Photon-Limited Object Detection using Non-local Feature Matching and Knowledge Distillation

Chengxi Li[1], Xiangyu Qu[1], Abhiram Gnanasambandam[1], Omar A. Elgendy[2], Jiaju Ma[2], and Stanley H. Chan[1]
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
[2]GigaJot Technology Inc., Pasadena, California, USA
{li2509, qu27, agnanasa, stanchan}@purdue.edu, {oelgendy, jiaju.ma}@gigajot.tech

Abstract

Robust object detection under photon-limited conditions is crucial for applications such as night vision, surveillance, and microscopy, where the number of photons per pixel is low due to a dark environment and/or a short integration time. While the mainstream "low-light" image enhancement methods have produced promising results that improve the image contrast between the foreground and background through advanced coloring techniques, the more challenging problem of mitigating the photon shot noise inherited from the random Poisson process remains open. In this paper, we present a photon-limited object detection framework by adding two ideas to state-of-the-art object detectors: 1) a space-time non-local module that leverages the spatial-temporal information across an image sequence in the feature space, and 2) knowledge distillation in the form of student-teacher learning to improve the robustness of the detector's feature extractor against noise. Experiments are conducted to demonstrate the improved performance of the proposed method in comparison with state-of-the-art baselines. When integrated with the latest photon counting devices, the algorithm achieves more than 50% mean average precision at a photon level of 1 photon per pixel.

1. Introduction

State-of-the-art object detection methods such as Faster R-CNN [64] and YOLO [63] are the backbones of many computer vision systems today, but their operating regimes have been limited to well-illuminated scenes with a sufficient amount of photons. As the number of photons decreases so that the signal-to-noise ratio becomes lower, the performance of these detectors will also degrade. For applications where photon-limited imaging is essential (e.g., night-time navigation, surveillance in an under-resourced environment, and microscopy with limited fluorescence dosage and cell exposures,) developing a more robust object detection algorithm presents a pressing need. The goal of this paper is to fill the gap by demonstrating object detection where existing methods fail to work.

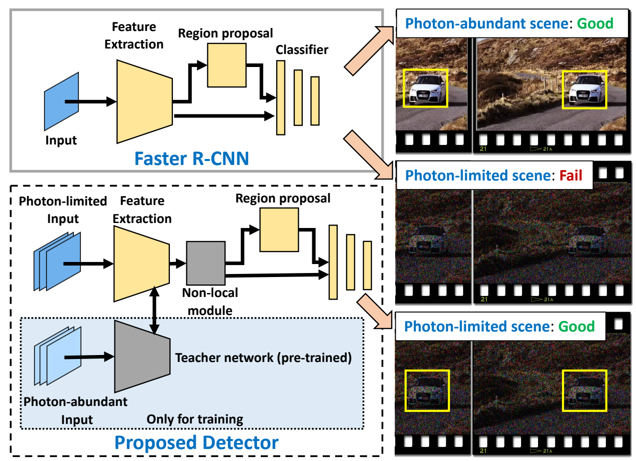


Figure 1: We present a new object detection method for photon-limited conditions. While traditional detectors fail because the signal is too weak, our method addresses the problem by proposing two improvements: (1) Space-time non-local module, and (2) Student-teacher learning.

Photon-limited imaging refers to image acquisition under a condition where the number of measured photons is very low. The fundamental limit is attributed to the Poisson process of the photon arrivals. This randomness is present even if the sensor is perfect – no read noise, no dark current, and has a uniform pixel response. Because the randomness is the nature of the problem, a photon-limited object detection algorithm must be able to extract the weak signal from the noise. Existing low-light enhancement algorithms have demonstrated promising results of improving the contrast of low-light images. In this paper, we are interested in pushing

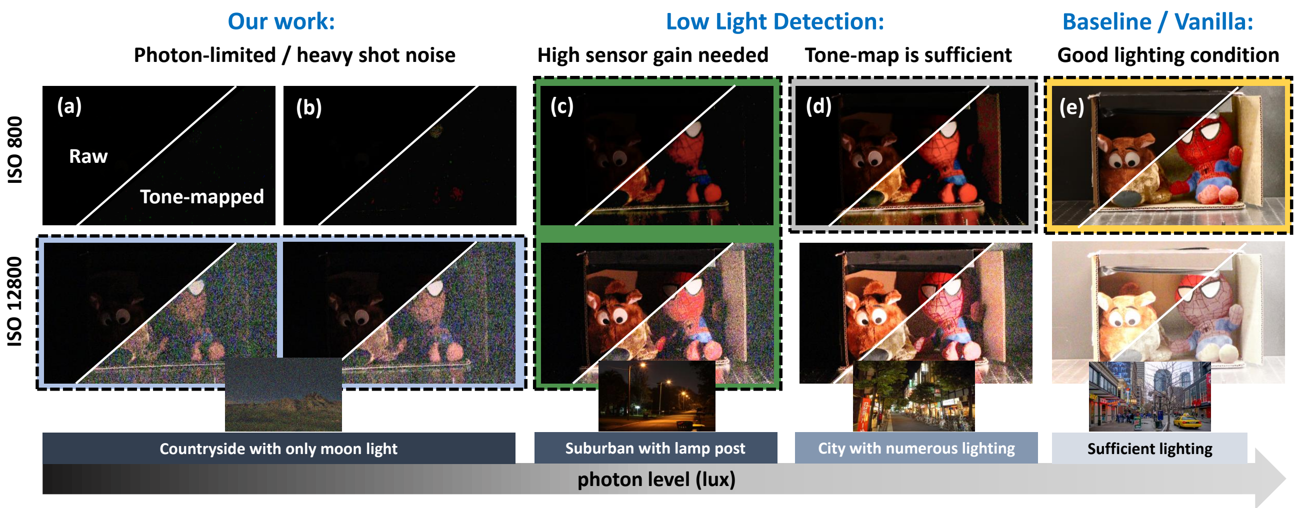


Figure 2: While baseline/vanilla methods [2, 8, 11, 13, 20, 36, 37, 46, 49, 51, 52, 63, 64, 70, 73, 80, 81] are designed to handle well-illuminated scenes, this paper focuses on the photon-limited regime where signals are very weak. Existing "low-light" methods [7, 53, 67, 78] typically do not operate in such an extreme condition where the signal is weak even after tone-map and/or adjusting the sensor's ISO.

the limit further by considering images that do not only have a low contrast but are also contaminated with shot noise.

