Exploring Image Enhancement for Salient Object Detection in Low Light Images

XIN XU and SHIQIN WANG, Wuhan University of Science and Technology, China
ZHENG WANG, National Institute of Informatics, Japan
XIAOLONG ZHANG, Wuhan University of Science and Technology, China
RUIMIN HU, Wuhan University, China

Low light images captured in a non-uniform illumination environment usually are degraded with the scene depth and the corresponding environment lights. This degradation results in severe object information loss in the degraded image modality, which makes the salient object detection more challenging due to low contrast property and artificial light influence. However, existing salient object detection models are developed based on the assumption that the images are captured under a sufficient brightness environment, which is impractical in real-world scenarios. In this work, we propose an image enhancement approach to facilitate the salient object detection in low light images. The proposed model directly embeds the physical lighting model into the deep neural network to describe the degradation of low light images, in which the environment light is treated as a point-wise variate and changes with local content. Moreover, a Non-Local-Block Layer is utilized to capture the difference of local content of an object against its local neighborhood favoring regions. To quantitative evaluation, we construct a low light Images dataset with pixel-level human-labeled ground-truth annotations and report promising results on four public datasets and our benchmark dataset.

CCS Concepts: • **Information systems** → *Information retrieval*;

Additional Key Words and Phrases: Low light images, salient object detection, images enhancement, physical lighting model, non-local-block layer

ACM Reference format:
Xin Xu, Shiqin Wang, Zheng Wang, Xiaolong Zhang, and Ruimin Hu. 2021. Exploring Image Enhancement for Salient Object Detection in Low Light Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 1s, Article 8 (March 2021), 19 pages.
https://doi.org/10.1145/3414839

This work was supported by the Natural Science Foundation of China (U1803262, 61602349, and 61440016).
Authors' addresses: X. Xu (corresponding authors), S. Wang, and X. Zhang, Wuhan University of Science and Technology, Wuhan, China; emails: xuxin@wust.edu.cn, wust_wangshiqin@163.com, xiaolong.zhang@wust.edu.cn; Z. Wang, National Institute of Informatics, Tokyo, Japan; email: wangz@nii.ac.jp; R. Hu, Wuhan University, Wuhan, China; email: hrm1964@163.com.

1551-6857/2021/03-ART8 $15.00
https://doi.org/10.1145/3414839

1 INTRODUCTION

Salient Object Detection (SOD) aims at localizing and segmenting the most conspicuous objects or regions in an image. As a pre-processing step in computer vision, SOD is of interest to urban surveillance and facilitates a wide range of visual applications, e.g., object re-targeting [2, 43, 53, 55, 58], semantic segmentation [66, 69], image synthesis [7, 48, 65], visual tracking [13, 44, 45], image retrieval [49, 52, 54, 56, 57].

Current SOD methods primarily utilize global and local features to locate salient objects on existing SOD datasets. Images in these datasets are usually captured in the environment with sufficient brightness. However, the effectiveness of current SOD methods in low light images is still limited. Images captured in low illumination conditions usually exhibit low contrast and low illumination properties. These properties cause severe object information loss in dark regions, where the salient object is hard to detect. As shown in the second and third columns of Figure 1, the results of R3Net lose detail information of salient objects and tend to contain non-saliency backgrounds in the degraded low light images. The reason mainly attributes to the fact that the environment light in low light image modality primarily consists of artificial light. Because of the influence of the artificial light, the environment light is ever-changing at different locals of the image. Thus, the environment light, working as noise, will degrade the image capturing process.

Different from existing SOD methods that conduct SOD directly on original degraded images, we eliminate the effect of low illumination by explicitly modeling the physical lighting of the environment for image enhancement. The detail information of the salient object can be retained to improve the SOD performance. To achieve this goal, it is natural to enhance low light image first. However, existing low light image enhancement mainly focuses on improving subjective visual quality, rather than facilitating subsequent high-level SOD task. To alleviate such a problem, we first embed the physical lighting model into the deep neural network to describe the degradation of low light images, in which the environment light is treated as a point-wise variate and changes with local content. Then a Non-Local-Block Layer is utilized to extract non-local features of salient objects. Moreover, a low light image dataset is built to evaluate the performance of SOD.

In summary, the main contribution of our work is threefold:

- We build a low light image dataset for the SOD community. Based on this dataset, we verify that low illumination can reduce the performance of SOD.
- We explore image enhancement for low illumination SOD. The effect of low illumination can be eliminated by explicitly modeling the physical lighting of the environment for image enhancement, and the detail information of the salient object can be retained to improve the SOD performance.
- To account for the non-uniform environment light, the physical lighting of low light images is analyzed to build the degradation model, where the environment light is treated as a point-wise variate and changes with the local light source. Moreover, a Non-Local-Block Layer is utilized to capture the difference of local content of an object against its local neighborhood favoring regions.

2 RELATED WORKS

SOD has achieved remarkable progress in recent years and is a hot topic in both academic and industrial communities. The main difficulty in SOD is how to separate salient object from its surroundings to resist the interference caused by variations in viewpoint, background, resolution, illumination, and so on. Inspired by current low light image enhancement approaches, this article focuses on the low illumination issue in the SOD task.

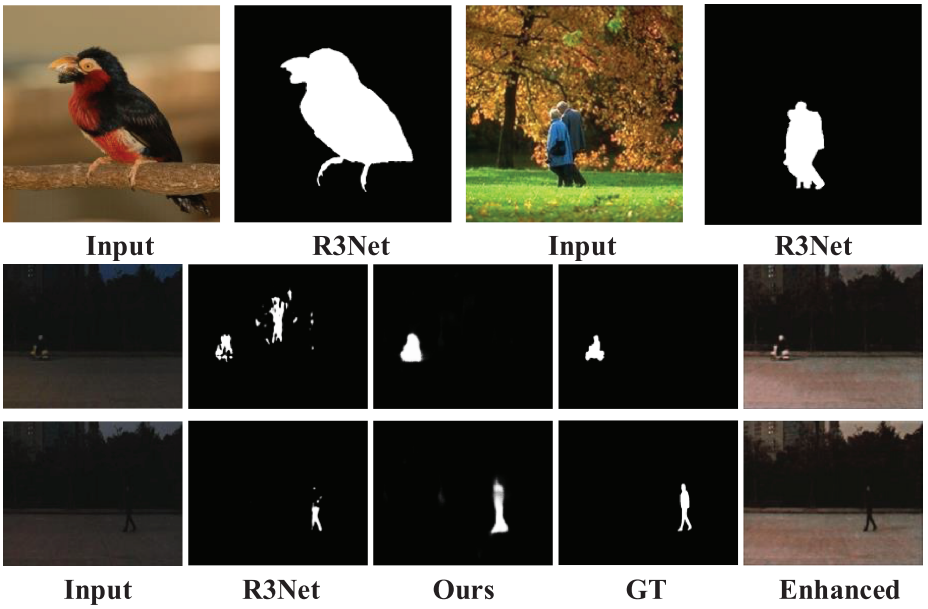

Fig. 1. Examples and their SOD results. The first row shows two examples in the general SOD task. The inputs are, respectively, the images from the DUT-OMRON [61] and PASCAL-S [30] datasets. Their corresponding results are generated by R3Net [10]. The SOD performances are perfect for these two images with sufficient brightness. The next two rows show the results of the low light images. From left to right are, respectively, the input images, the results by R3Net, the results by our approach, the ground truths, and the enhanced results of our approach. R3Net performs not so well, while our approach achieves considerable results.

2.1 Salient Object Detection

Traditional SOD models locate conspicuous image regions by computing the difference with their surroundings and primarily rely on hand-crafted features. These models do not require the training process, and extract saliency feature from color [9], contrast [28], contour [69], objectness [5], focusness [22], and backgroundness [37]. In recent years, deep learning–based SOD models extract high-level features in a data-driven way and have demonstrated their superior performance. These high-level features can be broadly divided into three categories: global feature, local feature, and global & local feature.

