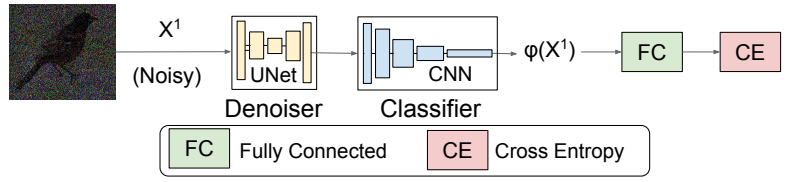
Photon-Starved Scene Inference using Single Photon Cameras Technical Report

Bhavya Goyal Mohit Gupta
University of Wisconsin-Madison
{bhavya, mohitg}@cs.wisc.edu

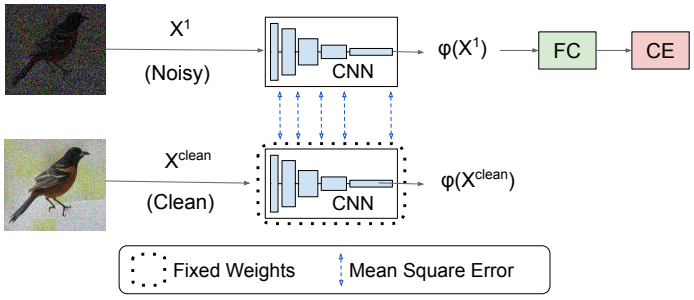
In this report, we provide technical details and results that are not included in the main paper due to space constraints.

1. Image Classification

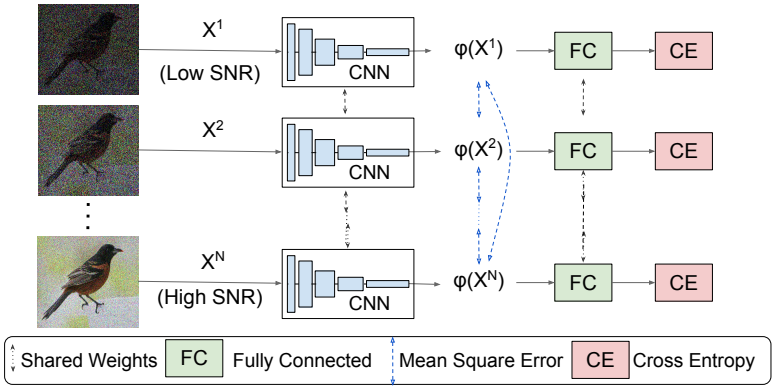
1.1. Architecture Overview



(a) Joint Denoising



(b) Student Teacher Learning



(c) Photon Net (Ours)

Figure 1: **Architecture Overview** for Image Classification

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

Joint Denoising Joint Denoising architecture [2] consists of a joint network with a denoiser (20 layer UNet) and a CNN classifier (Resnet-18 [4]). 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 1a). The denoiser is initialized with pretrained weights on noisy and clean images.

Student Teacher Learning Student Teacher architecture [3] 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 1b). Student Teacher learning uses double the 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.

1.2. Additional Results

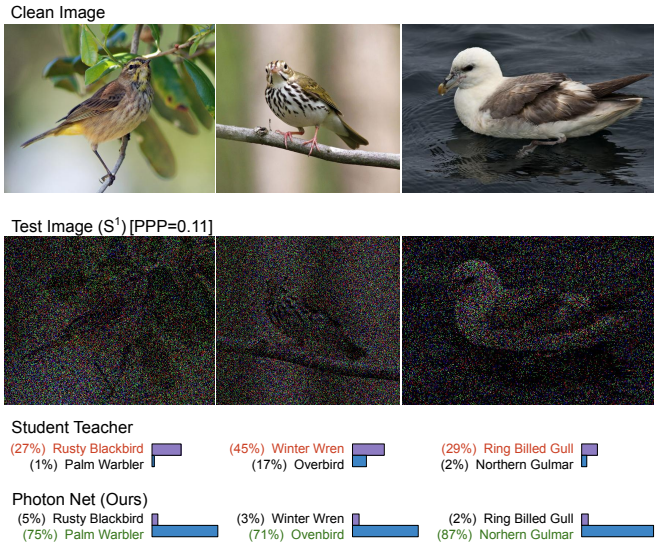


Figure 2: **Image Classification Results** using Photon Net on CUB-200-2011 Dataset for $\mathcal{S}^1$ test images.