The contributions of this paper are summarized in Figure 1. While conventional methods such as Faster R-CNN fail to detect objects under photon-limited conditions, we propose two improvements to overcome the difficulty:

- Leverage spatial-temporal redundancy. We assume that the input data is a burst of photon-limited frames. Although motion exists across the burst of frames, the total signal-to-noise ratio (SNR) of a burst is higher than a single frame. By borrowing ideas from the non-local neural network [71], we build a space-time non-local feature aggregation module to assemble neighboring space-time features.
- Regularize features via student-teacher knowledge distillation. The construction of the non-local features is based on feature matching. The success of feature matching depends on the SNR of the features. To maximize the SNR of the features, we employ a knowledge distillation technique where the feature extraction module of a student network is trained to mimic the features produced by a pre-trained teacher.

By incorporating the two improvements into Faster R-CNN, we offer improved detection performance. Our experimental results show that the new algorithm outperforms the baselines by more than 6% in mean accuracy precision (mAP). Given a desired mAP level, our system requires up to 50% fewer photons. When combined with the latest single-photon image sensors [55], we achieve object detection at 1 photon per pixels (PPP) or lower on real images.

2. Related Work

The taxonomy of the object detection methods is outlined in Figure 2, where we compare different detection tasks/methods against the photon-level (measured in lux) and the sensor gain (measured in ISO).

2.1. Baseline / Vanilla Methods

The mainstream object detection methods that are trained using large scale data set such as ILSVRC [66] and COCO [50] typically operate at the right most column of Figure 2 where the number of photons is sufficient. Depending on the input data format, the methods can be categorized into the following two groups:

Single-image detection methods that detect objects from a single image. Some of these methods focus on speed and real time processing capability [46, 49, 52, 63], whereas other methods based on region proposal focus on detection performance [11, 36, 37, 64]. On top of these methods, various work are proposed by leveraging multi-scale information [48], making network fully convolutional [11], utilizing multi-task training [36], tackling foreground-background imbalance [49], and improving bounding box prediction quality [39, 79].

Video detection methods that detect objects from multiple frames of a video. The premise of these methods is that the temporal information and the spatial-temporal redundancy provides valuable information for the detection. The aggregation of temporal cues are typically done at two levels: (i) feature level aggregation [2, 51, 70, 73, 80, 81], and (ii) box level aggregation [8, 13, 20, 70].

Despite the abundance of baseline methods, the networks and training are not designed for photon-limited conditions. As a result, directly applying these methods to our problem is ineffective (performance is limited even if one augment training data) and inefficient (pre-processing could be computationally expensive but does not necessarily lead to unparalleled performance), as demonstrated in [27, 78] and in our experiment.

2.2. Low-Light Detection Methods

Conventional low-light image processing methods can handle darker images than the baselines as shown in Figure 2(c) and (d). The easier case, as shown in Figure 2(d), happens when the lighting condition is not properly adjusted. However, information is mostly intact after tone-mapping and contrast enhancement. Image enhancement for this class of problem has been extensively studied [1, 10, 25, 29, 30, 35, 40, 43, 44, 54, 60, 65, 69, 74, 77]. For object detection, Loh et al. [53] and Yang et al. [78] created large-scale real low light detection data sets. The state-of-the-art detection systems in this scenario adopt Multi-Scale Retinex with Color Restoration (MSRCR) algorithm [43] for pre-processing and fine tune detectors on pre-processed data [78]. As will be shown in the experiment section, this strategy fails to work on photon-limited images; the strong photon shot noise will void the illumination smoothness assumption held by the Retinex model.

The harder case of the two, as shown in Figure 2(c), happens when the photon level is further reduced. In this operating regime, one needs to switch to a high sensor gain (higher ISO) so that the details can be observed. As far as object detection algorithms are concerned, to the best of our knowledge, no large scale detection dataset is available to date. Instead, Sasagawa et al. [67] treat detection in this scenario as a domain adaptation problem and use knowledge distillation to train a detector with normal lighting detection data and SID reconstruction data set [7]. In our study, we simulate the physical process of photon-limited image formation and demonstrate that our simulation enables our model to work on real photon-limited images.

2.3. Photon-Limited Imaging Methods

When the light level is extremely low or the exposure time is extremely short, each pixel only receives a handful of photons. Images captured under this condition are dominated by photon shot noise as shown in Figure 2(a)-(b), which are the cases of interest in this paper.

For object detection at this photon level, the pioneer study by Chen et al. [6] shows the feasibility of performing classification under such condition on MNIST [47] data set. Various new types of image sensors have been developed over the past few years, including the single-photon avalanche diodes (SPAD) [3, 5, 14–16, 34, 61, 62] and the quanta image sensors (QIS) [21–24, 56, 57]. A lot work has also been done in the signal processing side of both these sensors [4, 17, 18, 26, 28, 31, 32, 41, 42, 58, 75]. Specific to high-level computer vision tasks, Gyongy et al. demonstrated tracking and reconstruction of rigid planar object at this light level [33]. Gnanasambandam et al. [27] and Chi et al. [9] achieved image reconstruction and classification by combining student-teacher training scheme. The proposed idea is inspired by the student-teacher scheme. To further improve the performance, we introduce a spatial-temporal non-local module to leverage the information from neighbor frames. Our method generalizes the conventional detection methods by providing a more robust detection under photon-limited conditions.

3. Method

Given a sequence of photon-limited frames, our goal is to localize objects and identify their classes in *all* frames. Our proposed system is trained on data obtained from Sec 3.1 and consists of key components: the non-local module (Sec 3.2) and the student-teaching learning scheme (Sec 3.3).