Li et al. [29] proposed the multi-task (MT) neural network, which uses convolution to extract global features. Zeng et al. [63] formulated zero-shot learning to promote saliency detectors (LPS) to embedded DNN as a global feature extractor into an image-specific classifier. While global features can only roughly determine the location of salient object with incomplete information, Li and Yu [27] proposed the multi-scale deep features (MSDF) neural network, which decomposes input images into a set of non-overlapping blocks and then puts them into the three-scale neural networks to learn local features. Deng et al. [10] proposed a recurrent residual refinement network (R3Net) to learn local residual between the non-salient regions from intermediate prediction and saliency details from the ground truth. Similarly, Qin et al. [36] proposed a Boundary-aware salient object detection network (BASNet) to conduct a coarse prediction with the residual refinement to hierarchically extract local features.

However, multiple levels of local convolutional blur the object boundaries, and high-level features from the output of the last layer are spatially coarse to perform the SOD tasks. Recent work

attempted to combine the information from both global and local features. Yang et al. [61] utilized graph-based manifold ranking (MR) to evaluate the similarity of local image pixels with global foreground cues or background cues. Luo et al. [32] utilized the non-local deep features (NLDF) to connect each local feature and fused with the global features to output the saliency map. Hou et al. [19] proposed deeply supervised SOD with short connections (DSSC) from deep-side outputs with global location information to shallow ones with local fine details.

To the best of our knowledge, quite a few works attempt to address the issue of low illumination for SOD. Due to the effect of low illumination, images captured in a non-uniform illumination environment usually are degraded with the scene depth and the corresponding environment lights. This degradation results in severe object information loss, which makes the SOD more challenging. In Reference [59], we previously proposed to extract non-local features for SOD in low light images. However, conducting low illumination SOD directly on original degraded images may be non-optimal. In this work, we focus on the problem of low light image enhancement for SOD.

2.2 Low Light Image Enhancement

In the past decades, various enhancement techniques have been developed to improve the quality of low light images. Histogram equalization (HE) is a widely utilized approach due to its simplicity. Global HE approach [41] balances the histogram of the entire image and is suitable for lightening the overall low light images. However, global HE enables the gray levels with high frequency to dominate the other low-frequency gray levels and may degrade sharpness at the boundary. To tackle this problem, local HE approach [26] conducts the calculation inside a sliding window over the low light image. However, local HE may cause over-enhancement in bright regions.

Another choice is based on Retinex [18] and multi-scale Retinex (MSR) [24], which assume that an image can be composed of scene reflection and illumination. Fu et al. [14] proposed a weighted variational model to adjust the regularization terms by the fusion of multiple derivations of illumination map. However, this method ignores the structure of illumination and may lose realisticness in rich textures regions. Guo et al. [15] proposed a low light image enhancement (LIME) approach to estimate the illumination of each pixel in RGB channels, which is refined by structure prior. Retinex-based approaches are based on the Lambertian scene assumption and require illumination environment should be piece-wise smooth. However, low light images captured in low illumination environments usually contain regions with rapidly changing illumination due to artificial light interference, which may cause Halo effects in these regions.

To tackle the non-uniform illumination issue, an alternative way is to analyze the physical lighting of low light images. Dong et al. [11] assumed that the inverted low light images are similar to images captured in hazy conditions and applied dark channel prior to analysis of the image degradation. However, the image degradation model for the haze environment is inadequate to reflect the globally physical lighting and causes potential information loss in the dark regions of low light images. Ying et al. [62] and Ren et al. [39] utilized the illumination estimation technique to obtain the exposure ratio map and incorporated a camera response model to adjust image pixel according to the exposure ratio to solve lightness distortion.

Above conventional low light image enhancement approaches rely heavily on the parameter tuning to improve the subjective and objective quality of low light images. Recently, deep learning–based methods have been widely investigated, enhancing low light images directly in a data-driven way. Wei et al. [51] constructed a Retinex-based image decomposition network (RetinexNet) to learn the end-to-end mapping between low illumination and normal light image pairs. Wang et al. [46] proposed a GLobal illumination-Aware and Detail-preserving Network (GLADNet), including a global illumination estimation step and a detail reconstruction step. However, existing low light

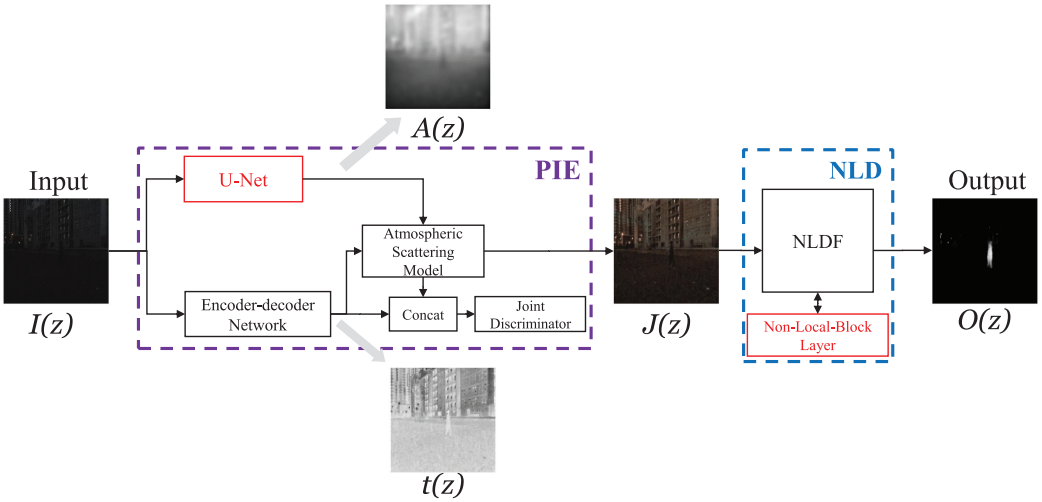

Fig. 2. An overview of the proposed method. The framework of our method consists of two sub-networks: Physical-based Image Enhancement (PIE) subnet and Non-local-based Detection (NLD) subnet. PIE attempts to generate a better image $J(z)$ from the given low light image $I(z)$ then benefit the SOD task. NLD aims to generate saliency map $O(z)$ from the enhanced image $J(z)$ by learning discriminant saliency features from the nighttime scene. For PIE, we treat the atmospheric light $A(z)$ as a point-wise random variate rather than a constant to follow the rules of nighttime light. For NLD, we reform the NLDF [32] by adding the Non-Local-Block Layer to provide a robust representation of saliency information towards low light images captured in non-uniform artificial light.

image enhancement mainly focuses on improving subjective visual quality, rather than facilitating subsequent high-level SOD task.

3 PROPOSED METHOD

The framework of our method consists of two sub-networks, i.e., Physical-based Image Enhancement (PIE) subnet and Non-Local-based Detection (NLD) Subnet. PIE enhances the image contrast by exploiting the relation among the atmosphere light $A(z)$ and the transmission map $t(z)$. NLD detects salient object from the enhanced images $J(z)$. Figure 2 shows the framework of our method. We explain these sub-networks in detail as follows.

3.1 PIE

Floating particles in the atmosphere greatly scatter the environment light in the nighttime scene, resulting in degradation in the image quality. This degradation causes severe object information loss in dark regions and in turn affects the performance of SOD. PIE aims at generating a better image J from the given low light image I then benefiting the SOD task.

Inspired by the dehazing method DCPDN [67], which is based on the atmospheric scattering model, we proposed the PIE for low light image enhancement. Although the atmospheric scattering model is utilized in the DCPDN for image dehazing, it is also capable of analyzing physical lighting of low light images, because of the existence of atmospheric particles in the nighttime scene. Therefore, following the DCPDN, the PIE also consists of four key modules, including a U-Net, an encoder-decoder network, an atmospheric scattering model, and a joint discriminator. However, different from the constant environment light in the typical hazy model, low light images are usually taken in non-uniform environmental light. The atmospheric light is treated as a

point-wise random variate in PIE rather than a constant in DCPDN to follow the rules of nighttime light.

