., Teng, S.W., Zhang, D., Lim, S., Lu, G.: Distortion robust image classification using deep convolutional neural network with discrete cosine transform. In: ICIP (2019)
36. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
37. Hu, Z., Cho, S., Wang, J., Yang, M.H.: Deblurring low-light images with light streaks. In: CVPR (2014)
38. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
39. Joshi, N., Cohen, M.: Seeing Mt. Rainier: Lucky imaging for multi-image denoising, sharpening, and haze removal. In: ICCP (2010)
40. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPR Workshop on Fine-Grained Visual Categorization (2011)
41. Kokkinos, F., Lefkimmiatis, S.: Iterative residual CNNs for burst photography applications. In: CVPR (2019)
42. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3D imaging with deep sensor fusion. ACM Trans. Graphics **37**(4), 1–12 (2018)
43. Liu, C., Freeman, W.: A high-quality video denoising algorithm based on reliable motion estimation. In: ECCV (2010)
44. Liu, D., Wen, B., Jiao, J., Liu, X., Wang, Z., Huang, T.S.: Connecting image denoising and high-level vision tasks via deep learning. IEEE Trans. Image Process. **29**, 3695–3706 (2020)
45. Liu, Z., Zhou, T., Shen, Z., Kang, B., Darrell, T.: Transferable recognition-aware image processing. arXiv preprint arXiv:1910.09185 (2019)

46. Liu, Z., Yuan, L., Tang, X., Uyttendaele, M., Sun, J.: Fast burst images denoising. ACM Trans. Graphics **33**(6), 232 (November 2014)
47. Lore, K.G., Akintayo, A., Sarkar, S.: LLNet: A deep autoencoder approach to natural low-light image enhancement. Patt. Recog. **61**, 650–662 (Jan 2017)
48. Ma, J., Fossum, E.: A pump-gate jot device with high conversion gain for a quanta image sensor. IEEE J. Electron Devices Soc. **3**(2), 73–77 (January 2015)
49. Ma, J., Masoodian, S., Starkey, D., Fossum, E.R.: Photon-number-resolving megapixel image sensor at room temperature without avalanche gain. Optica **4**(12), 1474–1481 (December 2017)
50. Ma, S., Gupta, S., Ulku, A.C., Brushini, C., Charbon, E., Gupta, M.: Quanta burst photography. ACM Transactions on Graphics (TOG) **39**(4) (Jul 2020)
51. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR (2015)
52. Makitalo, M., Foi, A.: Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. IEEE Trans. Image Process. **20**(1), 99–109 (July 2010)
53. Malm, H., Oskarsson, M., Warrant, E., Clarberg, P., Hasselgren, J., Lejdfors, C.: Adaptive enhancement and noise reduction in very low light-level video. In: ICCV (2007)
54. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: CVPR (2018)
55. O'Toole, M., Heide, F., Lindell, D.B., Zang, K., Diamond, S., Wetzstein, G.: Reconstructing transient images from single-photon sensors. In: CVPR (2017)
56. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. In: CVPR (2017)
57. Remez, T., Litany, O., Giryes, R., Bronstein, A.: Deep convolutional denoising of low-light images. arXiv preprint arXiv:1701.01687 (2017)
58. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR (2014)
59. FLIR Sensor Review: Mono Camera. https://www.flir.com/globalassets/iis/guidebooks/2019-machine-vision-emva1288-sensor-review.pdf
60. Talebi, H., Milanfar, P.: Learned perceptual image enhancement. In: ICCP (2018)
61. Wu, J., Timofte, R., Huang, Z., Van Gool, L.: On the relation between color image denoising and classification. arXiv preprint arXiv:1704.01372 (2017)
62. Xu, J., Li, H., Liang, Z., Zhang, D., Zhang, L.: Real-world noisy image denoising: A new benchmark. arXiv preprint arXiv:1804.02603 (2018)
63. Yang, C., Xie, L., Qiao, S., Yuille, A.: Training deep neural networks in generations: A more tolerant teacher educates better students. In: AAAI Conf. Artificial Intelligence (July 2019)
64. Yang, F., Lu, Y.M., Sbaiz, L., Vetterli, M.: An optimal algorithm for reconstructing images from binary measurements. In: Proc. SPIE. vol. 7533, pp. 158 – 169 (January 2010)
65. Yang, F., Lu, Y.M., Sbaiz, L., Vetterli, M.: Bits from photons: Oversampled image acquisition using binary poisson statistics. IEEE Trans. image process. **21**(4), 1421–1436 (2011)
66. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. ICML Deep Learning Workshop (2015)
67. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7) (February 2017)

68. Zhang, K., Zuo, W., Zhang, L.: FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9) (May 2018)
69. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)