3.1. Image Formation Model

Under a photon limited condition, the signal generated by the image sensor, $\boldsymbol{x}$, is modeled through a Poisson process [6, 9, 27]:

$$\boldsymbol{x} = \text{Poisson}(\alpha \cdot \text{CFA}(\boldsymbol{y}_{\text{RGB}}) + \boldsymbol{\eta}_{\text{dc}}) + \boldsymbol{\eta}_{\text{r}}, \tag{1}$$

where CFA stands for the color filter array. $\boldsymbol{y}_{\text{RGB}}$ is the clean RGB image in the range $[0, 1]$. α determines the average number of photons arriving at the sensor and therefore it depends on the exposure time and the average photon flux of the scene. $\boldsymbol{\eta}_{\text{dc}}$ is the dark current, and $\boldsymbol{\eta}_r \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{r}}\mathbb{I})$ is the readout noise with standard deviation σ_{r}.

The final output $\boldsymbol{x}$ is truncated at 3 standard deviation from mean pixel values and re-normalized to the range $[0, 1]$. All frames are assumed to be statistically independent, as the Poisson process and the noise are independent [68]. In our experiments, we used values listed in table 1, following [6, 27, 72]. The dark current parameter is set to 0 as it is insignificant compared to other noise sources on modern sensors when the exposure time is short.

α	$\boldsymbol{\eta}_{\text{dc}}$	σ_{r}
0.25 — 5	$\mathbf{0}$	0.25 or 2

Table 1: Data synthesis parameters used in our experiments

3.2. Space-Time Non-Local Module

The biggest challenge of detecting objects under photon-limited conditions is the presence of intense shot noise. Our solution to extract signals from the noise is to utilize the

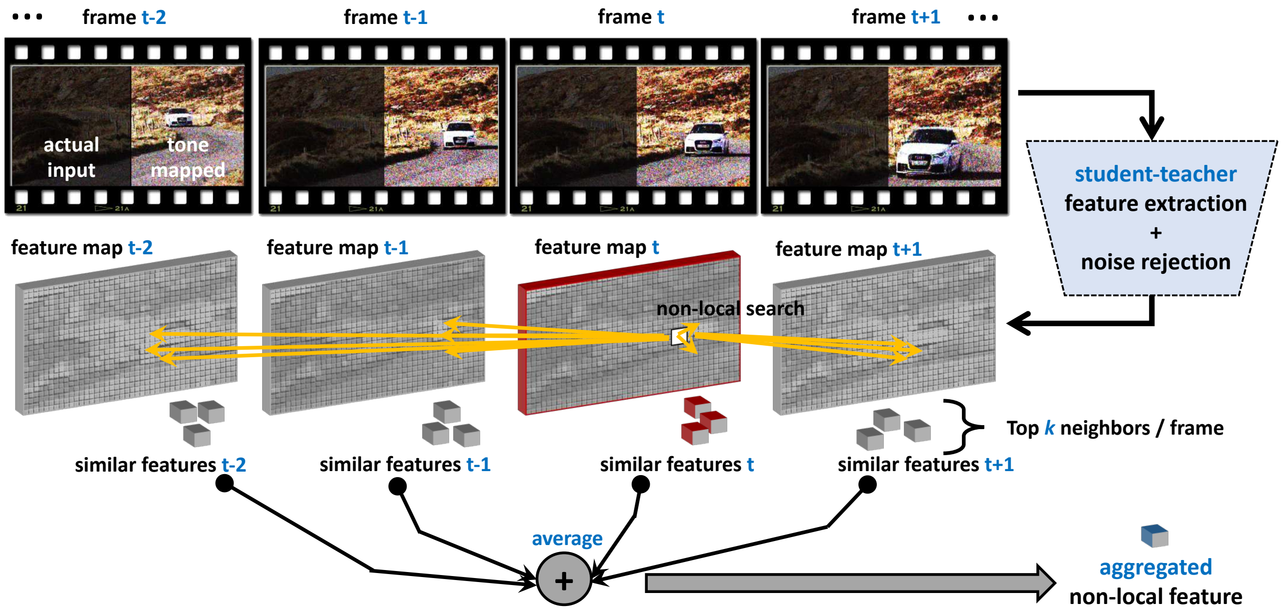


Figure 3: Our proposed non-local module and student-teacher training scheme. The teacher network is first pre-trained on photon-abundant data and it enforces the student to extract noise-rejected features of each input frame. By applying the non-local search in the feature space, similar spatial-temporal features are aggregated to update the key frame features.

spatial-temporal redundancy across a burst of frames. Our hypothesis is that if we are able to find similar patches in the space-time volume, we can take a non-local average to boost the signal. To achieve this goal, we design a non-local module as depicted in Figure 3.

Given an image sequence, each frame is fed into a feature extractor (the student-teacher module, which will be discussed in Section 3.3) to obtain the feature maps. For each feature vector at location (i, j, t), we conduct a non-local search for similar features by computing the inner-products of this feature and all the candidate features in the adjacent frames. This operation produces a set of scalars representing the similarities between the current feature and the features in the space-time neighborhood. Then for every time t, we select the top-k candidates with the highest inner product values. As shown in the Supplementary Materials, we find that $k = 2$ is an appropriate number for most of the experiments. After picking the top-k features, we take the average to generate the aggregated non-local feature.

Our proposed space-time non-local module differs from the traditional non-local neural networks [71] in the following two aspects:

- Before computing the similarity, [71] uses convolutional layers to first project features onto another feature space. This additional feature space is designed to represent high-level semantic meanings of the scene, such as interactions. For photon-limited imaging where the SNR is low, such semantic-level features are generally more corrupted and hence they are less reliable than low-level features. In addition, feature projection could cause confusion to our spatial-temporal feature matching step because the noise is heavy.
- [71] aggregates *all* space-time information via a softmax weighted average. We only average partially the space-time information from the top-k features because irrelevant features in the time-space can distract our model. In the Supplementary Material, we demonstrate that the top-2 features per frame are sufficient for our purpose.

3.3. Knowledge Distillation

The performance of the non-local feature matching depends heavily on the SNR of the features. If the features are contaminated by noise, finding correct feature correspondence would be difficult. Inspired by [9, 27], we introduce a knowledge distillation step known as the student-teacher learning scheme to regularize the features. The idea is to train the student feature extractor by minimizing its L_2 distance with a teacher pre-trained on clean data so that the features extracted by the student are denoised.

