Crafting Object Detection in Very Low Light

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

Abstract

Over the last decade, object detection, as a leading application in computer vision, has been intensively studied, heavily engineered and widely applicable to everyday life. However, existing object detection algorithms could easily break down under very dim environments, due to significantly low signal-to-noise ratio (SNR). Prepending a low-light image enhancement step before detection, as a common practice, increases the computation cost substantially, yet still does not yield satisfactory results. In this paper, we systematically investigate object detection in very low light and identify several design principles that are essential to the low-light detection system. Based upon these criteria, we design a practical low-light detection system that utilizes a realistic low-light synthetic pipeline as well as an auxiliary low-light recovery module. The former can transform any labeled images from existing object detection datasets into their low-light counterparts to facilitate end-to-end training, while the latter can boost the low-light detection performance without adding additional computation cost at inference. Furthermore, we capture a real-world low-light object detection dataset, containing more than two thousand paired low/normal-light images with instance-level annotations to support this line of work. Extensive experiments collectively show the promising results of our designed detection system in very low light, paving the way for real-world object detection in the dark. Our dataset are publicly available at https://github.com/ying-fu/LODDataset.

1 Introduction

Recent years have witnessed tremendous advancements in generic object detection [9, 39, 48, 50, 54, 62], a representative application in computer vision, which has then led to the boom of industry-level solutions with applications in a wide range of domains. The object detection functionality is now widely accessible in daily life, *e.g.,* smartphone cameras, surveillance systems, and autonomous driving [1, 10, 25, 36, 47, 59].

*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.

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

(a) (b) (c) (d)

Figure 1: **Low-light object detection results using different schemes**, including (a) direct detection on a low-light image; (b) detection on the image enhanced by the state-of-the-art extremely low-light enhancement algorithm REDI [13]; (c) detection using our proposed system on the amplified noisy image; and (d) detection on the corresponding long-exposure reference. All the results are given by the same detection architecture, *i.e.,* CenterNet [63].

Although existing detection systems can produce reliable results in daytime scenarios with sufficient illuminance, they often fail to accurately detect objects in pretty dim environments, where detailed contents in captured images are barely discernible. A common practice for low-light detection is to prepend a low-light enhancement step before detection [12, 40, 60]. In this way, the "buried" scene information is first restored by an enhancement algorithm and thus is more likely to be recognized by the subsequent detector. However, due to the scarcity of annotated low-light images, the two modules (*i.e.,* the enhancer and the detector) in these two-step methods are usually optimized separately with distinct goals. Using enhanced images as inputs of the detector, as shown in Figure 1, therefore often does not result in desired object detection performance. Besides, such a cascaded "enhance-then-detect" system incurs a larger computational burden owing to the additional enhancement step, which hinders its deployment on resource-constraint devices such as mobile phones.

Since low-light environment is an integral part of our everyday activities, building a robust low-light object detection system is definitely demanded. Such a system should meet the following criteria to ensure its practical use: 1) the system should be able to extract hidden scene information as much as possible, and 2) detect a large number of different objects in the wild; 3) its computation overhead should be lightweight, for potential real-time applications on low-budget computing devices.

Crafting a low-light detection system that fulfills these requirements could be quite non-trivial, owing to the intrinsic difficulties faced under very low-light settings. For example, due to limited photon count and inescapable noise, the weak signals present in typical 8-bit quantized JPEG camera outputs, are often severely distorted, and which sometimes even leads to permanent loss of scene information. Besides, the available low-light images with bounding-box annotations are rather limited, which substantially restricts the development of learning-based neural network detectors for low-light object detection. Collecting rich high-quality labeled training samples that cover a variety of object classes in the dark is tremendously labor-intensive and expensive.

