2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Dynamic Low-Light Image Enhancement for Object Detection via End-to-End Training

Haifeng Guo
Department of Computer Science and Technology
Nanjing University
Jiangsu, China
Email: guo-haifeng@outlook.com

Tong Lu†
Department of Computer Science and Technology
Nanjing University
Jiangsu, China
Email: lutong@nju.edu.cn

Yirui Wu
College of Computer and Information
Hohai University
Jiangsu, China
Email: wuyirui@hhu.edu.cn

***Abstract*—Object detection based on convolutional neural networks is a hot research topic in computer vision. The illumination component in the image has a great impact on object detection, and it will cause a sharp decline in detection performance under low-light conditions. Using low-light image enhancement technique as a pre-processing mechanism can improve image quality and obtain better detection results. However, due to the complexity of low-light environments, the existing enhancement methods may have negative effects on some samples. Therefore, it is difficult to improve the overall detection performance in low-light conditions. In this paper, our goal is to use image enhancement to improve object detection performance rather than perceptual quality for humans. We propose a novel framework that combines low-light enhancement and object detection for end-to-end training. The framework can dynamically select different enhancement subnetworks for each sample to improve the performance of the detector. Our proposed method consists of two stage: the enhancement stage and the detection stage. The enhancement stage dynamically enhances the low-light images under the supervision of several enhancement methods and output corresponding weights. During the detection stage, the weights offers information on object classification to generate high-quality region proposals and in turn result in accurate detection. Our experiments present promising results, which show that the proposed method can significantly improve the detection performance in low-light environment.**

***Index Terms*—Low-Light Image Enhancement; Object Detection**

I. Introduction

Object detection is one of the fundamental tasks in computer vision. Various applications based on object detection methods are hotspots in real world, such as autonomous driving [1] and pedestrian detection [2]. Breakthrough progress has been made in academic research of object detection especially after the emergence of deep learning. However, images captured in real world often have many quality problems, *e.g.*, low light, low resolution and color distortion, which significantly reduce the performance of various detection algorithms. For these problems, it is common practice to use some enhancement methods to recover a high-quality image from the original image first, and then detect objects in the recovered image.

Images captured in low-light conditions inevitably encounter the problem of quality degradation, such as low contrast, massive noise and blurred edges and texture. These problems will lead to object classification errors and inaccurate localization in low-light images. If using the paradigm of "enhancement first, detection later", we will encounter many difficulties. We summarize these difficulties of object detection tasks in low-light environments from three perspectives as follows:

1) Most of the existing low-light image enhancement methods are designed to improve the perceptual quality for human eyes. These methods which perform well on visual qualiy may not significantly improve the performance of the object detection task.
2) The diversity of low-light environments (see Fig. 1) makes it difficult for many enhancement methods to cover all situations. In other words, different enhancement methods may have different or even negative effects on different samples.
3) Quality problems in low-light images are interrelated. For example, the denoising method may simultaneously blur the edges and texture of the object to be detected, resulting in classification and localization errors.

To solve the first problem, We propose an *end-to-end*

Fig. 1. Different lighting types in low-light conditions, which demonstrates the diversity and complexity of low-light environments. Examples are collected from websites and search engines, where (a), (b) are extremely low light environments; (c), (d) are weak illumination environments with blurred or foggy objects; (e), (f) contain Visible light source and objects in dark place; (g), (h) are bright but objects are in shadows.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412802

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

framework that combines low-light image enhancement and object detection tasks. For the problem of environmental diversity in low-light conditions, we use a dynamic enhancement network to perform filtering on low-light images in a *sample-specific* way, which is capable of adapting to different low-light conditions. For the interrelated quality problems, we use multiple parallel subnetworks to simulate different enhancement methods, generating enhanced images. A sample-specific weight is used to selectively enhance the different features in low-light images. The weighted fusion of multiple enhancement methods can provide different views for the model and enhance its representation ability, thereby solving the interrelated image quality problems to a certain extent.

In this paper, we propose a novel method to combine the low-light image enhancement and object detection tasks in a end-to-end framework. In this way, our proposed method is capable of dynamically enhancing the low-light images to improve the object detection performance. Our quantitative evaluation shows that the proposed method significantly improves the detection performance without increasing the computational cost too much. We demonstrate the effectiveness of each part of the method and analyze the impact of hyperparameter selection through the ablation study.

The main contributions of this paper can be summarized as follows:

- We propose a novel framework for the end-to-end training of low-light image enhancement and object detection, which can significantly improve the detection performance in low illumination environment.
- We introduce dynamic filter networks and adaptive exposure module in enhancement stage, which can acquire stronger feature representation based on the supervision of existing enhancement methods.

II. Related Work

We will briefly introduce two topics related to our work in this section. The first part contains existing low-light image enhancement methods. Most of them are designed to improve perceptual quality but not object detection performance. The second part shows several works similar to ours, where the authors also seek to enhance images to improve the performance of downstream visual tasks such as object detection, image classification, etc.

