NOD: Taking a Closer Look at Detection under Extreme Low-Light Conditions with Night Object Detection Dataset (Supplement)

Igor Morawski National Taiwan University

Yu-An Chen

Yu-Sheng Lin

Winston H. Hsu

1 NOD: Night Object Detection Dataset

1.1 Collection

We collected our dataset with two cameras: Sony RX100 VII and Nikon D750, and because of that, the NOD dataset can be logically split into two subsets depending on the camera. In our dataset, there are 4143 images captured with Sony, and 4006 with Nikon. The resolution of images collected is 5472×3648 and 3936×2624 for Sony and Nikon, respectively. All photos were shot handheld, and most of them were shot in Full Auto mode. Some of them shot in Shutter Priority mode, especially when there were fast moving objects (*e.g.* cars) involved. Thus, the images in our dataset show all common culprits of low-light photography: motion blur, out-of-focus blur, and severe noise. Out of these, out-of-focus and motion blur occur sporadically, while the degree of noise is indicated by the high ISO shutter speed of the photos: ISO 6,400 in 89% of images in the Sony set, and ISO 12,800 in 89% of images in the Nikon set. In both subsets, we observe degradation common in low-light imaging: out-of-focus blur, motion blur, and intense noise. Moreover, some photos are severely underexposed. The variety of lighting conditions in our dataset is showed in Fig. 1.

1.2 Annotation

To ensure the high quality of bounding box annotation under challenging conditions, we outsourced data labeling to a company that annotated instances on images enhanced by MBLLEN [5] in their original resolution. 3210 out of 4143 images in the Sony subset, and all 4006 of images in the Nikon subset were labeled with bounding boxes for object: *person*, *bicycle*, *car*. The number of instances for each category is shown in Tab 1. Statistics

class	# instances
person	31,906
bicycle	9,589
car	5,246
in total	46,741

Table 1: Object statistics in our dataset.

of the images and bounding boxes can be found in Fig. 3, 4, 5. Sample images and bounding box annotation from the dataset can be seen in Fig. 2.

Objects were annotated according to our specifications, shortly summarized below.

- *Person* - real people, including people in posters, billboards, screens, mirrors; annotated whenever any part of the head from the chin up, front or back, is observed; excluding human-shaped objects, *e.g.*, mannequins; the bounding box should include: 1) hair and fake hair, 2) clothing (glasses, pants, dresses, hats); the bounding box should exclude: 1) any vehicles, *e.g.*, bicycles, motorbikes, 2) objects held by the person, 3) handbags, 4) backpacks.

- *Bicycle* - human-powered bicycles, excluding motor-powered bicycles and motorbikes; omitted if there is a doubt whether the object is a bicycle or a motorbike.

- *Car* - including minivans, vans, and ambulances; excluding trucks and buses; omitted if there is a doubt whether the object is a car or a motorbike: 1) only one tail light is visible, or 2) only one wheel is visible,

- For all - 1) if the boundary is not visible, *e.g.,* due to partial occlusion, the object should be still annotated around the most probable boundary, 2) if the object is truncated, the bounding box should be aligned with the image boundary.

We additionally provide 933 unnanotated images in the Sony dataset. For each set (Sony, Nikon), we randomly selected 80% of images for the training set, 10% for the validation set, and another 10% for the testing set, while maintaining the class distribution among the sets, as shown in Fig. 6.