Figure 4 depicts the idea of the proposed student-teaching learning scheme. In this figure, we have a teacher network and a student network. The teacher network is pre-trained using well-illuminated images. The student network has the same architecture but it is used to extract features from the photon-limited data (i.e., noisy). During training,

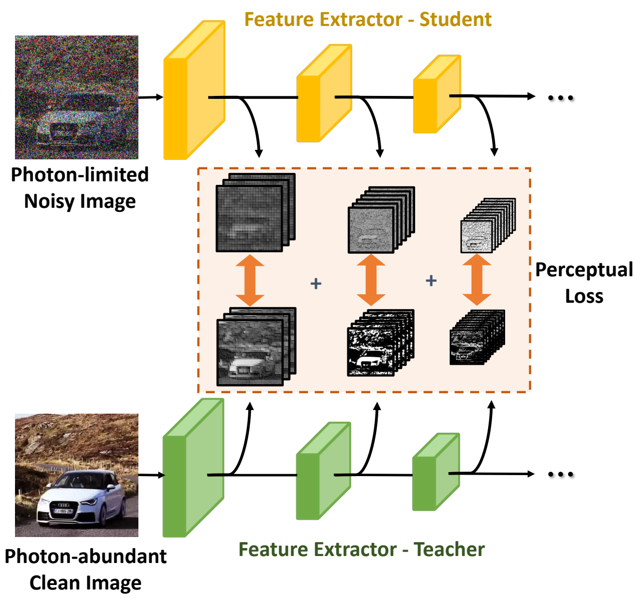


Figure 4: Knowledge distillation via student-teacher learning. The teacher network is pre-trained on clean images. We train the student network by minimizing the perceptual loss which measures the pixel-wise difference of the features.

the parameters of the teacher network are fixed and those of the student network are trainable. Because the teacher network is trained to handle clean images, it generates noise-free features when it is fed with clean images. We want features produced by the student network to be similar to those of the teacher. To this end, we introduce regularization to the student network by defining a **perceptual loss**:

$$\mathcal{L}_p = \sum_{i=1}^{N} \|\widehat{\phi}_i(x_{\text{clean}}) - \phi_i(x_{\text{noisy}})\|^2, \tag{2}$$

where $\widehat{\phi}_i(x_{\text{clean}})$ and $\phi_i(x_{\text{noisy}})$ are the i-th layer's feature of the teacher and student network, respectively. The perceptual loss is the Euclidean distance measuring the difference between the student's and the teacher's features. Minimizing the perceptual loss forces them to be close in the feature space. This further enforces the network to denoise the image and generate good representations before non-local feature matching.

The overall training loss of our detector consists of the perceptual loss $\mathcal{L}_p$, the standard cross-entropy loss, and the regression loss [64].

3.4. Rationale of Our Design

To illustrate the benefit of the proposed non-local module and the student-teacher learning scheme, we conduct an experiment in this section.

In Figure 5, we synthesize two independent and identically distributed (i.i.d.) copies of a photon-limited image at a photon level of 0.25 photons per pixel (ppp). We use this pair of images to check how the feature matching step performs. Three methods are compared: 1) Non-local search in the image space (i.e., the original non-local search), 2) non-local Search in the feature space, and 3) student-teacher + non-local Search in the feature space. In the image space, for each $h \times w$ patch, we compute its normalized cross-correlation (NCC) with all $h \times w$ patches in the other image and choose the one with the highest NCC as its matching patch. In the feature space, we use features trained with or without student-teacher training and find correspondence for every feature vector. The correspondence is visualized by the center of the receptive field of feature vectors.

The benefit of the proposed method can be seen in two aspects: accuracy and speed. As illustrated in Figure 5, the non-local search in the feature space has a much higher success rate of finding correct correspondence than the same method applied to the image space. The student-teacher training further increases the performance by enhancing the robustness of the feature extractor against noise. We performed the experiment for 100 images and we observed that the trend was consistent.

For the speed, non-local search in image space is computationally more expensive than in the feature space. Given an $H \times W$ image with desired patch size $h \times w$, the feature matching process takes approximately $(HW)^2 hw$ floating-point operations (FLOP) in the image space and $(\frac{HW}{S})^2 C$ FLOP's in the feature space, where C is feature vector dimension and S is spatial resolution compression ratio by the feature extractor. Reducing the patch size reduces the computation cost, but the matching quality deteriorates significantly. In our implementation, we use 64×64 for the image space search and it takes ~ 256 times more computation than in the feature space.

4. Experiments

4.1. Experimental Settings

Dataset. We use the procedure in Sec 3.1 to synthesize training data of the photon-limited images from the Pascal VOC 2007 dataset [19]. To synthesize motion across the frames, we introduce a random translation of image patches. The total movement varies from 7 to 35 pixels across 8 frames similar to [9]. For testing, we created a synthetic testing dataset and also collected a dataset of real images. The read noise of our model is assumed to be $0.25e^-$, based on the sensor reported in [55]. The average photon level we tested ranges from 0.1 to 5.0 photons per pixel (ppp). With an f/1.4 camera, 1.1μm pixel pitch, and 30ms integration, this range of photons roughly translates to 0.02 lux to 5 lux (typical night vision scenarios). For real data, we use the GJ01611 16MP photon counting Quanta Image Sensor developed by GigaJot Technology [55].