U-Net is exploited to predict the atmospheric light $A(z)$. The encoder-decoder network is used to estimate the transmission map $t(z)$. Combining the results of $A(z)$ and $t(z)$, the atmospheric scattering model generates the enhanced image $J(z)$. Since the enhanced image and its corresponding transmission map should have strong structural relationship, $t(z)$ and $J(z)$ are concatenated together and the joint discriminator is used to distinguish whether a pair of estimated $t(z)$ and $J(z)$ is a real or fake pair.

3.1.1 U-Net. We used an eight-block U-Net structure [40] to estimate the atmospheric light. The U-Net can preserve rich texture information and has achieved spectacular performance on image enhancement [8, 21, 23, 33]. Another advantage of using the U-Net lies in its efficient GPU consumption. The U-Net consists of an encoder and a decoder. They are connected like a "U" shape. The encoder is composed of four Conv-BN-Relu blocks, while the decoder is composed of symmetric Dconv-BN-Relu block (Con: Convolution, BN: Batch-normalization, and Dconv: Deconvolution).

As we know, images captured in a low illumination environment are degraded with the corresponding environment lights. It is impossible to describe the changes in the incident light for each image pixel at the same level. Different from the constant environment light in the typical hazy model [67], low light images are usually taken in non-uniform environmental light. Therefore, we treat the atmospheric light $A(z)$ as a point-wise variate, changing with the local scene light source. Thus, we synthesize the training samples for U-Net, where $A(z)$ is randomly valued to generate the corresponding atmospheric light maps. It can be formulated as:

$$A(z) = 1 - \alpha * uniform(0, 1), \tag{1}$$

where $uniform(0, 1)$ randomly generates real numbers between 0 and 1. To simplify our method, we set $\alpha = 0.5$ in this article.

3.1.2 Encoder-decoder Network. We used an encoder-decoder network to estimate the transmission map $t(z)$. The encoder-decoder architecture has achieved spectacular performance on image dehazing [68] and image enhancement [38, 51]. The encoder-decoder network can keep the structural information of the object and produce the high-resolution feature map from low-resolution saliency map. The encoder is composed of the first Conv layer and the first three Dense-Blocks with their corresponding down-sampling operations Transition-Blocks from a pre-trained dense-net121. The decoder consists of five dense blocks with the refined up-sampling Transition-Blocks. The function of the encoder is to leverage the pre-defined weights of the dense-net [20], and the function of the decoder is to reconstruct the transmission map into the original resolution.

3.1.3 Atmospheric Scattering Model. After estimating the atmospheric light $A(z)$ and transmission map $t(z)$, the target image $J(z)$ can be estimated via the atmospheric scattering model. The atmospheric scattering model is derived from McCartney's scattering theory, which assumes the existence of atmospheric particles and has been put into practice in haze removal [1, 50]. The atmospheric scattering model is also suitable for low light image enhancement, because there are similarities between low light images and hazy images. The scattering particles exist everywhere; even on clear sunny days [17], the scattering phenomenon caused by which is a cue to the aerial perspective [35]. Therefore, the existence of light scattering is necessary for low light images. Accordingly, the atmospheric scattering model for low light images can be composed of two items: the direct attenuation term and the scattering light term. The former represents the object reflects light that is not scattered by the scattering particles. While the latter is a part of the scattered

environment light that reaches the camera. The atmospheric scattering model for low light image can be mathematically expressed as:

$$I(z) = J(z)t(z) + A(z)(1 - t(z)), \tag{2}$$

where J is the enhanced target image, I is the observed low light image, z is the location of the pixel. Different from the constant environment light in the typical hazy model, A is a point-wise variate and changes with the local scene light source.

3.1.4 Joint Discriminator Learning. According to Zhang et al. [67], the structural information between the transmission map $t(z)$ and the enhanced image $J(z)$ is highly correlated. Therefore, we use the joint discriminator learning to refine the enhanced image $J(z)$. The joint discriminator learning aims to make sure that the estimated transmission map $t(z)$ and the enhanced image $J(z)$ are indistinguishable from their corresponding ground truths, respectively. It is formulated as:

$$\begin{aligned} \min_{G_t, G_d} \max_{D_{joint}} & \mathrm{E}_{I \sim Pdata(I)} \left[\log \left(1 - D_{joint} \left(G_t(I) \right) \right) \right] \\ & + \mathrm{E}_{I \sim Pdata(I)} \left[\log \left(1 - D_{joint} \left(G_d(I) \right) \right) \right] \\ & + \mathrm{E}_{t,J \sim Pdata(t,J)} \left[\log D_{joint}(t, J) \right)], \end{aligned} \tag{3}$$

where G_t and G_d denote the networks generating the transmission map and the enhanced result, respectively. The joint discriminator learning process exploits the structural correlation between the transmission map and the enhanced image.

3.2 NLD

NLD is a SOD model to learn discriminant saliency features and generate saliency map $O(z)$ from the enhanced image $J(z)$. As illustrated in Figure 2, NLD follows the architecture of our previous work [59]. Different from Reference [59] that conducts low illumination SOD directly on original degraded images, NLD detects salient object from the enhanced images. The Non-Local-Block Layer is utilized to capture the difference of each feature against its local neighborhood favoring regions in the enhanced image. Those regions are either brighter or darker than their neighbors, and the differences catch more details. Therefore, the extracted non-local feature can reflect both the local and global context of an image by incorporating the details of various resolutions. And the detail information of the salient object can be retained to improve the SOD performance.

Figure 3 illustrates the architecture of NLD subnet for SOD. The first row of NLD contains five convolutional blocks derived from VGG-16 (CONV-1 to CONV-5). The goal of these convolutional layers is to learn feature maps X_1-X_5. The second layer contains five convolutional blocks (CONV-6 to CONV-10). Each block changes the number of channels to 128. The goal of these convolutional layers is to learn multi-scale local feature maps B_1-B_5. Then, Non-Local-Block Layer obtains more useful features from enhanced images and learns feature maps C_1–C_5. The last row is a set of deconvolution layers (UNPOOL-2 to UNPOOL-5) to generate U_2-U_5. A 1×1 convolution is added after C_1 to sum the number of channels to 640, and then the local feature map is obtained. Finally, the SCORE block has 2 convolution layers and a softmax to compute the saliency probability by fusing the local and global features.

As illustrated in Figure 4, the proposed Non-Local-Block Layer consists of two operations: 1×1 convolution and softmax. The 1×1 convolution is used to generate feature maps, while the softmax is utilized to store the similarity of any two pixels. Motivated by Reference [47], the similarity of any two pixels is calculated by non-local mean [4] and bilateral filters [41], ensuring the feature

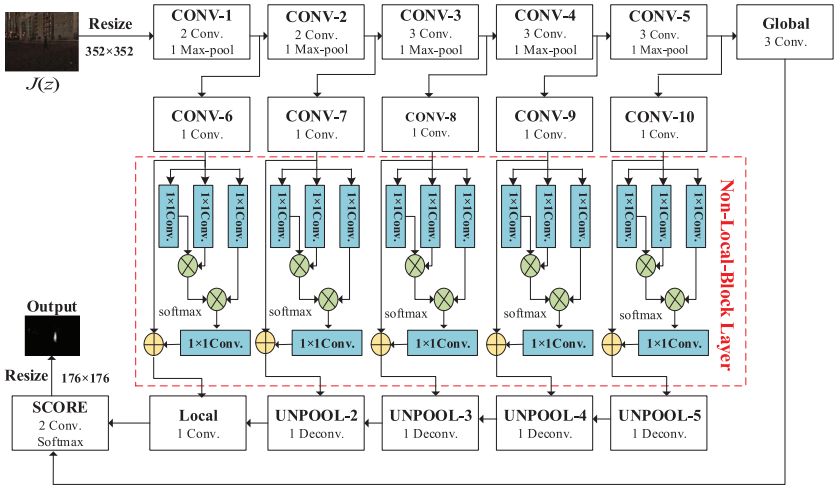

Fig. 3. Architecture of Non-Local-based Detection (NLD) subnet for salient object detection. $J(z)$ is the output of the former PIE and the input of NLD. The red region indicates the proposed Non-Local-Block Layer.