In this work, we aim to design a practical low-light object detection system in accordance with the aforementioned requirements. As shown in Figure 2, our crafted system makes full use of a realistic low-light synthetic pipeline and a “zero-cost” low-light recovery module to address the data scarcity and the low-light degradation issues respectively. The former consists of an unprocessing module and a noise injection module, which is capable of converting any labelled images from off-the-shelf object detection datasets into their low-light counterparts, thereby significantly alleviating the burden of labor requirements for capturing a massive amount of real-world training data. The latter could be integrated into a given detector to serve

as another output branch for low-light recovery, such that the whole network can be trained to simultaneously learn about visibility enhancement and object detection. Once trained, the low-light recovery module is no longer activated at inference, but indeed improves the detector's low-light accuracy thanks to the mutual benefits of these two tasks under low light. To systematically examine the performance of our system under real low-light environments, we additionally capture and label a low-light object detection (LOD) dataset covering more than two thousand real low-light scenes. Extensive experiments show the promising results of our system in very low light, consistently outperforming prior arts in terms of both detection accuracy and computation cost.

To summarize, our main contributions are as follows:

1. We formulate a realistic low-light synthetic pipeline that can transform annotated images from existing object detection datasets into their low-light counterparts. It facilitates end-to-end training meanwhile bypassing the labor-intensive real data collection.
2. We present an auxiliary low-light recovery module to help boost the detector accuracy under very low illuminance. The proposed module is only activated at the training phase and therefore does not impose additional computation cost at inference.
3. We collect a real-world low-light object detection dataset, which covers more than two thousand scenes with precise instance-level annotations, in order to systematically assess the low-light detection performance of existing schemes.

2 Related Work

Generic object detection. Existing object detectors can be roughly divided into two categories: two-stage and one-stage. A two-stage detector, *e.g.,* RCNN [17], usually generates foreground proposals first and then classifies them. In spite of high accuracy, its low inference speed motivates follow-up works, *e.g.,* [16, 22, 50] for acceleration, among which Faster RCNN [50] has become a solid baseline for successive researches [3, 4, 23, 38, 41]. Differing from two-stage methods, a one-stage detector, *e.g.,* YOLO [49], directly predicts class probabilities and bounding box offsets from full images with a single feed-forward CNN, which runs much faster at the cost of localization precision. Many improvements then have been suggested in the following works by *e.g.,* utilizing multiscale features [42], addressing sample imbalance problem [39] or exploiting keypoints representation [34, 53], to further boost the detection accuracy. All these generic object detection methods, however, are primarily designed for high-visibility inputs. Their performance almost always degenerates drastically under dark regimes with extremely low visibility.

Low-light image enhancement. Enhancing the visibility of images captured under low light has attracted many interests in the low-level vision community. Traditional methods rely on histogram mapping [11, 26, 35] or Retinex-theory-inspired optimization [15, 21] to adjust illumination adaptively to avoid over- and under-enhancement, but they do not consider the inherent noise issue raised naturally in low light. Modern learning-based methods employ convolutional neural networks to learn image brightening and noise suppression jointly from data [6, 8, 9, 27, 28, 43, 45, 55, 56, 60]. Despite promising results have been obtained by *e.g.,* leveraging synthetically darkened data [45], unpaired bright and low-light images [28], short- and long-exposure image pairs [6] and image-specific curve estimation [20], none of them is specifically optimized for the downstream object detection in the dark.

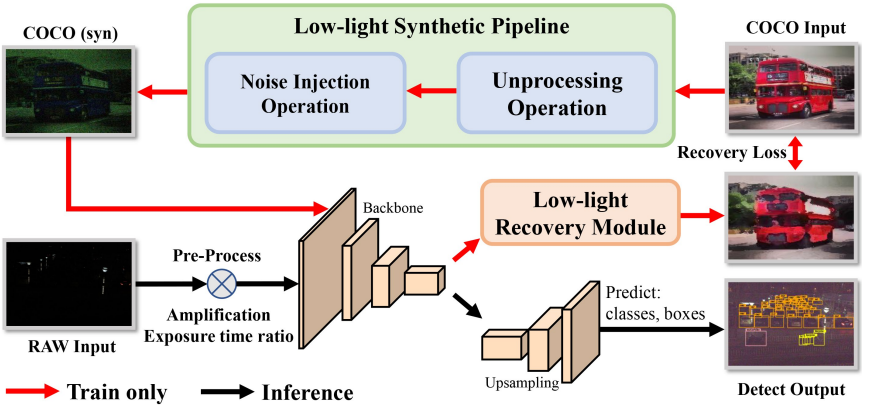

Figure 2: **Overview of our designed system**. On the basis of a given detector backbone, a low-light recovery module (LRM) is introduced for visibility enhancement. The whole system is trained with data generated by a low-light synthetic pipeline, and LRM is only activated during training.

