IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 24, 2020, accepted July 2, 2020, date of publication July 7, 2020, date of current version July 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007610

Making of Night Vision: Object Detection Under Low-Illumination

YUXUAN XIAO, AIWEN JIANG, (Member, IEEE), JIHUA YE, AND MING-WEN WANG
School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China

Corresponding authors: Aiwen Jiang (jiangaiwen@jxnu.edu.cn) and Ming-Wen Wang (mwwang@jxnu.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 61966018 and Grant 61876074.

ABSTRACT Object detection has so far achieved great success. However, almost all of current state-of-the-art methods focus on images with normal illumination, while object detection under low-illumination is often ignored. In this paper, we have extensively investigated several important issues related to the challenge low-illumination detection task, such as the importance of illumination on detection, the applicabilities of illumination enhancement on low-illumination object detection task, and the influences of illumination balanced dataset and model's parameters initialization, etc. We further have proposed a Night Vision Detector (NVD) with specifically designed feature pyramid network and context fusion network for object detection under low-illuminance. Through conducting comprehensive experiments on a public real low-illuminance scene dataset ExDARK and a selected normal-illumination counterpart COCO*, we on one hand have reached some valuable conclusions for reference, on the other hand, have found specific solutions for low-illumination object detection. Our strategy improves detection performance by 0.5%∼2.8% higher than basic model on all standard COCO evaluation criterions. Our work can be taken as effective baseline and shed light to future studies on low-illumination detection.

INDEX TERMS Object detection, image enhancement, low-illumination.

I. INTRODUCTION

Low illuminance environment is closely related to our life. We spend almost half of every day in low illuminance environment. It brings us many inconveniences, especially in the field of security where the obtained low-illumination image often contains valuable information. In low-illumination environment, due to the dim light or insufficient exposure, the low-illumination image has problems of low brightness, low contrast and noise. One of straightest solutions is to improve hardware, such as using infrared monitoring or increasing the aperture of the camera. However, these hardware improvements will make the cost too high. Therefore, much research is still focused on software algorithmic solutions.

Most of current works on low-illumination focus on image enhancement, aiming to improve low-level visual qualities of images. However, for high-level tasks, such as object detection in low-illumination environments, so far have not been received enough attention. Works on object detection under low-illumination as shown in Figure 1 are still scarce.

Object detection in low-illumination environments is very challenging. Due to insufficient reflected light under low-illumination, image captured contains a lot of dark areas and noise. Though many successful object detection algorithms have been proposed with developments of deep learning, most state-of-the-art object detectors, such as Faster-RCNN [1], SSD [2] etc., cannot perform their best performance under low-illumination conditions. Even with additional light source, due to the uneven distribution of brightness, it is still impossible to distinguish object's details. The underlying reason we think is that current mainstream detectors are designed for normal-illumination data. So far, there is no special solution for vision tasks in low-illumination environments.

As a natural idea, image enhancement can be straightforwardly taken as pre-processing step before low-illumination detection. Therefore, in this paper, we follow to attempt the idea to implement different low-illumination enhancement algorithms on ExDARK, a real low-illuminance scene dataset which was lately developed by Loh and Chan [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

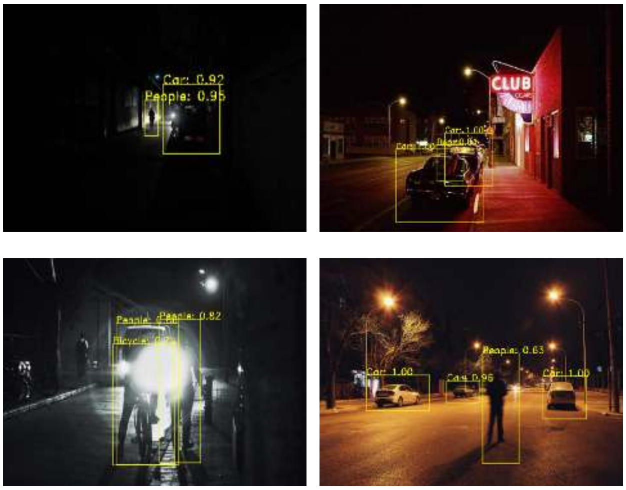


FIGURE 1. An example of low-illumination objects detection. Our detector have achieved amazing results in some common scenes of low-illumination.

We train state-of-the-art object detection models on the illumination-enhanced data. However, we are depressed to find that the detection performance of model trained on the enhanced data is not better than the performance of model trained directly on original dark data. The experiments frustratedly imply that most of current image enhancement algorithms though achieve visually pleasing results, could not meet low-illumination computer vision tasks. Therefore, we try to look into the causes and give some explanations through visualizing impacts of illumination on feature preservations at different convolution levels.

Going deeper, without loss of generality, we develop a night vision detector based on a representative state-of-the-art detector RFB-Net [4]. We introduce a specifically designed feature pyramid network into backbone layers of RFB-Net for hierarchical feature fusion. Context information are fused to compensate loss of low-level textural/contour information in deeper layers under low-illumination conditions. The experiments on ExDARK dataset demonstrate that our night vision detector achieves satisfied improvements on average detection precisions.

The contributions of this paper are summarized as followings:

- We have visualized impacts of illumination on feature preservations at different convolution levels and explained the underlying reasons that why illumination greatly affects the performance of detectors.
- We have evaluated a two-step strategy that performs detection on illumination-enhanced images. The experimental results give us some frustrated experiences that most of current low-illumination enhancement algorithms cannot meet real-time detection requirements, and do not bring substantial performance improvements on detectors.
- We have proposed a Night Vision Detector (NVD) with specifically designed feature pyramid network and context fusion network on basis of RFB-Net. We train and test the proposed detector directly on ExDARK, a lately published real low-light scene dataset. The experiments show that our scheme can amazingly improve the mean average detection precision performance in low-illumination scenes.
- We have further experimented and discussed the affects of data augmentation and parameters initialization. We reach an experimental conclusion that no matter which illumination bias is specified for initial model, the converged model trained on the illumination-balanced data consistently bias towards low-illuminance scenes. Fine-tuning from pre-trained parameters of normal-illumination model can benefit for low-illumination performance improvements.

In summary, our investigations in this paper can offer a good starting point for further studies on low-illumination object detection.

In the following parts, we will briefly survey some related work in Section II. Then we will discuss about the importance of illumination in SectionIII. In Section IV, we analysis the validation of current pre-enhancing strategies in low-illumination detection tasks. After that, in Section V, we describe the experiment settings of our developed night vision detector. In Section VI, we perform experiments and discussions on the affects of illumination-balanced data and model's parameter training. At last, we give our conclusions in final Section VII.

II. RELATED WORK

A. LOW-ILLUMINATION IMAGE ENHANCEMENT

1) TRADITIONAL PIPELINE

The most classic techniques to enhance the contrast of low-light images is the histogram equalization [5]. It is very simple, having low computational complexity. However, gray levels that characterize image details are easily lost due to excessive gray merge.

