Photon-Starved Scene Inference using Single Photon Cameras

Bhavya Goyal   Mohit Gupta

University of Wisconsin-Madison

{bhavya, mohitg}@cs.wisc.edu

Abstract

Scene understanding under low-light conditions is a challenging problem. This is due to the small number of photons captured by the camera and the resulting low signal-to-noise ratio (SNR). Single-photon cameras (SPCs) are an emerging sensing modality that are capable of capturing images with high sensitivity. Despite having minimal read-noise, images captured by SPCs in photon-starved conditions still suffer from strong shot noise, preventing reliable scene inference. We propose photon scale-space – a collection of high-SNR images spanning a wide range of photons-per-pixel (PPP) levels (but same scene content) as guides to train inference model on low photon flux images. We develop training techniques that push images with different illumination levels closer to each other in feature representation space. The key idea is that having a spectrum of different brightness levels during training enables effective guidance, and increases robustness to shot noise even in extreme noise cases. Based on the proposed approach, we demonstrate, via simulations and real experiments with a SPAD camera, high-performance on various inference tasks such as image classification and monocular depth estimation under ultra low-light, down to < 1 PPP. Project Page: `https://wisionlab.cs.wisc.edu/project/photon-net`

1. Scene Inference in Low Light

Over the past decade, deep learning has achieved unmatched accuracy on several complex, real-world scene inference tasks. As these techniques have matured, a new axis in the performance space is emerging, driven by applications (e.g., autonomous navigation), where reliable performance under non-ideal imaging conditions is as important as the overall accuracy. In such safety-critical applications, it is important to consider the *worst case* performance of the vision system to ensure robust *all-weather operation*. For example, for a vision system to be deployed on an autonomous car, it must perform reliably across the entire range of imaging scenarios, including night-time and poorly-lit scenes, and high-speed moving objects, all of which result in photon-starved images. Even state-of-the-art inference algorithms tend to fail for such images where the sensor has simply not collected sufficient light.

The goal of this paper is to develop vision systems that achieve high accuracy even in ultra low-light, when a camera pixel may receive even less than one photon per pixel. In such extreme conditions, images captured by conventional cameras get overwhelmed by noise, causing the signal-to-noise ratio (SNR) to dip below the threshold required for downstream inference algorithms to extract meaningful scene information. We propose a two-pronged approach to achieve these goals: (a) Leverage a class of highly sensitive single-photon detectors, and (b) Develop inference algorithms that are optimized for low-flux operation.

Single-Photon Sensors: Single-photon avalanche diodes (SPADs) [31, 37] are an emerging image sensor technology that are capable of detecting individual incident photons with high timing precision. In the past, these sensors were limited to single pixel or low-resolution devices (e.g., 32x32 pixels), and thus restricted to scientific applications [5, 33, 4]. But, recently, due to their compatibility with CMOS fabrication processes, high resolution cameras (up to 1 MPixel) have been developed based on SPADs [29], as well as the jots [27] technology. These single photon cameras are capable of capturing sequences of binary frames with minimal read noise [41], thus opening the possibility of capturing high-quality images even in low-flux conditions.

High-level Inference on Low-Flux Images: So far, SPADs have primarily been used for recovering image intensities [2, 21, 28] and low/mid-level scene information such as 3D shape [32, 36, 9, 22, 39, 16, 15] and motion [17]. Can we go beyond low-level imaging and signal processing, and develop algorithms for direct, high-level inference from SPAD cameras? Despite low read noise, the stochastic nature of photon arrival results in considerable shot noise in single-photon binary frames captured by SPAD cameras. Although there has been some recent work on joint denoising and classification [25, 26, 12], inference on ultra-low-light images where each pixel receives less than a photon on average still remains an intractable problem.
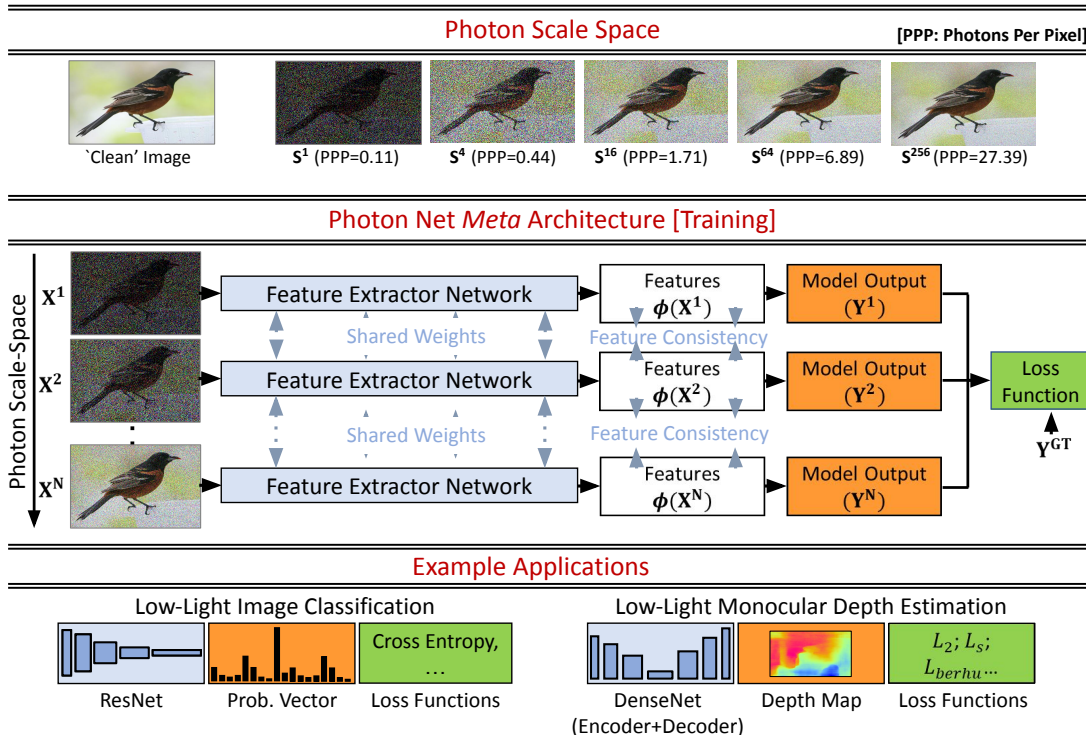
Photon Scale Space [PPP: Photons Per Pixel]

