UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Low Light Image Enhancement and Saliency Object Detection

by Yuanfang Zhang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor Xiangjian He

University of Technology Sydney
Faculty of Engineering and Information Technology

May 2022

Certificate of Authorship/Originality

I, Yuanfang Zhang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology, at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Northwestern Polytechnical University.

This research is supported by the Australian Government Research Training Program.

Signature Production Note: Signature removed prior to publication.

Date 5th May, 2022

ABSTRACT

Low Light Image Enhancement and Saliency Object Detection

by

Yuanfang Zhang

Low light images represent a series of image types with great potential. Their research focuses on images and videos of the environment at dusk and near darkness. It can be widely used in night safety monitoring, license plate recognition, night scene shot, special target recognition at dusk, and other emergency events that occur under light scenes. After the environment is enhanced and combined with other tasks in computer vision and pattern recognition, it can bring many results, such as saliency detection and object detection under low illumination, and abnormal detection in crowded places under low-light environment. For the enhancement of low light and low light scenes, using traditional methods often results in over-exposure and halo conditions. Therefore, using deep learning network technology can fix and improve these specific shortcomings. To achieve this goal, we have done several investigations about the current state-of-art researches on low-light enhancement and the relevant computer vision tasks. For low light image enhancement, a series of qualitative and quantitative experimental comparisons conducted on a benchmark dataset demonstrate the superiority of our approach, which overcomes the drawbacks of white and colour distortion. At present, most of the research works on visual saliency have concentrated on the field of visible light, and there are few studies on night scenes. Due to insufficient lighting conditions in night scenes, and relatively lower contrasts and signal-to-noise ratios, the effectiveness of available visual features is greatly reduced. Moreover, without sufficient depth information, many features and clues are lost in the original images. Therefore, the detection of salient targets in night scenes is also difficult and it is a focus of current research in the field of computer vision. The performance leads to vague effects when the existing methods are directly con-

ducted, so we adopt a new “enhance firstly detection secondly” mechanism that firstly enhances the low-light images in order to improve the contrast and visibility, and then combines it with relevant saliency detection methods with depth information. Furthermore, we concern about the feature aggregation schemes for deep RGB-D saliency object detection and propose novel feature aggregation methods. Meanwhile, for the monocular vision, of which the depth information is hard to acquire, a novel RGB-D image saliency detection method is proposed to leverage depth cues for enhancing the saliency detection performance but without actually using depth data. Both of the extra depth cues and the proposed “enhance firstly detection secondly” mechanism can improve saliency detection abilities, according to the experimental results. The model not only outperforms the state-of-the-art RGB saliency models, but also achieves comparable or even better results compared with the state-of-the-art RGB-D saliency models

Dissertation directed by Professor Xiangjian He, Professor Michael Blumenstein and Doctor Wenjing Jia
Faculty of Engineering and Information Technology

Dedication

To my parents, and those who always love me and support me along the way.

Acknowledgements

My doctor study at UTS in the past three years has been a life-changing and priceless experience for me. Sydney is a lovely place. It has a golden light harbour with white sails, delicate and charming beaches, and a mild Mediterranean climate. The streets are filled with wild scents, lush forests, and soaring seagulls. Its natural beauty is enhanced by golden beaches and unspoiled bush lands. Sydney is a fantastic location for scientists to investigate science mysteries.

First and foremost, I would like to express my heartfelt appreciation to Professor Xiangjian He, my Principal Supervisor, who helped me tremendously by supplying me with required tools, valuable guidance, and inspiration for new ideas with exceptional patience and constant encouragement. His recommendations have drawn my attention to a variety of flaws and clarified many questions for me.

I am also grateful to Professor Michael Blumenstein and Doctor Wenjing Jia, for being my co-supervisors, especially to Dr Jia for her thoughtfulness and generosity in arranging study meetings in Professor He's Computer Vision and Pattern Recognition (CVPR) group. In research, Dr Jia is very proactive and detail-oriented, and her passion for collaboration and ability to contribute has bonded all of the research students into a family. With Dr Jia and the CVPR study community, I am really enjoying studying sophisticated deep learning techniques in image analysis and pattern recognition.

Then, I would like to express my gratitude to my classmates for their invaluable help with the original manuscript. They kindly made important remarks and sound recommendations to the paper's outline.

I would like to thank the teachers, writers, and colleagues at UTS for their time and commitment. Special thanks to Xiaochen Fan, Yue Xi, Xudong Song,

Qingqing Wang, Edward Huang, Saeed Amirgholipour, Muhammad Usman, Hesam Hesamian, Haodong Chang, Farhan Mohammed, and Chengpei Xu, who provided me with a great deal of assistance at different times.

Finally, I would like to express my gratitude to my parents for their encouragement in allowing me to study abroad. It is very beneficial to me in completing this study and I always love them.

In one word, I would like to express my thanks to all who have assisted me in finishing this study.

List of Publications

Journal Papers

J-1. Zhang Y, Zheng J, Li L, Nian L, Wenjing Jia, Xiaochen Fan, Chengpei Xu, Xiangjian He. Rethinking feature aggregation for deep RGB-D salient object detection [J]. Neurocomputing, 2021, 423: 463-473.

J-2. Zhang Y F , Zheng J , Jia W , et al. Deep RGB-D Saliency Detection without Depth[J]. IEEE Transactions on Multimedia, 2021, PP(99):1-1. Doi: 0.1109/TMM.2021.3058788

J-3. Zhang Y, Zheng J, Fei Li, Wenjing Jia, Wenfeng Huang, Xiangjian He. Low Light Image Dedarking via Deep Semantic Fusion[J]. IEEE Signal Processing Letter (Under Review)

Contents

List of Figures

List of Tables

Abbreviation

ASPP - Atrous Spatial Pyramid Pooling

DASPP - Dense Atrous Spatial Pyramid Pooling

SOD - Saliency Object Detection

CGA - Convolutional Gated Attention

SA - Spatial Attention

CBAM- Convolutional Block Attention Module

DMSF- Dense MultiScale Fusion

NIQE- Natural image quality evaluator

PSNR- Peak Signal to Noise Ratio

SSIM- Structural Similarity

CNN - Convolutional Neural Networks

ExDARK - A Dataset for low light image enhancement

MSR - Multi-scale Retinex

LIME - A Dataset for low light image enhancement

LECARM - A Dataset for low light image enhancement

AAP - Adaptive Average Pooling

BN - Batch Normalization

ReLU - Rectified Linear Unit

UP - Upsampling

PCC - Pearson Correlation Coefficient

NL - Non Local

GT - Ground Truth

HA - Holistic Aggregation

EA - Early Aggregation

BU - Bottom-up

FGA - Factorized Gated Attention

SSF - A model for Saliency Object Detection

UCNet - A model for Saliency Object Detection

JLDCF - A model for Saliency Object Detection

NJUD - A dataset for Saliency Object Detection

NLPR - A dataset for Saliency Object Detection

SSD - A dataset for Saliency Object Detection

RGBD135 - A dataset for Saliency Object Detection

STEREO - A dataset for Saliency Object Detection

DUT-RGBD - A dataset for Saliency Object Detection

A2dele - A model for Saliency Object Detection

SOTA - State-of-the-art

Amulet - A model for Saliency Object Detection

DSS - A model for Saliency Object Detection

BMP - A model for Saliency Object Detection

PiCANet - A model for Saliency Object Detection

R3Net - A model for Saliency Object Detection

CPD - A model for Saliency Object Detection

EGNet - A model for Saliency Object Detection

PoolNet - A model for Saliency Object Detection

BASNet - A model for Saliency Object Detection

MINet - A model for Saliency Object Detection

ITSD - A model for Saliency Object Detection

DF - A model for Saliency Object Detection

AFNet - A model for Saliency Object Detection

CTMF - A model for Saliency Object Detection

MMCI - A model for Saliency Object Detection

PCF - A model for Saliency Object Detection

TANet - A model for Saliency Object Detection

CPFP - A model for Saliency Object Detection

DMRA - A model for Saliency Object Detection

S^2MA - A model for Saliency Object Detection

RGB-D - RGB and Depth

LFSD - A dataset for Saliency Object Detection

maxF - A performance index for Saliency Object Detection

STERE - A dataset for Saliency Object Detection

SIP - A dataset for Saliency Object Detection

SSD - A dataset for Saliency Object Detection

MAE - A performance index for Saliency Object Detection

PR Curve - A performance index for Saliency Object Detection

Chapter 1

Introduction

Low light images have an important impact on daily life and production. For example, in traffic safety, when vehicle cameras or surveillance equipment cannot detect inconspicuous objects at night, they pose a threat because the low visibility environment can largely mislead drivers, making them unable to understand road conditions and prone to accidents. In addition, the core components of most current imaging devices have little ability to process ambient lighting in dark environments and cannot accurately distinguish images due to noise. All of these disadvantages can be attributed to the interference of various conditions in low light environments. Some scientists and researchers have proposed a number of enhancement solutions based on traditional digital image processing in the study of low light images.

We can process low light images by a contextual correspondence learning theory approach. All of these phenomena require a complete framework of computer vision process to accomplish the task. Among them, the processing of low light environments is one of the most difficult problems. Low light scenes represent a range of image types such as images and videos of environments at dusk and near darkness, which may be seen in many applications. I including night-time security surveillance, license plate recognition, special target recognition at dusk, and recognition of other emergencies in low light scenes. The combination of environmental enhancement with other aspects of computer vision and pattern recognition can lead to many results, such as object detection and face detection in low light levels, and anomaly detection in crowded places in low light environments. For the enhance-

ment of low light and low light scenes, the use of traditional methods often results in overexposure and halo conditions, so these specific drawbacks can be fixed and improved using deep learning network techniques.

Compared to the daytime scene reconstruction, the night reconstruction is more difficult. The main challenges faced mainly include inadequate lighting and the low image contrast. Because of low brightness, image details are not visible, image quality decreases, the scene cannot be seen clearly, and the amount of visual information is low. Due to the impact of insufficient light at night for the visual field of view, it will cause the phenomenon of information loss in the process of light propagation, affecting the final imaging and resulting in the distortion of target boundaries, blurred boundaries between adjacent target objects and other problems. The imaging ability of the light-sensing device cannot contain more effective information in its limited working range under the dual effect of its own performance and external environment.

Above all, low light imaging technology has several major difficulties. The first one is that the traditional algorithms and technologies in enhancing low light images may appear with the halo effect, scene blur, and a series of defect problems. The second is that, the original image acquisition is always a pending problem. There is always the problem of obtaining a large amount of data and maintaining the stability of the image features. Third, for the objects that need to be detected and identified in low light environment, how to identify them based on a more refined approach, i.e., an approach for saliency detection.

Although research on image saliency is in full swing, its application is mainly for high-contrast environments such as daylight, and few research has been conducted on low light environments with insufficient light. In conditions such as darkness, both human visual recognition and machine recognition are difficult and cannot

accurately locate salient targets. This is also the current problem of salient target detection in night scenes. This problem can be solved by using the "enhance first, detect later" model.

1.1 Related Works

Humans can quickly perceive visually saliency targets from complex scenes such as night images, according to the visual attention mechanism. In practical applications, limited computing resources can be prioritized for processing the scene using this selective visual attention mechanism. The scene's saliency information is focused, and the scene's non-saliency information is selectively ignored. The visual saliency object detection model is now widely used in image segmentation, object recognition, image retrieval, and other applications. However, the majority of existing visual saliency object detection models are only suitable for visible light environments, which will present significant challenges at night. Because of the low signal-to-noise ratio and low contrast characteristics of the night-time images, the feature contrast measurement will be easily disturbed by noise, weak texture blur, and other factors. It is difficult to detect salient objects in night images due to the influence of various factors such as scene background changes and camera movement. It is difficult to accurately describe the salient information in night images using the visual saliency model proposed in recent years. Using preprocessing methods such as a self-square image transformation to improve image contrast and obtain a rich gray-value image can achieve a certain effect. However, self-square image equalization transforms all pixels in the image, including the image's edge pixels, smooth and non-noise pixels, noisy pixels, and so on, and they can easily lead to problems such as false edges and overexposed pixels. In this section, we try to give an overall view of the current research progress on Low Light Image Enhancement, Saliency Object Detection and Low light Saliency Object Detection.

1.1.1 Introduction to Low Light Image Enhancement

Many traditional methods have been proposed for enhancing low light images. A fast image enhancement method combined with a color space fusion was proposed by Xiao *et.al.* [115]. Zhu *et al.* [140] developed a method of cloud removal while local contrast is preserved. A method for the correction of colors and over-spectral images was developed by Artem *et al.* [75]. Although these traditional methods may achieve good results under certain conditions, they are not generally sufficiently robust. For reflection and light decomposition, no effective dependence mapping is established.

Researchers have developed a number of methods for improving the image quality and visibility based on cognitive perception principles of humans. The benefits of guided filtering and adaptive histogram equalization (e.g., Wang [110]) significantly improved scenic visibility and colored contrast. A tone mapping function in the images was learned to deal with the low light scenario in [56].

Unlike traditional linear and non-linear methods that only enhance specific image types, Retinex can balance compression, edges improvement, and color constancy with dynamic range. The results show that different types of images can be adjusted. Among the existing low light image enhancement methods, one group of works are built on cognitive perception theory, and they include MSR [39], LIME [27] and LECARM [124], which assume that the observed color image can be decomposed into reflectance and illumination. The initial Retinex model as a low light image supervision method has been determined in [44] to be equivalent to an advanced neural network with several Gaussian convolution kernels. The authors of [27] proposed a method to preserve the naturalness of illumination with lightness-order-error measure. Jiang *et al.* [39]proposed to use all useful information of the down-sampled paths to produce high-resolution enhancement results.

1.1.2 Introduction to Saliency Object Detection

The existing RGB saliency detection models usually extract low-level image features and then leverage the contrast mechanism [63, 11], background prior [121, 91], or objectness prior [30, 3] to detect salient objects. Recently, numerous investigating works have presented CNNs into the saliency detection field and have accomplished exceptionally promising outcomes. Most of these strategies straightforwardly fathom the saliency detection issue utilizing end-to-end CNNs. For instance, early models [96, 49, 135] generally use multi-scale CNNs, to extract Multiple Scalability features from the multiple local and global patches for each pixel and superpixel. The full convolutionary network (FCN) architecture [70] is used for all individual pixels simultaneously. For all individual pixels, simultaneously, subsequent models are implemented in the full convolutionary network (FCN) [70] architecture. Typically, an encoder and decoder model was a trend, to initially extract multi-level deep features with pretrained parameters like VGG or ResNet, to build a decoder that fuses those multi-level detection features. A certain set of projects is to progressively fuse multi-level features by using the [61, 72, 132] architecture, and another set of works [55, 33] adopt the HED network architecture [117] to fuse the features simultaneously. These methods all derive directly from the image output extracted from deep characteristics, without taking other knowledge into account.

A certain information can be added to improve the accuracy of saliency detection. Li *et al.* [52] introduced the semantic segmentation task to enhance the feature capability for object perception. Wang *et al.* [104] used eye fixation to guide the detection of salient objects. In [130], Zhang *et al.* leveraged image captioning to help capture semantic information of salient objects in visual scenes. Recently, many deep saliency models [53, 106, 25, 60, 134] have been proposed to simultaneously predict object contours and use the contour prior to enhance the object boundaries for salient objects. Neither of these works, however, has explored the possibility to

improve the output detection with depth awareness. In this paper, we propose to estimate the depth of each image and use the depth characteristics to add the RGB functions for the identification of salient.

In [116], Xiao *et al.* proposed to derive pseudo depths from the RGB images, and then leveraged the pseudo depths to boost the performance of RGB saliency models by computing depth-driven background priors and depth contrast features. In two respects, our approach varies greatly from theirs. Firstly, their model relies on conventional saliency features and structures, while our model is a profound saliency model that is much more efficient and speedy. Second, its model should first draw the pseudo-depth map and use this to calculate the depth-based function and the previous map, while it should be more efficient and effective to use intermediate depth features to enhance the RGB properties before the depth map is produced.

Based on the night vision enhancement technique, saliency object detection has also investigated a new way to explore the state-of-the-art performance without using depth. The identification of important objectives in night scenes is also the problem and subject of current vision science. If the previous approach is executed directly, the output results in the blurring effect. Therefore, we are taking a new mechanism that first enhances the picture of low light to increase contrast and visibility and then uses the object detection saliency method.

Most SOD models [11, 96, 61, 104, 134, 106, 65] typically detect salient objects from RGB images. In a pioneer work of [77], Ouerhani and Hugli have shown that depth can also provide a valuable evidence and increase the saliency detection efficiency in a large part. This is intuitive since humans reside in a genuine 3D world and have a significant influence on our sensory experience. Many subsequent models for output, like those in [45, 13, 12, 100], began to take RGB-D images for the detection of output. Recently, Convolutionary Neural networks (CNNs),

which also showed outstanding results in numerous computer vision activities, have been commonly seen in the computer vision community. Many works have also implemented two-stream CNNs for RGB-D SOD in order to leverage their efficient functionality in learning.

Besides SOD methods without using depth information, another UNet model principle gives a unique way to combine more information to get much more precise results.

Most of the other projects [5, 6, 98, 138], considering the multi-level function maps randomly acquired by CNNs, have taken on the two-way UNet [85] architecture to add multi-level features to RGB-D SOD. The two-stream UNet uses first two encoder networks to collect images from multi-levels in a bottom-up way. Then, one or two decoder networks are installed in the top-down processing and at the same time combining high level features with low-level ones. The characteristics of its symmetrical encoder module appears in any decoder module. As such, the top-down proliferation of discriminative semantic data in deep layers allows the precise localization and results in precision shapes and boundary segmentation with local structures in smaller layers.

UNet only aggregates top-down features once. Low-level details can only be added to enhance the depictive ability in the decoder, while the high-level functionality cannot be changed themselves. In this thesis, we suggest adding another bottom-up aggregation path to solve the problem by propagating enhanced low-level features from the top-down path to high-level layers again. If the combined functionality of both top-down and bottom-up is cascaded, the functions can be enhanced progressively at all stages.

Another challenge is to add functions on all two-neighboring levels over networks only progressively. Although it prevents big changes and is commonly used in previ-

ous articles, this feature aggregation system reduces the direct connections between multi-level features. We suggest systematic paths to aggregate multi-level features holistically after the bottom-up and top-down processing in order to mitigate this issue. The network will then benefit from them all at once, providing a rich cross-level functional convergence mechanism for SOD.

Given the two-stream architecture, it is typical that the writers of current work adopt only two-stream encoders independently and aggregate the function only in a [29, 138, 17] decoding method. It may otherwise reuse crossmodal encoder functionalities [137, 48] in decoders, without enhancing any other encoder feature. They use pretrained CNN models as encoders, requiring their configurations and pre-trained parameters to be maintained. In this study, we present the crucial parts, including the encoding process, the aggregation and propagation of cross-modal functions. We use a residual learning aggregation system to add cross-modal encoder features and propagate them back to the encoder's original route, improving the functionality right from the start.

1.1.3 Introduction to Low light Saliency Object Detection

Although research on image saliency is in full swing, its application is mainly for high-contrast environments such as daytime, and few research has been conducted on image saliency detection in narrow dynamic regions represented by low light environments with insufficient light.

There are some studies [118] trying to solve the low-illumination problem of saliency detection by doing low-illumination SOD directly on the original degraded images, and hence turning out to be inefficient. Researchers [118] also propose to extract non-local features in low-illumination images for SOD, since the low-illumination effect of images taken in uniformly illuminated environments tends to degrade with changes in the scene depth and the associated ambient lighting. Due

to this degradation, a large amount of object information is lost, making saliency detection in such scenes more challenging.

The deep learning-based methods trained using the dataset day-time images do not work properly under the night-time environments. This is because low light images contain flares, dense noise, glow/glare, etc., which are not present in the daytime training data. A possible solution is to train the network on low light images in a fully supervised manner. However, it is not easy to obtain disparity base data for low light images.

In conditions such as darkness, both human visual recognition and machine recognition struggle to accurately locate salient targets. This is the current problem of salient target detection in night scenes. This problem can be solvable by using the ”enhance first, detect later” model.

1.2 Research Objectives

Low light situations provide a diverse range of image types with enormous potentials. Their investigation is centered on photographs and recordings of the outdoors around dusk and near night. It has a wide range of applications, including night safety monitoring, licence plate identification, night scene photography, unique target recognition at dusk, and recognition of other emergency occurrences in low light situations. When the environment is upgraded and integrated with other tasks in computer vision and pattern recognition, it may provide a variety of outcomes, including saliency detection and object detection in low light environments, as well as anomalous detection in crowded areas. Due to the rising needs of severe visual tasks in many applications, low light image enhancement is quickly attracting interest from research community.