Another classic technique is the γ-correction. It assumes that the sensitivity value of human eyes to the external light is exponentially related to the input light intensity. Under low-illumination, it is easier for human eyes to recognize changes in brightness; with illumination increases, it becomes difficult for human eyes to distinguish changes in brightness. Through gamma correction, the contrast effect of image illuminance is more obvious. However, during image processing, it is difficult to automatically determine a reasonable gamma value to correct the original image.

Inspired by the dark channel prior based defogging algorithm [6], several low-illumination video enhancement algorithms were proposed [7], [8]. Łoza *et al.* [9] proposed a statistical modeling method based on image wavelet coefficients for images with low-illumination and uneven illumination.

2) RETINEX THEORY

The Retinex theory [10] from human vision system (HVS) believes that the observed brightness of objects is composed of illumination component and reflection component. It is formulated as Equation 1:

$$S = I \circ R \tag{1}$$

where, S represents observed image, I represents its illumination component, R represents its reflection component, and $\circ$ represents element-wise multiplication.

Inspired from retinex theory, single-scale retinex (SSR) [11], multi-scale retinex (MSR) [12] and multiscale retinex with color restoration (MSRCR) [13] were successively proposed. In recent, illumination map estimation like LIME [14] was proposed to construct illumination map through finding maximum intensity for image channels. Robust-Retinex [15] were similarly proposed based on retinex theory [10]. However, the decompose of observed brightness is an ill-posed problem that so far hasn't been solved well.

3) DEEP-LEARNING BASED MODEL

Benefiting from the popularity of deep learning methods, low-level image restoration research such as deblurring, denoising, and super-resolution have made great breakthroughs. However, for low-illumination enhancement, it is very difficult to obtain corresponding ground truth. Most of current low-illumination enhancement methods are therefore trained on synthetic data. LLNet [16] was the first deep auto-encoder-based approach to identify signals from low-light images and adaptively brighten images without over-amplifying the lighter parts in images. Tao *et al.* [17] proposed a two-step strategy to enhance low-light images based on atmospheric scattering lighting model.

Since Retinex theory is more suitable for human vision characteristics, it becomes popular to combine deep learning with retinex theory, such as LightenNet [18]. Shen *et al.* [19] proved that multi-scale retinex is equivalent to a feedforward convolutional neural network with different gaussian convolution kernels. They proposed a MSR-Net for directly learning the end-to-end mapping between dark and bright images. Wei *et al.* [20] proposed a Retinex-Net model that contains DecomNet for decomposition and EnhanceNet for lighting adjustment. Chen *et al.* [21] developed the first real low-illumination RAW dataset, and developed an enhanced network SID for processing low-light images. Jiang *et al.* [22] proposed an unsupervised generative adversarial network called EnlightenGAN to solve the training without low/normal illumination image pairs. In the most recent, a maximum entropy based retinex model [23] was proposed through self-supervised learning. Hong *et al.* [24] proposed to address spectral variability by applying a data-driven learning strategy in inverse problems of hyper-spectral unmixing.

B. OBJECT DETECTION

Following R-CNN [25] that use convolutional neural networks [26] for object detection, the object detection schemes are divided into two categories: (1) two-stage detectors such as R-CNN [25], Faster-RCNN [1], MS-CNN [27], Mask-RCNN [28],etc. and (2) one-stage detectors such as YOLO [29], SSD [2], DSSD [30], RFB-Net [4] and RetinaNet [31], etc.

The two-stage detection algorithm decomposes detection process into two steps in which candidate regions (region proposals) are first generated, then the candidate regions are classified and object locations are refined. The two-stage detection algorithm can achieve low recognition error rate, but cannot meet real-time detection scenarios. The one-stage detection algorithm does not require region proposal step. It can directly generate the category labels and coordinate positions of interested objects after a single detection, having much faster speed than most of two-stage detection algorithms. However, the detection accuracy of one-stage algorithms are mostly inferior to that of two-stage algorithms.

With the development of computer vision, both detection strategies have been greatly improved. The improvements are well-known attributed to public dataset such as PASCAL VOC[1] or COCO.[2] However, all these famous public data contain less than 1% of low-illuminance images. As a result, current detectors cannot fully exert their optimal performance in low-illuminance scenes.

In some application areas, such as in optical remote sensing imagery, specific auxiliary informations are utilized. Wu *et al.* [32] propose an optical remote sensing imagery detector through jointly considering the rotation-invariant channel features constructed in frequency domain and the original spatial channel features. Further, they proposed a fourier-based rotation-invariant feature boosting framework [33] to solve object deformation problem.

As we know, low-illumination scenes are closely related to our life. Currently, the development of corresponding intelligent vision systems is yet to be studied. In this paper, we are concentrated to shed light to providing some feasible solutions for improving the object detection performance under low-illumination.

III. IMPORTANCE OF ILLUMINATION

In order to study the impacts of illumination on feature extractions at different convolution layers, in this section, we propose to experiment and visualize the impacts on different illumination datasets.

Low-illumination data comes from ExDARK[3] [3], with a total of 7363 images, including 12 categories (*'bicycle', 'boat', 'bottle', 'bus', 'car', 'cat', 'chair', 'cup', 'dog', 'motorbike', 'people', 'table'*). Among these images, 4800 images are for training and 2563 images are for testing.

To construct normal-illumination counterpart, we randomly select 600 images for each category that is defined in ExDARK, from corresponding COCO category. As a result, a total of 7200 images are selected, from which 4800 images are used for training and 2400 images for testing. We denote

[1] http://host.robots.ox.ac.uk/pascal/VOC/

[2] http://cocodataset.org/

[3] https://github.com/cs-chan/Exclusively-Dark-Image-Dataset

TABLE 1. Models trained on different illumination datasets have limitations for different illumination scenarios. Train/Test represent training/testing datasets.