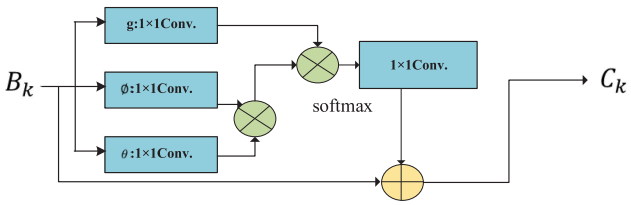

Fig. 4. Architecture of Non-Local-Block Layer. The softmax operation is performed on each row.

map can be embedded into Gaussian after 1×1 convolution. It is formulated as:

$$f(x_i, x_j) = e^{(W_\theta x_i)^T W_\phi x_j}, \tag{4}$$

where x_i and x_j represent two pixels of each feature map B_1-B_5. W_θ and W_ϕ are the weights of the convolution layers. A pairwise function f computes a scalar (representing relationship such as affinity) between i and all j. After the convolution, the number of channels becomes a half of the initial size.

The similarity calculated above is stored in the feature maps by the mean of self-attention. It is defined by $y_k = softmax(B_k^T W_\theta^T W_\phi B_k) g(B_k)$. For simplicity, we only consider g in the form of a linear embedding: $g(B_k) = W_g B_k$, where W_g is a weight matrix to be learned. Then, we use 1×1 convolutions to recover the number of channels. After that, the feature map C_k, $k = 1, \ldots, 5$ is obtained through a residual operation using y_k and B_k via:

$$C_k = W_B y_k + B_k, \tag{5}$$

where W_B is a weighting parameter to restore the same number of channels y_k as B_k. "$+B_k$" denotes a residual connection. The residual connection allows us to insert a new non-local block into any pre-trained model. After processing by the non-local network layer B_k, the size of the feature map C_k remains the same. By doing so, the pixel information of feature maps can be reserved.

3.3 Overall Loss Function

In PIE, the atmospheric light and transmission map are learned simultaneously, where a joint loss function is utilized to combine the atmospheric light estimation error and the transmission map estimation error. Different from DCPDN [67], which adopts the *L2* loss in predicting the atmospheric light, PIE minimizes the MSE loss function between the estimated value $A(z)$ and corresponding ground truth obtained from dark channel prior [16]. The MSE loss can be calculated as follows:

$$L^a = \frac{1}{NHW} \sum_{i=1}^{N} \|A(z) - A_{gt}\|^2, \tag{6}$$

where H and W are the height and the width of the image, respectively. And N is the total number of training batches.

In NLD, the local features L and global features G are linearly combined as follows:

$$\hat{y}(v) = p(y(v) = c) = \frac{e^{W_L^c L(v) + b_L^c + W_G^c G + b_G^c}}{\sum_{c' \in \{0,1\}} e^{W_L^{c'} L(v) + b_L^{c'} + W_G^{c'} G + b_G^{c'}}}, \tag{7}$$

where (W_L, b_L) and (W_G, b_G) are two linear operators. $y(v)$ represents the ground truth. The final saliency map is denoted as $\hat{y}(v_i)$.

The cross-entropy loss function can be formulated as follows:

$$H_j(y(v), \hat{y}(v)) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \{0,1\}} (y(v_i) = c) (\log(\hat{y}(v_i) = c)). \tag{8}$$

To make the boundary robust to background noise, the IoU boundary loss of NLDF [32] is utilized and can be calculated as follows:

$$\text{IoU}(C_i, C_j) = 1 - \frac{2|C_i \cap C_j|}{|C_i| + |C_j|}. \tag{9}$$

Finally, the overall loss function can be obtained by the combination of the cross-entropy loss function and the IoU boundary loss:

$$\text{Total Loss} \approx \sum_j \lambda_j \int H_j(y(v), \hat{y}(v)) + \sum_j \gamma_j (1 - \text{IoU}(C_i, C_j)). \tag{10}$$

4 NIGHTTIME IMAGE DATASET FOR SOD

We build a NightTime Image - V1 (NTI-V1) dataset for SOD. NTI-V1 contains 577 low light images, each image accompanied by pixel-level human-labeled ground-truth annotation. These images are captured at the nighttime of spring-summer and the autumn-winter from the indoor and the outdoor scene of our university. And we incorporate various challenges, such as viewpoint variation, changing illumination, and diverse scenes. The dataset was collected in two stages. In the first stage, 224 high-resolution images were captured by one surveillance camera from 7 PM to 9 PM. In the second stage, 353 images were captured by three smartphones from 9 PM to 11 PM. After the collection, five volunteers are invited to annotate the salient objects with bounding boxes. The shared image regions (with IoU > 0.8) of these bounding boxes are kept as the salient objects. To provide high-quality annotations, we further manually label the accurate silhouettes of the salient objects via the "LabelMe" software. Figure 5 shows some examples of the NTI-V1 dataset. The dataset includes three types of objects: single person (Figure 5(a)), multiple persons (Figure 5(b)), and vehicle (such as bicycle, car, etc.) (Figure 5(c)). Figure 6 shows the data collection division of

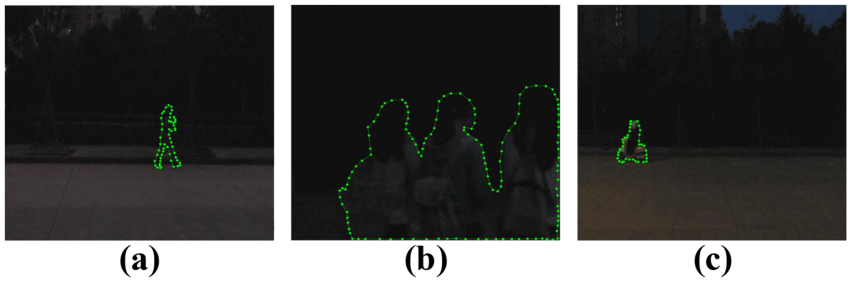

Fig. 5. Example samples of the NTI-V1 dataset.

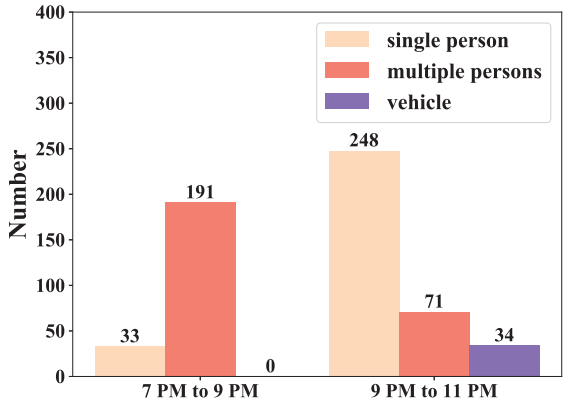

Fig. 6. The data collection division of the NTI-V1 dataset.

Table 1. Comparing the NTI-V1 Dataset with Existing Low Light Image Datasets and SOD Datasets

Dataset	#Img.	#DT	#NT	#Obj.	Annotation
SID	5,518	424	5,094	–	✗
LOL	1,000	500	500	–	✗
MSRA-B	5,000	4,964	36	1-2	✓
SOD	300	299	1	1-4+	✓
ECSSD	1,000	998	2	1-4+	✓
DUT-OMRON	5,168	5,166	2	1-4+	✓
PASCAL-S	850	842	8	1-4+	✓
NTI-V1 (Ours)	577	0	577	1-4+	✓

the NTI-V1 dataset. In our evaluation protocol, 457 images are used for training, and 120 images are used for testing.