A. Low-Light Image Enhancement

Low-light images taken in low-light conditions always have poor quality. Researchers have been trying to recover high-quality images from low-light images through image enhancement methods. Traditional methods can be divided into two categories: histogram-based methods and Retinex-based methods. Histogram-based methods, such as adaptive histogram equalization (AHE), map the histogram of the entire image to a new distribution pixel by pixel. The idea of the Retinex theory [3] is to separate the illumination from the reflectance and the Retinex-based methods use the illumination map to enhance the image. However, the performance of traditional methods is poor due to the large amount of noise in low-light images.

Recently, methods based on deep learning have shined on many low-level vision tasks, such as denoising [4]. The performance of deep learning in the field of low-light image enhancement is also exciting. Lore *et al.* [5] proposed a deep autoencoder-based approach to perform contrast enhancement and denoising simultaneously for low-light images. Wei *et al.* [6] combined the deep neural networks and Retinex theory and proposed an end-to-end framework for decomposition and illumination enhancement. EEMEFN in [7] employed a multi-exposure fusion module and an edge enhancement module for extremely low-light image enhancement.

Unlike the methods whose purpose is to improve image quality for better human visual perception, in this paper, we propose an end-to-end framework to enhance low-light images dynamically in order to improve the performance of object detection tasks. We mainly use the performance of object detection to quantitatively evaluate the effectiveness of our method, rather than considering it from the perspective of perceptual quality.

B. Image Enhancement for Downstream Vision Tasks

In recent years, some studies have pointed out the impact of image quality degradation on downstream visual tasks. Karahan *et al.* [8] analyzed the impact of several image quality degradations on the performance of CNN-based deep face recognition methods. They showed that blur, noise and occlusion will cause a significant performance degradation, and the deep CNN model is robust to distortions such as color distortion and color balance changes. Dodge *et al.* [9] reached a similar conclusion after considering five types of quality distortion. In real life scenarios, the images we captured cannot always be assumed of high quality. Therefore, using image enhancement techniques is a straightforward idea to improve the performance of high-level vision tasks.

Similar to our work, there are some researchers aim to use image enhancement techniques to help improve downstream vision tasks, *e.g.*, image classification [10], action recognition [11] and object detection [12]. Although the methods designed in this way have limited improvement in the perceptual quality of the image, they can significantly improve the performance of downstream vision tasks. Costa *et al.* [10] considered several types of noise and noise level and proved that these denoising methods can improve classification accuracy for noisy data.Kvyetnyy *et al.* [12] proposed an image denoising method based on bilateral filtering and wavelet thresholding, and a boosting method for object detection. Bai *et al.* [13] focused on small object detection and proposed an end-to-end generative adversarial network to improve the detection performance for small-sized objects.

In this paper, we focus our attention on subfields of low-light images and adopt a end-to-end framework. By combining low-light image enhancement task with object detection task, our proposed method can selectively learn good representations that can help the detector improve performance.

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

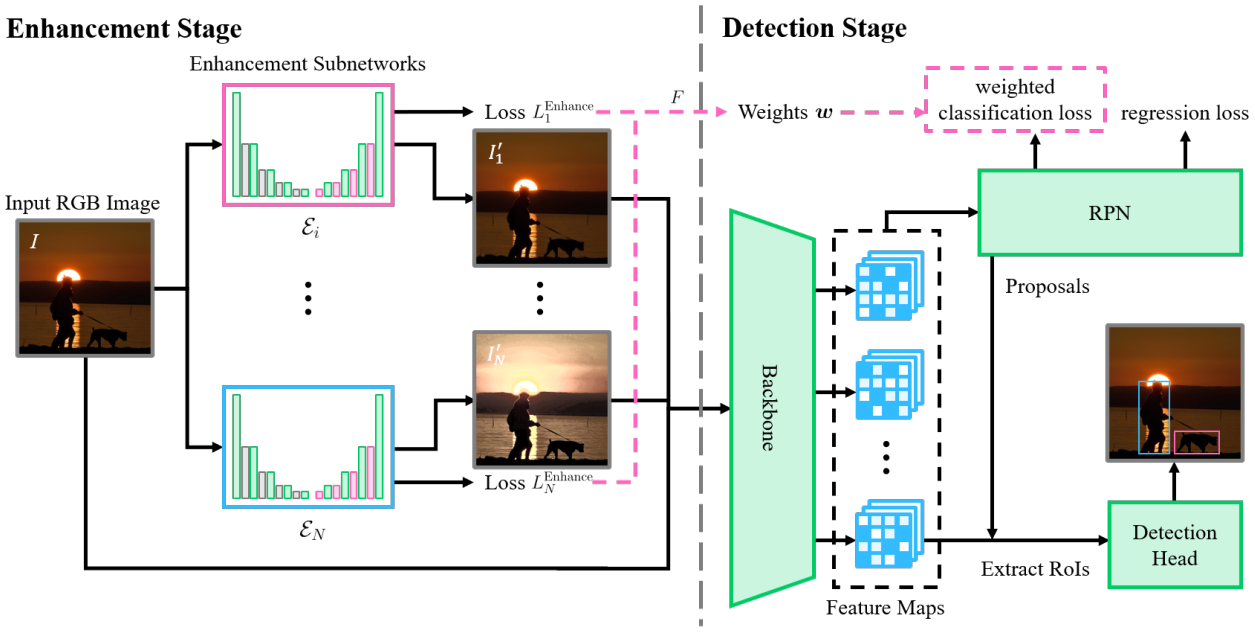

Fig. 2. **Our proposed end-to-end framework**. We show the end-to-end framework of multiple enhancement subnetworks and the detector. I refers to the input of the enhancement stage. $I'_1, \ldots, I'_N$ denotes the output of each enhancement subnetwork, which is supervised by different enhancement methods. The architecture of the enhancement subnetworks is shown in Fig. 3. Note that the parameters of each enhancement subnetwork are not shared.