Test	Train	AP	$AP^{.50}$	$AP^{.75}$	AP^{s}	AP^{m}	AP^{l}	AR^{1}	AR^{10}	AR^{100}	AR^{s}	AR^{m}	AR^{l}
COCO*	COCO*	13.4	26.1	12.1	1.6	10.1	22.7	18.7	26.8	30.3	7.8	28.9	46.9
	ExDARK	9.4	20.4	7.3	1.6	8.2	15.8	16.4	23.9	25.5	4.4	24.7	45.0
ExDARK	ExDARK	34.0	64.1	32.5	3.6	15.7	40.1	31.9	45.1	47.2	11.0	31.2	52.7
	COCO*	23.9	48.9	20.7	1.0	6.8	29.3	24.2	34.2	36.6	5.3	17.4	42.7

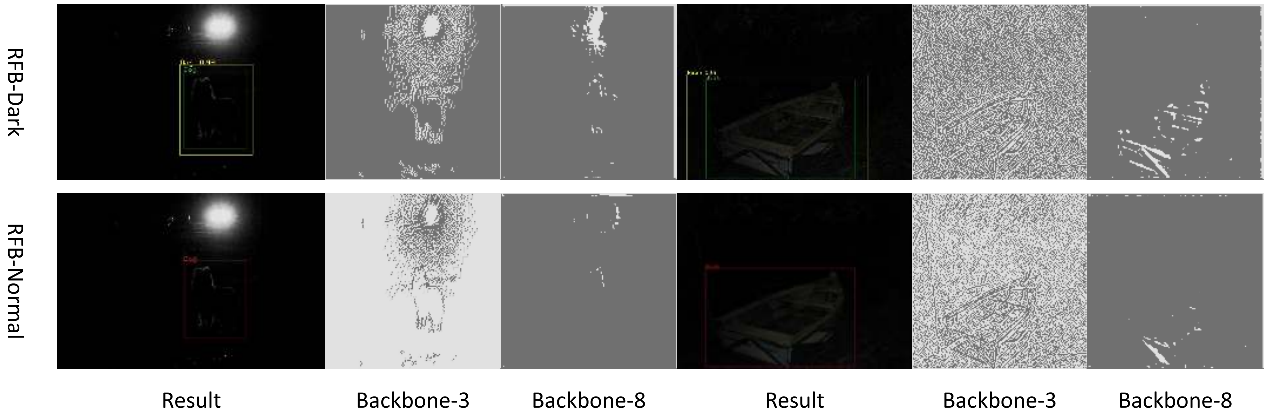


FIGURE 2. Shallow illumination information affects high-level feature representation learning, which will also affect the final detect results. 'Backbone-3' represents the third layer of the backbone network output. (Red: undetected groundtruth, Green: detected groundtruth, Yellow: proposed box).

the new collected dataset as COCO* to distinguish it from its original COCO dataset.

Without loss of generality, we employ a lightweight one-stage objection detection algorithm, representatively the RFB-Net [4] for experiment. One of reasons for selecting one-stage strategy is that we can easily visualize the intra-layer feature mappings of whole network from bottom to top, which can benefit and simply our analysis process. We train RFB-Net on ExDARK and COCO* respectively with the same experiment setting, and obtain two variants: **RFB-Dark** and **RFB-Normal**.

The detection performances are evaluated by using standard COCO evaluation APIs.[4] Specifically, in COCO, 10 IoU thresholds $IoU = [.50 : .05 : .95]$ are use. Object is taken as *small* if its area is less than 32^2, taken as *medium* if its area is between 32^2 and 96^2, and taken as *large* if its area is more than 96^2. The AP (Average Precision) is averaged over multiple Intersection over Union (IoU) values and all categories. The $AP^{.50}$ and $AP^{.75}$ are computed at a single IoU of 0.50 and 0.75 respectively. The AP^s, AP^m, AP^l are AP for small objects, medium objects, and large objects respectively. The AR (Average Recall given a fixed number of detections per image) is also averaged over all categories and IoUs. In COCO, three thresholds [1, 10, 100] on max detections per image are given in default for AR^1, AR^{10} and AR^{100} respectively. Similarly, AR^s, AR^m and AR^l are computed across scales for objects of different size respectively.

In order to observe the impacts of different illumination data on model's robustness, we perform a cross-dataset testing. Specifically, we train and test model on different illumination data. The cross-testing results are shown in Table 1. As expected, performances of detector severely degrades when it is tested on a dataset having different illumination from its training data.

For deeper insight into the influence of illumination on feature learning, we visualize feature maps at different convolution levels. The comparisons between **RFB-Dark** and **RFB-Normal** models are shown in Fig 2. It is obvious that, when dealing with low-illumination image, model **RFB-Dark** and **RFB-Normal** extract very different features at the same convolution layers. Features from **RFB-Dark** is much richer and more semantic complete than ones from **RFB-Normal**. Especially on high-level layers, unexpected less information is extracted by **RFB-Normal**. That's why the low-illumination model **RFB-Dark** can successfully detect objects in low-illumination environment, while normal-illumination model **RFB-Normal** can not. Therefore, illumination as low-level information should be paid more specific attention for model's robustness.

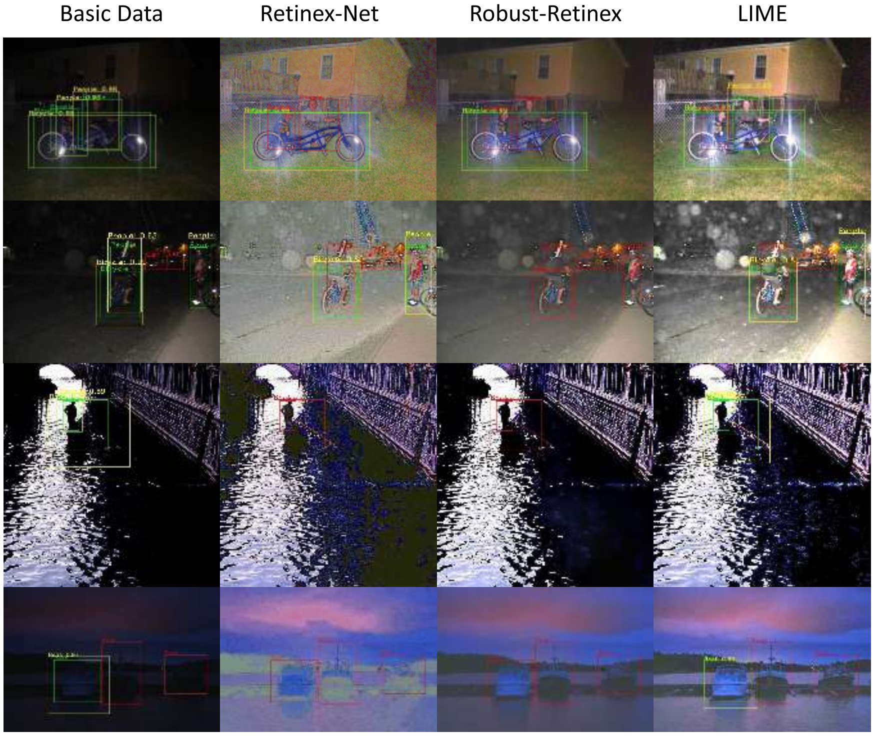
IV. PRE-ENHANCEMENT MAYBE INVALID

Existing object detectors under normal-conditions are powerless in face of extreme adverse environments such as fogy, rainy, and night. A natural idea is to employ image enhancement as preprocess stage before proceeding to high-level vision task. This seems in line with the requirements

[4] https://github.com/cocodataset/cocoapi

TABLE 2. Although the enhanced data has improved visually, it does not mean that the performance of computer vision tasks can be improved.