`Clean' Image S^1 (PPP=0.11) S^4 (PPP=0.44) S^{16} (PPP=1.71) S^{64} (PPP=6.89) S^{256} (PPP=27.39)

Photon Net *Meta* Architecture [Training]

Photon Scale-Space

X^1 — Feature Extractor Network → Features $\phi(X^1)$ → Model Output (Y^1)

Shared Weights — Feature Consistency

X^2 — Feature Extractor Network → Features $\phi(X^2)$ → Model Output (Y^2) → Loss Function

Shared Weights — Feature Consistency — Y^{GT}

X^N — Feature Extractor Network → Features $\phi(X^N)$ → Model Output (Y^N)

Example Applications

Low-Light Image Classification

ResNet   Prob. Vector   Cross Entropy, … Loss Functions

Low-Light Monocular Depth Estimation

DenseNet (Encoder+Decoder)   Depth Map $L_2; L_s; L_{berhu}\cdots$ Loss Functions

Figure 1: **Inference in Low-Light using Photon Scale-Space. (Top)** Photon scale-space is a hierarchy of images, each with a different flux level, but sharing the same scene content. Successive images in the hierarchy have similar flux, so that high-flux images can guide the low-flux images during a training procedure. **(Middle)** We use photon-scale space to develop a meta network architecture called the photon net, where a network is trained with multiple input images with the same scene content but with varying noise levels in order to push them together in the feature space. **(Bottom)** The proposed approach is modular and versatile, lending itself to a wide range of inference tasks such as classification and depth estimation.

To address this problem, we design inference techniques based on the notion of *guided training*, where a high-quality image is used as a guide for training low-quality images. This is similar in spirit to the classical guided filtering [18] where a guidance image is used for low-level image processing tasks, such as denoising [18] and super-resolution [10]. More recently, the idea of guided training has been explored in the context of student-teacher training [14] where a teacher network pre-trained on high-quality images guides a student network operating on low-flux images. These approaches rely on underlying similarities in the inputs of the student and teacher networks to aid the guidance process, and therefore, are not very effective in the extreme case where the student's and teacher's input images may have a huge difference in the number of photons-per-pixel (e.g., < 1 vs. > 1000). These images may have no structural similarity despite representing the same scene. If the guide and the *guidee* images have no common content and features, how can one perform guided training?

Photon-Scale Space: We propose using a hierarchy of guide images from a wide spectrum of photon levels, each having the same scene content, but varying number of av-

erage photons-per-pixel (PPP), from as low as PPP ≈ 0.1, going up to PPP> 100. The key idea is that although all the images taken together span a large range of SNR values (including high SNR images at the top which provide most accurate labels), *successive* images in the hierarchy have similar number of photons (and thus, similar features) so that guidance percolates down effectively to the lowest levels, to the images with the minimum PPP. We call this hierarchy of images the *photon scale-space* (Fig. 1), reminiscent of the classical image-size scale-space [24] which is used in many computer vision algorithms.

Photon-Net Guided Training: Based on photon scale-space, we propose *Photon Net*, a *meta* architecture and training techniques for performing inference on low-flux input images (Fig. 1). The key idea is to train a given network architecture with different images from a photon scale-space, so that the images having the same scene content (but different flux level) are trained together leading to effective guidance from the highest SNR training images to the low SNR test images. We do this by enforcing feature consistency for high-level features (e.g., the final feature vector) of the network. Since frames at different levels

in the photon scale space share the same scene content, we encourage similarity of high-level features, despite having large differences in the low-level image statistics (low/mid-level features).

We perform empirical analysis on various design choices for creating the photon scale-space (e.g., the number of levels), and suggest rules-of-thumb for good performance. Due to the known forward model of the single photon imaging process (Poisson sampling), the photon scale-space can be created using images captured from conventional cameras, making the proposed approach amenable to training using existing large-scale image datasets. We demonstrate, via extensive simulations as well as real experiments on a 1/8 megapixel SPAD array (SwissSPAD2 [41]), considerable (up to 10%) improvement for various inference tasks in extreme low-light conditions (~0.1 PPP).

Scope and Implications: The proposed approach is modular and versatile — it is possible to use a wide range of network architectures, loss functions, and model outputs in the same framework — thus lending itself to a variety of inference tasks including low-light image classification and even regression tasks such as monocular depth estimation in the dark (Fig. 1). SPADs remain a nascent imaging modality, and cannot yet directly compete with conventional sensors which have been optimized over decades. However, given their sensitivity, high speed and dynamic range [2, 21, 28], they have the potential to provide capabilities (e.g., vision in ultra low-light and rapid motion) that were hitherto considered impossible. This work takes the first steps towards exploring SPADs as all-purpose sensors capable of not just low-level imaging, but also high-level inference across a wide gamut of challenging imaging conditions.