To the best of our knowledge, although there are some datasets for the low light image enhancement, there is no related dataset for evaluating the performance of low light image SOD. In Table 1, we make a comparison of related datasets, including See-In-the-Dark (SID) [6], LOw Light image dataset (LOL) [51], SOD [34], MSRA-B [31], ECSSD [60], DUT-OMRON [61], and PASCAL-S [30]. For the former two datasets, they do not contain salient object segmentation, thus are inappropriate for SOD. Similar to the following five datasets, each image of the NTI-V1 dataset is accompanied

with pixel-level ground-truth annotation. However, those datasets are generally constructed at daytime, containing very few low light images. Hence, our dataset is the first available benchmark dataset for the low light image SOD.

To facilitate the research of low light SOD, we collect a dataset called NTI-V1 Dataset with following distinct features: (1) It contains 577 images captured in nighttime, and each of which is accompanied by pixel-level human-labeled ground-truth annotation. (2) The dataset is captured by one surveillance camera from 7 PM to 9 PM and three smartphones from 9 PM to 11 PM, which covers a large area of districts and at different times. (3) It contains multiple salient objects per image including three types of objects: single person, multiple persons, and vehicle (such as bicycle and car). And (4) the capture conditions involve various viewpoints, illumination changes, and different scenes.

5 EXPERIMENTS

5.1 Datasets and Experimental Settings

We conduct extensive experiments on five SOD datasets, including DUT-OMRON, ECSSD, PASCAL-S, SOD, and our proposed NTI-V1 dataset. The former four are generally built in a bright environment and are widely used in the SOD field. While the latter NTI-V1 dataset is built in the nighttime scene.

DUT-OMRON. The DUT-OMRON dataset consists of 5,168 high-quality images. Images in this dataset have one or more salient objects and relatively complex backgrounds. Thus, it is challenging for saliency detection.

ECSSD. The ECSSD dataset contains 1,000 images with semantic meaning in their ground truth segmentation. It also contains images with complex structures.

PASCAL-S. The PASCAL-S dataset contains 850 challenging images (each composed of several objects), all of which are chosen from the validation set of the PASCAL VOC 2010 segmentation dataset.

SOD. The SOD dataset contains 300 images designed for image segmentation. This dataset provides ground truth for the boundaries of salient objects perceived by humans in natural images.

NTI-V1. We constructed the NTI-V1 dataset, which contains 577 natural scene images under low illumination. This dataset contains three types of objects hand-labeled as the ground truth, including single person, multiple persons, and vehicle (such as bicycle, car, etc.).

Saliency Evaluation Metrics. We adopt three widely used metrics to measure the performance of all algorithms, the Precision-Recall (PR) curves, F-measure, and Mean Absolute Error (MAE) [3]. The precision and recall are computed by thresholding the predicted saliency map and comparing the binary map with the ground truth. The PR curve of a dataset indicates the mean precision and recall of saliency maps at different thresholds. The F-measure is a balanced mean of average precision and average recall and is calculated by $F_\beta = \frac{(1+\beta^2)\times \text{Precision}\times \text{Recall}}{\beta^2 \times \text{Precision}+\text{Recall}}$, where β^2 is set to 0.3 to emphasize the precision over recall [42]. The maximum F_β (max F_β) of each dataset is reported in this article. We also calculate the MAE for fair comparisons as suggested by Reference [3]. The MAE evaluates the saliency detection accuracy by $\text{MAE} = \frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y) - L(x,y)|$, where $S(x,y)$ is the predicted salient map and $L(x,y)$ is the ground truth. The parameters W and H represent the width and height of the image, respectively.

Implementation Details. Our model was built on PyTorch. We set related hyper-parameters of PIE following Reference [67]. During training, we utilized the Adam optimization with the learning rate of 2×10^{-3} for both generator and discriminator. All the training samples were resized to 512×512. For the NLD, the weights of CONV-1 to CONV-5 were initialized by the VGG-16 network. All weights in the network were initialized randomly by a truncated normal distribution ($\sigma = 0.01$)

Table 2. Quantitative Results of Our Method on Three Types of Objects in Two Stages

Type	Criteria	7 PM–9 PM	9 PM–11 PM
single person	MAE↓	0.004	0.024
	max F_β ↑	0.921	0.695
multiple persons	MAE↓	0.011	0.035
	max F_β ↑	0.795	0.706
vehicle	MAE↓	-	0.018
	max F_β ↑	-	0.797

and the biases were initialized to zero. We also used Adam optimization with the learning rate of 10^{-6}. 457 images from the NTI-V1 dataset were fed into the network for training, and in turn, the other 120 images were used for testing.

5.2 Comparison with State-of-the-art Methods

To evaluate the proposed algorithm, extensive tests have been performed using a set of SOD methods, including MR [61], BSCA [37], DSSC [19], LPS [63], R3Net [10], and BASNet [36]. Qualitative and quantitative evaluations are explored to comprehensively evaluate the performance of PIE for SOD with other seven low light image enhancement methods, including Gamma [12], LIME [15], LECARM [39], Dong [11], Ying [62], RetinexNet [51], and GLADNet [46].

To investigate the influence of low illumination and object type, we conducted experiments on three types of objects in two stages separately. Table 2 summarizes the results in terms of MAE and max F_β. We can observe that the results of 7 PM to 9 PM is superior to the results of 9 PM to 11 PM, due to the illumination conditions around 7 PM to 9 PM being better. However, there is no obvious phenomenon for different types of objects.

To evaluate the effectiveness of our image enhancement methods for SOD as well as to promote further research on this new problem, we adopt three types of performance evaluation: (1) To verify the effectiveness of our proposed methods on low light images, we compare our proposed method with several SOD methods on low light images. (2) To verify the effectiveness of our PIE for low light SOD, we enhance low light images by PIE and then compare our proposed method with several SOD methods on the enhanced images. (3) To verify the appropriateness of PIE for NLD, we compare our proposed method on enhanced images generated via different image enhancement methods.

5.2.1 Comparison with State-of-the-art SOD Methods on Low Light Images. We compared our method with several state-of-the-art SOD methods, including MR [61], BSCA [37], DSSC [19], LPS [63], R3Net [10], and BASNet [36] on five datasets. Table 3 shows the comparison results in terms of MAE and max F_β for all datasets. We can observe that our method does not obtain the best performances on the four public daytime SOD datasets; while our method beats the state-of-the-art methods on the NTI-V1 dataset. To further evaluate the quality of SOD methods, we compared their PR curves on the NTI-V1 dataset, as shown in Figure 7(a). Our method achieves a better PR curve than all the other methods. It shows that our method achieves the best performance on the NTI-V1 dataset with respect to both two metrics. It also indicates that our method is more effective to detect salient objects in low light images, although not better than the others on the images with sufficient light.

In Figure 8, we show the qualitative results. The non-deep learning–based methods MR and BSCA perform badly on the low light images. The DSSC method is difficult to learn useful features

Table 3. Benchmarking Results of Six State-of-the-art SOD Models on Five Datasets

Dataset	Criteria	MR	BSCA	DSSC	LPS	R3Net	BASNet	Ours
DUT-OMRON	MAE↓	0.187	0.191	0.065	0.064	0.063	0.056	0.069
	max F_β ↑	0.610	0.616	0.720	0.635	0.795	0.805	0.667
ECSSD	MAE↓	0.189	0.183	0.062	0.087	0.040	0.037	0.122
	max F_β ↑	0.736	0.758	0.873	0.814	0.934	0.942	0.775
PASCAL-S	MAE↓	0.223	0.224	0.103	0.041	0.092	0.076	0.133
	max F_β ↑	0.666	0.666	0.773	0.694	0.834	0.854	0.761
SOD	MAE↓	0.273	0.266	0.126	0.169	0.125	0.114	0.118
	max F_β ↑	0.619	0.634	0.787	0.707	0.850	0.851	0.849
NTI-V1	MAE↓	0.355	0.326	0.027	0.029	0.033	0.028	0.026
	max F_β ↑	0.138	0.136	0.481	0.678	0.591	0.557	0.745

DUT-OMRON, ECSSD, PASCAL-S, SOD, and our Newly Constructed NTI-V1. Three top results are highlighted in red, blue, and green, respectively. The up-arrow ↑ shows the larger the value achieves, the better the performance is. The down-arrow ↓ has the opposite meaning.