Data	AP	$AP^{.50}$	$AP^{.75}$	AP^{s}	AP^{m}	AP^{l}	AR^{1}	AR^{10}	AR^{100}	AR^{s}	AR^{m}	AR^{l}
Retinex-Net	31.0	59.8	29.2	1.9	13.7	36.7	30.1	42.0	44.3	7.8	27.7	49.7
LIME	32.7	62.6	31.6	2.5	14.7	38.7	31.2	43.7	45.6	11.3	28.7	51.2
Retinex-Rob	33.8	63.8	32.9	2.4	15.3	39.8	31.7	44.2	46.2	12.2	29.8	51.7
ExDARK	34.0	64.1	32.5	3.6	15.7	40.1	31.9	45.1	47.2	11.0	31.2	52.7



Basic Data Retinex-Net Robust-Retinex LIME

FIGURE 3. Comparison of object detection results before and after using image enhancement. (Red: undetected groundtruth, Green: detected groundtruth, Yellow: predicted bounding-box).

of human vision. Therefore, we use traditional method LIME[5] [14], Retinex-Robust[6] [15] and deep-learning based method Retinex-Net[7] [20] to respectively enhance original low-illumination data, and obtain three resulted enhanced datasets. We name them as the same as their respectively constructor, *LIME*, *Retinex-Rob*, *Retinex-Net* for simplicity.

We train detection model respectively on these newly constructed illumination-enhanced data. The detection performances are shown in Table 2. We are frustrated to find that, compared with performance directly training and testing on original ExDARK data, all pre-enhanced data do not improve detection performance, while decrease the performance instead. Some visualization results are show in Fig 3. More examples are shown in Appendix VII.

From these visual samples, we can easily find that image enhancement algorithms can improve image's brightness visually. However, noises are inevitably introduced, especially in Retinex-Net scheme. The Retinex-Robust and LIME show relatively better than Retinex-Net. In spired of that, their final detection performances are still worse than detection performance on original ExDARK. Furthermore, the enhancement process of Retinex-Robust takes too much time (e.g. processing a 1080P high-resolution image spent more than one minutes), which is impossible to serve as a preprocessing step for real-time detection task. Therefore, according to our experiment results, we reach a frustrated conclusion that current image enhancement algorithms seem helpless for low-level illumination detection task, except improving image's visual qualities. Pre-enhancement maybe invalid.

V. MAKING OF NIGHT VISION

In this section, we describe and experiment our developed model which is specific for low-illumination detection task. It is on basis of the state-of-the-art RFB-Net model. We name our proposed detector as Night Vision Detector(NVD). As demonstrated in previous section III, valuable informations are easily lost in deeper layers. Especially under

[5] https://github.com/Sy-Zhang/LIME

[6] https://github.com/martinli0822/Low-light-image-enhancement

[7] https://github.com/weichen582/RetinexNet

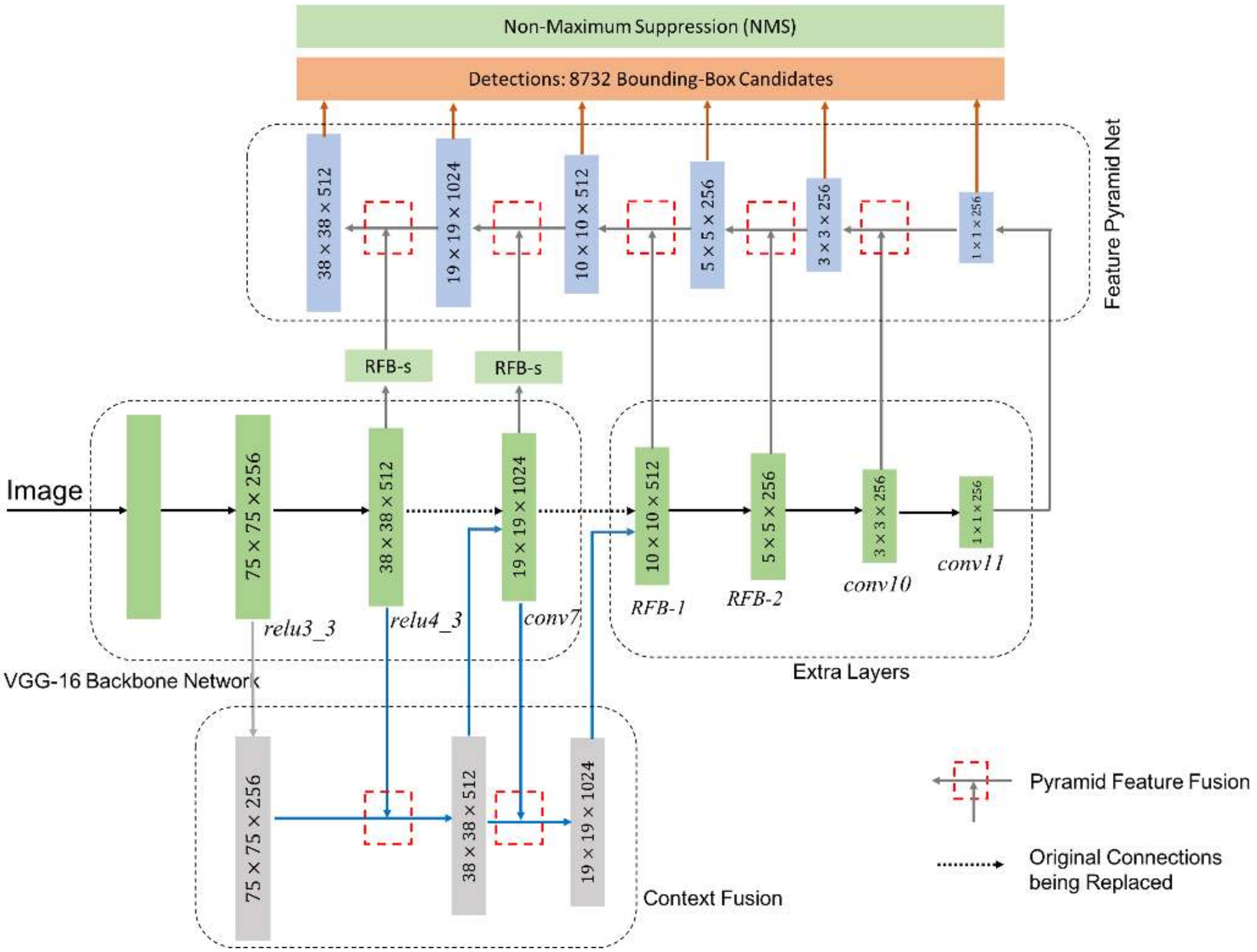


FIGURE 4. The architecture of our developed Night Vision Detector (NVS). The specific structure is based on RFB-Net. Feature pyramid net is introduced for improvements on small object detection. Contextual information fusion is introduced for maximumly retain the limited and weak object information in low-illumination image.

low-illumination, parts of the object are easily merged into dark background during performing convolutions. Therefore, in order to improve the detection performance under low-illumination, we introduce feature pyramid fusion network into detection layers. Context informations are fused into the detection backbone to compensate for loss of low-level textural and contour features. The architecture details are shown in Fig 4.