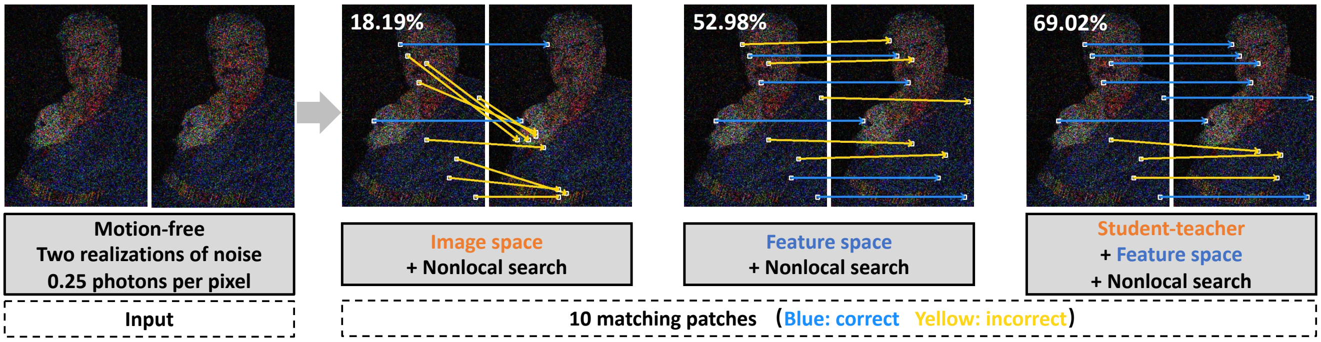
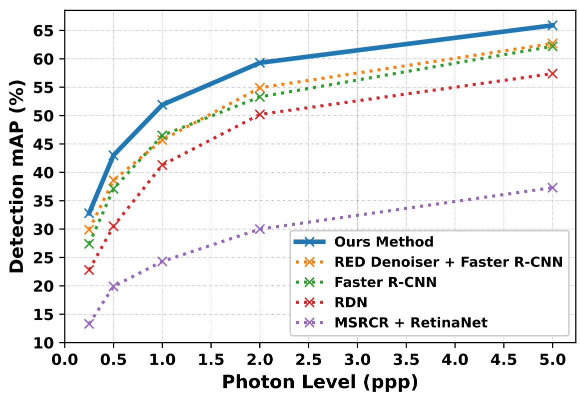
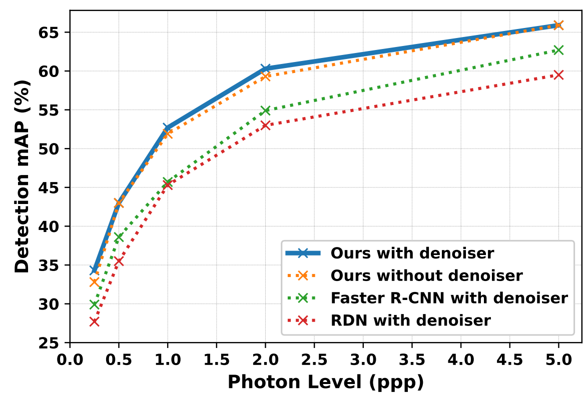


Figure 5: Comparison of different non-local patch matching methods. We synthesize two i.i.d. copies of a photon-limited image. For each competing configuration, we visualize 10 matching patch examples. The blue and yellow arrows indicate correct and incorrect matching, respectively. As the image pair is motion-free, the correct matches should be indicated by horizontal arrows. The combination of non-local search and student-teacher learning demonstrates the best performance.



(a) Comparison with baselines



(b) Comparison with image denoisers.

Figure 6: **Experiments on synthetic data.** (a) Compare different object detection methods: Faster R-CNN [64], RED [59] + Faster R-CNN [64], RDN [13], and MSRCR [43] + RetinaNet [49]. (b) Compare methods that use image denoising as a pre-processing step.

Implementation Details. Our method is implemented in Pytorch based on [76]. The framework takes a T-frame image sequence as input ($T = 1, 3, 5$ and 8 in the following experiments). We adopt ResNet-101 [38] pretrained on ImageNet [12] as the backbone. The perceptual loss is applied to the features from `block_1`, `block_2` and `block_3` of ResNet-101 and the non-local module is processed on the features from `block_3`. We utilize RoIAlign [36] to extract the features from object proposals and `block_4` is further applied to the extracted proposal features before the final classifier. The model is trained for 20 epochs and we use Adam [45] optimizer with default parameters, learning rate 0.001, and weight decay 0.1 every 5 epochs.

Competing Methods. We compare our method with four baselines. (a) A generic image object detector: Faster R-CNN [64]; (b) A video object detector: Relation Distillation Network (RDN) [13]; (c) A low-light detection framework: color restoration algorithm (MSRCR) [43] plus a detection RetinaNet [49], which is one of the winning solutions of 2019 UG2$^+$ low-light face detection challenge; (d) A two-stage pre-denoised detection framework: REDNet [59] plus Faster R-CNN [64]. (a) and (b) are fine-tuned using the synthesized photon-limited data.

4.2. Main Results

Our first experiment is conducted on synthetic data. We use 8-frame inputs with the number of features for non-local aggregation set to 2 per frame in the following experiments.

Comparison with the baselines. Figure 6a shows the detection rate, measured in mean average precision (mAP), as a function of the photon level, measured in photons per pixel (ppp). The proposed method consistently outperforms the competing methods across the tested photon levels from 0.25 ppp to 0.5 ppp. The difference between our method