III. Proposed Method

In this section, we describe our proposed low-light image enhancement method for object detection. It is important to emphasize that our method does not pursue perceptual image enhancement but aims to dynamically enhance the low-light image features that is conducive to improving the performence of a high-level vision task, *i.e.*, object detection. We will first introduce the general framework of our method, then we will detail the two components of the framework. Finally, we will demonstrate how to train the overall framework end-to-end.

A. Overall Framework

As is illustrated in Fig. 2, the proposed method mainly consists of two stages: the enhancement stage and the detection stage. We unify these two stages in one framework for end-to-end joint optimization. In the enhancement stage, inspired by [14], we use a dynamic filter networks to generate sample-specific convolution kernels. These convolution kernels are used to dynamically enhance low-light images. We employ common enhancement methods to constrain the behavior of each enhancement subnetwork so that the model can adaptively choose the most effective enhancement methods.

In the detection stage, we use a variant of Faster R-CNN [15] to perform object detection based on the enhanced images generated by the enhancement stage. We assign weights to the classification losses of RPN to improve the classification performance at this stage. The weight is calculated from the losses in the enhancement stage, which represents the importance of each enhancement subnetwork for each sample.

The network in the enhancement stage is composed of N subnetworks of the same architecture in parallel. We denotes the input RGB image as I, and the enhancement network will output N images, *i.e.*,

$$I'_i = \mathcal{E}_i\left(I\right) \ , \quad 1 \leqslant i \leqslant N \ . \tag{1}$$

where I'_i is generated by the i-th subnetwork. Each subnetwork $\mathcal{E}_i$ is supervised by a specific enhancement method. We train the subnetwork by calculating the loss L_i^{Enhance} between the illumination images enhanced by the specific enhancement method and the enhancement subnetwork $\mathcal{E}_i$.

At the same time, we can get the weight corresponding to each subnetwork in some way, such as softmax after negation or the method we use later. Let the weight calculation method be F, then the weight vector $\boldsymbol{w}$ can be calculated as:

$$\boldsymbol{w} = F\left(L_1^{\text{Enhance}}, L_2^{\text{Enhance}}, \ldots, L_N^{\text{Enhance}}\right) \ . \tag{2}$$

The i-th component $\boldsymbol{w}_i$ of the weight $\boldsymbol{w}$ indicates the importance of the i-th enhancement method. A larger weight indicates that the corresponding enhancement method is more conducive to improving the performance of the detector, and vice versa.

The detector receives the output of the enhancement stage as input, and finally generates bounding boxes and the corresponding object categories. The detection stage is similar to a standard two-stage object detection algorithm. The difference is that we use the weights calculated in the enhancement stage to produce higher quality proposals. In this way, we can make the gradient in the detection stage backpropagate to the enhancement stage and train end-to-end.

B. Stage I: Dynamic Enhancement

In essence, the enhancement stage actually simulates various enhancement methods, and on this basis, enhance those