A. NIGHT VISION DETECTOR

1) FEATURE PYRAMID FUSION NETWORKS

Feature Pyramid Network (FPN) [34] was first introduced as an extension of Mask R-CNN [28] for better representing objects at multiple scales. FPN improves the standard feature extraction pyramid by adding a second pyramid that takes the high level features from the first pyramid and passes them down to lower layers. It is a general strategy that combines top-down fusion with skip layer and pyramid predictions at multi-scales.

The motivation of FPN is very suitable to adapt itself to low-illumination object detection task, since it can generate feature pyramids with strong semantic information without obvious computation cost on all scales. Different from original FPN [34], we give specific modification on the structure of pyramid feature fusion process. The structure differences are shown in Fig 5.

For the pyramid feature fusion modules illustrated by dashed red rectangles in the Feature Pyramid Net of Figure 4, from top to down, lower spatial resolution feature map (e.g. the *"conv11"* layer) is interpolated to the same spatial size of higher spatial resolution feature map (e.g. the *"conv10"* layer), and concatenated with the higher spatial resolution feature map (e.g. herein the *"conv10"* layer), then fed into a 1 × 1 convolution layer. The output feature map is with the same spatial and channel sizes as the considered higher spatial resolution feature map (e.g. herein the *"conv10"* layer).

The motivation of our pyramid feature fusion process is that we aim to maximumly utilize pre-trained channel informations. Through concatenation, semantic information in the

TABLE 3. The ablation performance comparisons of NVD for low-illumination object detection on ExDARK.

FPN	CF	AP	$AP^{.50}$	$AP^{.75}$	AP^{s}	AP^{m}	AP^{l}	AR^{1}	AR^{10}	AR^{100}	AR^{s}	AR^{m}	AR^{l}
✗	✗	34.0	64.1	32.5	3.6	15.7	40.1	31.9	45.1	47.2	11.0	31.2	52.7
✓	✗	34.2	64.1	33.2	**5.8**	16.8	40.1	32.2	44.9	46.9	11.3	31.4	52.4
✗	✓	34.7	64.5	34.3	3.5	**17.3**	40.6	32.3	45.8	48.1	11.6	**32.7**	53.4
✓	✓	**35.3**	**65.5**	**34.8**	4.4	**17.3**	**41.2**	**32.7**	**46.2**	**48.2**	**13.2**	31.7	**53.7**

TABLE 4. Detector trained on ExDARK+COCO* except RFB-Normal and RFB-Dark. 'Init R' means that no interpolation parameters are used. When 'R' is a number, it represents the value of α.

Method	Test	AP	$AP^{.50}$	$AP^{.75}$	AP^{s}	AP^{m}	AP^{l}	AR^{1}	AR^{10}	AR^{100}	AR^{s}	AR^{m}	AR^{l}
RFB-Normal	COCO*	13.4	26.1	12.1	1.6	10.1	22.7	18.7	26.8	30.3	7.8	28.9	46.9
RFB-Dark	ExDARK	34.0	64.1	32.5	3.6	15.7	40.1	31.9	45.1	47.2	11.0	31.2	52.7
Init R	COCO*	14.7	26.6	14.4	2.9	12.0	24.9	21.0	31.6	34.9	11.1	34.1	55.1
	ExDARK	37.1	66.3	37.6	3.0	18.3	43.2	33.8	47.6	50.2	13.6	33.1	55.7
Init .5	COCO*	14.8	27.0	14.7	2.9	12.1	24.7	21.3	31.9	35.0	11.0	34.8	54.2
	ExDARK	36.3	65.3	36.3	3.3	17.4	42.4	33.1	46.9	49	13.4	31.1	53.7
Init .7	COCO*	15.0	27.8	14.7	3.0	12.8	25.0	21.4	32.0	34.9	10.2	34.8	55.0
	ExDARK	34.3	62.9	33.7	3.0	14.5	40.4	32.0	45.1	47.7	12.5	29.2	53.3
Init 1	COCO*	15.9	27.6	16.1	3.4	12.9	26.3	22.5	33.8	37.0	12.4	36.7	56.8
	ExDARK	38.3	67.0	39.1	4.4	18.6	44.6	34.2	48.7	51.2	20.3	34.1	56.8

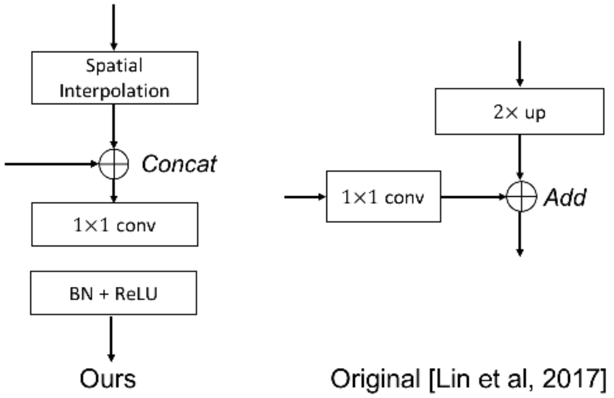


FIGURE 5. The structures of our proposed pyramid feature fusion module and original module [34]. 'BN+ReLU' means ReLU after BatchNormalization.

channels of lower spatial resolution feature maps and textural/contour information in the channels of higher spatial resolution maps are redundantly preserved. 1 × 1 convolution operations learn optimal linear combination relationships among these concatenated complementary informations. Our process can be easily incorporated into any multi-scale network structure, not limited to herein RFB-Net, as auxiliary modules without any backbone structure changing.

In contrast, the 'add' operation in original FPN has limitation that channel size of higher spatial feature map must be compromised to the same as the one of lower spatial feature map. We has tried the use of original FPN too, during our many experimental attempts. Frustratedly, the training process cannot converge, if we use original FPN structure.

In our later ablation experiments, we validate that the proposed feature pyramid fusion network improve the performances especially on small objects detection by 2.2%, when compared basic RFB-Net.

2) CONTEXT FUSION (CF) NET

We observe that there are a lot of dark areas in the image captured under low-illumination. The information of objects in the image is often covered by dark areas or merged with dark background. Affected by uneven light source, the sensor often can only capture limited parts of objects' contours on image. In most cases, the captured contours are weak. Conventional hierarchical convolutions inevitably lose valuable informations what there are little, such as informative texture/contour details. Therefore, we introduce a context fusion net in bottom-up way into backbone network for feature compensation during the lower-level to higher-level convolution process. The fusion process has similar structure with pyramid feature fusion process.