The comprehensive perceptual processing of low light images represented by

night-time is primarily concerned with accurate target detection and recognition. However, since low illumination results in sparse visual information, the visual range of night-time images is significantly reduced compared to those captured in daytime with ambient natural light. The lack of illumination can result in that the recognition person and recognition systems have a natural blind area (area of interest). Using some anti-interference technology and enhancement technology can effectively suppress the interference of background noise.

When high light images produce degradation, it can damage the statistic properties and structural information of image pixels, and the dynamic range will become narrow, leading to feature drift of the network. By superimposing them together, the errors caused by feature drift may slow down the vision system. Using actual degraded datasets with manual annotations is a straightforward approach to improving the robustness and accuracy of computer vision systems in real-world applications. As a result, it is difficult to collect large-scale degraded datasets with semantic annotations, and it is more challenging to comprehensively cover all forms of degradation. Annotating corrupted images is also very costly, making it difficult to for this approach to be scaled up and adopted practically.

In addition, human can quickly perceive visually salient targets from complex scenes, according to the visual attention mechanism. In practical applications, this selective visual attention mechanism allows us to prioritize limited computational resources for processing the salient information of the scene and then selectively ignore the non-salient information of the scene. Visual saliency object detection models are now widely used for image segmentation, object recognition, image retrieval and other applications. However, saliency detection in extreme environments, such as low light scenes at night, is still a challenging topic.

As a result of night pictures' poor signal-to-noise ratio and low contrast, feature

contrast measurements are required to be performed. However, it is sensitive to noise interference, poor texture blurry backdrop and camera movement, among other variables, which makes it difficult to discern prominent items in night-time images. However, it is challenging for the visual saliency model that has been developed in the last several years to correctly characterize the saliency in night images. For example, histogram equalization transforms all pixels in an image, including edge pixels and noisy pixels, and hence can easily cause issues such as false edges and overexposure.

The aims of the research are to:
1) conduct research on low light image enhancement using CNN-based methods and 2) conduct research on saliency object detection from the enhanced low light images using estimated depth information and UNet-based feature aggregation.

1.3 Main Contributions

Aiming to accomplish the above objectives, in this project the following four algorithms are developed for low light image enhancement and saliency object detection from the enhanced low light images.

1) We proposed a deep learning dedarking network based on the cognitive perception model of Retinex theory, which effectively combines the inception network with high-level semantic information of the foreground and the background of the images.

2) Towards saliency object detection on the enhanced low light images, we propose to simultaneously estimate the depth and detect saliency for RGB images in a unified deep CNN. Intermediate depth features can be fused with RGB saliency features to supply complementary information for improving the saliency detection performance. We further propose to fuse multiscale depth and RGB features and

also introduced global contexts.

3) Last but not the least, we reconsider the feature aggregation schemes for deep RGB-D SOD and propose novel feature aggregation methods. Based on the widely used two-stream UNet architecture, we first propose to add early aggregation and holistic aggregation paths to propagate cross-modal information in an early stage and learn abundant feature interactions among all multi-level features. Then, we propose to cascade the top-down decoder network in UNet with a bottom-up decoder network, thus enabling to improve the high-level features with the already improved low-level features. Furthermore, we propose a factorized gated attention model to modulate the feature aggregation actions for each feature node with reduced computational costs and boosted model performance.

Qualitative and quantitative experimental comparisons conducted on the benchmark ExDARK dataset [69] demonstrate that our approaches have improved the quality of the dedarking images and overcome the drawbacks of white and color distortion that are shown in current state-of-the-art techniques.

1.4 Thesis Organization

This rest of the thesis is organized as follows in the following sub-sections:

1.4.1 Chapter 2

As our first attempt, in this chapter, we propose a deep semantic brightening network based on the cognitive perception model of Retinex theory to combine the inception network with the high-level semantic information about foreground and background effectively. In order to train a model towards this goal, we introduce the structure loss and perceptual loss so as to integrate the high-level semantic information to improve the enhancement result. Qualitative and quantitative comparison experiments conducted on benchmark datasets demonstrate the superior

performance of our approach, which overcomes the drawbacks of white and color distortion with existing techniques.

1.4.2 Chapter 3

For the low light image enhancement, we can use the UNet based architecture for this research with multiple input information.

UNet based architectures are widely used in deep RGB-D salient object detection models. However, UNet only adopts a top-down decoder network to progressively aggregate high-level features with low-level ones. In this chapter, we propose to enrich feature aggregation via holistic aggregation paths and an extra bottom-up decoder network. The former aggregates multi-level features holistically to learn abundant feature interactions while the latter aggregates improved low-level features with high-level features, thus promoting their representation ability. We also propose a factorized attention module to efficiently modulate the feature aggregation action for each feature node with multiple learned attention factors. Experimental results on seven widely used benchmark datasets demonstrate that all of the proposed components can gradually improve RGB-D salient object detection results. Consequently, our final saliency model performs favorably against other state-of-the-art methods.

1.4.3 Chapter 4

Although many RGB-D saliency models have been proposed, they require to get depth data, which is expensive and not easy to get. For instance, it is very practical for the dealing of the low light images with less depth information.

In this chapter, we propose to estimate depth information from monocular RGB images and leverage the intermediate depth features to enhance the saliency detection performance in a deep neural network framework. Specifically, we first use an

encoder network to extract common features from each RGB image and then build two decoder networks for depth estimation and saliency detection, respectively.

The depth decoder features can be fused with the RGB saliency features to enhance their capability. Furthermore, we also propose a novel dense multiscale fusion model to densely fuse multiscale depth and RGB features based on the dense ASPP model.

A new global context branch is also added to boost the multiscale features. Experimental results demonstrate that the added depth cues and the proposed fusion model can both improve the saliency detection performance. Finally, our model not only outperforms state-of-the-art RGB saliency models, but also achieves comparable results compared with state-of-the-art RGB-D saliency models.

Chapter 2

Low Light Image Enhancement via CNN-based Models

Although there are various ways for improving low light image quality, it is still unclear how to balance human observation perception with computer vision processing. The existing solutions can result in over-exposure and a halo effect. As our first attempt, in this chapter, we propose a deep semantic brightening network based on the cognitive perception model of Retinex theory to combine the inception network with the high-level semantic information about foreground and background effectively.

2.1 Introduction

Many traditional methods have been proposed for enhancing low light images. Xiao *et. al.* [115] put forward a rapid image enhancement approach combining the space fusion of color. Zhu *et al.* [140] developed a method for removing cloud while preserving local contrast. Li *et al.* [50] developed an on-line detection method based on a single scale retinex. Lee *et al.* [47] proposed a new method to satisfy the multi-scale morphology. Artem *et al.* [75] designed a method to correct color and hyper-spectral images. Zohair *et al.* [1] developed a new method to use a single-scale method to meet the needs of the image enhancement. Although these traditional methods can produce some good results under specific conditions, they are generally not robust enough. For the decomposition of reflectance and illumination, no effective dependence diagram was established.

Based on the cognitive perception principles of humans, researchers have developed a series of methods to enhance the image quality and visibility. For instance, Wang [110] took the advantages of guided filtering and adaptive histogram equalization, and greatly enhanced the visibility and color contrast of the scenes. In [56], a tone mapping function of the images was learned to deal with the low light scenario. Michael [90] created a model, which could effectively conduct object detection and recognition, especially in dark scenes using low light enhancement methods. In [124] and [123], fusion networks, which represented the light reflection model principles, were proposed.

Different from the traditional linear and nonlinear methods that only enhance specific types of images, the Retinex model can balance dynamic range compression, edge enhancement, and color constancy. The results show that it can be adapted to enhance different types of images. In [44], it was concluded that the Retinex model as a low light image supervision method was equivalent to a feed-forward neural network with different Gaussian convolution kernels.

However, the basic Retinex model cannot establish an effective mapping between the R, G, and B channels in the RGB color space. As a result, the enhancement level of each color channel is inconsistent, resulting in color distortion and loss of edge information. This problem is more pronounced when the image is rich in depth-of-field information, and is especially prominent in low light images.

Concerned with low light image enhancement, conventional methods such as MSR [40], LECARM [126] and LIME [28] have been proposed, but they can result in over-exposure and halo effects.

With the development of deep neural networks, researchers have applied convolutional neural networks to many areas of image processing, including image restoration, image super-resolution, and image denoising. Some of the methods have been

proposed for low light image enhancement. These methods can help with the information loss during the imaging process, preserve edge information and enhance color channels while reducing color distortion. For example, LLnet [71] used a stacked sparse denoising auto-encoder for simultaneous low light enhancement and noise reduction. Tao *et al.* [93] proposed an MSR-based image enhancement approach with modified luminance for fast and efficient processing. Lore *et al.* [71] put forward a deep auto-encoder approach for natural low light image enhancement. Lv *et al.* [73] proposed MBLLEN for low light image/video enhancement using CNNs. Experiments show that these methods towards the de-darking problems have better performance than the traditional methods.

Unlike the existing de-darking methods based on convolutional neural networks, which often use pooling operations, our method uses an efficient method to extract high-level and low-level multi-scale semantic information about objects and backgrounds from the scene. With this information, our method is more adaptive to the brightening sub-tasks, *i.e.*, feature extraction, inception and feature fusion. Moreover, different from the existing widely used loss functions, we introduce three loss functions to achieve the above goal and design a joint loss to focus on the semantic information. Last but not the least, when it comes to performance evaluation, existing research works mostly use PSNR and MMSE (Minimum Mean Square Error), which require paired images of both the dark images and their ground truth. However, for many real-world applications, the ground truth images are not easy to acquire. Thus, Naturalness Image Quality Evaluator (NIQE), which does not require paired daytime images, becomes a more objective evaluation metric and can produce more convincing results for low light image enhancement tasks.

The main contributions made in this chapter are as follows. 1) We introduce the Perceptual Loss function into the joint loss function to effectively fuse high-level semantic information and low-level image information, and adding perceptual

constraints to reduce the training time. 2) We propose a new method that treats the front and back scenes at different processing levels, so that it can eliminate background interference and focus on enhancing the region-of-interests areas of the foreground. 3) We adopt the NIQE metric to evaluate the algorithm performance and compare it with the state-of-the-arts. Our proposed approach has achieved better performance on the ExDARK datasets [69] tested.

2.2 Cognitive Perception Retinex Theory

In this section, we present the related Retinex theory and its mathematical relation with CNN.

The Retinex theory was proposed by Land in the 20th century and it has been widely applied in the image processing field. The Retinex theory states that people can perceive the color component that reflects the incident component.

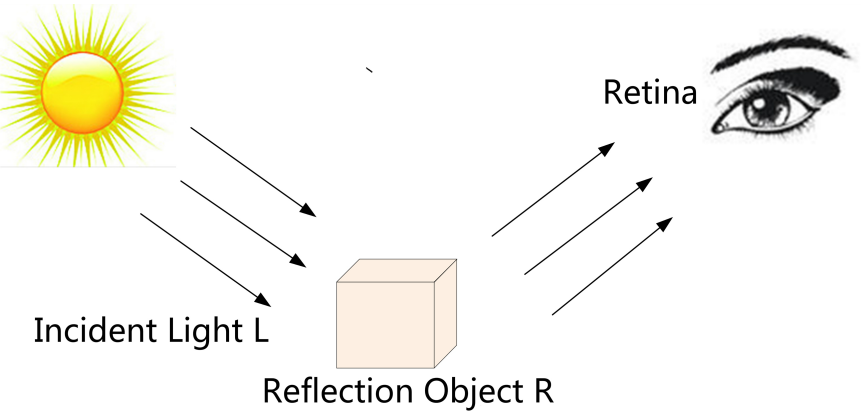


Figure 2.1 : The Retinex model

The Retinex model is based on the consistency of color perception (color constancy). Its basic theory is that the color of an object is not affected by the illumination's non-uniformity, but is determined by the ability of the object to reflect light from long waves (red), medium waves (green) and short waves (blue), rather

than the absolute value of the intensity of the reflected light.

Researchers have taken several different image estimation methods to develop various Retinex based algorithms, such as Single Scale Retinex (SSR), Multi-Scale Retinex (MSR), and Multi-Scale Retinex using Color Restoration (MSRCR). SSR constrains the illumination map to be smooth by a Gaussian filter. MSRCR extends SSR with multi-scale Gaussian filters and color restoration. The main idea of MSR is to choose different Gaussian surrounding scales to calculate based on the SSR algorithm [101][93], and then generate the output results. MSR can be described as:

$$\log R_i(x,y) = \sum_{i=1}^{k} w_k\{\log S_i(x,y) - \log[F_k(x,y) * S_i(x,y)]\}, \tag{2.1}$$

where (x,y) is the coordinate of the pixel, i is the color channel, $i \in \{R,G,B\}$, $S(x,y)$ is the original image, $R(x,y)$ is the reflection component, $F_k(x,y)$ is the Gaussian surround function, k is the number of scales, and w_k is the value of the Gaussian surround function such that $\sum_{k=1}^{K} w_k = 1$ and K is the number of the scales.

A large number of experimental tests show that there is a standard to set the number of the scales K, from a small, middle to a large range, dependent on the results that the actual application needs.

Multi-scale Retinex usually acts as the traditional method towards the low light images [86], but it usually has the weakness of color halos and over enhancement in local areas. With the specialty of the neural networks that can simulate the working procedure, this process can be more efficient. According to [86], the multi-scale Retinex algorithm can be regarded as a feedforward convolutional neural network with a residual structure. Therefore, our method can be regarded as a new CNN based method from this perspective.

The multi-scale Retinex model reveals that images have the potential to be

further improved, for the physical structure of the retina can be considered as a residual convolutional neural network. With the improvement of neural networks, researchers have discovered that the calculation in cognitive recognition can perform better than the previous algorithms.

The input of the network is a low light color image, and the output is the enhanced image of the same size.

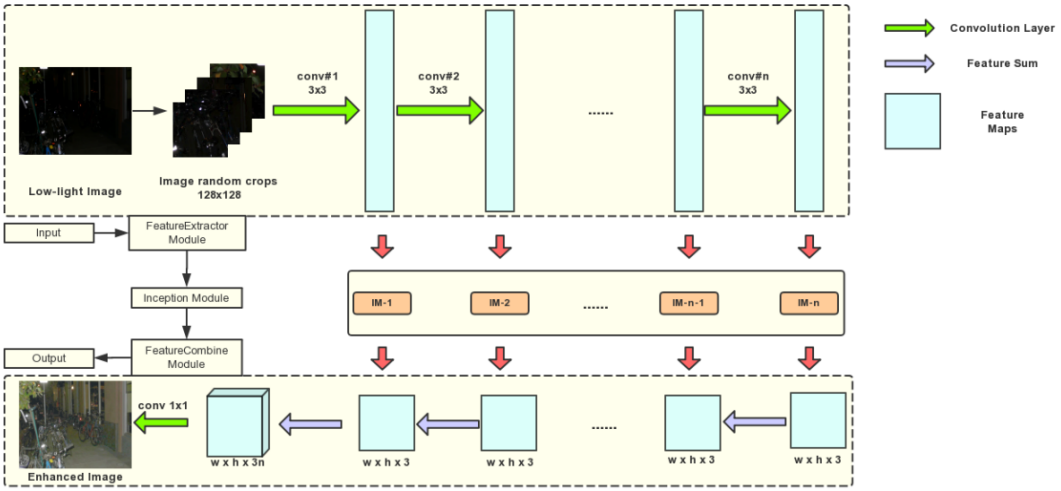


Figure 2.2 : The overall view of our proposed network architecture.

As is shown, the proposed network consists of three modules, *i.e.*, the feature extraction module, the inception module and the feature fusion module. Each of the three modules plays a different role, detailed as below.

Feature Extraction Module. This module contains 10 convolutional layers, each of which using kernels of size 3×3, a stride of 1, a padding of 1 and ReLU nonlinearity. The output of each layer is the input to both the next convolutional layer and the corresponding inception module.

Inception Module. The module is inspired by the inception module and residual learning [32]. The left is two 3×3 convolutional layers and the right one is a

1×1 convolutional layer. The function of this module is to reconstruct the features and form the enhanced image that matches the size of input image. In addition, we obtain features from each feature extractor module at various positions, which contain image information of different scales at both low and high levels, and combine such information to produce the final result.

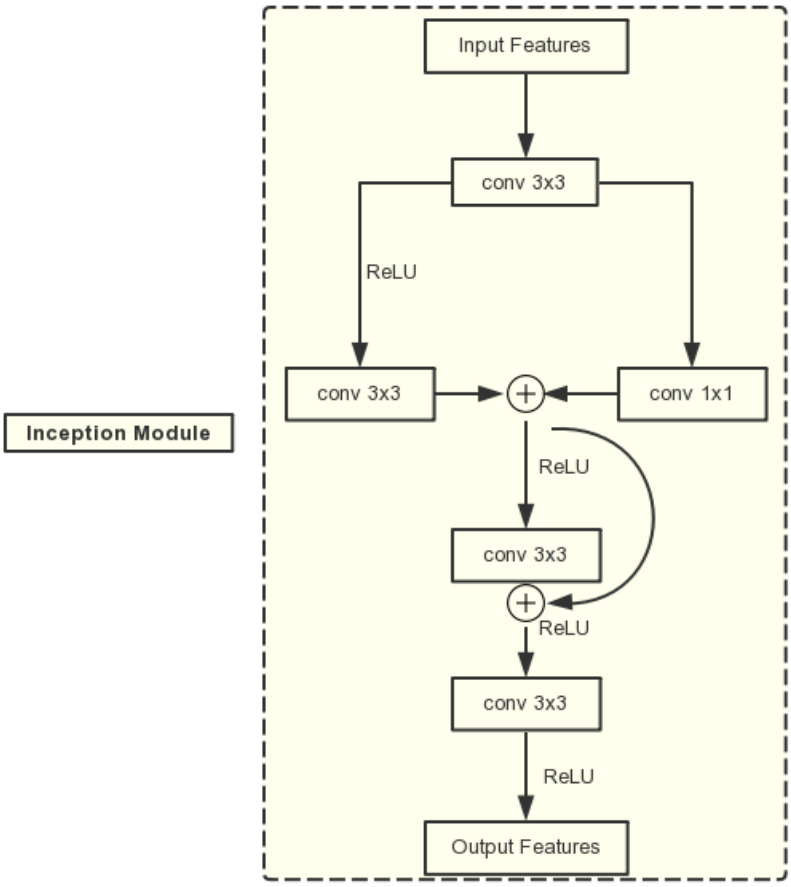


Figure 2.3 : The Inception Module in our proposed network

Feature Fusion Module. The input of this module is the combined features from each inception module. In this chapter, we concatenate all of the outputs from the inception module and use a 1×1 convolution kernel to reduce the feature dimension from $w \times h \times 3 \times n$ to $w \times h \times 3$ and obtain the enhanced color image.

2.3 Loss Functions

In order to guide the model, instead of using the common error metrics such as MSE as the loss functions, we introduce the Structure Loss and Perceptual Loss so as to combine the high-level semantic information to improve the enhancement result.

Our joint loss contains three types of loss functions in total, *i.e.*, Similar Structure Loss (SSIM) (denoted as L_{SSIM}), Region Loss (denoted as L_{reg}) and Perceptual Loss (denoted as L_{per}), each focusing on a different aspect, as detailed below.

Similar Structure Loss. This loss function is designed to facilitate the quality of the image enhancement through the network processing. This chapter aims at choosing SSIM loss to compute the model loss as well as the final evaluation metric. Low light images with structure distortion tend to yield color blur, color distortion, and foreground and background mixture. We can directly minimize these drawbacks through SSIM loss.

In this chapter, we introduce a structure loss to measure the difference between the enhanced image and the ground-truth image. A simplified form of SSIM is used to compute for a pixel p by:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)^2 + (\sigma_x^2 + \sigma_y^2 + C_2)^2}, \tag{2.2}$$

where μ and σ represent the mean and variance of the input image, C_1 and C_2 are two constants used to avoid division by zero, and the subscripts x and y represent input images.

Thus, the SSIM Loss can be defined by:

$$L_{SSIM} = -\frac{1}{N}\sum_{p} SSIM(x,y), \tag{2.3}$$

where p is the pixels.

Region Loss. As shown in [47], the low light enhancement has been regarded as a significant issue, which needs to be dealt with in computer vision field. In human's usual cognitive processes, human tends to pay much more attention to the brighter regions and often ignore, or pay less attention, to the low light regions. This chapter introduces the region loss, which pays attention to the MSE of the low light areas so as to improve the enhancement of low light regions.

The region loss function is extended from MSE and is defined by:

$$L_{reg} = \beta_l \frac{1}{w_l h_l} MSE(RL_l, RG_l) + \frac{1}{w_{all} h_{all}} MSE(RL_{all}, RG_{all}), \tag{2.4}$$

where β_l is the parameter assigned to the low light region, and w and h represent the width and height of the features, respectively. In Eq. 2.4, the subscript l means the low light area and all means all-pixel regions while RL and RG mean the recovered images and ground truth, respectively. RL_l and RG_l describe the low light part of the image being processed and RL_{all} and RG_{all} represent all-pixel regions of the reference sample. Note that the latter part of the region loss is the standard MSE.