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

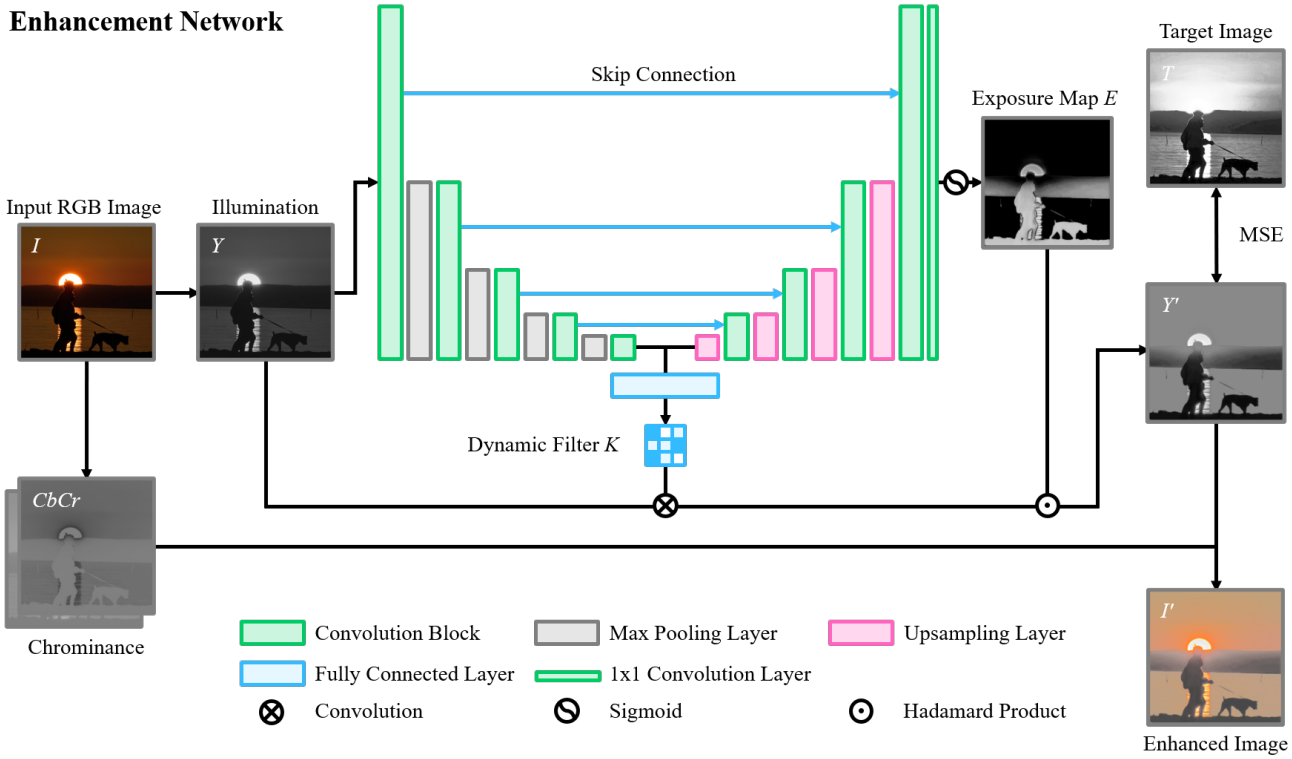

Fig. 3. **The architecture of the enhancement networks**. The heights of the rectangles corresponding to different layers indicate the scale of the corresponding feature maps. For clarity, we have omitted the reshape operations of the input and output of U-Net-like architecture. Convolution block consists of two consecutive convolutional layers activated by ReLU. During the downsampling process, the number of channels of the feature map is continuously increased, from 1 to 512, and the upsampling process is the opposite. The output image is finally obtained through a 1×1 convolution layer, *i.e.*, the exposure map.

features that can improve detection performance. We propose a dynamic enhancement network composed of several subnetworks which are independent of each other. In Fig. 3, we illustrate the detailed architecture of the enhancement network. The enhancement network includes two components: the dynamic filter generator and the adaptive exposure module (AEM). The dynamic filter is designed to simulate a specific enhancement method, and AEM is to further activate areas in the image that are critical to improve detection performance. We proposed a convolutional architecture based on U-Net [16] to combines the two parts in one subnetwork. In this way, the two modules can share the feature maps to reduce the computation cost.

Given a RGB image $I \in \mathbb{R}^{h \times w \times 3}$, the enhancement network output the dynamic filter $K \in \mathbb{R}^{m \times m \times 1}$ and the exposure map $E \in \mathbb{R}^{h \times w \times 1}$. Hereby, h and w indicate the original height and width of the input image, respectively, and m denotes the size of the dynamic convolution kernel. Concretely, we first transform the input image into the luminance-chrominance color space, obtaining the luminance component Y and the chrominance components Cb and Cr, and the luminance component $Y \in \mathbb{R}^{h \times w \times 1}$ is resized into a fixed shape. The resized luminance component of image I is fed into the downsampling part of U-Net and output feature maps of a fixed size.

Based on these shared feature maps, the filter generator use a fully connected layer to dynamically generate the filter K. Through the upsampling part of the enhancement network, AEM normalizes the output of U-Net with sigmoid function and resize it into the original shape using bilinear interpolation. The output exposure map E can be considered as the pixel-wise exposure intensity. The significance of AEM is that it provides more non-linearities for the enhancement network to simulate those non-linear enhancement methods. Additionally, AEM can alleviate the problem of edge blur caused by the convolution operation in the dynamic filtering. In this way, the enhancement network can even retain the potential to surpass the performance of the baseline methods. Finally, we use the enhanced luminance component Y, the dynamic fitler K and the exposure map E to obtain the enhanced output, *i.e.*,

$$Y' = (Y * K) \circ E \quad , \tag{3}$$

in which $*$ is the convolution operation and $\circ$ denotes the Hadamard product. Note that Y' must not be clipped (*e.g.*, 0.0 to 1.0 or 0 to 255) because it will make the initial gradient too small, which cause the enhancement network to difficult to converge. After that, we combine Y' and the two chrominance components and transform it back into RGB color space. To train the enhancement network, we use the mean squared error (MSE) between the filtered image Y' and the target image T, where T is obtained from Y through a specific image enhancement method, *e.g.*, bilateral filter and histogram equalization. We use N parallel enhancement

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

networks to get N enhanced images, *i.e.*, $I'_1, \ldots, I'_N$. The loss of the enhancement stage is the sum of the losses of all the subnetworks, *i.e.*,

$$L^{\text{Enhance}} = \gamma \cdot \sum_{i=1}^{N} L_i^{\text{Enhance}} = \gamma \cdot \sum_{i=1}^{N} \text{MSE}\left(Y'_i, T_i\right) \quad , \quad (4)$$

where γ is a hyperparameter we have to adjust. At the end of the stage, we assemble the enhanced images and the original image I into a minibatch and feed them into the detector in the second stage.