Specifically, we select the '*relu*3_3' feature map in the backbone network (VGG-16) [35] and spatially interpolated it down-scale to the same spatial size of its successive '*relu*4_3'. We denote the resulted feature map as '*relu*3_3_*d*'. The size of '*relu*3_3' is 75 × 75 × 256, and therefore the size of '*relu*4_3_*d*' is 38 × 38 × 256. Feature map '*relu*4_3_*d*' is concatenated with feature map '*relu*4_3' whose size is 38 × 38×512, and then fed into a convolution block which contains a 1 × 1 convolution, batch-normalization and ReLU activation. The output feature map, denoted as '*relu*4_3_*new*', has the same size with feature map '*relu*4_3'. We on one hand replace the connection (dashed in Figure 4) between

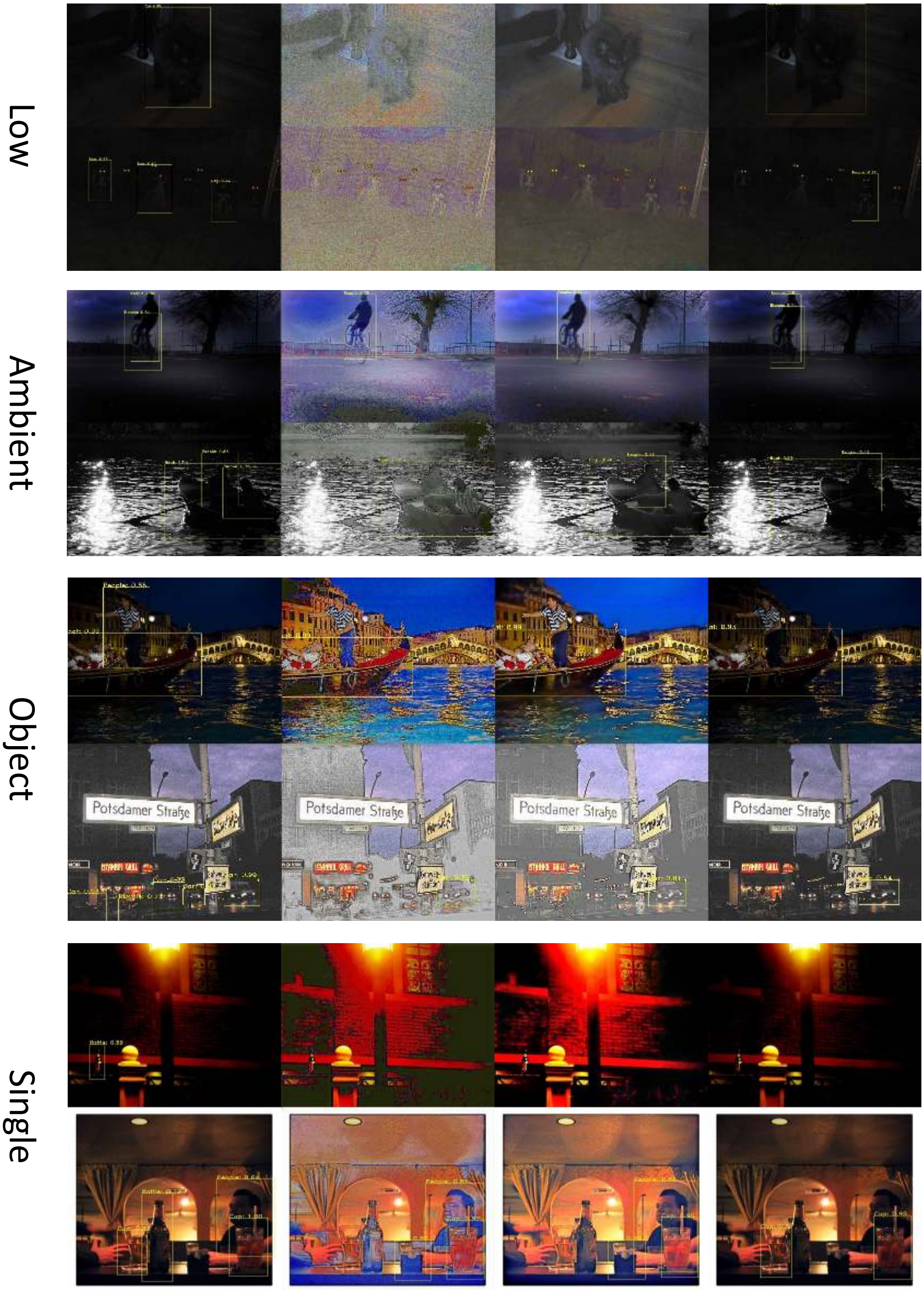


FIGURE 6. Comparisons on detection results for images of 10 listed low-illumination types. From left to right, original image, image enhanced by Retinex-Net, Robust-Retinex and image enhanced by LIME.

*'relu*4_3*'* and *'conv_7'* with the new connection between the *'relu*4_3_*new'* and *'conv_7'*. On the other hand, we feed the *'relu*4_3_*new'* to the next context fusion process with feature map *'conv_7'*. At last, the again obtained feature map, denoted as *'conv_7_new'* is fed back to backbone network, replacing the original connection between *'conv_7'* and *'RBF-1'*. These fusion processes are illustrated in Figure 4.

B. EXPERIMENT

Since currently there is no special solution for low-illuminance detectors, the experiments in this paper have to be compared with our basic model RFB-Net. We directly train model on ExDARK dataset and study the contributions of each proposed component to the model. The experimental results are shown in the Table 3.