2. Related Work

Single-Photon (Quanta) Sensors: SPADs and jots are two current major technologies for large single-photon camera arrays. Jots amplify the single-photon signal by using an active pixel with high conversion gain [13]. By avoiding avalanche, jots achieve smaller pixel pitch, higher quantum efficiency and lower dark current, but have lower temporal resolution [27]. Although we demonstrate our approaches using SPADs, the computational techniques are applicable generally to a single-photon sensors, including jots.

Inference on Single Photon Sensors: Starting with the early (primarily theoretical) work [6, 7] which proposed the idea of directly performing computer vision tasks on stream of photons instead of forming an image, there has been a growing trend of using quanta sensors for various scene inference applications. This includes high-speed tracking using quanta sensors [17], and more recently, object identification [3] and image classification [14]. Our work is a next step in this direction, providing a general and versatile

approach capable of achieving high performance across a wide variety of scene inference tasks.

Low-light Classification: There has been a lot of work on inference in low-light using conventional cameras as well. The most notable in this line of work are recent approaches that perform joint denoising and inference on noisy images [25, 26, 12]. Although such joint denoising and inference techniques outperform conventional sequential denoising and inference approaches, they do not have the benefit of effective guidance from high SNR images, and thus are unable to achieve high performance in extreme low-light conditions ($\sim$ 0.1 PPP).

Image-size Scale Space: A recent work [43] proposes techniques that use image-size scale space, i.e., images at multiple resolutions, for designing pose estimation techniques that can perform well for very low resolution images. We borrow numerous insights from this work, as we create photon scale-space and photon net family of architectures for inference on very low-light images.

3. Passive Single Photon Imaging Model

For a single photon camera, the number $Z(x, y)$ of photons arriving at pixel (x, y) during an exposure time of τ seconds is modeled as a Poisson random variable [44], whose distribution is given as:

$$P\{Z = k\} = \frac{(\phi\tau\eta)^k e^{-\phi\tau\eta}}{k!} , \qquad (1)$$

where $\phi(x, y)$ is the photon flux (photons/seconds) incident at (x, y), and $0 \leq \eta \leq 1$ is the quantum efficiency of the pixels. In the binary mode, each pixel detects at most one photon during the exposure time and returns a binary value $B(x, y)$ such that $B(x, y) = 1$ if $Z(x, y) \geq 1$; $B(x, y) = 0$ otherwise.[1] Due to the randomness in photon arrival, the binary measurements $B(x, y)$ are also random variables with Bernoulli distribution:

$$P\{B = 0\} = e^{-(\phi\tau\eta + r_q\tau)},$$
$$P\{B = 1\} = 1 - e^{-(\phi\tau\eta + r_q\tau)} \qquad (2)$$

where r_q is the dark count rate (DCR), which is the rate of spurious photon detections.

Sources of Image Noise: Conventional sensors measure incident photons as analog current, which is then converted to a discrete number. This analog-to-digital conversion (ADC) results in a fixed read noise per frame, which leads to low signal to noise ratio (SNR) in dark scenes. In contrast, SPCs directly measure the photon counts, skipping the intermediate ADC, thereby avoiding read noise.

---

[1] After each photon detection, a SPAD pixel experiences a dead time during which it cannot detect any further photons [38]. For modern SPAD pixels, the dead time is significantly smaller than the exposure time τ, and therefore is not considered in the following analysis.

Although SPCs have minimal read noise, binary frames still have extremely low signal-to-noise ratio (SNR) in low flux environments due to shot noise. Fig. 1 shows an example of a clean image, with the corresponding binary image ($\mathcal{S}^1$). The shot noise in the binary image (Eq. 2) causes extreme degradation. While it is possible to increase the SNR by temporally averaging a large number of binary frames, this approach is not applicable in the presence of scene / camera motion due to motion blur or large computational requirements of motion compensation algorithms [8, 28]. This raises the following question: Is it possible to extract meaningful scene information from a single (or a small number of) single-photon binary frames?

4. Photon Scale Space

To address this question, we develop a *guided training* approach, where high SNR images act as a guide for training low SNR images. To facilitate such guided training, we propose the concept of *photon scale-space*, a hierarchy of guide images with varying flux levels, but each having the same scene content. The key idea is that although all the images taken together span a large range of SNR values (including high SNR and most informative images at the top), *successive* images in the hierarchy have similar SNR levels (and thus, similar features) so that guidance percolates down effectively to the lowest levels.

How to generate a photon scale space? Consider a 'clean image' as captured by a camera in high-flux conditions. Assuming the pixel intensities in the clean image to be the ground-truth flux values for the corresponding scene points[2], we can generate multiple stochastic binary images as captured by a single photon camera using the image formation model described in Section 3. Assuming the scene is stationary, i.e. there is no motion between binary frames, we can simulate a series of images with different flux levels by summing a sequence of N binary frames (for various values of N) to get N-sum images ($\mathcal{S}^N$), defined as follows:

Definition 4.1 (N-Sum image $\mathcal{S}^N$). The average of N binary frames

$$\mathcal{S}^N(x,y) = \sum_{i=1}^{N} B_i(x,y) \ . \tag{3}$$

Using the definition of N-Sum images, we define Photon Scale Space as a hierarchy of images with successively higher flux levels as follows:

Definition 4.2 (Photon Scale Space $PSS(K, L, n)$). A set of n N-Sum images, starting from the lowest-SNR image S^K (noisiest), to the highest-SNR image S^L, with $K < L$.