C. Stage II: Object Detection

The detection stage is based on a variant of the original Faster R-CNN [15] framework, which extracts feature pyramids using FPN [17] and generates RoIs through RoIAlign operation. The minibatch formed by the enhanced images in the first stage will be fed into the backbone to extract features. We train RPN on the $N+1$ groups of different feature maps and use the weighted classification loss for backpropagation to improve the classification performence. The weights are calculated based on the losses of the enhancement stage as follows:

$$\boldsymbol{w}_i = \left(1 - \frac{L_i^{\text{Enhance}}}{\sum_{k=1}^{N} L_k^{\text{Enhance}}}\right) \cdot \frac{N}{N-1} \quad , \quad (5)$$

where L_i^{Enhance} is the loss of the i-th enhancement subnetwork. The more the output of the enhancement network deviates from the basic enhancement method, the smaller the weight is set. Note that, the parameters of RPN are shared by these groups. In addition, there is a problem from which group of feature maps should we extract RoIs. In our method, we simply use the feature maps corresponding to the original image. We try the different operations, but we finally found that there are only marginal performance differences between them and the original one provides the most stable results.

D. End-to-end Joint Optimization

For the end-to-end training of the N enhancement networks and the detector, we add the enhancement loss L^{Enhance} to the detection losses and obtain a total loss as:

$$\begin{aligned} L = L^{\text{Enhance}} &+ \frac{1}{N+1} \cdot \sum_{i=0}^{N} \boldsymbol{w}_i L_i^{\text{rpn\_cls}} \\ &+ \frac{1}{N+1} \cdot \sum_{i=0}^{N} L_i^{\text{rpn\_reg}} + L^{\text{cls}} + L^{\text{reg}} \end{aligned} \quad (6)$$

Note that, $\boldsymbol{w}_0$ is the weight corresponding to the original RGB image and is invariably equal to 1. The regression losses of RPN are not weighted because a weighted regression loss will cause inaccurate object localization. Through the joint optimization, the enhancement network can learn a sample-specific transformation for improving detection performance.

IV. Experiments

To quantitatively evaluate our proposed method, we conduct all our experiments on a challenging low-light image dataset named Exclusively Dark [18].

A. Dataset

The Exclusively Dark (ExDark) dataset is the only open-access collection consisting entirely of low-light images with object level annotations. The dataset contains 7363 images with 12 object categories (3000 images for training, 250 images per class; 1800 image for validation, 150 images per class; 2563 images for testing). These images are captured in 10 different low-light conditions from the extremely low illumination to twilight. Note that, this dataset does not provide paired high-low exposured images, so it is difficult to apply supervised enhancement methods.

The metrics used here are the same as the MS COCO [19] to fairly evaluate the performance of the detector. *i.e.*, AP is the mean value of mean average precision (mAP) over 10 IoU thresholds from 0.5 to 0.95. AP_{50} and AP_{75} are the mAP over the IoU thresholds of 0.5 and 0.75, respectively. AR_{100} is the average recall given 100 detections per image. AP_{S}, AP_{M} and AP_{L} are the AP for small, medium and large objects, respectively.

B. Implementation Details

All our experiments were conducted on a server with two Intel Xeon E5-2620 v4 (@2.1GHz) CPUs and four NVIDIA GTX 1080Ti graphics cards. Our experimental codes are mainly based on PyTorch framework,

We train all our models for 12 epochs using SGD optimizer with an initial learning rate of 0.01. Weight decay is set to 0.0001 and the momentum is 0.9. Due to the linear warm up mechanism, the learning rate increases from $1/3 \times 0.01$ to 0.01 in the first 500 iterations. The learning rate is decreased to 0.001 after 8 epochs and 0.0001 after 11 epochs. We choose ResNet-50 as the backbone of Faster R-CNN and the 5-level feature pyramid extracted by FPN. We use an image scale of 800 and 512 RoIs per image as set in [17]. We apply data augmentation by horizontal flip with 0.5 probability for both baselines and our method.

C. Ablation Study

TABLE I
Comparison of Three Ways to Fuse Enhancement Methods.

	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
Selection Network[1]	30.6	61.6	27.3	**4.4**	18.9	35.2
Raw Input[2]	31.6	61.6	28.5	3.7	18.6	36.4
Ours	**32.1**	**62.1**	**29.9**	3.6	**20.0**	**36.9**

[1] Train a network to select a specific enhancement method;
[2] Use the target images T and original image I as inputs of detector.

1) Fuse Various Enhancement Methods: We compare three different ways to use the existing enhancement methods for improving the performance of detector. We simply choose histogram equalization and Image sharpening filtering as the enhancement methods.

