Robust Perception under Adverse Conditions

by

Bhavya Goyal

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2025

Date of final oral examination: 04/08/2025

The dissertation is approved by the following members of the Final Oral Committee:

Mohit Gupta, Professor, Computer Sciences
Yin Li, Professor, Biostatistics & Medical Informatics
Yong Jae Lee, Professor, Computer Sciences
Jean-François Lalonde, Professor, ECE, Université Laval

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor, Professor Mohit Gupta. Mohit has been an incredible mentor throughout my time at UW–Madison. He has provided an outstanding research environment that encourages independent thinking and personal growth. I have gained immensely from his constructive feedback, vast knowledge, and great patience.

I am also deeply grateful to Professor Yin Li. Yin has been an amazing collaborator and has provided invaluable insights across numerous projects. I was also fortunate to spend a summer at Cruise AI, where I collaborated with many exceptional researchers. I especially want to thank Professor Yong Jae Lee for his guidance during the internship and for being part of my PhD committee. Ankit Laddha and Zhe Wang were outstanding hosts during my internship, helping me learn many valuable lessons and experience the industry research environment. I also want to thank Professor Jean-François Lalonde for both serving on my PhD committee and being part of my research journey through collaborations that contributed to this thesis.

Many of the concepts and ideas in this thesis came to life with the help of the amazing members of our lab. In particular, I want to thank Felipe, Wei, and Sizhuo, whose expertise in simulations and hands-on knowledge of camera sensors were crucial. I have also appreciated the camaraderie, support, and feedback from Atul, Jongho, Shantanu, Matt, Varun, Sacha, Alex, Talha, Carter, Avery, Hadi, and many other fellow graduate students.

I feel incredibly fortunate to have had a wonderful group of people around me in Madison, especially during the early years of graduate school and the uncertain times of the pandemic. To my friends — Vishnu, Shantanu, Aarati, Nils, Bhumesh, Muni, Ashish, Ambarish, Rishabh, Ankit, Shashank, Sourav, Shubham, Abhay, Megh, Angel, Tara, Wendy and Tom — thank you for your unwavering friendship and support. Your presence made all the difference.

The research in this thesis was generously funded by the National Science Foundation CAREER Award (#1943149, #2003129), the Sony Faculty Innovation Award, and the Intel MLWiNS Award. I also want to thank Edoardo Charbon for the SwissSAPAD2 array and Andreas Velten for the FLIMera sensor used in the experiments.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A perception system is a crucial component of several applications, including autonomous driving, surveillance, embodied perception, consumer devices like smartphones, and many more. Deep learning techniques have matured significantly over the past years and are the primary choice for many perception and scene inference tasks. Although these models show high performance on the *overall accuracy* of the perception tasks, it is often not sufficient for their safe deployment in real-world applications, as *worst-case* performance is also an important consideration. For instance, the performance on challenging scenarios such as low-light, motion, and extreme weather is considered extremely crucial in determining the viability of a perception system for many safety-critical applications like autonomous driving.

My thesis is that *we can unlock robust perception under adverse real-world conditions by improving the worst-case performance of the perception models and employing well-designed learning approaches that are tied to its sensing modalities.*

To prove this thesis, we propose a perception framework that is robust under a variety of adverse scenarios. We consider numerous sensing modalities like single photon cameras (SPCs), which are based on single-photon avalanche diodes (SPADs) technology, LiDARs, and conventional CMOS cameras, under a variety of challenging conditions such as low light, camera or scene motion, and extreme weather.

First, we consider a SPAD LiDAR and introduce a new 3D scene representation called probabilistic point cloud (PPC), which allows us to do robust 3D object detection for distant low albedo objects. Next, we demonstrate robustness with the passive imaging mode of a single photon camera under extremely low light conditions, which results in low signal-to-noise ratio (SNR) captures. We introduce photon scale space, a collection of varying SNR images with the same scene content for training perception tasks. Finally, we show robust scene inference

using conventional cameras under low light and motion conditions. We discuss the tradeoff between two kinds of image degradations, *i.e*, motion blur and noise, which we refer to as dual corruption.

The work in this dissertation shows that the key to unlocking robust perception under various adverse conditions lies in (A) emerging sensing technologies such as SPADs because of the high timing precision capability, and (B) effective learning techniques designed for these sensors to train scene inference models.

1 INTRODUCTION

Deep learning based techniques have matured significantly for complex, real-world scene inference and perception tasks. With these advancements, a new axis in the performance space is emerging, driven by applications such as 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 robustness across all conditions. 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 nighttime and poorly-lit scenes, high-speed moving objects, and extreme weather conditions like fog, snow, rain, etc. Even the state-of-the-art inference algorithms tend to fail for such scenarios where the sensor has not collected sufficient light or the captured image has been degraded due to rain, fog, or snow.

Learning based methods for perception: Most perception models are learned from large amounts of data collected under common operating conditions. Challenging scenarios mentioned above form a small fraction of the overall scenes that are captured for training. This causes such scenarios to be *long-tail* cases for the deep learning models, where most learning based models tend to struggle. Thus, the usual training approaches lack designs that explicitly improve the performance on such long-tail cases resulting in unreliable worst-case perception.

Sensing Modalities for perception: Perception pipelines are designed for a variety of sensors, such as cameras, depth sensors like LiDARs, and radars. Many perception stacks often include multiple such sensing modalities. Each sensor has different operating conditions that are considered adverse for its usage. For instance, RGB cameras struggle at nighttime due to an insufficient amount of light during the capture. On the other hand, Time of Flight (ToF) depth sensors like LiDARs are not impacted by nighttime but suffer from large noise under strong ambient lighting conditions. Thus, adverse conditions for perception can

take different forms for each modality and hence require careful consideration for each individual sensor. In this chapter, we introduce each sensing modality that is considered in this dissertation, along with their corresponding challenging scenarios, namely SPAD LiDARs, single photon cameras, and conventional RGB cameras.

1.1 Active Imaging with SPADs

LiDARs are a prominent 3D imaging modality driving several applications, from embodied perception and autonomous vehicles (Wang et al., 2024; Li and Ibanez-Guzman, 2020), to surveillance (Guo et al., 2024), and more recently, deployed even in consumer devices (e.g., Apple iPhones). LiDARs are based on the time-of-flight (ToF) principle; a typical LiDAR consists of a laser source that emits short sub-nanosecond light pulses into the scene and a sensor that captures the reflected pulses; scene depths are estimated by computing the time delay between emitted and received pulses (Lange, 2000) (Fig. 3.1a). Increasingly, single photon avalanche diodes (SPADs) (Cova et al., 1996) are becoming the sensor-of-choice in LiDARs due to their high sensitivity (Pellegrini et al., 2000), and amenability to fabrication of high-resolution arrays (Morimoto et al., 2021). It is not a surprise that the next generation of LiDAR devices is dominated by solid-state, high-resolution, and low-cost SPAD technology.

A typical single-photon LiDAR builds a histogram of photon counts over time (Fig. 3.1b). Under ideal imaging conditions, the peak in the timing histogram can be reliably detected, resulting in an accurate estimation of the time delay and hence the scene depth (Fig. 3.1b). This estimated depth at each pixel can then be used to construct a 3D point cloud-based scene representation, a canonical input to various downstream 3D perception tasks (Li and Ibanez-Guzman, 2020).

Adverse conditions for SPAD LiDARs: Under several real-world conditions, the raw timing histograms do not have a single, clearly discernible peak (Fig. 3.1b). This could be due to a variety of factors, including (a) distant objects or low scene

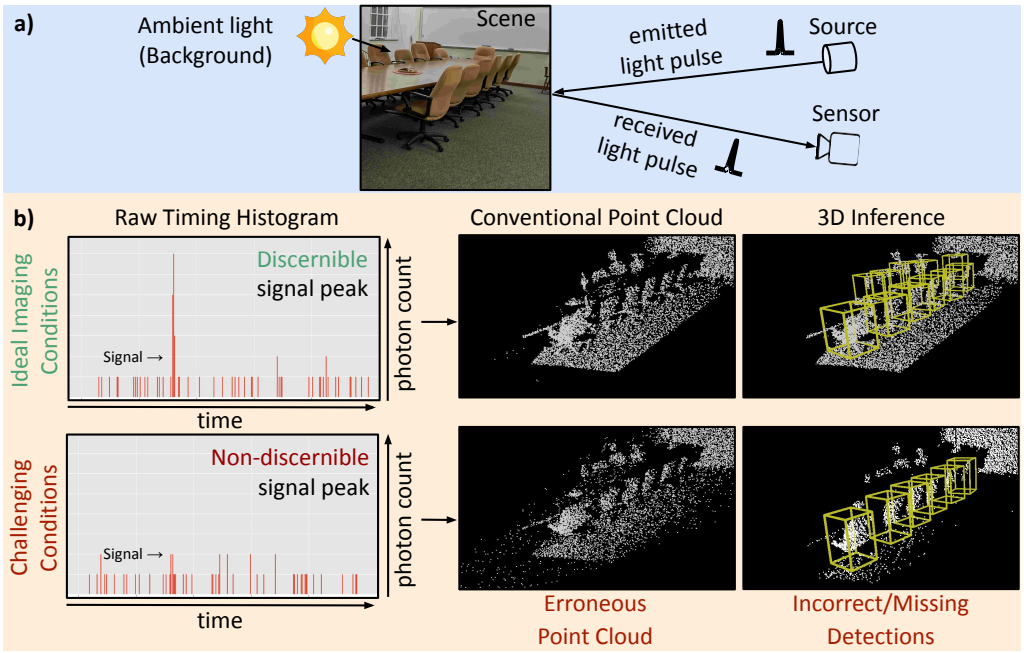


Figure 1.1: **3D Inference with SPAD LiDARs:** a) A LiDAR measures depth by emitting a pulsed laser and measuring time delays between emitted and received pulses. b) The raw sensor data at each LiDAR pixel is a temporal histogram (photon counts vs. time) where the time location of the peak corresponds to the scene distance, which is then converted to a 3D point in a point cloud.

albedo, (b) strong ambient illumination which increases the background level and noise, (c) other sources *e.g* non-diffuse and specular materials, multi-path or multi-camera interference, weather phenomena such as rain, snow, fog (Pellegrini et al., 2000; Pediredla et al., 2018; Beer et al., 2018; Satat et al., 2018) etc. Imagine a LiDAR mounted on an autonomous vehicle needing to safely navigate the world, not just under fair lighting and weather, but across all operating conditions. Under these scenarios, it is often challenging to detect the correct peak location, resulting in large depth errors.

This challenge is typically addressed by filtering techniques that retain only the measurements with prominent peaks, while discarding the rest (Zhang et al., 2013; Chen et al., 2020; Goudreault et al., 2023). Therefore, in challenging scenarios

mentioned above, most raw histograms that have small or ambiguous peaks either get filtered out, resulting in incorrect removal of scene components, or introduce significant noise in the final point cloud if they are retained. Such filtering steps severely affect the inference performance by: (1) removing critical scene content in low signal regimes due to over-aggressive filtering, or (2) propagating excessive measurement noise to downstream inference models. These issues are exacerbated for distant, small, or low albedo objects, which is indeed where a LiDAR sensor is needed the most. Fig. 1.1b shows an example where distant low albedo chairs are not detected by a downstream 3D detection model due to a sparse and noisy point cloud capture.

1.2 Passive Imaging with SPADs

Single-photon avalanche diodes (SPADs) 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 (Buttafava et al., 2015; O'Toole et al., 2018; Bruschini et al., 2019). But, recently, due to their compatibility with CMOS fabrication processes, high-resolution cameras up to 1 MP have been developed based on SPADs (Morimoto et al., 2020), as well as the jots (Ma et al., 2017) technology. SPADs have primarily been used for recovering image intensities (Antolovic et al., 2018; Ingle et al., 2019; Ma et al., 2020) and low/mid-level scene information such as 3D shape (O'Connor and Phillips, 1984; Renker, 2006; Dautet et al., 1993; Kirmani et al., 2014; Shin et al., 2016; Gupta et al., 2019c,a) and motion (Gyongy et al., 2018). High-level inference using single photon cameras is the next frontier for single photon imaging technology.

Adverse conditions of low light: One of the most critical scenarios of perception pipelines with cameras is low-light, due to the presence of a large amount of noise, primarily due to read noise and shot noise. Figure 1.2 shows a scene captured at nighttime and its brightened image for reference, which highlights the amount of

Figure 1.2: **Low Light Imaging:** Captures in low light suffer from extreme noise.

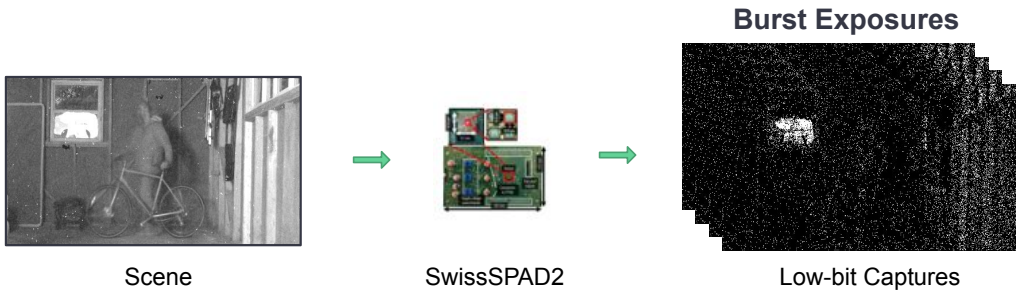


Figure 1.3: **Single Photon Cameras:** Low bit captures from SPCs contain large amounts of shot noise.

noise in low light. 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.

When operating in passive imaging mode, single photon cameras are capable of capturing sequences of binary frames with minimal read noise (Ulku et al., 2018), and hence a prime candidate for low light sensing. Despite low read noise, the stochastic nature of photon arrival results in considerable shot noise in single-photon binary frames captured by SPAD cameras. Figure 1.3 includes a

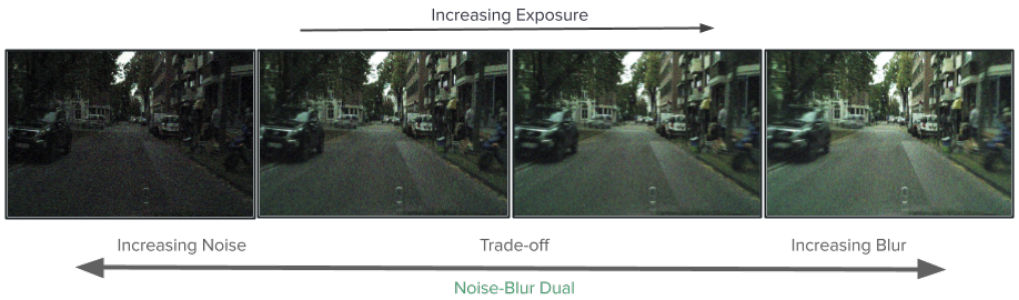


Figure 1.4: **Noise Blur Dual:** Under low light and motion, there exists a tradeoff between noise and blur.

scene captured under low light by a SwissSAPAD2 Ulku et al. (2018) camera. Low bit captures (binary frames in this case) suffer from severe shot noise, making it extremely challenging to do robust inference. Although there has been some recent work on joint denoising and classification (Liu et al., 2020a, 2019; Diamond et al., 2017), inference on ultra-low-light images where each pixel receives less than a photon on average still remains an intractable problem.

1.3 Imaging with conventional CMOS cameras

Conventional RGB cameras have matured over decades and are still the go-to choice for many perception stacks because of many factors such as low cost, low power usage, small form factor, high resolution, etc. Despite all these advantages, their inability to capture high-quality images under challenging imaging conditions like low light, motion, extreme weather, etc., makes them a liability for many perception systems.

Adverse conditions of low light and motion: One challenging condition that is fairly common for cameras is low-light and motion (scene or camera). In such conditions, multiple types of corruption are bound to be present in the image. As discussed earlier, low light could cause the images captured by the camera to exhibit strong noise. While it is possible to mitigate noise by capturing longer

exposures (or larger apertures), this often results in strong motion (or defocus) blur, leading to another kind of image quality degradation. Hence, noise and blur represent "dual corruptions", reducing one (*e.g.*, by adjusting the exposure) necessarily increases the other. Figure 1.4 shows an example of a scene with increasing exposure times, which illustrates this tradeoff between noise and blur. As we can observe, each capture suffers from dual corruption, making it challenging to do inference with any of these captures.

1.4 Thesis

It is my thesis that:

> *We can unlock robust perception under adverse real-world conditions by improving the worst-case performance of the perception models and employing well-designed learning approaches that are tied to its sensing modalities.*

To prove this thesis, I propose a perception framework that is robust under a variety of challenging conditions. We consider numerous sensing modalities like single photon cameras (SPCs), LiDARs and conventional CMOS cameras, under a variety of challenging conditions such as low light, camera or scene motion, and extreme weather.

I start with a SPAD sensor in active imaging mode to demonstrate robust 3D object detection for distant low albedo objects. We introduce a new 3D scene representation called probabilistic point cloud (PPC), which retains the uncertainty information in each LiDAR measurement (Chapter 3). PPCs allow robust 3D object detection under challenging scenarios of low signal-to-background ratio (SBR) in raw LiDAR measurements. Numerous real-world scenarios can lead to low SBR measurements, such as (a) distant objects or low scene albedo, (b) strong ambient illumination and noise, (c) other sources, e.g. non-diffuse and specular materials, multi-path or multi-camera interference, weather phenomena

such as rain, snow, fog, etc. I then demonstrate scene inference using a single photon camera in passive imaging mode under extremely low light conditions. We introduce photon scale space, a collection of varying signal-to-noise ratio (SNR) images with the same scene content, which are used for training perception tasks (Chapter 4). Our proposed training procedure with feature consistency of photon scale images is able to train monocular depth estimation and image classification models that are robust to shot noise present in low light conditions. Finally, I show robust scene inference using conventional cameras under low light and motion conditions. We propose to utilize multiple exposures of the same scene and their ensemble output prediction for scene inference and perception tasks like object detection (Chapter 5).

The main contributions of this dissertation can be summarized as follows:

- A new 3D scene representation called probabilistic point clouds (PPCs) that retains confidence information of each depth measurement from a LiDAR.
- 3D inference approaches designed to utilize probabilistic point clouds that are robust to distant low albedo objects.
- A learning procedure that utilizes photon scaled images, which is a set of varying SNR images of the same scene, to train robust perception models for extremely low light inference.
- A multi exposure ensemble approach for robust 3D object detection under low light and motion, that utilizes multiple captures of varying exposures.

Dissertation Overview: Here is an overview of the dissertation:

- Chapter 2 provides the necessary background on imaging models for the variety of sensing modalities considered in this dissertation. It also discusses the noise models for various challenging conditions considered in our work.
- Chapter 3 details a new 3D scene representation called Probabilistic Point Cloud (PPC), introduces inference approaches to utilize PPCs, and shows its benefits on 3D inference for distant low-albedo objects.
- Chapter 4 introduces photon scale space, a training procedure using feature

consistency and experiments to show its effectiveness under low light.

- Chapter 5 proposes a multi exposure ensemble approach for scene inference, a camera setup for simultaneously capturing with varying exposure times, and its performance under low light and moving conditions.
- Chapter 6 concludes this dissertation, provides limitations of our work, and its future outlook.

2 IMAGING MODEL BACKGROUND

In this section, we discuss the imaging model used for the range of sensors that are considered in our work. We include the background details for both active and passive imaging with SPAD sensors, as well as imaging with conventional CMOS cameras. We also discuss major sources of noise in these sensing modalities and their respective noise models.

2.1 Active Imaging: SPAD LiDAR 3D Sensing Model

A LiDAR imaging system typically consists of a synchronized pulsed laser source and a high-speed time-resolved detector such as a single-photon avalanche diode (SPAD) (Cova et al., 1996). The laser source transmits a pulsed signal $s(t)$, *e.g* a Gaussian pulse with a repetition period of $\mathcal{T}_r$. The scene is illuminated in a raster-scan manner or flood-illuminated to cover the detector's field of view, and the sensor observes the reflected light from the scene. For each pixel location (i, j), the incident photon flux $r_{i,j}$ is modelled (Lindell et al., 2018; Peng et al., 2020) by the following equation:

$$r_{i,j}[n] = \int_{n\Delta t}^{(n+1)\Delta t} \Phi_{i,j} \cdot s(t - \frac{2d_{i,j}}{\mathcal{C}}) dt + b_\gamma, \tag{2.1}$$

where Δt is the time duration of each discrete time-bin, $n \in \{1, 2, ...N\}$ where N is the number of time bins over the duration of a laser period, $\Phi_{i,j}$ is a term that accounts for the distance fall-off, scene reflectance, and Bidirectional Reflectance Distribution Function (BRDF), $s(t)$ denotes the pulsed signal from the laser, $\mathcal{C}$ denotes the speed of light, $d_{i,j}$ is the distance of the scene at the point of illumination (i, j), and b_γ denotes the photon flux from the ambient light.

Suppose the sensor has a quantum efficiency $\eta \in [0, 1)$, which describes the probability that the sensor can detect an incident photon. The detector's dark

count is modeled as b_d, the number of spurious photon detections. Then, the number of photons measured by the sensor at each pixel can also be represented as a 1D timing histogram (raw sensor measurement) $h_{i,j}$ as follows:

$$h_{i,j}[n] \sim \mathcal{P}\{[\eta r_{i,j}[n] + b_d]\}. \tag{2.2}$$

The measurements are modeled as a Poisson process $\mathcal{P}$ with a time-varying arrival function $r_{i,j}[n]$ as the mean rate. Fig. 3.1 shows an example histogram captured under low and high noise. Distance estimate $\hat{d}_{i,j}$ for each location is computed using the bin with the highest photon count (peak location) from each timing histogram[1] and by converting the time of flight to distance using the following equation:

$$\hat{d}_{i,j} = (\Delta t * c/2) * \arg\max_n \; h_{i,j}[n], \tag{2.3}$$

Finally, the distance estimates can be converted to coordinates in a 3D point cloud using camera intrinsic parameters.

2.2 Passive Imaging: Single Photon Imaging Model

Passive imaging is a common operating mode of single photon cameras because of their capability of capturing high-speed binary frames. In this section, we look at the image formation model of a single photon camera and also discuss its noise characteristics.

Image Formulation Model: The number $Z(x, y)$ of photons arriving at pixel (x, y) of a single photon camera during an exposure time of τ seconds is modeled as a Poisson random variable (Yang et al., 2012), whose distribution is given as:

---

[1] Although we focus on SPAD histograms, the proposed analysis and techniques are compatible with other direct ToF time-resolved sensors, including avalanche photodiodes (APDs) and Silicon photomultiplier (SiPMs), which follow a similar peak detection approach.

$$P\{Z = k\} = \frac{(\phi\tau\eta)^k e^{-\phi\tau\eta}}{k!}\ , \tag{2.4}$$

where $\phi(x, y)$ is the photon flux (photons/seconds) incident at (x, y), and $0 \leqslant \eta \leqslant 1$ 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 $Z(x, y) \geqslant 1$; $B(x, y) = 0$ otherwise.[2] Due to the randomness in photon arrival, the binary measurements $B(x, y)$ are also random variables with a Bernoulli distribution:

$$\begin{aligned} P\{B = 0\} &= e^{-(\phi\tau\eta + r_q\tau)}, \\ P\{B = 1\} &= 1 - e^{-(\phi\tau\eta + r_q\tau)} \end{aligned} \tag{2.5}$$

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 an 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 a 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.

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. 1.3 shows an example of a clean image, with the corresponding binary image ($\mathcal{S}^1$). The shot noise in the binary image (Eq. 2.5) 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

[2] After each photon detection, a SPAD pixel experiences a dead time during which it cannot detect any further photons (Rochas, 2003a). For modern SPAD pixels, the dead time is significantly smaller than the exposure time τ, and therefore is not considered in the following analysis.

algorithms (Chi et al., 2020; Ma et al., 2020).

2.3 Imaging with Conventional CMOS cameras

In this section, we discuss the image degradations for conventional CMOS cameras that are observed under low light and motion conditions. We describe the noise and blur degradation model that is considered for our work, then introduce *noise-blur dual* corruption that accounts for the presence of both degradations.

Noise Blur Trade-off in Image Formation: The number of photons incident at a given pixel during the exposure is small under low-light conditions. Because of this, noise becomes dominant in the captured images and has to be properly modeled. In the presence of scene/camera motion, let the photon flux (photons/second) at a pixel p on time t be $\phi_{\mathrm{p,t}}$. The key is to consider that the incident flux at each pixel changes over time t, since the pixel may image different scene points due to scene/camera motion, resulting in an image x with motion blur. Assuming an exposure time $\Delta\mathrm{t}$ and a linear camera with quantum efficiency η, the raw reading at pixel p (without quantization) is given by

$$\mathrm{I_p} = \int_0^{\Delta \mathrm{t}} \phi_{\mathrm{p,t}} \eta \, \mathrm{dt} + z_{\mathrm{p}} \tag{2.6}$$

where z_{p} is the noise at pixel p. Here we ignore the non-uniformity of photon response and noise (Granados et al., 2010), and consider three sources of noise.

- *Shot noise* $z_{\mathrm{p}}^{\mathrm{s}}$ refers to the inherent natural variation of the incident photons due to the Poisson process of photon arrival $\mathcal{P}$ and is modelled as the square root of the signal. Therefore, $z_{\mathrm{p}}^{\mathrm{s}} \sim \mathcal{P}\left(\int_0^{\Delta \mathrm{t}} \phi_{\mathrm{p,t}} \eta \, \mathrm{dt}\right)$.
- *Readout noise* $z_{\mathrm{p}}^{\mathrm{r}}$ comes from the process of quantizing the electronic signal as well as electrical circuit noise, which is modelled as a zero mean Gaussian with variance σ_{r}^2 at each readout. Namely, $z_{\mathrm{p}}^{\mathrm{r}} \sim \mathcal{N}(0, \sigma_{\mathrm{r}}^2)$.
- *Dark current* $z_{\mathrm{p}}^{\mathrm{d}}$ arises due to thermally generated electrons and also follows a square root relationship with signal with a variance of σ_{d}. We thus have

$z_p^d \sim \mathcal{P}(\sigma_d \Delta t)$.

We further assume that z_p^s, z_p^r, and z_p^d are independent of each other, and follow an additive noise model (Hasinoff et al., 2010b), such that $z_p = z_p^s + z_p^r + z_p^d$ (Granados et al., 2010). Thus, $\text{Var}(z_p) = \text{Var}(z_p^s) + \text{Var}(z_p^r) + \text{Var}(z_p^d)$. This leads to the derivation of the signal-to-noise ratio (SNR) for the captured images, given by

$$\text{SNR} = \frac{\left(\int_0^{\Delta t} \phi_{p,t} \eta \, dt\right)^2}{\int_0^{\Delta t} \phi_{p,t} \eta \, dt + \sigma_r^2 + \sigma_d \Delta t}. \tag{2.7}$$

We now discuss the relationship of noise and blur with the exposure time under the presence of both low-light and scene (or camera) motion. The longer exposure time during capture leads to improved SNR, as the noise increases more slowly than the signal. This, however, comes at a cost of increased motion blur in the captured images due to the integral of the incoming flux $\phi_{p,t}$. Hence, the exposure time allows us to trade off noise and blur in the image degradation space, which we term as *Dual Corruption Space*.

Dual Corruption: We define the spectrum of dual-corruption images by varying the camera parameters, resulting in a set $\mathcal{I} = \{x_1, ...x_N\}$ of images with different low-level characteristics (*e.g*, different amounts of blur and noise). For example, varying exposure time Δt creates a sequence of images where noise gradually decreases but the amount of blur increases. An example such sequence is shown in Figure 1.4. Since these images are captured simultaneously (or in rapid succession), we can assume that they have similar semantic content.

An Image without Noise and Blur. A special and theoretically interesting case in the dual corruption space is an *ideal clean image* x_{clean} captured using a very short exposure time ($\Delta t \to 0$) and without noise corruption ($z_p = 0$). Such an image is free of noise and blur. Despite being physically implausible, this construct is sometimes convenient for our derivations.

3 ROBUST 3D OBJECT DETECTION FOR DISTANT LOW-ALBEDO OBJECTS

LiDARs are a prominent 3D imaging modality driving several applications, from embodied perception and autonomous vehicles (Wang et al., 2024; Li and Ibanez-Guzman, 2020), to surveillance (Guo et al., 2024), and more recently, deployed even in consumer devices (e.g., Apple iPhones). LiDARs are based on the time-of-flight (ToF) principle; a typical LiDAR consists of a laser source that emits short sub-nanosecond light pulses into the scene and a sensor that captures the reflected pulses; scene depths are estimated by computing the time delay between emitted and received pulses (Lange, 2000), Fig. 3.1a). Increasingly, single photon avalanche diodes (SPADs) (Cova et al., 1996) are becoming the sensor-of-choice in LiDARs due to their high sensitivity (Pellegrini et al., 2000), and amenability to fabrication of high-resolution arrays (Morimoto et al., 2021). It is not a surprise that the next generation of LiDAR devices is dominated by solid-state, high-resolution, and low-cost SPAD technology.