---

[2]In general, the pixel intensities have a non-linear relationship to incident flux due to sensor's radiometric response and image compression algorithms. Although we do not explicitly model these effects, they can be accounted for in the following discussion.

We choose the parameters (K, L, n) so that the images span a large gamut of SNR levels (i.e., $K \ll L$). The choice of the number of levels n presents a trade-off: To ensure effective guidance from high SNR to low SNR images, the successive images in the hierarchy should have similar flux levels, thus requiring a large n. On the other hand, a large n would increase computational cost of the training algorithms. In our implementations, we choose images with N increasing as a geometric series $N \in [K, K(L/K)^{\frac{1}{n-1}}, K(L/K)^{\frac{2}{n-1}} ..., L]$ so that the approximate ratio of the flux level between successive images is a constant. We round the values of N to the nearest integer if it is a fraction.

For instance, suppose we want to train inference models for $\mathcal{S}^1$ images (1 binary frame), but use high flux images up to $\mathcal{S}^{256}$ (256 binary frames) for guidance during training. The photon scale space for this setting with, say, 5 levels would consist of $\mathcal{S}^1, \mathcal{S}^4, \mathcal{S}^{16}, \mathcal{S}^{64}$ and $\mathcal{S}^{256}$ images. Fig. 1 shows an example of images from photon scale space with $K = 1$, $L = 256$ and $n = 5$, thereby spanning a broad range of SNR levels, while ensuring that successive images have similar SNR and features.

What is the range of flux values spanned by a photon scale space? Since each binary frame is independent, the expected value of the sum image $\mathcal{S}^N(x,y)$ is:

$$\begin{aligned} E[\mathcal{S}^N(x,y)] &= N * E[B(x,y)] \\ &= N(1 - e^{-(\phi\tau\eta + r_q\tau)}) . \end{aligned} \tag{4}$$

The maximum likelihood estimate (MLE) of the incident flux (ϕ) is therefore given as

$$\hat{\phi}(x,y) = -\ln(1 - \mathcal{S}(x,y)/N)/\tau\eta - r_q/\eta . \tag{5}$$

This non-linear relationship between $\mathcal{S}$ (number of photons detected by the camera), and ϕ (number of photons incident of camera) has an asymptotic behavior [28, 2] — $\mathcal{S}^N$ keeps increasing with increasing number of incident photons ϕ, allowing us to span a large range of incident flux levels in the photon scale space, even with a finite range of N values.

5. Guided Training with Photon Scale Space

In this section, we design a guided training technique that leverages photon scale space images for developing high-performance low-light inference algorithms.

Photon Net: The key enabling component of the proposed technique is a *meta* architecture called Photon Net that uses photon scale space images as input, along with a feature consistency loss that encourages similar feature representations for all the images belonging to the same photon scale space (thus having the same scene content), despite having a large variation in brightness levels.

Fig. 1 shows the overview of the architecture, which consists of several identical network branches. During training, each branch takes as input an image from the photon scale space with a certain PPP level (ranging from low SNR to high SNR images). All the branches are trained with shared weights, so that gradient updates from high PPP branches can guide low PPP branches. In order for high SNR images to act as a guide to low SNR images, all the photon scale space images with different PPP levels from the same original image are trained together by sampling them in the same mini-batch. Since, the weights are shared, there is no additional overhead of network parameters as we do not keep multiple copies of the network.

Encouraging Feature Consistency: In order to encourage consistency in the learned feature representations for different inputs of the network (images with the same scene content but different noise levels), we use feature consistency loss during training. It is possible to use a variety of loss functions such as contrastive loss, L2, or L1 loss for consistency. In our implementation, we used an L2 loss function (Mean Squared Error loss), to push features from the same image with different PPP levels closer to each other.

$$L_{MSE}(\{x_i\}) = \frac{1}{N} \sum_{i,j} \|\varphi(x_i) - \varphi(x_j)\|_2^2 \qquad (6)$$

where $\{(x_i)\}$ is set of all training images, N is the total number of training pairs in the mini-batch with same scene content (i.e. x_i and x_j are images from the same original image with different PPP level) and $\varphi(.)$ is the feature output from network. We use feature vector from the final layer of the CNN (after global pooling layer) as our feature representation.

Overall loss function is the combination of L_{MSE} and the primary loss function for the inference task. For the case of image classification: $L_{overall} = L_{CE} + \lambda L_{MSE}$, where L_{CE} is cross entropy loss and λ is the weighting factor to control the contribution of both losses. Please see the supplementary report for details.

6. Applications: Low-Light Scene Inference

The guided training approach based on photon scale space and photon net is modular since it is possible to use a wide range of network architectures, loss functions, and tasks in the same framework (Fig. 1). We demonstrate the effectiveness and versatility of the proposed techniques via two low-flux inference tasks: image classification and monocular depth estimation.

6.1. Image Classification

Datasets: We first show the performance of our approach using simulated images using two datasets: CUB-200-2011 image classification dataset [42] and CARS dataset [23].
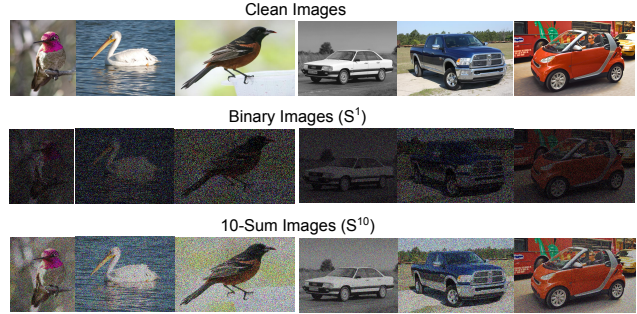

Figure 2: **Simulated Single-Photon Images**: Clean images and simulated noisy images from CUB-200-2011 and CARS-196 dataset. SPCs captures sequence of binary images like ($\mathcal{S}^1$) with heavy shot noise. 10-sum images ($\mathcal{S}^{10}$) are average of 10 binary images.