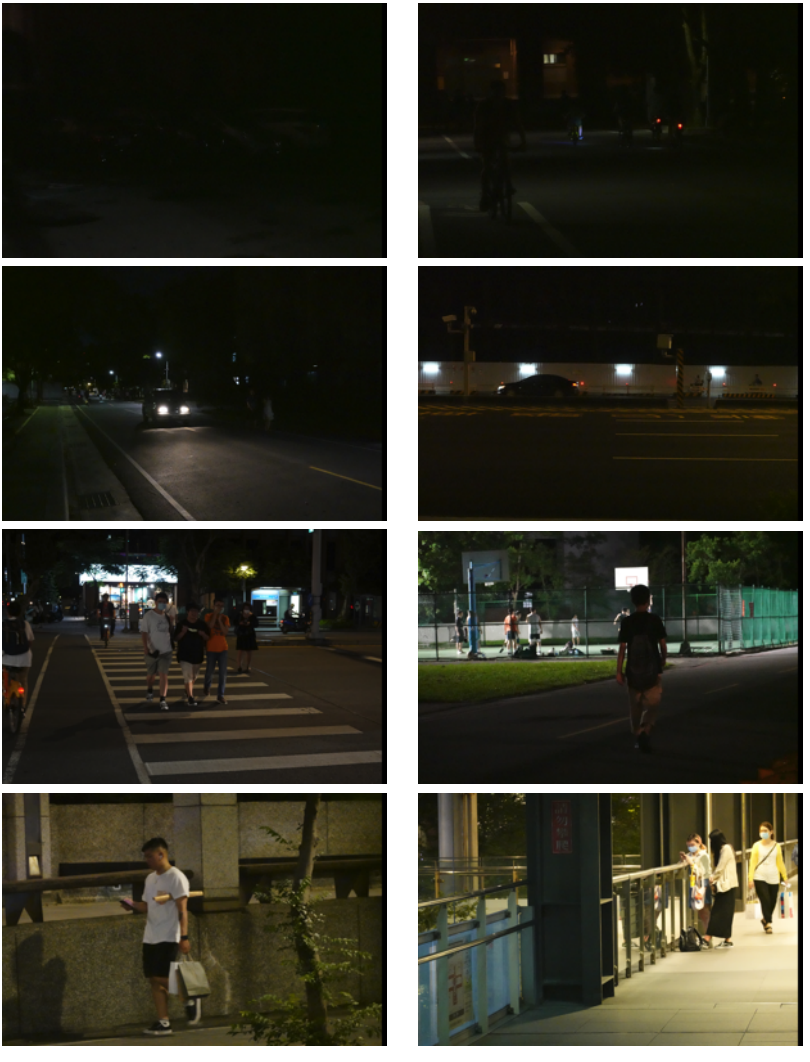


Figure 1: Variety of lighting conditions in our dataset.

Figure 2: Sample images and bounding box annotation from the dataset.

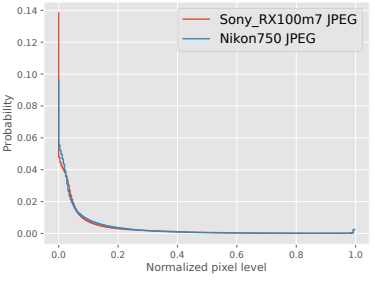


Figure 3: Histogram of images in the dataset.

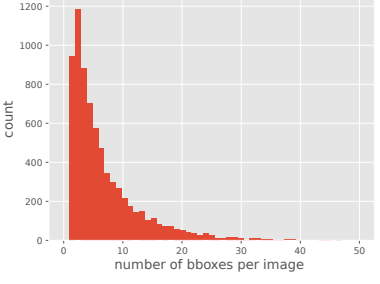
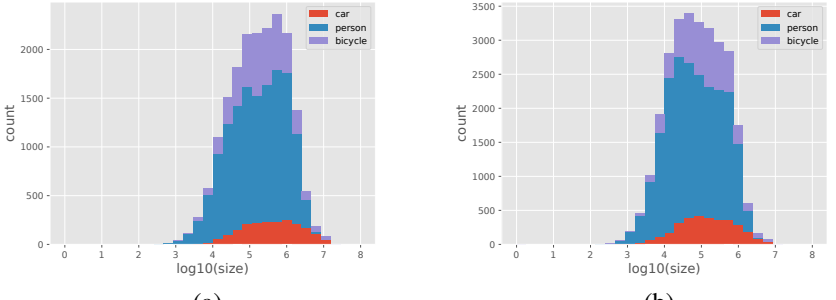


Figure 4: Bounding box count per image.