A typical single-photon LiDAR builds a histogram of photon counts over time (Fig. 3.1b). Under ideal imaging conditions, the peak in the timing histogram can be reliably detected, resulting in an accurate estimation of the time delay and hence the scene depth (Fig. 3.1b). This estimated depth at each pixel can then be used to construct a 3D point cloud-based scene representation, a canonical input to various downstream 3D perception tasks (Li and Ibanez-Guzman, 2020).

However, under several real-world conditions, the raw timing histograms do not have a single, clearly discernible peak (Fig. 3.1b). This could be due to a variety of factors, including a) distant objects or low scene albedo, b) strong ambient illumination which increases the background level and noise, c) other sources *e.g* non-diffuse and specular materials, multi-path or multi-camera interference, weather phenomena such as rain, snow, fog (Pellegrini et al., 2000; Pediredla et al., 2018; Beer et al., 2018; Satat et al., 2018) etc. Imagine a LiDAR mounted on an autonomous vehicle needing to safely navigate the world, not just under fair

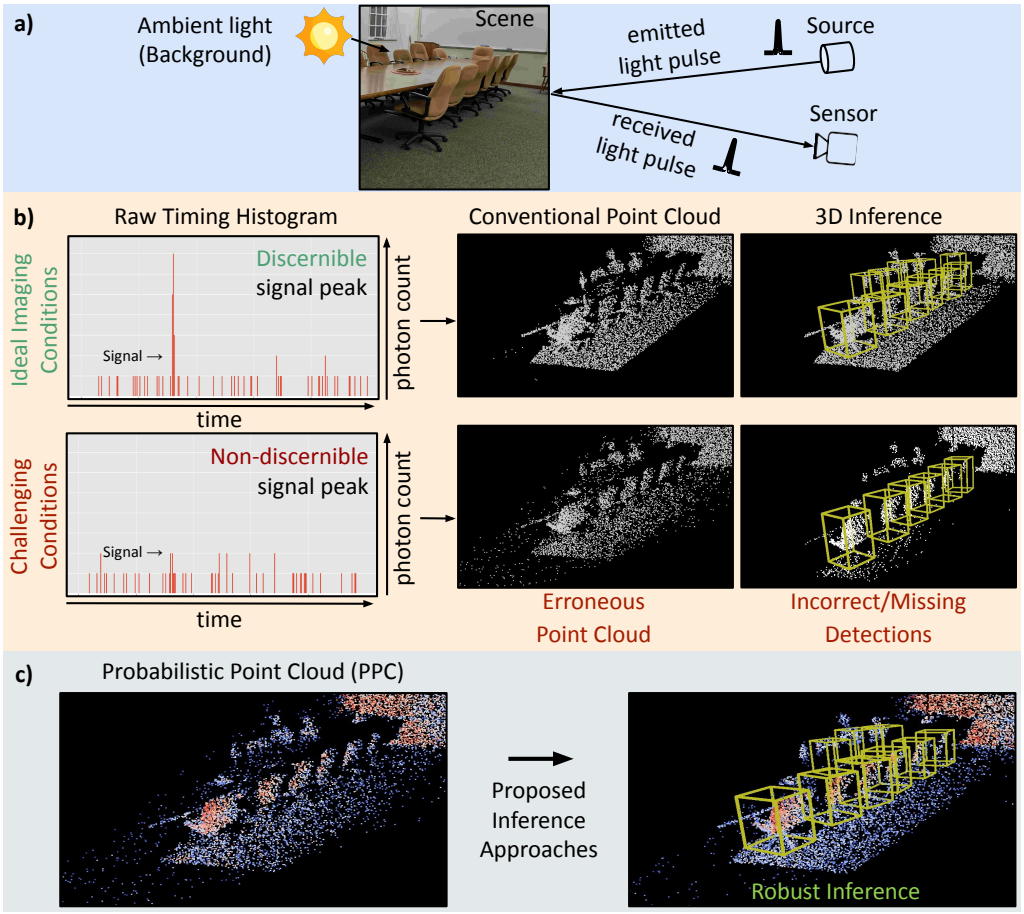


Figure 3.1: **Robust 3D Inference under Challenging Real-World Conditions:** a) A LiDAR measures depth by emitting a pulsed laser and measuring time delays between emitted and received pulses. b) The raw sensor data at each LiDAR pixel is a temporal histogram (photon counts vs. time) where the time location of the peak corresponds to the scene distance, which is then converted to a 3D point in a point cloud. Under challenging conditions, the histogram peaks cannot be easily discerned from the background, resulting in large errors in the point clouds and incorrect inference. c) We propose Probabilistic Point Clouds (PPC), a scene representation with a probability attribute for each point. The figure shows real LiDAR captures, where PPC is visualized with heatmap colors. Leveraging PPC, our proposed approaches achieve robust 3D inference even under challenging scenarios.

lighting and weather, but across all operating conditions. Under these scenarios, it is often challenging to detect the correct peak location, resulting in large depth errors.

This challenge is typically addressed by filtering techniques that retain only the measurements with prominent peaks, while discarding the rest (Zhang et al., 2013; Chen et al., 2020; Goudreault et al., 2023). Therefore, in challenging scenarios mentioned above, most raw histograms that have small or ambiguous peaks either get filtered out resulting in incorrect removal of scene components, or introduce significant noise in the final point cloud if they are retained. Such filtering steps severely affect the inference performance by: (1) removing critical scene content in low signal regimes due to over-aggressive filtering, or (2) propagating excessive measurement noise to downstream inference models[1].

We present a different approach. Our key observation is that raw SPAD histograms encode rich scene information, which traditional LiDAR signal-processing approaches overlook by relying solely on peak locations to estimate point clouds. Instead of *deterministically* removing/retaining depth measurements, we propose augmenting point clouds with meaningful "confidence" features that encode physics-based information about the raw sensor measurements that could be valuable for downstream inference under challenging scenarios. These confidence features require only lightweight compute operations, which is a critical consideration since these operations need to be performed on/close to the sensor where compute and memory resources are extremely scarce.

Confidence. If you have it, you can make anything look good.

Diane Von Furstenberg

[1]This filtering is a pre-processing step before outputting the final point-cloud. Consequently, such noise is not observed in widely-used point cloud datasets as they consist of processed point clouds where low-confidence observations have already been removed, e.g., a distant pedestrian wearing dark clothing. Hence, a considerable amount of useful scene information is irretrievably lost from such point clouds.

We use these confidence measures to create *probabilistic point clouds (PPC)*, a novel 3D scene representation where each scene point is augmented with the confidence (or probability) attribute. Going further, we leverage this PPC representation to design computationally cheap inference approaches that are plug-and-play compatible with a wide range of existing 3D perception models, without needing to alter the architectures. We demonstrate the effectiveness of our approach using widely employed point cloud-based 3D object detection models across both indoor and outdoor scenarios, for both LiDAR as well as camera-LiDAR fusion models. Despite its simplicity, the proposed approach achieves significant performance gains under challenging conditions, such as severe noise encountered in low albedo or distant scenes and intense ambient illumination, outperforming complex point cloud denoising methods.

While the above methods can be inserted as drop-in modules in any 3D inference network, we also explore the integration of the PPC representation directly *within* point cloud based 3D object detectors. Our findings indicate that architectures tailored for PPC can provide further performance gains under challenging scenarios, thereby pointing towards future development of PPC-aware 3D object detectors.

This work takes the first steps toward designing an *end-to-end 3D inference pipeline* that is capable of propagating the uncertainty in depth measurements, starting from raw SPAD data, to downstream 3D inference models. Our contributions can be summarised as:

- Designing lightweight physically meaningful confidence features from raw SPAD histograms.
- Developing a geometric scene representation (PPC) that propagates sensor uncertainty to point clouds.
- Designing computationally low-cost approaches that utilize PPCs for robust 3D inference.
- Demonstrating the performance of PPCs for 3D inference tasks using both simulations and real LiDAR captures, under a variety of challenging scenarios.

3.1 Related Work

3D Inference using Point Clouds: Point-based feature extraction networks (Qi et al., 2017a,b) have become a standard building block for 3D inference (Qi et al., 2019; Shi et al., 2019; Qi et al., 2018; Yang et al., 2020; Yin et al., 2021). These methods largely evaluate using clean point clouds on benchmark datasets, but not on noisy point clouds under challenging conditions; in fact, several works (Qi et al., 2017a; Zhang et al., 2024; Liu et al., 2020b) explicitly mention that network backbones are *not robust* to noise in the point clouds. Although a lot of work has been done for 2D image-based inference under challenging scenarios (Gnanasambandam and Chan, 2020; Goyal and Gupta, 2021; Diamond et al., 2021; Hendrycks and Dietterich, 2019; Goyal et al., 2022), there is surprisingly little prior work on analyzing the 3D inference performance under physically-accurate sensor noise, especially for LiDARs. We show the performance degradation of widely used 3D detection models under real-world noise, and design approaches that perform robustly under such scenarios.

Confidence Attribute in 3D Inference: While methods have been proposed to predict depth uncertainty from a camera image as part of monocular depth estimation (Bae et al., 2022; Xia et al., 2020), there are no prior works that leverage confidence or uncertainty in depth from 3D LiDAR sensor data. This is largely because of limited access to raw sensor data. Our work is a first in designing and propagating such confidence attributes from raw LiDAR sensor data, and accompanying processing techniques, for robust 3D inference.

Point Cloud Denoising: Denoising could potentially be used to reduce the noise in point clouds, and multiple algorithmic (Digne and De Franchis, 2017; Wolff et al., 2016) and learning-based (Ma et al., 2023; Rakotosaona et al., 2020; Luo and Hu, 2021; Hermosilla et al., 2019; Luo and Hu, 2020) solutions have been proposed. These methods often introduce an extra, often significant, computation step to the 3D inference pipeline. Perhaps more critically, these methods do not consider physically realistic sensor noise models and achieve limited success

under challenging conditions with large depth errors, especially for 3D inference. In this work, we evaluate and compare the performance and computational cost of these denoising methods under a wide gamut of challenging conditions.

LiDAR Data Denoising: Emerging high-resolution LiDAR sensor arrays are an exciting new platform for 3D computer vision (Hutchings et al., 2019; Gyongy et al., 2020; Della Rocca et al., 2020; Morimoto et al., 2020; Milanese et al., 2023). It has been shown that directly denoising raw LiDAR timing histograms can increase 3D reconstruction quality (Lindell et al., 2018; Peng et al., 2020; Tachella et al., 2019; Rapp and Goyal, 2017; Lee et al., 2023), albeit at the expense of significant computation. In this work, we bypass the expensive denoising step and perform 3D inference on the lighter-weight (but potentially noisy) PPC representation extracted from the raw LiDAR data.

3D Sensing in Challenging Conditions: Challenging conditions during sensing discussed in Section 1 result in a timing histogram of low signal-to-background ratio (SBR). Fig. 3.1 shows an example histogram where in ideal high SBR conditions, the signal peak is clear, but in challenging low SBR conditions, there are multiple small peaks. Consequently, the estimated depths will be noisy due to the lack of a dominant peak. Another challenge faced by SPADs in ambient illumination is that of photon pileup (Heide et al., 2018; Gupta et al., 2019b) which distorts the histograms. There are approaches that computationally mitigate the structured pileup distortions (Coates, 1968), but end up amplifying the noise in the histograms (Lee et al., 2023), making it challenging to detect low SBR peaks. In that regard, the proposed approaches are complementary to (and can be applied in conjunction with) these pileup mitigation methods.

Noise in Low SBR Point Clouds: Noisy depth measurements under challenging conditions result in point clouds that are sparse or prone to severe noise. Since background peaks are uniformly distributed over the timing bins, the noise points are often arbitrarily far away from the ground truth (GT) depth and along the camera ray axis. Therefore, physically realistic noise under such scenarios is

anisotropic and often more severe than traditional isotropic Gaussian noise often considered in the existing literature.

3.2 Probabilistic Point Clouds

We introduce a confidence measure that can be derived from raw timing histograms using light-weight compute operations. The following equation shows a probability $\Pr(.)$ of a point defined as the ratio of photon detections for the peak bin to the total photon detections in a histogram:

$$\begin{aligned} \Pr(p^{ij}) &= \frac{h_{i,j}[m]}{\sum_{n=1}^{N} h_{i,j}[n]}, \\ \text{where} \quad m &= \arg\max_{n} \; h_{i,j}[n] \end{aligned} \tag{3.1}$$

and p^{ij} is the point corresponding to the sensor pixel (i, j) and N is the number of timing bins in the histogram. We augment points with this probability attribute to create a Probabilistic Point Cloud (PPC). Under ideal conditions and with no background photons, all photons would be detected in a single bin, resulting in a point with a probability of 1. Fig. 3.1c shows a PPC captured under a challenging scenario. Our probability measure is a simple yet effective statistical estimate of the confidence in sensor depth measurements.

Inference with Probabilistic Point Clouds

The point-wise probability attribute in PPC provides vital information for robust inference. Normally, one would expect most spurious points to have a low probability (due to smaller peaks from background photons). However, under challenging scenarios, even ground truth points may have low probability values. Hence, simply filtering low-probability points would remove noise, but also remove true scene points with low signal. To this end, we propose the following approaches to leverage the probability measure of points for robust inference. A

key feature of these approaches is that they do not require significant modifications to the inference model or its training procedures.

Neighbor Probability Density (NPD) Filtering

Our key observation is that most spurious points in a point cloud under challenging conditions have low probability and/or low spatial density. In contrast, points that belong to true scene objects typically have a high local spatial density of points due to neighboring points being on the same surface or object. To leverage this, we compute a score called the Neighbor Probability Density (NPD) score for each point, which encapsulates both the spatial density and the average probability of its neighbors as follows:

$$\mathrm{NPD}(p_i) = \Sigma_{p_j \in \mathcal{BQ}_{L,r}(p_i)} \mathrm{Pr}(p_j)/L, \tag{3.2}$$

where $\mathrm{Pr}(\cdot)$ is the probability of a point and $\mathcal{BQ}_{L,r}(\cdot)$ returns up to L points that are in the local neighborhood ball of radius r around a point. We aggregate the probability of these neighbors and normalize it with L to get the final score. Since the ball query returns up to L neighbors, the score for spatially dense points with $>$ L neighbors is the average probability of its L neighbors in the local neighborhood ball, whereas sparse points with $<$ L neighbors get penalized with a lower score, as we normalize the score with L which is greater than its number of neighbors.

Fig. 3.2 shows the distribution of NPD scores of all points in scenes under different SBR levels. A large peak of noise points (red) on the left of each plot has much lower NPD scores than the GT points (green). The points below a certain NPD score (α) can be filtered out without removing many GT points. As SBR decreases, GT points also have lower NPD scores, as expected, but the separation between the peak of noise points and the rest remains. Another smaller peak towards the right also contains some noise points with higher NPD scores. NPD filtering cannot remove these points as it would also remove many GT points.

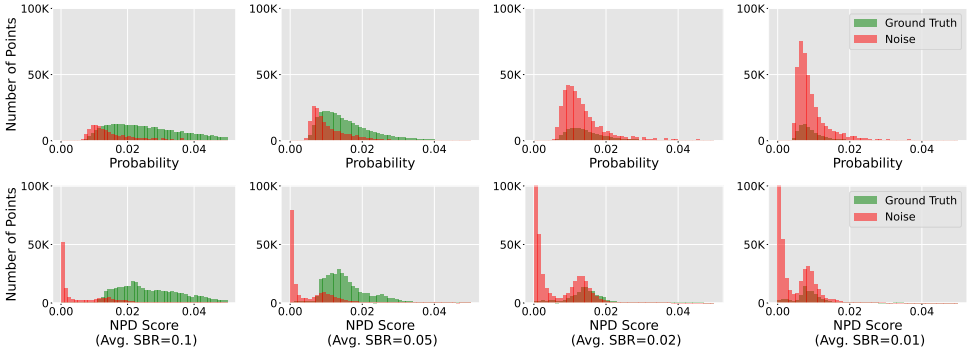


Figure 3.2: **Distribution of Point Probability and NPD Score:** Point probabilities (top row) and NPD scores (bottom row) of all points in scenes under different SBR levels. NPD Scores for the noise (red) are lower as they have lower spatial density than ground truth points (green). We use an NPD threshold to filter out a large number of noise points from our point cloud. Point probability (proportional to the number of photon detections) alone is not a good indicator for separating noise from GT as many GT points also have low probability.

NPD score is a simple but effective way to filter out noise points as it leverages information from multiple sensor measurements by considering neighboring points, whereas filtering approaches based on low photon counts only rely on the timing histogram of the same pixel. Note that we do not make any assumptions regarding the object/scene surface, and the score is easy to calculate without much computational overhead.

Farthest Probable Point Sampling (FPPS)

A key component in many point cloud inference models (*e.g*, PointNet++ (Qi et al., 2017b) and Point Transformer (Zhao et al., 2021)) is the Farthest Point Sampling (FPS) which is used for keypoint sampling. FPS samples the points that are farthest from each other. This ensures that sampled keypoints cover all scene regions, including objects with sparse or few points.

However, FPS is not robust to strong noise as it prioritizes points that are distant from each other. Since noise points can be spread out far from object

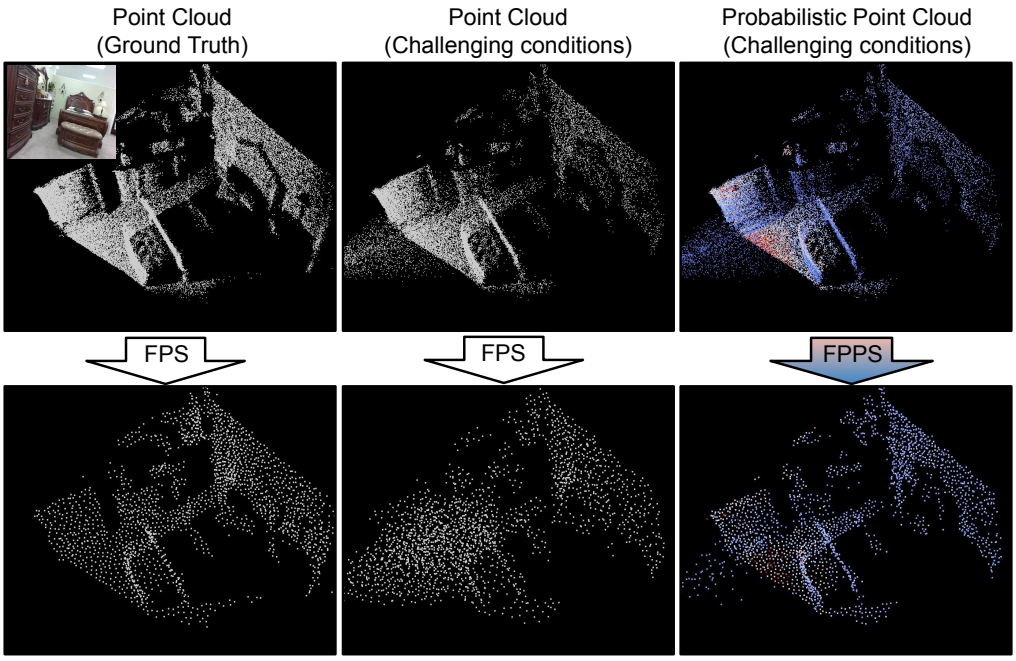


Figure 3.3: **Farthest Probable Point Sampling (FPPS):** Sampled points using Farthest Point Sampling (FPS) are not robust to the noise present under challenging conditions. FPS samples a large number of noise points as it prioritizes farther points (column 2). We propose FPPS which utilizes the point probability to build a candidate set of high-probability points for sampling which ensures most sampled points are on the object surface while still covering the entire scene (column 3).

surfaces, this results in a large number of noise points being sampled and thus a significant drop in performance. Fig. 3.3 shows a point cloud of a scene under both ideal (column 1) and challenging (column 2) conditions along with its sampled keypoints using FPS. While FPS is effective in high SBR scenarios and covers most surfaces and objects, it suffers in challenging scenarios by sampling a lot of noise.

To address this issue, we propose Farthest *Probable* Point Sampling (FPPS), which only considers high-probability points as candidates for sampling. We build a candidate set of points above a certain probability value (β) and perform FPS on this candidate set. This ensures that the sampled centers have fewer spurious

points and more surface or ground truth points, which results in more effective inference output (Fig. 3.3).

We should note that the low-probability points are still part of the point cloud and are included in the rest of the network operations like feature aggregation. This allows the network backbone to still utilize low-probability points for feature extraction if they are in the neighborhood of a sampled keypoint. FPPS is not needed for the network backbones that do not include a keypoint sampling operation.

3.3 3D Object Detection Results

Datasets: We evaluate our approach on 3D object detection benchmarks of SUN RGB-D (Song et al., 2015) and KITTI (Geiger et al., 2012). SUN RGB-D consists of ~10K RGB-D scans of *indoor* scenes annotated with 3D bounding boxes. We use the standard evaluation protocol that considers 10 common object categories. KITTI dataset is a widely used *outdoor* autonomous driving dataset containing ~7.4k annotated scenes with LiDAR point clouds. We follow the standard evaluation protocol using three categories: car, pedestrian, and cyclist.

Point Clouds under Challenging Conditions: We use ground truth depth maps to simulate a physically realistic photon timing histogram for each pixel using the simulation model and code provided by (Lindell et al., 2018). Each histogram has 1024 bins and a temporal bin width of 97ps. Our pulsed signal has a repetition period of 100ns and the detected illumination pulse has a full-width half maximum of ~350ps. We simulate histograms with various levels of mean Signal to Background Ratio (SBR) for the scene to cover a variety of scenarios. Our benchmark consists of the following SBR levels {0.1 (5-50), 0.05 (5-100), 0.02 (1-50), 0.01 (1-100), *Clean*} where *Clean* denotes the ground truth point cloud.

Implementation: We implement our method using the MMDetection3D framework (Contributors, 2020) provided by OpenMMLab. For the SUN RGB-D

dataset, we evaluate using VoteNet (Qi et al., 2019) architecture, which is a LiDAR-based 3D object detector and uses PointNet++ (Qi et al., 2017b) as the point processing network backbone. For the KITTI dataset, we use PV-RCNN (Shi et al., 2020) architecture, which is a LiDAR-based 3D object detector and uses 3D Voxel CNN with sparse convolutions (Graham et al., 2018) as a backbone. Our NPD filtering step uses $L(= 64)$ neighbors within radius $r(= 0.2)$ to calculate the NPD score. We use $\alpha = 0.003$ and $\beta = 0.01$ as hyperparameters for NPD Filtering and FPPS respectively for our model. Please refer to the supplement for ablation studies on hyperparameters.

Baselines: We compare our approach with the following set of baselines. All methods use the same 3D detection model architecture and backbone for a fair comparison.

- *Matched Filtering* (Turin, 1960): This method uses matched filtering output of the timing histograms. We convolve the histograms with the signal pulse in Eq. 2.3 before calculating the depth estimate. This provides a strong baseline for temporal denoising.
- *Thresholding:* This method uses a thresholding approach where depth estimates corresponding to small bin values are ignored. This removes a large number of spurious points from the point cloud. We select the optimal value of the threshold for our evaluation.
- *PointClean Net* (Rakotosaona et al., 2020): This is a point cloud denoising network that uses a combination of outlier removal and denoising steps.
- *Score-based Denoising* (Luo and Hu, 2021): This is a state-of-the-art point cloud denoising approach, which denoises each point in the point cloud by updating them to their estimated local surface based on a calculated score.
- *PathNet* (Wei et al., 2024): This is a point cloud denoising method based on reinforcement learning using a noise and geometry-based reward function.

We use a matched filtering step for all methods (including PPC) for temporal denoising of the histograms. We retrain all denoising networks (PointClean

Avg. SBR	Clean		0.1		0.05		0.02		0.01	
	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50
Matched Filtering	51.34	27.45	42.43	20.49	38.77	17.57	16.95	5.05	11.34	2.73
Thresholding	57.11	33.21	51.27	28.62	46.44	24.86	29.58	14.81	16.47	6.45
PointClean Net	54.58	31.89	45.65	26.44	40.19	19.15	18.24	8.05	12.78	3.01
Score Denoising	57.38	34.02	53.19	29.45	48.61	25.78	26.35	13.73	14.55	4.73
PathNet	57.16	33.87	52.16	28.79	47.11	24.89	25.45	12.96	13.87	4.56
PPC (Ours)	**58.61**	**34.99**	**54.29**	**31.15**	**52.46**	**30.20**	**38.49**	**16.47**	**29.42**	**13.16**

Table 3.1: **3D Object Detection Comparison**: Table shows AP@0.25 and AP@0.50 results on the SUN RGB-D dataset using VoteNet architecture. Our approach outperforms all baselines and shows large gains under very low SBR conditions.

Net, PathNet, and Score-denoising) using the SUN RGB-D dataset. Denoised point clouds are used for training and testing 3D detection models for these baselines. Other approaches for timing histogram denoising are also discussed in the supplement.

Benchmark Results: For every method, we train a joint model using all SBR levels and evaluate the performance at each SBR level. Table 3.1 shows AP@25 and AP@50 comparisons on the SUN RGB-D dataset with VoteNet. Matched Filtering and Thresholding suffer due to a large amount of noise. PointClean Net, PathNet, and Score-denosing show improvement for higher SBR levels, but are not as effective under low SBR conditions. Our approach performs significantly better even under extremely low SBR conditions. Please see supplementary text for per-category results. Table 3.2 shows mAP on moderate difficulty of KITTI val split calculated with 11 recall positions for PV-RCNN architecture (Shi et al., 2020). Our approach shows significant gains for pedestrian and cyclist categories under low SBR conditions.

Observations: Fig. 3.4 and 3.5 visualize results on the SUN RGB-D and KITTI

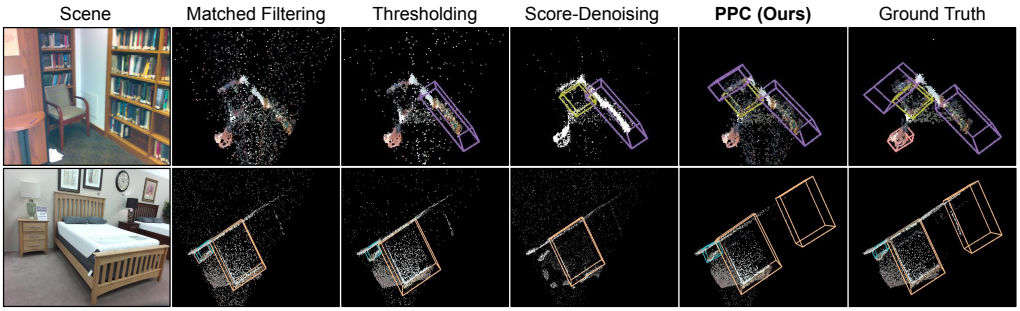


Figure 3.4: **3D Object Detection Visualization:** Figure shows results from the SUN RGB-D dataset using VoteNet under low SBR (0.05) conditions. Matched Filtering struggles with noise and misses a lot of objects. Thresholding misses smaller (chair) and farther (bookshelf) objects. Score-Denoising removes noise closer to the surface but still misses a few objects. PPC (Ours) outperforms all baselines.

Avg. SBR	Clean			0.05			0.02			0.01			0.005		
	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc
Matched Filtering	82.48	60.11	71.36	**73.14**	55.76	61.84	68.17	50.03	52.85	59.95	47.06	43.74	50.68	37.01	35.01
Thresholding	82.81	58.63	71.55	72.80	57.72	60.44	68.05	54.80	52.71	59.40	49.23	44.96	50.35	38.62	35.74
PPC (Ours)	**83.56**	**60.62**	**73.35**	73.03	**59.12**	**64.14**	**68.42**	**59.04**	**53.18**	**60.29**	**55.39**	**47.76**	**51.30**	**49.51**	**36.44**

Table 3.2: **KITTI 3D Detection Comparison**: Table shows mAP for car, pedestrian, and cyclist category on moderate difficulty of KITTI val split calculated with 11 recall positions for PV-RCNN architecture. Our method shows significant gains under low SBR conditions.

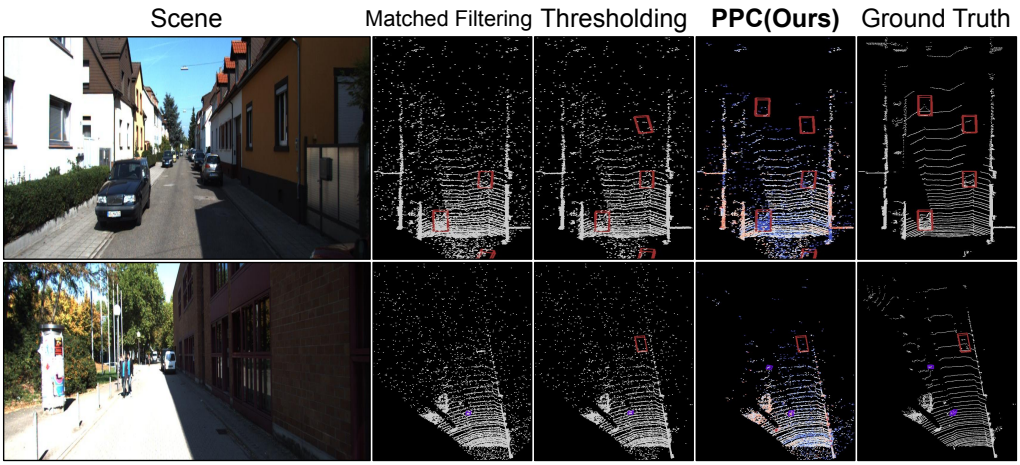


Figure 3.5: **3D object detection visualization** on the KITTI dataset under low SBR (0.02) conditions. Baselines miss small and farther objects: car and pedestrian (zoom-in). PPC detects most objects.

datasets respectively using LiDAR-only 3D detectors. Matched filtering misses many objects due to strong noise. Thresholding removes a large number of noise points but also removes ground truth points that have low photon detections. Thus, it performs better on larger and closer objects as they include points with high incident photon flux but misses farther objects like nightstands and chairs. Score-based Denoising is able to denoise points closer to a surface. However, the recognition is still affected by noise points that are far from the surface as they are not denoised effectively. Our approach can detect smaller and farther objects more consistently. Color information is only used for visualization and not for inference. Please refer to the supplementary material for more visualization results.