Object detection in the dark. General object detection is a well-developed topic in high-level vision. However, due to the limited datasets and domain gap between synthetic and real data, object detection in the dark still faces relatively slow progress. ExDark [44] collects a low-light image dataset with object-level annotations, but it only contains limited classes and is too small to train the data-hungry deep learning-based detectors in general. [53] proposes a method of domain adaptation for merging pretrained models in both low-light enhancement and detection domains, and reports results of detecting objects from short-exposure low-light RAW images. Zero-DCE [20] demonstrates its potential benefits to face detection in the dark. These methods are all two-step and therefore are suboptimal for low-light detection. In this work, we propose an end-to-end low-light detection system that is optimized directly for the low-light detection performance, meanwhile circumventing the unwieldy two-step pipeline.

3 Low-light Detection System Design

Figure 2 summarizes our proposed low-light detection system. It augments a common detection pipeline by a low-light synthetic pipeline and a low-light recovery module to help address the challenges (*i.e.,* the limited training data and the degraded image feature) of object detection in very low light. Unlike common detectors that usually execute detection on 8-bit sRGB–JPEG camera outputs, our system advocates start from RAW, which is an uncompressed, minimally processed image format representing the response from the camera sensor directly [31, 46]. Starting from (typically 14-bit) RAW enables the system to extract signals which are otherwise severely distorted or destroyed at 8-bit quantized sRGB–JPEG camera outputs due to pretty low SNR [1] [6, 7]. Besides, our system is *detector-agnostic*, suggesting that our system design is applicable to various neural network detectors. In the following, we mainly prototype the proposed system using the CenterNet [53], one of the most commonly used one-stage detectors with low computation complexity.

3.1 Low-light synthetic pipeline

[1] We refer interested readers to *suppl. material* for a more detailed discussion of why we use RAW as input.

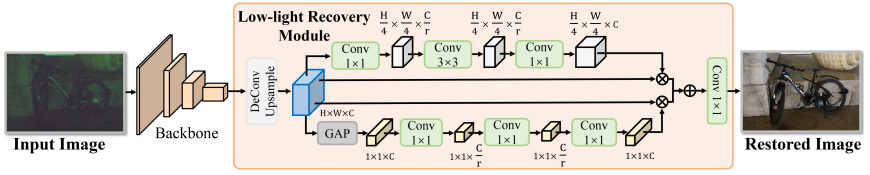

Figure 3: **Illustration of our auxiliary low-light recovery module**, which is implemented by an upsampling layer and a low-light enhancement attention submodule.

One of the major bottlenecks of low-light detection is the scarcity of high-quality labeled training data that are necessary for learning-based neural network detectors. To mitigate this issue, we formulate a low-light synthetic pipeline to synthesize realistic low-light images based upon annotated images from existing object detection datasets. As shown in Figure 2, this pipeline is composed of two operations: an unprocessing operation and a noise injection operation, which are implemented by the methods in [2] and [53] respectively.

Unprocessing. Given an input sRGB image, the pipeline first "inverts" the on-camera image signal processing (ISP) to synthesize its RAW format counterpart. This would involve sequentially canceling the effects of image processing transformations including tone mapping, gamma correction, color correction, white balance, and demosaicking. Note that the internet images in the real world are almost processed by intractable unknown ISPs, which renders the accurate inversion extremely difficult if not impossible. Fortunately, our study suggests that there is no need to precisely retrieve the original RAW image given the processed sRGB input, instead, a rough approximation (*i.e.,* the unprocessing operation [2]) could reach sufficiently good results for training RAW-input detectors.

Noise injection. After the images are unprocessed into RAW, realistic noise is injected into the "unprocessed" raw images to simulate the noisy images captured under very low illuminance. In contrast to adopting the widely used Poissonian-Gaussian noise model (*a.k.a.* heteroscedastic Gaussian model) [13, 14], we find employing a recently proposed physics-based noise model [53] could yield more accurate results in simulating the complex real noise structure in very low light. The noise model in [53] derives from the inherent process of electronic imaging by considering how photons go through several stages (*i.e.,* electrons, voltage, and digital number) to accurately characterize the real noise structures. It takes photon shot noise, read noise, banding pattern noise, and quantization noise into account, which are modeled by Poisson distribution, long-tailed Tukey-lambda distribution [30], Gaussian distribution, and uniform distribution respectively.