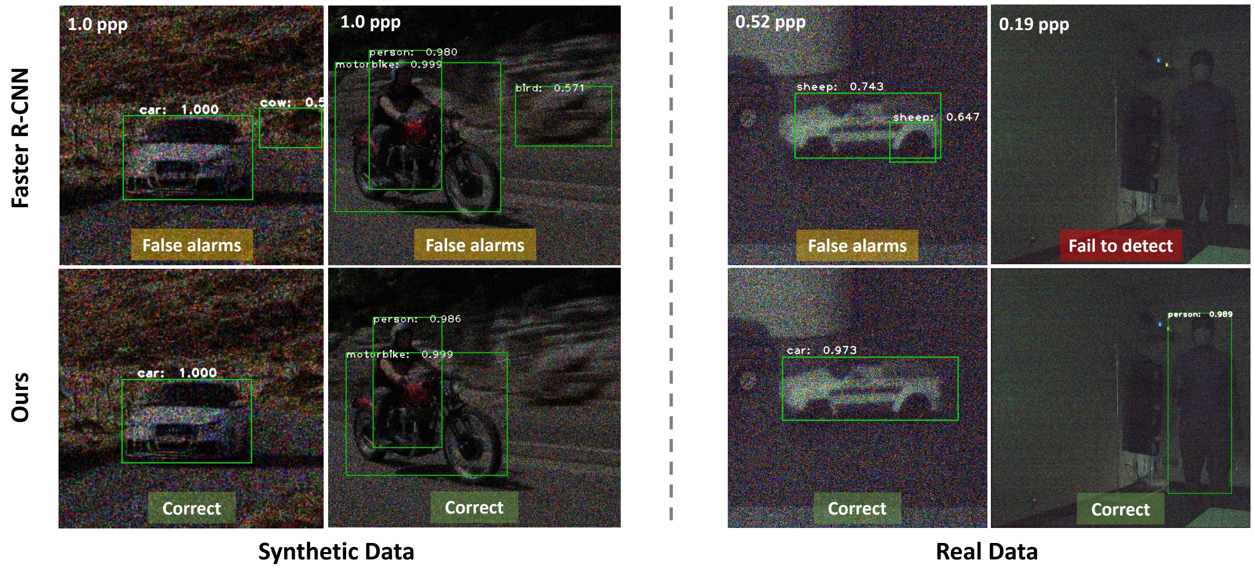


Figure 7: **Detection results on synthetic and real data.** The top row is the Faster R-CNN [64]. The bottom row is our method. The photon level is shown in the top-left corner. The real data is captured by Gigajot Technology 16 MP Photon Counting Quanta Image Sensor (GJ01611).

and the second-best method is as large as 6% in terms of mAP when the photon level is 2.0 ppp.

Comparison with image denoisers. When handling noisy images, a natural solution is to first run a denoiser and feed the denoised images into a standard object detector. Figure 6b depicts the comparisons with such baseline methods. The denoiser we use is the RED-Net [59] previously used in other photon-limited imaging papers such as [9] and [27]. As the figure indicates, the proposed method outperforms the baselines by a big margin. In addition, adding a denoiser to the proposed method offers almost no additional benefit. Therefore, the proposed method has effectively executed the denoising task without requiring another network for denoising.

Different network designs. Table 2 demonstrates the importance of the space-time non-local module and the student-teacher learning module. In this table, we present the relative performance gain compared with Faster R-CNN baseline [64]. The addition of the non-local module and the student-teacher training shows improvement upon the baseline. We observe that the performance gain shrinks when the photon level increases, as detection becomes easier. The combination of both designs shows the best performance across all photon levels, especially in extremely low light, where the relative gain is 20.07%.

Real data. We collected 225 real images in low light and annotate objects from 3 categories: `person`, `sheep`, and `car`. We train our model using the synthetic data and verify the results using the real data. The results are shown in Table 3. On average, our proposed method achieves an mAP of 87.9% while the baseline method achieves 66.9%.

Photon Level (ppp)	0.25	0.5	1.0	2.0	5.0
ST	9.12	6.20	4.52	5.44	2.57
NL	16.06	14.56	9.89	10.13	5.14
ST+NL	**20.07**	**15.90**	**11.61**	**11.26**	**5.95**

Table 2: **Comparison of different network designs**. Relative mAP increase are reported with respect to Faster R-CNN baseline. The unit is %. ST: student-teacher learning; NL: non-local module; ST+NL:student-teacher learning + non-local module.

	person	car	sheep	mAP (%)
Faster R-CNN	54/105	58/60	60/60	66.9
Ours	73/105	60/60	60/60	**87.9**

Table 3: **Detection results of real data**. Each class column shows the number of correct detections versus ground truth. The last column is the overall mAP.

Figure 7 shows a qualitative comparison between our method and the baseline Faster R-CNN. The result shows that the baseline suffers from either false alarms or missed detection. In contrast, the proposed method is able to detect the static toy car and moving person on the real data when the photon level is 0.52 ppp and 0.19 ppp, respectively.

4.3. Performance comparison with CIS and QIS

We evaluate the proposed method with a conventional CMOS image sensor (CIS) from Google Pixel 3XL and a

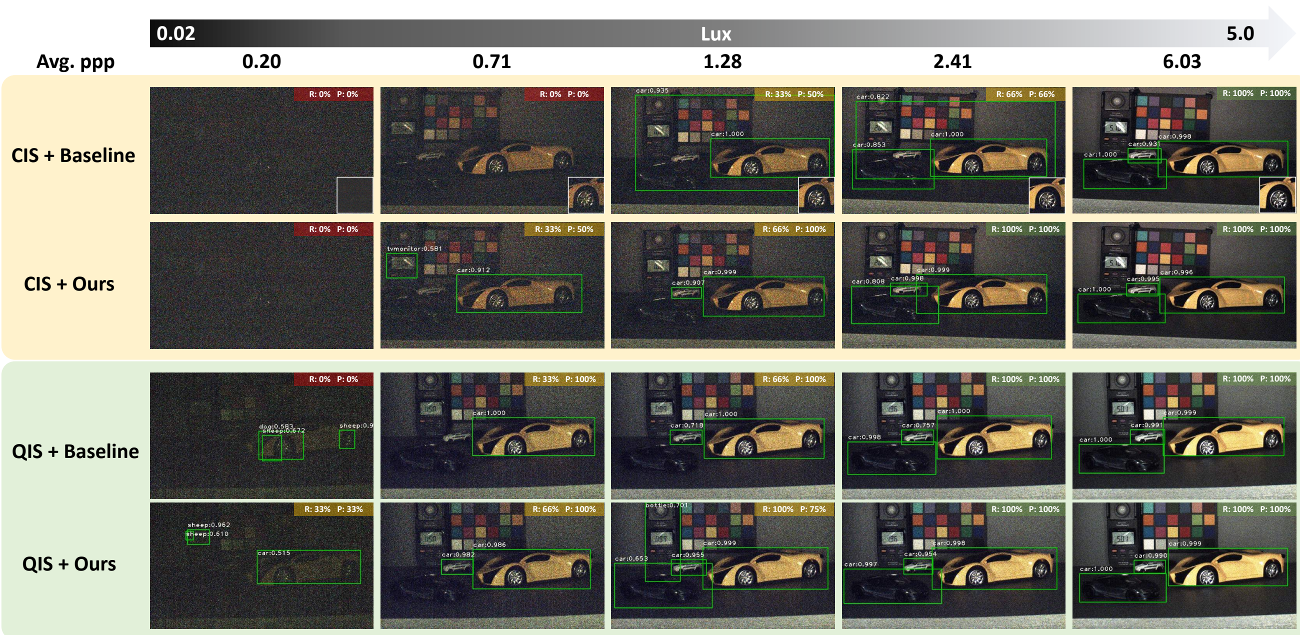


Figure 8: **Comparison of different sensors and different methods on real data.** The visualized figures are tone mapped and the baseline method is Faster R-CNN. We choose 5 different lux levels ranging from 0.02 to 5.0, equivalent to Avg. ppp ranging from 0.20 to 6.03. In the right-top corner of images, the recall (R) and precision (P) are computed, enclosed in frames with different colors. Red/Yellow/Green indicates totally failed/partially correct/totally correct, respectively. In the first row, we zoom into the left-front side of the yellow car and show details in the right-bottom box. We can see that in the extremely low light condition, the images suffer from the high-noise problem.