CUB-200-2011 is commonly used for fine-grained image classification benchmarks and consists of 200 species of birds with 5,994 training images and 5,794 test images. The CARS dataset contains images of 196 classes of cars (with different make, model and year) with 8,144 training images and 8,041 testing images.

We generate sequences of binary frames from the original images in the dataset (images captures by CMOS cameras) for training using the image formation model in Equation 2. $\phi\tau\eta$ in the model corresponds to the poisson parameter for the model. We assume $\eta * \tau = 1/1000$ to be constant for the dataset to simulate low flux setting and ϕ proportional to pixel value of original image. We then sum N binary frames together to generate $\mathcal{S}^N$ images for the complete dataset. Figure 2 shows sample clean images from the dataset and sample noisy images generated using imaging model for Single Photon Cameras.

Comparisons and Baselines: We compare our method with two approaches that are designed for inference on low-SNR images. Our baseline approach is Joint Denoising [12], which uses a denoiser for noisy images coupled with a conventional image classification architecture. Both denoiser and classifier are trained together on noisy data (Fig. 3a), with a combined loss consisting of cross entropy loss for the classification and mean squared error for the denoiser. We also compare our method with the Student-Teacher learning approach [14] where clean images are used to train a *fixed* teacher network and noisy images are used for training the student network (Fig. 3b). This approach encourages consistency between feature representation of clean image and noisy images by minimizing a mean squared error between the feature outputs of the student and the teacher networks. For more comparisons, please refer to the technical report.

Experiments: We perform all of our experiments with ResNet-18 [20] as the backbone architecture provided by

(a) Joint Denoising
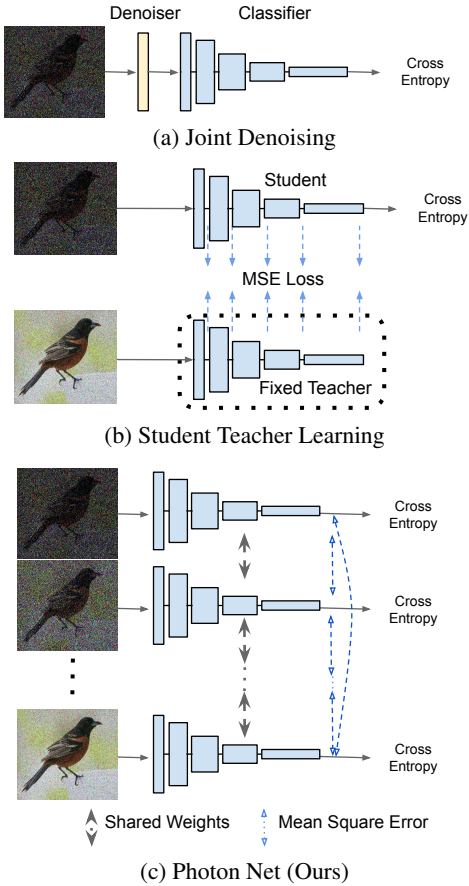

(b) Student Teacher Learning

(c) Photon Net (Ours)

Figure 3: **Comparison of Architectures with Existing Approaches for Low-Light Inference**: (a) *Joint Denoising* consists of a denoiser jointly trained with an inference network. (b) *Student Teacher Learning* uses a fixed teacher model trained on clean images and a trainable student model for noisy images. (c) *Photon Net (Ours)* uses multiple images with different PPP level as input to the network. Different branches of network corresponds to different PPP levels, and all branches share weights with each other. A loss function such as mean squared error between feature representations is used to push images with different PPP level closer to each other in feature space.

Pytorch [34] for all the baselines. As shown in Figure 3, output of global pooling layer of size 512 is used as our feature extractor. We choose 5 levels of Photon Scale Space images for training Photon Net in our experiments. This choice of number of levels is analysed later in the paper as part of an ablation study. All photon scale space images corresponding to the same image are sampled together in the same minibatch during training. We initialize our network with pre-trained weights from the model trained on clean images. Stochastic gradient descent with momentum optimizer with momentum as 0.9, base learning rate of 0.1 with

cosine decay and batch size of 80 is used for fine-tuning.

Results and Implications: Table 1 shows the results of our approach on CUB-200-2011 and CARS-196 dataset for different illumination levels. The proposed approach significantly outperforms Joint Denoising since denoising in the image space is not very effective for extreme noise levels (e.g., PPP < 0.1). With as few as ∼ 0.1 Photons Per Pixel, our approach is able to get top-1 accuracy comparable to what conventional denoising approaches can achieve with 1 Photon Per Pixel (1 magnitude higher). Student Teacher Learning performs better than Joint Denoising as it enforces feature consistency between noisy and clean images. However, since it uses a fixed teacher network with only clean images, the guidance is not as effective. Photon Net trains a wide gamut of SNR images together in the same network with noisy images.