We should emphasize though we leverage existing tools to build our low-light synthetic pipeline, *the underlying motivation/objectives are entirely different*: the methods in [2] and [53] are originally designed for learned RAW image denoising in low-level vision, in contrast, we combine and repurpose them for low-light object detection in high-level vision. Our approach bridges the knowledge from low-level and high-level vision communities, which sheds light on a new way towards practical object detection in very low light.

3.2 Auxiliary low-light recovery module

Once the data is synthesized from the aforementioned low-light synthetic pipeline, any off-the-shelf detectors can be employed to detect objects at night via retraining or finetuning on new

data. Nevertheless, we find the feature extractors (backbone) of such detectors often struggle to retrieve discriminative features for accurate prediction, since the image content is largely overwhelmed in the dark, sometimes even human observers cannot recognize anything from exclusively dark. To restore distinguishable semantic features produced at feature extractor from low-light images, we propose an auxiliary low-light recovery module (LRM), which is augmented into the detector backbone as an additional output branch. The system thus as in Figure 2 has two heads: the original detection head for bounding-box and object class prediction and the low-light recovery head for visibility enhancement. These two tasks are then jointly learned through the training phase.

Note that the design motivation of LRM is derived from the spirit of *multitask learning* [5, 52], which is a subfield of machine learning in which multiple related tasks are learned in parallel while using a shared representation for mutual benefits. In our circumstance, intuitively, the learning of low-light visibility enhancement task would encourage the network to find a representation robust to the low-light degradation, such that the detection head could use restored "clean" features for better prediction.

In principle, the LRM can be implemented by any decoder-style network architecture connected to the detector backbone (*i.e.,* the encoder). Here, we design a simple yet effective architecture whose key component is a low-light enhancement attention submodule. It promotes the network to use global information to selectively emphasize informative semantic responses and suppress less useful ones. Specifically, as shown in Figure 3, the features from the upsampling layer would be fed into two parallel attention branches, *i.e.,* pixel and channel attention [24] branches. On the pixel side, we use three cascaded convolutions to learn pixel-wise attention weights for each pixel of the input feature map. Next, each pixel of the input feature map is multiplied by the corresponding pixel attention weight to get the final feature map of the pixel attention branch. On the channel side, the input feature map is first compressed into a one-dimensional vector by global average pooling, and the channel-wise attention weights are then generated and multiplied into each input feature channel to obtain the branch output. The outputs from these two branches are finally aggregated and followed by a 1×1 convolution to reconstruct high-visibility images.

The LRM is trained by a combination of L_1 loss and perceptual loss $\mathcal{L}_{per}$ [29, 57][2], *i.e.,*

$$\mathcal{L}_{LRM} = \left\| X - \widehat{X} \right\|_1 + \mathcal{L}_{per}(X - \widehat{X}), \tag{1}$$

where X denotes clean images, and $\widehat{X}$ indicates recovered images. Together with the classification and localization losses attached in the detection head, it provides useful mixed gradients back-propagated into the backbone, making it a robust representation for low-light prediction. The overall loss for training our low-light detection system is thus defined by:

$$\mathcal{L}_{total} = \lambda_{LRM}\mathcal{L}_{LRM} + \lambda_{cla}\mathcal{L}_{cla} + \lambda_{loc}\mathcal{L}_{loc}, \tag{2}$$

where we set λ_{LRM}=1 and λ_{cla}=λ_{loc}=10 to balance loss magnitudes empirically. Once trained, the LRM is no longer triggered at inference, thus incurring "*zero cost*" in terms of running time, but increases the low-light detection accuracy implicitly and appreciably.

[2]The specifics of the adopted perceptual loss is the same as the one in [57]. Other choices of loss functions for training LRM are discussed in the *suppl. material*.