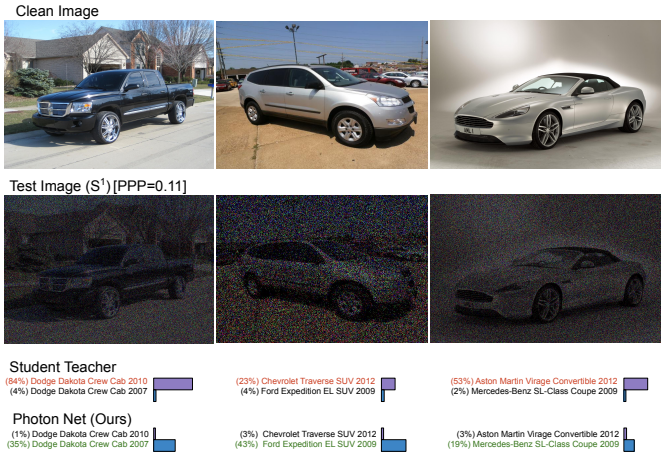
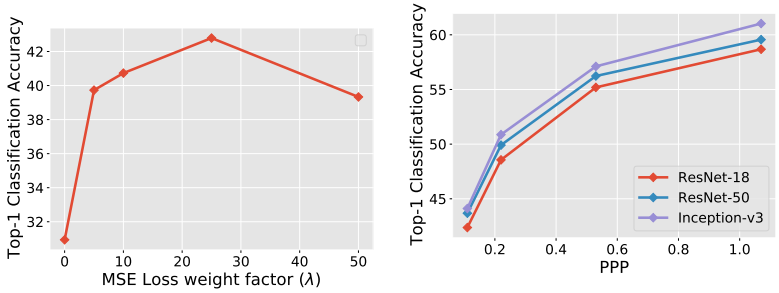


Figure 3: **Image Classification Results** using Photon Net on CARS Dataset for $\mathcal{S}^1$ test images.

Figure 4: **Few Failure cases examples** of Photon Net on CUB-200-2011 and CARS dataset for $\mathcal{S}^1$ test images.

Figure 2 and 3 shows results of image classification on CUB-200-2011 [8] and CARS [5] 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 example of few failure cases where Photon Net architecture fails to get the correct prediction. As we can observe, these cases are extremely challenging.

1.3. More ablation studies



(a) MSE Loss weight factor (λ)

(b) Different CNN architectures for Photon Net

Figure 5: **Ablation Studies**: Performance of Photon Net training while varying: (a) MSE loss weight factor (λ) , (b) base architecture

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 5 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. [6]. Figure 5 shows increase in the performance of Photon Net with deeper CNN architectures. This shows the versatility and ease to extend Photon Net to different CNN architectures.

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**

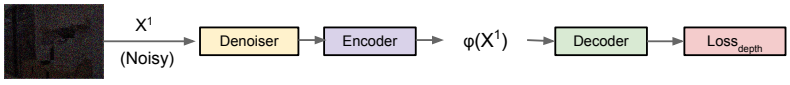
Table 1: **Ablation Study**: Top-1 Accuracy results of image classification on CUB-200-2011 dataset

We perform an ablation study to analyse the individual contribution of Photon Net training and using Photon Scaled Images in the final performance. Table 1 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.