Generalization: We also evaluate using ImVoteNet (Qi et al., 2020) (fusion of camera and LiDAR), Uni3DETR (Wang et al., 2023) (LiDAR only with Transformer-based architecture), and PointPillars (Lang et al., 2019) (LiDAR only with pillar-based representation) in the supplementary section which shows the effectiveness of our method on a wide range of 3D detectors.

	Matched Filtering	Thresh-olding	PointClean Net	Score-Denoising	PathNet	PPC (Ours)
Runtime (ms)	87	89	755	1345	867	95

Table 3.3: **Comparison of Inference Time:** Our method adds no significant computational cost while outperforming all baselines.

Should we use denoising networks for point cloud inference? Current denoising networks are designed for surface reconstruction but are not very effective for downstream inference tasks like 3D object detection. This is because, under challenging real-world conditions considered in this work, noise is usually more severe due to spurious points with large depth errors. Further, these denoising networks add a significant computational overhead, whereas our method performs inference robustly without the extra computational cost of denoising. Table 3.3 shows per scene runtime for each method. This suggests that training a robust inference model to handle noise is more beneficial than denoising point clouds under challenging conditions.

Modeling with Probabilistic Points

Moving forward, we explore the integration of the PPC representation directly into the point cloud models. To this end, we focus on VoteNet (Qi et al., 2019) — a classic model built on PointNet++ (Qi et al., 2017b) for 3D object detection. VoteNet processes a point cloud by taking points and their attributes as input, leveraging PointNet++ to extract point-wise features. These features are passed through a voting module that groups the point cloud into local clusters. Each cluster generates 3D boxes as the object proposal. The proposals are then classified to produce the final detection results.

This design allows integrating point probability at various stages of the model, including the input, point-wise features, and object proposals. We explore these options through the experiments detailed below. While alternative approaches to integrate probability into point cloud networks may exist, our goal is to delineate

Avg. SBR	Clean	0.1	0.05	0.02	0.01
PPC (Baseline)	58.61	54.29	52.46	38.49	29.42
PPC + A	58.35	55.73	53.67	39.10	30.58
PPC + B	59.29	55.86	52.95	37.68	30.42
PPC + C	58.53	55.18	53.15	38.76	30.17
PPC + A+B+C	**59.35**	**56.11**	**53.45**	**39.87**	**30.81**

Table 3.4: AP@25 results on SUN RGB-D with PPC variants.

evident design choices to build PPC-aware 3D inference models.

- *Probability as a point attribute (A):* Point probability can be treated as an attribute for the point and used as an input for the network. This design tasks the network to learn effective features from the probability.
- *Probability weighted point feature vectors (B):* Point-wise features from PointNet++ can be weighted by the average neighborhood probability. This allows the network to prioritize features from high-probability points.
- *Probability weighted objectness scores (C):* The objectness score for a proposal can be weighted by the average probability of points within the corresponding 3D box. This allows the network to assign higher objectness scores for proposals with high confidence points.

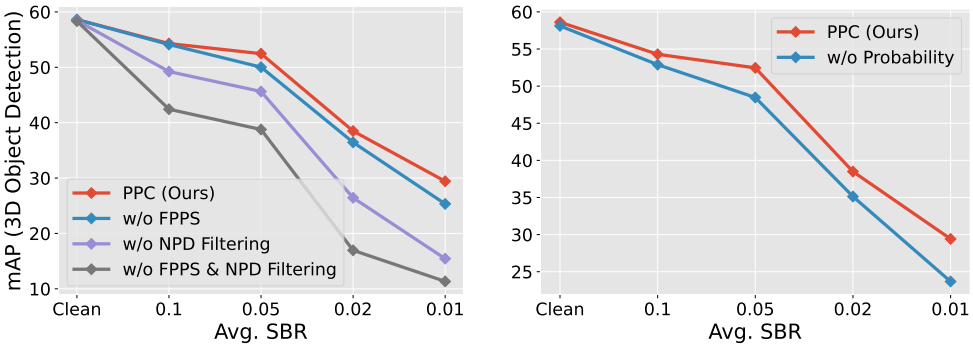
Results and Discussion: Table 3.4 presents the AP@25 results on the SUN RGB-D dataset. By employing individual options, the model attains a further gain of 1-2%. Using all three options leads to a modest boost of ~2%. We note that this boost is on top of our already strong baseline. Our results show the potential of integrating PPC with deep models and further call for innovative approaches in this area.

3.4 Ablation Studies

In this section, we include the ablation studies of our proposed method. We evaluate various design choices on the SUN RGB-D benchmark using the VoteNet

architecture.

First, we analyze the effectiveness of both NPD Filtering and FPPS individually. Fig. 3.6a shows the results of our approach without NPD and FPPS. NPD filtering shows a significant gain in performance, especially under very low SBR scenarios. FPPS shows an additional 2-4% improvement in mAP under very low SBR conditions.



(a) (b)

Figure 3.6: **Ablation Study of PPC Components:** Performance of our approach (a) without FPPS and NPD Filtering, and (b) without probability attribute.

Second, we show the performance of our method on point clouds without the probability attribute in Fig. 3.6b. This is equivalent to using our approach with a conventional point cloud (*i.e*, all points with probability 1). The probability attribute accounts for about a 4-5% gain in mAP performance and is significant in very low SBR conditions.

We also show the performance of our approach by varying the hyperparameters of NPD filtering (α) and FPPS (β). Fig. 3.7a and 3.7b show the mAP on the complete SUN RGB-D test set of all SBR levels. We chose the best performing value of $\alpha = 0.003$ and $\beta = 0.01$ for our models.

We also analyze our method by varying the hyperparameters of NPD score calculation, i.e. max ball neighbors (L) and ball radius (r). Fig. 3.7c and 3.7d show the mAP on the complete SUN RGB-D test set of all SBR levels. We find an optimal NPD score value (α) for each experiment. Increasing the radius too much starts to hurt the performance as noisy sparse points have more neighbors if

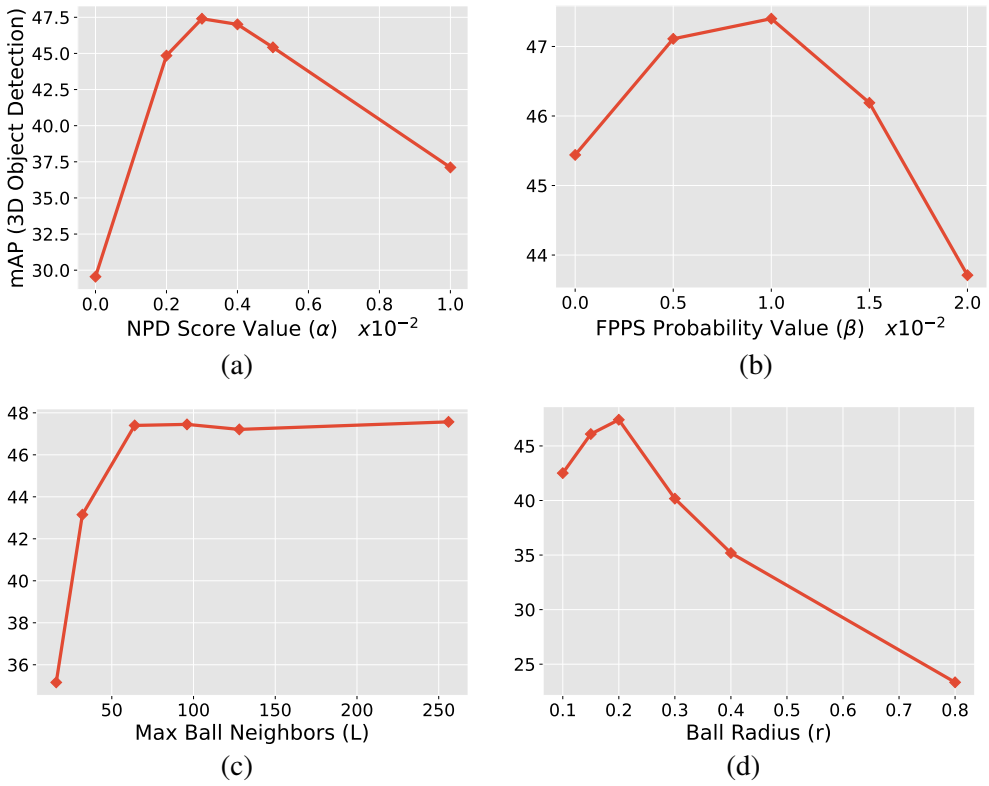


Figure 3.7: **Ablation Study of PPC Hyperaparametes:** Performance of our approach with varying (a) NPD Score Value, (b) FPPS Value, (c) Max Ball Neighbors, and (d) Ball Radius.

the ball radius is bigger. Performance improves as the value of L increases but saturates around 64. We chose the values of r = 0.2 and L = 64 for our models.

Table 3.5 analyzes the total per-scene runtime of our method on the SUN RGB-D dataset. We use a single RTX2070 Super GPU machine for inference time calculation. Adding FPPS adds no computational overhead, whereas adding NPD filtering adds less than 8% of runtime with our implementation.

	Inference Time (ms)
PPC	95
PPC w/o FPPS	95
PPC w/o NPD	88
PPC w/o FPPS & NPD	88

Table 3.5: **Ablation Study of Runtime:** Our method adds no significant computational cost to runtime.

We analyze the performance of the Thresholding baseline by varying the threshold used for the model. Fig. 3.8 shows AP@25 results on the complete SUN RGB-D test set of all SBR levels. We select the best-performing threshold (=1.1) for evaluating this baseline.

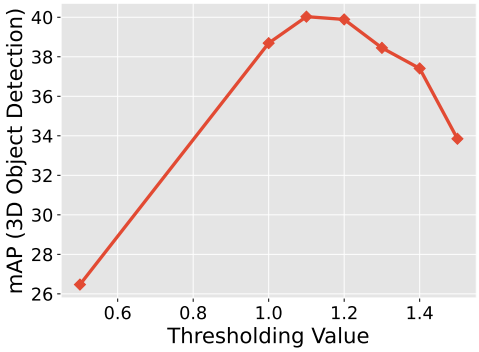


Figure 3.8: **Ablation Study of Thresholding baseline:** Performance of the Thresholding baseline with varying threshold used for the model.

3.5 Experiments with Real Hardware

Finally, we demonstrate our approach using real PPCs captured by our prototype Single-photon LiDAR systems.

Indoor Capture Setup: We use a HORIBA FLIMera (Henderson et al., 2019) SPAD camera (Fig. 3.9) since it allows access to raw timing histograms for each pixel. The camera consists of a 192x128 pixel SPAD array; each pixel has a

Figure 3.9: **Camera Setup:** Figure shows our complete LiDAR setups with FLIMera (left) and Adaps (right) sensors.

quantized 12-bit (4096 bins) time axis, with each bin having a width of 41.1ps. We use a flash illumination setup using a diffuser to illuminate the field of view of the sensor. We consider several indoor scenes (*e.g*, conference rooms, lecture halls, and living rooms) under ambient light ranging from 200-800 lux under varying exposure times from 0.1s to 1s to simulate various signal levels.

Outdoor Capture Setup: For outdoor captures, we use Adaps ADS6311 (Adaps, 2024) sensor which is a commercial medium-range Single-photon LiDAR. It has a spatial resolution of 256x192 pixels and allows raw temporal histogram read-out for each pixel. Each pixel has 672 temporal bins with a bin width of 297ps. We consider outdoor scenes (*e.g*, parking lots, traffic stops, and busy roads) for captures. Please refer to the supplement for more details on both setups.

3D Object Detection Results: Fig. 3.10 and Fig. 3.11 show the results of our approach compared with the baselines using real captures for indoor and outdoor scenes, respectively. PPC can detect most objects with accurate bounding boxes. More results on real captures are in the supplement.

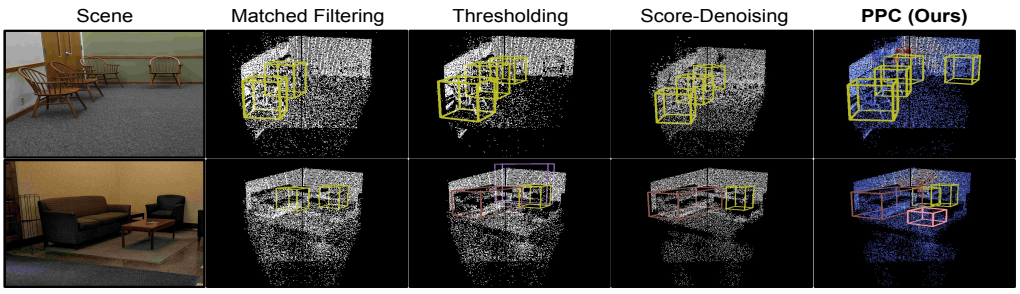


Figure 3.10: **3D Detection Results on Indoor Real Captures:** Figure shows scenes under challenging conditions captured by our FLIMera LiDAR system. Baselines fail to detect many small distant objects. PPC detects all objects (*e.g* chair, table and couch).

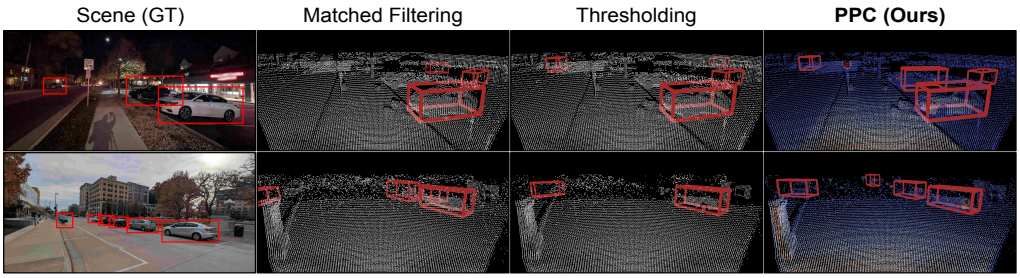


Figure 3.11: **3D Detection Results on Outdoor Real Captures:** Figure shows scenes under challenging conditions captured by our Adaps LiDAR system. Baselines fail to detect distant cars, whereas PPC detects farther cars with accurate bounding boxes.

3.6 Conclusion

In this work, we demonstrate numerous real-world challenging scenarios of 3D object detection using LiDAR point clouds, such as long-distance or low-albedo objects. These conditions produce sparse or erroneous point clouds, resulting in severe loss of accuracy. To mitigate this, we propose a novel 3D scene representation called Probabilistic Point Cloud (PPC), where each point is augmented with a probability attribute that encapsulates the measurement uncertainty (confidence) in raw data. We further introduce inference approaches that leverage PPC for robust 3D object detection; these methods are versatile

and can be used as computationally lightweight drop-in modules in 3D inference pipelines. We demonstrate, via both simulations and real captures, that PPC-based 3D inference methods outperform several baselines with LiDAR as well as camera-LiDAR fusion models, across challenging indoor and outdoor scenarios involving small, distant, and low-albedo objects, as well as strong ambient light.

3.7 Supplementary Section: Recognition with Real Captures

In this section, we provide further details about our LiDAR setups and more results on real PPC captures.

Camera Setups

We use a LiDAR sensor with an external laser for our indoor captures. This allows us to control various camera and scene parameters (*e.g*, exposure time, laser power, and ambient illumination) over a wide range. We use a commercial LiDAR sensor for our outdoor captures. This allows us to have a portable low-power LiDAR setup for outdoor environments. Here are the details of both setups:

Indoor Camera Setup: Our indoor setup uses a SPAD-LiDAR sensor with an external class 4 laser. Fig. 3.12 shows the front view of our setup with the HORIBA FLIMera (Henderson et al., 2019) camera. The temporal resolution of the camera is about 380ps, which is in line with the full-width at half-maximum (FWHM) of the instrument response function (IRF) of the device. We set up our camera system with the Katana laser (Katana, 2024) which is a high-powered pulsed picosecond laser system by OneFive. The laser has a wavelength of 532nm (green). We operate the laser under low power settings (ranging from 50-100 mW). We use a flash illumination setup with a diffuser to illuminate the field of view of the sensor. We use a 3.8mm focal length lens for a wider field of view of the scene. The laser system, as well as the FLIMera sensor, are connected to

Figure 3.12: **Indoor Camera Setup:** The Figure shows the front view of our FLIMera camera setup (left) and the sensor (right).

Figure 3.13: **Outdoor Camera Setup:** The Figure shows the front view of our Adaps camera setup (left) and the sensor (right).

an external computer to receive the synchronization signal and trigger for the capture.

Outdoor Camera Setup: Our outdoor camera setup uses a commercial LiDAR

sensor which is a more portable camera and uses low input power. Fig. 3.13 shows the front view of the setup with Adaps (Adaps, 2024) camera. The camera is rated for an accuracy of more than 5cm up to a range of 30m. The setup has a wide FOV (120°horizontal and 90°vertical). We also vary the exposure time from 0.1s to 1s to simulate various signal levels. The camera can operate at a very low power input (<10W) and is connected to a small portable AC power source. The camera also saves low-resolution (256x192) grayscale images, which are used for visualizations only. We also mount a smartphone camera in our setup to simultaneously capture high-res RGB images, used for visualizations only for some static scenes shown in the main paper and supplementary report.

3D Object Detection Results

Fig. 3.14 and 3.15 show a comparison of our approach with the baselines using real indoor and outdoor captures. Matched Filtering baseline suffers from noise and often detects false positives. Thresholding frequently misses small or farther objects in the scene. Baselines struggle with farther chairs in indoor captures and farther cars and pedestrians in outdoor captures. Our approach detects most objects with tight bounding boxes.

3.8 Supplementary Section: Additional 3D Object Detection Results

In this section, we include additional results and analysis that supplement the experiments in the main paper.

Implementation Details

We implement our method using the MMDetection3D framework (Contributors, 2020) provided by OpenMMLab and use the same evaluation procedure as the previous literature. For the SUN RGB-D dataset, the color information from the

Figure 3.14: **3D Detection Results on Real Indoor Captures:** Figure compares our method with the baselines under challenging low SBR conditions. Baselines fail to detect many small and farther objects, *e.g* chairs in the back of many scenes. PPC detects most objects (*e.g* chair, table, and couch) with tight bounding boxes.

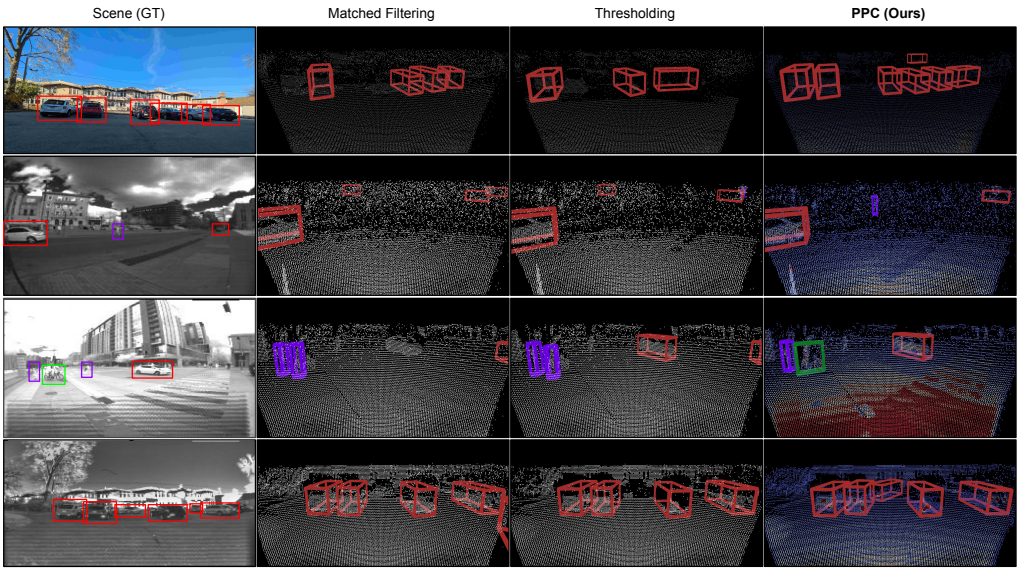


Figure 3.15: **3D Detection Results on Real Outdoor Captures:** Figure compares our method with the baselines under very challenging low SBR conditions. The first scene includes 6 cars and baselines fail to detect farther cars. The second scene shows 2 cars and a pedestrian, and baselines struggle to detect the distant pedestrian. The third scene includes a car, two pedestrians, and a cyclist. The ground truth objects are marked in the camera scene images for easier visualization. PPC detects most objects (*e.g* cars, pedestrians, and cyclists) in all scenes with accurate bounding boxes.

point clouds is not used for inference and is only used for visualization. For the KITTI dataset, the reflection intensity information from LiDAR point clouds is used as input for all methods.

Category-wise Performance

Table 3.6 shows per-category AP@25 results on the SUN RGB-D dataset for all methods under low SBR (0.02) conditions. Our approach shows significant gains for all categories, particularly larger gains for *small-sized* object categories (*e.g* chair, nightstand, dresser) which typically suffer the most under low SBR conditions. Table 3.7 shows mAP on the KITTI val split for PV-RCNN architecture

Category	Matched Filtering	Thresh-olding	PointClean Net	Score Denoise	PathNet	**PPC (Ours)**
Bed	54.37	67.33	53.59	68.57	67.53	**72.97**
Sofa	16.20	28.15	22.00	38.25	38.03	**45.19**
Table	29.61	37.99	30.20	33.93	32.17	**40.47**
Bathtub	6.04	25.14	2.71	14.04	13.46	**54.32**
Desk	7.97	14.16	7.78	8.33	9.18	**17.37**
Bookshelf	2.12	4.85	0.67	1.01	0.89	**9.80**
Chair	22.05	34.27	22.92	27.45	25.43	**47.47**
Night Stand	3.10	14.45	7.41	10.23	9.13	**30.49**
Dresser	2.58	4.64	2.80	2.44	2.09	**13.73**

Table 3.6: **Category-wise 3D Object Detection Results**: Table shows per category AP@25 results on SUN RGB-D dataset under low SBR (0.02) conditions. Our approach outperforms all baselines and shows large gains for smaller object categories (below the line) like chairs and nightstands.

Avg. SBR	Car			Pedestrian			Cyclist		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Matched Filtering	79.02	59.95	57.67	50.85	47.06	43.51	68.78	43.74	41.10
Thresholding	78.73	59.40	55.35	54.45	49.23	45.38	68.13	44.96	42.64
PPC (Ours)	**79.10**	**60.29**	**59.08**	**60.42**	**55.39**	**50.82**	**71.99**	**47.76**	**44.84**

Table 3.7: **KITTI 3D Detection Comparison**: Table shows mAP for easy, moderate, and hard difficulty levels on KITTI val split calculated with 11 recall positions for PV-RCNN architecture under low SBR (0.01) conditions.

under low SBR (0.01) conditions calculated with 11 recall positions in a standard format similar to previous works. PPC outperforms the baselines in all categories.

Visualizations and Observations

Fig. 3.16 to 3.19 show visualizations of 3D object detection on the SUN RGB-D dataset for all methods under different SBR conditions using VoteNet architecture. Fig. 3.16 and 3.17 show complex scenes with a large number of small objects (*e.g* chairs). Baselines fail to detect a lot of small and farther objects (last row of chairs) whereas PPC detects the most number of objects accurately. Fig. 3.18

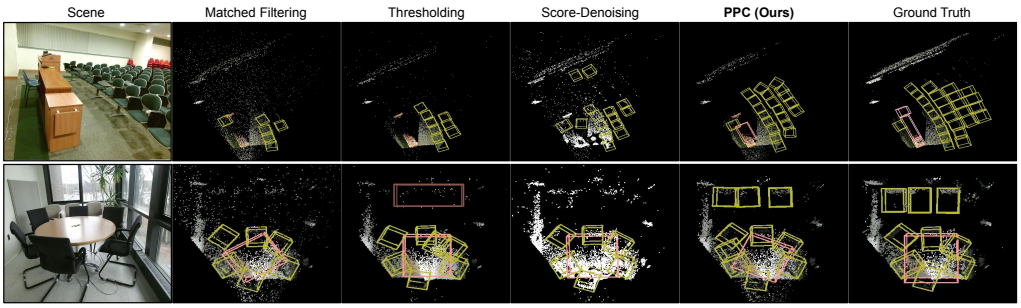


Figure 3.16: **3D Object Detection Results:** Figure shows scenes from the SUN RGB-D dataset under medium SBR (0.1) conditions. The first scene contains multiple rows of small objects (chair). Baselines fail to detect farther rows of chairs. Our approach detects most chairs and table.

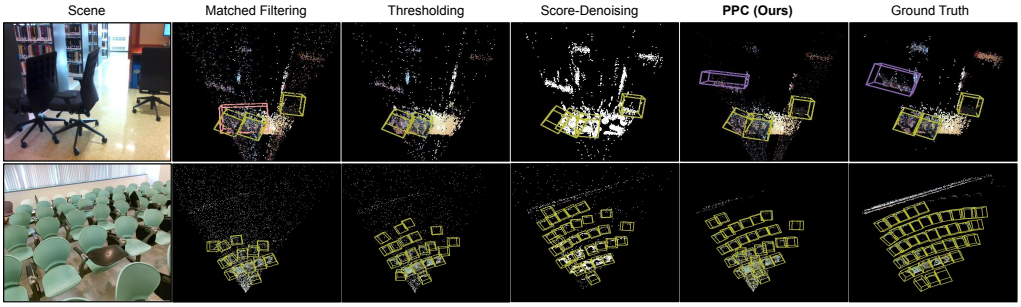


Figure 3.17: **3D Object Detection Results:** Figure shows scenes from SUN RGB-D dataset under medium SBR (0.05) conditions. Scenes consist of small (chair) and farther objects (bookshelf). Our approach detects most objects with no false detections.

and 3.19 show scenes with very low SBR conditions. Baselines fail to detect many objects whereas our approach performs significantly better even under the presence of a large amount of noise. Fig. 3.20 shows a few failure cases for our method. PPC can sometimes detect multiple overlapping bounding boxes for the same object under noise. Couch or single-sitter couches are often detected as chairs by PPC or other baselines. Fig. 3.21 to 3.24 show scenes from the KITTI val dataset under varying SBR conditions using PV-RCNN architecture. Baselines fail to detect many objects like farther cars and pedestrians, whereas our approach

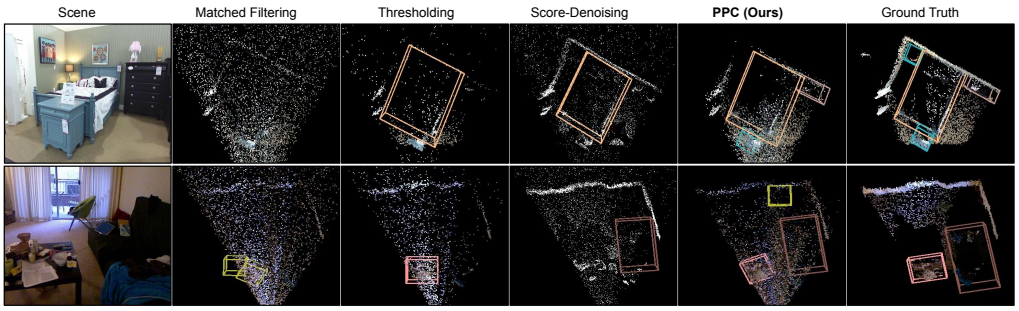


Figure 3.18: **3D Object Detection Results:** Figure shows scenes from the SUN RGB-D dataset under low SBR (0.02) conditions. Scenes consist of small (nightstand) and occluded objects (table). Our approach performs better than all baselines.