Ablation Study: We study the effect of the parameters of the Photon Scale Space (PSS) on the performance by varying the number of levels of PSS during training. We start with 2 levels of PSS (only noisy and clean image) and increase up to 9 (noisy, clean and 7 more intermediate levels). Fig. 4 shows results of image classification on $\mathcal{S}^1$ test images of CUB-200-2011 and CARS-196 dataset. For these datasets, the performance of the model increases with increasing number of PSS levels, but saturates around 5 levels, thus informing the choice of parameters in our experiments. An important next step is to perform a similar empirical analysis for a wider range of datasets and tasks.

6.2. Monocular Depth Estimation

We also show the application of our approach to monocular depth estimation, a regression task.

Depth Estimation Overview: For this application, we use the DenseDepth [1] base architecture, consisting of an Encoder and a Decoder. The Encoder is a deep CNN (ResNet-34 pretrained on ImageNet [11]) which extracts the feature maps and the Decoder is a series of upsampling layers with skip connections to construct the depth map from the feature maps. Loss function used is a combination of point-wise L1 loss and Structural Similarity loss between predicted and ground truth depth values. We use the same training procedure as described in [1].

Photon Net training for Depth Estimation: We train our Depth Estimation architecture with photon scale space images. Mean Squared Error Loss is used for feature consistency of the feature outputs of the images from different PPP levels. We use output of the Encoder network (after global pooling layer) for our feature representation. We provide more details on the architecture in the technical report.

Experiments and Results We evaluate our approach on NYUV2 dataset [40]. Same training and testing split is used as [1] which includes 50K training and 654 testing

Test Data	PPP	CUB-200-2011			CARS-196		
		Joint Denoising	Student-Teacher Learning	Photon Net (Ours)	Joint Denoising	Student-Teacher Learning	Photon Net (Ours)
$\mathcal{S}^1$	0.11	27.21	35.43	**42.37**	34.51	57.81	**64.23**
$\mathcal{S}^2$	0.22	31.33	39.50	**48.56**	43.14	65.85	**70.51**
$\mathcal{S}^5$	0.53	39.41	44.46	**55.19**	57.11	71.13	**75.23**
$\mathcal{S}^{10}$	1.07	44.17	48.08	**58.68**	65.78	73.51	**78.97**

Table 1: **Image Classification Results**: Top-1 Accuracy results on CUB-200-2011 and CARS-196 dataset. Photon Net outperforms both Joint Denoising [12] and Student-Teacher Learning [14] on all noise levels.

Test DataSet	PPP	Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	rel$\downarrow$	rms$\downarrow$	$log_{10} \downarrow$
$\mathcal{S}^1$	0.11	Joint Denoising	0.671	0.896	0.967	0.209	1.412	0.087
		Photon Net (Ours)	**0.713**	**0.917**	**0.976**	**0.183**	**1.275**	**0.078**
$\mathcal{S}^{10}$	1.07	Joint Denoising	0.763	0.941	0.984	0.162	1.177	0.069
		Photon Net (Ours)	**0.793**	**0.953**	**0.987**	**0.149**	**1.104**	**0.063**

Table 2: **Monocular Depth Estimation Results:** on NYUV2 dataset.

images. We simulate SPC images using the same procedure described earlier in Section 6.1. The following six standard evaluation metrics are used:

- average relative error (rel): $\frac{1}{n}\Sigma_p^n \frac{y_p - \hat{y}_p}{y}$,
- root mean square error (rms): $\sqrt{\frac{1}{n}\Sigma_p^n (y_p - \hat{y}_p)^2}$,
- average (log_{10}) error: $\frac{1}{n}\Sigma_p^n |log_{10}(y_p) - log_{10}(\hat{y}_p)|$ and
- threshold accuracy (δ_i): % of y_p s.t. $\max(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}) = \delta_i < thr$ for $thr = 1.25, 1.25^2, 1.25^3$
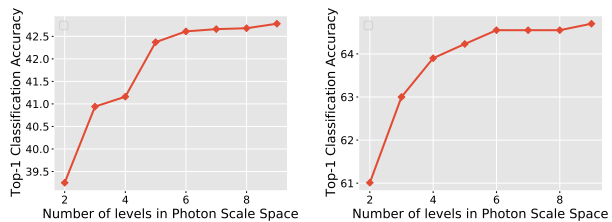
where y_p is a pixel in depth image y, $\hat{y}_p$ is a pixel in the predicted depth image $\hat{y}$, and n is the total number of pixels for each depth image.

We compare our method with Joint Denoising, which uses a denoiser with Depth Estimation architecture. Table 2 shows results on NYUV2 dataset and Fig. 5 shows example depth output results of our approach and the baseline. Photon Net outperforms baseline approach both qualitatively and quantitatively for multiple noise levels.

7. Experiments on Real SPAD Images

In order to evaluate the validity of the SPC image simulation model and the proposed approaches on real SPAD images, we collect a data-set of SPAD images using a SwissSPAD2 camera [41] (Fig. 6).

Camera Setup: We operate the camera in the binary mode where it captures binary frames at a spatial resolution of 512×256 with maximum frame rate at 96.8kHz. Currently, the sensor is not equipped with Bayer filters, so only grayscale (single channel) frames are captured. The captured images contain hot pixels which we correct in post processing. We capture an image of a black scene to identify the location of the hot pixels and then filter them by using spatial neighborhood information.