(a) (b)

Figure 5: Bounding box size in the (a) Sony, and (b) Nikon subsets.

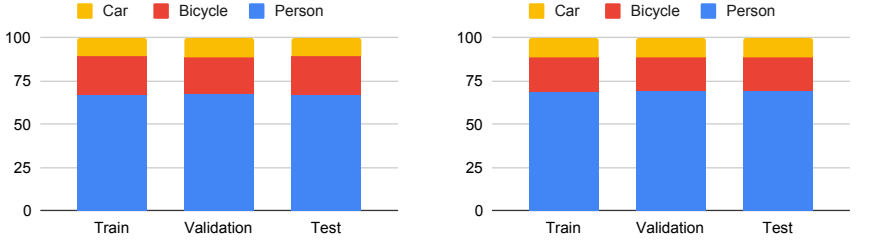


Figure 6: Proportion of classes in train, validation and test subsets of (a) Sony, and (b) Nikon subsets.

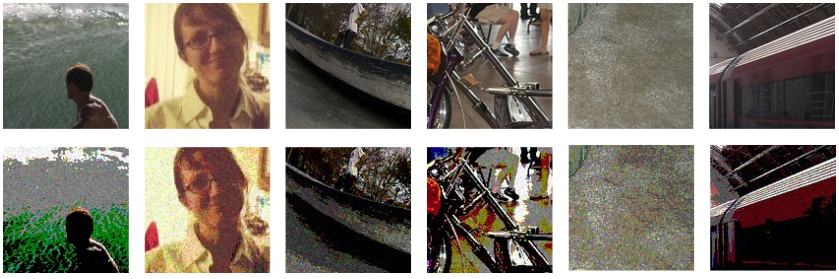


Figure 7: First row: patches from the COCO [3] dataset. Second row: patches corruped by posterization and shot noise.

2 Pre-Training Examples

W define the pre-training task as an image restoration task. In our implementation, we extracted random patches from images in the COCO [3] dataset, and corrupt them by applying posterization, from 2 to 8 gray levels, and adding shot noise. Examples of images used in training are shown in Fig. 7

3 Implementation Details

We implement all models with PyTorch and Open MMLab Detection Toolbox [1] on 2 Tesla-V100 32GB GPUs with SyncBN. We use SGD optimizer, and apply a batch size of 8. We set the learning rate to 1e−4, and use linear warmup policy with warmup ratio 1e−4 for 4 epochs. All models are initialized with COCO weights and trained for at most 90 epochs depending on the augmentation methods used, unless otherwise specified. In all experiments, we resize images to 1333×800 pixels while keeping aspect ratio and padding to size divisible by 32, except for experiments on ExDark [4] where we resize to 1000×600 pixels. In our experiments on [4], we initialize UNet from a checkpoint fine-tuned on our dataset, because of the small size of the training set in [4].

As for the U-Net [7], we modify the original network by replacing ReLU activatations with Mish [6] layers and adding Batch Norm layers before every activation layer. We randomly extract patches from images in COCO [3] dataset, and corrupt them by applying posterization, from 2 to 8 gray levels, and adding shot noise. We use Adam [2] as an optimizer, apply a batch size of 64, set the learning rate to 1e−4, and train for 130,000 steps, on two Teslas K80 12GB.

4 Qualitative Results

In our paper, we propose to incorporate an image enhancement module into the object detection. In Fig. 8-10, we present examples of images enhanced by the enhancement module. The images come from the Sony subset.

Figure 8: Left: input image. Right: intermediate representation enhanced by the proposed image enhancement module.