Perceptual Loss. Similar to those low-level vision approaches, such as image denoising and image reconstruction [41], researchers always adopt MSE and SSIM [109] as the metrics to measure the quality of the output results. Besides, there are many relevant articles [71, 73, 18, 69] about combining different levels of information to improve the performance of their approaches for image denoising and super reconstruction. It is also essential to combine high-level information to solve this issue.

In this chapter, we adopt the Perceptual Loss [41], which employs a content extractor. In other words, if the enhanced image is the output, the features extracted should be similar to those of the ground truth. In particular, this chapter considers the perceptual loss based on the output of the different ReLU layers of the VGG-16

network as:

$$L_{per} = \frac{1}{C_j H_j W_j} ||\phi_j(x') - \phi_j(x)||_2^2, \tag{2.5}$$

where C_j is a constant, W_j and H_j represent the width and height of the feature vector, ϕ_j represents the j-th convolutional layer, and $\phi_j(x')$ and $\phi_j(x)$ are the features generated from the j-th layer.

Without loss of generality, the final loss can be computed by:

$$Loss = L_{SSIM} + L_{per} + L_{reg}. \tag{2.6}$$

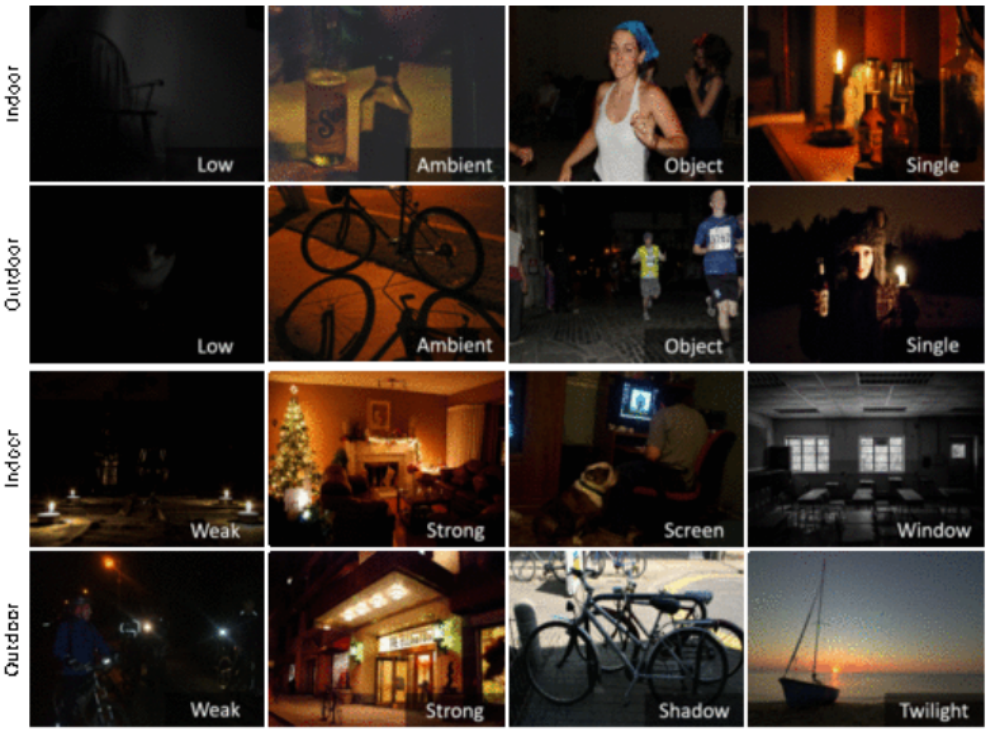


Figure 2.4 : Example indoor and outdoor images of the ExDARK dataset [69]

2.4 Experiments

In this section, we report the experimental results on benchmark datasets to validate the effectiveness of our proposed model.

2.4.1 Dataset

In order to establish a baseline performance and facilitate the new direction of research for object detection and image enhancement, we evaluate our proposed method on the newly published Exclusively Dark (ExDARK) dataset [18, 69]. Previously, there are several small datasets including LIME [28], NPE [99] and DCIM [46] in the de-darking field. The sizes of these datasets are not appropriate for further research. Compared with these datasets, ExDARK [69] has advantages in both the quantity and quality of the data in the dataset.

The ExDARK dataset [69] is the largest collection of natural low light images taken in visible light to date, and it contains object level annotations. It has a collection of 7,363 low light images from very low light environments to twilight with 12 object classes (similar to PASCAL VOC) annotated at both image class level and local object bounding box level.

2.4.2 Implementation Details

Training of the Model. During the training stage, all convolutional layers are initialized randomly using the Gaussian distribution, which is provided from the framework, and the biases are set false. First, the initial learning rate is 0.0001 and our strategy of lowering the learning rate depends on the loss results of the current model. If the loss continues decreasing, it means that the current learning rate is appropriate and will remain unchanged in the next epoch. Otherwise, if the loss remains stable or starts increasing, we should reduce the current learning rate by half until it reaches 0.00001. In our work, the model training is completed on a PC with two NVIDIA 1080Ti GPU cards.

Data Collection. In previous work, we also test our approach on synthesized low light images so as to evaluate the PSNR results. Each artificial low-light image is randomly generated by the original image and non-linear degradation func-

tion. They are processed by the controlled and added gamma noise method. Note that, in most real-world de-darking applications, low light images often do not come with ground truth and it is difficult to obtain the corresponding ground truth. To overcome the difficulties in gathering paired data, we use the same approach as in previous research [71, 73] and synthesized low light images from the MS COCO and Pascal VOC image datasets [18, 69] for this research. As described in [28], we compare the similarities between the de-darking algorithm based on atmospheric physical scattering model and the enhanced algorithm based on Retinex theory, and make a series of transformations towards the original images. Then, we apply a random gamma adjustment to each channel of the common images to produce the low light images. Moreover, in order to simulate low light images in the real scenes and improve the model's robustness, we add random level Gaussian noise to low light images.

2.4.3 Results and Analysis

Our proposed method is evaluated and compared quantitatively and qualitatively with several existing methods, including MSR [40], LIME [28] and LECARM [126], where the published codes are used for these comparative methods. Moreover, for datasets where there is no ground truth available, we adopt NIQE for evaluation. Note that, higher values of PSNR and SSIM indicate better quality of the de-darking images, whereas for NIQE, the lower the better.

Comparing the enhancement results obtained using MSR [40], LIME [28] and LECARM [126] and the results using our proposed method, similar conclusions can be drawn. For the dark regions in testing images, the results of LIME and LECARM are still not clear enough for recognition. Apart from dark regions and highlighted areas, the most significant point that matters is the color distortion shown from the results obtained with the other three methods. Hence, enhancement results obtained

Figure 2.5 : Examples of image enhancement results obtained on the synthetic dataset with benchmark approaches and our proposed approach. (a) Ground truth. (b) low light images. (c) MSR [40] results. (d) LIME [28] results. (e) LECARM [126] results. (f) Our results.

using our approach have boosted the best rehabilitation from low light images to ground truth.

In addition to the drawbacks with the MSR approach when dealing with low light images, the results obtained using LIME and LECARM present an overall reddish background.

Figure 2.6 visually compares the image enhancement results with the results of non-reference image enhancement approaches, which operate without the need of the referential image. Comparing the results obtained using the four methods comprising MSR [40], LIME [28], LECARM [126] and our proposed de-darking method, it is obvious that the MSR results shown in Figure 2.6(b) show a general enhancement to the whole image but still have a weakness in color space distortion. The LIME results are shown in Figure 2.6(c) improve the general recovery effect during the processing procedure but still exhibit certain weaknesses. The LECARM results shown in Figure 2.6(d) achieve a good enhancement result in general but still

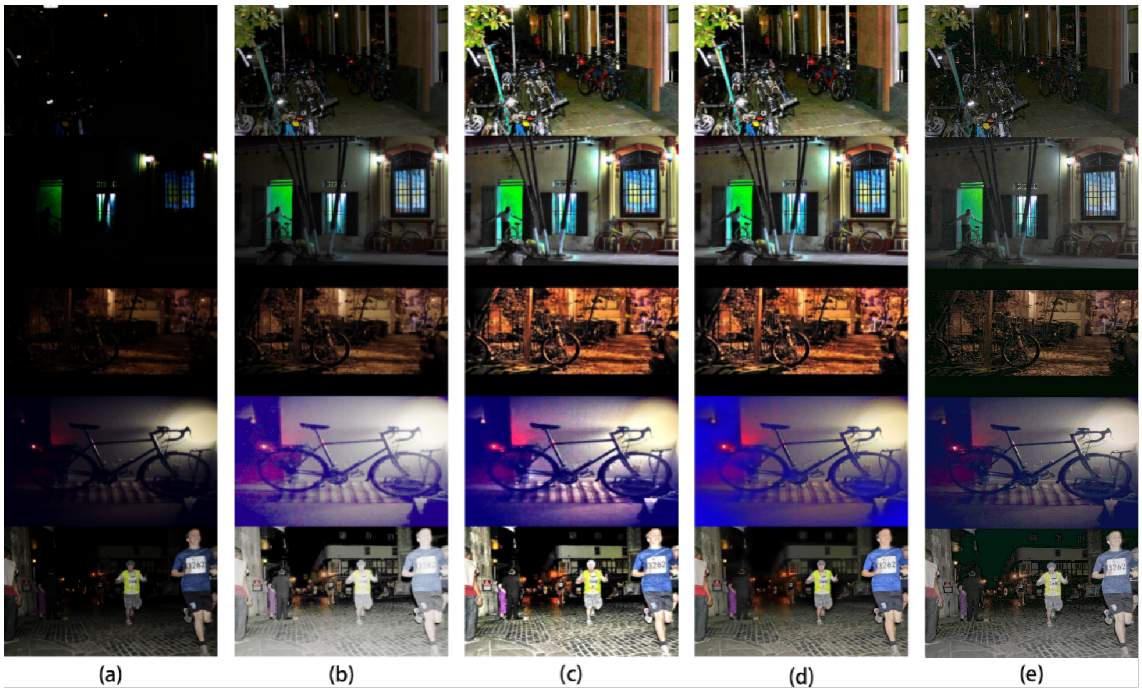


Figure 2.6 : Examples of image enhancement results obtained on natural low light dataset. (a) low light images. (b) MSR [40] results. (c) LIME [28] results. (d) LECARM [126] results. (e) Our results

have the difficulties similar to those in Figure 2.6(c). The results obtained using our proposed method (displayed in Figure 2.6(e)) show a better performance both in detail and in general.

2.4.4 Object Theme Enhancement Analysis

In contrast to other research works on image enhancement [43, 69, 57, 39] and those approaches using different channel algorithms and learning techniques [107, 94, 92], in this section, we discuss this issue from both pixel and object levels, and design several experiments to demonstrate the effectiveness of our approach for enhancing object themes. The results are shown as both a comparison chart in Figure 2.7 and visualisation results in Figure 2.9.

In this discussion, we use 12 object themes, including Bicycle, Boat, Bottle, Bus,

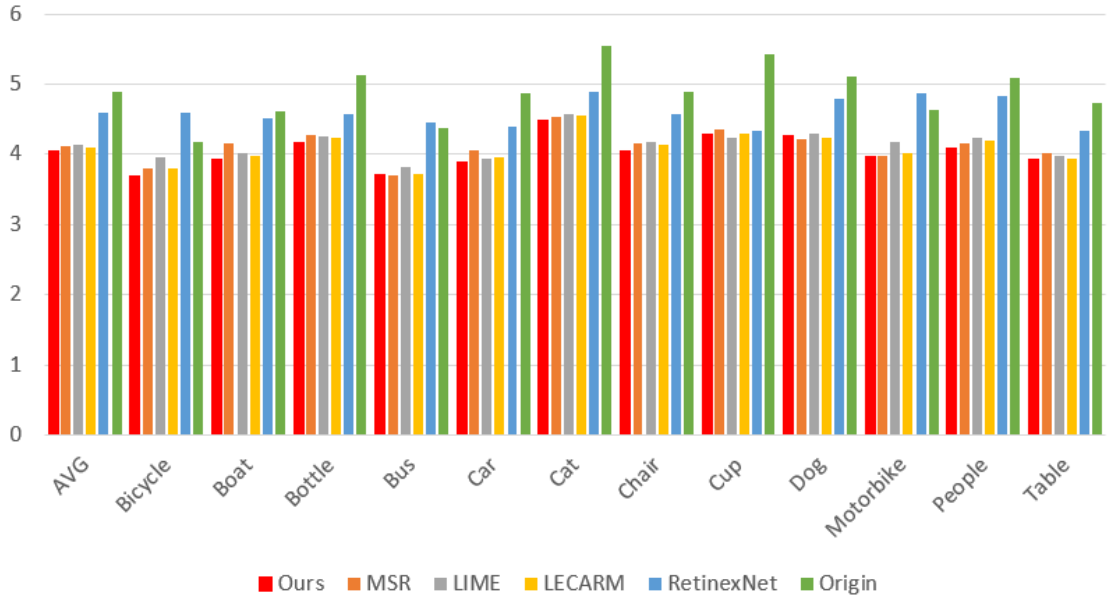


Figure 2.7 : Comparison of the NIQE results of the enhanced images obtained with benchmark and our approaches.

Chair, Cup, *etc*, from the ExDARK dataset [69], which are naturally dark scenes and have no references, to conduct the analysis. Since the ground truth daytime images are not available for these datasets, we use NIQE as the metric to evaluate the enhancement performance, the same as the above-published works. Figure 2.8 shows the results of two exemplar images from the dataset.

Please notice the difference in these results for the regions in the red bounding boxes. Row 1 shows the results of comparing the scene depth and enhancement details. Row 2 shows that the details of the background and foreground are mixed. We can clearly see that the details including the scene depth in Row 1 and the tire detail of the bike in Row 2 are clearer in our results.

Figure 2.7 shows the NIQE results obtained using the benchmark approaches and our approaches for the 12 classes. Some classes, such as Cat, Dog or People, appear in the foreground, while other classes appear in the background. As it can

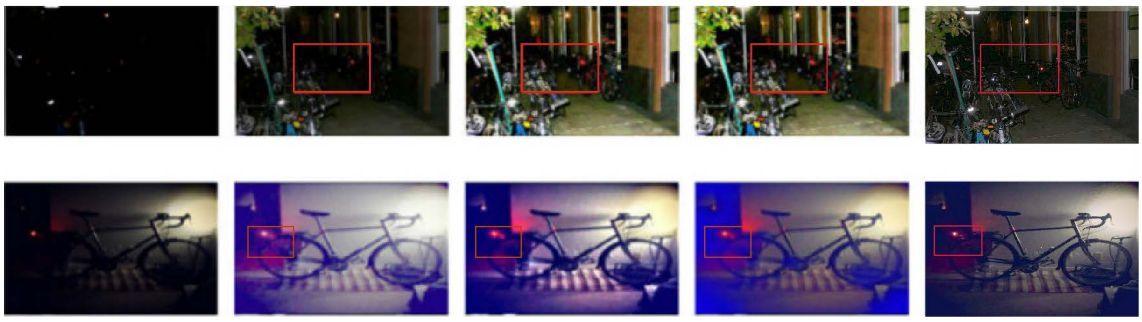


Figure 2.8 : Image enhancement results of the two example images from the Ex-DARK Dataset[69] obtained with the comparison methods. Note the details shown in the red bounding boxes. The first column represents the source images without enhancement. The second to fifth columns represent the image enhancement results obtained with MSR [40], LIME [28], LECARM [126] and our approach, respectively.

be seen from the chart that, our proposed de-darking algorithm prevails on eight classes (Bicycle, Boat, Bottle, Car, Chair, Motorbike, People, and Table) and the average NIQE score is lower than those obtained using other comparison methods. If we further investigate the results, the eight classes considered in the comparison obviously exist in the foreground and many of them have been enhanced by our method. From these results, we can conclude that, although the enhancement varies for different object classes, using the combination of pixel and object information to guide the model training is effective.

The comparable NIQE results of the 12 classes in Figure 2.9 demonstrate the advantages of our proposed method.

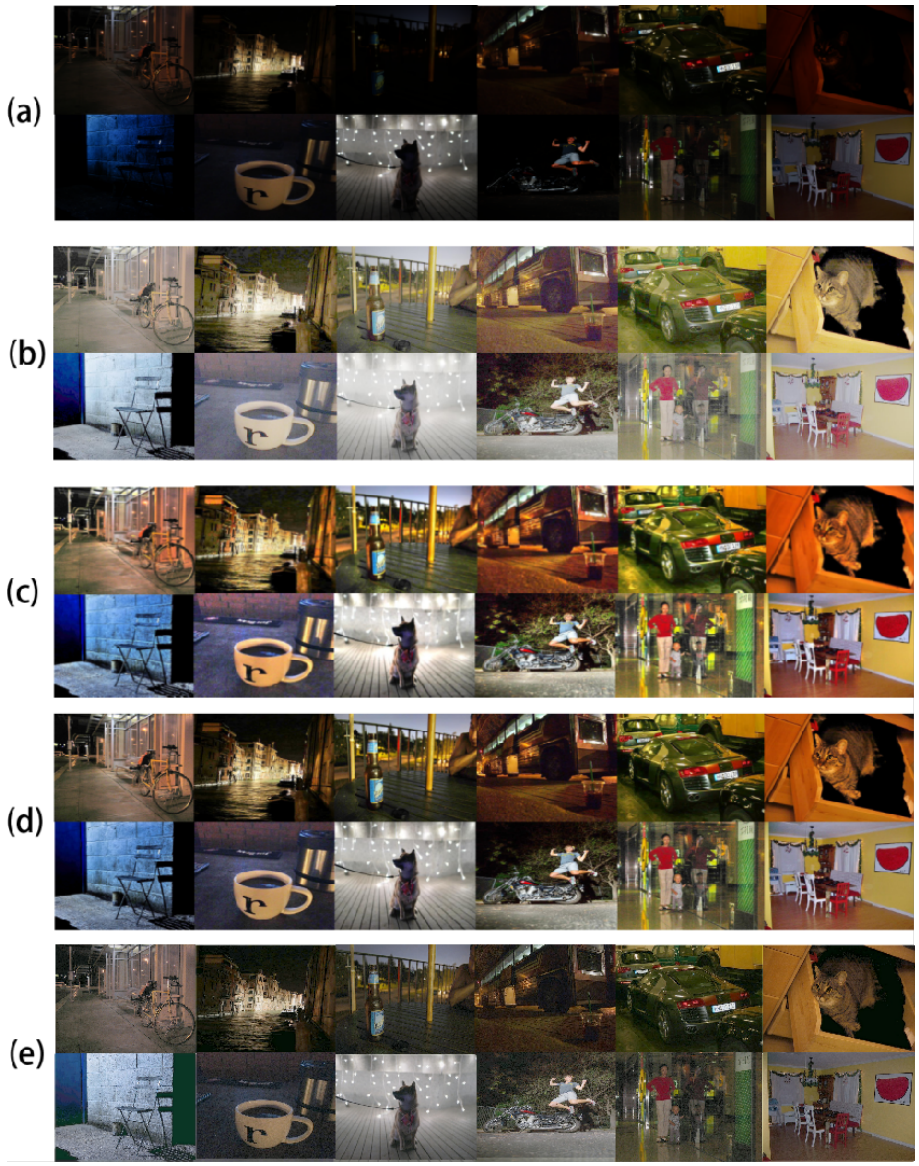


Figure 2.9 : Comparison of the dedarking results of the 12-object themes from the ExDARK Dataset [69] obtained with the four comparison benchmark methods. (a) Original images. (b) MSR [40] results. (c) LIME [28] results. (d) LECARM [126] results. (e) Our results.

2.4.5 Comparison on Other Datasets

The other approaches that we have compared with, including LIME [28], MSR [40] and LECARM [126], reported their results on other datasets in their publications.

However, the sizes of these datasets, containing less than 100 or hundreds of images in each dataset, are much smaller compared to that of ExDARK [69]. Hence, the statistical results tend to be affected by individual, extreme samples.

2.4.6 Ablation study

To understand why our model has performed so well, we conduct ablation studies to show the impact of using our proposed Perceptual Loss and SSIM Loss on the quality of the dedarked images. The ablation study was conducted on the ExDARK dataset[69] and we compared the NIQE results of the resultant images enhanced by models when different loss functions were considered in training. The results are presented in Table 2.1.

In this table, the 2nd column lists the NIQE results of the 12 classes without using our proposed Perceptual Loss L_{per} and SSIM Loss L_{SSIM}, the 3rd column lists the NIQE results without using our Perceptual Loss L_{per}, and the last column shows the results using all three losses as we proposed.