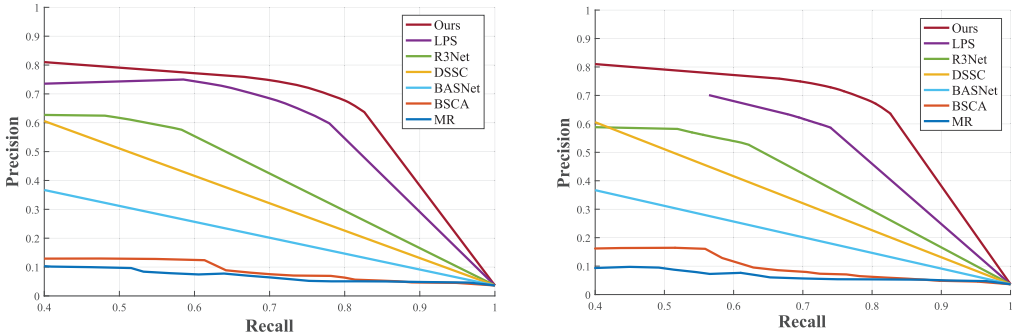

Fig. 7. PR curves on the NTI-V1 dataset. (a) PR curves for our method compared to LPS [63], R3Net [10], DSSC [19], BASNet [36], BSCA [37], and MR [61] on NTI-V1 dataset. (b) PR curves for our method compared to LPS [63], R3Net [10], DSSC [19], BASNet [36], BSCA [37], and MR [61] on enhanced NTI-V1 dataset via PIE.

for the nighttime scene. The pixel-based method LPS produces a lot of false detection due to noise interference. The methods BASNet and R3Net lost many saliency details and tend to contain non-saliency backgrounds. Comparing the results from the $2nd$ to the $8th$ columns, we can observe that our method exhibits sharper and uniformly highlighted salient objects, and the saliency maps are closer to the ground truth (the $9th$ column).

5.2.2 Comparisons with Several SOD Methods on Low Light Images Enhanced Via PIE. To verify the effectiveness of our PIE for low light SOD, we first enhanced the NTI-V1 datasets by our PIE, then the enhanced images were trained and tested by our NLD and the other state-of-the-art SOD methods. Table 4 shows the comparison results in terms of MAE and max F_β on the NTI-V1 dataset. It is obvious that PIE can improve the performance of SOD compared to Table 3. To further evaluate the quality of SOD methods, we compared their PR curves on the NTI-V1 dataset, as shown in Figure 7(b). Our method achieves a better PR curve than all the other methods on the NTI-V1 dataset. It shows that our method achieves the best performance on the NTI-V1 dataset

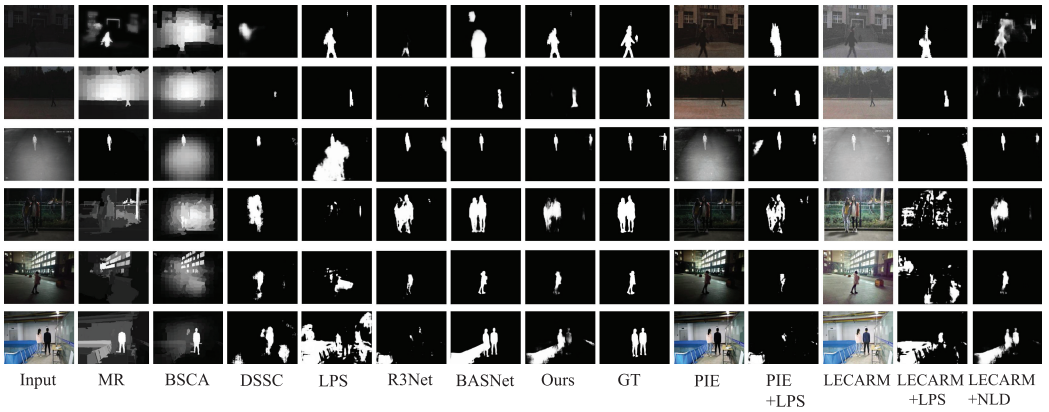

Fig. 8. Selected qualitative evaluation results on the NTI-V1 dataset. The 1*st* column shows the input images. From the 2*nd* column to the 8*th* column are, respectively, the SOD results of MR [61], BSCA [37], DSSC [19], LPS [63], R3Net [10], BASNet [36], and the proposed method. The 9*th* column is the ground truth. The 10*th* and 12*th* columns are the enhanced images by our PIE and the LECARM method, respectively. The 11*th* and 13*th* columns are the SOD results of the LPS method on the enhanced images of the 10*th* and 12*th* columns, respectively. The 14*th* column shows the SOD results of our NLD on the enhanced images of the 12*th* column.

Table 4. Benchmarking Results of Six State-of-the-art SOD Methods on Enhanced NTI-V1 Dataset Via PIE

Method	MR	BSCA	DSSC	LPS	R3Net	BASNet	**Ours**
MAE↓	0.351	0.306	0.029	0.036	0.031	0.028	0.026
max F_β ↑	0.140	0.144	0.458	0.689	0.549	0.551	0.745

The dataset was first enhanced by our PIE, then evaluated by different saliency detection models.

with respect to both two metrics. Moreover, our image enhancement method PIE not only has effectiveness for our SOD model (NLD), but also other existing SOD models.

In Figure 8, we show the qualitative results. Comparing the "PIE" and "Input," we can observe that PIE improves the brightness and contrast of the low light images obviously, outstanding the salient object. Furthermore, it is observed that "PIE+LPS" (11*th* column) achieved better SOD results than the method "LPS" (5*th* column), which shows the effectiveness of PIE for SOD. Furthermore, from these results, we can obverse that "PIE+LPS" (11*th* column) includes non-saliency backgrounds compared to "Ours" (8*th* column). Our method (8*th* column) is more close to the ground truth (the 9*th* column).

5.2.3 Comparisons Among PIE and Several Image Enhancement Methods. To verify the appropriateness of PIE for NLD, low light images were, respectively, enhanced by Gamma [12], LIME [15], LECARM [39], Dong [11], Ying [62], RetinexNet [51], GLADNet [46], and our PIE. Then, the enhanced images were trained and tested by our NLD. Table 5 shows the comparison results in terms of MAE and max F_β on NTI-V1 dataset. It is clear that our method achieves the best results with respect to both two metrics, verifying the appropriateness of PIE for NLD.

In Figure 8, we show the qualitative results. The "PIE" (10*th* column) and "LECARM" (12*th* column) can improve the brightness and contrast of the low light images obviously compared with the "Input" (1*st* column). Furthermore, the results of "PIE+LPS" (11*th* column) and "LECARM+LPS" (13*th* column), "Ours"(8*th* column) and "LECARM+NLD" (14*th* column) indicate that the PIE's

Table 5. Benchmarking Results of Seven State-of-the-art Image Enhancement Methods on NTI-V1 Dataset for SOD by NLD

Method	Gamma	LIME	LECARM	Dong	Ying	RetinexNet	GLADNet	**Ours**
MAE↓	0.060	0.050	0.052	0.067	0.064	0.041	0.052	0.026
max F_β ↑	0.575	0.484	0.577	0.469	0.494	0.424	0.465	0.745

The dataset was first enhanced by different image enhancement models, then evaluated with our saliency detection model NLD.