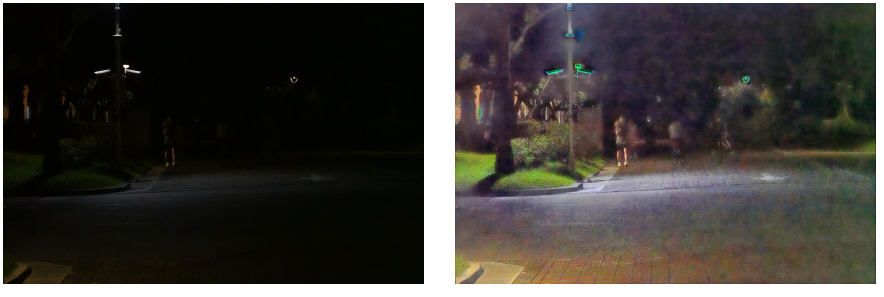
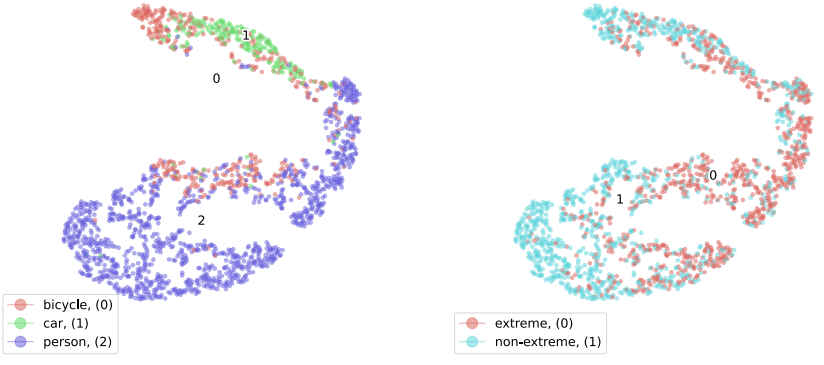


Figure 9: Left: input image. Right: intermediate representation enhanced by the proposed image enhancement module.

Figure 10: Left: input image. Right: intermediate representation enhanced by the proposed image enhancement module.

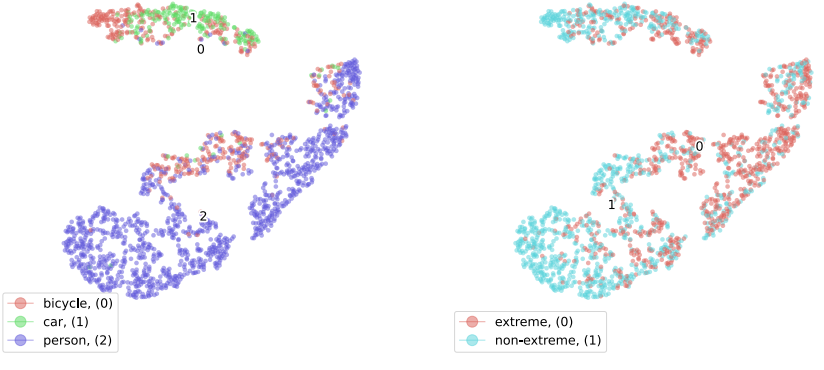
5 t-SNE embeddings

To investigate whether differentiating between the extreme and non-extreme low-light conditions in the way discussed in the paper is meaningful, we visualized t-SNE embedding of the features extracted by the models in our paper. In the paper, we shown t-SNE emdeddings of the features extracted by the baseline model trained on the COCO [3] dataset. We further visualized t-SNE embeddings for all the models in our paper, and found out that our findings hold for all of them. We show the results for the baseline model and proposed method trained on the Sony subset, in Fig. 11 and 12, respectively.



(a) (b)

Figure 11: t-SNE embeddings of the features extracted by the baseline model trained on the Sony subset.



(a) (b)

Figure 12: t-SNE embeddings of the features extracted by the proposed detection-with-enhancement model trained on the Sony subset.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[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] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. doi: https://doi.org/10.1016/j.cviu.2018.10.010.

[5] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, page 220, 2018.

[6] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.