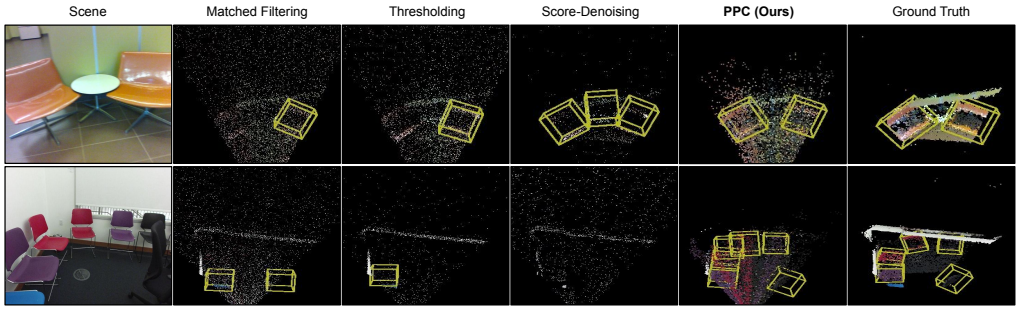


Figure 3.19: **3D Object Detection Results:** The figure shows scenes from the SUN RGB-D dataset under low SBR (0.01) conditions. Baselines fail to detect numerous objects (chair) due to noise whereas our approach detects most objects in the scene.

detects most objects.

3.9 Supplementary Section: More 3D Detection Architectures

In this section, we evaluate our PPC approach using a variety of 3D object detection architectures. First, we evaluate a camera-LiDAR fusion approach ImVoteNet

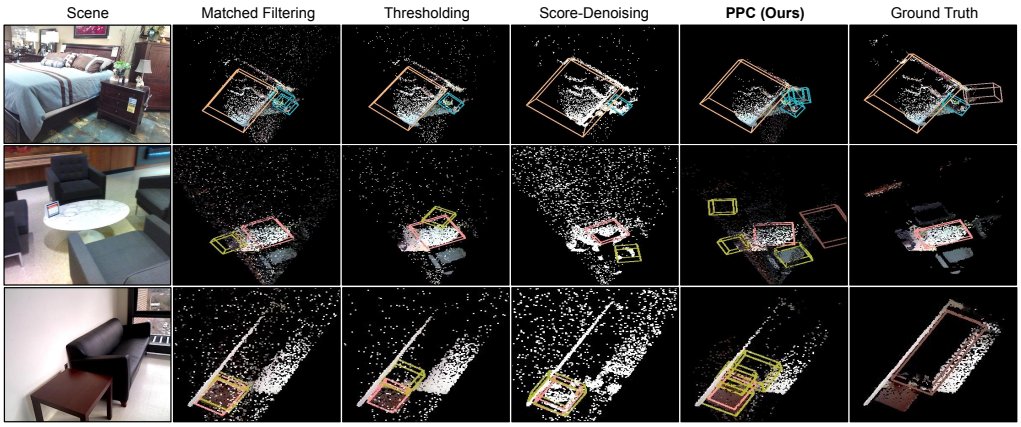


Figure 3.20: **3D Detection Failure Cases:** The first scene shows a scenario where PPC detects multiple boxes for the same object (nightstand). The second and third scenes show scenarios where a couch is detected as a chair by PPC and the baselines. Single-sitter couches are often detected as chairs by this model.

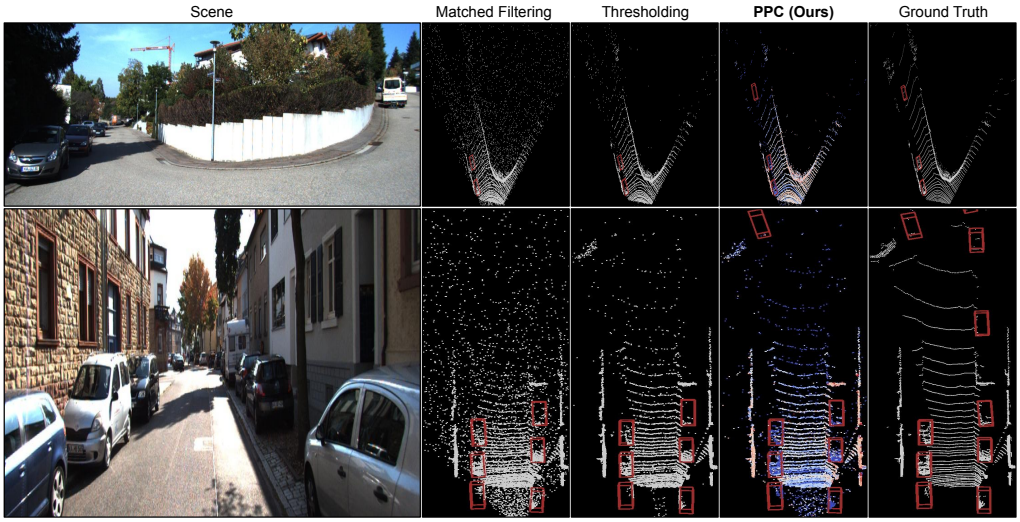


Figure 3.21: **3D Object Detection Results:** Figure shows scenes from the KITTI dataset under medium SBR (0.05) conditions. Baselines fail to detect farther cars. PPC is more robust for distant objects.

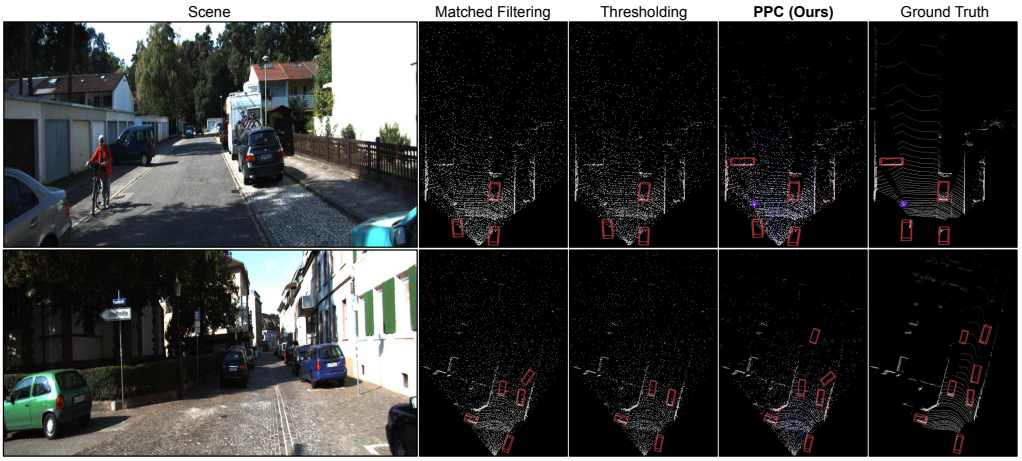


Figure 3.22: **3D Object Detection Results:** Figure shows scenes from the KITTI dataset under low SBR (0.02) conditions. The first scene shows a scenario where baselines fail to detect objects like farther cars and pedestrian. PPC is more robust for distant objects.

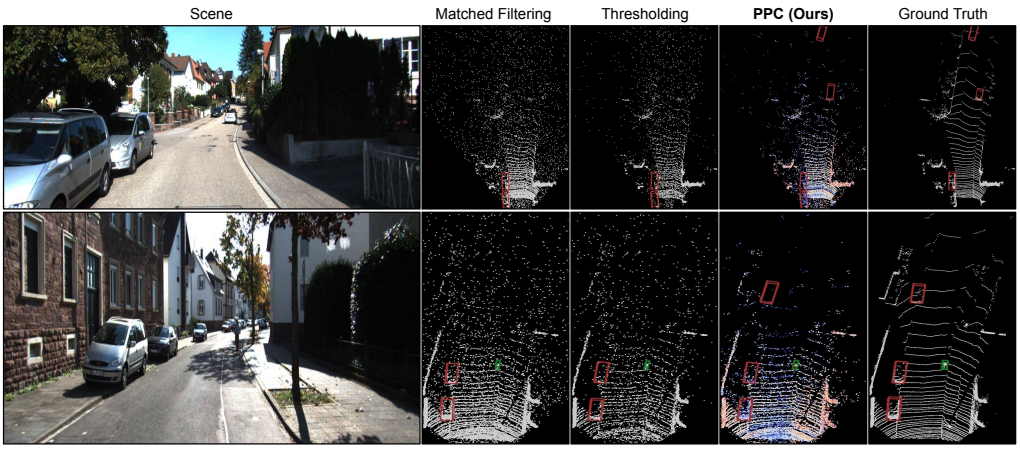


Figure 3.23: **3D Object Detection Results:** Figure shows scenes from the KITTI dataset under low SBR (0.01) conditions. Baselines struggle with farther cars, whereas PPC detects most objects.

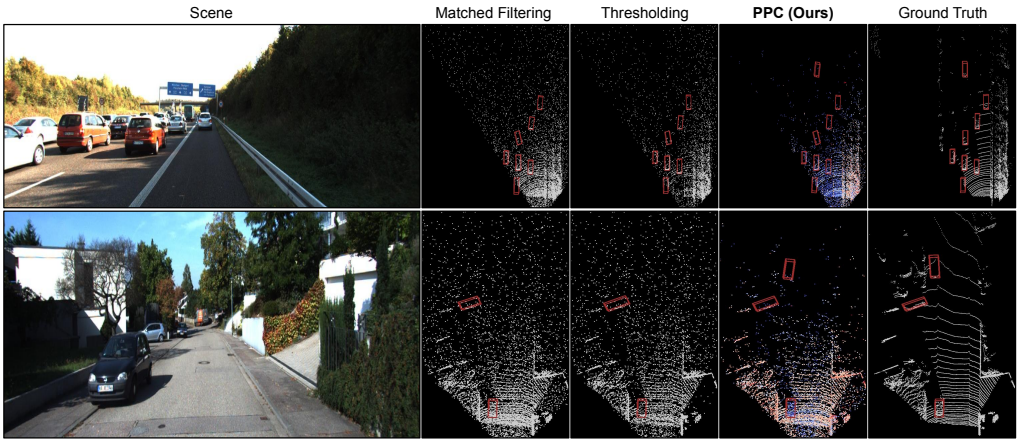


Figure 3.24: **3D Object Detection Results:** Figure shows scenes from the KITTI dataset under low SBR (0.005) conditions. Baselines fail to detect dark and farther objects like black cars. PPC is more robust for distant objects.

Avg. SBR	Clean		0.1		0.05		0.02		0.01	
	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50
Matched Filtering	63.37	35.51	53.89	27.64	53.23	24.67	37.54	10.99	33.17	7.98
Thresholding	64.25	36.17	59.57	33.44	58.82	32.45	42.43	18.17	39.51	12.61
PPC (Ours)	**64.36**	**36.94**	**61.51**	**35.69**	**60.19**	**31.38**	**53.21**	**25.37**	**46.84**	**20.14**

Table 3.8: **3D Detection Comparison using camera-LiDAR fusion architecture**: AP@0.25 and AP@0.50 results on the SUN RGB-D dataset using ImVoteNet show significant gains for PPC for all SBR levels.

(Qi et al., 2020). Table 3.8 includes the comparison on the SUN RGB-D dataset, which shows significant improvement for all SBR levels. Second, we evaluate using a recent LiDAR-only transformer-based architecture Uni3DETR (Wang et al., 2023). Table 3.9 includes mAP comparison on the SUN RGB-D dataset which shows performance improvement for low SBR levels. Lastly, we evaluate using a Pillar-based architecture PointPillars (Lang et al., 2019). Table 3.10 includes mAP for car, pedestrian, and cyclist categories on moderate difficulty of KITTI val split calculated with 11 recall positions. Our method shows significant improvements under low SBR conditions for pedestrian and cyclist categories.

PPC shows significant gains under low SBR for all detection architectures,

Avg. SBR	Clean		0.1		0.05		0.02		0.01	
	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50	AP@25	AP@50
Matched Filter	64.98	48.28	61.52	45.09	60.82	43.92	51.12	34.09	45.29	27.97
Thresholding	64.50	47.94	61.08	44.71	61.19	44.29	51.97	34.87	48.13	28.64
PPC (Ours)	**65.53**	**49.35**	**62.58**	**46.71**	**61.98**	**48.28**	**56.46**	**38.03**	**51.21**	**31.16**

Table 3.9: **3D Detection Comparison using LiDAR-only tranformer-based architecture**: AP@0.25 and AP@0.50 results on the SUN RGB-D dataset using Uni3DETR show significant gains for PPC under low SBR conditions.

Avg. SBR	Clean			0.05			0.02			0.01			0.005		
	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc	Car	Ped	Cyc
Matched Filtering	77.08	**52.78**	64.49	68.25	49.52	58.96	64.13	47.67	46.45	54.43	41.61	41.76	45.03	32.46	31.06
Thresholding	**77.34**	52.09	64.81	68.06	49.63	59.09	63.87	47.92	46.96	54.18	40.88	42.18	45.11	32.79	31.89
PPC (Ours)	77.19	52.12	**65.21**	**69.12**	**50.23**	**62.44**	**65.63**	**49.27**	**48.09**	**56.39**	**45.77**	**44.46**	**47.24**	**38.74**	**34.89**

Table 3.10: **3D Detection Comparison using LiDAR-only pillar-based architecture**: mAP results for car, pedestrian, and cyclist category on moderate difficulty of KITTI val split calculated with 11 recall positions for PointPillars. Our method shows significant gains under low SBR conditions.

which shows its versatility for a wide range of 3D detection models. The gain is large for point-net or transformer-based architectures (Uni3DETR, VoteNet, and ImVoteNet) as they suffer the most from the low SBR noise. The gain is significant but comparatively smaller for voxel or pillar-based architectures (PointPillars and PV-RCNN). Intuitively, this could be because the spurious points with large depth errors are away from the surface, and are not part of the same voxel or pillar as the surface points. Hence, their impact on the performance is also smaller.

3.10 Supplementary Section: Comparison with Denoised 3D Temporal Histograms

An effective approach for removing noise in 3D sensing systems described in this work is to denoise the 3D temporal histograms. Current state-of-the-art denoising methods for temporal histograms denoising (Peng et al., 2020) show

high performance on depth reconstruction tasks. Hence, we also evaluate our approach and baselines using denoised temporal histograms. We use a 3D-CNN denoising model (Peng et al., 2020) to denoise the temporal bins which are then used to construct point clouds for inference. We compare our method with the baselines under low SBR (0.02) conditions in Table 3.11. As expected, all methods perform better with denoised temporal histograms. Our method shows a further gain in AP@0.25 of about 3-4% which shows that 3D inference can benefit from our PPC approach with denoised temporal histograms as well.

Histogram Denoising Method	Thresholding	PointClean Net	Score-Denoising	PathNet	**PPC (Ours)**
-	29.58	18.24	26.35	25.45	**38.49**
3D-CNN	50.30	51.03	50.85	51.07	**54.25**
Gaussian Filter	38.79	40.12	43.36	43.25	**50.93**

Table 3.11: **3D Object Detection Results**: Comparison of AP@0.25 results using denoised temporal histograms.

We also compare the effectiveness of our method with a non-learning-based histograms denoising method. We use a 5x5 Gaussian filter in the spatial dimension to denoise histograms. Matched filtering is still used over the temporal dimension. Table 3.11 shows our method has significant gains with Gaussian denoised histograms and the performance is comparable to the results with 3D-CNN denoising.

Should we denoise 3D temporal histograms for inference? Denoising methods like (Peng et al., 2020) require compute and memory-intensive 3D-CNN operations, which makes it infeasible for real-time applications. It is thus not suitable for sensor on-chip processing. It also requires a read-out of full 3D temporal histograms, which has a significantly high data-bandwidth cost (compared to only reading out a point cloud as considered in earlier approaches in the main paper). However, it is an effective approach for denoising in non-real-time applications.

Gaussian filter is a computationally cheap non-learning-based operation and is feasible with sensor on-chip processing. Table 3.12 shows the inference time

of our complete recognition pipeline (including denoising temporal histograms, point cloud processing, and inference) with different histogram denoising methods that are discussed. This suggests that, given a computational budget, a simple histogram denoising approach like a Gaussian filter is a good candidate for 3D recognition.

Histogram Denoising Method	Runtime (ms)
-	95
3D-CNN	7200
Gaussian Filter	98

Table 3.12: **Runtime Time for 3D Detection:** Comparison of per scene runtime time of our method using different histogram denoising methods.

Comparison with Compressed 3D Timing Histograms

Recently, compression techniques have been proposed to read out compressed representations (Gutierrez-Barragan et al., 2022) of the temporal histograms to reduce data bandwidth requirements. We also show the performance of our approach on such decompressed histograms in Table 3.13. We use a lightweight Truncated Fourier (k=32) representation from (Gutierrez-Barragan et al., 2022) for evaluation. Our approach is effective even under the data loss incurred due to compression and shows significant gains over the Thresholding baseline for 3D detection.

	Threshoding	**PPC (Ours)**
Decompressed Histograms	16.50	**29.77**

Table 3.13: **3D Object Detection Results**: Comparison of AP@0.25 results using decompressed histograms.

4 ROBUST SCENE INFERENCE UNDER 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 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 nighttime 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 work 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) (Niclass et al., 2005; Rochas, 2003b) are an emerging image sensor technology that is 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 (Buttafava et al., 2015; O'Toole et al., 2018; Bruschini et al., 2019). But, recently, due to their compatibility with CMOS fabrication processes, high-resolution cameras (up to 1 MPixel) have been developed based on SPADs (Morimoto et al., 2020), as well as the

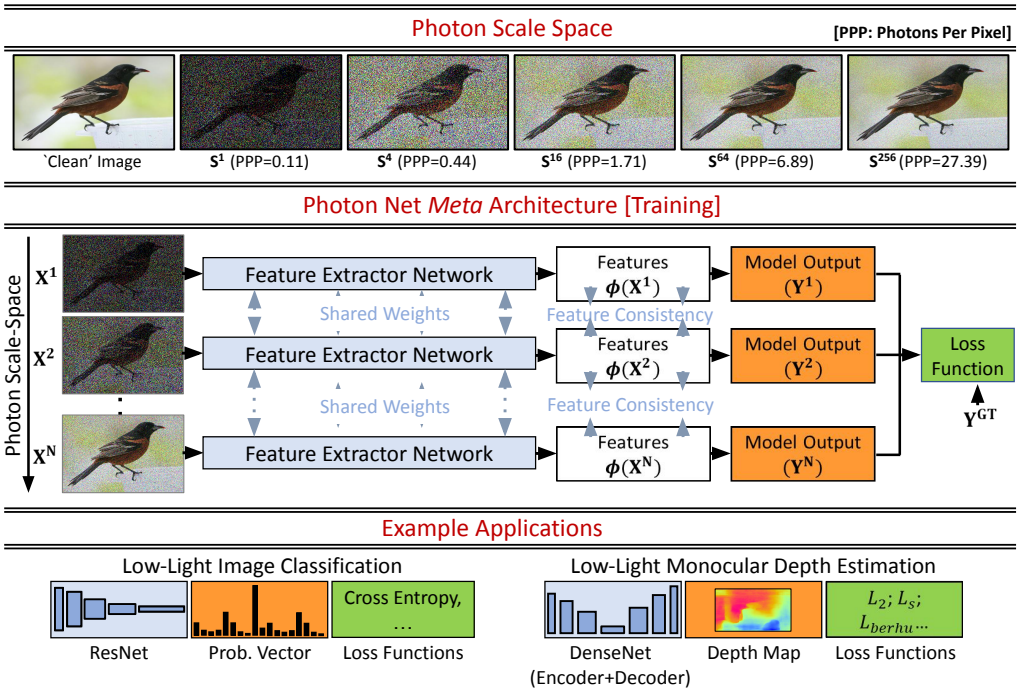


Figure 4.1: **Inference in Low-Light using Photon Scale-Space. (Top)** Photon scale-space is a hierarchy of images, each with a different flux level, but sharing the same scene content. Successive images in the hierarchy have similar flux so that high-flux images can guide the low-flux images during 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.

jots (Ma et al., 2017) technology. These single photon cameras are capable of capturing sequences of binary frames with minimal read noise (Ulku et al., 2018), 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 (Antolovic et al., 2018; Ingle et al., 2019; Ma et al., 2020) and low/mid-level scene information such as 3D shape (O'Connor and Phillips, 1984; Renker, 2006; Dautet et al., 1993; Kirmani et al., 2014; Shin et al., 2016; Gupta et al., 2019c,a) and motion (Gyongy et al., 2018). 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 (Liu et al., 2020a, 2019; Diamond et al., 2017), inference on ultra-low-light images where each pixel receives less than a photon on average still remains an intractable problem.

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 (He, 2010) where a guidance image is used for low-level image processing tasks, such as denoising (He, 2010) and super-resolution (de Lutio et al., 2019). More recently, the idea of guided training has been explored in the context of student-teacher training (Gnanasambandam and Chan, 2020) 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 average 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 the most accurate labels), *successive* images in the hierarchy have a 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. 4.1), reminiscent of the classical image-size scale-space (Lindeberg, 1994) 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. 4.1). The key idea is to train a given network architecture with different images from a photon scale, 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 the 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 (Ulku et al., 2018)), 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. 4.1). 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 (Antolovic et al., 2018; Ingle et al., 2019; Ma et al., 2020), 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.

4.1 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 (Fossum, 2005). By avoiding avalanche, jots achieve smaller pixel pitch, higher quantum efficiency and lower dark current, but have lower temporal resolution (Ma et al., 2017). 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 (Chen and Perona, 2016, 2017) 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 sen-

sors (Gyongy et al., 2018), and more recently, object identification (Antsiperov, 2019) and image classification (Gnanasambandam and Chan, 2020). 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 (Liu et al., 2020a, 2019; Diamond et al., 2017). 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 (~ 0.1 PPP).

Image-size Scale Space: A recent work (Xu et al., 2020) 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.

4.2 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, 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[1], we can generate multiple stochastic binary images as captured by a single photon camera using the image formation model described in Section 2.2. 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*

$$\mathcal{S}^N(x, y) = \sum_{i=1}^{N} B_i(x, y) \ . \tag{4.1}$$

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* $K < L$.

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 tradeoff: 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 the computational cost of the training algorithms. In our implementation, 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.

---

[1] In general, the pixel intensities have a non-linear relationship to incident flux due to the 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.

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. 4.1 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:

$$\begin{aligned} E[\mathcal{S}^N(x, y)] &= N * E[B(x, y)] \\ &= N(1 - e^{-(\phi\tau\eta + r_q\tau)}) \,. \end{aligned} \tag{4.2}$$

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

$$\hat{\phi}(x, y) = -\ln(1 - \mathcal{S}(x, y)/N)/\tau\eta - r_q/\eta \,. \tag{4.3}$$

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 Ma et al. (2020); Antolovic et al. (2018) — $\mathcal{S}^N$ 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.

4.3 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. 4.1 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.

$$L_{MSE}(\{x_i\}) = \frac{1}{N} \sum_{i,j} \|\varphi(x_i) - \varphi(x_j)\|_2^2 \tag{4.4}$$

where $\{(x_i)\}$ is a set of all training images, N is the total number of training pairs in the mini-batch with the 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 the network. We use a feature vector from the final layer of the CNN (after the global pooling layer) as our feature representation.

The 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

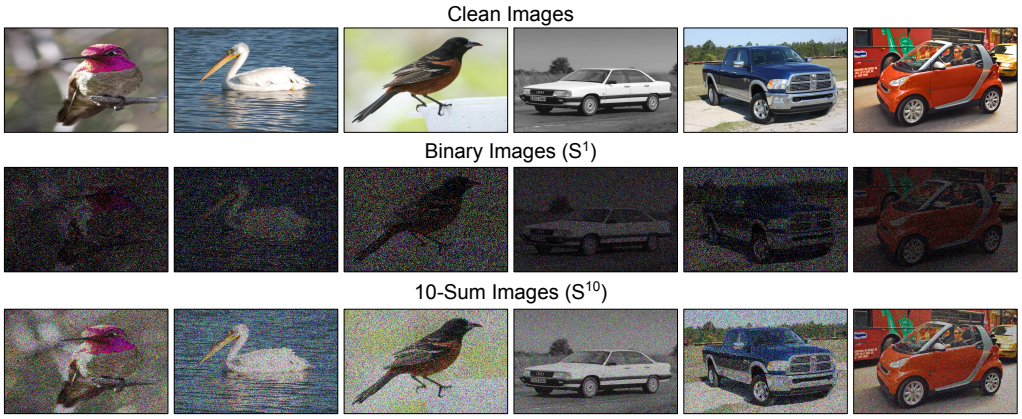


Figure 4.2: **Simulated Single-Photon Images**: Clean images and simulated noisy images from CUB-200-2011 and CARS-196 dataset. SPCs capture a sequence of binary images like ($\mathcal{S}^1$) with heavy shot noise. 10-sum images ($\mathcal{S}^{10}$) are average of 10 binary images.

control the contribution of both losses. Please see the supplementary report for details.

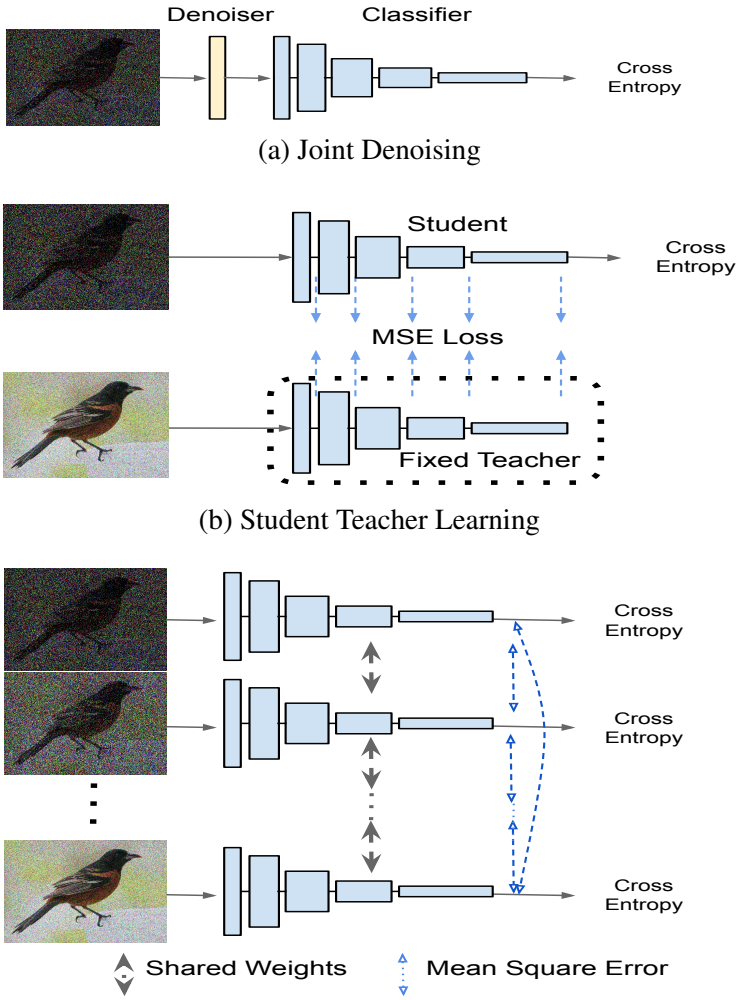
4.4 Low-Light Scene Inference

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. 4.1). We demonstrate the effectiveness and versatility of the proposed techniques via two low-flux inference tasks: image classification and monocular depth estimation.

Image Classification

We first show the application of our approach to image classification task.

Datasets: We first show the performance of our approach using simulated images using two datasets: CUB-200-2011 image classification dataset (Wah et al., 2011) and CARS dataset (Krause et al., 2013). CUB-200-2011 is commonly used for



(a) Joint Denoising

(b) Student Teacher Learning

(c) Photon Net (Ours)

Figure 4.3: **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.

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 2.5. $\phi\tau\eta$ 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 4.2 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 (Diamond et al., 2017), 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. 4.3a), 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 (Gnanasambandam and Chan, 2020) where clean images are used to train a *fixed* teacher network and noisy images are used for training the student network (Fig. 4.3b). 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 (He et al., 2016) as the backbone architecture provided by Pytorch (Paszke et al., 2019) for all the baselines. As shown in Figure 4.3, 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

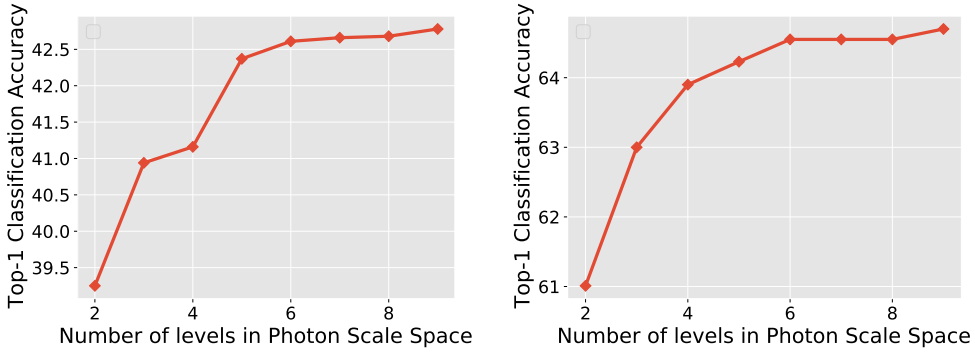
Test Data	PPP	CUB-200-2011 Joint Denoising	CUB-200-2011 Student-Teacher Learning	CUB-200-2011 Photon Net (Ours)	CARS-196 Joint Denoising	CARS-196 Student-Teacher Learning	CARS-196 Photon Net (Ours)
$\mathcal{S}^1$	0.11	27.21	35.43	**42.37**	34.51	57.81	**64.23**
$\mathcal{S}^2$	0.22	31.33	39.50	**48.56**	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 4.1: **Image Classification Results**: Top-1 Accuracy results on CUB-200-2011 and CARS-196 dataset. Photon Net outperforms both Joint Denoising (Diamond et al., 2017) and Student-Teacher Learning (Gnanasambandam and Chan, 2020) on all noise levels.

of levels is analyzed 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 4.1 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. 4.4 shows results of image



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

Figure 4.4: **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 the increasing number of levels in PSS and saturates at 5-6 levels for both datasets.