As it can be seen from this comparison, adding the SSIM Loss L_{SSIM} as we proposed has reduced the NIQE figures for all categories, and adding the Perceptual Loss L_{per} has further reduced the NIQE results for all but two categories (Car and People). This shows the effectiveness of considering our proposed loss functions for dedarking.

Table 2.1 : The NIQE results obtained on the ExDARK dataset [69] using and without using our proposed Perceptual Loss and SSIM Loss.

Items	w/o L_{per} & L_{SSIM}	w/o L_{per}	w all losses
Bicycle	3.714251	3.677626	3.616516
Boat	4.037668	3.959704	3.901504
Bottle	4.13207	4.085261	4.077717
Bus	3.746283	3.690841	3.67074
Car	3.830932	3.756289	3.770743
Cat	4.486879	4.464398	4.407415
Chair	4.006253	3.960921	3.919645
Cup	4.328764	4.262844	4.229619
Dog	4.294867	4.251049	4.240586
Motorbike	3.903596	3.883475	3.791888
People	4.12894	4.084356	4.090785
Table	3.898326	3.848223	3.812703
AVERAGE	4.042402	3.993748	3.960821

2.5 Summary

In this chapter, aiming to address the limitations of the existing dedarking solutions, we have proposed a deep learning dedarking network based on the cognitive perception model of Retinex theory. Our model effectively combines the inception network with high-level semantic information of the foreground and the background. Qualitative and quantitative experimental comparisons conducted on the benchmark ExDARK dataset [69] have demonstrated that our approach has improved the qual-

ity of the dedarked images and overcome the drawbacks of white and color distortion that are shown in the current state-of-the-art techniques.

Chapter 3

Saliency Detection from Low Light RGB-D Images

In the previous chapters, we have investigated for low-light image enhancement. Salient object detection (SOD) focuses on localizing and segmenting the most distinctive object(s) in a visual scene. It mimics the human visual attention mechanism to efficiently allocate visual processing resources on informative visual elements. Thus, SOD has been used as a pre-processing technique to supply informative cues for many other computer vision tasks, such as object detection [127], video object segmentation [81], semantic segmentation [111, 4], image editing [105] and intelligent vision surveillance in smart city applications [24]. Although the research on SOD is in full swing, the scope of its applications is mainly for the environment with relatively high contrast like in the daytime, and there is less research on the low light environment with poor lighting conditions like a dark night. In night scenes, human visual discrimination and machine recognition are difficult, and salient targets cannot be accurately located. This is also a problem faced by salient target detection in current night scenes. Figure 3.1 shows that if we directly conduct the saliency detection methods on the original low light images, there are lots of drawbacks, including low visibility, as well as the poor acquisition of the detailed information. The consequence of the enhanced low light image can lead to better effect of saliency dectection.

Therefore, in this research, as an important extension of low-light image analysis, we investigate low-light SOD. Experimental results show the universality of the model towards the high light images and low light images.

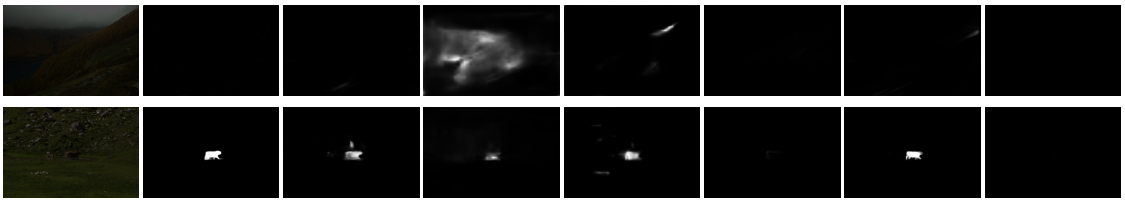


Figure 3.1 : Comparison of the performance of different SOD methods on low-light images.

3.1 Introduction

Video surveillance is an important area for operations such as analysis, detection or segmentation of targets based on images or videos. Intelligent video surveillance has an urgent need for transportation, finance, communities, shopping malls and other fields. It is also the gradual shift of security strategy from passive defense, a key element of active defense. At present, most of the analysis and processing of night video surveillance use manual observation. The visual attention mechanism of the human eye ensures that humans can still quickly pay attention to objects of visual interest when facing complex perceptual environments such as night. How to extend this characteristic of human vision to surveillance cameras and give full play to its active monitoring function of night scenes is an important task of current intelligent visual surveillance systems. It is important to improve the visibility of night surveillance video, give the computer the ability to accurately understand and perceive scene information in the night environment, and improve the security of sensitive occasions in social life (such as banks, shops, residential areas, parking lots, etc.) at night. Security has great practical significance.

Most SOD models [11, 96, 61, 104, 134, 106, 65] typically detect salient objects from RGB images. In a pioneer work [77], Ouerhani and Hugli showed that depth could also supply useful cues and largely boost the performance for saliency detection. This is also intuitive since human beings live in a real 3D environment and

depth largely impacts our perception of visual scenes. Many subsequent saliency models, *e.g.*, those in [45, 13, 12, 100], have started to leverage RGB-D images for saliency detection. Recently, Convolutional Neural Networks (CNNs) have been widely seen in the computer vision community and have also shown excellent performance on various computer vision tasks. Hence, many works have also introduced two-stream CNNs for RGB-D SOD to exploit their powerful feature learning capability.

Some deep models [29, 23] applied the two-stream Fully Convolutional Network (FCN) [70] architecture to feedforward each input RGB-D image pair into two CNN streams and directly obtained the saliency map by fusing their final feature maps, as shown in Figure 3.2(a). FCN processes the input image pair in a bottom-up manner, progressively extracting low-level features in shallow layers and high-level features in deep layers. Although it is simple and straightforward, the single path of the bottom-up information flow heavily limits the model's performance since usually the final feature map of a CNN is very coarse, thus the obtained saliency map lacks object details.

Considering the multi-level feature maps spontaneously obtained by each CNN, most of other works [5, 6, 98, 138] have adopted the two-stream UNet [85] architecture to aggregate multi-level features for RGB-D SOD. As shown in Figure 3.2(b), the two-stream UNet first uses two encoder networks to extract multi-level image features in a bottom-up manner. Then, there are one or two decoder networks to successively aggregate high-level features with low-level ones in a top-down processing and simultaneously fuse cross-modal features. In each decoder module, the features of its symmetric encoder module at the same level are reused through a skip connection and fused with previous decoder features. As such, discriminative semantic information in deep layers can be effectively integrated with local structures in shallow layers through the top-down propagation, thus enabling both accurate object

localization and precise shape and boundary segmentation.

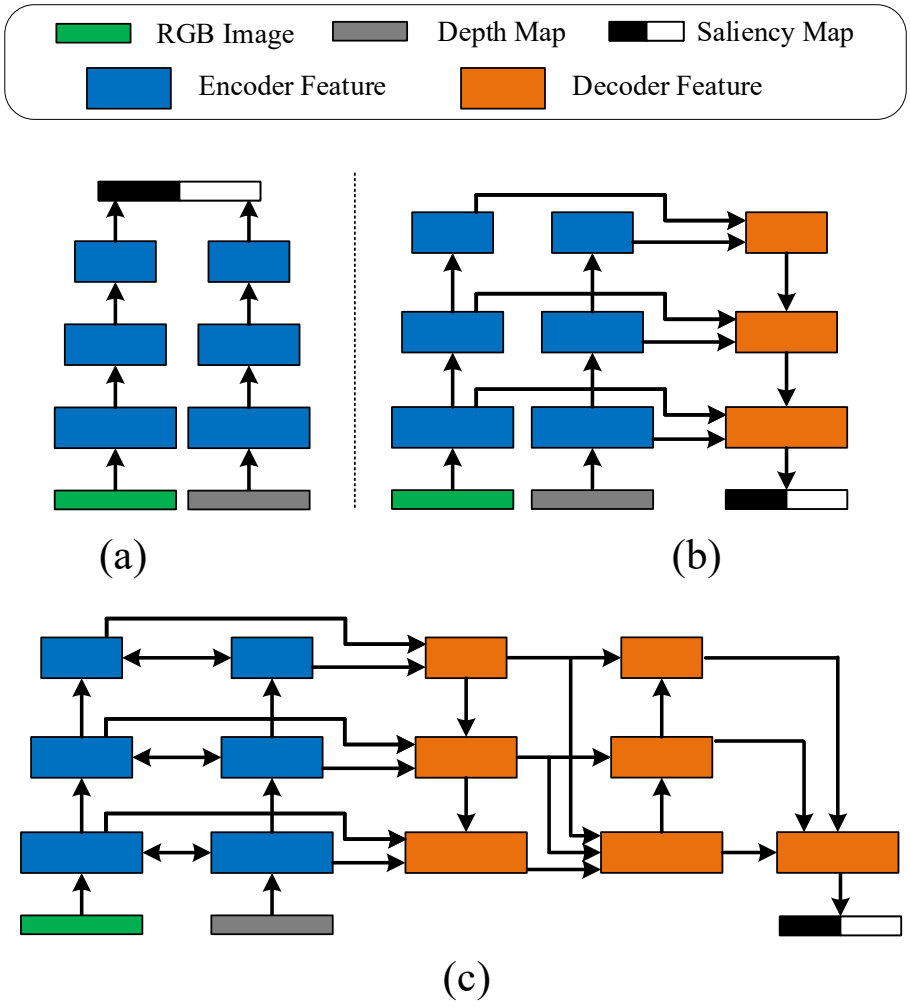


Figure 3.2 : Comparison of different network architectures. (a) Two-stream FCN [70]. (b) Two-stream UNet [85]. (c) Our proposed network. We cascade both top-down and bottom-up feature aggregation for deep RGB-D SOD to further leverage improved low-level features for promoting high-level features. We also propose to holistically aggregate features across all levels to learn plentiful multi-level feature interactions. Early aggregation paths are also presented to aggregate and propagate cross-modal encoder features.

However, UNet carries out top-down feature aggregation only once. Only high-level information can be aggregated with low-level features to improve their repre-

sentation ability in the decoder, while the high-level features themselves cannot be improved. To solve this problem, in this thesis, we propose to add an additional bottom-up aggregation path, in which the improved low-level features from the top-down path are propagated again to high-level layers, as shown in Figure 3.2(c). As we cascade both bottom-up and top-down feature aggregation, the features across all levels can be gradually improved.

Another problem is that the above networks only gradually aggregate features at every two adjacent levels. Although this feature aggregation scheme avoids large scale changes and is widely used in existing works, we argue that it limits the direct feature interactions among multi-level features. To alleviate this issue, we further propose holistic aggregation paths to holistically aggregate multi-level features after the bottom-up and top-down processing. Thus, the network can learn abundant cross-level feature fusion mechanism for SOD by considering them all at the same time.

Considering the two-stream architecture, the existing works usually simply adopt the two-stream encoders independently and only conduct feature aggregation in the decoding phase [29, 138, 17], or they fuse cross-modal encoder features to reuse them in decoders [59, 137, 48], without improving other encoder features. This is because they use pretrained CNN models as encoders, which require preserving their network structures and pretrained parameters. In this thesis, we present to aggregate and propagate cross-modal features at the early stage, i.e., in the encoding phase. We adopt a residual-learning based aggregation scheme to aggregate cross-modal encoder features and propagate them back to the original encoder paths, hence enhancing the feature capability from the very beginning.

Furthermore, the previous work usually aggregates features by directly concatenating [5, 6] or adding [81] them together. However, not all aggregated features

are helpful for the final SOD task. We propose to generate a gated attention for all of the involved features to modulate the aggregation flow at every node. To reduce the amount of the required gated attention weights and the computation and memory costs, we propose to factorize the gate matrix into the multiplication of channel-wise and spatial gates with multiple factors. This proposed multi-factored gated attention mechanism learns different gates in different factors and thus can ensemble multiple attention models to make a better decision.

To summarize, the main contributions of this chapter are as follows.

1) We propose a novel feature aggregation architecture for RGB-D SOD. We cascade both bottom-up and top-down feature aggregation paths and also introduce holistic aggregation paths, so we promotes both low-level and high-level features and boosts multi-level feature interactions. An early aggregation scheme is also presented to enhance the two-stream encoders.

2) We propose a novel factorized gated attention model for modulating the feature aggregation actions. We factorize the gated attention weight matrix of each feature map as the multiplication of two multi-factored channel-wise and spatial gate matrices. As such, both computational costs and model effectiveness are improved.

3) We conduct experiments on eight widely used RGB-D SOD benchmark datasets and low light image datasets. Experimental results demonstrate that all of the proposed model components can gradually improve the model's performance. Consequently, from the perspective of visualization and quantitative performance, our final model outperforms other state-of-the-art methods on both low light images and common images.

In Section 3.2, we first discuss our model with related work. Then, we present our model in Section 3.3 and report the experimental results in Section 3.4. Finally, in Section 3.5 we draw our conclusion.

3.2 Related Work

CNNs have been widely used for RGB SOD and RGB-D SOD. For the former, please refer to [102] for a comprehensive survey. We focus on the later in this thesis. In the two early pioneering deep RGB-D SOD works [84, 87], the authors used superpixels as the computational units and combined both traditional handcrafted features and CNNs to classify them as salient or non-salient. However, such schemes are usually computationally inefficient and therefore limit the model performance. Subsequent models start to adopt CNNs to directly process each input image and obtain the saliency map. Specifically, Han *et al.* [29] adopted two-stream CNNs to process RGB and depth images respectively, and then used fully connected layers to predict global saliency maps. Chen *et al.* [7] further combined this method with FCNs to fuse global and local contextual reasoning. In [23], Fan *et al.* first developed depth maps and then used single-stream FCNs with Pyramid Dilated Convolution modules [89] to predict saliency maps. These models directly predict saliency maps from the last layer of a CNN without considering multi-level features.

Most of the other works use the UNet architecture to gradually aggregate multi-level deep features. For instance, Chen *et al.* [6] first used two encoder networks to extract multi-level features from an RGB image and a depth image, respectively. Then, they proposed to densely fuse multi-level cross-modal features in a top-down decoder network. Zhao *et al.* [133] first proposed to leverage depth-based contrast to enhance the RGB encoder features, and then fused multi-level features using a top-down decoder with dense short connections. Liu *et al.* followed the work in [61] to embed recurrent convolutional layers into top-down decoder modules for fusing encoder and decoder features with the depth map. Li *et al.* [48] fused RGB and depth encoder features first and then also adopted a UNet style decoder to aggregate the multi-level features. All of these models only considered a top-down feature aggre-

gation path for RGB-D SOD, without exploring other feature aggregation schemes. In contrast, we cascade both top-down and bottom-up processing to promote features at all levels. Furthermore, most previous works, except the work in [7],directly use pretrained two-stream encoder networks without fusing and improving encoder features. The authors of [7] only propagated depth encoder features to RGB ones. In this chapter, we perform the bidirectional feature aggregation and propagation via the proposed early aggregation scheme.

Attention models are also widely used in RGB-D SOD models. Chen *et al.* [5] adopted the SENet [34] style channel attention in decoder modules to modulate feature channels. In [81], channel attention and spatial attention were separately adopted in a recurrent attention module for generating the final saliency maps. Liu *et al.* [66] proposed to selectively fuse self-mutual attention for fusing cross-modal information at the beginning of the decoder network. Different from the existing models, we propose to modulate the whole feature map in each decoder module with a gated attention and further present a multi-factored factorization mechanism to save computational costs and enhance the model capability.

In [14, 125, 64], a gated attention was also used in the convolution operation for language modeling, image inpainting, and RGB-D SOD, respectively. Different from them, we propose the multi-factored factorization operation for the gated attention to reduce computational costs and boost the model capability.

Two works are closely related to our proposed model. Chen and Li [10] also used both top-down and bottom-up decoders. The difference between our model and theirs are as follows. Firstly, they adopted the bottom-up decoder first to fuse cross-modal features and then used the top-down decoder to obtain coarse-to-fine saliency maps, while we build our model based on UNet and use the top-down decoder first. Secondly, we also propose to use the holistic aggregation paths to aggregate all-level

features simultaneously, while they only linearly fused the side output saliency maps. Thirdly, they used the existing SENet [34] style channel attention in the top-down decoder while we propose a novel factorized gated attention model and employ it in all aggregation paths. Fourthly, we also propose an early aggregation scheme to promote the two-stream encoders.

Wang *et al.* [103] proposed to iterate top-down and bottom-up decoders for multiple steps for RGB SOD. Different from them, although we only cascade top-down and bottom-up decoding paths once, our model's performance is already saturated. Furthermore, they adopt RNN in each decoder module to enhance the decoder capability while we use the proposed gated attention mechanism. We also propose the holistic aggregation paths to more effectively leverage multi-level features and present the early aggregation scheme for the two-stream architecture nature of the RGB-D SOD models.

3.3 Proposed Method

In this section, we articulate the proposed network for RGB-D SOD. Its detailed network architecture is shown in Figure 3.3.

We have reconsidered the feature aggregation schemes for deep RGB-D SOD and proposed novel feature aggregation methods. Based on the widely used two-stream UNet architecture, we have first proposed to add early aggregation and holistic aggregation paths to propagate cross-modal information in an early stage and learn abundant feature interactions among all multi-level features. We have also proposed to cascade the top-down decoder network in U-Net with a bottom-up decoder network, thus enabling to improve the high-level features with the already improved low-level features. Furthermore, we have proposed a factorised gated attention model to modulate the feature aggregation actions for each feature node with reduced computational costs and boosted model performance.

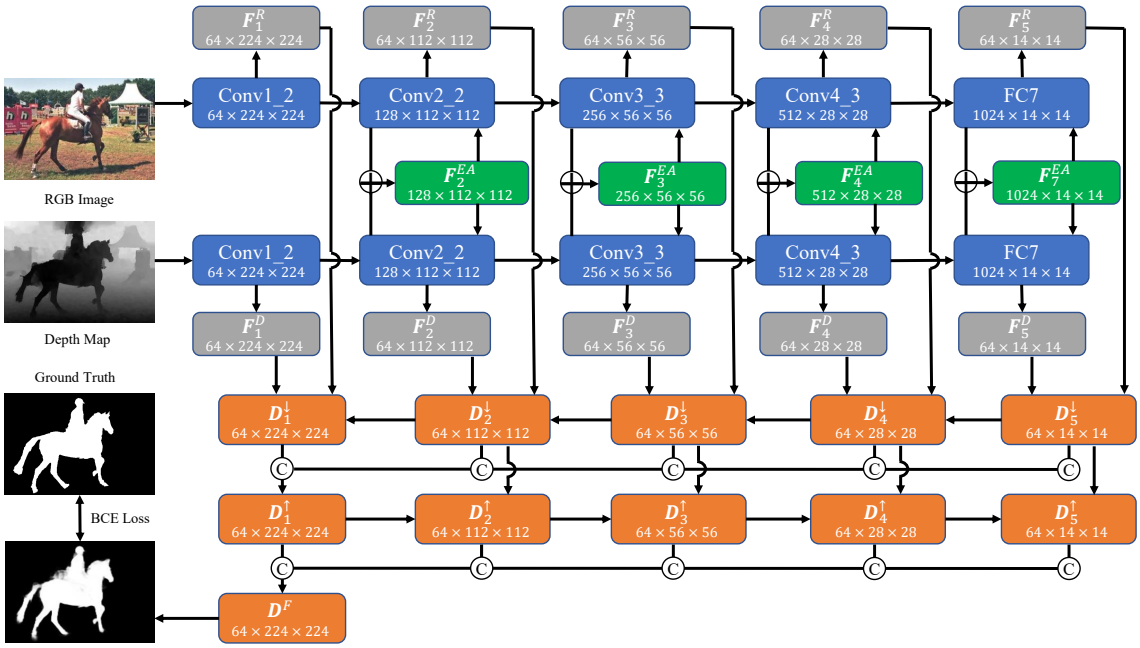


Figure 3.3 : Network architecture of the proposed RGB-D SOD model. We first use two encoder branches for the RGB and depth inputs to extract multi-level encoder features ($\boldsymbol{F}_*^R$ and $\boldsymbol{F}_*^D$). Within the two-stream encoders, we adopt early aggregation paths ($\boldsymbol{F}_*^{EA}$) to propagate cross-model information from the very beginning. Here, the early aggregation path for the two Conv5_3 layers is not shown. Then, we successively adopt a top-down decoder network ($\boldsymbol{D}_*^{\downarrow}$) and a bottom-up one ($\boldsymbol{D}_*^{\uparrow}$) to aggregate multi-level features. We also use holistic aggregation paths to directly aggregate features across all levels. The size of each feature map is also given and denoted by *channel* $\times$ *height* $\times$ *width*. Ⓒ denotes concatenation and $\oplus$ means element-wise summation.

