Crafting Object Detection in Very Low Light (Supplementary Material)

Yang Hong *
hongyang@bit.edu.cn
Kaixuan Wei *
kaixuan_wei@bit.edu.cn
Linwei Chen
chenlinwei@bit.edu.cn
Ying Fu †
fuying@bit.edu.cn

School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

This supplementary material provides more details and results that are not included in the main paper due to the space limitations. The contents are organized as follows:

Section A gives more insights and discussions on our system design; Section B presents additional quantitative results; more examples of the LOD dataset are provided in Section C, and Section D shows more visual results on our LOD dataset. It's worth noting that we have also *created a video* to help the readers better perceive the performance of our low-light detection system against other representative competing methods. **The video file has been attached with this supplementary document.**

A More Discussion on System Design.

A.1 Why RAW?

More detailed comparison results about analysis of RAW-input for low-light detection are shown in Table I and Figure I, which extends Table 1 in the main paper. By further comparing the performance of CenterNet [13] trained on either brightened RGB-dark (by digital gains) or RAW-dark images from LOD dataset respectively, we can see nearly +5% AP gain can still be obtained by using RAW images as inputs, further verifying the advantage of the RAW-input detector design for low-light object detection.

A.2 Noise injection in the low-light synthetic pipeline.

Based on our low-light synthetic pipeline in Figure II, we further compare different kinds of noise models (*i.e.,* Gaussian, Poissonian-Gaussian [2, 3] and Physics-based noise model [13]) for noise injection. As shown in Table II, we can find that the physics-based noise model could

*Indicates equal contributions. † Corresponding author.

This work was supported by the National Natural Science Foundation of China under Grants No. 62171038, No. 61827901, No. 61936011, and No. 62088101.

yield better results than other noise models, suggesting the effectiveness of our implementation for noise injection.

Table I: The performance of CenterNet on the short-exposure sRGB and RAW images from our LOD, in which the training and testing are executed on the same data type. “RGB-dark" is the short-exposure sRGB data type, “RGB-dark*" is the brightened short-exposure sRGB data type, “RAW-dark" is the short-exposure RAW data type.

Data type	AP	$AP_{0.5}$	$AP_{0.75}$
RGB-dark	37.6	59.0	40.2
RGB-dark*	40.3	59.7	44.0
RAW-dark	**44.7**	**67.9**	**49.0**

A.3 Low-light recovery module (LRM) design.

Here, we discuss the architecture choice of our LRM. To verify the superiority of the LRM architecture adopted in the main paper, we replace it with another commonly used architecture for low-light enhancement (*i.e.,* UNet-style architectures), and then augment it into the CenterNet [15] backbone (DLA-34 [14]). As shown in Table III, the result shows that our proposed architecture outperforms all other architectures for LRM in terms of detection accuracy, which further endorses the superiority of our designed system. This extends Table 2 of the main paper.

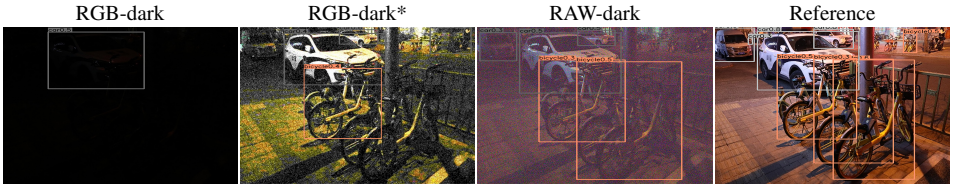
A.4 Evaluation of loss functions for training LRM.

We also evaluate some classical alternative loss functions to determine the best choice in the LRM module. As shown in Table IV, our baseline is the widely used L_2 (MSE) loss, but replacing the L_2 loss by L_1 produces better image enhancement results (higher PSNR/SSIM) and further improves the detection performance (higher AP) of our overall low-light detection system. Moreover, we find that combining L_1 loss and *perceptual* loss leads to the best result. Similar findings are also established in the Ours-cascade approach, but its performance is always worse than the LRM counterpart. These results also imply there is a positive correlation between the auxiliary recovery quality (in terms of PSNR/SSIM) and the final detection accuracy. This extends Table 2 in the main paper.