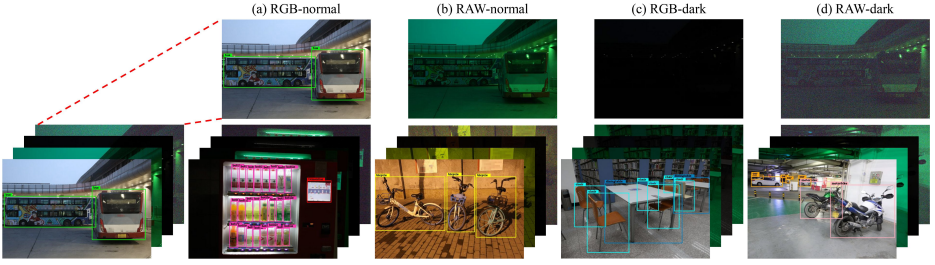

Figure 4: **Example scenes in our LOD dataset**. Four image instances (long-exposure normal-light and short-exposure low-light images in both RAW and sRGB formats) are captured for each scene.

4 Low-light Object Detection (LOD) Dataset

To systematically study the effectiveness of the proposed system, we collect a new Low-light-Object-Detection (**LOD**) dataset using a Canon EOS 5D Mark IV camera. LOD is recorded in both RAW and sRGB-JPEG formats, containing paired short-exposure and corresponding long-exposure reference images. The camera was mounted on a sturdy tripod and controlled remotely via a mobile app to avoid vibration. At each data collection point, a long-exposure reference image at the base ISO was firstly taken, followed by short-exposure images whose exposure time was deliberately decreased by low-light factors (*e.g.,* $\times$50) to simulate very low-light conditions. We choose four ISO levels (800, 1600, 3200, 6400) and six low-light factors (10, 20, 30, 40, 50, 100) in data capture, resulting in **2230** image pairs in total.

Furthermore, to support low-light object detection applications, we provide precise instance-level annotations carried out by professional annotators for each collected image, yielding **9726** labeled instances of 8 common object classes (car, motorbike, bicycle, chair, diningtable, bottle, tvmonitor, bus) in summary[3]. Some examples of annotated images in our dataset are shown in Figure 4. We believe the LOD dataset can not only serve as a real-world benchmark for method comparisons but also facilitate the research in low-light object detection domain.

5 Experiments

5.1 Implementation details

For prototyping purposes, our system is instantiated by CenterNet [63] as detector whose backbone is a modified DLA-34 [61]. The system is trained by synthetic low-light RAW-RGB[4] images generated from COCO [37] dataset using our low-light synthetic pipeline[5]. During training, we follow [63] to use a fixed input resolution of 512$\times$512, and use random flip, random scaling, cropping, and color jittering as data augmentation. We choose Adam [32] optimizer and train with a batch-size of 32 and learning rate 5e-4 for 40 epochs, with learning rate dropped by 10$\times$ at 10 and 30 epochs, respectively. To make the detector quickly adapt to low-light settings, we use COCO pretrained model (generic detection model) as initialization.

[3]The labeling process is conducted on long-exposure reference images, which is then transferred into corresponding aligned short-exposure ones directly.

[4]We use demosaicked 3-channel RAW-RGB images as inputs instead of the Bayer RAW images to make the detector compatible to sRGB inputs. In the following, we refer "RAW" and "RAW-RGB" interchangeably.

[5]We use COCO samples belonging to the same 8 object classes in the LOD dataset.

Table 1: Quantitative results of different training schemes on our LOD testing set. "UP" and "NI" indicate *unprocessing* and *noise injection* operations in the low-light synthetic pipeline, respectively.

Data Type	Training Set	Testing Set	Operation		AP	AP_{50}	AP_{75}
			UP	NI			
REAL	LOD RGB-dark	LOD RGB-dark	-	-	37.6	59.0	40.2
	LOD RAW-dark	LOD RAW-dark	-	-	**44.7**	**67.9**	**49.0**
SYNTHETIC	LOD RGB-normal	LOD RAW-dark	-	-	31.7	49.8	33.4
			√	-	34.5	52.5	36.1
			-	√	39.8	62.1	41.8
			√	√	**42.3**	**66.2**	**46.0**
	COCO RGB-normal	LOD RAW-dark	-	-	23.0	38.5	24.4
			√	-	25.2	41.1	26.5
			-	√	28.8	48.4	31.3
			√	√	**30.7**	**49.4**	**34.2**

Table 2: Quantitative evaluation of our auxiliary low-light recovery module. The models are trained on synthetic data from either LOD or COCO using the complete low-light synthetic pipeline.