3.3.1 Encoder Network

We first follow most previous methods and adopt a two-stream encoder network for extracting multi-level RGB and depth features. In order to learn common features for cross-modality, we share the network structure and parameters for the two encoder branches. To leverage better image features, we use an ImageNet [15] pre-

trained network as the encoder. The VGG 16-layer network [88] is adopted for a fair comparison with previous works. It has five convolutional (Conv) blocks and pooling layers, and two fully connected (FC) layers. For better adapting the network to SOD, we enhance the original VGG network by keeping large scale feature maps and preserving high-level FC layers. Concretely, we first reduce the stride of the pool5 layer to 1. Then, we convert the FC6 layer to a Conv layer with 1024 channels and 3×3 kernels, and adopt the dilated convolution algorithm [8] with $dilation = 6$. Similarly, the FC7 layer is also converted to a Conv layer with 1024 channels and 1×1 kernels. As such, the stride of the encoder network is reduced from 32 to 16 and high-level FC features are also preserved in the encoder.

To propagate cross-modal information from an early stage, we introduce early aggregation (EA) into the two encoders, specifically for the last Conv feature maps of the last four Conv blocks, which are Conv2_2, Conv3_3, Conv4_3, Conv5_3, and the FC7 layers. We do not use EA for the first Conv block since its low-level features may be quite different in the two modalities while the other higher layers can learn more common semantics. Given an RGB encoder feature map denoted by $\boldsymbol{E}_i^R$ and a depth one denoted by $\boldsymbol{E}_i^D$ from the same level, our EA path first aggregates them by element-wise summation and then taking an average to obtain the EA feature map:

$$\boldsymbol{F}_i^{EA} = \frac{\boldsymbol{E}_i^R + \boldsymbol{E}_i^D}{2}. \tag{3.1}$$

Then, we propagate $\boldsymbol{F}_i^{EA}$ back to the two encoder features using residual learning:

$$\begin{aligned} \boldsymbol{E}_i^R &= \boldsymbol{E}_i^R + \alpha \cdot Conv(\boldsymbol{F}_i^{EA}), \\ \boldsymbol{E}_i^D &= \boldsymbol{E}_i^D + \alpha \cdot Conv(\boldsymbol{F}_i^{EA}), \end{aligned} \tag{3.2}$$

Where two $Conv$ operations perform on two 1×1 Conv layers and α is a learnable parameter. We initialize α to 0 to make sure that the EA path brings no impact to the pretrained encoder networks at the beginning of the model training. As such,

the EA path boosts the encoder representation ability by leveraging cross-modal information and can also leverage pretrained model parameters.

Finally, we pick out the output feature maps of the Conv1_2, Conv2_2, Conv3_3, Conv4_3, and FC7 layers as the multi-level features and reuse them in the decoders later. Since these features have diverse channel numbers, we first use 3×3 Conv layers to convert each of them to 64 channels, thus making them compatible with each other in the subsequent feature aggregation. For representation simplicity, we denote these multi-level features by $\boldsymbol{F}_1^R$ to $\boldsymbol{F}_5^R$ and $\boldsymbol{F}_1^D$ to $\boldsymbol{F}_5^D$ for the RGB and the depth branches, respectively, as shown in Figure 3.3. The input scales of each RGB image and the depth map are fixed to 224×224 for simplicity. Hence, the sizes of the multi-level feature maps can be easily inferred, as marked in Figure 3.3.

3.3.2 Decoder Networks

After obtaining the ten multi-level features from both of the RGB and the depth branches, we aggregate them for RGB-D SOD. First, we follow UNet [85] to progressively aggregate features at every two adjacent levels in a top-down (denoted as $\downarrow$) decoder network. Specifically, in the i^{th} top-down decoder module, where $i \in \{1, 2, 3, 4\}$, we obtain its decoder feature $\boldsymbol{D}_i^{\downarrow}$ by aggregating the previous decoder feature $\boldsymbol{D}_{i+1}^{\downarrow}$ with the RGB and depth features $\boldsymbol{F}_i^R$ and $\boldsymbol{F}_i^D$ at this level. Since $\boldsymbol{D}_{i+1}^{\downarrow}$ has a smaller spatial size, we first upsample it by bilinear interpolation. For the 5^{th} decoder module, we directly aggregate $\boldsymbol{F}_5^R$ and $\boldsymbol{F}_5^D$. The top-down feature aggregation process can be summarized by:

$$\boldsymbol{D}_i^{\downarrow} = \begin{cases} Conv(BR([\boldsymbol{F}_i^R, \boldsymbol{F}_i^D])), & i = 5, \\ Conv(BR([UP(\boldsymbol{D}_{i+1}^{\downarrow}), \boldsymbol{F}_i^R, \boldsymbol{F}_i^D])), & i \in \{1, 2, 3, 4\}, \end{cases} \tag{3.3}$$

where $[\cdot]$ means the concatenation operation, BR means the batch normalization [36], $Conv$ denotes a 3×3 Conv layer with 64 channels and UP means the bilinear upsampling.

After the top-down feature aggregation, low-level features can be enhanced by high-level features. Thus, the final output feature map $\boldsymbol{D}_1^{\downarrow}$ simultaneously preserves local details and contains high-level semantics. Most of the previous works directly use this layer to predict the saliency maps. We further construct a bottom-up (denoted by $\uparrow$) decoder network to use the enhanced low-level features to improve the high-level features. To be concrete, we first use holistic aggregation paths to aggregate the features $\boldsymbol{D}_i^{\downarrow}$ at all levels to obtain the first feature map $\boldsymbol{D}_1^{\uparrow}$. Then, in the subsequent $i \in \{2, 3, 4, 5\}$ bottom-up decoder modules, we generate the decoder features $\boldsymbol{D}_i^{\uparrow}$ by aggregating the previous bottom-up decoder feature $\boldsymbol{D}_{i-1}^{\uparrow}$ with the top-down decoder feature $\boldsymbol{D}_i^{\downarrow}$ at this level. Since $\boldsymbol{D}_{i-1}^{\uparrow}$ has a larger spatial size, we downsample it using a max-pooling layer with stride of 2. The bottom-up feature aggregation process can be represented by:

$$\boldsymbol{D}_i^{\uparrow} = \begin{cases} Conv(BR([\boldsymbol{D}_1^{\downarrow}, UP(\boldsymbol{D}_2^{\downarrow}), \cdots, UP(\boldsymbol{D}_5^{\downarrow})])), & i = 1, \\ Conv(BR([DW(\boldsymbol{D}_{i-1}^{\uparrow}), \boldsymbol{D}_i^{\downarrow}]), & i \in \{2, 3, 4, 5\}, \end{cases} \tag{3.4}$$

where DW means down-sampling with a max-pooling layer.

After the bottom-up feature aggregation, high-level features can also perceive better low-level features thus generating better semantic information. Hence, by cascading both of the top-down and bottom-up decoder networks, we can simultaneously enhance all low-level and high-level features. Finally, we adopt the holistic aggregation again at the finest scale to obtain the final decoder feature map as:

$$\boldsymbol{D}^F = Conv(BR([\boldsymbol{D}_1^{\uparrow}, UP(\boldsymbol{D}_2^{\uparrow}), \cdots, UP(\boldsymbol{D}_5^{\uparrow})])). \tag{3.5}$$

A 1×1 Conv layer with 1 channel and the Sigmoid activation function can be used on top of $\boldsymbol{D}^F$ to obtain the final saliency map. During training, we also generate an intermediate saliency map from $\boldsymbol{D}_1^{\uparrow}$ in the same way. Then, we compute two binary cross entropy losses between the two saliency maps and the ground truth to train the whole network.

3.3.3 Factorized Gated Attention

It is worth noting that SOD is a challenging dense prediction task, and usually not all features are useful for the final decision. Thus, we propose to introduce the gated attention for the feature aggregation operations to adaptively select informative features for each decoder module. Specifically, for the *Conv* layers in (3.2), (3.3), (3.4), (3.5), considering an input feature map $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$, in which C, H, and W respectively denote its channel number, height, and width, we predict an gated attention matrix $\boldsymbol{G}$ of the same size with each of its elements in the range of $[0, 1]$. Then, we use $\boldsymbol{G}$ to modulate each node of $\boldsymbol{X}$ to control the aggregation flow in each decoder module as:

$$\boldsymbol{X}^G = \boldsymbol{G} \odot \boldsymbol{X}, \tag{3.6}$$

where $\odot$ is the element-wise multiplication. As such, $\boldsymbol{G}$ serves as a modulator and can retain informative features and suppress useless ones in $\boldsymbol{X}$. Then, we use $\boldsymbol{X}^G$ as the input for the *Conv* layers.

However, predicting $\boldsymbol{G}$ requires predicting all of the $C \times H \times W$ gate weights. A straightforward way is using a Conv layer with C channels on $\boldsymbol{X}$. Nevertheless, this scheme only uses local information, which equals to generating channel-wise gates for each pixel with shared parameters. Another way is to use an FC layer. This design is computationally prohibitive since it requires a large number of parameters to learn. We propose to learn a factorized form of $\boldsymbol{G}$ for reducing the number of attention weights to predict. Concretely, we factorize $\boldsymbol{G} \in \mathbb{R}^{C\times H\times W}$ into the multiplication of two low-rank matrices $\boldsymbol{G}^c \in \mathbb{R}^{C\times r}$ and $\boldsymbol{G}^s \in \mathbb{R}^{r\times(H\times W)}$. In this way, when a small number is used for r, the number of gate weights to predict can be reduced to $(C + H \times W) \times r$. For example, for $\boldsymbol{D}_2^{\downarrow}$ where $C = 192, W = H = 112$, using our factorization scheme with 2 factors, we can decrease the computational costs by 94.6 times.

Using the factorized attention, Equation (3.6) can be rewritten to:

$$\begin{aligned} \boldsymbol{X}^G &= \boldsymbol{G} \odot \boldsymbol{X} \\ &= (\boldsymbol{G}^c \boldsymbol{G}^s) \odot \boldsymbol{X} \\ &= \sum_{j=1}^{r} (\boldsymbol{G}_j^c (\boldsymbol{G}_j^s)^\top) \odot \boldsymbol{X}, \end{aligned} \tag{3.7}$$

where $\boldsymbol{G}_j^c \in \mathbb{R}^C$ and $\boldsymbol{G}_j^s \in \mathbb{R}^{(H \times W)}$ are the j^{th} factors of $\boldsymbol{G}^c$ and $\boldsymbol{G}^s$, respectively. We can respectively regard $\boldsymbol{G}_j^c$ and $\boldsymbol{G}_j^s$ as the traditional channel and spatial gated attention. In this way, $\boldsymbol{G}$ can be seen as being spanned by the outer product of channel attention and spatial attention. As such, we efficiently generate the attention weights for the entire feature map and leads to cheaper computation and memory costs. Furthermore, we generate r factors for both channel and spatial gated attentions, similar to the multi-head attention in [95]. Thus, our proposed factorized gated attention (FGA) mechanism can help to select different channels and spatial locations in different factors.

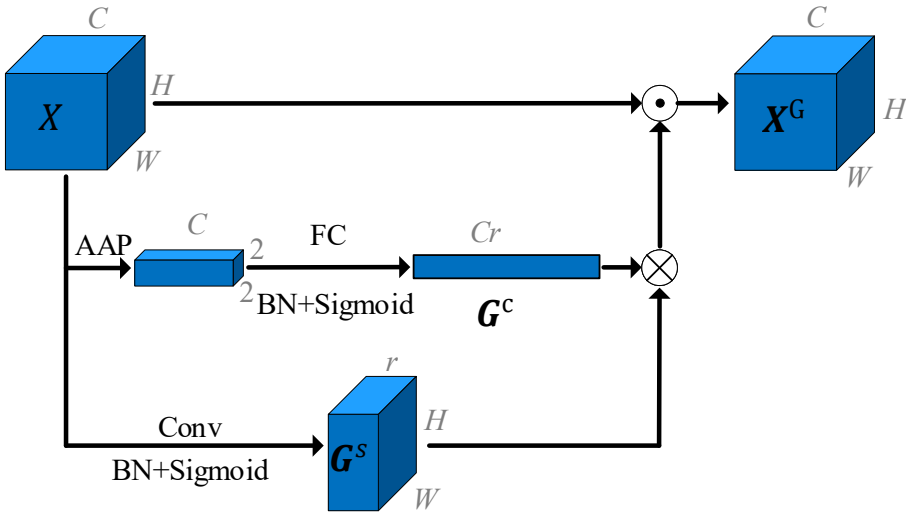


Figure 3.4 : Architecture of the proposed factorized gated attention module. We factorize the gated attention of the feature map $\boldsymbol{X}$ as the multiplication of multi-factored channel-wise gate weights $\boldsymbol{G}^c$ and spatial gate weights $\boldsymbol{G}^s$ to reduce computation and memory costs and introduce attention ensemble. AAP: adaptive average pooling. $\odot$: element-wise multiplication. $\otimes$: matrix multiplication. Sizes of some crucial features are marked by gray font.

Motivated by the SENet model [34], we use average pooling and an FC layer to predict $\boldsymbol{G}^c$. Specifically, we first adopt the adaptive average pooling on $\boldsymbol{X}$ to pool the entire feature map to the spatial size of 2×2. The resultant feature map represents the mean activation value of each channel in a $\frac{H}{2} \times \frac{W}{2}$ window. Then, we use an FC layer with BN and the Sigmoid activation function to generate $\boldsymbol{G}^c$, which is a vector of $C \times r$ dimensions. For generating $\boldsymbol{G}^s$, we first use a 7×7 Conv layer with r channels on $\boldsymbol{X}$. Then, BN and the Sigmoid activation function are used to obtain $\boldsymbol{G}^s$. Figure 3.4 shows the detailed architecture of the proposed FGA module.

Since each element of $\boldsymbol{G}^c$ and $\boldsymbol{G}^s$ is in the range of $[0, 1]$ and the summation over r factors in (3.7) will magnify the value range of the elements of $\boldsymbol{G}$, we further divide $\boldsymbol{G}$ by r to shrink its value range back to $[0, 1]$. The final formulation of the

proposed FGA module is:

$$\boldsymbol{X}^G = \frac{1}{r}(\boldsymbol{G}^c\boldsymbol{G}^s) \odot \boldsymbol{X}. \tag{3.8}$$

We write a new layer for this operation to implement it efficiently. Given $\partial L/\partial \boldsymbol{X}^G$ being the gradient of the loss function L with respect to $\boldsymbol{X}^G$, the gradients with respect to the three inputs can be easily obtained by the chain rule as:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{X}} &= \frac{1}{r}(\boldsymbol{G}^c\boldsymbol{G}^s) \odot \frac{\partial L}{\partial \boldsymbol{X}^G}, \\
\frac{\partial L}{\partial \boldsymbol{G}^c} &= \frac{1}{r}(\frac{\partial L}{\partial \boldsymbol{X}^G} \odot \boldsymbol{X})(\boldsymbol{G}^s)^\top, \\
\frac{\partial L}{\partial \boldsymbol{G}^s} &= \frac{1}{r}(\boldsymbol{G}^c)^\top(\frac{\partial L}{\partial \boldsymbol{X}^G} \odot \boldsymbol{X}).
\end{aligned} \tag{3.9}$$

Thus, the proposed FGA module can be trained along with other layers of the network simultaneously via existing gradient based optimizers.

We adopt FGA for all decoder modules and the generation of the multi-level encoder features $\boldsymbol{F}_*^R$ and $\boldsymbol{F}_*^D$. Experimental results in Section 3.4.4 demonstrate that the feature aggregation effectiveness for RGB-D SOD is further improved.

3.4 Experiments

In this section, we report the experimental results on benchmark datasets to validate the effectiveness of our proposed model.

3.4.1 Datasets

We evaluate the effectiveness of the proposed model on eight widely used RGB-D SOD benchmark datasets. The first one is the **NJUD** [42] dataset, which has 1985 stereo images. The images are selected from the Internet, 3D movies, and stereo photographs. The salient objects are labeled in a 3D display environment. The second one is the **NLPR** [80] dataset with 1000 RGB-D images collected by Microsoft Kinect. Most of them are indoor images with simple salient objects. The

third one is the **RGBD135** [12] dataset, which has 135 RGB-D indoor images captured by Kinect. The fourth one is the **LFSD** [51] dataset. It consists of 100 challenging images captured by the Lytro light field camera, including 60 indoor scenes and 40 outdoor scenes. The fifth one is the **STERE** [76] dataset, which has 1000 stereoscopic images. Many of the images include complex scenes and various objects. **SSD** [139] is the sixth dataset that has 80 images selected from three stereo movies. **DUT-RGBD** [81] dataset is the seventh one. It includes 800 indoor and 400 outdoor images with challenging scenes and generated depth maps. The last one **SIP** [22] dataset is a newly released one with 1000 human activities oriented images.

3.4.2 Implementation Details

We follow the previous work [81] to select 1400, 650, and 800 images from the **NJUD**, **NLPR**, and **DUT-RGBD** datasets, respectively, to train the proposed SOD network. To alleviate overfitting, we conduct data augmentation by first resizing each training image pair to 288×288 pixels and then randomly cropping 224×224 image patches and also using random horizontal flipping. The input image pairs are pre-processed by subtracting the mean RGB and depth pixels computed on the training set. We adopt the stochastic gradient descent (SGD) algorithm with momentum to train our network, where we set the batchsize, momentum, and weight decay to 4, 0.9, and 0.0005, respectively. We set the initial learning rate of the VGG part of the two encoder branches to 0.001 and train the other part of the network with random initialization and the initial learning rate of 0.01. We train the network with totally 60,000 steps and reduce the learning rates by 10 times at the $40,000$th and $50,000$th steps, respectively.

Our code is implemented based on an improved Caffe [38] library* to save the

*https://github.com/yjxiong/caffe

GPU memory. We use a GTX 1080 Ti GPU to accelerate network training and testing. During testing, we directly resize each image pair to 224×224 pixels as the input and get the network output as the predicted saliency map, without any post-processing technique. The testing process costs 0.089 seconds for each image.

3.4.3 Evaluation Metrics

We adopt four widely used SOD metrics. The first one is the max F-measure score. Concretely, for each image, we first use a series of thresholds, which vary from 0 to 1 to binarize the predicted saliency map. Then, we compare the binarized saliency maps with the ground truth saliency map, thus obtaining a series of precision-recall value pairs. F-measure comprehensively considers both precision and recall as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{3.10}$$

where β^2 is set to 0.3 as suggested in the previous work to emphasize more on precision. Max F-measure F_β^{max} is obtained by selecting the highest F-measure score under the optimal threshold.

The second metric is the Mean Absolute Error (MAE), which computes the average absolute difference between the predicted saliency map $\boldsymbol{S}$ and the ground truth saliency map $\boldsymbol{G}$ as:

$$MAE = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left|\boldsymbol{G}(w,h) - \boldsymbol{S}(w,h)\right|. \tag{3.11}$$

Although being widely used in previous work, the above two mentioned metrics are all based on pixel-wise errors and ignore structural information, and they are shown to be highly sensitive for the human visual system. Thus, we use the Structure-measure S_m [20] as our third metric to evaluate the structural similarity between the predicted saliency maps and the ground truth maps.

Fan *et al.* [21] recently simultaneously evaluate image-level statistics and local

pixel matching with the proposed Enhanced-alignment measure E_ξ, which demonstrated superiority over other existing measures. Thus, we also follow recent work to adopt this measure as the fourth metric.

3.4.4 Component Analysis

In this part, we analyze the effect of each proposed model component on four large datasets to verify their effectiveness. We use the two-stream UNet [85] as the baseline model, as shown in Row(I) of Table 3.1.

Table 3.1 : Ablation study on the effectiveness of the holistic aggregation paths (HA), the bottom-up aggregation (BU), the factorized gated attention (FGA), and the early aggregation (EA). **Bold** indicates the best performance.