GJ01611 Quanta Image Sensor (QIS) from Gigajot Technology [57] under different illumination levels. By combining the proposed algorithm with the QIS device, we demonstrate the performance of the proposed detection method under extremely photon-limited conditions (0.02 lux and only 0.20 ppp).

To ensure a fair comparison, we note that the CIS has a pixel pitch of 1.4μm and read noise of 2.14e^-, while the QIS has 1.1μm pixels and read noise of 0.22e^-. In the experiments, the f-number of the lens is adjusted to balance the difference of pixel sizes (f/1.8 for CIS and f/1.4 for QIS) in the two sensors and 30msec exposure time is used for both sensors.

The comparison results are shown in Figure 8. The images were taken under illumination levels from 0.02 lux to 5.0 lux. Under strong illumination conditions such as 5.0 lux, all the compared methods show high detection accuracy without any false alarms. However, as the illumination level decreases, the proposed algorithm shows significant advantages over the baseline methods. This performance improvement is further enhanced with the QIS compared to the CIS because of its ultra-low read noise. For example, under 0.02 lux and an average photon level of 0.20 ppp, only the combination of the proposed algorithm and the QIS device can successfully detect the yellow car in the scene.

5. Conclusion

We proposed a photon-limited object detection framework. Our solution integrates a new non-local feature aggregation method and a knowledge distillation technique with the state-of-the-art detector networks. The two new modules offer better feature representations for photon-limited images. In comparison with the baselines, the proposed detector demonstrated superior performance in synthetic and real experiments. When applied to the latest photon counting devices, we demonstrated object detection at a photon level of 1 photon per pixel or lower, significantly surpassing the existing CMOS image sensors and algorithms. It is envisioned that the new detection framework will enable a variety of applications, such as security, defense, life science, and consumer, as well as the emerging medical applications.

6. Acknowledgment

The work is supported, in part, by the National Science Foundation under the grants CCF-1718007 and ECCS-2030570.

References

[1] Yousef Atoum, Mao Ye, Liu Ren, Ying Tai, and Xiaoming Liu. Color-wise attention network for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2130–2139, 2020. 3

[2] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, September 2018. 2

[3] Claudio Bruschini, Samuel Burri, Scott Lindner, Arin C Ulku, Chao Zhang, I Michel Antolovic, Martin Wolf, and Edoardo Charbon. Monolithic SPAD arrays for high-performance, time-resolved single-photon imaging. In *IEEE International Conference on Optical MEMS and Nanophotonics*, pages 1–5. IEEE, 2018. 3

[4] Stanley H. Chan, Omar A. Elgendy, and Xiran Wang. Images from bits: Non-iterative image reconstruction for quanta image sensors. *Sensors*, 16(11), 2016. 3

[5] Paramanand Chandramouli, Samuel Burri, Claudio Bruschini, Edoardo Charbon, and Andreas Kolb. A bit too much? high speed imaging from sparse photon counts. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9, 2019. 3

[6] Bo Chen and Pietro Perona. Vision without the image. *Sensors*, 16(4), 2016. 3

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, June 2018. 2, 3

[8] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10337–10346, June 2020. 2

[9] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H. Chan. Dynamic low-light imaging with quanta image sensors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 122–138, 2020. 3, 4, 5, 7

[10] D. Coltuc, P. Bolon, and J.-M. Chassery. Exact histogram specification. *IEEE Transactions on Image Processing*, 15(5):1143–1152, 2006. 3

[11] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 379–387, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *Proceedings of the IEEE International Conference on Computer Vision and pattern Recognition (CVPR)*, 2009. 6

[13] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7023–7032, October 2019. 2, 6

[14] Neale Dutton, Tarek Al Abbas, Istvan Gyongy, Francesco-paolo Mattioli Della Rocca, and Robert Henderson. High dynamic range imaging at the quantum limit with Single Photon Avalanche Diode based image sensors. *MDPI Sensors*, 18(4):1166, 2018. 3

[15] Neale AW Dutton, Istvan Gyongy, Luca Parmesan, Salvatore Gnecchi, Neil Calder, Bruce R Rae, Sara Pellegrini, Lindsay A Grant, and Robert K Henderson. A SPAD-based QVGA image sensor for single-photon counting and quanta imaging. *IEEE Transactions on Electron Devices*, 63(1):189–196, 2015. 3

[16] Neale AW Dutton, Istvan Gyongy, Luca Parmesan, and Robert K Henderson. Single photon counting performance and noise analysis of CMOS SPAD-based image sensors. *Sensors*, 16(7):1122, 2016. 3

[17] Omar A Elgendy and Stanley H Chan. Color Filter Arrays for Quanta Image Sensors. *IEEE Transactions on Computational Imaging*, 6:652–665, 2020. 3

[18] Omar A Elgendy, Abhiram Gnanasambandam, Stanley H Chan, and Jiaju Ma. Low-light Demosaicking and Denoising for Small Pixels using Learned Frequency Selection. *IEEE Transactions on Computational Imaging*, 7:137–150, 2021. 3

[19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010. 5

[20] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3038–3046, Oct 2017. 2

[21] Eric R Fossum. Gigapixel digital film sensor (DFS) proposal. *Nanospace Manipulation of Photons and Electrons for Nanovision Systems*, 2005. 3