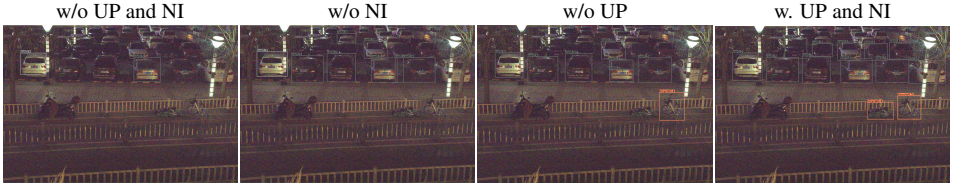
B More Quantitative Results.

B.1 Upper bound of our proposed approach on the LOD dataset.

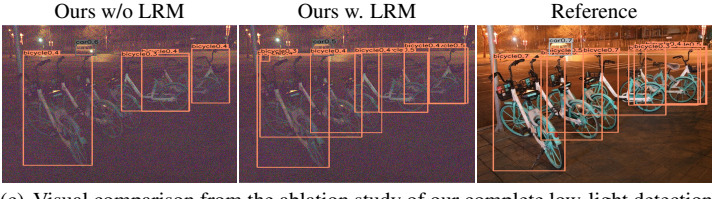
The LOD dataset contains long and short-exposure images in RAW and sRGB formats with annotations. Hence we could evaluate the results of detecting objects directly based on the long-exposure normal-light images to see the performance gap between short-exposure low-light inputs and long-exposure normal-light inputs. Table V (row 2 and 4) shows the AP gap between detecting on low-light and normal-light images is pretty large (11.9% AP), indicating the difficulty of low-light detection. Besides, we find (in row 1 and 2) that RAW-input also



(a) Visual analysis of RAW-input on our LOD dataset. *RGB-dark** is that the brightened short-exposure sRGB format image.



(b) Visual comparison from the ablation study of low-light synthetic pipeline that trained with and without *unprocessing* and *noise injection* operation(s) (based on COCO dataset).



(c) Visual comparison from the ablation study of our complete low-light detection system that trained with and without LRM (based on COCO dataset).

Figure I: Visual results of ablation study.

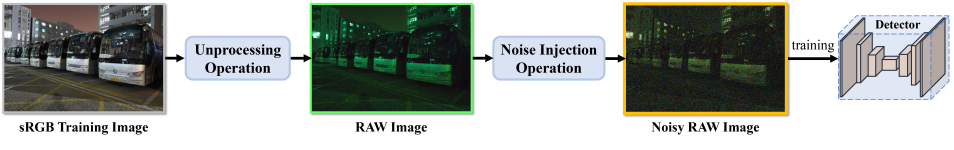


Figure II: Illustration of our low-light synthetic pipeline.

Table II: Quantitative analysis of noise injection model using CenterNet. “Gaussian" is the Gaussian noise model, “Poissonian-Gaussian" is the Poissonian-Gaussian noise model [3], “Physics-based" is the physics-based noise model [13].

Training Set	Noise Injection Model	AP	$AP_{0.5}$	$AP_{0.75}$
LOD	None	34.5	52.5	36.1
	Gaussian	38.0	60.5	39.8
	Poissonian-Gaussian	39.7	62.9	42.3
	Physics-based	**42.3**	**66.2**	**46.0**
COCO [9]	None	25.2	41.1	26.5
	Gaussian	26.4	50.2	25.5
	Poissonian-Gaussian	27.2	50.9	27.8
	Physics-based	**30.7**	**49.4**	**34.2**

Table III: Quantitative analysis of low-light recovery module architecture. The CenterNet with various LRM architectures are trained on LOD dataset to assess performance improvements.