Table 6. Ablation Study Results of $A(z)$ and Non-Local-Block Layer on the NTI-V1 Dataset

Method	MAE↓	max F_β ↑
Ours w/ Constant $A(z)$	0.040	0.741
Ours w/o NLB Layer	0.047	0.704
Ours	**0.026**	**0.745**

Note that our method is with a random value $A(z)$ and the NLB Layer.

performance improvement for LPS is lower than LECARM's, but the PIE's performance improvement for NLD is better than LECARM's. Visualization results also show this phenomenon that our method achieved better SOD results than the methods "PIE+LPS" ($11th$ column) and "LECARM+NLD" ($14th$ column). From these results, we can obverse that "PIE+LPS" ($11th$ column) and "LECARM+NLD" ($14th$ column) tend to lose many saliency details and include non-saliency backgrounds. Our method can accurately segment salient objects in low light.

5.3 Ablation Study

The ablation experiments are conducted on the NTI-V1 dataset.

5.3.1 The Ablation Study of $A(z)$. Our PIE differs from DCPDN [67]. We believe that the atmospheric light $A(z)$ is a point-wise random value rather than a constant. So, we treat $A(z)$ as a random value instead of a constant. Here, we compared our method with a constant $A(z)$. Table 6 shows that our method improves MAE by 1.4% and max F_β by 0.4% on the NTI-V1 dataset. It validates our designing for the atmospheric light $A(z)$.

5.3.2 The Ablation Study of Non-Local-Block Layer. The structure of NLD is similar to NLDF [32]. However, our NLD utilizes additional Non-Local-Block (NLB) Layers to calculate the similarity among different pixels. Table 6 shows that our method improves MAE by 2.1% and max F_β by 4.1% on the NTI-V1 dataset, compared with our method without the NLB Layer. It validates our designing for the NLB Layer.

6 CONCLUSION

In this work, we propose an image-enhancement-based SOD for low light images, which is critical for CV applications in low light conditions [25, 64]. This method directly embeds the physical lighting model into the deep neural network to describe the degradation of low light images, and in turn, utilizes a Non-Local-Block Layer to extract non-local features of salient objects. Further, we construct an NTI-V1 dataset containing 577 low light images with pixel-wise object-level annotations for the SOD community. With extensive experiments, we verify that low illumination can actually reduce the performance of SOD. The proposed method is effective to enhance local content in low light images to facilitate the SOD task.

REFERENCES

[1] Codruta Orniana Ancuti and Cosmin Ancuti. 2013. Single image dehazing by multi-scale fusion. *IEEE Trans. Image Proc.* 22, 8 (May 2013), 3271–3282. DOI: https://doi.org/10.1109/TIP.2013.2262284

[2] Sten Andler. 2018. Depth-aware stereo video retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 6517–6525.

[3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE Trans. Image Proc.* 24, 12 (Dec. 2015), 5706–5722. DOI: https://doi.org/10.1109/TIP.2015.2487833

[4] Antoni Buades, Bartomeu Coll, and J.-M. Morel. 2005. A non-local algorithm for image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE Computer Society, 60–65. DOI: https://doi.org/10.1109/CVPR.2005.38

[5] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. 2011. Fusing generic objectness and visual saliency for salient object detection. In *Proceedings of the International Conference on Computer Vision (ICCV'11)*. IEEE Computer Society, 914–921. DOI: https://doi.org/10.1109/ICCV.2011.6126333

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 3291–3300.

[7] Lei Chen, Le Wu, Zhenzhen Hu, and Meng Wang. 2019. Quality-aware unpaired image-to-image translation. *IEEE Trans. Multimedia* 21, 10 (Oct. 2019), 2664–2674. DOI: https://doi.org/10.1109/TMM.2019.2907052

[8] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. 2018. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 6306–6314. DOI: https://doi.org/10.1109/CVPR.2018.00660

[9] Guang Deng. 2010. A generalized unsharp masking algorithm. *IEEE Trans. Image Proc.* 20, 5 (Nov. 2010), 1249–1261. DOI: https://doi.org/10.1109/TIP.2010.2092441

[10] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. 2018. R3Net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press, 640–690. DOI: https://doi.org/10.24963/ijcai.2018/95

[11] Xuan Dong, Guan Wang, Yi Pang, Weixin Li, Jiangtao Wen, Wei Meng, and Yao Lu. 2011. Fast efficient algorithm for enhancement of low lighting video. In *Proceedings of the ACM SIGGRAPH 2010 Posters (SIGGRAPH'10)*. Association for Computing Machinery, 1–6. DOI: https://doi.org/10.1145/1836845.1836920

[12] Hany Farid. 2001. Blind inverse gamma correction. *IEEE Trans. Image Proc.* 10, 10 (Oct. 2001), 1428–1433. DOI: https://doi.org/10.1109/83.951529

[13] Wei Feng, Ruize Han, Qing Guo, Zhu, and Song Wang. 2019. Dynamic saliency-aware regularization for correlation filter-based object tracking. *IEEE Trans. Image Proc.* 28, 7 (Jan. 2019), 3232–3245. DOI: https://doi.org/10.1109/TIP.2019.2895411

[14] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. 2016. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, 2782–2790. DOI: https://doi.org/10.1109/CVPR.2016.304

[15] Xiaojie Guo, Yu Li, and Haibin Ling. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Proc.* 26, 2 (Feb. 2016), 982–993. DOI: https://doi.org/10.1109/TIP.2016.2639450

[16] Kaiming He, Sun Jian, and Xiaoou Tang. 2011. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 12 (Dec. 2011), 2341–2353. DOI: https://doi.org/10.1109/TPAMI.2010.168

[17] Kaiming He, Jian Sun, and Xiaoou Tang. 2011. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 12 (Dec. 2011), 2341–2353. DOI: https://doi.org/10.1109/TPAMI.2010.168

[18] Edwin H. Land. 1986. An alternative technique for the computation of the designator in the Retinex theory of color vision. *Proc. Nat. Acad. Sci. United States Amer.* 83, 10 (May 1986), 3078–3080. DOI: https://doi.org/10.1073/pnas.83.10.3078

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2019. Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (Jan. 2019), 815–828.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE, 4700–4708. DOI: https://doi.org/10.1109/CVPR.2017.243

[21] Jie Huang, Pengfei Zhu, Mingrui Geng, Jiewen Ran, Xingguang Zhou, Chen Xing, Pengfei Wan, and Xiangyang Ji. 2018. Range scaling global U-net for perceptual image enhancement on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer-Verlag, Munich, Germany, 230–242.

[22] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng. 2013. Salient region detection by UFO: Uniqueness, focusness and objectness. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'13)*. IEEE, 1976–1983. DOI: https://doi.org/10.1109/ICCV.2013.248

[23] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2019. EnlightenGAN: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972* (2019).

[24] Daniel J. Jobson, Zia ur Rahman, and Glenn A. Woodell. 1997. A multiscale Retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Proc.* 6, 7 (Aug. 1997), 965–976. DOI: https://doi.org/10.1109/83.597272

[25] Kajal Kansal, A. V. Subramanyam, Zheng Wang, and Shin'Ichi Satoh. 2020. SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification. *IEEE Trans. Circ. Syst. Vid. Technol.* DOI: https://doi.org/10.1109/TCSVT.2019.2963721

[26] Chulwoo Lee, Chul Lee, and Chang-Su Kim. 2013. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Trans. Image Proc.* 22, 12 (Sept. 2013), 5372–5384. DOI: https://doi.org/10.1109/ICIP.2012.6467022

[27] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE, 5455–5463. DOI: https://doi.org/10.1109/CVPR.2015.7299184

[28] Lin Li, Ronggang Wang, Wenmin Wang, and Wen Gao. 2015. A low-light image enhancement method for both denoising and contrast enlarging. In *Proceedings of the International Conference on Image Processing (ICIP'15)*. IEEE, 3730–3734. DOI: https://doi.org/10.1109/ICIP.2015.7351501

[29] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. 2016. DeepSaliency: Multi-task deep neural network model for salient object detection. *IEEE Trans. Image Proc.* 25, 8 (Aug. 2016), 3919–3930. DOI: https://doi.org/10.1109/TIP.2016.2579306

[30] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. IEEE Computer Society, 280–287. DOI: https://doi.org/10.1109/CVPR.2014.43

[31] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. 2011. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2 (Feb. 2011), 353–367. DOI: https://doi.org/10.1109/CVPR.2007.383047

[32] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. 2017. Non-local deep features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE, 6609–6617. DOI: https://doi.org/10.1109/CVPR.2017.698

[33] Feifan Lv and Feng Lu. 2019. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *arXiv preprint arXiv:1908.00682* (2019).

