Supplementary Materials

1. Training Details

In order to train the Faster R-CNN[5] model with ResNet101[2] backbone on our generated COCO raw dataset, we use an initial learning rate of 0.002, batch size of 2, momentum of 0.9, and weight decay of 0.0001. For Yolov3[4], the initial learning rate is changed to 0.0002 and the batch size is 24 while the other hyperparameters are kept the same. All the models are trained for 110,000 iterations, and the learning rate is divided by 10 after 85,000 and 100,000 iterations.

For MobileNetV2-0.35X on VWW, we use the RMSProp optimizer with an initial learning rate of 0.003. We train for 50 epochs, and the learning rate is divided by 10 each after 35 and 45 epochs.

In the few-shot learning experiments on PASCALRAW [1] datset, the initial learning rate is chosen to be 0.001 for the Faster R-CNN model and 0.0001 for YOLOv3 model. We train both the models for 30,000 iterations and divide the learning rate by 10 after 24,000 iterations.

2. Visualization

As shown in Figure 1, the first column shows the results of the model pre-trained on the COCO[3] RGB dataset. It demonstrates that the model can precisely detect the location and class of the objects in most cases.

The second and third column shows images from our demosaiced COCO raw dataset. The second column, with the results obtained by our model that operates on the demosaiced images (without the in-pixel convolution) shows that the model still can detect most objects but also incur some mistakes, especially when similarly colored objects obscure each other more severely.

The third column shows the performance of the model with embedded in-pixel convolution, and the performance is generally similar to that of the traditional model shown in the second column. The last column shows the mosaiced COCO raw images with only one channel. As we can see, for most of the objects detected correctly, they have lower confidence than the other three columns, which is consistent with the fact that training on mosaiced images yields the lowest mAP for the COCO dataset as shown in Table 2 of the paper.

Table 1. mAP of normal convolution and in-pixel convolution models on the COCO raw dataset

	mean average precision					
model	0.5:0.95	0.5	0.75	S	M	L
normal Conv	**42.8**	**64.1**	**47.1**	**25.6**	**46.9**	55.0
in-pixel Conv	42.3	62.8	46.2	24.0	46.3	**55.5**

Table 2. test mAP of models with our proposed demosaicing approach coupled with normal and in-pixel convolution with the Faster R-CNN model on the PASCALRAW dataset. Note, FT denotes fine-tuned and FS denotes few-shot learned.

	mean average precision					
model	0.5:0.95	0.5	0.75	S	M	L
normal conv.+FT	12.4	37.4	3.6	1.2	14.4	24.7
in-pixel conv.+FT	12.8	39.8	3.6	2.1	15.9	23.9
normal conv.+FT+FS	**31.7**	60.3	**29.7**	8.4	**30.5**	42.0
in-pixel conv.+FT+FS	31.4	**60.6**	28.7	**8.6**	28.4	**43.8**

3. Impact of In-Pixel Convolution on Object Detection

For the object detection experiments, we evaluate the in-pixel demosaicing approach coupled with in-pixel convolution with the Faster R-CNN model in Table 2. As we can see, it leads to similar mAP values with that of the normal convolution with only our proposed demosaicing approach. The difference is only 0.35% on average across the two cases (only fine-tuning and fine-tuning coupled with few-shot learning) for IoU ranging from 0.5 to 0.95, which implies that the custom function in the in-pixel convolution can be re-trained to recover any drop in performance due to the analog non-idealities.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[2] 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.

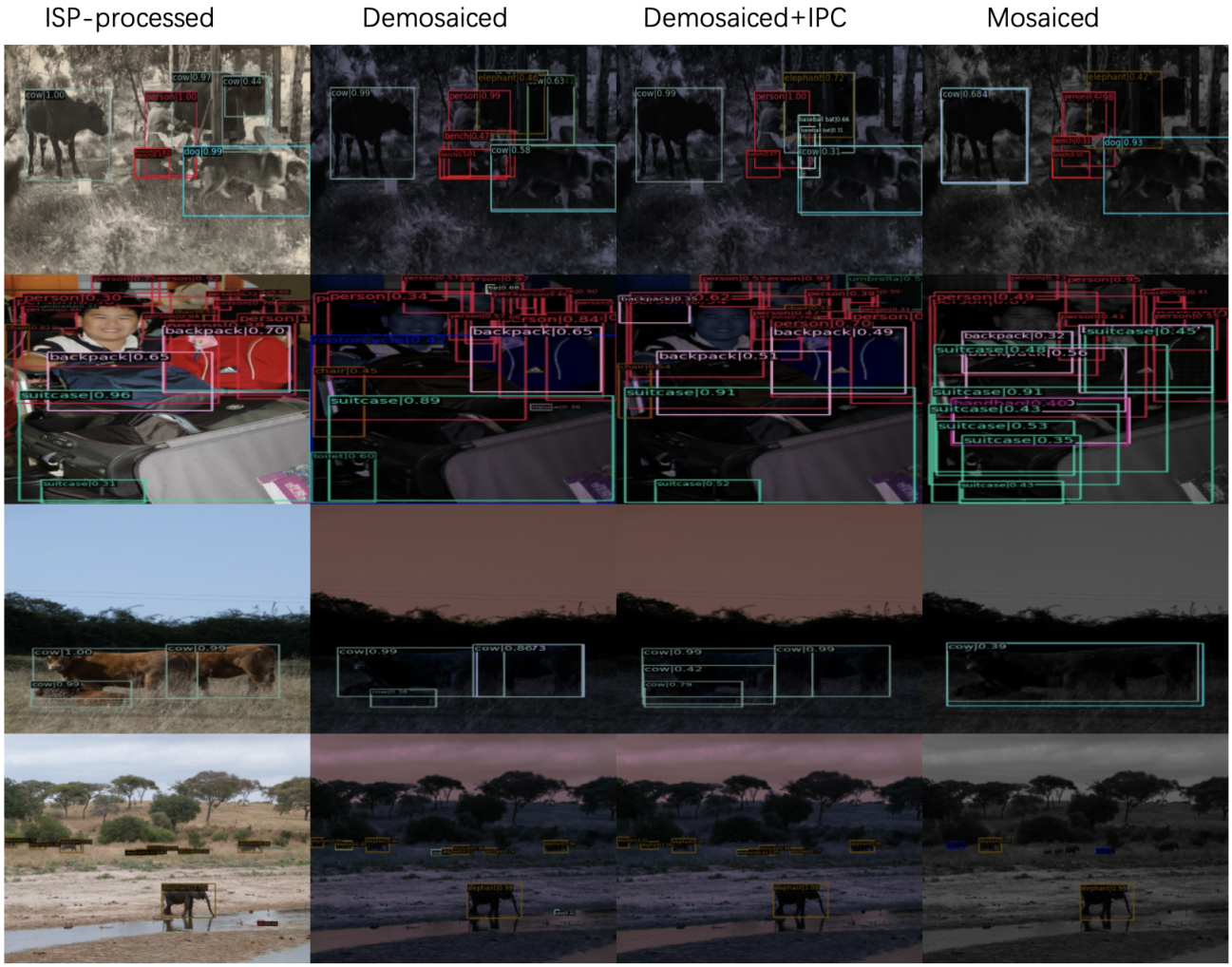

Figure 1. The visualization of the results on different versions of the COCO raw dataset.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.