The first method is to train a network to select a specific enhancement method or use the original image. However, some technical issues prevent it from performing well. First,

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

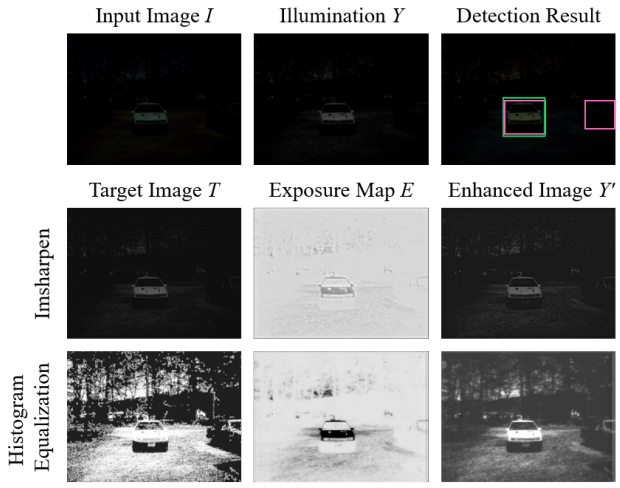

Fig. 4. **Quanlitative results of the enhancement networks**. The first row mainly shows the detection result on the input image I, in which the red box represents the ground truths and the green one is the result.

TABLE II
COMPARISON OF TWO USAGE OF WEIGHTS.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Feature Fusion[1]	30.8	**62.4**	26.8	**8.0**	19.1	35.1
Weighted Loss[2]	**32.1**	62.1	**29.9**	3.6	**20.0**	**36.9**

[1] Feature Fusion: weighted sum of feature maps;
[2] Weighted Loss: weighted RPN classification loss.

the detection performance of an object is easy to evaluate but the one of an image which contains various objects is not. For example, one method has poor localization performance, and the other has good localization but misses objects. We can hardly judge which method performs better. Second, it causes a serious imbalance of the ground-truth labels for the training of selection network. For example, in our experiments, the two enhancement methods only perform better that the original image on few samples.

The second one is to directly use the images enhanced by the existing methods as the input of the detector. In other words, we do not use enhancement networks to simulate these methods. The classifation loss is simply averaged, so the enhancement network will not get feedback on which enhancement method is better.

Compared to the two methods, ours can learn better representations thanks to the well-designed enhancement subnetworks, which are more cirital the improve detection result. The end-to-end joint training helps our method dynamically set the weights of different enhancement methods for different samples, so that our method is more reasonale and outperforms the two methods above (see Table I).

2) Feature Fusion or Weighted Loss: We use two methods to explore the application of weights derived from enhancement network. One is feature fusion. We use weighted sum of the extracted feature maps, and then send them to the RPN to generate proposals. Table II shows the results using this method. The overall performance of feature fusion is worse than the method using weighted loss, but the detection performance of small objects is improved. In addition, it can also reduce the computational cost required to generate region proposals, although the cost of computation for RPN is not large in absolute term. The other is to perform forward propagation on these feature maps respectively, and calculate the weighted average of the RPN classification loss. In this paper, we use the second approach because it can bring about a greater overall performance improvement.

TABLE III
QUANTITATIVE RESULT OF THE IMPACT OF AEM.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
with AEM	**32.1**	**62.1**	**29.9**	3.6	**20.0**	**36.9**
without AEM	31.5	61.4	29.0	**3.7**	19.7	36.1

TABLE IV
COMPARISON OF DIFFERENT FILTER SIZES.

Filter Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
3×3	31.5	61.9	29.1	5.3	18.9	36.3
5×5	31.7	61.7	29.3	4.0	19.1	36.5
$7 \times 7^*$	**32.1**	**62.1**	29.9	3.6	**20.0**	**36.9**
9×9	31.8	61.5	**30.5**	4.1	18.9	36.4
11×11	31.7	61.9	28.8	**5.6**	18.7	36.3

3) Impact of Adaptive Exposure Module: We investigate the impact of adaptive exposure module by removing the upsampling stage of the enhancement network, which means that it only generates filters and does not outputs exposure maps. Quantitative result is shown in Table III. Results with AEM are better than the one without it.

The filter generated by the dynamic filter generator is the same for every position in the image, which is disadvantageous for some scenes (*e.g.*, shadow scene) with sharp luminance contrast. Adding the AEM can alleviate this problem for two reasons. First, this module provides more nonlinearity for the enhancement process to better simulate the nonlinear enhancement methods, while the convolution is a linear operation. In the experiment, we find that after adding this module, the loss of the enhancement stage has decreased significantly. Second, this module can avoid texture and edge blurring problems that occur caused by convolution operations.

D. Hyperparameter Selection

We investigate two critical hyperparameters that have an effect on the experimental results, *i.e.*, the size of the filter and the weight γ of the loss of enhancement stage. As claimed in [24] and [25], the choice of filter size depends on the specific application. Intuitively, an excessively large filter size will make the enhanced image too smooth, which reduce the detection performance. Too small filter size will make it difficult to fit the basic enhancement method. We modify the last fully-connected layer of the filter generator so that it can output filters of different sizes Too large or too small filter

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

TABLE V
QUANTITATIVE EVALUATION OF THE PROPOSED METHOD.