Module Architecture	AP	$AP_{0.5}$	$AP_{0.75}$
None	42.3	66.2	46.0
U-net [10]	42.6	67.8	46.4
U-net + Channel attention [6]	43.5	69.1	47.0
LRM (**Ours**)	**44.9**	**71.5**	**46.7**

Table IV: Quantitative evaluation of different loss functions used for training LRM and our cascade approach.

Training Set	Method	Loss Function	PSNR	SSIM	AP
LOD	Ours-cascade	L_1	27.36	0.738	44.0
		L_1 + *perceptual*	27.49	0.740	44.4
	Ours (LRM)	L_2	27.27	0.730	44.3
		L_1	27.58	0.744	44.7
		L_1 + *perceptual*	**27.73**	**0.747**	**44.9**

helps improve the detection precision (+1.5% AP) when performing detection on normal-light images.

Table V: Upper bound of detection precision on our LOD dataset using CenterNet.

Dataset	Training Format	Testing Format	AP	AP_{50}	AP_{75}
LOD	RGB-normal	RGB-normal	55.1	77.9	62.2
	RAW-normal	RAW-normal	56.6	78.5	62.5
	RAW-normal	RAW-dark	35.8	57.6	38.5
	RAW-dark	RAW-dark	44.7	67.9	49.0

B.2 Method comparisons under the finetuning setting.

In Table 3 of the main paper, we assume the LOD dataset is never seen by any methods during training to unbiasedly justify the low-light detection performance in the uncontrolled real world. Here, we present the results under a new setting where the LOD training data is available for finetuning. The proposed approach is compared against the best two-step "enhance-then-detect" method (REDI [8]) in the main paper under this new setting. Specifically, for REDI, we finetune the detector based upon the enhanced low-light images from LOD training set; for ours, the detector is trained end-to-end using LOD training set. As shown in Table VI, our method (Ours*) still outperforms the two-step approach (REDI*) by a wide margin.

C More examples of LOD Dataset.

More examples of annotated images in our newly collected Low-light Object Detection (LOD) dataset are shown in Figure III. We emphasize that our dataset contains both indoor and outdoor images, and most images contain multiple instances and are a mixture of different objects. This extends Figure 4 in the main paper.

Table VI: Quantitative comparison of our method with the best two-step method (REDI) when the LOD training set is used for finetuning. The results finetuned on LOD are denoted by *, and the results using COCO data only (as in the main paper) are also shown as reference (w/o *).

Detector	Method	AP	AP_{50}	AP_{75}
CenterNet	None	12.7	19.4	14.1
	REDI [8]	26.2	35.1	28.8
	REDI*	29.9	45.8	31.1
	Ours	30.7	49.4	34.2
	Ours*	**44.7**	**67.9**	**49.0**

Figure III: More visual examples of our low-light object detection (LOD) dataset.

D More Qualitative Results.

D.1 Visual results of ablation studies.

Some visual results of ablation studies are provided in Figure I to further assess the performance of each operation and component. It can be seen that all operations and components in our low-light detection system help improve the detection performance and detect more objects. This complements Table 1 and Table 2 of the main paper.

D.2 Visual results of competing methods.

More visual comparison results on our LOD dataset images are shown in Figure IV. We compare our methods against representative two-step "Enhance-then-Detect" methods in very low light. It can be seen that all compared methods fail to restore the very low-light images and further provide perfect detection results, but our method accurately detects more objects. This extends Figure 5 of the main paper.