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.

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 (Alhashim and Wonka, 2018) base architecture, consisting of an Encoder and a Decoder. The Encoder is a deep CNN (ResNet-34 pretrained on ImageNet (Deng et al., 2009)) 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

Test Set	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 4.2: **Monocular Depth Estimation Results:** on NYUV2 dataset.

Structural Similarity loss between predicted and ground truth depth values. We use the same training procedure as described in (Alhashim and Wonka, 2018).

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 (Silberman et al., 2012). Same training and testing split is used as (Alhashim and Wonka, 2018) which includes 50K training and 654 testing images. We simulate SPC images using the same procedure described earlier in Section 4.4. The following six standard evaluation metrics are used:

- average relative error (rel): $\frac{1}{n}\sum_p^n \frac{y_p - \hat{y}_p}{y}$,
- root mean square error (rms): $\sqrt{\frac{1}{n}\sum_p^n (y_p - \hat{y}_p)^2}$,
- average ($\log_{10}$) error: $\frac{1}{n}\sum_p^n |\log_{10}(y_p) - \log_{10}(\hat{y}_p)|$ 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 the 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 a Depth Estimation architecture. Table 4.2 shows results on the NYUV2 dataset,

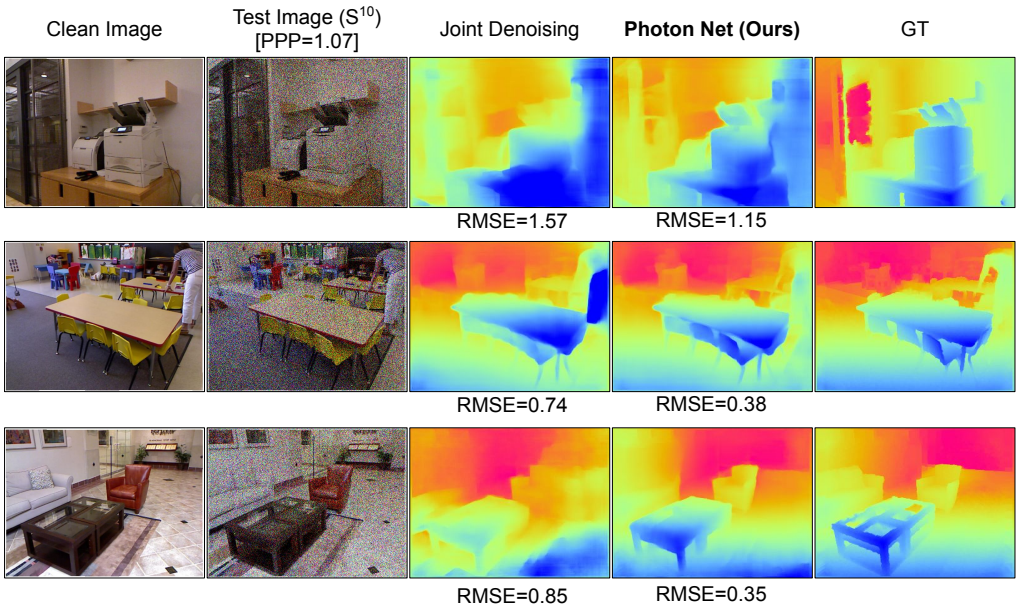


Figure 4.5: **Estimated depth maps**: Comparison of depth maps from Photon Net (Ours) and the baseline on NYUV2 test images $\mathcal{S}^{10}$ (PPP~1.07).

and Fig. 4.5 shows example depth output results of our approach and the baseline. Photon Net outperforms the baseline approach both qualitatively and quantitatively for multiple noise levels.

4.5 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 (Ulku et al., 2018) (Fig. 4.6).

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 gray-scale (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

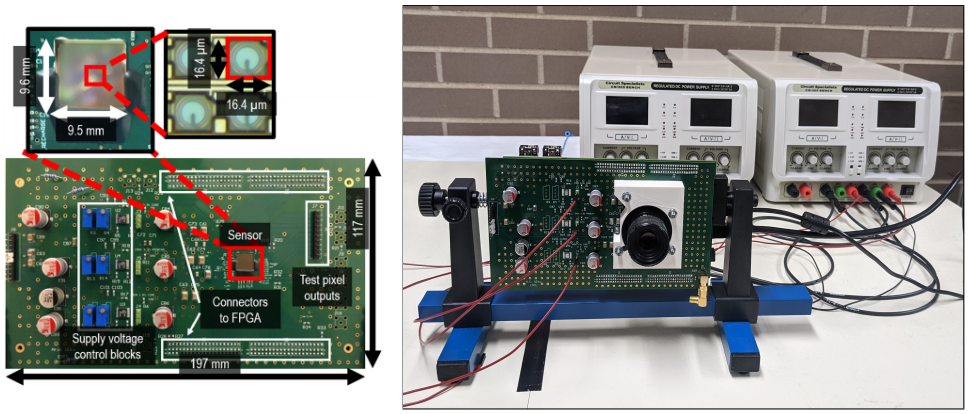


Figure 4.6: **Camera Setup:** SwissSPAD2 board (Left) and the setup for image capture (Right).

the location of the hot pixels and then filter them by using spatial neighborhood information.

Dataset: For the dataset collection for image classification task, we displayed the original RGB images on a monitor 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. 4.7 shows examples of N-Sum images captured by the camera.

Experiments and Results: We follow the same procedure for training as described in Section 4.4. Table 4.3 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. 4.8 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

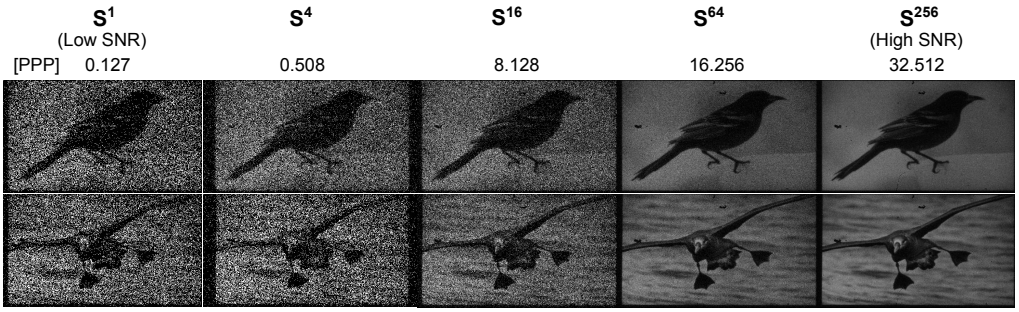


Figure 4.7: **Real SPAD images** captured from CUB-200-2011 dataset using the SwissSPAD2 camera.

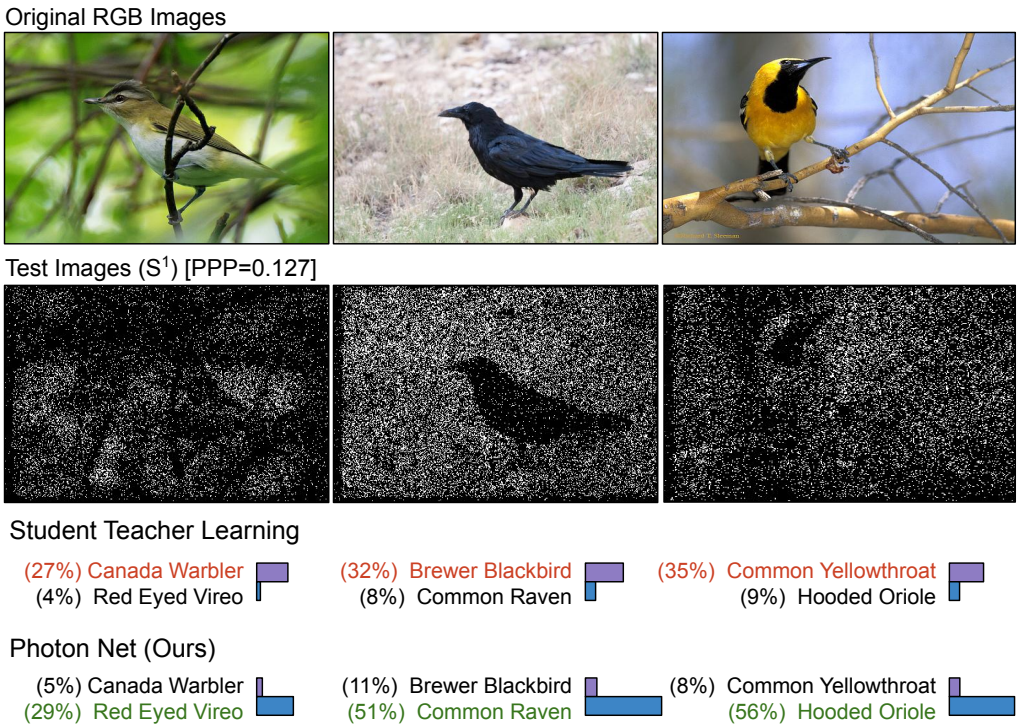


Figure 4.8: **Results with Real SPAD Sensor** of image classification on CUB-200 dataset for $\mathcal{S}^1$ test images with prediction probabilities output by both Student Teacher Learning and Photon Net (Ours). Classification output is highlighted in red for wrong predictions and green for correct predictions.

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

~ 0.1, the proposed photon net approach is able to recover correct class labels.

4.6 Discussion

In this work, we present an approach for scene inference using single photon cameras, which shows significant improvement under low light conditions. Our proposed method uses higher SNR images during training and encourages the model to use them as a guide for low SNR images. This improves robustness to shot noise originating from the stochastic nature of photon arrival during image capture. We demonstrate our method on multiple image classification and monocular depth estimation tasks, and show significant gains in extreme low light conditions ($\sim$0.1 photons per pixel).

Low-light inference beyond classification and depth estimation: So far in this work, we 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 (Ren et al., 2015), instance segmentation (He et al., 2017), and key-point detection (Newell et al., 2016).

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 (An-

tolovic et al., 2018; Ingle et al., 2019; Ma et al., 2020), the proposed techniques can be adapted for inference in extremely bright scenes where conventional sensors get saturated.

4.7 Supplementary Section: Additional Image Classification Results

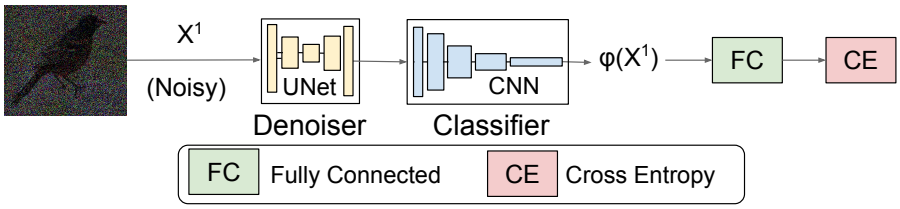
In this section, we include further technical details and results that complement our main results for image classifiation.

Architecture Overview

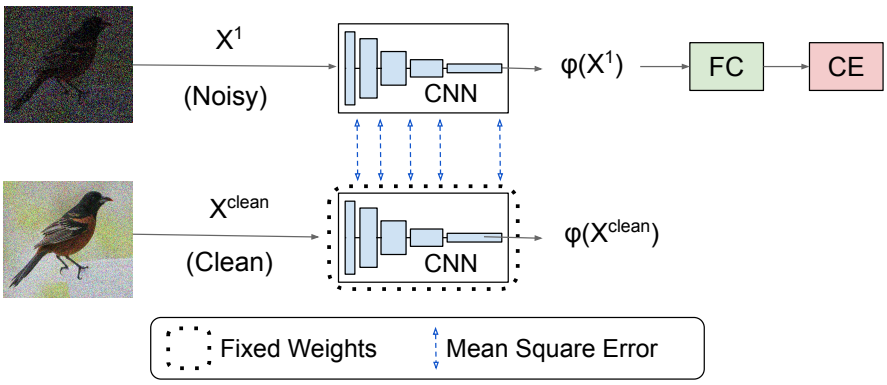
We provide more detailed overview of the architectures used for the image classification task.

Joint Denoising: Joint Denoising architecture (Diamond et al., 2017) consists of a joint network with a denoiser (20 layer UNet) and a CNN classifier (Resnet-18 (He et al., 2016)). We use Mean Squared Error loss for the denoiser which uses noisy and clean images. Cross Entroy Loss is used for the classifer with uses the class label of the image. The joint network is trained with sum of both the losses (Figure 4.9a). The denoiser is initialized with pretrained weights on noisy and clean images.

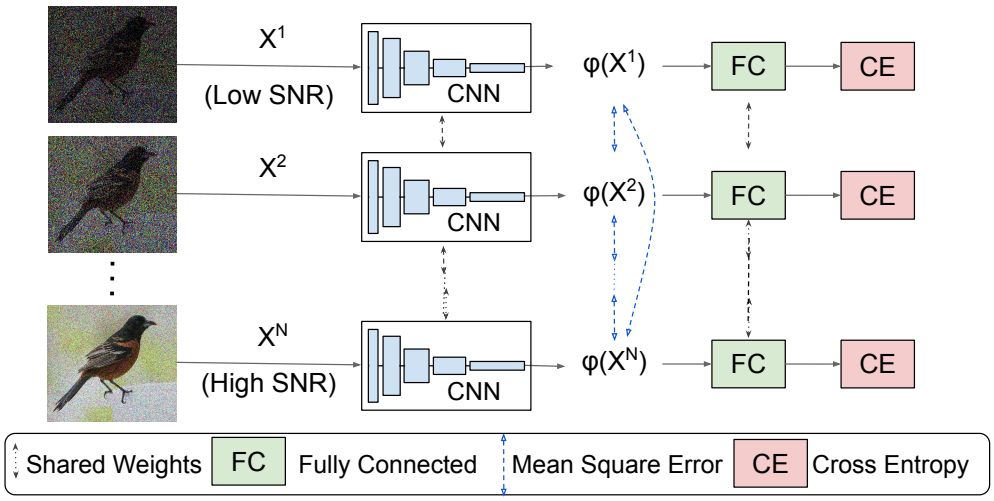
Student Teacher Learning: Student Teacher architecture (Gnanasambandam and Chan, 2020) is composed of a teacher network and a student network. Teacher network (ResNet-18) is a pre-trained classifier on clean images. Student Network uses the same network architecture as the teacher network (ResNet-18). Intermediate feature output maps ('relu', 'layer1', 'layer2', 'layer3', 'layer4' from pytorch's implementation) from the CNN Network of both student and teacher network is used for feature consistency. Final training consists of training the student network with cross entropy loss and mean squared error loss while teacher network is kept fixed (Figure 4.9b). Student Teacher learning uses double the



(a) Joint Denoising



(b) Student Teacher Learning



(c) Photon Net (Ours)

Figure 4.9: **Architecture Overview** for Image Classification

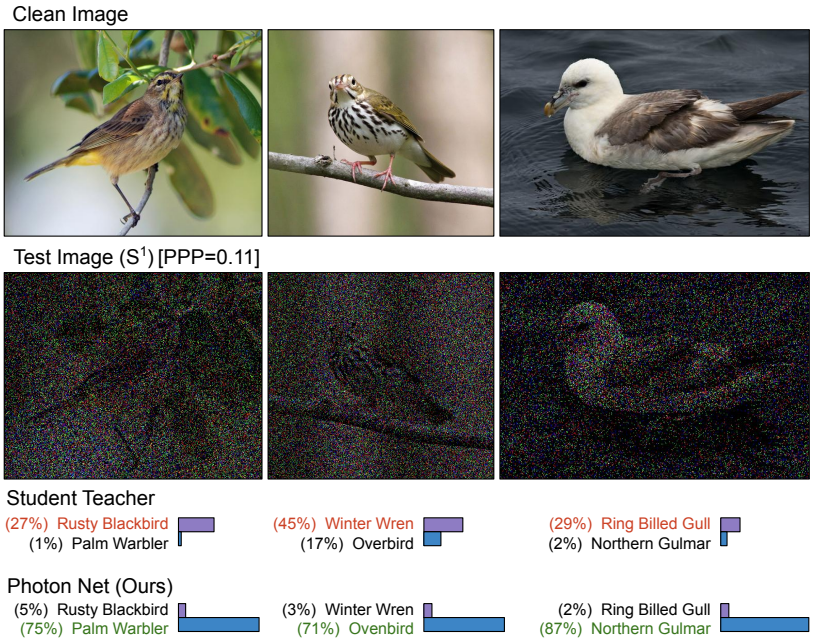


Figure 4.10: **Image Classification Results** using Photon Net on CUB-200-2011 Dataset for $\mathcal{S}^1$ test images.

network parameters for classifier during training but only uses student network for testing.

Photon Net (Ours): Photon Net training uses multiple images with different PPP level as input to the network. Different branches of the network are CNN architectures (ResNet-18) which share weights with each other and act as a feature extractor. Images with different PPP levels are sampled together in the same mini-batch so gradients from high SNR image branches can guide the low SNR images. The feature output from the final layer (after global pooling layer) is used for the feature consistency of different PPP levels using Mean Square Error Loss. Cross Entropy Loss is used for the training the image classifier which uses the classification label.

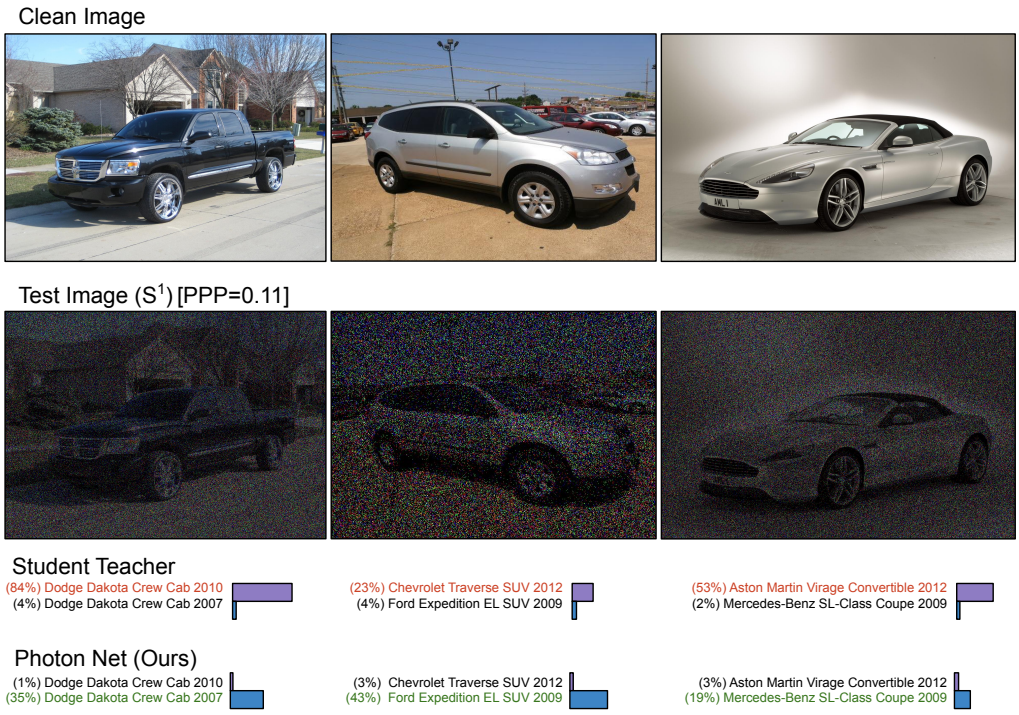


Figure 4.11: **Image Classification Results** using Photon Net on CARS Dataset for $\mathcal{S}^1$ test images.

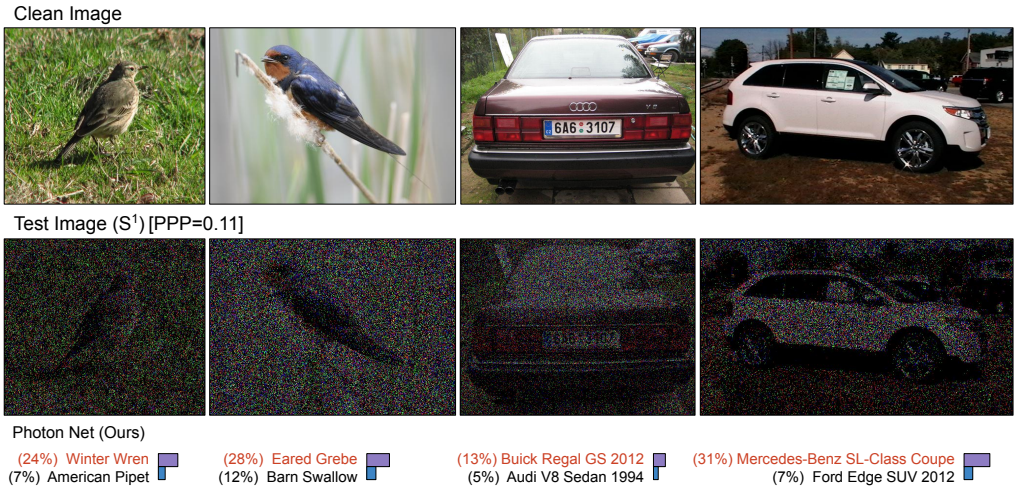
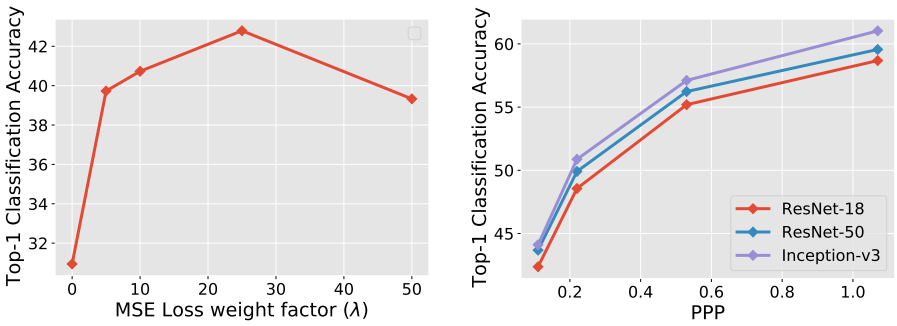


Figure 4.12: **Few Failure cases examples** of Photon Net on CUB-200-2011 and CARS dataset for $\mathcal{S}^1$ test images.



(a) MSE Loss weight factor (λ) (b) Varying CNN backbones

Figure 4.13: **Ablation Studies**: Performance of Photon Net training while varying: (a) MSE loss weight factor (λ) , (b) CNN backbones

Additional Results

Figure 4.10 and 4.11 shows results of image classification on CUB-200-2011 (Wah et al., 2011) and CARS (Krause et al., 2013) dataset $\mathcal{S}^1$ test images using Photon Net. Probability output of incorrect class is highlighted in red and correct class is highlighted in green. Even in the case of extreme low light (PPP 0.1), Photon Net is able to recover the correct output label. Figure 4.12 example of few failure cases where Photon Net architecture fails to get the correct prediction. As we can observe, these cases are extremely challenging.

More Ablation Studies

We study the effect of the hyper parameter of the Photon Net training on the performance. We vary the weighting factor of the MSE loss in the overall loss for image classification. We start with λ=0 and increase upto $\lambda = 50.0$. Figure 4.13 shows Photon Net performs best for λ=25.0.

We also analyse the performance of Photon Net using different base architecture for the feature extractor. We compare ResNet-18 with deeper CNN architectures such ResNet-50 and InceptionV3. (Szegedy et al., 2016). Figure 4.13 shows increase in the performance of Photon Net with deeper CNN architectures. This

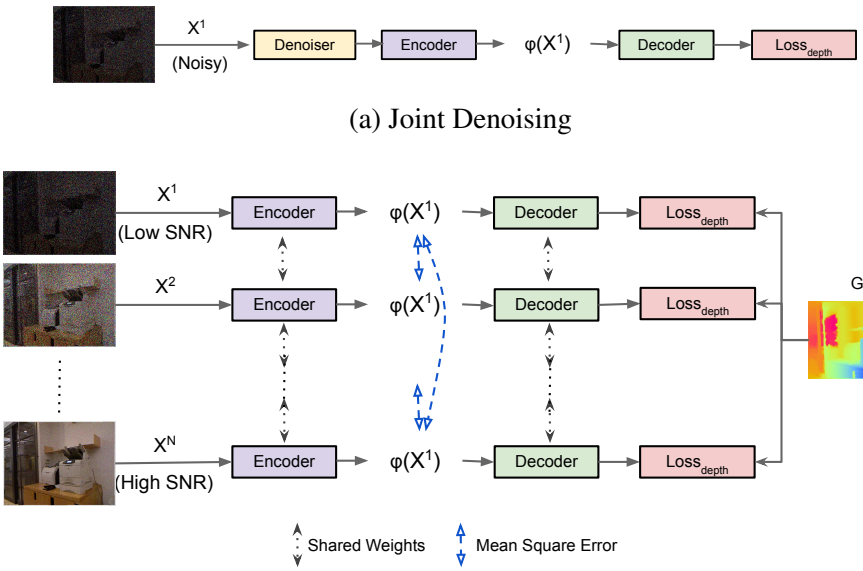
Test Data	PPP	Vanilla Net	Vanilla Net w/ Photon Scaled Images	BM3D Denoising	Curriculum Learning	Student Teacher Learning (N-steps)	Photon Net (Ours)
$\mathcal{S}^1$	0.11	21.35	28.92	25.52	33.72	35.79	**42.37**
$\mathcal{S}^2$	0.22	25.61	34.51	29.15	39.44	42.16	**48.56**
$\mathcal{S}^5$	0.53	37.14	43.26	38.81	44.99	46.91	**55.19**
$\mathcal{S}^{10}$	1.07	42.99	44.63	43.34	48.65	48.86	**58.68**

Table 4.4: **Ablation Study**: Top-1 Accuracy results of image classification on CUB-200-2011 dataset.

shows the versatility and ease to extend Photon Net to different CNN architectures.

We perform an ablation study to analyse the individual contribution of Photon Net training and using Photon Scaled Images in the final performance. Table 4.4 shows Top-1 accuracy on CUB-200-2011 dataset. 'Vanilla Net' represents the training procedure where a conventional image classification CNN model (ResNet-18) is trained with cross entropy loss using only noisy images. 'Vanilla Net w/ Photon Scaled Images' trains the Vanilla Net with photon scaled images. As we an see, adding Photon Scale Space images increases the performance by about 8-9% on all noise levels and shows the effectiveness of high SNR images in training. Photon Net training further improves the model by more than 13% as feature consistency loss increases the robustness to noise. 'BM3D denoising' shows the performance of Vanilla Net training on denoising training and testing images using BM3D algorirhtm.

We also compare our model to Curriculum Learning technique, where the Vanilla Net is trained in N steps, starting with only the clean images first step and successively finetuning the model by adding images with higher noise levels in next steps. Photon Net outperforms Curriculum Learning as it uses the high SNR as a guide more effectively by adding the feature consistency loss. We also do Student Teacher Learning in N-steps (N is number of photon scaled levels) using the Photon Scaled Images. We use successive levels of photon scaled images for student and teacher network. Photon Net performs better Student Teacher Learning by significant margin.



(a) Joint Denoising

(b) Photon Net (Ours)

Figure 4.14: **Arechitecture Overview** of Monocular Depth Estimation with Photon Net:

4.8 Supplementary Section: Additional Monocular Depth Estimation Results

In this section, we include further technical details and results that complement our main results for monocular depth estimation.

Architecture Overview

We provide more detailed overview of the architectures used for the monocular depth estimation task.

Joint Denoising: Joint Denoising consists of a depth estimation architecture based on DenseDepth (Alhashim and Wonka, 2018) coupled with a denoiser for noisy images. Denoiser is a UNet network (20 layers) which is pretrained on noisy and clean images using Mean Square Error Loss. DenseDepth architecture for

depth estimation consists of an encoder network (Deep CNN network pretrained on Imagenet) and a decoder network (upsampling layers with skip connects) that generates the output depth maps. Loss function for depth estimation is a combination of point wise L1 loss and Structural Similarity loss between predicted and ground truth depth values. Overall Loss is the sum of losses from denoiser and depth estimation.

Photon Net: Photon Net architecture takes multiple images with different PPP levels as the input to the network. Different branches of the network are the encoder networks with shared weights. We use the same encoder and decoder as baseline for fair comparison. Different images are sampled together in the same mini-batch in order for high SNR images to guide the low SNR images. Final feature output map from the encoder (after global pooling layer) is used for the feature consistency of different PPP levels (using Mean Square Error Loss). Overall Loss is the combination of Mean Square Error loss (for feature consistency) and depth estimation loss (point wise L1 loss and Structural Similarity Loss).