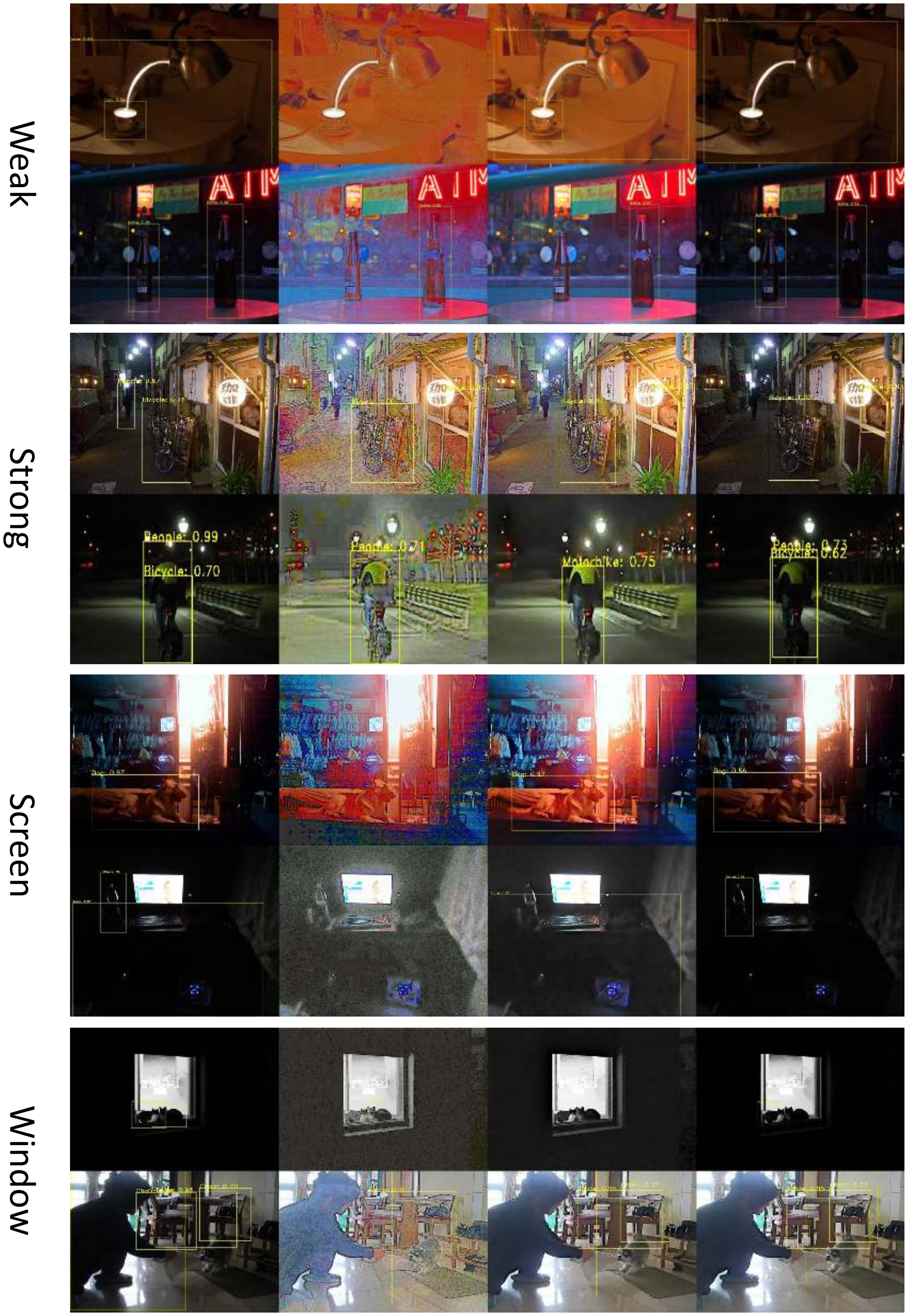


FIGURE 7. Comparisons on detection results for images of 10 listed low-illumination types. From left to right, original image, image enhanced by Retinex-Net, Robust-Retinex and image enhanced by LIME.

Compared with basic model RFB-Net, our proposed new FPN component can improve the detection performance AP^s of small objects greatly by 2.2% without decrease on all other detection performances. The contribution of component **CF** achieves performance improvements on almost all average precisions. Finally, after incorporating both components, our

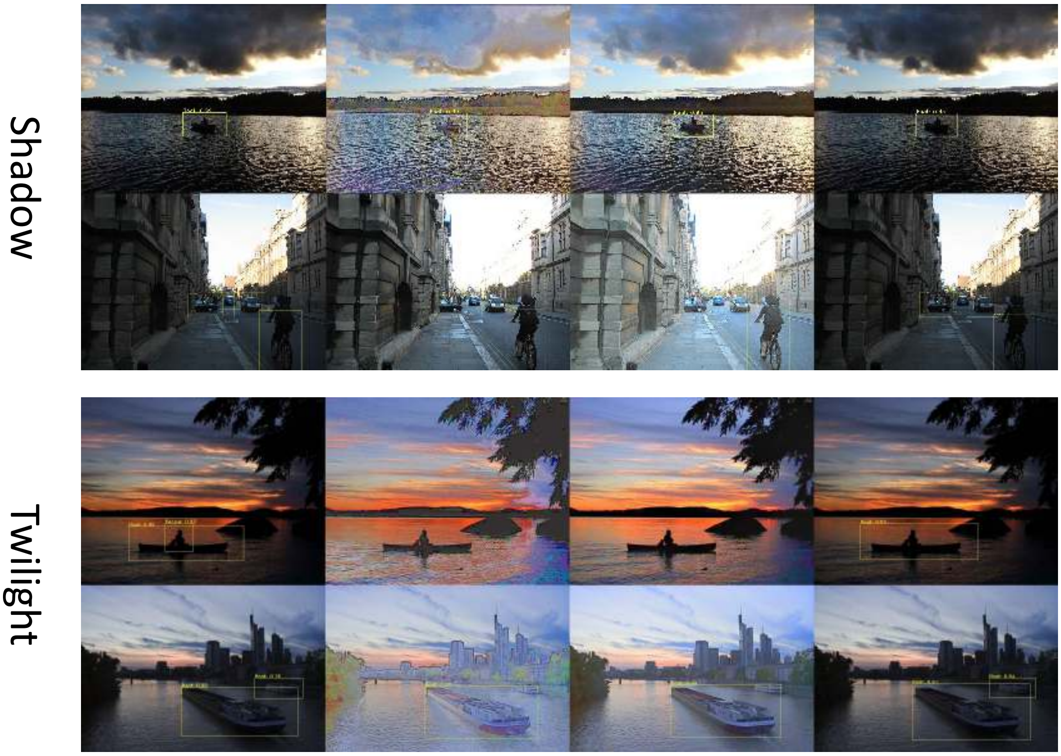


FIGURE 8. Comparisons on detection results for images of 10 listed low-illumination types. From left to right, original image, image enhanced by Retinex-Net, Robust-Retinex and image enhanced by LIME.

proposed NVD model achieves improvements of 0.5%∼2.8% according to the COCO standard evaluation criterions.

It should be point out that both our proposed components are general that can be easily applied to any multi-scale detection network. Our proposed scheme has effectively improved the performance of object detection under low-illumination.

VI. THE EFFECTS OF ILLUMINATION-BALANCED DATA AND PARAMETERS INITIALIZATION

As we know, the performance of data-driven deep model is generally benefited from rich data and well-initialized parameters. Therefore, in this section, we discuss the effects of data augmentation and parameters initialization. Since the discussed factors are not related specific model, we experiment the discussions on basic RBF-Net for simplicity.

We augment training data by collecting training data from both the two compared datasets COCO* and ExDARK, a total of 9,600 training images.