[22] Eric R Fossum. Some thoughts on future digital still cameras. *Image sensors and signal processing for digital still cameras*, page 305, 2006. 3

[23] Eric R Fossum. Modeling the performance of single-bit and multi-bit quanta image sensors. *IEEE Journal of the Electron Devices Society*, 1(9):166–174, 2013. 3

[24] Eric R. Fossum, Jiaju Ma, and Saleh Masoodian. Quanta Image Sensor: Concepts and Progress. In Mark A. Itzler and Joe C. Campbell, editors, *Advanced Photon Counting Techniques X*, volume 9858, pages 1–14. International Society for Optics and Photonics, SPIE, 2016. 3

[25] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016. 3

[26] Abhiram Gnanasambandam and Stanley H Chan. HDR Imaging with Quanta Image Sensors: Theoretical Limits and Optimal Reconstruction. *IEEE Transactions on Computational Imaging*, 6:1571–1585, 2020. 3

[27] Abhiram Gnanasambandam and Stanley H. Chan. Image classification in the dark using quanta image sensors. In *Pro-*

ceedings of the European Conference on Computer Vision (ECCV), pages 502–519, 2020. 3, 4, 7

[28] Abhiram Gnanasambandam, Omar Elgendy, Jiaju Ma, and Stanley H. Chan. Megapixel photon-counting color imaging using quanta image sensor. *Opt. Express*, 27(12):17298–17310, Jun 2019. 3

[29] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2511–2520, October 2019. 3

[30] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1780–1789, June 2020. 3

[31] Anant Gupta, Atul Ingle, and Mohit Gupta. Asynchronous single-photon 3d imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7909–7918, October 2019. 3

[32] Anant Gupta, Atul Ingle, Andreas Velten, and Mohit Gupta. Photon-flooded single-photon 3d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6770–6779, June 2019. 3

[33] Istvan Gyongy, Neale A.W. Dutton, and Robert K. Henderson. Single-photon tracking for high-speed vision. *Sensors*, 18(2), 2018. 3

[34] Istvan Gyongy, Germán Mora-Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, and Jonathan Leach. High-speed Vision with a 3D-stacked SPAD Image Sensor. In Mark A. Itzler, Joshua C. Bienfang, and K. Alex McIntosh, editors, *Advanced Photon Counting Techniques XV*, volume 11721, pages 1–7. International Society for Optics and Photonics, SPIE, 2021. 3

[35] 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. Graph.*, 35(6), nov 2016. 3

[36] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020. 2, 6

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 2

[38] 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 (CVPR)*, pages 770–778, 2016. 6

[39] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2888–2897, 2019. 2

[40] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 3

[41] Atul Ingle, Trevor Seets, Mauro Buttafava, Shantanu Gupta, Alberto Tosi, Mohit Gupta, and Andreas Velten. Passive inter-photon imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8585–8595, June 2021. 3

[42] Atul Ingle, Andreas Velten, and Mohit Gupta. High flux passive imaging with single-photon sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6760–6769, June 2019. 3

[43] D.J. Jobson, Z. Rahman, and G.A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997. 3, 6

[44] Yeong-Taeg Kim. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE Transactions on Consumer Electronics*, 43(1):1–8, 1997. 3

[45] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 6

[46] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, September 2018. 2

[47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[48] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 2

[49] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2, 6

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, Cham, 2014. Springer International Publishing. 2

[51] Mason Liu and Menglong Zhu. Mobile video object detection with temporally-aware feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5695, June 2018. 2

[52] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[53] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 2, 3

[54] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light sim-

ulation dataset. *International Journal of Computer Vision*, 129:2175–2193, Jul 2021. 3

[55] Jiaju Ma, Saleh Masoodian, Dakota A Starkey, and Eric R Fossum. Photon-number-resolving Megapixel Image Sensor at Room Temperature without Avalanche Gain. *Optica*, 2017. 2, 5

[56] Jiaju Ma, Dakota Starkey, Arun Rao, Kofi Odame, and Eric R. Fossum. Characterization of quanta image sensor pump-gate jots with deep sub-electron read noise. *IEEE Journal of the Electron Devices Society*, 3(6):472–480, 2015. 3

[57] Jiaju Ma, Dexue Zhang, Omar A Elgendy, and Saleh Masoodian. A 0.19 e-rms Read Noise 16.7 Mpixel Stacked Quanta Image Sensor With 1.1 μm-Pitch Backside Illuminated Pixels. *IEEE Electron Device Letters*, 42(6):891–894, 2021. 3, 8

[58] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Trans. Graph.*, 39(4), July 2020. 3

[59] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 6, 7

[60] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2510, June 2018. 3

[61] Germán Mora-Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, Jonathan Leach, and Istvan Gyongy. High-speed Object Detection using SPAD Sensors. In Yakov Soskind and Lynda E. Busse, editors, *Photonic Instrumentation Engineering VIII*, volume 11693, pages 73–82. International Society for Optics and Photonics, SPIE, 2021. 3

[62] 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. *OSA Optica*, 7(4):346–354, 2020. 3

[63] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1, 2

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1, 2, 5, 6, 7

[65] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 3

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[67] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark - domain adaptation method for merging multiple models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 345–359, 2020. 2, 3

[68] Donald L. Snyder, Carl W. Helstrom, Aaron D. Lanterman, Mohammad Faisal, and Richard L. White. Compensation for readout noise in ccd images. *J. Opt. Soc. Am. A*, 12(2):272–283, Feb 1995. 3

[69] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6849–6857, June 2019. 3

[70] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 542–557, September 2018. 2

[71] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[72] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2767, June 2020. 3

[73] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, September 2018. 2

[74] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2281–2290, June 2020. 3

[75] Feng Yang, Yue M. Lu, Luciano Sbaiz, and Martin Vetterli. Bits from photons: Oversampled image acquisition using binary poisson statistics. *IEEE Transactions on Image Processing*, 21(4):1421–1436, 2012. 3

[76] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017. 6

[77] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3063–3072, June 2020. 3

[78] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan,

Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 2, 3

[79] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 516–520, New York, NY, USA, 2016. Association for Computing Machinery. 2

[80] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7218, June 2018. 2

[81] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 408–417, Oct 2017. 2