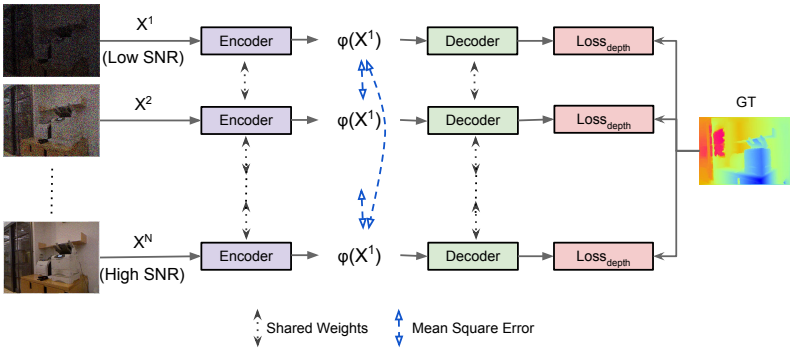
We also compare our model to Curriculum Learning technique, where the Vanilla Net is trained in N steps, starting with only the clean images first step and successively finetuning the model by adding images with higher noise levels in next steps. Photon Net outperforms Curriculum Learning as it uses 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.

2. Monocular Depth Estimation

2.1. Architecture Overview



(a) Joint Denoising



(b) Photon Net (Ours)

Figure 6: **Overview of the Depth Estimation with Photon Net**:

Joint Denoising Joint Denoising consists of a depth estimation architecture based on DenseDepth [1] 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).