(a) Evaluation of low-light detection performance.

Train set	Method	AP	AP_{50}	AP_{75}
LOD	Ours w/o LRM	42.3	66.2	46.0
	Ours	**44.9**	**71.7**	**48.2**
	Ours-cascade	44.4	70.5	46.7
COCO	Ours w/o LRM	30.7	49.4	34.2
	Ours	**35.6**	**57.1**	**42.9**
	Ours-cascade	34.4	55.3	37.9

(b) Evaluation of computation cost at inference.

Method	GFLOPs	Time(ms)
Ours w/o LRM	26.2	84
Ours	26.2	84
Ours-cascade	53.6	224

The real low-light detection performance is evaluated on the LOD dataset, in which the total 2230 image pairs are randomly split into a train set of 1830 pairs and a test set of 400 pairs. We note though the LOD is used for training/finetuning in some following experiments, this mainly serves the purpose of system design validation. To unbiasedly justify the system performance in real-world low-light environments, *we assume the LOD dataset is never seen by any methods during training*, such that the results accurately reflect the practical use in the real world. As for evaluation metrics, we adopt the average precision over all IOU thresholds (AP), AP at IOU thresholds 0.5 (AP_{50}) and 0.75 (AP_{75}) as in COCO standard.

5.2 Ablation study

In this section, experiments are conducted to evaluate the RAW-input detection system design, the low-light synthetic pipeline, as well as the auxiliary low-light recovery module. To this end, we compare the performance of a given detector (CenterNet) trained with different schemes utilizing either pure real data or synthetic data. For real data, we use short-exposure low-light images from our LOD dataset in RAW/sRGB format (denoted by "RAW/RGB-dark"). For synthetic data, we generate synthetic low-light images from long-exposure normal-light sRGB ("RGB-normal") images from LOD/COCO dataset. The results of these training schemes are provided in Table 1.

RAW-RGB *v.s.* sRGB. Table 1 (row 1 and 2) shows simply using RAW images instead of sRGB images can significantly improve the detection accuracy (+7.1% AP) in low-light conditions. It indicates that RAW-input design indeed enables the detector to extract signals which are otherwise destroyed at the commonly used sRGB inputs due to pretty low SNR.

Low-light synthetic pipeline. Next, we assess the effectiveness of the proposed low-light

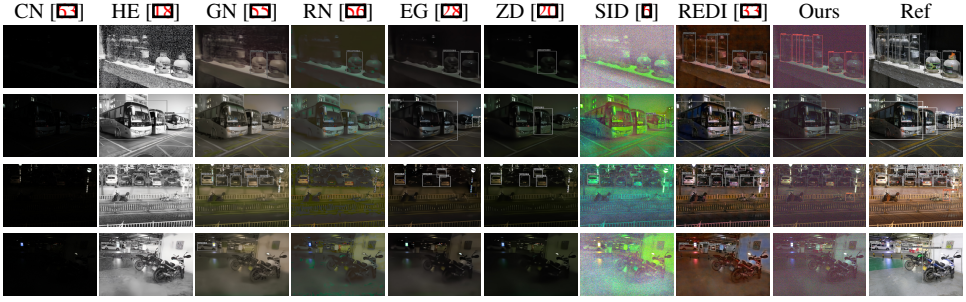

Figure 5: Visual comparison on our LOD dataset. **(Please zoom in to see details.)**

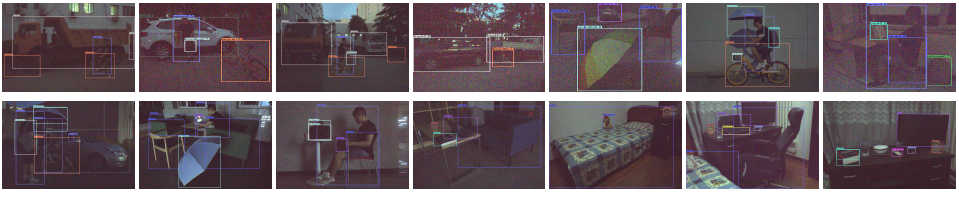

Figure 6: Visual examples of low-light object detection using our system in the wild (beyond eight classes in the LOD dataset).