Additional Results

Figure 4.15 shows examples of output depth maps from the Photon Net and the baseline for more sample test cases, which shows significant gains both qualitatively and quantitatively. Figure 4.16 shows output depth maps while using increasing SNR test image of the same scene.

4.9 Supplementary Section: Additional Real SPAD Captures

In this section, we provide further details regarding real SPAD dataset collectoin method and additional real captures for visualization. To collect dataset of real captures from SPAD sensors, we displayed the original RGB images on a monitor

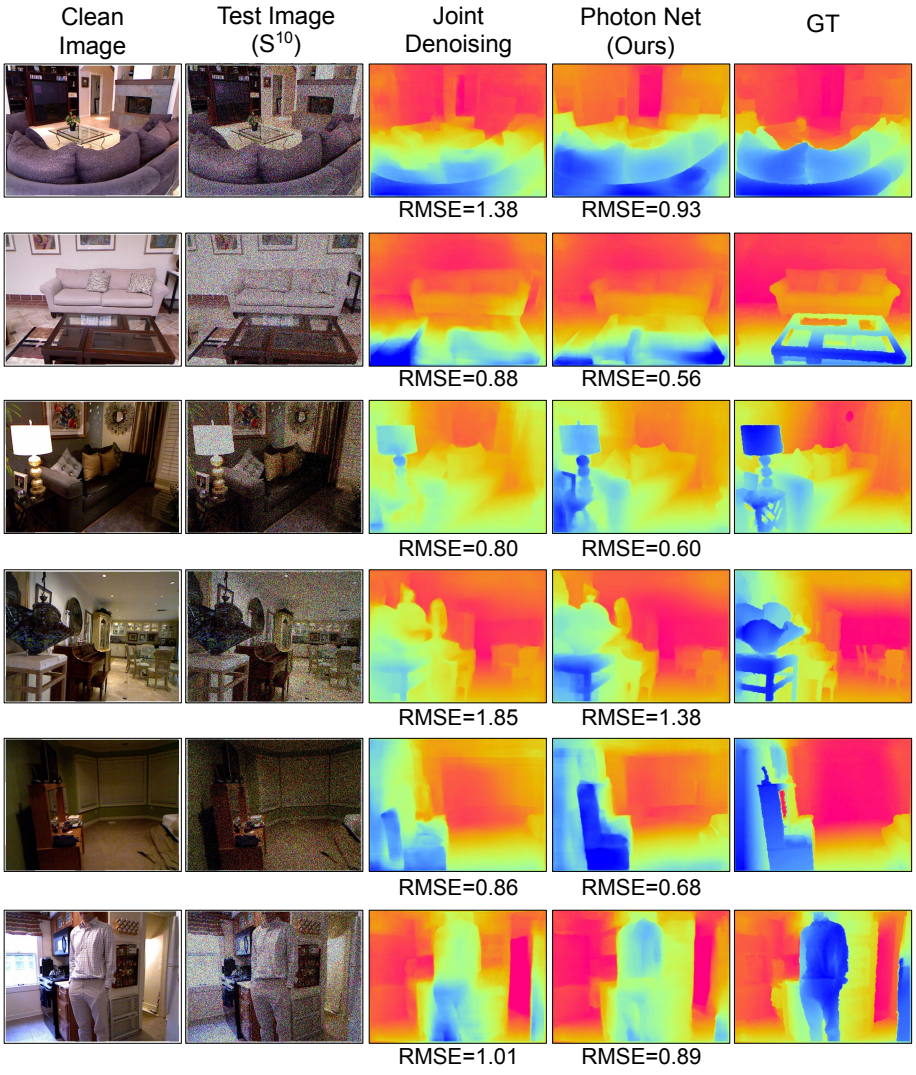


Figure 4.15: **Monocular Depth Estimation Results** on NYUV2 dataset of $\mathcal{S}^{10}$ test images

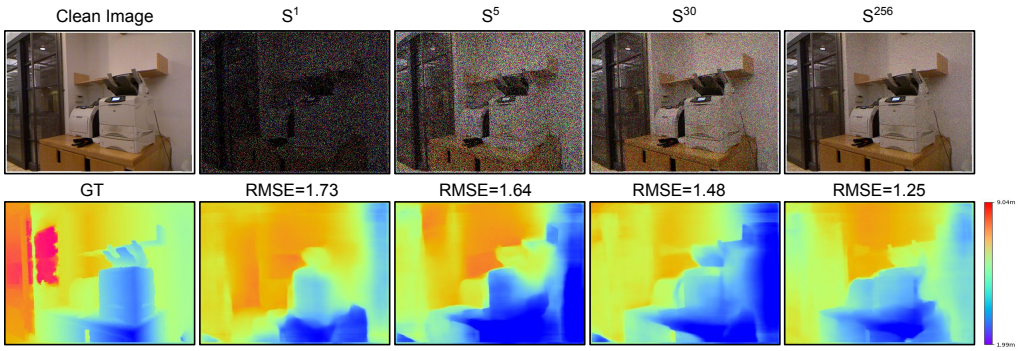


Figure 4.16: **Monocular Depth Estimation Results** on NYUV2 dataset with increasing PPP level in the testing image

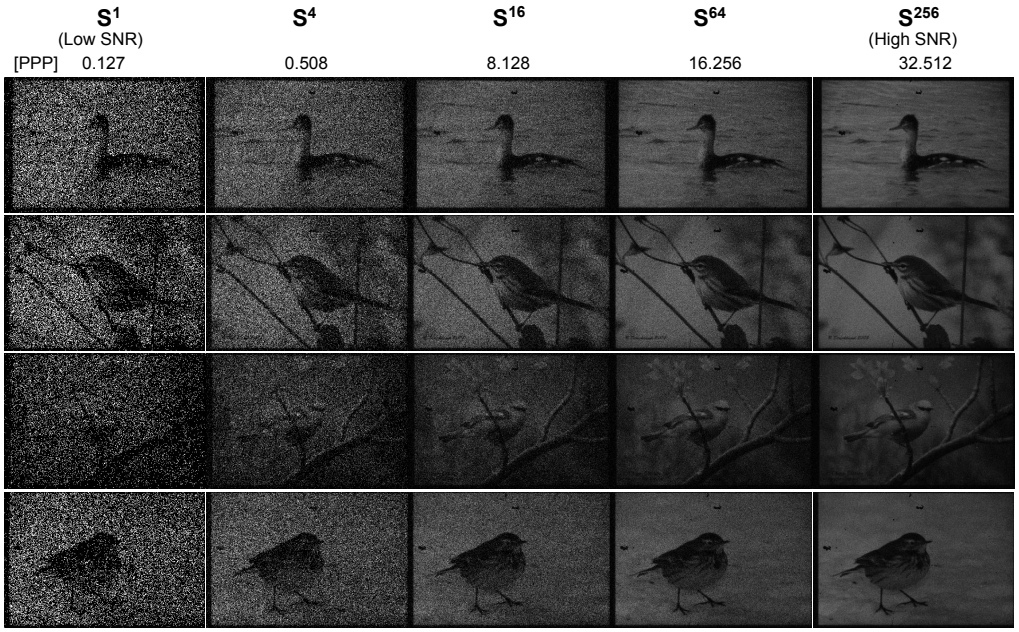


Figure 4.17: **Real Captures:** Sample of images from SPAD cameras

Figure 4.18: **Artifacts** in Real Captures from SwissSPAD2 camera

screen (full screen while maintainting the aspect ratio) and captured it using SPAD sensors. The camera is positioned to cover the monitor display in its field of view. Since the monitor has the aspect ratio of 16:9 and camera has the resolution 512x256, captured frames have black padding outside the screen area. We crop all the captured frames based on the size of the original images to remove all the padding. Frames are grayscale and contain hot pixels. We correct these hot pixels by capturing an image of a black scene to identify the locations and then filter them using spatial neighborhood information. Figure 4.17 shows example of images captured using SwissSPAD2 camera (Ulku et al., 2018) as described in Section 7 of the main text. Images formed from the sensor contain a few artifacts (in form of black patches) as shown in Figure 4.18.

5 ROBUST SCENE INFERENCE UNDER LOW-LIGHT AND MOTION

Consider a set of images captured under low-light at varying exposures (Figure 5.1), thereby spanning the space of noise-blur "dual corruptions". Each image, being corrupted in its own way, offers a different "window" on the scene: moving objects will appear sharper when the exposure is lower, while static low-contrast regions will be more easily perceptible in longer exposures. In other words, while any single image from the set might never be optimal in challenging scenarios, the *set of images spanning the dual corruption space* contains much richer and complementary information that can be leveraged for performing robust scene inference even under challenging imaging scenarios.

> *All happy families are alike;*
> *each unhappy family is unhappy in its own way.*
>
> Leo Tolstoy

In this work, we propose the idea of performing scene inference in the space of noise-blur corruption. Our *key observation* is that by utilizing the "persistence of prediction" across differently degraded images of the same scene, significantly higher accuracy can be achieved as compared to performing inference on individual images. Figure 5.1 shows an example. Although differently degraded images have different low-level features, the semantic content remains the same across all images. We develop techniques that encourage similar predictions from individual captures, and aggregate the predictions across individual images for robust visual recognition.

We demonstrate the proposed approaches on two visual recognition tasks, namely image classification and object detection. We perform experiments on large scale datasets of real images with synthetic corruptions and show that performing inference on a set of dual corruption images outperforms conventional

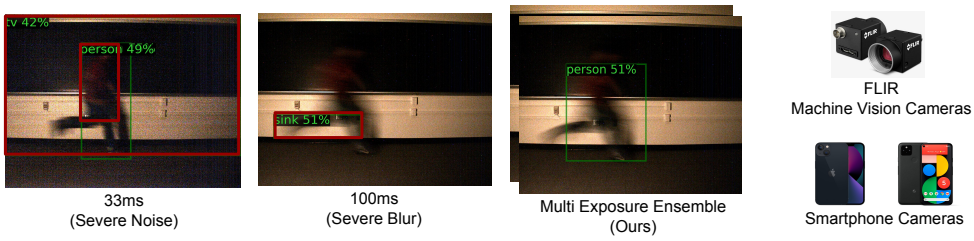


Figure 5.1: **Multi Exposure Ensemble:** Figure shows a scene containing a fast moving object under low-light. Images with short exposure (a) and long exposure (b) suffer from dual corruption: noise and/or blur. Inference tasks like object detection on these images are severely affected: numerous false positives and wrong bounding boxes. (c) Our approach leverages multiple captures of varying exposures for robust inference: accurate and tight bounding boxes. (d) Such multi-exposure images are easy to capture with machine vision cameras or modern smartphone cameras (*e.g*, Google Pixel and iPhone) that use burst photography for HDR imaging.

baselines in extreme low-light and motion conditions. Finally, we also show improved performance on real-world experiments using machine vision sensors.

Scope and Limitations: While implementing this idea requires capturing multiple exposures, most modern cameras already allow varying imaging parameters (*e.g*, exposure, aperture) in rapid succession. For example, modern cell phone cameras can take multiple snaps with a variety of exposures and fuse them to create an aesthetically pleasing image (Hasinoff et al., 2016). Increasingly, machine vision sensors (Comma, 2022) are also starting to perform exposure bracketing to capture high dynamic range (HDR) images for autonomous driver assist systems, while others go further and offer the capability of *simultaneously* capturing different exposure images via a spatially varying exposure sensor for HDR imaging (Nayar and Mitsunaga, 2000) and motion-deblurring (Nguyen et al., 2022). These ongoing developments in camera technology, coupled with the proposed computational techniques can lead to the next generation of computer vision systems which will perform reliably even in non-ideal real-world scenarios (*e.g*, imagine an

autonomous car driving on a dark night attempting to detect pedestrians) where it is extremely challenging for conventional algorithms to extract meaningful information reliably.

5.1 Related Work

Image Corruptions and Benchmarks: There has been some recent interest in simulating common image corruptions and benchmarking their adversarial effect on the performance of computer vision models, especially those relying on deep models (Hendrycks and Dietterich, 2018; Michaelis et al., 2019). In parallel, developing robust visual inference methods has also received much attention. For example, a teacher-student framework was proposed (Xie et al., 2020) to improve image classification performance. Several noise and corruption models have been considered, including both physics-based (Wei et al., 2020) and learning-based (Abdelhamed et al., 2019). Efforts in capturing real datasets of noisy images have also been pursued. A dataset of images captured in low light with annotations for object detection (Loh and Chan, 2019) has been collected. Another example is the dataset containing low-light and corresponding well-lit cellphone images for denoising (Abdelhamed et al., 2018), which has recently been extended to videos in (Wang et al., 2021b). Most previous works simulate or collect real captures with image degradations like noise in low-light, but we consider a more challenging and practical setting where both low-light and motion are present, and hence dual image degradation comes into play.

Noise Removal and Deblurring: Due to its importance in image processing, denoising and/or deblurring degraded images has been a very popular topic for decades. Recently, numerous works have been proposed using neural networks for deblurring (Xu et al., 2014; Schuler et al., 2015; Zhang et al., 2017a) and denoising (Zhang et al., 2017b). For example, a sparse denoising auto-encoder was considered for robust denoising (Agostinelli et al., 2013). A recent line of work proposes to perform joint denoising and inference on noisy images (Liu et al.,

2019, 2020a; Diamond et al., 2017). While existing image restoration methods can obtain high quality reconstructions, performing inference directly on the corrupted images does not require any pre-processing and is thus more efficient and as we demonstrate, can achieve increased robustness under severe image degradation. Alternatively, other methods aim to design cameras that produce better images directly, either by optimizing the hyperparameters of existing image signal processors (ISP) (Tseng et al., 2019) or, by designing novel ISPs (Heide et al., 2014; Gharbi et al., 2016; Chen et al., 2017, 2018). These methods may, however, not entirely remove noise in challenging low-light situations, due to the fundamental limitation of the optics and sensors.

Inference on Corrupted Images: Many recent works tackle different inference tasks *directly* on images with common corruptions. Rozumnyi *et al.* (Rozumnyi et al., 2021) proposed a matting and deblurring network for faster inference for the detection of fast moving objects in videos. Cui *et al.* (Cui et al., 2021) designed a multitask auto-encoder for image enhancement, which leverages a physical noise model and ISP setting in a self-supervised manner to improve detection performance. Wang *et al.* (Wang et al., 2021a) presented a framework for monocular depth estimation under low-light using self-supervised learning and demonstrate their results on nighttime datasets. Others have used knowledge distillation techniques for image classification under low-light (Gnanasambandam and Chan, 2020), or for object detection by leveraging bursts of short exposure frames (Li et al., 2021). Photon Net (Goyal and Gupta, 2021) used a single photon camera and proposed to train on a wide spectrum of images at various SNR, with encouraging results on image classification and monocular depth estimation. Song *et al.* (Song et al., 2021) introduced a technique for image matching using local descriptors and initial point-matching methods for extremely low-light images in RAW format. Wang *et al.* (Wang et al., 2020) proposed to learn the mapping relationship between representations of low and high quality images, and used it as a deep degradation prior (DDP) for image classification on degraded images. Adversarial Logit Pairing (Kannan et al., 2018) also provides

some robustness to the inference on noise and blur corruptions (Hendrycks and Dietterich, 2018) by matching the logits output of a clean image with an adversarial perturbed image.

Our goal is different from all previous approaches. We propose techniques that leverage the space of noise-blur dual corruptions rather than looking at a single image corruption. We show that our approach is versatile for several downstream tasks, including image classification and object detection.

Leveraging Multiple Captures: Multiple exposures can be used to reconstruct high dynamic range (HDR) images (Debevec and Malik, 1997), even in the presence of motion (Sen et al., 2012; Kalantari and Ramamoorthi, 2019). Hasinoff *et al.* (Hasinoff et al., 2009, 2010a) proposed ways to select settings for these multiple captures, like ISOs and focus settings. The popularity of mobile photography has led to the further development of burst photography (Hasinoff et al., 2016), which has been used for denoising (Mildenhall et al., 2018), deblurring (Delbracio and Sapiro, 2015; Aittala and Durand, 2018), and super-resolution (Wronski et al., 2019). In sharp contrast, we exploit multiple exposures for high-level inference tasks such as classification and detection, rather than low-level image reconstruction.

5.2 Scene Inference under Noise-Blur Dual Corruptions

We consider scene inference tasks represented as an inference module $f(x) \equiv g \circ \varphi(x)$, where, without loss of generality, $\varphi(x)$ is a feature extractor, and g is a prediction module. Here, $\circ$ is the composition operator. $f(x)$, oftentimes represented by a neural network, maps an input image x into its semantic label y. This generic formulation covers several vision recognition tasks, including image classification, where y is a categorical label, and object detection, where y is a set of labeled bounding boxes. We further assume that this function $f(\cdot)$ is

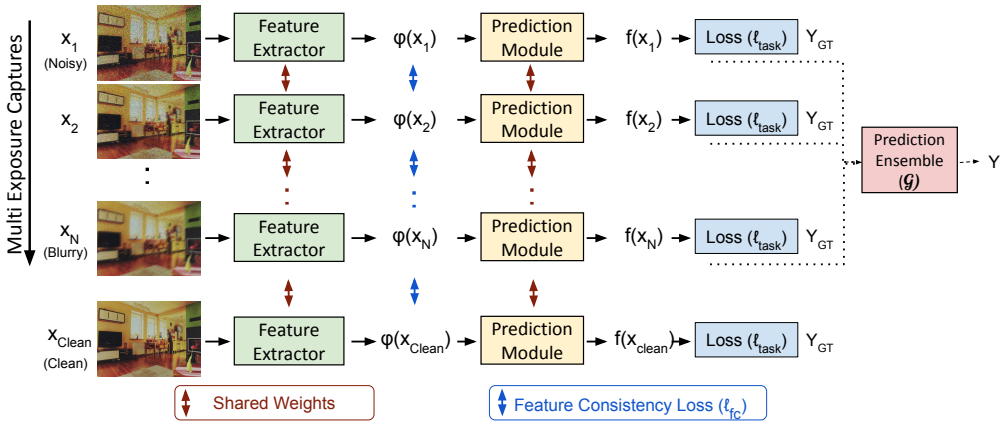


Figure 5.2: **Architecture Overview:** Our approach trains an inference model using multiple captures of varying exposures, all containing the same semantic content but different amounts of noise-blur dual corruptions. We introduce feature consistency loss during training to enforce the consistency of feature outputs from individual captures. During testing (dashed lines), our model returns the ensemble prediction using each individual capture to produce the final output for a more robust prediction.

learned from data by minimizing a certain loss function.

Given a set of N noise-blur dual corruption images $\mathcal{X} = \{x_1, ...x_N\}$ capturing the same scene, our key intuition is that despite differences in low-level *image* features (*e.g*, pixel values), their *latent* features should remain similar. In what follows, we formulate this intuition as a data prior, devise the training and inference schemes, and demonstrate interesting properties of the resulting method.

Robust Inference with Multiple Exposures

A simple prior is to assume that the latent features $\{\varphi(x_1), ...\varphi(x)_N\}$ follow a Gaussian distribution, centered at the ideal clean image x_{clean} and with a small variance ϵ^2. This prior ensures that with high probability the distance between any pair of latent features will stay in a small ℓ_2 radius controlled by ϵ^2. With such assumption, we arrive at the following conditional probability $p(y|x)$ for

scene inference.

$$p(y|x) \propto p(y|\varphi(x))p(\varphi(x)|x), \tag{5.1}$$

where $p(\varphi(x)|x) \sim \mathcal{N}(\varphi(x_{clean}), \epsilon^2)$ represents the data prior and $p(y|\varphi(x))$ is given by the prediction module g. We now describe the training and inference schemes based on this formulation, as illustrated in Figure 5.2.

Training with Multiple Exposures: Given the ground-truth label y, minimizing the negative log likelihood of Equation 5.1 on a training sample (a set of images $\{x_i\}$ spanning the dual corruption space) leads to the following loss function

$$\ell = \sum_{i}^{N} \ell_{task}(p(y|\varphi(x_i)), y) + \frac{1}{\epsilon^2} \sum_{i}^{N} \|\varphi(x_i) - \varphi(x_{clean})\|_2^2. \tag{5.2}$$

Here, we slightly abuse the notation to replace the first term - $\log(p(y|\varphi(x)), y)$ with a more general task-specific loss $\ell_{task}(p(y|\varphi(x)), y)$. It is easier to consider the case of image classification, where the target y is a categorical variable. The term of $-\log(p(y|\varphi(x)), y)$ becomes the cross-entropy loss, commonly used for classification. When y moves beyond simple categorical or scalar outputs (*e.g*, for the object detection task), Equation 5.2 allows to plug in any loss function ℓ_{task} tailored for the task. On the other hand, the second term can be interpreted as a feature consistency loss, re-weighted by a coefficient as the reciprocal of the Gaussian variance ($1/\epsilon^2$).

Our loss function in Equation 5.2 assumes that a reference clean image is available during training, as is often the case in our experiments. When such a clean image is not presented, we simply replace the second term with its equivalent form that only involves the summation of pairwise distances between $\varphi(x_i)$ and $\varphi(x_j)$, *i.e*, $\frac{1}{2N\epsilon^2} \sum_{i,j} \|\varphi(x_i) - \varphi(x_j)\|_2^2$.

Inference with Model Aggregation: At inference time, the maximum likelihood estimation of Equation 5.1 is not viable without the clean image x_{clean}. Instead,

we resort to using the ensemble of the predictions from individual multi-exposure images as the final output prediction. Our key intuition is that no individual capture in the dual corruption space captures all the necessary information that may be required for robust inference, but the ensemble output is more effective as it uses the predictions from individual images that contribute the relevant information individually. This is given by

$$f(\mathcal{X}) = \mathcal{G}(f(x_1), f(x_2)...f(x_N)), \tag{5.3}$$

where $\mathcal{G}$ is an aggregate function to get the ensemble prediction. $\mathcal{G}$ is highly flexible and often task-relevant. For example, for the image classification task, $\mathcal{G}$ could be a simple average operator over the probability outputs. For object detection, $\mathcal{G}$ might be a voting scheme of detected objects. By aggregating multiple model outputs, Equation 5.3 is conceptually similar to the well-known model ensemble (Rokach, 2010).

Certified Robustness: When considering a classification problem with c categories (*e.g*, image classification), we notice an interesting link between our inference scheme and a well-known robust classifier (Cohen et al., 2019). Specifically, when $\mathcal{G}$ is an average operator and the decision is made by taking the category with the highest confidence from $f(\mathcal{X})$, our inference defines a "smoothed" classifier with certified robustness (Cohen et al., 2019) under the Gaussian distribution

$$\begin{aligned} &\arg\max\ p(g(\hat{\varphi}(x)) = c), \\ &\text{where}\quad \hat{\varphi}(x) \sim \mathcal{N}(\varphi(x_{clean}), \epsilon^2). \end{aligned} \tag{5.4}$$

(Cohen et al., 2019) showed that such a classifier, if it passes additional certification, is robust within a certain ℓ_2 radius around $\varphi(x_{clean})$. Intuitively, this indicates that our model will produce consistent results (the same as ones given by the clean image) for all corrupted images spanning the dual corruption space, should the Gaussian assumption be satisfied. We deem a theoretical investigation

in this direction as our future work.

5.3 Evaluation of Robust Scene Inference

We demonstrate the effectiveness of our method on two important scene inference tasks: object detection and image classification.

Object Detection

Instantiation: Figure 5.2 shows the overview of our approach using a multi-exposure ensemble for the object detection task. We implement our approach using the single-stage FCOS architecture (Tian et al., 2019b). The output prediction of the FCOS model for image of size $H \times W$ consists of pixel-wise classification scores ($H \times W \times C$) for C object categories, centerness scores ($H \times W \times 1$) and bounding box coordinates regression outputs ($H \times W \times 4$). During inference, our ensemble predictor ($\mathcal{G}$), takes the pixel-wise classification scores, centerness scores, and box coordinates, and returns their average at each FPN level. Loss function for the inference task (ℓ_{task}) is the same as defined in FCOS architecture (i.e., sum of focal loss, regression loss for bounding boxes, and centerness loss). Refer (Tian et al., 2019b) for details. Our feature consistency loss (ℓ_{fc}) is the L2 distance between feature outputs from the CNN network (final layer after global average pooling).

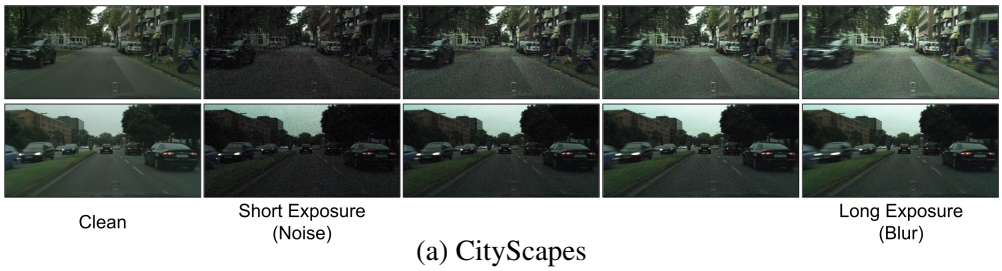
Datasets and Metrics: We evaluate our approach using three object detection datasets: Cityscapes (Cordts et al., 2016), MS-COCO (Lin et al., 2014), and REDS (Nah et al., 2019). Cityscapes consists of street scenes captured from a vehicle and consists of 8 categories related to autonomous driving, with 2975 training and 500 test images. MS-COCO consists of 80 categories for general object detection with 118k training and 5k validation images. REDS consists of 120fps video sequences of 270 scenes captured by a high-speed camera. The dataset represents images with common objects (like person, car, chair etc.).

The ground truth annotations provided in Cityscapes and MS-COCO are used for evaluation. We follow common conventions, train our models on their training sets, and report results on the validation sets. In contrast, REDS does not have object annotations. We thus use a pretrained Faster R-CNN object detector model (Ren et al., 2015) available in the Detectron2 platform (Wu et al., 2019) to obtain pseudo-ground truth annotations to create our evaluation benchmark containing 270 images with 2160 box annotations.

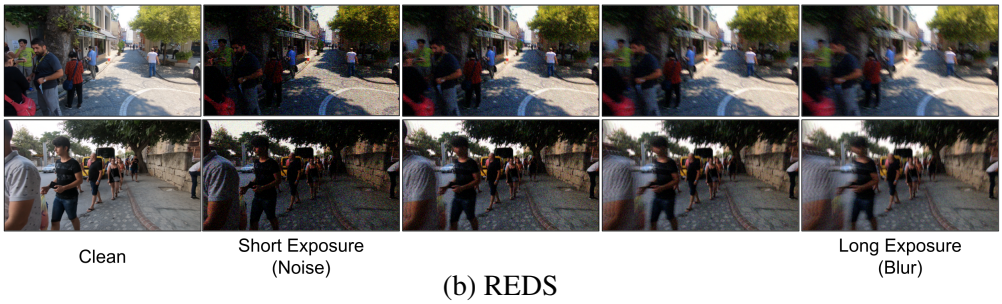
All results are reported using mean average precision (mAP) across multiple intersection-over-union (IoU) thresholds, following the COCO evaluation protocol (Lin et al., 2014).

Low-light and Motion Blur Dataset Generation: All three datasets mentioned above contain images captured in sufficient light and no noticeable motion blur (scene or camera). Since there is no publicly available large-scale annotated dataset containing images captured in low-light and motion blur conditions, we simulate such conditions using various strategies, as described below.

- *REDS*: Since the REDS dataset contains video sequences captured by a 120fps camera, we first simulate low-light conditions for each individual frame of the sequence by adding Poisson noise (shot noise) and read noise. Multiple frames are then averaged together to generate images with motion blur that capture realistic motion conditions of the camera or scene. In practice, we select a random frame from each video sequence, select a varying number of adjacent frames (from 0 to 3 on each side of the frame), and compute their average (after adding noise) to simulate blurry images with motion. This generates images with different exposures, examples of which are shown in Figure 5.3b.
- *CityScapes*: CityScapes provides low-fps video sequences around each annotated frame in the dataset (30-frame sequence captured at 17fps). We use a pretrained video interpolation network (Sim et al., 2021) to synthesize a high-fps video sequence by increasing the frame rate by a factor of 4x. A motion-blurred image is then generated as with the REDS dataset, that is adding noise to each individual frame, and averaging a number of adjacent frames.



(a) CityScapes



(b) REDS

(c) MS-COCO

(d) CUB-200-2011

Figure 5.3: **Simulated Images:** Few examples of images with simulated noise and blur. CityScapes and REDS dataset images are generated by simulating low-light frames from high-speed video sequences. MS-COCO and Birds dataset images are generated using a single frame by adding noise (shot noise and read noise) and blur (random motion blur kernel) of varying amounts.

Figure 5.3a shows examples of simulated low-light and motion-blur frames used for training and evaluation on the CityScapes dataset. The resulting images indeed represent realistic motion conditions under autonomous driving scenarios (like fast moving camera/car or moving pedestrians, other vehicles etc.).