(a) CUB-200-2011 (b) CARS-196

Figure 4: **Effect of Photon Scale Space Parameters on Inference Performance:** Top-1 classification accuracy of Photon Net on $\mathcal{S}^1$ test images (PPP~0.11) with increasing number of levels in Photon Scale Space. Performance increases with increasing number of levels in PSS and saturates at 5-6 levels for both datasets.
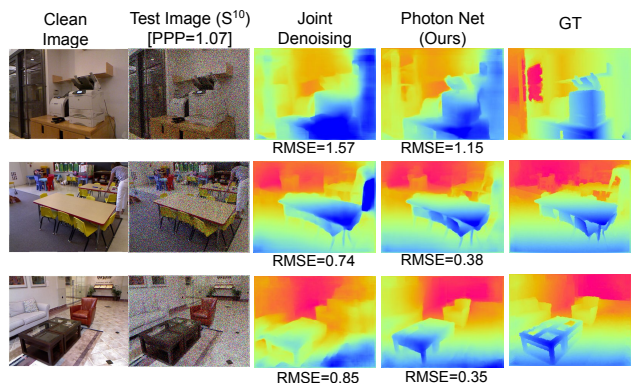

Figure 5: **Estimated depth maps**: Comparison of depth maps from Photon Net (Ours) and the baseline on NYUV2 test images $\mathcal{S}^{10}$ (PPP~1.07).

Dataset: For the dataset collection for image classification task, we displayed the original RGB images on a moni-
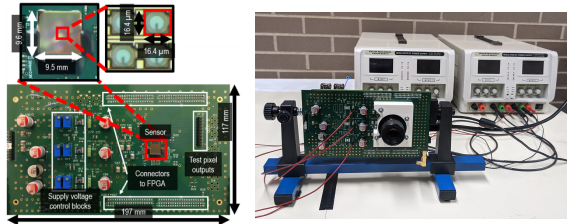
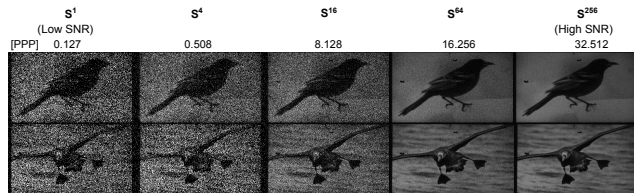Figure 6: **Camera Setup:** SwissSPAD2 board (Left) and the setup for image capture (Right).

Figure 7: **Real SPAD images** captured from CUB-200-2011 dataset using the SwissSPAD2 camera.
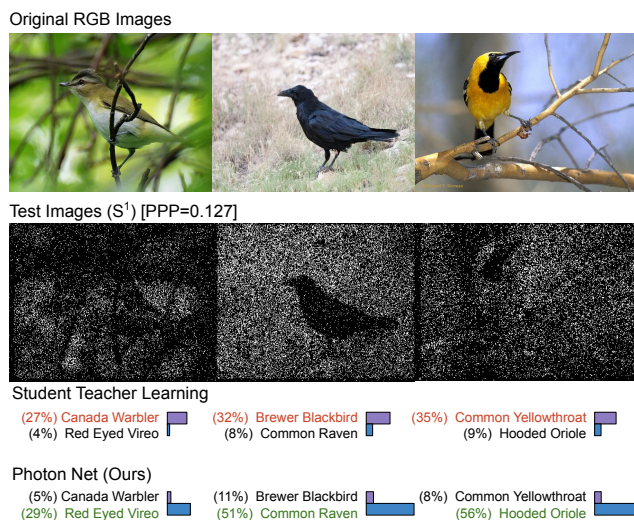

Figure 8: **Results with Real SPAD Sensor** of image classification on CUB-200 dataset for $\mathcal{S}^1$ test images with prediction probabilities output by both Student Teacher Learning and Photon Net (Ours). Classification output is highlighted in red for wrong prediction and green for correct prediction.

tor screen (Dell P2419H, 60Hz) and then captured it using SPAD sensors. The camera is placed at around 1m distance from the screen and positioned to cover the display in its field of view. We selected a subset of images from CUB-200-2011 dataset (CUB-subset) for the data collection, including 3656 training images and 3518 testing images from a randomly collected subset of 122 categories. Fig. 7 shows examples of N-Sum images captured by the camera.

Experiments and Results: We follow the same procedure

Test Data	PPP	Joint Denoising	Student-Teacher Learning	Photon Net (Ours)
$\mathcal{S}^1$	0.127	13.34	17.54	**21.78**
$\mathcal{S}^2$	0.254	16.57	20.67	**26.74**
$\mathcal{S}^4$	0.508	18.82	24.55	**32.33**
$\mathcal{S}^8$	1.016	21.07	28.34	**35.79**
$\mathcal{S}^{16}$	2.032	24.91	29.82	**39.14**

Table 3: **Experiments with real SPAD data.** Top-1 image classification results on CUB-subset images captured using a SPAD camera.

for training as described in Section 6.1. Table 3 shows results of our approach on real images from SPADs. Although the overall accuracy levels are lower (for all approaches) than those with simulated images due to the real images having a lower resolution and only gray-scale intensities (no Bayer filter on the real SPAD sensor), Photon Net outperforms both baselines on all noise levels.

Fig. 8 shows output probabilities of the predicted classes with ground truth for a few samples. Even in extreme low-light conditions with PPP as low as ~ 0.1, the proposed photon net approach is able to recover correct class labels.

8. Discussion and Limitations

Low-light inference beyond classification and depth estimation: So far, we have demonstrated the benefits of the proposed approaches for image classification and depth estimation tasks. A natural direction is to extend these ideas to inference models for a larger gamut of image inference and scene understanding tasks, including object detection [35], instance segmentation [19] and key-point detection [30]

Inference in high-flux scenarios: Although the primary focus of this paper is on low-light inference, due to the high dynamic range capabilities of SPADs [2, 21, 28], the proposed techniques can be adapted for inference in extremely bright scenes where conventional sensors get saturated.