2.2. Results

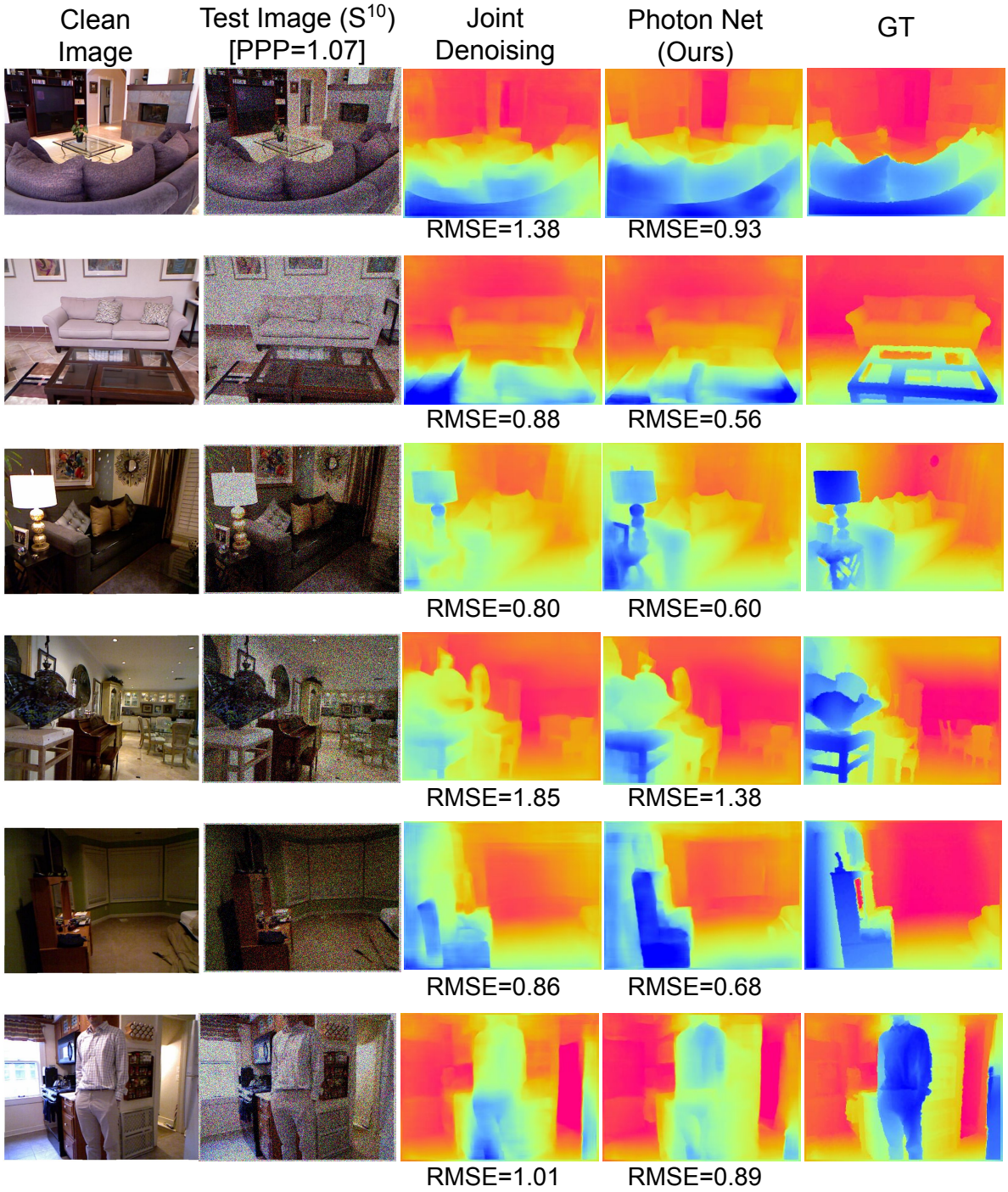


Figure 7: **Monocular Depth Estimation Results** on NYUV2 dataset of $\mathcal{S}^{10}$ test images

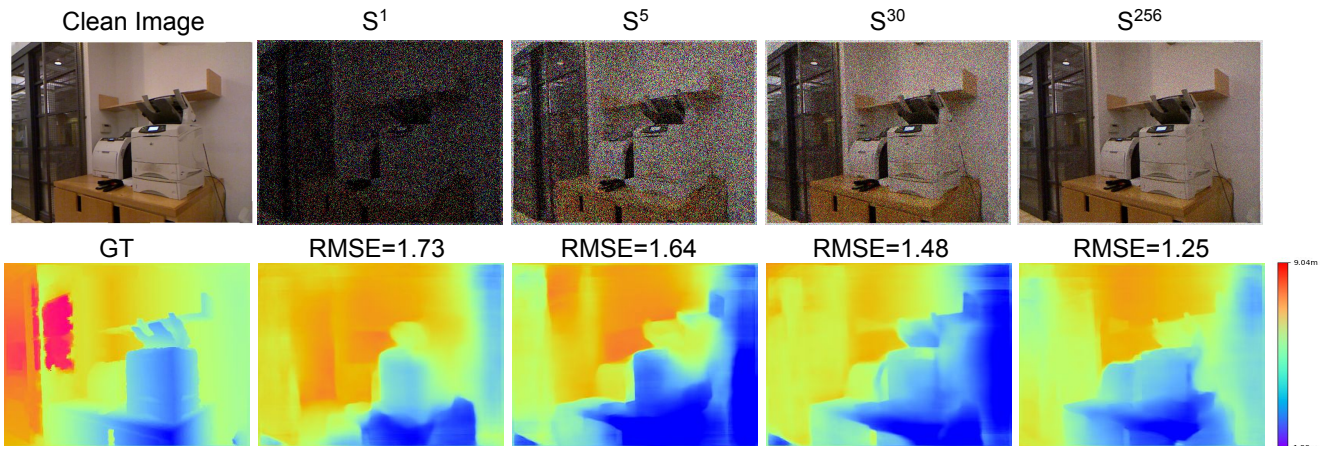


Figure 8: **Monocular Depth Estimation Results** on NYUV2 dataset with increasing PPP level in the testing image

Figure 7 shows examples of output depth maps from the Photon Net and the baseline. Figure 8 shows output depth maps while using higher SNR image for testing.