synthetic pipeline. As shown in Table 1 (row 3-6), by invoking unprocessing operation, we can obtain RAW images from sRGB images for training, this leads to 2.8% AP performance improvement on LOD RAW-dark testing set. Given noise is a key difference between normal images and low-light images, we employ noise injection operation to help detectors better deal with it. This boosts the performance by 8.1% AP. Furthermore, calling these two operations together can produce realistic low-light RAW images, which increases the detection accuracy by 10.6% AP. Notice that this result is even comparable to the detector trained with real low-light data (42.3% v.s. 44.7% in AP). Such trends are consistently observed when adopting the COCO dataset, which implies the generality of our low-light synthetic pipeline.

Low-light recovery module (LRM). Finally, we test the efficacy of our proposed LRM. As shown in Table 2, utilizing LRM could introduce up to +3.8% AP gain, which clearly demonstrates its benefits. We also design a cascade two-step method (Ours-cascade) as a competitive baseline against our LRM-based approach leveraging low-light domain knowledge. For this cascade method, a commonly used UNet-style [6, 51] architecture for low-light enhancement is connected to CenterNet for joint training. The results suggest that the LRM approach outperforms the cascade approach in terms of both detection accuracy and computation cost, which further endorses the superiority of our designed system.

5.3 Comparisons against other methods

Here, we compare our approach against existing two-step "enhance-then-detect" methods. We choose representative traditional (histogram equalization [18]) and learning-based (GLAD-Net [65], Retinex-Net [56], EnlightenGAN [28], Zero-DCE [20], SID [6], REDI [33]) methods as enhancers and adopt the same detector (either CenterNet [63] or FCOS [64]) for fairness. Note some competing methods (except SID [6] and REDI [33]) have no explicit denoising mechanism, which might not be well-suited for extreme low-light conditions. For fairer comparisons, we additionally perform low-light denoising by the state-of-the-art

Table 3: Numerical comparisons on our LOD dataset. **The detectors are all trained using COCO data only in order to assess the low-light detection performance in the uncontrolled real world.**

DETECTOR	METHOD	AP on different Exposure Ratios						
		×10	×20	×30	×40	×50	×100	Avg
CenterNet [63]	None	23.3	16.7	14.4	14.1	13.4	3.6	14.3
	HE [43] + Denoising [49]	32.1	25.7	22.6	20.4	18.2	8.4	21.2
	GLADNet [55] + Denoising	28.9	20.7	14.8	12.1	11.9	7.7	16.0
	Retinex-Net [58] + Denoising	18.5	12.9	12.7	9.8	7.1	3.6	10.8
	EnlightenGAN [28] + Denoising	33.1	26.5	22.7	17.9	16.2	8.1	20.8
	Zero-DCE [20] + Denoising	32.5	25.3	23.4	21.5	17.8	8.9	21.6
	SID [6]	25.8	20.0	16.4	15.1	13.2	6.7	16.2
	REDI [33]	33.6	30.2	26.1	24.6	23.4	14.1	25.4
	Ours	**38.5**	**31.7**	**29.3**	**27.8**	**27.1**	**18.1**	**28.8**
FCOS [62]	None	20.0	15.1	13.2	12.7	12.1	3.8	12.8
	HE [43] + Denoising [49]	28.1	21.7	19.8	16.7	16.3	8.8	18.6
	GLADNet [55] + Denoising	23.8	17.1	11.8	10.3	9.1	5.1	12.9
	Retinex-Net [58] + Denoising	15.8	11.1	10.8	9.5	6.5	3.8	9.6
	EnlightenGAN [28] + Denoising	29.4	23.1	21.1	16.5	14.9	7.3	18.7
	Zero-DCE [20] + Denoising	28.1	21.2	20.1	18.9	16.1	8.7	18.9
	SID [6]	21.8	18.1	15.7	14.3	12.1	6.5	14.8
	REDI [33]	30.7	25.4	22.3	19.3	18.1	14.7	21.8
	Ours	**34.1**	**29.1**	**27.5**	**25.3**	**24.2**	**14.8**	**25.8**

self-guided network [49] after the enhancement, before the final detection[6]. This denoising module is attributed as a part of the enhancement step in these two-step methods.