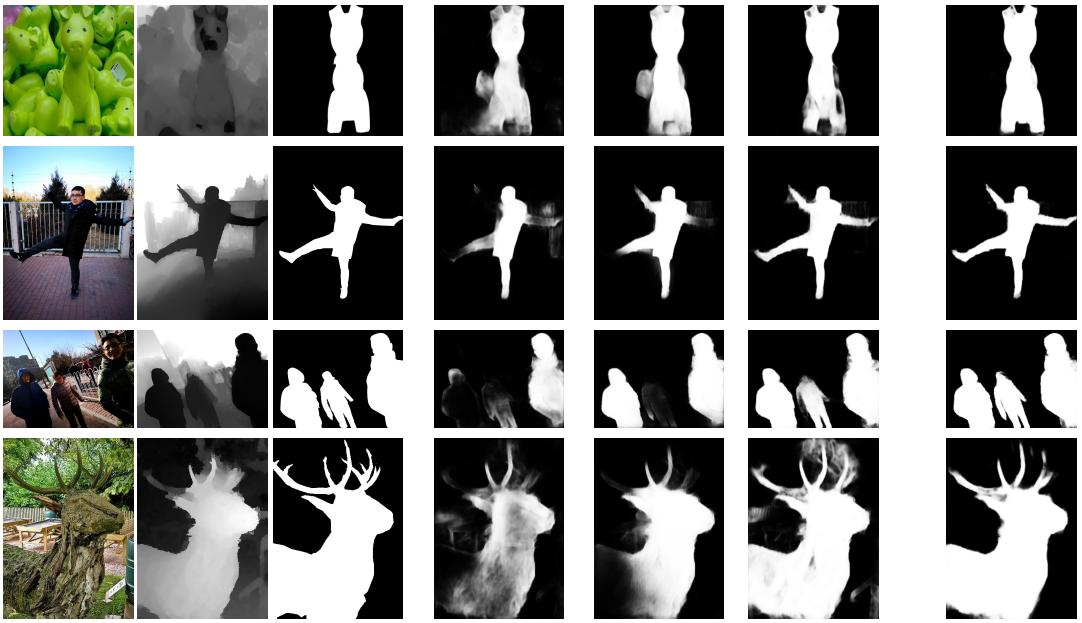
ID	Settings				NJUD [42]				NLPR [80]				DUT-RGBD [81]				STERE [76]			
	HA	BU	FGA	EA	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
I					0.888	0.889	0.930	0.059	0.908	0.894	0.951	0.036	0.898	0.906	0.937	0.052	0.891	0.888	0.936	0.055
II	✓				0.894	0.892	0.933	0.053	0.911	0.902	0.953	0.035	0.912	0.915	0.948	0.046	0.889	0.890	0.937	0.055
III	✓	✓			0.897	0.890	0.929	0.051	0.917	0.901	0.950	0.030	0.915	0.914	0.944	0.041	0.897	0.887	0.932	0.048
IV	✓	✓	1		0.899	0.890	0.928	0.048	0.914	0.894	0.944	0.031	0.918	0.921	0.949	0.042	0.897	0.887	0.934	0.049
V	✓	✓	2		0.901	0.893	0.933	0.047	0.920	0.901	0.953	0.029	0.921	0.926	0.952	0.037	**0.905**	**0.897**	**0.941**	0.043
VI	✓	✓	3		0.903	0.894	0.934	0.047	0.919	0.903	0.953	0.029	0.919	0.919	0.946	0.040	0.902	0.892	0.938	0.046
VII	✓	✓	2	✓	**0.906**	**0.902**	**0.936**	**0.045**	**0.927**	**0.912**	**0.961**	**0.025**	**0.926**	**0.927**	**0.954**	**0.034**	0.904	0.896	0.940	**0.042**

Holistic Aggregation Paths.

To evaluate the effectiveness of the proposed holistic aggregation paths, we directly aggregate decoder features across all levels of the UNet model on the finest level and use the obtained feature map (i.e., $\boldsymbol{D}_1^{\uparrow}$) to generate saliency maps. The results are shown in row (II) of Table 3.1. By comparing them with the results in row (I), we can see that aggregating multi-level features holistically can improve the performance of UNet, especially on the DUT-RGBD [81] dataset.

Bottom-up Aggregation.

We further add the bottom-up decoder network to promote high-level features using low-level features from the top-down decoder network of UNet. The results in row (III) show obvious performance gains based on the model setting in row (II), and demonstrate the effectiveness of an additional bottom-up feature aggregation path.



(a)Image (b)Depth (c)GT Saliency (d)Two-stream UNet (e)+HA+BU (f)+HA+BU+FGA (g)+HA+BU+FGA+EA

Figure 3.5 : Visual comparison of different model settings. We compare the results of the baseline Two-stream UNet (d), adding the holistic aggregation paths and the bottom-up aggregation (e), and further adding the factorized gated attention (f).

(a1)Image (b1)Depth (c1)Att 1 (d1)Att 2 (a2)Image (b2)Depth (c2)Att 1 (d2)Att 2

Figure 3.6 : Visualization of two learned two spatial attention factors for $\boldsymbol{D}_2^{\uparrow}$. "Att 1" and "Att 2" denote the two spatial attention maps, respectively.

Factorized Gated Attention.

We further adopt our proposed factorized gated attention in all decoder modules to verify its effectiveness. We have tried different settings with the factor number r varying from 1 to 3 and show the results in rows (IV) to (VI) of Table 3.1. We can see that when using 1 factor to factorize the gated attention, the model does not bring obvious performance gains when compared with the results in row (III). However, when we increase the factor number to 2 and 3, the model performance can be obviously improved. We also observe that the model performance saturates when r is greater than 2. Thus, we do not try other settings for r and select $r = 2$ as the best setting.

Early Aggregation.

The above model settings follow most previous works to use the original VGG network as encoders. Then, we add early aggregation paths between our two-stream encoders to introduce early cross-modal information interaction. The results are given in row (VII) of Table 3.1. We can see that adding early aggregation paths can effectively improve the model performance on most datasets. Thus, we select this model setting as our final SOD model.

Qualitative Comparison.

To further demonstrate the effectiveness of the proposed model components, we show a visual comparison in Figure 3.5. We can see that adopting the proposed holistic aggregation, the bottom-up aggregation, the factorized gated attention, and the early aggregation can gradually improve the SOD results. We observe that the proposed model components can help to not only recover missing salient regions, but also filter out redundant detected regions. As a result, the final model can obtain better saliency maps that are close to the ground truth.

What do the multi-factored attention learn?

Since we factorize the gated attention into the multiplication of channel-wise gated attention $\boldsymbol{G}^c$ and a spatial gated attention $\boldsymbol{G}^s$ with multiple factors, what do these multiple attention factors learn? To answer this question, we show the learned two spatial attention maps of our final SOD model in Figure 3.6 for the $\boldsymbol{D}_2^{\uparrow}$ feature map. We can see that the spatial attention maps mainly focus to highlight object boundaries. The two attention maps in each example are slightly different. Thus, our proposed multi-factored attention model can be seen as an ensemble of multiple submodules, and it has been widely proved to be useful in various machine learning algorithms. We also observe similar phenomena for the spatial attention in other layers and the channel-wise gated attention.

Table 3.2 : Comparison between FGA and existing attention models, including convolutional gated attention (CGA), spatial attention (SA), and the Convolutional Block Attention Module (CBAM). We report both RGB-D SOD performance and computational costs, which include both memory costs and running times during testing. Here we only test the network forwarding time and ignore the time for reading and writing images for rigorous comparisons. **Bold** indicates the best performance.

Attention	Mem	Time	NJUD [42]				NLPR [80]				DUT-RGBD [81]				STERE [76]			
	(Mb)	(s)	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
SA	**2247**	**0.057**	0.901	0.896	0.933	0.049	0.916	0.898	0.946	0.032	0.918	0.921	0.948	0.040	0.902	0.893	0.937	0.047
CGA	4139	0.226	**0.908**	**0.903**	**0.941**	0.044	0.916	0.898	0.951	0.031	0.924	**0.927**	**0.954**	0.036	0.903	0.894	0.939	0.046
CBAM[112]	2813	0.139	0.907	0.901	0.937	**0.043**	0.922	0.907	0.958	0.027	0.922	0.926	0.952	0.036	**0.904**	**0.896**	**0.941**	**0.042**
FGA	3033	0.066	0.906	0.902	0.936	0.045	**0.927**	**0, 0, 1**0.912	**0.961**	**0.025**	**0.926**	**0.927**	**0.954**	**0.034**	**0.904**	**0.896**	0.94	**0.042**

Comparison between FGA and existing attention models.

We compare our proposed FGA with the conventional convolutional gated attention (CGA), spatial attention (SA), and the Convolutional Block Attention Module (CBAM) [112], in terms of both model performance and computational costs. For CGA, we simply use a 7×7 Conv layer to generate the gated attention weights with the same size with each input feature map. The attention generation for SA is similar, except that we generate a single channel attention map. For CBAM, we use the default settings to incorporate cascaded channel and spatial attention. We substitute FGA in our SOD model with these three attention models and report the comparison results in Table 3.2. The results clearly show that our proposed FGA model achieves the best RGB-D SOD performance. In terms of computational costs, we can see that FGA costs much less GPU memory than CGA and is much faster than CGA and CBAM. Compared with CGA, FGA predicts much fewer attention weights. Compared with CBAM, FGA only needs to carry out the attending operation once while CBAM needs to do it twice. Compared with SA, FGA costs a little more inference time but achieves better model performance.

3.4.5 Comparison with State-of-the-art Models

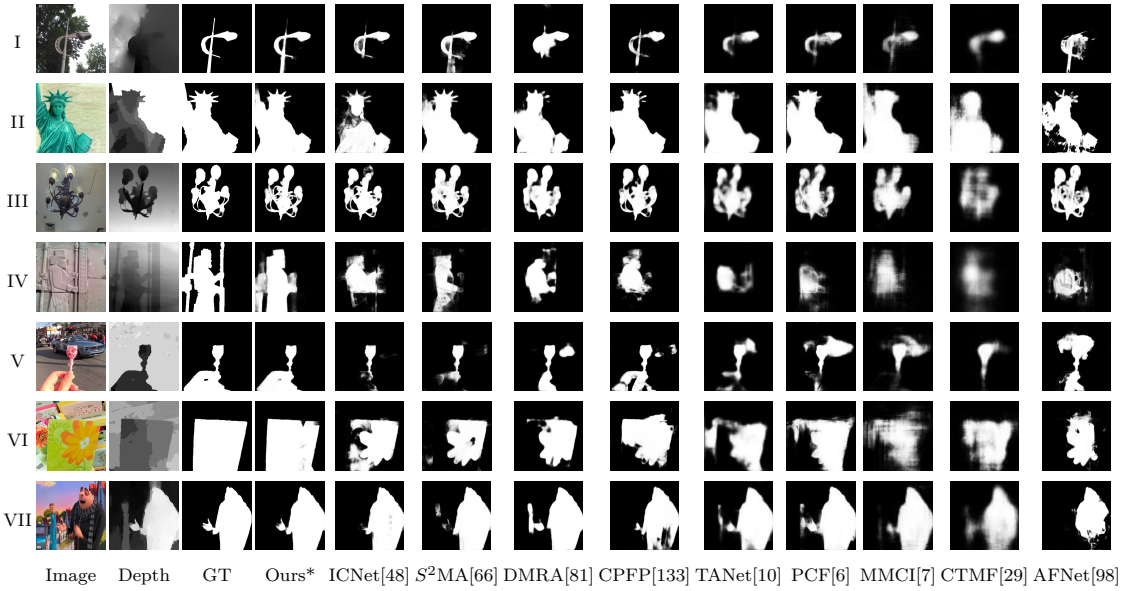


Figure 3.7 : Visualization of the saliency maps of our SOD model and other state-of-the-art RGB-D SOD models.

The universality of the model.

More experimental results illustrate the universality of the model towards the high light images and low light images.

We use our designed CNN-based model in chapter 2 to improve the ExDARK [69] sample images, which comprise more than 7000 original low-light shots. Then, after the previous improvement, we perform seven separate SOD techniques on these samples. Figure 3.8 represents the CNN-based enhanced results.

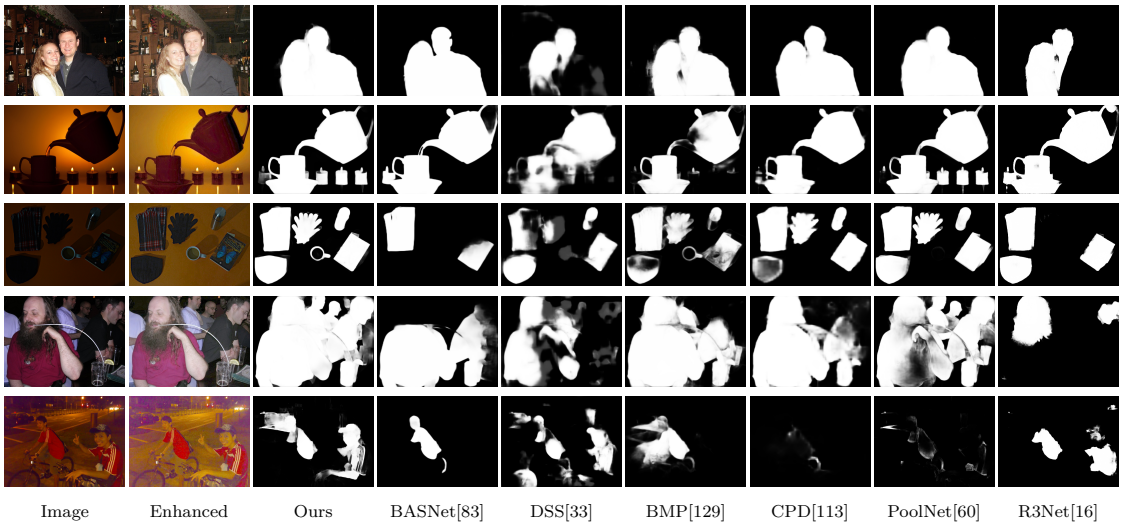


Figure 3.8 : Comparison of different SOD methods conducted on enhanced low light images from ExDARK Dataset [69] with our designed CNN-based model in chapter 2.

To verify the effectiveness of our final model for RGB-D SOD, we conduct a performance comparison with other 11 state-of-the-art RGB-D SOD methods. We consider recently published deep-learning-based models, including CTMF [29], MMCI [7], PCF [6], TANet [10], CPFP [133], DMRA [81], S^2MA [66], ICNet [48], UCNet [128], and JL-DCF [26].

The quantitative comparison in terms of the above mentioned four metrics is reported in Table 3.3. Since most compared models, except DMRA and S^2MA, were trained on only two datasets, i.e., NJUD and NLPR, we report the comparison results of the methods using both 2 and 3 training datasets for fair comparisons. The results show that, when using 2 training datasets, our proposed model achieves a comparable performance with the SOTA UCNet. When trained on 3 datasets, our model obviously outperforms all other methods, including all of those trained on either 2 or 3 datasets.

On the other hand, we show a qualitative model comparison of the saliency

Table 3.3 : Quantitative comparison of our proposed model with state-of-the-art RGB-D SOD methods. We report comparison results under two settings, i.e., training with 2 datasets (NJUD and NLPR) and training with 3 datasets (NJUD, NLPR, and DUT-RGBD). Underline and **Bold** indicate the best and the second best performance under each setting, respectively. **Underline** means the best performance under both settings. Note that, for fair comparisons, we show the results of the JL-DCF [26] model with the VGG backbone, whose results are only reported on 6 datasets in their paper.

Dataset	Metric	Training with 2 Datasets									Training with 3 Datasets		
		CTMF [29]	MMCI [7]	PCF [6]	TANet [10]	CPFP [133]	ICNet [48]	UCNet [128]	JL-DCF [26]	Ours	DMRA [81]	S^2MA [66]	Ours*
NJUD	$S_m \uparrow$	0.849	0.858	0.877	0.878	0.878	0.894	**0.897**	**0.897**	**0.908**	0.886	**0.894**	0.906
	maxF $\uparrow$	0.845	0.852	0.872	0.874	0.877	0.891	0.895	**0.899**	0.901	0.886	**0.889**	**0.902**
	$E_\xi \uparrow$	0.913	0.915	0.924	0.925	0.923	0.926	0.936	**0.939**	**0.943**	0.927	**0.930**	0.936
[42]	MAE $\downarrow$	0.085	0.079	0.059	0.060	0.053	0.052	**0.043**	0.044	**0.040**	**0.051**	0.053	0.045
NLPR	$S_m \uparrow$	0.860	0.856	0.874	0.886	0.888	0.923	0.920	0.920	**0.922**	0.899	**0.915**	**0.927**
	maxF $\uparrow$	0.825	0.815	0.841	0.863	0.867	0.908	0.903	0.907	**0.908**	0.879	**0.902**	**0.912**
	$E_\xi \uparrow$	0.929	0.913	0.925	0.941	0.932	0.952	0.956	0.959	**0.957**	0.947	**0.953**	**0.961**
[80]	MAE $\downarrow$	0.056	0.059	0.044	0.041	0.036	0.028	**0.025**	**0.026**	**0.026**	0.031	**0.030**	**0.025**
RGBD135	$S_m \uparrow$	0.863	0.848	0.842	0.858	0.872	0.920	0.933	0.913	**0.925**	0.900	**0.941**	**0.943**
	maxF $\uparrow$	0.844	0.822	0.804	0.827	0.846	**0.913**	**0.930**	0.905	0.910	0.888	**0.935**	**0.937**
	$E_\xi \uparrow$	0.932	0.928	0.893	0.910	0.923	0.960	0.976	0.955	**0.963**	0.943	**0.973**	**0.978**
[12]	MAE $\downarrow$	0.055	0.065	0.049	0.046	0.038	0.027	**0.018**	**0.026**	0.018	0.030	**0.021**	**0.016**

Dataset	Metric	Training with 2 Datasets									Training with 3 Datasets		
		CTMF [29]	MMCI [7]	PCF [6]	TANet [10]	CPFP [133]	ICNet [48]	UCNet [128]	JL-DCF [26]	Ours	DMRA [81]	S^2MA [66]	Ours*
LFSD	$S_m \uparrow$	0.796	0.787	0.794	0.801	0.828	0.868	**0.864**	0.833	0.860	**0.847**	0.837	**0.879**
	maxF $\uparrow$	0.791	0.771	0.779	0.796	0.826	0.871	0.864	0.840	**0, 0,**	**0.856**	0.835	**0.881**
										10.867			
	$E_\xi \uparrow$	0.865	0.839	0.835	0.847	0.872	0.903	0.905	0.877	**0.904**	**0.900**	0.873	**0.914**
[51]	MAE $\downarrow$	0.119	0.132	0.112	0.111	0.088	**0.071**	0.066	0.091	0.078	**0.075**	0.094	**0.062**
STERE	$S_m \uparrow$	0.848	0.873	0.875	0.871	0.879	1, 0, 00.903	0.903	0.894	**0.897**	0.886	**0.890**	**0.904**
	maxF $\uparrow$	0.831	0.863	0.860	0.861	0.874	**0.898**	**0.899**	0.889	0.887	**0.886**	0.882	0.896
	$E_\xi \uparrow$	0.912	0.927	0.925	0.923	0.925	**0.942**	**0.944**	0.938	0.934	**0.938**	0.932	0.940
[76]	MAE $\downarrow$	0.086	0.068	0.064	0.060	0.051	**0.045**	**0.039**	0.046	0.048	**0.047**	0.051	0.042
SSD	$S_m \uparrow$	0.776	0.813	0.841	0.839	0.807	0.848	**0.865**	-	**0.880**	0.857	**0.868**	0.876
	maxF $\uparrow$	0.729	0.781	0.807	0.810	0.766	0.841	**0.855**	-	**0.871**	0.844	**0.848**	0.852
	$E_\xi \uparrow$	0.865	0.882	0.894	0.897	0.852	0.902	**0.907**	-	0.926	0.906	**0.909**	0.915
[139]	MAE $\downarrow$	0.099	0.082	0.062	0.063	0.082	0.064	**0.049**	-	**0.045**	0.058	**0.052**	0.049
DUT-	$S_m \uparrow$	0.831	0.791	0.801	0.808	0.818	0.852	0.897	-	**0.870**	0.889	**0.903**	**0.926**
RGBD	maxF $\uparrow$	0.823	0.767	0.771	0.790	0.795	0.850	0.895	-	**0.860**	0.898	**0.901**	**0.927**
	$E_\xi \uparrow$	0.899	0.859	0.856	0.861	0.859	0.899	0.936	-	**0.901**	0.933	**0.937**	**0.954**
[81]	MAE $\downarrow$	0.097	0.113	0.100	0.093	0.076	0.072	0.043	-	**0.066**	0.048	**0.043**	**0.034**
SIP	$S_m \uparrow$	0.716	0.833	0.842	0.835	0.850	0.854	**0.875**	0.866	0.881	0.806	**0.872**	**0.889**
	maxF $\uparrow$	0.694	0.818	0.838	0.830	0.851	0.857	**0.879**	0.873	0.884	0.821	**0.877**	**0.889**
	$E_\xi \uparrow$	0.829	0.897	0.901	0.895	0.903	0.903	**0.919**	0.916	0.926	0.875	**0.919**	**0.930**
[22]	MAE $\downarrow$	0.139	0.086	0.071	0.075	0.064	0.069	**0.051**	0.056	0.049	0.085	**0.057**	**0.047**

maps in Figure 3.7. The results show that the saliency maps of our model can not only highlight salient objects more accurately, but also recover object details more precisely (see Row III). Our model can also cope with various challenging scenarios, *e.g.*, the large statue in Row II, the very challenging relief in Row IV, and the book in row VI, where most other SOTA models fail to completely highlight the salient objects. For Rows V and VII, although the backgrounds are very cluttered, our model can successfully separate the salient objects from the backgrounds despite that other SOTA models are largely distracted by the backgrounds.

3.4.6 Failure Analysis

Figure 3.9 : Visualization of common failure patterns.