[34] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'01)*. IEEE, 416–423. DOI: https://doi.org/10.1109/ICCV.2001.937655

[35] Arcot J. Preetham, Peter Shirley, and Brian Smits. 1999. A practical analytic model for daylight. In *Proceedings of the 26th Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'99)*. ACM Press/Addison-Wesley Publishing Co., 91–100. DOI: https://doi.org/10.1145/311535.311545

[36] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, 7479–7489.

[37] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. 2015. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE, 110–119. DOI: https://doi.org/10.1109/CVPR.2015.7298606

[38] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. 2019. Low-light image enhancement via a deep hybrid network. *IEEE Trans. Image Proc.* 28, 9 (Apr. 2019), 4364–4375. DOI: https://doi.org/10.1109/TIP.2019.2910412

[39] Yurui Ren, Zhenqiang Ying, Thomas H. Li, and Ge Li. 2018. LECARM: Low-light image enhancement using the camera response model. *IEEE Trans. Circ. Syst. Vid. Technol.* 29, 4 (Apr. 2018), 968–981. DOI: https://doi.org/10.1109/TCSVT.2018.2828141

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI'15)*. Springer-Verlag, Munich, Germany, 234–241. DOI: https://doi.org/10.1007/978-3-319-24574-4_28

[41] Carlo Tomasi and Roberto Manduchi. 1998. Bilateral filtering for gray and color images. In *Proceedings of the 6th International Conference on Computer Vision (ICCV'98)*. IEEE, 839–846. DOI: https://doi.org/10.1109/ICCV.1998.710815

[42] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2018. Salient object detection with recurrent fully convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 7 (June 2018), 1734–1746. DOI: https://doi.org/10.1109/TPAMI.2018.2846598

[43] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. Multimedia* 14, 4 (Aug. 2012), 975–985. DOI: https://doi.org/10.1109/TMM.2012.2185041

[44] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and Tat-Seng Chua. 2012. Movie2Comics: Towards a lively video content presentation. *IEEE Trans. Multimedia* 14, 3 (June 2012), 858–870. DOI: https://doi.org/10.1109/TMM.2012.2187181

[45] Meng Wang, Xueliang Liu, and Xindong Wu. 2015. Visual classification by ℓ_1 -hypergraph modeling. *IEEE Trans. Knowl. Data Eng.* 27, 9 (Sept. 2015), 2564–2574. DOI: https://doi.org/10.1109/TKDE.2015.2415497

[46] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. 2018. GLADNet: Low-light enhancement network with global awareness. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG'18)*. IEEE, 751–755.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 7794–7803.

[48] Yang Wang. 2020. Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion. *arXiv preprint arXiv:2006.08159* (2020).

[49] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. 2017. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Trans. Image Proc.* 26, 3 (Mar. 2017), 1393–1404. DOI: https://doi.org/10.1109/TIP.2017.2655449

[50] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2016. Nighttime haze removal with illumination correction. *arXiv preprint arXiv:1606.01460* (2016).

[51] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep Retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference (BMVC'18)*. BMVA Press, Newcastle, UK, 1–12.

[52] Shikui Wei, Lixin Liao, Jia Li, Qinjie Zheng, Fei Yang, and Yao Zhao. 2019. Saliency inside: Learning attentive CNNs for content-based image retrieval. *IEEE Trans. Image Proc.* 28, 9 (May 2019), 4580–4593. DOI: https://doi.org/10.1109/TIP.2019.2913513

[53] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. 2019. Where-and-when to look: Deep Siamese attention networks for video-based person re-identification. *IEEE Trans. Multimedia* 21, 6 (June 2019), 1412–1424. DOI: https://doi.org/10.1109/TMM.2018.2877886

[54] Lin Wu, Yang Wang, Junbin Gao, Meng Wang, Zheng-Jun Zha, and Dacheng Tao. 2020. Deep co-attention based comparators for relative representation learning on person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* (Apr. 2020), 1–14. DOI: https://doi.org/10.1109/TNNLS.2020.2979190

[55] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. 2018. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Trans. Cyber.* 49, 5 (May 2018), 1791–1802. DOI: https://doi.org/10.1109/TCYB.2018.2813971

[56] Lin Wu, Yang Wang, and Ling Shao. 2018. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Proc.* 28, 4 (Apr. 2018), 1602–1612. DOI: https://doi.org/10.1109/TIP.2018.2878970

[57] Lin Wu, Yang Wang, Ling Shao, and Meng Wang. 2019. 3-D PersonVLAD: Learning deep global representations for video-based person reidentification. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 11 (Nov. 2019), 3347–3359. DOI: https://doi.org/10.1109/TNNLS.2019.2891244

[58] Lin Wu, Yang Wang, Hongzhi Yin, Meng Wang, and Ling Shao. 2020. Few-shot deep adversarial learning for video-based person re-identification. *IEEE Trans. Image Proc.* 29, 1 (Mar. 2020), 1233–1245.

[59] Xin Xu and Jie Wang. 2018. Extended non-local feature for visual saliency detection in low contrast images. In *Proceedings of the European Conference on Computer Vision (ECCV'18) Workshops*. Springer-Verlag, Munich, Germany, 580–592.

[60] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*. IEEE Computer Society, 1155–1162. DOI: https://doi.org/10.1109/CVPR.2013.153

[61] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*. IEEE, 3166–3173. DOI: https://doi.org/10.1109/CVPR.2013.407

[62] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. 2017. A new low-light image enhancement algorithm using camera response model. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17) Workshops*. IEEE, 3015–3022. DOI: https://doi.org/10.1109/ICCVW.2017.356

[63] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. 2018. Learning to promote saliency detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 1644–1653.

[64] Zelong Zeng, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. 2020. Illumination-adaptive person re-identification. *IEEE Trans. Multimedia*. DOI: https://doi.org/10.1109/TMM.2020.2969782

[65] Fangneng Zhan, Shijian Lu, and Chuhui Xue. 2018. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer-Verlag, Munich, Germany, 249–266.

[66] Dingwen Zhang, Junwei Han, Yu Zhang, and Dong Xu. 2019. Synthesizing supervision for learning deep saliency network without human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 8 (Feb. 2019), 1–14. DOI: https://doi.org/10.1109/TPAMI.2019.2900649

[67] He Zhang and Vishal M. Patel. 2018. Densely connected pyramid dehazing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, 3194–3203.

[68] He Zhang, Vishwanath Sindagi, and Vishal M. Patel. 2018. Multi-scale single image dehazing using perceptual pyramid deep network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18) Workshops*. IEEE, 902–911. DOI: https://doi.org/10.1109/CVPRW.2018.00135

[69] Jun Zhang, Meng Wang, Shengping Zhang, Xuelong Li, and Xindong Wu. 2016. Spatiochromatic context modeling for color saliency analysis. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 6 (June 2016), 1177–1189. DOI: https://doi.org/10.1109/TNNLS.2015.2464316

Received February 2020; revised July 2020; accepted July 2020