- *MS-COCO*: As the MS-COCO dataset does not contain any video sequences, we simulate the blur and noise from a single image using the same procedure as the image corruptions benchmark in (Hendrycks and Dietterich, 2018) by selecting varying severity of shot noise and motion blur. Specifically, the noisiest image has a shot noise level of 4 and a motion blur level 1. Subsequent levels in the dual corruptions are simulated by increasing the motion blur and decreasing the shot noise successively to generate 4 levels of dual corruptions. Figure 5.3c shows a few examples of simulated images. We note that, contrary to the other two datasets above, the blur simulated by this approach is not spatially varying.

Baselines: We compare our approach with the following set of baselines. All approaches use the same backbone for fair comparison. We evaluate all the methods using all four exposures and report the results for the best exposure settings.

- *Clean Model*: This baseline model is trained only on clean images and evaluated on noisy images.
- *Stylized Training*: We follow the data augmentation approach of (Michaelis et al., 2019), who propose to augment training images with stylization for robustness.
- *Single Exposure*: We train a model on a dataset containing varying exposures and clean images, essentially considering distortions as a way to perform data augmentation (Hendrycks and Dietterich, 2018). For evaluation, we select the single exposure setting yielding the best performance and report those results. This baseline acts as an oracle for selecting the best-performing exposure time at inference time.

Method	REDS				CityScapes				MS-COCO			
	mAP	APs	APm	APl	mAP	APs	APm	APl	mAP	APs	APm	APl
Clean Model	16.36	17.96	18.46	16.45	2.72	0.22	2.47	7.02	3.35	0.21	2.51	7.69
Stylized Training	19.13	18.11	21.64	23.71	6.75	0.24	3.32	20.00	7.89	0.25	3.13	17.07
Single Exposure	30.17	20.27	25.75	36.88	18.07	3.96	15.77	35.54	21.25	6.58	22.39	33.88
Denoising (BM3D)	30.25	20.28	25.90	37.08	18.01	3.82	15.53	35.97	21.78	6.76	22.78	34.43
Denoising (MPRNet)	25.67	18.97	23.47	31.84	15.26	2.97	13.89	34.11	18.78	5.12	17.45	27.13
Deblurring	30.68	18.82	26.36	36.02	17.67	3.63	15.90	34.67	12.42	2.52	11.77	21.11
Denoising + Deblurring	29.45	18.46	26.35	34.46	17.91	4.01	15.34	35.09	22.03	6.79	22.89	34.63
Short Exposures (N = 4)	30.81	18.41	26.53	36.02	18.46	4.33	15.97	35.86	22.17	6.87	23.91	35.11
Ours (N = 2)	33.76	14.67	27.64	40.81	19.36	5.11	17.23	37.66	23.11	8.01	25.87	36.09
Ours (N = 4)	**36.17**	14.15	**29.04**	**42.17**	**20.97**	**5.38**	**19.46**	**38.95**	**24.71**	**9.13**	**27.08**	**37.79**

Table 5.1: **Object Detection Results**: AP results on REDS, MSCOCO, and CityScapes datasets. Our approach of Multi-Exposure Ensemble (Ours) outperforms all baselines.

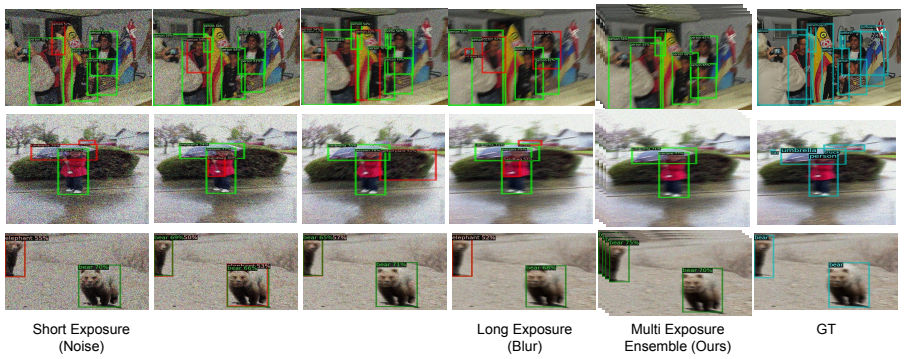
- *Denoising*: This baseline represents the conventional approach of denoising the noisy images under low-light conditions. We perform both training and inference on denoised images. Here, we experiment with the BM3D (Dabov et al., 2007) and MPRNet (Zamir et al., 2021) approaches for denoising the images.
- *Deblurring*: We also compare with the approach of deblurring the images for scene inference, where we use a deblurring model (Carbajal et al., 2021). We perform both training and evaluation of our model using deblurred images.
- *Denoising + Deblurring*: As the test images in low-light and motion blur have both noise and blur, we also compare with the approach of denoising (BM3D) followed by deblurring. The model is trained and evaluated using Denoised+Deblurred images.
- *Short Exposures*: This baseline compares with the approach of evaluating using multiple short exposures by using the ensemble prediction from N short exposure images. The model is trained with short exposure images.

Implementation Details: We used the official implementation of the FCOS architecture (Tian et al., 2019a) for the object detection experiments, which is based on the Detectron2 framework (Wu et al., 2019). ResNet-50 (He et al., 2016)

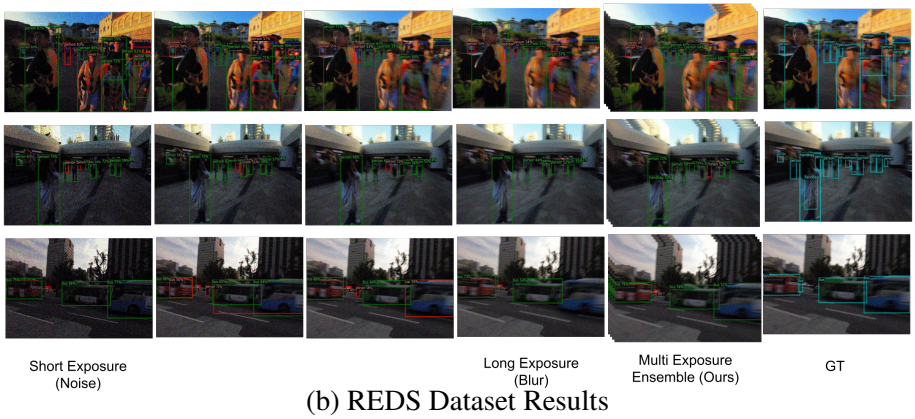
with FPN was used as the backbone for training and initialized with ImageNet pretraining weights for all our models. We followed the hyperparameters from Detectron2 to train our models. MS-COCO models were trained with a learning rate of 0.01, batch size of 16 for 90k iterations, whereas CityScapes models were trained with a learning rate of 0.005, batch size of 8 for 24k iterations. REDS is used only for evaluation; in this case, we use the model trained on MS-COCO.

Results and Discussions: Table 5.1 shows the results (in mAP along with AP of small, medium and large objects) of our approach on all three datasets. Our method outperforms all baselines by a significant margin. Our approach beats Single Exposure baseline by 6% in REDS, 2.9% in CityScapes, and 3.5% in MS-COCO with four exposures. In other words, it is best to leverage all the dual-corruption images even if we knew the best possible single exposure ahead of time. Denoising provides improvements over Single Exposure baseline in some cases but is not as effective. Deblurring approaches does not show performance improvement over Single Exposure baseline in most cases. This is because images contain both noise and blur and deblurring models are specialized to handle only blur. Deblurring+Denoising baseline also shows relatively minor performance gain. We see significant gain with Short Exposures (with 4 exposures) baseline, highlighting the benefit of ensemble prediction. However, since all the exposures are short, they all suffer from sever noise and have similar errors, and hence outperformed by our method. Our method provides large improvements even with two exposures, and increasing the number of exposures (from two to four) further increases the performance. This highlights that our approach benefits with more number of exposures as different exposures have a wide variety of dual corruption level.

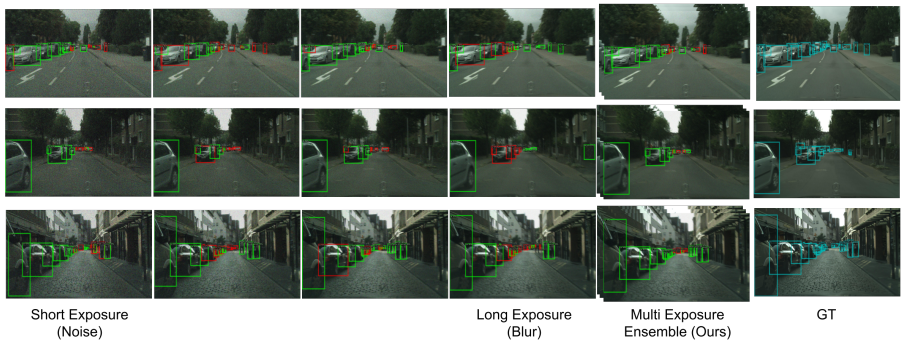
Figure 5.4 shows representative qualitative examples of our approach for object detection and shows direct comparison with each individual exposure and its predictions. The correct/incorrect bounding boxes are highlighted in green/red and ground truth bounding boxes are highlighted in blue on the clean image (right). Our approach makes fewer false positive predictions (red) compared to the



(a) MS-COCO Dataset Results



(b) REDS Dataset Results



(c) CityScapes Dataset Results

Figure 5.4: **Object Detection Results** for MS-COCO, REDS and CityScapes Dataset. Correct/Incorrect predictions are highlighted with green/red, and ground truth boxes are highlighted with blue in the clean image. The first 4 columns show results on single captures, followed by a column with results from multi-exposure captures using our approach. Single Captures have a lot more false positives (red) while our approach effectively removes those cases (Better viewed on screen).

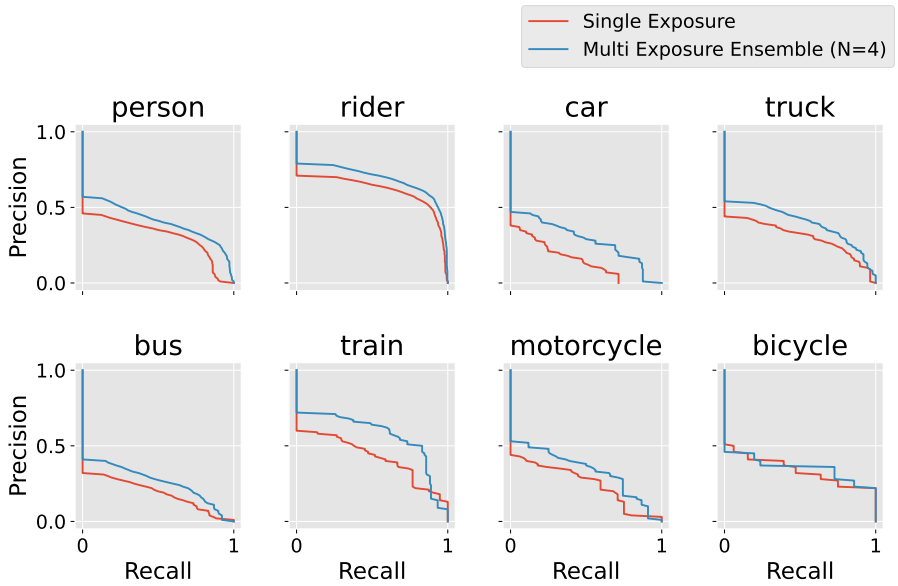


Figure 5.5: **Precision Recall Curve** of our approach and baselines on CityScapes Dataset for all 8 categories with IOU threshold of 0.5. We see significant improvement for 'person' and 'car' categories, which are the most common in the dataset.

Single Exposure. Since individual single captures make different false positive predictions, the ensemble is able to remove those false positive boxes. Figure 5.5 shows the precision recall curve on CityScapes dataset for IOU threshold of 0.5 for all 8 categories in the dataset. We see a significant improvement in area under the curve for person and car category, which is the most common in the dataset.

Image Classification

Instantiation: Similar to object detection, our approach uses a shared CNN architecture as a feature extractor. In particular, we used a ResNet-18 (He et al., 2016) as the image classification architecture. The feature consistency loss ℓ_{fc} is defined as the L2 distance between the feature map output of the final layer (after global average pooling) to encourage consistent predictions. The model returns the average of the predictions (probability output) from multiple degraded images

Method	Top-1	Top-5
Clean Model	6.13	13.45
Stylized Training [(Michaelis et al., 2019)]	9.51	17.83
Single Exposure	41.18	64.13
Denoising (BM3D) [(Dabov et al., 2007)]	43.34	67.11
Deblurring [(Carbajal et al., 2021)]	39.13	60.45
Denoising [(Dabov et al., 2007)] + Deblurring [(Carbajal et al., 2021)]	42.95	66.59
Short Exposures (N = 4)	45.16	69.84
Multi Exposure Ensemble (N = 2)	52.10	74.13
Multi Exposure Ensemble (N = 4)	**55.27**	**79.34**

Table 5.2: **Image Classification Results**: Top-1 and top-5 accuracy results on CUB-200-2011 dataset. Our approach of Multi-Exposure Ensemble outperforms all the baselines.

(as the ensemble operator $\mathcal{G}$) for the final output.

Datasets, Metrics, and Baselines: We use simulated images from the CUB-200-2011 image classification dataset (Wah et al., 2011). 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. All results are reported using top-1/5 accuracy on the test set, following the standard evaluation protocol for image classification. A set of baselines similar to the ones used in the experiments on object detection (Section 5.3) is considered here.

Simulating Noise and Blur: Since CUB only contains single images, we employ the same strategy to generate dual corruption images as for the MS-COCO dataset in the object detection experiments (Section 5.3). Figure 5.3d shows a few examples of simulated images.

Implementation Details: The model is trained with SGD with the momentum of 0.9, a base learning rate of 0.1 with cosine decay, and a batch size of 32 is used to train for 100 epochs.

Results and Discussions: Table 5.2 shows top-1 and top-5 accuracy of our approach on the simulated CUB dataset. We report the results of our model using

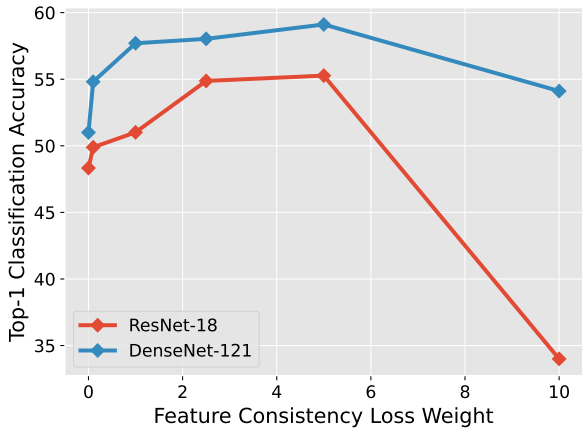


Figure 5.6: **Ablation Studies:** Image classification results of our approach on CUB-200-2011 while varying feature consistency loss weight and backbone architecture.

two and four exposure settings. Our method outperforms both baselines using a single exposure by a significant margin. Compared to choosing the single best exposure, our approach, with N $= 4$, attains an overall gain of 14.1% and 15.2% in top-1 and top-5 accuracy, respectively. Our approach shows significant gains with only two exposures; however, having more number of exposures (from 2 to 4) further helps the overall performance.

Ablation Studies: We study the performance of our approach with varying weights for feature consistency loss. Figure 5.6 shows that our approach performs best for the weight factor of 5 in image classification on the CUB-200-2011 Dataset. We also evaluate the performance of our approach with another backbone architecture. Figure 5.6 shows similar performance gain using DenseNet-121 (Huang et al., 2017), which highlights the versatility of our approach as it can extend to different CNN feature extractors.

Figure 5.7: **Camera Setup** for capturing multiple exposure images

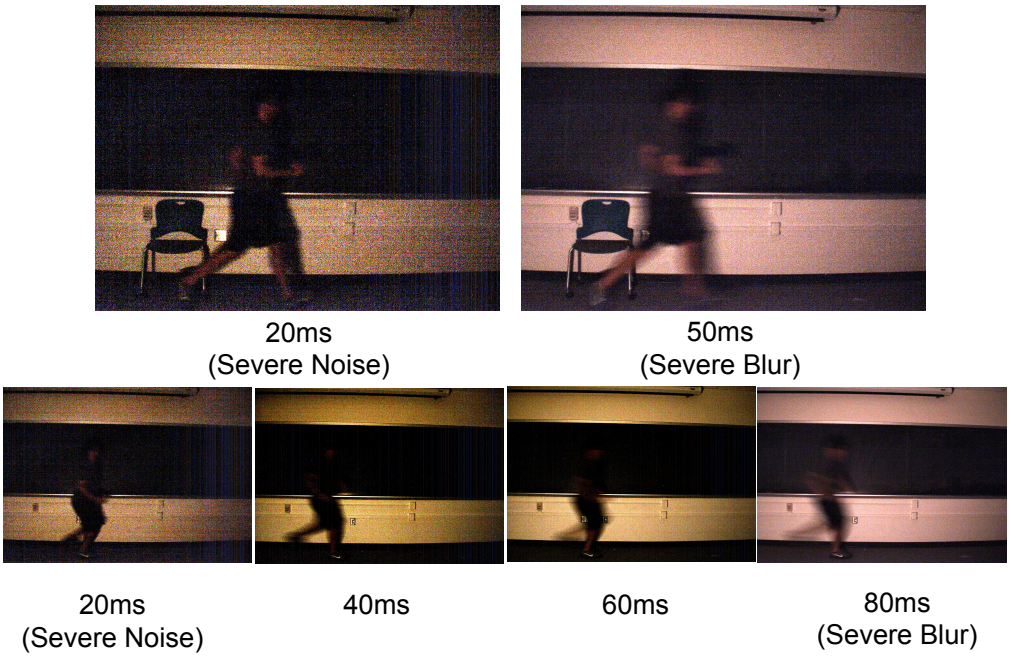


Figure 5.8: **Examples of Real Captures:** Images captured with varying exposure settings with our multi-camera setup. Images with shorter exposure have severe noise, while images with longer exposure contain motion blur for the moving objects.

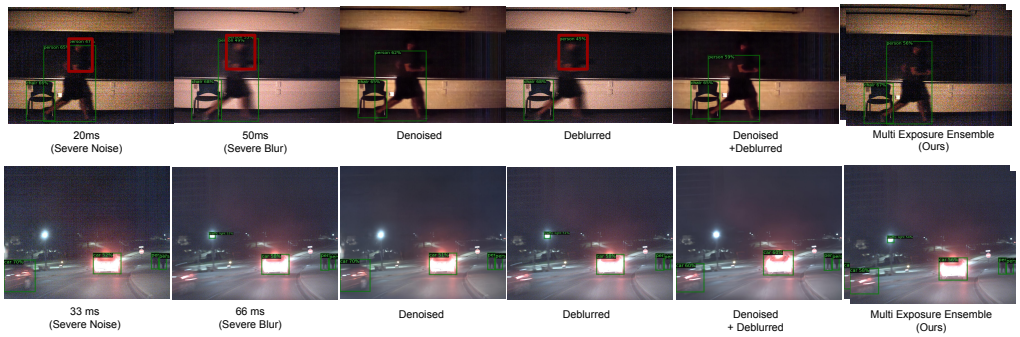


Figure 5.9: **Object Detection Results on Real Captures:** Scene in the first row contains an indoor scenario with two objects: a person (moving) and a chair (stationary). Single Exposures are severely affected by noise and/or blur: detect false positives or inaccurate bounding boxes. Scene in the second row contains a driving scenario with a car (moving) on the left and a traffic light (stationary) in the front. Single Exposures fail to detect the moving car or the stationary traffic light. The Multi-Exposure Ensemble approach (right) leverages multiple exposures and detects all objects with correct labels and tight bounding boxes in both scenes.

5.4 Experiments with Real Captures

Finally, we evaluate our approach on real images by capturing multiple simultaneous exposures of the same scene.

Camera Setup: Our setup includes four BlackflyS USB3 cameras (Teledyne, 2022) by Teledyne Flir. These are machine vision cameras that can capture colored images with a resolution of 1280×1024 with up to 175 frames per second. Same lenses (Tamron 8mm) are used for all cameras, which are stacked together to get similar (overlapping) fields-of-view. Aside from an approximate physical alignment of the cameras, no further alignment of the captured images is done as all cameras have similar fields-of-view, and the scene is sufficiently far away. Cameras are connected to a computer that triggers the simultaneous captures (software sync). Our complete setup is shown in Figure 5.7.

We use Spinnaker SDK (Spinnaker, 2022) provided by Teledyne to capture RAW images. Maximum available gain (18dB) for the camera is used, and a

gamma correction ($\gamma = 2.2$) is applied on the captures to get the final images. We set different exposure times for each camera and synchronously capture images using all the cameras.

Exposure Selection: We manually select the exposure times in order to span a wide range of exposures while ensuring that images are not too under- nor over-exposed. Our indoor scenes consist of fast-moving objects in a very dark environment (~0.25lux) lit by a single light source. We experiment with multiple settings depending on the lightning conditions, including A) 20-30-40-50ms, B) 20-40-60-80ms, and C) 16-33-66-100ms. When evaluating our approach, we use two or four exposures, examples of which are shown in Figure 5.8.

Results and Discussions: We train our object detection models with the simulated images from the MS-COCO dataset and evaluate the trained model on real captures. Figure 5.9 shows sample results with the real captures on two scenes. Both scenes consist of both fast-moving and stationary objects under low light. The prediction output from the individual exposure contains several false positives and inaccurate boxes. By leveraging the multiple exposures across the space of dual corruptions, our method is able to correctly detect all the objects with tight bounding boxes and remove false positive boxes.

Our approach performs better inference even with two exposures ($N = 2$). As we increase the number of exposures, the prediction improves as long as the exposures are not too noisy or blurry for inference (as that can deteriorate the performance of the ensemble prediction). One simple heuristic that performs well with our approach is to select exposure times around the *auto-exposure* value, as this ensures the frames are not too under- or over-exposed. We show more examples in the supplementary text with two and four exposures, including failure cases.

5.5 Conclusion

Our work demonstrates the challenges in scene inference under low-light and motion conditions. We discuss the trade-off between two kinds of image degradations: motion blur (due to long exposure) vs. noise (due to short exposure), also referred to as a dual image corruption pair in this paper. To this end, we propose a method to leverage multi exposure captures for robust inference under low-light and motion. Our method builds on a feature consistency loss to encourage similar results from these individual captures and uses the ensemble of their final predictions for robust visual recognition. We demonstrate the effectiveness of our approach on simulated images as well as real captures with multiple exposures, and across the tasks of object detection and image classification.

5.6 Supplementary Section: Additional Object Detection Results

Comparison to Baselines: We compare our approach with additional baselines. Table 5.3 shows the performance of the model trained and evaluated on clean images. We also show the results of training and testing with a single corruption level. Results are included for four different noise-blur dual corruption levels (from 1 to 4) with increasing motion blur and decreasing shot noise. Comparing with clean images shows the impact of noise and blur degradation as the mAP drops significantly. Our approach utilizes clean images and corrupted images with feature consistency that helps the model learn robust features. Our model outperforms these baselines by a significant margin using the same model capacity.

Results Visualization for Object Detection: Figure 5.10 shows examples where our approach outperforms the baselines. The first row of Figure 5.10 shows an example where one baseline predicts correct bounding boxes, and our approach is as good as the best single-exposure baseline. Figure 5.11 shows a few result images with real captures using our approach and the baseline. Our method

Method	REDS				CityScapes				MS-COCO			
	mAP	APs	APm	APl	mAP	APs	APm	APl	mAP	APs	APm	APl
Clean Training & Testing	78.21	52.94	73.91	84.33	33.36	10.40	32.26	54.70	38.59	22.9	42.28	49.56
Corruption Level 1 (Severe Noise)	23.46	16.14	24.08	26.27	14.06	1.82	11.80	30.98	20.26	6.18	21.18	32.73
Corruption Level 2	30.20	20.27	25.75	36.88	17.19	3.71	15.36	33.84	20.29	5.59	21.19	32.21
Corruption Level 3	27.85	19.75	23.80	35.09	17.07	3.26	15.45	32.89	20.21	6.35	20.94	32.70
Corruption Level 4 (Severe Blur)	26.78	15.51	20.13	33.53	15.94	4.21	14.73	30.39	20.47	6.45	20.94	32.32
Multi-Exposure Ensemble ($\mathrm{N} = 4$)	**36.17**	14.15	**29.04**	**42.17**	**20.97**	**5.38**	**19.46**	**38.95**	**24.71**	**9.13**	**27.08**	**37.79**

Table 5.3: **Object Detection Results**: AP results on REDS, MSCOCO, and CityScapes datasets.

is more effective in predicting the correct bounding boxes and predicts fewer false-positive boxes.

Failure Cases for Object Detection: Figure 5.12 and 5.13 show some failure cases where our approach performs worse than a single-exposure baseline. Since our approach relies on the average of output predictions, it fails to perform well when one of the exposures has too much degradation.

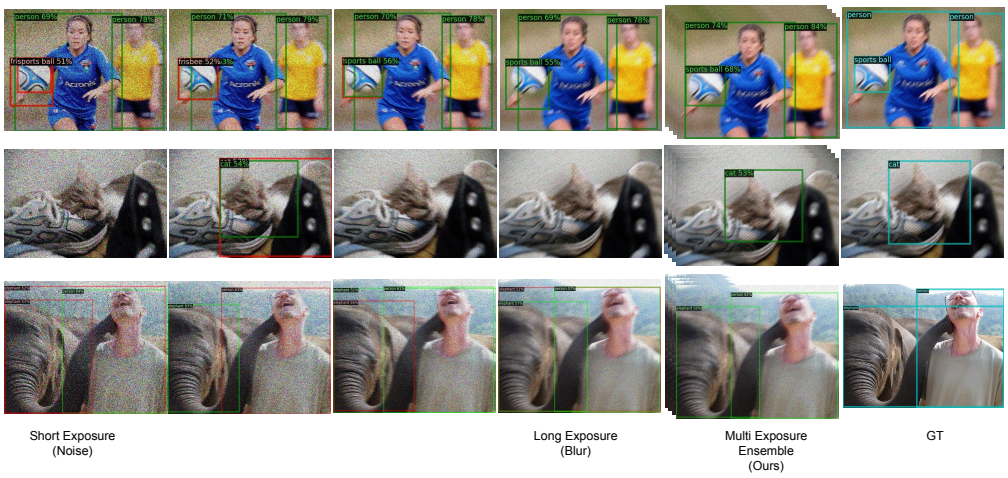


Figure 5.10: **Object Detection Results on MS-COCO dataset:** Correct/Incorrect predictions are highlighted with green/red, and ground truth boxes are highlighted with blue in the clean image. Single Exposures have a lot more false positives (red) while our approach effectively removes those cases. For the first scene, our approach produces tighter bounding boxes than individual predictions (Better viewed on screen).

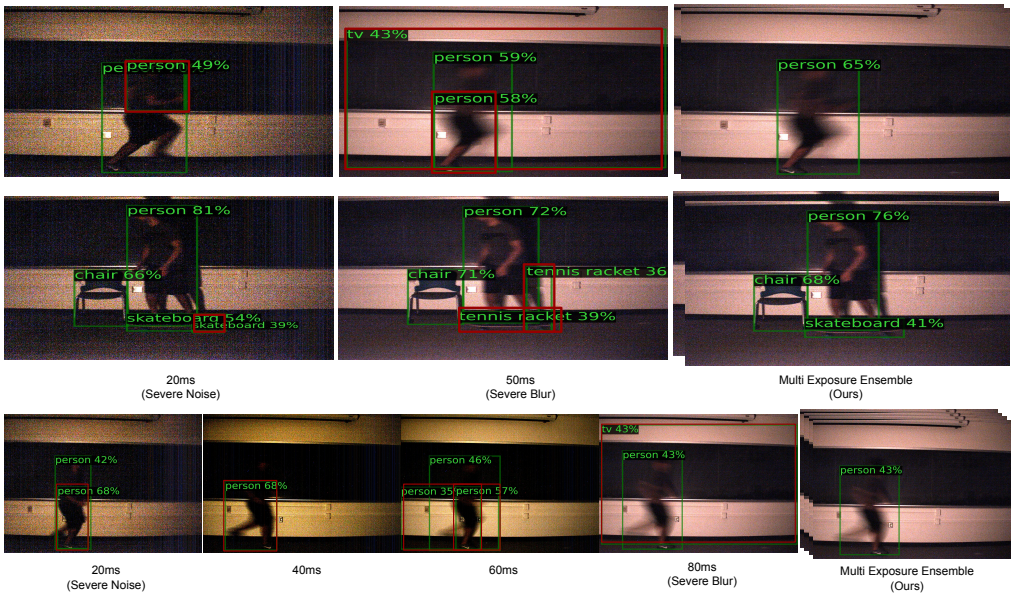


Figure 5.11: **Object detection results with Real Captures:** Single Exposures are severely affected by noise and/or blur. The model detects false positives and inaccurate bounding boxes. The Multi-Exposure Ensemble approach (right) leverages multiple exposures and detects all objects with correct labels and tight bounding boxes.