We show some common failure patterns in Figure 3.9. We observe that our RGB-D SOD model mainly fails in four cases. The first row of Figure 3.9 demonstrates that it is hard to perceive low-level (*e.g.*, color) contrast thus may incorrectly localize salient objects. The left example in the second row shows that extreme illumination condition is a challenge for our model. The right example shows that it may be distracted by cluttered backgrounds. The last row indicates that it may also fail when facing images with no obvious salient objects. All of these four cases are challenging for all deep learning based SOD models. Solving these problems can be our future work.

3.5 Summary

We learned in this chapter that UNet-based architectures are frequently used in Deep RGB-D salient object detection models, although UNet only uses a top-down decoder network to progressively combine high-level functions with low-level functions. In this chapter, we proposed employing holistic aggregation pathways and a bottom-up decoder network to improve the function of aggregation. The former combines multi-level features in a holistic way to learn a large number of function interactions, whereas the latter combines improved low-level features with high-level features to improve their representation capacity. When compared to relatively current state-of-the-art approaches, experimental results have shown that our final RGB-D SOD model is more effective.

Chapter 4

Low-Light Saliency Detection via Deep CNN without Depth

In the previous chapter, we have investigated saliency detection on low-light RGB-D images, where RGB and depth saliency cues are fused to obtain the saliency detection results. In this way, depth maps provide complementary information for appearance cues and thus promote the saliency detection performance, especially for challenging scenarios. Nevertheless, 3D sensors are not popular and usually expensive, making RGB-D images much more difficult to obtain than RGB images. Moreover, how to solve the depth information loss under low illumination is another key problem. In this chapter, we propose a novel deep learning framework to detect RGB-D saliency without actually requiring input depth data. Specifically, we predict depth maps for RGB images and simultaneously fuse depth features with RGB features to detect salient objects. Experimental results illustrate the universality of the model towards the high light images and low light images.

4.1 Introduction

The existing saliency models usually detect salient objects from RGB signals, which can be easily captured by modern cameras or cell phones. They usually use the contrast mechanism [37] to find the regions different from others and extract semantic features to find the regions that are most likely to be objects. Although the recent CNN-based RGB saliency models [61, 33, 62, 25, 134] and other models [58, 119, 2, 35, 19] have achieved very promising results, they can still easily fail to detect salient objects in challenging scenarios since RGB data can only provide visual cues

for saliency detection, thus greatly limits the model capability. In this chapter, we talk about how low-light images can be dealt with in the saliency detection after images are enhanced.

Figure 4.1 shows two examples of natural images. We can see that for the images with cluttered backgrounds (the top row), saliency models that use RGB cues only (the column (c)) can be easily distracted by backgrounds. For salient objects with complex appearance (the bottom row), using only the RGB cues (the column (c)) easily leads to incomplete detection.

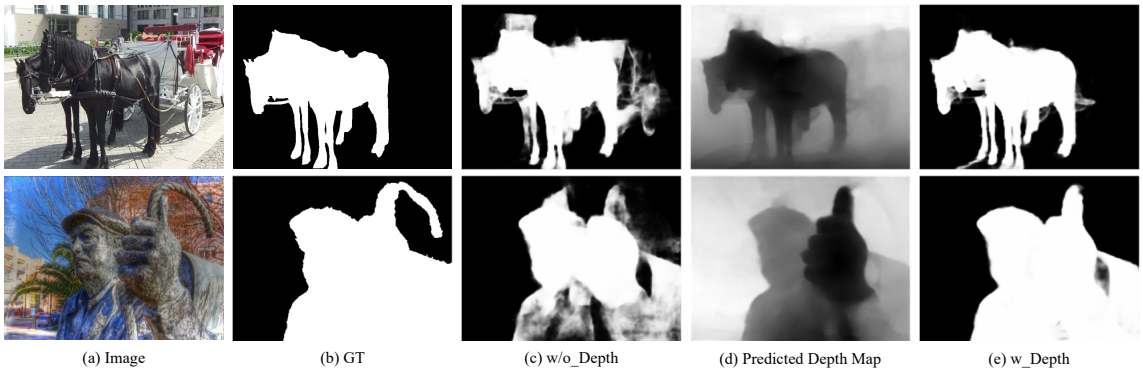


Figure 4.1 : Comparison of the saliency detection results without ("w/o_Depth") and with ("w_Depth") using depth cues. (a) and (b) show two example images and their ground truth (GT) saliency maps. (c) shows the saliency detection results of a baseline deep saliency model without using depth cues. (d) shows our predicted depth maps. (e) shows our predicted saliency maps with using depth cues.

On the other hand, the human visual system can easily leverage 3D information from the real world, in which the depth information plays a very important role in visual perception. For example, in the human visual attention mechanism, salient objects usually have different depths with the backgrounds. To this end, some researchers propose to detect saliency using RGB-D data. First, RGB-D images are captured using 3D sensors such as Microsoft Kinect, stereo cameras, light field cameras, etc. Then, they fuse RGB and depth saliency cues to obtain the saliency

detection results. As such, depth maps can provide complementary information for appearance cues and thus promote the saliency detection performance, especially for challenging scenarios.

Nevertheless, 3D sensors are not popular and usually expensive, making RGB-D images much more difficult to obtain than RGB images. For the common vision, which is usually taken by the dual cameras instead of the monocular devices, how to solve the depth information loss under low illumination is another key problem. In this chapter, we propose a novel deep learning framework to detect RGB-D saliency without actually requiring input depth data. Specifically, we predict depth maps for RGB images and simultaneously fuse depth features with RGB features to detect salient objects. By using the predicted depth information, our model can filter out the distraction from backgrounds (see the top row of Figure 4.1) and highlight the whole salient object more uniformly (see the bottom row of Figure 4.1).

On one hand, we leverage depth information to detect saliency more accurately. On the other hand, we do not require testing images to have depth maps and only use common RGB images. In Figures 4.1 (c) and (e), we show the comparison of the saliency detection results between our proposed model (column (e)) and a baseline model without using depth cues (column (c)). The results show that our model can obviously improve saliency detection performance for RGB images.

Furthermore, the previous RGB-D saliency detection models usually fuse RGB features with depth features by using simple feature concatenation, addition [29, 6], or attention models [81, 133]. Different from them, inspired by the DenseASPP (DASPP) model [122], we propose a novel multimodal feature fusion model by densely fusing RGB and depth DASPP features, thus greatly enriching the feature fusion paths across multiple scales. Considering that the original DASPP model only incorporates multiscale local features, we also propose to enhance it

with an additional global context propagation module [108]. Experimental results demonstrate that this proposed feature fusion model can better improve multimodal features. Finally, our saliency model outperforms previous RGB based saliency detection methods and also achieves comparable or even better results compared with state-of-the-art RGB-D saliency algorithms.

We have proposed to simultaneously estimate the depth and detect saliency for RGB images in a unified deep CNN. Intermediate depth features can be fused with RGB saliency features to supply complementary information for improving the saliency detection performance. We have further proposed to fuse multiscale depth and RGB features, and introduced global contexts.

To summarize, our contributions of this chapter are as follows.

1) We propose the first deep saliency model to leverage depth cues for enhancing the saliency detection performance but without actually using depth data.

2) We propose a novel depth feature fusion model by introducing dense fusion paths in DASPP and also enhance it by incorporating global contexts.

3) Experimental results demonstrate the effectiveness of our proposed model on both low light images and common images. It not only outperforms the previous RGB saliency models, but also can obtain comparable results with the state-of-the-art RGB-D saliency methods.

In the rest of the chapter, we first discuss our model with related work in Section 4.2. Then, we present our proposed model in Section 4.3 in detail and report the experimental results in Section 4.4. Finally, we draw conclusion in Section 4.5.

4.2 Related Work

Early RGB saliency detection models usually extract low-level image features and then leverage the contrast mechanism [63, 11], background prior [121, 91], or

objectness prior [30, 3] to detect salient objects. Recently, many researches works introduced CNNs into the saliency detection field and have achieved very promising results. Most of these methods directly solve the saliency detection problem using end-to-end CNNs. For example, early models [96, 49, 135] usually use multi-scale CNNs to extract multi-scale features for each pixel or superpixel from its multiple local and global patches, and then combine the multi-scale deep features to classify or regress its saliency value. Subsequent models adopt the fully convolutional network (FCN) [70] architecture to perform saliency classification for individual pixels, simultaneously. Typically, encoder and decoder model has been a trend for researchers [24]. An encoder with pretrained parameters is first used, such as the work of VGG [88] or ResNet [31], to extract multi-level deep features, and then a decoder is built to fuse these multi-levels features for saliency detection. Some works [61, 72, 132, 9] use the U-Net [85] architecture to progressively fuse multi-level features, and some other works [55, 33] adopt the HED [117] network architecture to fuse them simultaneously. The above-mentioned methods all directly infer image saliency from extracted deep features, without considering other knowledge.

Some complementary knowledge has been introduced to enhance the saliency detection performance. Li *et al.* [52] introduced the semantic segmentation task to enhance the feature capability for object perception. Wang *et al.* [104] used eye fixation to guide the detection of salient objects. In [130], Zhang *et al.* leveraged image captioning to help to capture semantic information of salient objects in visual scenes. Recently, many deep saliency models [53, 106, 25, 60, 134] have been proposed to simultaneously predict object contours and use the contour prior to enhance the object boundaries for salient objects. However, none of these works has explored to use depth knowledge to enhance the saliency detection performance. In this work, we propose to simultaneously predict the depth map for each image and use the depth features to supplement the RGB features for saliency detection.

In [116], Xiao *et al.* also proposed to derive pseudo depth from an RGB image, and then leverage the pseudo depth to boost the performance of RGB saliency model by computing a depth-driven background prior and a depth contrast feature. Our method significantly differs from theirs in two aspects. First, their model is based on traditional saliency features and frameworks, while ours is an end-to-end deep saliency model, thus having much better model performance and much faster speed. Second, their model needs to derive the pseudo depth map first and then use it to compute the depth-based feature and prior map, while ours can use the intermediate depth features to boost the RGB features before the generation of the depth map, thus is more effective and efficient.

4.3 RGB-D Saliency Detection

Traditional RGB-D saliency models usually use the depth map as another channel and follow RGB saliency models to derive some saliency cues, such as depth-based contrast [12] or background priors. Finally, RGB and depth cues are combined to obtain the final saliency detection results. Some other models propose some special saliency cues from the depth data, such as the shape and 3D layout priors proposed in [13], to supplement the RGB saliency cues. Recently, many works adopted CNNs in the RGB-D saliency detection task and have obtained much better results than traditional models. Some of the models [84, 67] regard the depth map as the fourth channel besides the RGB image and then train a deep saliency model with four-channel input images. Some other models [98] adopt two CNNs on the RGB image and the depth map separately to generate two saliency maps and then fuse them to obtain the final saliency map. Most works use two-stream CNNs to extract RGB and depth features from the two modalities, respectively, and then fuse the multimodal features with various methods, such as feature concatenation and addition [29, 6], spatial channel attention [81, 133], and mutual attention [66]. All of these methods

require to input the captured depth maps into the saliency models for enhancing their performance. In contrast, we propose to infer depth maps from the input RGB images and simultaneously leverage the intermediate depth features to enhance the RGB features, thus exploiting the depth knowledge for saliency detection without actually requiring depth data. Furthermore, different from previous feature fusion schemes, we propose a novel feature fusion model by adding dense fusion paths and a global context propagation branch in DASPP.

A contemporary work in [82] also proposed to eliminate the dependency on depth maps for RGB-D saliency detection. They trained a saliency detection branch based on depth data, and then used the result to perform knowledge distillation for promoting the model capability of the RGB saliency branch. Although our model and theirs try to achieve the same goal, the implementation mechanisms are totally different. First, they aimed at designing a light-weight saliency detection model and adopted the knowledge distillation technique, while we aim to build a powerful saliency model and also eliminate the dependency on the input depth maps, hence performing multi-task learning. As a result, they can only leverage RGB-D saliency detection data to train their model, while ours can exploit large-scale external RGB saliency detection data and depth estimation data. Second, their model was implemented based on knowledge distillation, hence the model capacity was theoretically limited by the teacher network, i.e., the depth saliency detection branch. In contrast, our method explicitly fuses RGB information with the inferred depth feature. Thus, the model capacity is an ensemble of both modalities. As a result, our model is much more effective than theirs, although theirs may be more efficient.

In this part, we articulate the proposed deep saliency model in detail. As shown in Figure 4.2, given each image, we first use an encoder network to extract multi-level encoder features. Then, we follow the U-Net [85] architecture to progressively

fuse the multi-levels features to predict the depth map and the saliency map via two decoder networks, respectively. We also fuse the depth features with the RGB features to leverage depth cues for enhancing the saliency detection performance. Specifically, we adopt a DASPP [122] model at the beginning of the depth decoder branch, and then fuse the depth DASPP features with RGB features in a novel Dense MultiScale Fusion (DMSF) module. Subsequently, we fuse each depth decoding feature map with each corresponding RGB decoding feature map.

4.3.1 Encoder Network

Following the work in[62], our encoder network is based on the VGG-16 network [88]. It is an FCN and has seven convolutional (Conv) blocks. The first five blocks are based on the five Conv blocks of VGG-16, i.e., Conv1-Conv5, each of which is composed of two or three consecutive Conv layers. The last two blocks are based on the two fully connected (FC) layers of VGG-16, i.e., FC6 and FC7. Since the original VGG-16 network has five pooling layers following the five Conv blocks, respectively, thus downsampling the input image by a factor of 32, which is too large for saliency detection. To enlarge the spatial sizes of the feature maps, we change the strides of the last two pooling layers to 1, and also use atrous Conv layers [8] with rate $r = 2$ in the Conv5 block. We also transform the two FC layers of VGG-16 to Conv layers for taking advantage of the plentiful features learned in them. Concretely, the FC6 layer is transformed to a 3×3 atrous Conv layer with $r = 12$ and 1024 channels, while the FC7 layer is transformed to a 1×1 Conv layer with the same channel number. Finally, we obtain the final FC7 feature map with a downsampling factor of 8, and also get five multi-level feature maps from Conv1-Conv5.

4.3.2 Decoder Networks

Next, we construct two decoder branches for predicting the depth map and the saliency map. We name them the depth branch and the RGB saliency branch,

respectively. We first use DASPP and DMSF to extract and fuse multiscale features from the FC7 encoder features, and then follow the U-Net [85] architecture to progressively fuse multi-level encoding features in subsequent decoding modules.

DASPP and DMSF

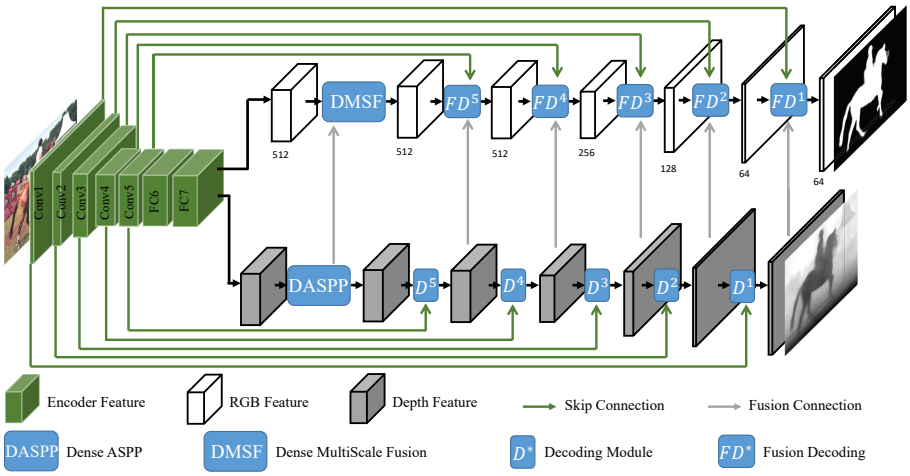


Figure 4.2 : Network architecture of the proposed model. The whole model has an encoder network (green cubes) and two decoder networks (white and gray cubes). The encoder network is used to extract multi-level encoder features, while the two decoder networks are used for predicting the depth map and the saliency map, respectively. We use the VGG-16 network [88] as our encoder, and its multi-level features are marked on the cubes. Each decoder progressively fuses the multi-level features by using skip-connections. The depth features are also fused with the RGB features via fusion connections for enhancing the saliency detection performance. The channel numbers of the decoding modules are also marked under the cubes.

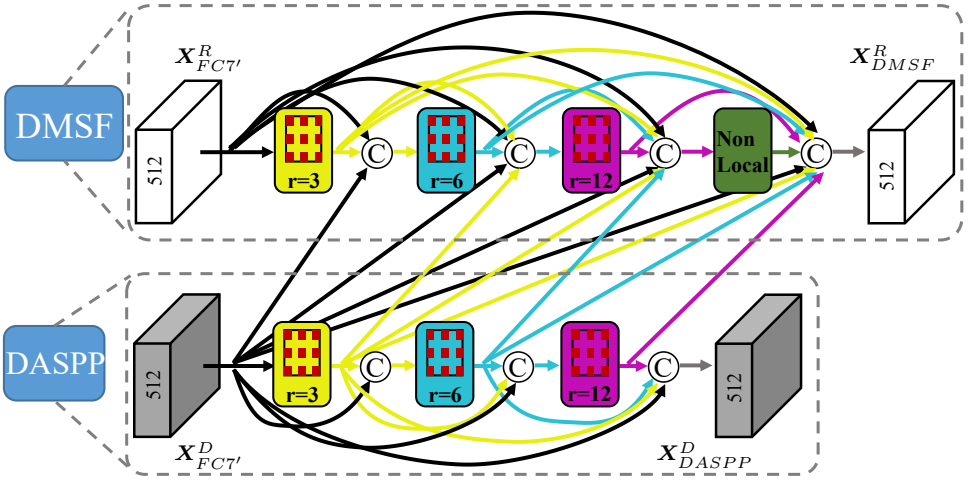


Figure 4.3 : Network architecture of the DASPP model and the proposed DMSF model.

For each decoder branch, we first use a Conv layer to reduce the channel number of the FC7 feature map to 512 channels. We denote the two feature maps by $\boldsymbol{X}^R_{FC7'}$ for the RGB saliency branch and $\boldsymbol{X}^D_{FC7'}$ for the depth branch. Then, we extract and fuse multiscale features based on the DASPP [122] model. Specifically, we directly use DASPP for the depth feature $\boldsymbol{X}^D_{FC7'}$. DASPP deploys several atrous Conv layers on the input feature map with increasing rates, thus obtaining multiscale features with different receptive fields. Meanwhile, it also introduces dense connections among the multiscale atrous layers, connecting each layer to all subsequent layers with larger rates, thus covering scale ranges densely.

As shown in Figure 4.3, we use three atrous layers with rates $r = \{3, 6, 12\}$. In each layer $i \in \{0, 1, 2\}$, we first concatenate all previous features. Then, a 1×1 Conv layer is used to reduce the channel number to 512. Finally, we use an atrous Conv layer with rate r_i to generate the atrous feature $\boldsymbol{X}^D_i$ with a large receptive field. We follow the work in [122] to set the channel number of each atrous Conv layer to $\lfloor \frac{512}{3} \rfloor$ for reducing the computational cost. The whole process can be written as:

$$\boldsymbol{X}^D_i = \mathbb{AC}_{r_i}(\mathbb{C}([\boldsymbol{X}^D_{FC7'} | \boldsymbol{X}^D_0 | \cdots | \boldsymbol{X}^D_{i-1}])), \tag{4.1}$$

where $\mathbb{C}$ means a Conv layer, $\mathbb{AC}_{r_i}$ means an atrous Conv layer with rate r_i, and $[\cdot]$ denotes the concatenation operation.

Finally, the three multiscale features and the original feature are concatenated to form the final depth DASPP feature via a 1×1 Conv layer with 512 channels as:

$$\boldsymbol{X}^D_{DASPP} = \mathbb{C}([\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]). \tag{4.2}$$

Nevertheless, DASPP is only designed for a single modality. To adapt it to multimodal features, we propose a novel dense multiscale fusion (DMSF) model by extending DASPP with dense fusion connections to fuse cross modality features. In our case, we densely fuse the depth multiscale features with the RGB ones. Specifically, we deploy the same three atrous Conv layers on the input RGB feature map. At the same time, we not only densely connect each RGB atrous feature $\boldsymbol{X}^R_i$ to all subsequent atrous layers, but also densely connect each depth atrous feature $\boldsymbol{X}^D_i$ to all RGB atrous layers with larger rates as:

$$\begin{aligned}\boldsymbol{X}^R_i = \mathbb{AC}_{r_i}(\mathbb{C}([&\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\cdots|\boldsymbol{X}^R_{i-1}| \\ &\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\cdots|\boldsymbol{X}^D_{i-1}])).\end{aligned} \tag{4.3}$$

Furthermore, since DASPP only uses atrous Conv layers to construct multiscale features, these features all have local contexts (since the atrous Conv operation is a local operation). Based on the basic idea of the DASPP model to construct multi-scale features with small to large scales, we propose to incorporate global contexts at the end of DMSF as the largest scale. Specifically, we adopt the non-Local network [108] as the global context model since its effectiveness has been widely verified. For the input feature map $\boldsymbol{X} \in \mathbb{R}^{W \times H \times C}$, the non-local network computes three feature embeddings $\theta(\boldsymbol{X}), \phi(\boldsymbol{X}), g(\boldsymbol{X}) \in \mathbb{R}^{WH \times C'}$ first. Then, it uses $\theta(\boldsymbol{X})$ and $\phi(\boldsymbol{X})$ to compute an global attention matrix with size $WH \times WH$, which can be used to propagate global contexts from $g(\boldsymbol{X})$. Finally, the global contexts are transformed to C'' channels by a Conv layer.