3. Real Captures from SPADs

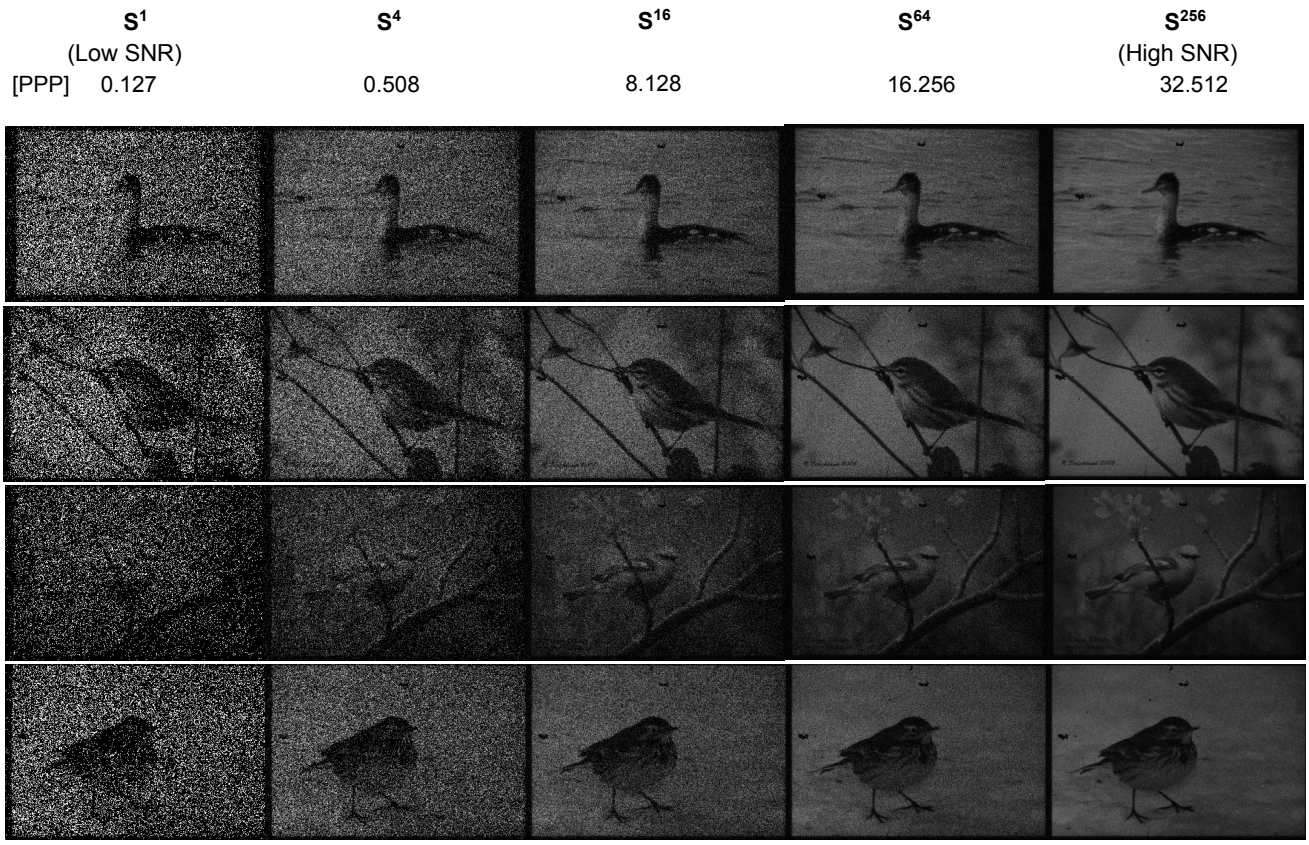


Figure 9: **Real Captures:** Sample of images from SPAD cameras

Figure 10: **Artifacts** in Real Captures from SwissSPAD2 camera

To collect dataset of real captures from SPAD sensors, we displayed the original RGB images on a monitor 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 9 shows example of images captured using SwissSPAD2 camera [7] 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 10.

References

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

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

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

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

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

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

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