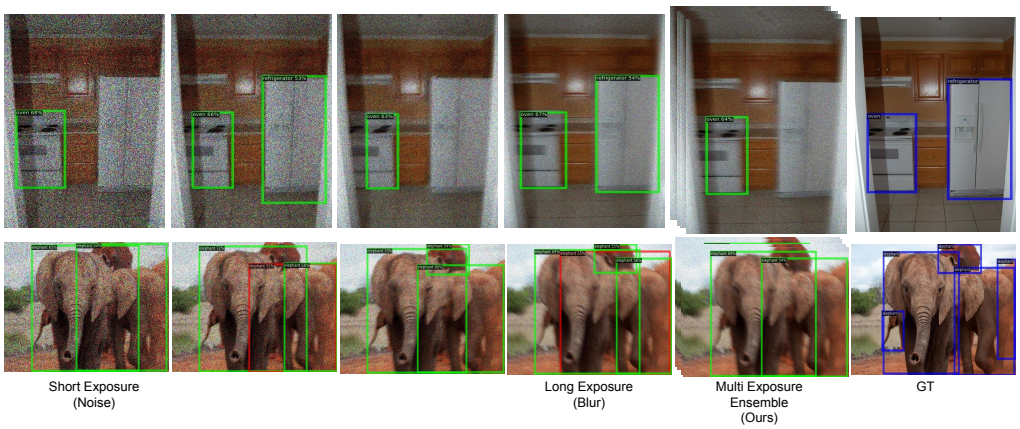


Figure 5.12: **Object Detection Failure Cases on MS-COCO dataset:** Figure shows examples where single exposure performs better than our approach. The first scene contains two objects, and our approach fails to detect the second object. The second scene contains a lot of overlapping ground truth bounding boxes, and our approach fails to detect a few bounding boxes.

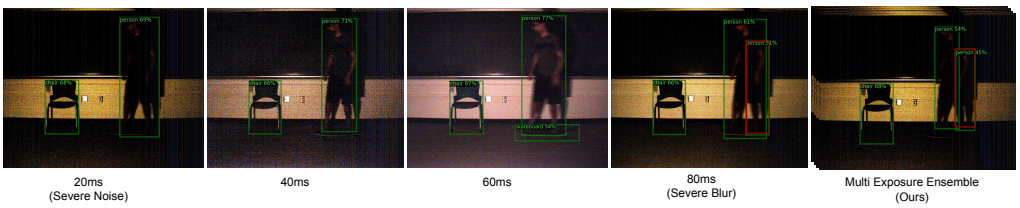


Figure 5.13: **Object Detection Failure Cases on Real Captures:** Figure shows a failure case with the real captures where a single exposure (60ms) detects all three objects correctly, whereas our model a detects false positive box and fails to detect skateboard object. Our model performs worse in cases when any single exposure has too much degradation.

6 CONCLUSION AND FUTURE OUTLOOK

The robustness of perception models on challenging scenarios such as low-light, motion, and extreme weather is an important factor and often a bottleneck to the safe deployment of vision systems. This dissertation provides a perception framework that is capable of performing robust inference under a variety of challenging conditions and for multiple sensing modalities. We discuss numerous scenarios, such as autonomous driving, where the reliable performance under these non-ideal imaging conditions is critical. Our proposed framework improves the *worst-case* performance of such vision systems significantly for sensors including SPAD LiDARs, single photon cameras, and conventional RGB cameras.

I would now conclude this dissertation by discussing some limitations of our work and providing some future outlook on approaches that could further improve this goal.

6.1 Scope and Limitations

Learned Confidence Measure

In our proposed method, we handcrafted an effective confidence measure from raw LiDAR measurements. Although our proposed confidence measure shows significant performance gains, our work does not explore other possible confidence attribute formulations. Finding an optimal confidence measure for downstream inference could be helpful in further performance improvements.

Generalizability of our confidence attribute across different LiDAR sensors is also an important consideration. A learned representation of the confidence attribute could be useful for better generalizability as well.

Exposure Selection for Multiple Captures

Our proposed approach of multiple captures uses a manually chosen set of exposure times while capturing. Most modern cameras, however, have the functionality of *auto-exposure* that selects the exposure setting based on the lighting and motion conditions (light and motion metering) of the scene for the best image quality. Determining the optimal exposure for inference automatically (for a single exposure) is an active area of research (Onzon et al., 2021). With the ability to capture multiple exposures, an important research problem is to develop generalized auto-exposure techniques for *multiple captures* that result in the best performance for the inference tasks under these challenging conditions.

Computational Considerations

Capturing, processing, and performing inference on multiple exposures incurs a linear increase in computational cost. However, since many of these computations can be done in parallel, the increase in latency is small, which is important for safety-critical applications like autonomous driving. Our approach is agnostic to the number of exposures during inference, which allows inference systems to switch between multi-exposure settings (during challenging conditions of low-light and/or motion) and single-exposure settings during less challenging conditions (daytime driving or slow/no motion). In practice, the inference system can operate at no computational overhead by using a single exposure setting during most of the time (*e.g*, daytime driving) and use a multi-exposure setting during more challenging conditions (*e.g*, nighttime driving).

6.2 Future Outlook for Robust Perception

All-Weather Perception Model

We have considered various adverse conditions in this dissertation, but a crucial scenario of extreme weather, such as rain, fog, and snow, has not been explored in

detail. Although our approach of photon scaled images is also applicable to such types of image degradation, more thorough analysis and inference approaches for extreme weather are very promising research directions.

The future goal should be to design a single perception framework that is robust to all such image degradations. Such a robust *all-weather* perception system would be extremely valuable for a huge number of outdoor applications, including autonomous driving.

Sensor Fusion Approaches

This dissertation aims to improve the robustness of inference models using individual sensing modalities like SPAD cameras and LiDARs. A promising direction is to tackle this problem using a fusion approach for multiple sensors. Raw sensor measurements from multiple sensing modalities, like RGB images and raw LiDAR data, could provide additional benefits in conditions where different modalities perform better or worse in the same conditions.

MultiModal Foundation Models

Recent advances in foundation models for various sensing modalities (Bachmann et al., 2022; Girdhar et al., 2023; Zhang et al., 2023) have shown great improvements in model pre-training for downstream inference performance. Fusion models mentioned above can benefit greatly from such pre-training. Multi-modal foundation model that can (a) learn from multiple modalities in a self-supervised manner, and (b) utilize raw sensor measurements is a promising future direction.

Confidence Measure from Other Depth Sensors

Although our work considers LiDARs, other 3D sensors such as stereo (Zaarane et al., 2020), structured light (Gupta et al., 2013), and indirect time-of-flight, e.g. Azure Kinect (Qiu et al., 2019), also suffer from noise and low fidelity in challenging imaging scenarios, and could benefit from similar probability

attributes. Since the proposed inference approaches do not make any assumption about the nature of the probability attributes or the noise characteristics of the underlying sensing modality, it could be extended to a wide range of 3D sensors. This is a promising future research direction.

Inference on Time-Varying Inputs

In their current form, the proposed approaches assume static single-frame input. However, most current single-photon sensors (Ma et al., 2017; Ulku et al., 2018) 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 a high-speed camera or scene motion on a temporal sequence of such low bit-depth frames, while exploiting temporal correlations.

Multi-Exposure Cameras

We demonstrated our approach of multi-exposure captures by utilizing multiple cameras with similar or overlapping fields-of-view. With cameras that are capable of capturing multiple images with varying exposures simultaneously (Nayar and Mitsunaga, 2000; Nguyen et al., 2022), multiple exposure images could be captured with a single camera, thus making it easier to perform spatio-temporal alignment. Our work can be considered as a preliminary proof-of-concept for an eventual implementation where a single camera captures multiple exposure images. Demonstrating our approach on such images is an important next step.

Dual Image Degradations

So far, we have considered the dual corruptions of noise and blur. In principle, a similar dual relationship exists between several other image degradation pairs, such as rain and defocus blur (Garg and Nayar, 2005), and snow and motion blur (Barnum et al., 2008). A promising research direction is to evaluate the

proposed approach on other such dual pairs of image degradations, toward the goal of achieving ‘all-weather’ computer vision systems.

REFERENCES

Abdelhamed, Abdelrahman, Marcus A Brubaker, and Michael S Brown. 2019. Noise flow: Noise modeling with conditional normalizing flows. In *Int. conf. comput. vis.*

Abdelhamed, Abdelrahman, Stephen Lin, and Michael S Brown. 2018. A high-quality denoising dataset for smartphone cameras. In *Ieee conf. comput. vis. pattern recog.*

Adaps. 2024. Adaps photonics ads6311. `https://www.adapsphotonics.com/product-55669-218658.html`.

Agostinelli, Forest, Michael R Anderson, and Honglak Lee. 2013. Adaptive multi-column deep neural networks with application to robust image denoising. In *Adv. neural inform. process. syst.*, 1493–1501.

Aittala, Miika, and Frédo Durand. 2018. Burst image deblurring using permutation invariant convolutional neural networks. In *Eur. conf. comput. vis.*

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

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

Antsiperov, VE. 2019. Target identification for photon-counting image sensors, inspired by mechanisms of human visual perception. In *Journal of physics: Conference series*, vol. 1368, 032020. IOP Publishing.

Bachmann, Roman, David Mizrahi, Andrei Atanov, and Amir Zamir. 2022. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, 348–367. Springer.

Bae, Gwangbin, Ignas Budvytis, and Roberto Cipolla. 2022. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2842–2851.

Barnum, P., S. G. Narasimhan, and T. Kanade. 2008. Analysis of rain and snow in frequency space. In *Ijcv*.

Beer, Maik, Jan F Haase, Jennifer Ruskowski, and Rainer Kokozinski. 2018. Background light rejection in spad-based lidar sensors by adaptive photon coincidence detection. *Sensors* 18(12):4338.

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

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

Carbajal, Guillermo, Patricia Vitoria, Mauricio Delbracio, Pablo Musé, and José Lezama. 2021. Non-uniform blur kernel estimation via adaptive basis decomposition. *arXiv preprint arXiv:2102.01026*.

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

———. 2017. Seeing into darkness: Scotopic visual recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 3826–3835.

Chen, Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Ieee conf. comput. vis. pattern recog.*

Chen, Qifeng, Jia Xu, and Vladlen Koltun. 2017. Fast image processing with fully-convolutional networks. *Int. Conf. Comput. Vis.*

Chen, Zhen, Bo Liu, and Guangmeng Guo. 2020. Adaptive single photon detection under fluctuating background noise. *Optics express* 28(20):30199–30209.

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

Coates, PB. 1968. The correction for photonpile-up'in the measurement of radiative lifetimes. *Journal of Physics E: Scientific Instruments* 1(8):878.

Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *Icml*, 1310–1320. PMLR.

Comma. 2022. Comma ai. `https://comma.ai`.

Contributors, MMDetection3D. 2020. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`.

Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Ieee conf. comput. vis. pattern recog.*

Cova, Sergio, Massimo Ghioni, Andrea Lacaita, Carlo Samori, and Franco Zappa. 1996. Avalanche photodiodes and quenching circuits for single-photon detection. *Applied optics* 35(12):1956–1976.

Cui, Ziteng, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. 2021. Multitask aet with orthogonal tangent regularity for dark object detection. In *Int. conf. comput. vis.*, 2553–2562.

Dabov, Kostadin, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. 2007. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing* 16(8):2080–2095.

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

Debevec, Paul, and Jitendra Malik. 1997. Recovering high dynamic range radiance maps from photographs. In *Acm siggraph*, 369–378.

Delbracio, Mauricio, and Guillermo Sapiro. 2015. Burst deblurring: Removing camera shake through fourier burst accumulation. In *Ieee conf. comput. vis. pattern recog.*

Della Rocca, Francesco Mattioli, Hanning Mai, Sam W Hutchings, Tarek Al Abbas, Kasper Buckbee, Andreas Tsiamis, Peter Lomax, Istvan Gyongy, Neale AW Dutton, and Robert K Henderson. 2020. A 128$\times$ 128 spad motion-triggered time-of-flight image sensor with in-pixel histogram and column-parallel vision processor. *IEEE Journal of Solid-State Circuits* 55(7):1762–1775.

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

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

Diamond, Steven, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. 2021. Dirty pixels: Towards end-to-end image processing and perception. *ACM Transactions on Graphics (TOG)* 40(3):1–15.

Digne, Julie, and Carlo De Franchis. 2017. The bilateral filter for point clouds. *Image Processing On Line* 7:278–287.

Fossum, Eric R. 2005. 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*, 214–217.

Garg, K., and S.K. Nayar. 2005. When Does a Camera See Rain? In *Iccv*, vol. 2, 1067–1074.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 ieee conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Gharbi, Michaël, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. 2016. Deep joint demosaicking and denoising. *ACM Trans. Graph.* 35:1 – 12.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*, 15180–15190.

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

Goudreault, Félix, Dominik Scheuble, Mario Bijelic, Nicolas Robidoux, and Felix Heide. 2023. Lidar-in-the-loop hyperparameter optimization. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13404–13414.

Goyal, Bhavya, and Mohit Gupta. 2021. Photon-starved scene inference using single photon cameras. In *Proceedings of the ieee/cvf international conference on computer vision*, 2512–2521.

Goyal, Bhavya, Jean-François Lalonde, Yin Li, and Mohit Gupta. 2022. Robust scene inference under noise-blur dual corruptions. In *2022 ieee international conference on computational photography (iccp)*, 1–12. IEEE.

Graham, Benjamin, Martin Engelcke, and Laurens Van Der Maaten. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 9224–9232.

Granados, Miguel, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. 2010. Optimal hdr reconstruction with linear digital cameras. In *Ieee conf. comput. vis. pattern recog.*, 215–222. IEEE.

Guo, Wenxuan, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Feng, and Jie Zhou. 2024. Lidar-based person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*, 17437–17447.

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

———. 2019b. Asynchronous single-photon 3d imaging. In *Proceedings of the ieee international conference on computer vision*, 7909–7918.

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

Gupta, Mohit, Qi Yin, and Shree K Nayar. 2013. Structured light in sunlight. In *Proceedings of the ieee international conference on computer vision*, 545–552.

Gutierrez-Barragan, Felipe, Atul Ingle, Trevor Seets, Mohit Gupta, and Andreas Velten. 2022. Compressive single-photon 3d cameras. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 17854–17864.

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

Gyongy, Istvan, Sam W Hutchings, Abderrahim Halimi, Max Tyler, Susan Chan, Feng Zhu, Stephen McLaughlin, Robert K Henderson, and Jonathan Leach. 2020. High-speed 3d sensing via hybrid-mode imaging and guided upsampling. *Optica* 7(10):1253–1260.

Hasinoff, Samuel W, Frédo Durand, and William T Freeman. 2010a. Noise-optimal capture for high dynamic range photography. In *Computer vision and pattern recognition (cvpr), 2010 ieee conference on*, 553–560. IEEE.

Hasinoff, Samuel W., Frédo Durand, and William T. Freeman. 2010b. Noise-optimal capture for high dynamic range photography. In *Ieee conf. comput. vis. pattern recog.* IEEE.

Hasinoff, Samuel W, Kiriakos N Kutulakos, Frédo Durand, and William T Freeman. 2009. Time-constrained photography. In *Int. conf. comput. vis.*

Hasinoff, Samuel W, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *TOG* 35(6): 1–12.

He, Kaiming. 2010. Guided image filtering. In *Proc. eccv*.

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

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

Heide, Felix, Steven Diamond, David B Lindell, and Gordon Wetzstein. 2018. Sub-picosecond photon-efficient 3d imaging using single-photon sensors. *Scientific reports* 8(1):1–8.

Heide, Felix, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen O. Egiazarian, Jan Kautz, and Kari Pulli. 2014. Flexisp: a flexible camera image processing framework. *ACM Trans. Graph.* 33.

Henderson, Robert K., Nick Johnston, Francescopaolo Mattioli Della Rocca, Haochang Chen, David Day-Uei Li, Graham Hungerford, Richard Hirsch, David Mcloskey, Philip Yip, and David J. S. Birch. 2019. A 192×128 time correlated spad image sensor in 40-nm cmos technology. *IEEE Journal of Solid-State Circuits* 54(7):1907–1916.

Hendrycks, Dan, and Thomas Dietterich. 2018. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. conf. learn. represent.*

———. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hermosilla, Pedro, Tobias Ritschel, and Timo Ropinski. 2019. Total denoising: Unsupervised learning of 3d point cloud cleaning. In *Proceedings of the ieee/cvf international conference on computer vision*, 52–60.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 4700–4708.

Hutchings, Sam W, Nick Johnston, Istvan Gyongy, Tarek Al Abbas, Neale AW Dutton, Max Tyler, Susan Chan, Jonathan Leach, and Robert K Henderson. 2019. A reconfigurable 3-d-stacked spad imager with in-pixel histogramming for flash lidar or high-speed time-of-flight imaging. *IEEE Journal of Solid-State Circuits* 54(11):2947–2956.

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

Kalantari, Nima Khademi, and Ravi Ramamoorthi. 2019. Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, vol. 38, 193–205. Wiley Online Library.

Kannan, Harini, Alexey Kurakin, and Ian Goodfellow. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*.

Katana. 2024. Onefive katana-05 hp laser. `https://www.laserlabsource.com/Solid-State-Lasers/solid-state-and-fiber-lasers/Picosecond-Laser-532nm-5--J-Onefive`.

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

Krause, Jonathan, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th international ieee workshop on 3d representation and recognition (3drr-13)*. Sydney, Australia.

Lang, Alex H, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 12697–12705.

Lange, Robert. 2000. 3d time-of-flight distance measurement with custom solid-state image sensors in cmos. *CCD-Technology [PhD dissertation]*.

Lee, Jongho, Atul Ingle, Jenu V Chacko, Kevin W Eliceiri, and Mohit Gupta. 2023. Caspi: collaborative photon processing for active single-photon imaging. *Nature Communications* 14(1):3158.

Li, Chengxi, Xiangyu Qu, Abhiram Gnanasambandam, Omar A Elgendy, Jiaju Ma, and Stanley H Chan. 2021. Photon-limited object detection using non-local feature matching and knowledge distillation. In *Proceedings of the ieee/cvf international conference on computer vision*, 3976–3987.

Li, You, and Javier Ibanez-Guzman. 2020. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine* 37(4):50–61.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Eur. conf. comput. vis.* Springer.

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

Lindell, David B, Matthew O'Toole, and Gordon Wetzstein. 2018. Single-photon 3d imaging with deep sensor fusion. *ACM Transactions on Graphics (ToG)* 37(4):1–12.

Liu, Ding, Bihan Wen, Jianbo Jiao, Xianming Liu, Zhangyang Wang, and Thomas S Huang. 2020a. Connecting image denoising and high-level vision tasks via deep learning. *IEEE Transactions on Image Processing* 29:3695–3706.

Liu, Zhe, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. 2020b. Tanet: Robust 3d object detection from point clouds with triple attention. In *Proceedings of the aaai conference on artificial intelligence*, vol. 34, 11677–11684.

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

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

Luo, Shitong, and Wei Hu. 2020. Differentiable manifold reconstruction for point cloud denoising. In *Proceedings of the 28th acm international conference on multimedia*.

———. 2021. Score-based point cloud denoising. In *Proceedings of the ieee/cvf international conference on computer vision*, 4583–4592.

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

Ma, Baorui, Yu-Shen Liu, and Zhizhong Han. 2023. Learning signed distance functions from noisy 3d point clouds via noise to noise mapping. In *International conference on machine learning (icml)*.

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

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

Michaelis, Claudio, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*.

Milanese, Tommaso, Claudio Bruschini, Samuel Burri, Ermanno Bernasconi, Arin C Ulku, and Edoardo Charbon. 2023. Linospad2: an fpga-based, hardware-reconfigurable 512$\times$ 1 single-photon camera system. *Optics Express* 31(26): 44295–44314.

Mildenhall, Ben, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. 2018. Burst denoising with kernel prediction networks. In *Ieee conf. comput. vis. pattern recog.*

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

Morimoto, Kazuhiro, J Iwata, M Shinohara, H Sekine, A Abdelghafar, H Tsuchiya, Y Kuroda, K Tojima, W Endo, Y Maehashi, et al. 2021. 3.2 megapixel 3d-stacked charge focusing spad for low-light imaging and depth sensing. In *2021 ieee international electron devices meeting (iedm)*, 20–2. IEEE.

Nah, Seungjun, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. 2019. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Ieee conf. comput. vis. pattern recog. worksh.*

Nayar, S.K., and T. Mitsunaga. 2000. High dynamic range imaging: spatially varying pixel exposures. In *Ieee conf. comput. vis. pattern recog.*, vol. 1, 472–479 vol.1.

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

Nguyen, Cindy M, Julien NP Martel, and Gordon Wetzstein. 2022. Learning spatially varying pixel exposures for motion deblurring. *arXiv preprint arXiv:2204.07267*.

Niclass, Cristiano, Alexis Rochas, P-A Besse, and Edoardo Charbon. 2005. 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.

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

Onzon, Emmanuel, Fahim Mannan, and Felix Heide. 2021. Neural auto-exposure for high-dynamic range object detection. In *Ieee conf. comput. vis. pattern recog.* IEEE.

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

Paszke, Adam, 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. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, 8024–8035. Curran Associates, Inc.

Pediredla, Adithya K, Aswin C Sankaranarayanan, Mauro Buttafava, Alberto Tosi, and Ashok Veeraraghavan. 2018. Signal processing based pile-up compensation for gated single-photon avalanche diodes. *arXiv preprint arXiv:1806.07437*.

Pellegrini, Sara, Gerald S Buller, Jason M Smith, Andrew M Wallace, and Sergio Cova. 2000. Laser-based distance measurement using picosecond resolution time-correlated single-photon counting. *Measurement Science and Technology* 11(6):712.

Peng, Jiayong, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. 2020. Photon-efficient 3d imaging with a non-local neural network. In *European conference on computer vision*, 225–241. Springer.

Qi, Charles R, Xinlei Chen, Or Litany, and Leonidas J Guibas. 2020. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 4404–4413.

Qi, Charles R, Or Litany, Kaiming He, and Leonidas J Guibas. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the ieee/cvf international conference on computer vision*, 9277–9286.

Qi, Charles R, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. 2018. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 918–927.

Qi, Charles R, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 652–660.

Qi, Charles Ruizhongtai, Li Yi, Hao Su, and Leonidas J Guibas. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30.

Qiu, Di, Jiahao Pang, Wenxiu Sun, and Chengxi Yang. 2019. Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In *Proceedings of the ieee international conference on computer vision*, 9994–10003.

Rakotosaona, Marie-Julie, Vittorio La Barbera, Paul Guerrero, Niloy J Mitra, and Maks Ovsjanikov. 2020. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer graphics forum*, vol. 39, 185–203. Wiley Online Library.

Rapp, Joshua, and Vivek K Goyal. 2017. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Transactions on Computational Imaging* 3(3):445–459.

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

Renker, D. 2006. 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. Proceedings of the 4th International Conference on New Developments in Photodetection.

Rochas, Alexis. 2003a. Single photon avalanche diodes in cmos technology. Ph.D. thesis, EPFL.

———. 2003b. Single photon avalanche diodes in cmos technology. Tech. Rep., Citeseer.

Rokach, Lior. 2010. Ensemble-based classifiers. *Artificial intelligence review* 33(1):1–39.

Rozumnyi, Denys, Jiri Matas, Filip Sroubek, Marc Pollefeys, and Martin R. Oswald. 2021. Fmodetect: Robust detection of fast moving objects. In *Int. conf. comput. vis.*, 3541–3549.

Satat, Guy, Matthew Tancik, and Ramesh Raskar. 2018. Towards photography through realistic fog. In *2018 ieee international conference on computational photography (iccp)*, 1–10. IEEE.

Schuler, Christian J, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. 2015. Learning to deblur. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(7): 1439–1451.

Sen, Pradeep, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. 2012. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.* 31(6):203–1.

Shi, Shaoshuai, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 10529–10538.

Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 770–779.

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

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

Sim, Hyeonjun, Jihyong Oh, and Munchurl Kim. 2021. Xvfi: extreme video frame interpolation. In *Int. conf. comput. vis.*

Song, Shuran, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 567–576.

Song, Wenzheng, Masanori Suganuma, Xing Liu, Noriyuki Shimobayashi, Daisuke Maruta, and Takayuki Okatani. 2021. Matching in the dark: A dataset for matching image pairs of low-light scenes. In *Int. conf. comput. vis.*, 6029–6038.

Spinnaker. 2022. Spinnaker sdk. `https://www.flir.com/products/spinnaker-sdk/`.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In

Proceedings of the ieee conference on computer vision and pattern recognition, 2818–2826.

Tachella, Julián, Yoann Altmann, Nicolas Mellado, Aongus McCarthy, Rachael Tobin, Gerald S Buller, Jean-Yves Tourneret, and Stephen McLaughlin. 2019. Real-time 3d reconstruction from single-photon lidar data using plug-and-play point cloud denoisers. *Nature communications* 10(1):4984.

Teledyne. 2022. Teledyne flir blackfly s machine vision camera. `https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-13Y3C-C`.

Tian, Zhi, Hao Chen, Xinlong Wang, Yuliang Liu, and Chunhua Shen. 2019a. AdelaiDet: A toolbox for instance-level recognition tasks. `https://git.io/adelaidet`.

Tian, Zhi, Chunhua Shen, Hao Chen, and Tong He. 2019b. Fcos: Fully convolutional one-stage object detection. In *Int. conf. comput. vis.*, 9627–9636.

Tseng, Ethan, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. 2019. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.* 38(4):27–1.

Turin, G. 1960. An introduction to matched filters. *IRE Transactions on Information Theory* 6(3):311–329.

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

Wah, C., S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology.

Wang, Kun, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. 2021a. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Int. conf. comput. vis.*, 16055–16064.

Wang, Ruixing, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. 2021b. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Int. conf. comput. vis.*, 9700–9709.

Wang, Tai, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. 2024. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 19757–19767.

Wang, Yang, Yang Cao, Zheng-Jun Zha, Jing Zhang, and Zhiwei Xiong. 2020. Deep degradation prior for low-quality image classification. In *Ieee conf. comput. vis. pattern recog.*, 11049–11058.

Wang, Zhenyu, Ya-Li Li, Xi Chen, Hengshuang Zhao, and Shengjin Wang. 2023. Uni3detr: Unified 3d detection transformer. In *Advances in neural information processing systems*, ed. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, vol. 36, 39876–39896. Curran Associates, Inc.

Wei, Kaixuan, Ying Fu, Jiaolong Yang, and Hua Huang. 2020. A physics-based noise formation model for extreme low-light raw denoising. In *Ieee conf. comput. vis. pattern recog.*

Wei, Zeyong, Honghua Chen, Liangliang Nan, Jun Wang, Jing Qin, and Mingqiang Wei. 2024. Pathnet: Path-selective point cloud denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46(6):4426–4442.

Wolff, Katja, Changil Kim, Henning Zimmer, Christopher Schroers, Mario Botsch, Olga Sorkine-Hornung, and Alexander Sorkine-Hornung. 2016. Point

cloud noise and outlier removal for image-based 3d reconstruction. In *2016 fourth international conference on 3d vision (3dv)*, 118–127. IEEE.

Wronski, Bartlomiej, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. 2019. Handheld multi-frame super-resolution. *ACM Trans. Graph.* 38(4).

Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.

Xia, Zhihao, Patrick Sullivan, and Ayan Chakrabarti. 2020. Generating and exploiting probabilistic monocular depth estimates. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 65–74.

Xie, Qizhe, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Ieee conf. comput. vis. pattern recog.*

Xu, Li, Jimmy S Ren, Ce Liu, and Jiaya Jia. 2014. Deep convolutional neural network for image deconvolution. *NIPS* 27.

Xu, Xiangyu, Hao Chen, Francesc Moreno-Noguer, Laszlo A. Jeni, and Fernando De la Torre. 2020. 3d human shape and pose from a single low-resolution image with self-supervised learning. `2007.13666`.

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

Yang, Zetong, Yanan Sun, Shu Liu, and Jiaya Jia. 2020. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 11040–11048.

Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. 2021. Center-based 3d object detection and tracking. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 11784–11793.

Zaarane, Abdelmoghit, Ibtissam Slimani, Wahban Al Okaishi, Issam Atouf, and Abdellatif Hamdoun. 2020. Distance measurement system for autonomous vehicles using stereo camera. *Array* 5:100016.

Zamir, Syed Waqas, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-stage progressive image restoration. In *Ieee conf. comput. vis. pattern recog.*

Zhang, Jiaming, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. 2023. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems* 24(12):14679–14694.

Zhang, Jiawei, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. 2017a. Learning fully convolutional networks for iterative non-blind deconvolution. In *Ieee conf. comput. vis. pattern recog.*, 3817–3825.

Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017b. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* 26(7):3142–3155.

Zhang, Zhongyuan, Lichen Lin, and Xiaoli Zhi. 2024. R-pointnet: Robust 3d object recognition network for real-world point clouds corruption. *Applied Sciences* 14(9):3649.

Zhang, Zijing, Yuan Zhao, Yong Zhang, Long Wu, and Jianzhong Su. 2013. A real-time noise filtering strategy for photon counting 3d imaging lidar. *Optics express* 21(8):9247–9254.

Zhao, Hengshuang, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. 2021. Point transformer. In *Proceedings of the ieee/cvf international conference on computer vision*, 16259–16268.