Both numerical and visual results are presented in Table 3 and Figure 5 respectively. It can be seen that our method achieves the best performance under all low-light settings among all competing methods using either CenterNet or FCOS detector. Obviously, our method can accurately detect more objects in very low light, *e.g.,* motorbike and bicycle in the third row of Figure 5, while other competing methods often fail to detect these two object classes. This not only demonstrates the advantages of our end-to-end system design over two-step approaches but also supports the versatility of our system for a rich pool of object detectors.

Furthermore, we train our system using the synthetic low-light COCO data covering all 80 object classes and investigate its low-light detection performance in the wild. The visual results in Figure 6 demonstrate our system can effectively detect varieties of objects in low light, which is not limited to the eight object classes in our LOD dataset. *This stands out as a major merit that is not available when training on real low-light data from LOD*. Thanks to the low-light synthetic pipeline, our system could easily leverage abundant annotated images from existing detection datasets for training, thereby supporting the development of a practical system that is capable of detecting miscellaneous objects in the dark.

6 Conclusion

We have crafted the first end-to-end low-light object detection system that can precisely detect diverse objects in the dark. We also provide a large-scale real-world low-light object detection benchmark and demonstrate our system can consistently surpass existing approaches by a large margin. We hope that our dataset and the experimental findings can stimulate more works and open new opportunities for future research.

[6]We find adding such a denoising module generally improves the performance of two-step methods.

References

[1] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] 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.

[7] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3185–3194, 2019.

[8] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 354–363, 2021.

[9] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018.

[10] Seoyoung Choi, Eli Salter, Xuyun Zhang, and Burkhard C Wünsche. Bird eyes: A cloud-based object detection system for customisable surveillance. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2018.

[11] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE Transactions on Image Processing*, 15(5):1143–1152, 2006.

[12] 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.

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

[14] 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.

[15] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2782–2790, 2016.

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

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

[19] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019.

[20] 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.

[21] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

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

[25] Ling Hu and Qiang Ni. Iot-driven automated object detection algorithm for urban surveillance systems in smart cities. *IEEE Internet of Things Journal*, 5(2):747–754, 2017.

[26] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[28] 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.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[30] Brian L Joiner and Joan R Rosenblatt. Some properties of the range in samples from tukey's symmetric lambda distributions. *Journal of the American Statistical Association*, 66(334):394–399, 1971.

[31] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2289–2302, 2012.

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

[33] 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.

[34] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision*, pages 734–750, 2018.

[35] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22 (12):5372–5384, 2013.

[36] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2019.

[37] 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.

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[40] Ding Liu, Bihan Wen, Jianbo Jiao, Xianming Liu, Zhangyang Wang, and Thomas S Huang. Connecting image denoising and high-level vision tasks via deep learning. *IEEE Transactions on Image Processing*, 29:3695–3706, 2020.

[41] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[43] Yihao Liu, Jingwen He, Xiangyu Chen, Zhengwen Zhang, Hengyuan Zhao, Chao Dong, and Yu Qiao. Very lightweight photo retouching network with conditional sequential modulation. *arXiv preprint arXiv:2104.06279*, 2021.

[44] 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.

[45] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[46] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. *International journal of computer vision*, 126(6):637–650, 2018.

[47] Hazem Rashed, Eslam Mohamed, Ganesh Sistu, Varun Ravi Kumar, Ciaran Eising, Ahmad El-Sallab, and Senthil Yogamani. Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2272–2280, 2021.

[48] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[51] 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.

[52] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[53] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *European Conference on Computer Vision*, pages 345–359. Springer, 2020.

[54] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9627–9636, 2019.

[55] 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.

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

[57] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[58] 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.

[59] Marie Yahiaoui, Hazem Rashed, Letizia Mariotti, Ganesh Sistu, Ian Clancy, Lucie Yahiaoui, Varun Ravi Kumar, and Senthil Yogamani. Fisheyemodnet: Moving object detection on surround-view cameras for autonomous driving. *arXiv preprint arXiv:1908.11789*, 2019.

[60] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[61] 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.

[62] Xiaosong Zhang, Fang Wan, Chang Liu, Xiangyang Ji, and Qixiang Ye. Learning to match anchors for visual object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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