	Enhanced Channel(s)*	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [20]	-	27.6	52.7	25.9	4.5	16.0	31.8
Faster R-CNN w FPN	-	30.4	61.0	27.3	4.3	**18.8**	35.2
Bilateral Filter (BF)	Y	27.8	57.6	23.2	2.2	17.3	32.0
Guided Filter (GF) [21]	Y	25.8	53.9	21.7	1.9	14.6	30.2
Histogram Equalization (HE)	Y	28.4	57.4	25.6	2.6	17.6	32.7
Image Sharpening (IS)	Y	29.0	59.2	25.7	2.9	17.2	34.0
Loh *et al.* [22]	Y	29.0	58.4	25.4	5.2	17.4	33.2
EnlightenGAN (EGAN) [23]	Y	29.2	59.7	25.5	4.5	18.1	33.6
Loh *et al.* [22]	RGB	27.5	55.8	23.3	4.3	16.6	31.6
EnlightenGAN (EGAN) [23]	RGB	29.4	58.8	26.1	6.8	18.6	33.8
Proposed Method (based on Loh *et al.*'s & EGAN)	Y	31.6	61.7	28.8	**7.4**	18.5	36.3
Proposed Method (based on HE & IS)	Y	**32.1**	**62.1**	**29.9**	5.4	**18.8**	**36.4**

* Enhanced channel means which channel we apply enhancement methods. Y denotes the illumination component in YCbCr color space and RGB indicates all the channels in RGB color space.

size can affect the detection results to varying degrees, mainly due to differences in the detection results of medium-scale objects. We found that the filter of 7×7 achieves the best overall performance.

For the choice of γ, our main goal is to make the loss of enhancement stage on the same order of magnitude as other losses. In our experiments, we found that γ in a certain range (from 0.05 to 0.2) has little effect on the experimental results. However, if γ is too large (*e.g.*, $\gamma = 1$), it will cause the model to be untrainable due to gradient explosion. Too small γ will make it difficult for the enhancement network to converge. As a result, we choose $\gamma = 0.1$ in all our experiments.

E. Qualitative and Quantitative Evaluation

Fig. 4 shows the qualitative results of the enhancement stage. We demonstrate the intermediate results of the enhancement network corresponding to two different enhancement methods (Image sharpening Filtering and Histogram Equalization), *i.e.*, exposure map E, enhanced luminance image Y' and the results of the specific enhancement method (target image T). Fig. 5 shows the curated examples of the detection results. All the quantitative results are shown in Table V.

First, we show the results of two classic one-stage and two-stage object detection algorithms, *i.e.*, RetinaNet [20] and Faster R-CNN [15]. Without any enhancement to low-light images, both of them have low detection performance.

Second, we independently compare the effects of several image enhancement methods on the detection results. Specifically, we feed all the images enhanced by specific enhancement methods into the detector to evaluate the impact of each method on detection performance. As is illustrated, the classic filtering methods reduce the performance of the detector to varying degrees, which shows that they are actually not good for object detection tasks. Among them, the two methods for denoising (Bilateral Filter and Guided Filter [21]) greatly reduce AP. It may be due to the two filters blurring objects in the images in low-light conditions, although they can suppress noise well. We also compare our method with several trainable methods [22], [23] proposed recently. It should be noted that we can only compare the unsupervised methods, because the dataset does not provide well-exposed images as supervision. Intuitively, these methods have no explicit denoising process, resulting in considerable noise in the detection results, which is detrimental to localization.

Finally, we apply our method to untrainable and trainable methods mentioned before. For the untrainable methods (HE and IS), despite their poor overall performance, they still performed well on some samples. The advantage of untrainable methods is that they usually do not have a large computational cost and do not require complicated adjustment for hyperparameters. Experimental results show that our method has good compatibility with these methods and can further improve the detection performance on the basis of them.

V. CONCLUSION

This paper presents an end-to-end solution for the object detection in low-light conditions. The proposed method consists of two stages: the enhancement stage and the detection stage. The enhancement stage uses a convolutional neural network to generate filters and exposure maps and obtain the dynamically enhanced images. The detection stage is based on a variant of Faster R-CNN, which is trained jointly with the enhancement network. We demonstrate the effectiveness of our proposed method on a new low-light image dataset named ExDark. Experimental results show that the proposed method significantly improves detection performance without making too many modifications to the detector. We hope our work can inspire more research on object detection in low-light environments in the future.

VI. ACKNOWLEDGEMENTS

This work is supported by Scientific Foundation of State Grid Corporation of China (Research on Ice-wind Disaster Feature Recognition and Prediction by Few-shot Machine Learning in Transmission Lines).

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.