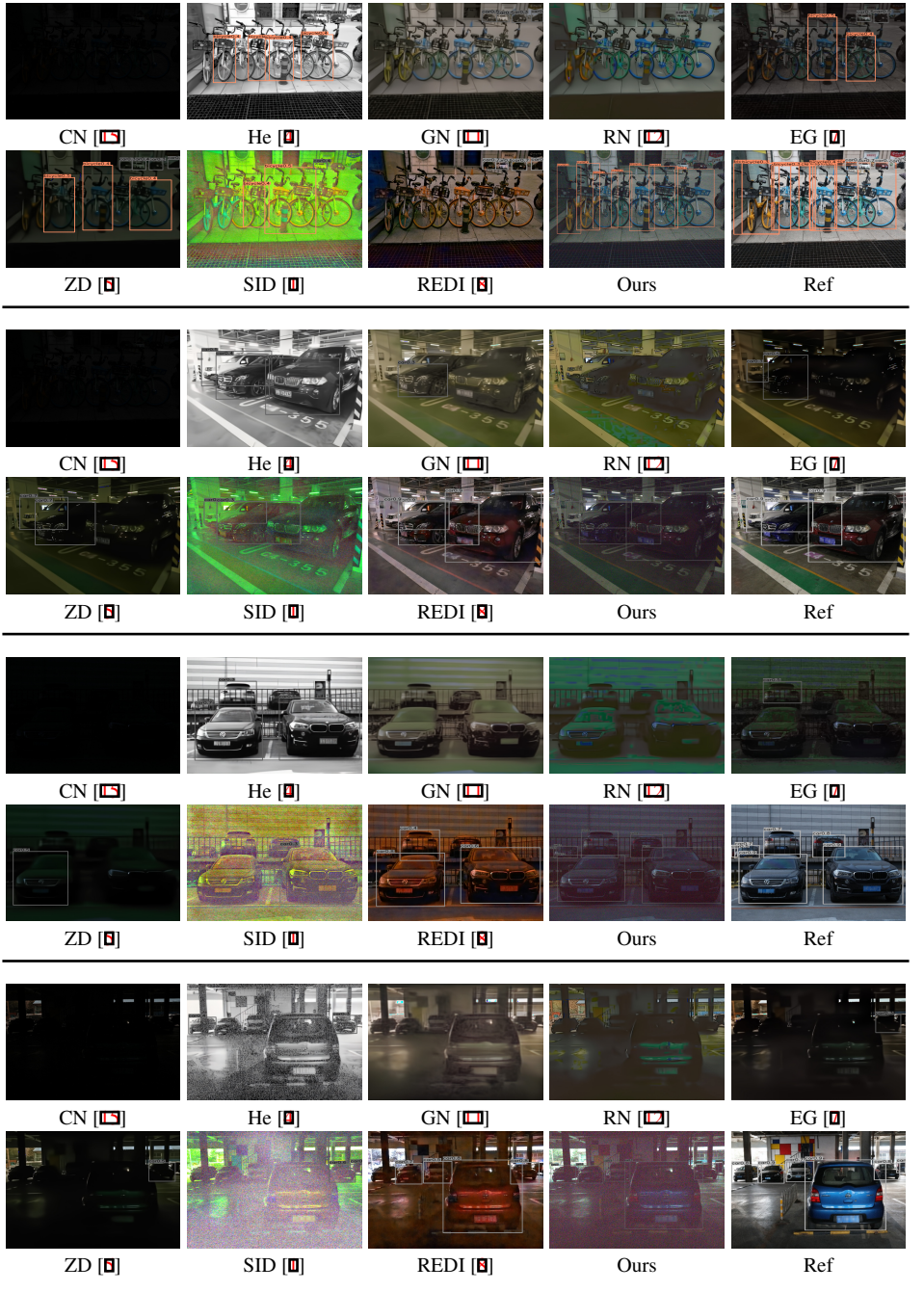


Figure IV: Visual comparison on our LOD dataset. **(Please zoom in to see details.)**

References

[1] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

[2] Alessandro Foi. Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing*, 89(12):2609–2629, 2009.

[3] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008.

[4] Rafael C Gonzalez, Richard E Woods, et al. Digital image processing, 2002.

[5] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[7] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[8] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3487–3497, 2021.

[9] 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.

[10] 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.

[11] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 751–755. IEEE, 2018.

[12] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[13] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020.

[14] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018.

[15] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.