Inference on time-varying inputs: In their current form, the proposed approaches assumes static single-frame input. However, most current single-photon sensors [27, 41] can capture binary frames at high speeds, up to several thousand frames per second. A promising future research direction is to perform inference in the presence of high-speed camera / scene motion on a temporal sequences of such low bit-depth frames, while exploiting temporal correlations.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.

[2] Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Dynamic range extension for photon counting arrays. *Optics Express*, 26(17):22234, Aug. 2018.

[3] VE Antsiperov. Target identification for photon-counting image sensors, inspired by mechanisms of human visual perception. In *Journal of Physics: Conference Series*, volume 1368, page 032020. IOP Publishing, 2019.

[4] Claudio Bruschini, Harald Homulle, Ivan Michel Antolovic, Samuel Burri, and Edoardo Charbon. Single-photon avalanche diode imagers in biophotonics: review and outlook. *Light: Science & Applications*, 8(1):1–28, 2019.

[5] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–21011, 2015.

[6] Bo Chen and Pietro Perona. Vision without the image. *Sensors*, 16(4):484, 2016.

[7] Bo Chen and Pietro Perona. Seeing into darkness: Scotopic visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3826–3835, 2017.

[8] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H. Chan. Dynamic low-light imaging with quanta image sensors. *European Conference on Computer Vision*, 2020.

[9] Henri Dautet, Pierre Deschamps, Bruno Dion, Andrew D. MacGregor, Darleene MacSween, Robert J. McIntyre, Claude Trottier, and Paul P. Webb. Photon counting techniques with silicon avalanche photodiodes. *Appl. Opt.*, 32(21):3894–3900, Jul 1993.

[10] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *Proc. ICCV*, 2019.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[12] 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.

[13] Eric R Fossum. What to do with sub-diffraction-limit (sdl) pixels?—a proposal for a gigapixel digital film sensor (dfs). In *IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors*, pages 214–217, 2005.

[14] Abhiram Gnanasambandam and Stanley H. Chan. Image classification in the dark using quanta image sensors. *European Conference on Computer Vision*, 2020.

[15] Anant Gupta, Atul Ingle, and Mohit Gupta. Asynchronous single-photon 3d imaging. *IEEE ICCV*, 2019.

[16] Anant Gupta, Atul Ingle, Andreas Velten, and Mohit Gupta. Photon flooded single-photon 3d cameras. *IEEE CVPR*, 2019.

[17] Istvan Gyongy, Neale AW Dutton, and Robert K Henderson. Single-photon tracking for high-speed vision. *Sensors*, 18(2):323, 2018.

[18] Kaiming He. Guided image filtering. In *Proc. ECCV*, 2010.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Atul Ingle, Andreas Velten, and Mohit Gupta. High flux passive imaging with single-photon sensors. *IEEE CVPR*, 2019.

[22] Ahmed Kirmani, Dheera Venkatraman, Dongeek Shin, Andrea Colaço, Franco N. C. Wong, Jeffrey H. Shapiro, and Vivek K Goyal. First-photon imaging. *Science*, 343(6166):58–61, 2014.

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[24] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *J. of Applied Statistics*, 1994.

[25] 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 Transactions on Image Processing*, 29:3695–3706, 2020.

[26] Zhuang Liu, Tinghui Zhou, Hung-Ju Wang, Zhiqiang Shen, Bingyi Kang, Evan Shelhamer, and Trevor Darrell. Transferable recognition-aware image processing. *arXiv preprint arXiv:1910.09185*, 2019.

[27] Jiaju Ma, Saleh Masoodian, Dakota A Starkey, and Eric R Fossum. Photon-number-resolving megapixel image sensor at room temperature without avalanche gain. *Optica*, 4(12):1474–1481, 2017.

[28] Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics (TOG)*, 39(4):79–1, 2020.

[29] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated spad image sensor for 2d and 3d imaging applications. *Optica*, 7(4):346–354, 2020.

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[31] Cristiano Niclass, Alexis Rochas, P-A Besse, and Edoardo Charbon. Design and characterization of a cmos 3-d image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9):1847–1854, 2005.

[32] D.V. O'Connor and D. Phillips. *Time-correlated single photon counting*. Academic Press, 1984.

[33] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[36] D. Renker. Geiger-mode avalanche photodiodes, history, properties and problems. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 567(1):48 – 56, 2006. Proceedings of the 4th International Conference on New Developments in Photodetection.

[37] Alexis Rochas. Single photon avalanche diodes in cmos technology. Technical report, Citeseer, 2003.

[38] Alexis Rochas. *Single Photon Avalanche Diodes in CMOS Technology*. PhD thesis, EPFL, 2003.

[39] Dongeek Shin, Feihu Xu, Dheera Venkatraman, Rudi Lussana, Federica Villa, Franco Zappa, Vivek K. Goyal, Franco N. C. Wong, and Jeffrey H. Shapiro. Photon-efficient imaging with a single-photon camera. *Nature Communications*, 7:12046, Jun 2016.

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[41] Arin Can Ulku, Claudio Bruschini, Ivan Michel Antolović, Yung Kuo, Rinat Ankri, Shimon Weiss, Xavier Michalet, and Edoardo Charbon. A 512×512 spad image sensor with integrated gating for widefield flim. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–12, 2018.

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[43] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A. Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning, 2020.

[44] F. Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli. Bits from photons: Oversampled image acquisition using binary poisson statistics. *IEEE Transactions on Image Processing*, 21(4):1421–1436, 2012.