We first train models on COCO*, ExDARK and COCO*+ExDARK respectively with random initialized parameters. The testing performances are shown in the first three rows in Table 4. It is worth noting that in this experiment, we haven't made any model settings to favor any kind of illumination data. Augmented training data benefits the performance boosting. However, compared with the performance improvement on normal illumination data COCO*, for low-illumination scenes, the detection performance is improved much greatly. We explain that model tend to learn hard samples during training and the knowledge learned from normal-illumination data benefits more for low-illumination learning case.

Afterwards, we try to specify model bias through weighted combination of pre-trained model's parameters with different types of lighting data. We interpolate the model parameters by parameters pre-trained with normal-illumination data and parameters pre-trained on low-illumination data. The formal expression follows Equation 2.

$$\theta_{init} = (1-\alpha)\theta_{dark} + \alpha\theta_{coco} \qquad \alpha \in [0,1] \tag{2}$$

where, θ_{init} is the interpolated parameters. θ_{dark} represents the **RFB-Dark**'s parameters, θ_{coco} represents the **RFB-Normal**'s parameters. α is to control the initial bias of model to some specific illumination type data. $\alpha \in [0, 1]$. The bigger the α is, the more bias the initial mode is towards **RFB-Normal**.

We discretely take three α cases, $\alpha = \{0.5, 0.7, 1\}$ for experiments. We train model on COCO*+ExDARK by using the interpolated parameters as model's initial parameters and test the detector on COCO* and ExDARK test data respectively. The performances are shown in the fourth to sixth rows in Table 4. The results demonstrate that performance gains on ExDARK have always been presented, when compared with **RFB-Dark** (the second row in Table 4). It indicates that no matter which illumination bias is specified for initial model, the converged model trained on the illumination-balanced data consistently bias towards low-illuminance scenes. When $\alpha = 1$, in other words, fine-tuning from pre-trained parameters of normal-illumination model achieves the best detection performance. We explain that this result is reasonable, since it is in line with our experience that we generally rely on prior knowledge (e.g. shape, texture of object, etc) learned at daytime to reason about unknown objects in dark. Moreover, during experiments, we additionally find the training process with interpolated pre-trained parameters is more stable, having better speed-up convergence.

VII. CONCLUSION

In this work, we have investigated several important issues on object detection under low-illumination environment and without loss of generality proposed a Night Vision Detector (NVD) based on RFB-Net for low-illumination environments. We find that illumination information has great impacts on feature learning. Different illumination data should be modeled separately, since they would interfere with each other during training. We further suggest that current image quality driven enhancement methods could be employed to improve visual quality, while helpless for high-level real-time object detection tasks. Therefore, our proposed NVD framework introduced feature pyramid fusion net and context fusion net. The two components take comprehensive considerations of informations that affect low-illumination detection. The experiments demonstrate that the proposed NVD achieves low-illumination detection performance by 0.5%∼2.8% higher than basic RFB-Net on all standard COCO evaluation criterions. Our work can provide baseline strategies and shed light to future studies on low-illumination detection.

APPENDIX

More results are shown in Fig. 6, Fig. 7 and Fig. 8. Ten different low-illumination scenes (**Low**, **Ambient**, **Object**, **Single**, **Weak**, **Strong**, **Screen**, **Window**, **Shadow**, **Twilight**, please refer to [3] for their specific definition), and corresponding detection results after low-illumination enhancement are listed. The results suggest that visually improvements after enhancement show little benefits for high-level vision tasks.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 91–99.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[3] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Understand.*, vol. 178, pp. 30–42, Jan. 2019.

[4] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 385–400.

[5] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. T. H. Romeny, and J. B. Zimmerman, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, Sep. 1987.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[7] H. Malm, M. Oskarsson, E. Warrant, P. Clarberg, J. Hasselgren, and C. Lejdfors, "Adaptive enhancement and noise reduction in very low light-level video," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[8] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Seven Springs, PA, USA, Jul. 2011, pp. 1–6.

[9] A. Łoza, D. R. Bull, P. R. Hill, and A. M. Achim, "Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients," *Digit. Signal Process.*, vol. 23, no. 6, pp. 1856–1866, Dec. 2013.

[10] E. Land, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 58, p. 1428A, 1967.

[11] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 451–462, Mar. 1997.

[12] Z.-U. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," Tech. Rep., 1996.

[13] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[14] X. Guo, "Lime: A method for low-light image enhancement," in *Proc. 24th ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, 2016, pp. 87–91.

[15] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[16] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[17] L. Tao, C. Zhu, J. Song, T. Lu, H. Jia, and X. Xie, "Low-light image enhancement using CNN and bright channel prior," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3215–3219.

[18] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: A convolutional neural network for weakly illuminated image enhancement," *Pattern Recognit. Lett.*, vol. 104, pp. 15–22, Mar. 2018.

[19] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-Net: Low-light image enhancement using deep convolutional network," 2017, *arXiv:1711.02488*. [Online]. Available: http://arxiv.org/abs/1711.02488

[20] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, U.K., 2018, pp. 1–12.

[21] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3291–3300.

[22] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep light enhancement without paired supervision," 2019, *arXiv:1906.06972*. [Online]. Available: https://arxiv.org/abs/1906.06972

[23] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," 2020, *arXiv:2002.11300*. [Online]. Available: http://arxiv.org/abs/2002.11300

[24] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[27] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 354–370.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Honolulu, HI, USA, Oct. 2017, pp. 2961–2969.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[30] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and C. B. Alexander, "DSSD: Deconvolutional single shot detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[32] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, ''ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[33] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, ''Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[35] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–14.

YUXUAN XIAO was born in 1995. He received the Bachelor of Engineering degree from Jiangxi Normal University, China, in 2017, where he is currently pursuing the master's degree. His research interests include computer vision and image processing.

AIWEN JIANG (Member, IEEE) was born in Jingdezhen, Jiangxi, China, in 1984. He received the B.S. degree in electronic engineering from the Nanjing University of Posts and Telecommunications, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, China, in 2010. Since 2010, he has been an Associate Professor with the School of Computer and Information Engineering, Jiangxi Normal University, China. From 2015 to 2016, he was a Visiting Researcher with NICTA, Australia National University. His research interests include computer vision and machine learning.

JIHUA YE received the M.S. degree in electronic engineering from the South China University of Technology, in 2006. He is currently a Full Professor with the School of Computer and Information Engineering, Jiangxi Normal University, China. His research interests include computer vision and the Internet of Things.

MING-WEN WANG received the B.S. and M.S. degrees from the School of Computer and Information Engineering, Jiangxi Normal University, China, in 1988, and the Ph.D. degree in computer science from Shanghai Jiaotong University, China, in 2001. From 2002 to 2003, he was a Visiting Professor with the University of Montreal, Canada. From April 2009 to October 2009, he was a Senior Visiting Professor with Yale University. He is currently the Director and a Full Professor with the School of Computer and Information Engineering, Jiangxi Normal University. His research interests include machine learning and information retrieval. He is a Senior Member of CCF and a member of council of CAAI.