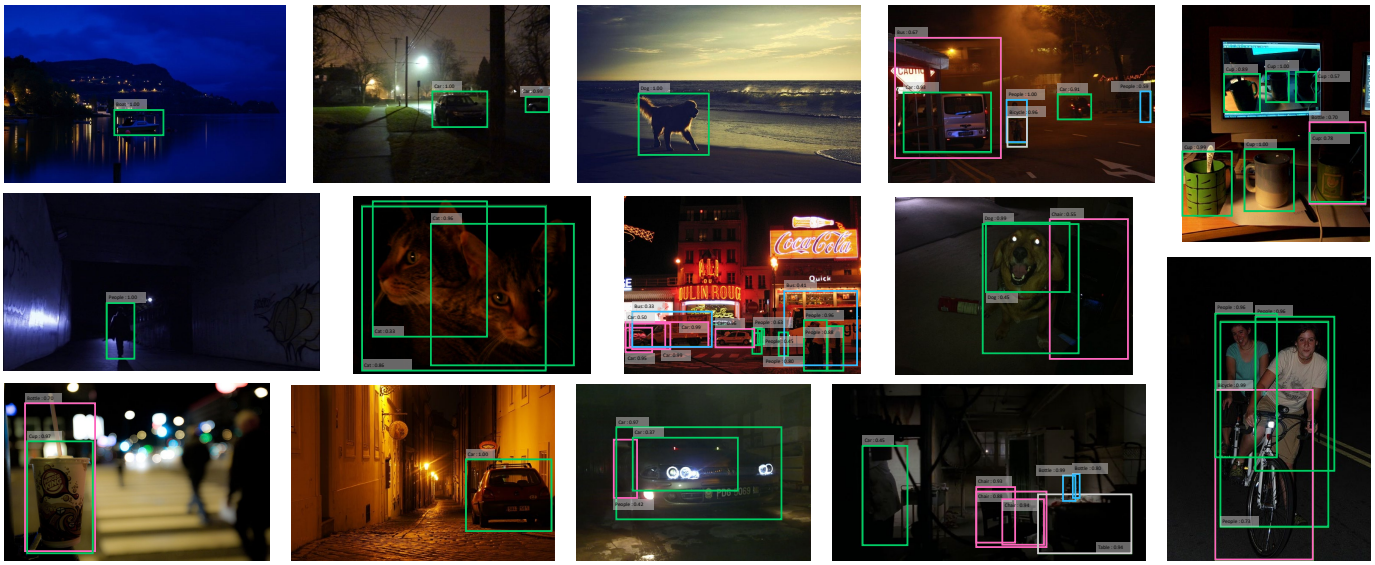

Fig. 5. **Example detections on the ExDark dataset** (test set). The bounding box of the same color in the same image represents the detection result of the same type of object. A score threshold of 0.3 is used for demonstration. Zoom for the best viewing experience.

REFERENCES

[1] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 129–137.

[2] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-Aware Fast R-CNN for Pedestrian Detection," *IEEE transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[3] E. H. Land, "The Retinex Theory of Color Vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[4] C. Godard, K. Matzen, and M. Uyttendaele, "Deep Burst Denoising," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 538–554.

[5] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A Deep Autoencoder Approach to Natural Low-Light Image Enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," in *Proceedings of British Machine Vision Conference*, 2018.

[7] M. Zhu, P. Pan, W. Chen, and Y. Yang, "EEMEFN: Low-Light Image Enhancement via Edge-Enhanced Multi-Exposure Fusion Network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[8] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How Image Degradations Affect Deep CNN-based Face Recognition?" in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–5.

[9] S. Dodge and L. Karam, "Understanding How Image Quality Affects Deep Neural Networks," in *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[10] G. B. P. da Costa, W. A. Contato, T. S. Nazare, J. E. Neto, and M. Ponti, "An Empirical Study on the Effects of Different Types of Noise in Image Classification Tasks," *arXiv preprint arXiv:1609.02781*, 2016.

[11] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-Task Sparse Learning with Beta Process Prior for Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 423–429.

[12] R. Kvyetnyy, R. Maslii, V. Harmash, I. Bogach, A. Kotyra, Z. Gradz, A. Zhanpeisova, and N. Askarova, "Object Detection in Images with Low Light Condition," in *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments*, vol. 10445. International Society for Optics and Photonics, 2017, p. 104450W.

[13] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 206–221.

[14] V. Sharma, A. Diba, D. Neven, M. S. Brown, L. Van Gool, and R. Stiefelhagen, "Classification-Driven Dynamic Image Enhancement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4033–4041.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," in *Proceeding of 29th Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[18] Y. P. Loh and C. S. Chan, "Getting to Know Low-Light Images with the Exclusively Dark Dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[21] K. He, J. Sun, and X. Tang, "Guided Image Filtering," in *European Conference on Computer Vision*. Springer, 2010, pp. 1–14.

[22] Y. P. Loh, X. Liang, and C. S. Chan, "Low-Light Image Enhancement using Gaussian Process for Features Retrieval," *Signal Processing: Image Communication*, vol. 74, pp. 175–190, 2019.

[23] 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," *arXiv preprint arXiv:1906.06972*, 2019.

[24] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial Transformer Networks," in *Proceeding of 29th Conference on Neural Information Processing Systems*, 2015, pp. 2017–2025.

[25] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic Filter Networks," in *Proceeding of 30th Conference on Neural Information Processing Systems*, 2016, pp. 667–675.

Authorized licensed use limited to: Hohai University Library. Downloaded on May 16,2022 at 11:37:18 UTC from IEEE Xplore. Restrictions apply.