Supplementary material

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

This supplementary document summarizes the following experimental results:

- Required photon levels for detection (Section 1).
- Choice of frame numbers and K (Section 2).
- More qualitative results (Section 3).

1. Required Photon Levels for Detection

In Figure 1, we discuss how many photons are needed for each pixel in order to achieve the target detection performance. The x-axis represents the detection accuracy we want to achieve and the y-axis is the minimal numbers of photons per pixel needed in the images. We compare four settings by switching the inputs from synthetic CIS to QIS images and changing the baseline method to our method. When the target mAP is 50%, QIS data only needs half photons of CIS data to reach the same accuracy by just using Faster R-CNN. By introducing our method, we can further decrease the required photon level by half on average.

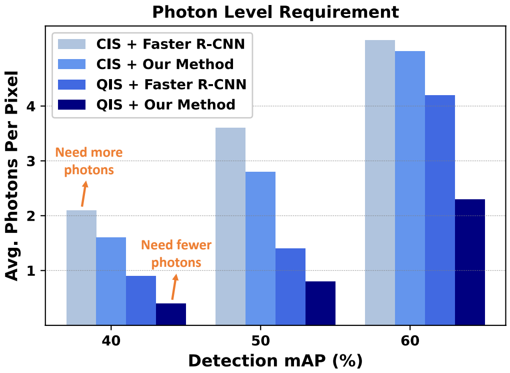

Figure 1: Photon level requirement vs. detection performance.

2. Choice of Frame Numbers and K

mAP (%)	ppp = 0.25		ppp = 0.5		ppp = 1.0		ppp = 2.0		ppp = 5.0	
	$T = 3$	$T = 8$	$T = 3$	$T = 8$	$T = 3$	$T = 8$	$T = 3$	$T = 8$	$T = 3$	$T = 8$
$K = 1$	32.3	**33.3**	41.5	42.8	49.6	51.9	58.4	59.0	65.1	**66.0**
$K = 2$	**32.7**	33.2	**41.6**	**43.0**	**50.0**	51.9	**58.7**	**59.3**	**65.6**	**66.0**
$K = 3$	32.4	33.2	41.5	42.8	49.9	**52.1**	58.6	59.2	65.4	65.9
$K = 4$	32.5	33.0	41.5	**43.0**	**50.0**	**52.1**	58.6	59.1	65.4	65.9

Table 1: A study of frame numbers and searched similar feature numbers. T is the number of frames input to our model and K is the number of searched features per frame for feature aggregation. We test our model under different photon levels from 0.25 to 5.0. For each column, the best mAP is shown in bold.

Non-local module is applied to multi-frame input and searches for K similar features in each frame. Thus, we study the best and practical settings for our designed network. In Table 1, we find that using 8 frames is always better than 3 frames no matter which photon levels. It is easy to interpret this result because more frames provide more information and the proposed Non-local module is able to associate similar patches across multiple frames. However, more input frames require more computations and processing time. When we set the frame number larger than 8, it will exceed the GPU memory. Thus we use 8-frame sequences as input for practical usage. Moreover, we discover that K=2 is the best choice for the number of searched similar features per frame. Too many selected features could be a distractor for the denoising purpose.

3. More Qualitative Results

In Figure 2, we show more qualitative examples comparing our method with the baseline method: Faster R-CNN [1]. All of the four scenarios are dynamic scenes. The first two are synthetic data and the photon levels are set to 2.0 ppp and 1.0 ppp. The last two scenes are real data captured by the BostonQIS camera at photon levels of 0.28 ppp and 0.19 ppp. We observe that the presence of heavy shot noise results in false alarms detected in the background, such as the sheep and the bird in scene 2. Also, the baseline method fails to detect the moving person for most of the time in scene 3.

References

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

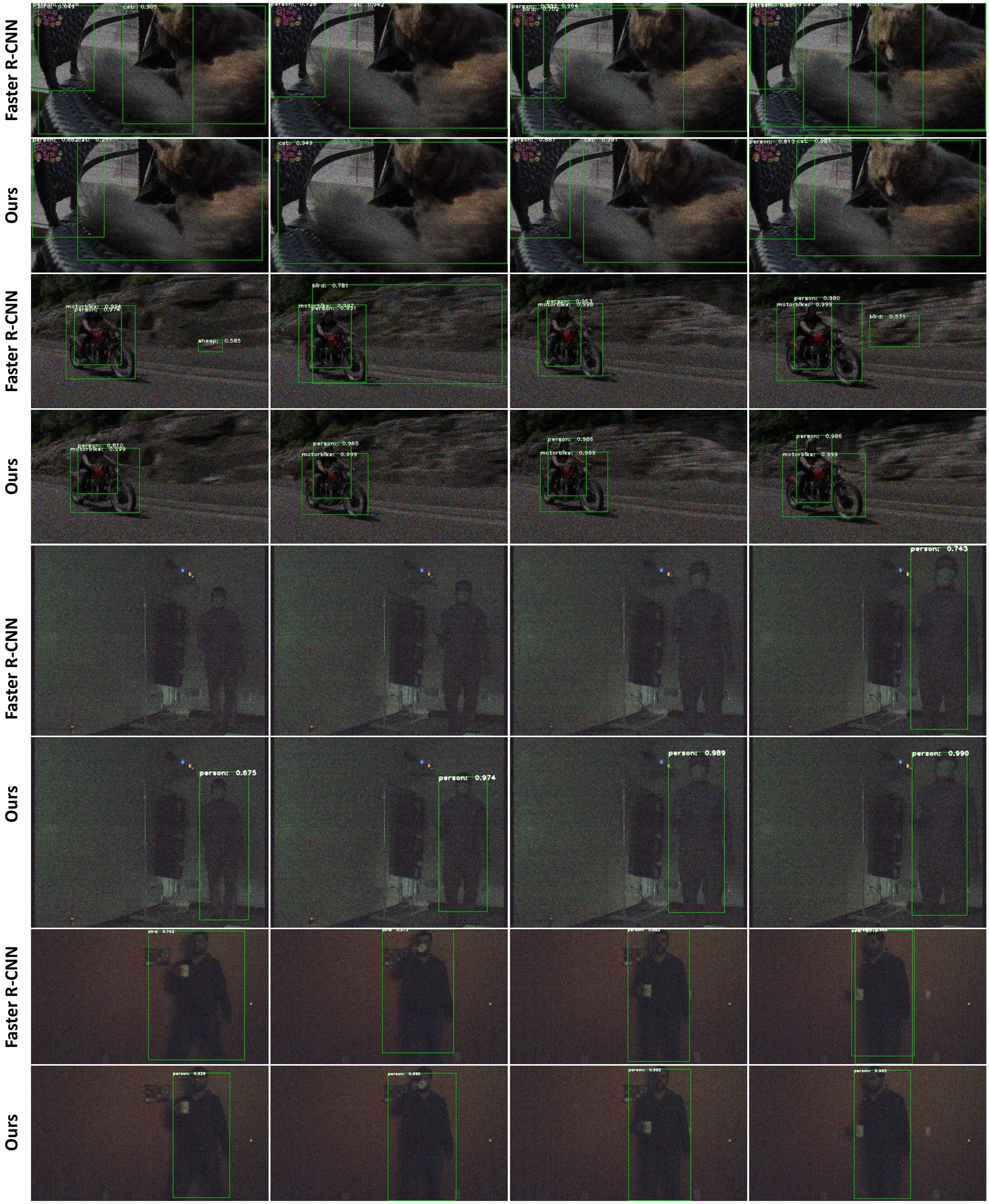

Figure 2: Detection results on synthetic and real data.