The whole model can be formulated as:

$$\mathbb{NL}(\boldsymbol{X}) = \mathbb{C}(softmax(\theta(\boldsymbol{X})\phi(\boldsymbol{X})^{\top})g(\boldsymbol{X})), \tag{4.4}$$

where $softmax$ operates on each row of the attention matrix.

As shown in Figure 4.3, we adopt the non-local network as the fourth multiscale layer in DMSF. It takes all previous RGB multiscale features and depth ones as input and propagates global contexts for them. We set the number of channels C' to 512 and the number of C'' to $\lfloor \frac{512}{3} \rfloor$ to keep them consistent with those in the previous layers. The obtained feature is:

$$\boldsymbol{X}^R_{NL} = \mathbb{NL}([\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\boldsymbol{X}^R_1|\boldsymbol{X}^R_2|\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]). \tag{4.5}$$

Finally, similar to DASPP, all previous RGB and depth layers are concatenated and a 1×1 Conv layer is used to obtain the final DMSF feature with 512 channels as:

$$\begin{aligned} \boldsymbol{X}^R_{DMSF} = \mathbb{C}([&\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\boldsymbol{X}^R_1|\boldsymbol{X}^R_2|\boldsymbol{X}^R_{NL}| \\ &\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]). \end{aligned} \tag{4.6}$$

The final features $\boldsymbol{X}^D_{DASPP}$ $\boldsymbol{X}^R_{DMSF}$ extract and fuse multiscale features from the two FC7 features, thus supplying good starting points for the depth prediction branch and the saliency detection branch.

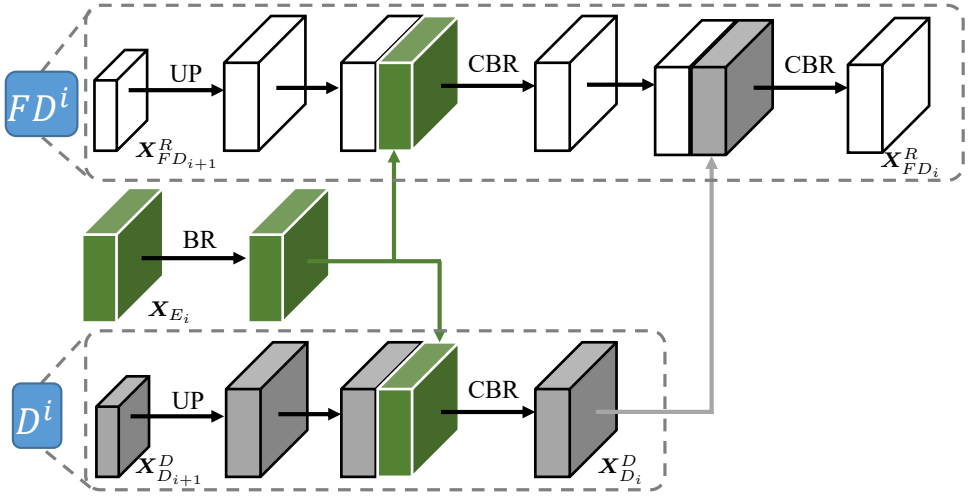


Figure 4.4 : Network architecture of the decoding module and the fusion decoding module. "BR" means BN [36] and ReLU, "CBR" means Conv, BN and ReLU. "UP" means upsampling.

Inspired by U-Net, we design five decoding modules for each decoder branch to progressively fuse multi-level encoding features via skip-connections. At the same time, we fuse each depth decoding feature with the corresponding RGB one to enhance it for saliency detection. We name the decoding modules in the depth branch D^i, where i is in inverse order from 5 to 1, as shown in Figure 4.2. For the RGB branch, we name them fusion decoding modules FD^i since they simultaneously fuse encoding features and depth decoding ones. The encoding feature of the i^{th} Conv block is denoted by $\boldsymbol{X}_{E_i}$, which is the last Conv feature before using the ReLU activation function.

As shown in Figure 4.4, in the depth branch, each decoding module D^i fuses $\boldsymbol{X}_{E_i}$ with the previous depth decoding feature $\boldsymbol{X}^D_{D_{i+1}}$ to obtain the current depth decoding feature $\boldsymbol{X}^D_{D_i}$:

$$\boldsymbol{X}^D_{D_i} = \mathbb{CBR}([\mathbb{BR}(\boldsymbol{X}_{E_i})|\mathbb{UP}(\boldsymbol{X}^D_{D_{i+1}})]), \tag{4.7}$$

where $\mathbb{BR}$ means batch normalization (BN) [36] and ReLU,

$\mathbb{UP}$ denotes upsampling since $\boldsymbol{X}^D_{D_{i+1}}$ is smaller than $\boldsymbol{X}_{E_i}$ for $i \leq 3$, $\mathbb{CBR}$ means 1×1 Conv, BN, and ReLU, and the channel numbers of $\boldsymbol{X}^D_{D_i}$ are set to 512, 256, 128, 64, 64 for $i \in \{5, 4, 3, 2, 1\}$, respectively. Here, BN is used to normalize $\boldsymbol{X}_{E_i}$ for making it compatible with $\boldsymbol{X}^D_{D_{i+1}}$. We use $\boldsymbol{X}^D_{DASPP}$ as the first decoding feature $\boldsymbol{X}^D_{D_6}$.

In the RGB saliency branch, each fusion decoding module FD^i fuses $\boldsymbol{X}_{E_i}$ with the previous RGB decoding feature $\boldsymbol{X}^R_{FD_{i+1}}$ to obtain the current RGB decoding feature $\boldsymbol{X}^R_{D_i}$. Then, $\boldsymbol{X}^D_{D_i}$ is also fused to obtain $\boldsymbol{X}^R_{FD_i}$:

$$\begin{aligned} \boldsymbol{X}^R_{D_i} &= \mathbb{CBR}([\mathbb{BR}(\boldsymbol{X}_{E_i})|\mathbb{UP}(\boldsymbol{X}^R_{FD_{i+1}})]), \\ \boldsymbol{X}^R_{FD_i} &= \mathbb{CBR}([\boldsymbol{X}^R_{D_i}|\boldsymbol{X}^D_{D_i}]), \end{aligned} \tag{4.8}$$

where $\boldsymbol{X}^R_{DMSF}$ is used as $\boldsymbol{X}^R_{FD_6}$.

Finally, we directly use a 1×1 Conv layer with 1 channel on $\boldsymbol{X}^D_{D_1}$ to obtain the predicted depth map. The same Conv layer with the Sigmoid activation function is also used on $\boldsymbol{X}^R_{FD_1}$ to predict the saliency map.

4.3.3 Loss Functions

For saliency prediction, we use a simple binary cross entropy loss function. Supposing that we have a predicted saliency map $\bar{S} \in [0, 1]^{W \times H}$ and the corresponding ground truth $S \in \{0, 1\}^{W \times H}$, the saliency loss can be computed by:

$$L_s(\bar{S}, S) = \frac{1}{WH} \sum_{w,h=1}^{W,H} (S_{wh} log \bar{S}_{wh} + (1 - S_{wh}) log(1 - \bar{S}_{wh})). \tag{4.9}$$

To ease the network training, we also predict a saliency map from $\boldsymbol{X}^R_{FD_i}$ and adopt the supervision with the saliency loss for each fusion decoding module.

For depth prediction, we adopt the depth ranking loss in [114], which optimizes the ordinal relation of each pair of pixels. First, for each predicted depth map Z

and the corresponding ground truth depth map G, we sample N point pairs. For pair k, we denote the pair of points as (i_k, j_k), where i_k and j_k are the coordinates of the two points. Its ordinal relation label ℓ_k can be defined by:

$$\ell_k = \begin{cases} +1, & \frac{G_{ik}}{G_{jk}} > 1 + \delta, \\ -1, & \frac{G_{jk}}{G_{ik}} > 1 + \delta, \\ 0, & otherwise, \end{cases} \tag{4.10}$$

where δ is an empirical threshold.

We follow the work in [114] to set $N = 3000$ and $\delta = 0.02$. Then, the depth ranking loss is defined by:

$$L_r(Z, G) = \frac{1}{\sum \omega_k} \sum_{k=1}^{N} \omega_k \psi(i_k, j_k, \ell_k, Z), \tag{4.11}$$

where

$$\psi = \begin{cases} log(1 + exp[(-Z_{ik} + Z_{jk})\ell_k]), & \ell_k \neq 0 \\ (Z_{ik} - Z_{jk})^2, & \ell_k = 0, \end{cases} \tag{4.12}$$

and $\omega_k \in \{0, 1\}$ is the loss weight for pair k. We follow [114] to sort the losses ψ for all training pairs at each iteration, and set the pairs with the smallest 25% losses to have $\omega_k = 0$. In this way, we can increase the ratio of equal pairs and avoid keeping optimizing pairs with large difference.

However, using the ranking loss will lead to slow convergence. Thus, we also adopt a normalized ℓ_2 loss between Z and G. Specifically, it is the ℓ_2 loss between the normalized Z and normalized G:

$$L_n(Z, G) = \frac{1}{WH} \sum_{w,h=1}^{W,H} (\frac{Z_{wh} - \mu_Z}{\sqrt{\sigma_Z}} - \frac{G_{wh} - \mu_G}{\sqrt{\sigma_G}})^2, \tag{4.13}$$

where μ_* and σ_* are the mean and variance, respectively.

For each decoding module in the depth decoder, we predict a depth map from $\boldsymbol{X}_{D_i}^D$ and use this loss to accelerate the network training.

4.4 Experiments

In this section, we report the experimental results on seven RGB-D saliency benchmark datasets to validate the effectiveness of our proposed model.

4.4.1 Datasets and Evaluation Metrics

We conduct experiments on seven widely used RGB-D saliency detection datasets and four evaluation metrics.

The first dataset is the **NJUD** [42] dataset with 1985 stereo images. The second and the third ones are the **NLPR** [80] and the **RGBD135** [12] dataset with 1000 and 135 RGB-D image pairs, respectively. These two datasets are both collected by using Microsoft Kinect. The fourth dataset is the **SSD** [139] dataset with 80 stereo images collected from movies. The fifth dataset is **DUT-RGBD**, which contains 800 training images and 400 testing images with real life scenarios. The sixth dataset is **STEREO** with 1,000 stereo images collected from the Internet. The last one is the **LFSD** dataset. It has 100 images captured by a light field camera.

As for the evaluation metrics, the first one is the maximum F-measure (maxF) score. By binarizing the predicted saliency map with a threshold, we can compare it with the corresponding ground truth saliency map and obtain the precision and recall. Then, the F-measure score can be computed by:

$$F_m = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{4.14}$$

where we follow the previous work and set $\beta^2 = 0.3$ for emphasizing more on precision. By varying the threshold, we can find the maximum F-measure score. The second metric is the Mean Absolute Error (MAE). It computes the absolute difference between the predicted saliency map $\bar{S}$ and the ground truth S by:

$$MAE = \frac{1}{WH}\sum_{w,h=1}^{W,H} |\bar{S}_{wh} - S_{wh}|. \tag{4.15}$$

The above two metrics both evaluate individual pixels, separately, without considering high-level statistics. Thus, we also adopt the structure-measure S_m [20] score as the third metric. It computes and combines a region-aware structural similarity S_r and an object-aware one S_o between each saliency map and the ground truth by:

$$S_m = \alpha * S_o + (1 - \alpha) * S_r, \tag{4.16}$$

where α is set to 0.5 following the advice from [20].

The last metric is the recently proposed enhanced-alignment measure E_ξ [21]. It considers both global statistics and local pixel matching information. We use it for a more comprehensive evaluation. Our designed algorithm performance is displayed in Table 5.1.

4.4.2 Implementation Details

We train our model in two stages using the stochastic gradient descent (SGD) algorithm. In the first stage, we leverage a depth estimation dataset, i.e., ReDWeb [114], and a saliency detection dataset, i.e., DUTS [97], to pretrain the model iteratively. The ReDWeb dataset contains 3600 stereo images collected from the Internet with various scenes, such as street, office, hill, park, etc. The DUTS dataset consists of 10553 training images collected from the ImageNet DET's training and validation sets [15] with human-annotated saliency maps. We first initialize the encoder part using the VGG-16 parameters pretrained on Imagenet and randomly initialize the two decoder branches. For each iteration, we first use the ReDWeb data to train the encoder and the depth decoder branch with the depth losses L_r and L_n, and then use the DUTS data to train the whole network using the saliency loss L_s. We set the batch size and momentum to 10 and 0.9, respectively. The learning rates of the encoder part and the two decoders are set to 0.001 and 0.01, respectively. The whole training step is set to 40000 and we divide the learning rates by 10 at the

20000^{th} and 30000^{th} step.

In the second stage, we follow most previous works [29, 6, 133] to train the whole network using 1400 images of the NJUD dataset and 650 images of the NLPR dataset and using L_s and L_r losses. Since many depth maps of the NJUD dataset are very noisy, we do not use a deep supervision with the L_n losses for the depth branch. We initialize the network parameters from the pretrained model in the first stage. The learning rates of the encoder part and the depth decoder are set to 0.0001 and the learning rate of the saliency decoder is set to 0.001 to fine-tune the network. Other training settings are set to the same as those in the first stage.

We use the scale of 224×224 to train and test the network. Specifically, for training, we first resize each RGB-D image pair to a random size from 224×224 to 272×272 and then randomly crop a 224×224 patch from it for network training. Random horizontal-flipping is also used for data augmentation. For testing, we directly resize each RGB-D image pair to 224×224 as the network input and then obtain the saliency map from the last layer of the RGB saliency branch. Each image is also pre-processed by subtracting the mean pixel value. The whole model is implemented using Pytorch [79]. A GTX 1080Ti GPU is used for acceleration and the inference time for each testing image is only 0.019 seconds.

4.4.3 Comparison with State-of-the-art Models

We compare our method with nine state-of-the-art saliency detection models, i.e., Amulet [132], DSS [33], BMP [129], PiCANet [62], R3Net [16], CPD [113], EGNet [134], MINet [78], and ITSD [136]. We also include 12 state-of-the-art RGB-D saliency detection models for comparison, including DF [84], AFNet [98], CTMF [29], MMCI [7], PCF [6], TANet [10], CPFP [133], DMRA [81], SSF [131], UCNet [128], JLDCF [26], and A2dele [82]. Note that all these models are deep learning based models and the last four were published in 2020. Since our model uses the

VGG-16 network as the backbone, we use the results of these models with the same backbone for fair comparisons.

The quantitative comparison results are shown in Table 4.1 and visual comparisons are shown in Figure 4.6. We can see that our model outperforms the state-of-the-art RGB saliency models on five out of the seven datasets. The comparison in terms of PR curves on four datasets are also given in Figure 4.5. These results demonstrate the effectiveness of the depth features that we use and the importance of introducing depth information to deep saliency detection on these datasets.

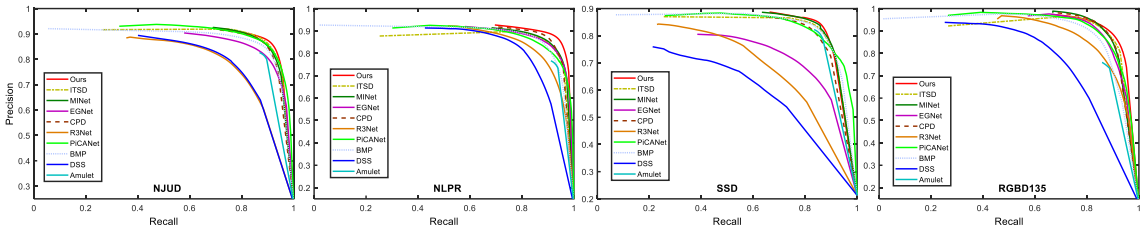


Figure 4.5 : Comparison with state-of-the-art RGB saliency models in terms of PR curves on four datasets. The compared models are Amulet [132], DSS [33], BMP [129], PiCANet [62], R3Net [16], CPD [113], EGNet [134], MINet [78], and ITSD [136].

Table 4.1 : Quantitative comparison between our proposed model and state-of-the-art RGB and RGB-D salient object detection models. We compare our model with nine state-of-the-art (SOTA) CNN-based RGB saliency models and twelve SOTA deep learning based RGB-D saliency models on seven datasets in terms of four evaluation metrics. "Train w D" means training with depth while "Test w D" means test with depth. The number in **bold** indicates the best performance in each group (i.e., RGB and RGB-D). The number in *italic* indicates the cases where our model outperforms RGB SOTA models, while * indicates the cases where our model outperforms the A2dele model. "-" means the results are unavailable since the authors did not release them

	Train	Test	NJUD [42]				NLPR [80]				SSD [139]				RGBD135 [12]			
	w D	w D	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
RGB Saliency Detection Models																		
Amulet [132]	✗	✗	0.827	0.819	0.879	0.079	0.838	0.779	0.885	0.055	0.822	0.808	0.876	0.077	0.823	0.761	0.872	0.065
DSS [33]	✗	✗	0.767	0.762	0.83	0.116	0.832	0.797	0.892	0.057	0.634	0.589	0.729	0.167	0.752	0.741	0.871	0.076
BMP [129]	✗	✗	0.860	0.850	0.905	0.068	0.880	0.848	0.917	0.045	0.849	0.817	0.895	**0.060**	0.878	0.854	0.920	0.040
PiCANet [62]	✗	✗	0.872	0.860	0.910	0.068	0.871	0.830	0.900	0.054	0.846	0.810	0.889	0.069	0.890	0.866	0.922	0.039
R3Net [16]	✗	✗	0.770	0.752	0.827	0.116	0.846	0.812	0.899	0.056	0.679	0.656	0.773	0.148	0.855	0.814	0.911	0.052
CPD [113]	✗	✗	0.863	0.858	0.905	0.060	**0.893**	**0.866**	**0.925**	**0.034**	0.833	0.804	0.878	0.067	**0.896**	**0.882**	**0.932**	**0.028**
EGNet [134]	✗	✗	0.840	0.826	0.883	0.079	0.880	0.847	0.917	0.045	0.740	0.701	0.802	0.126	0.888	0.872	0.919	0.036
MINet [78]	✗	✗	0.870	0.859	0.906	**0.057**	0.886	0.854	0.914	0.041	**0.856**	0.827	0.902	**0.054**	0.894	0.880	0.924	0.029
ITSD [136]	✗	✗	**0.873**	**0.867**	**0.911**	**0.057**	0.884	0.850	0.919	0.039	0.850	**0.829**	**0.904**	0.057	**0.896**	0.879	0.930	0.031
A2dele [82]	✓	✗	0.871	0.874	0.916	0.051	0.898	0.882	0.944	0.029	0.802	0.776	0.862	0.070	0.886	0.872	0.920	0.029
Ours	✓	✗	*0.886**	*0.876**	*0.927**	*0.050**	*0.906**	0.882	0.936	0.038	*0.861**	*0.832**	*0.917**	*0.049**	*0.906**	*0.886**	*0.943**	*0.027**

Table 4.1 (continued): Quantitative comparison between our proposed model and state-of-the-art RGB and RGB-D salient object detection models. We compare our model with nine state-of-the-art (SOTA) CNN based RGB saliency models and twelve SOTA deep learning based RGB-D saliency models on seven datasets in terms of four evaluation metrics. "Train w D" means training with depth while "Test w D" means test with depth. The number in **bold** indicates the best performance in each group (i.e., RGB and RGB-D). The number in *italic* indicates the cases where our model outperforms RGB SOTA models, while * indicates the cases where our model outperforms the A2dele model. "-" means the results are unavailable since the authors did not release them

	Train	Test	NJUD [42]				NLPR [80]				SSD [139]				RGBD135 [12]			
	w D	w D	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
RGB-D Saliency Detection Models																		
DF [84]	✓	✓	0.763	0.804	0.864	0.141	0.802	0.778	0.880	0.085	0.747	0.735	0.828	0.142	0.752	0.766	0.870	0.093
AFNet [98]	✓	✓	0.772	0.775	0.853	0.100	0.799	0.771	0.879	0.058	0.714	0.687	0.807	0.118	0.770	0.729	0.881	0.068
CTMF [29]	✓	✓	0.849	0.845	0.913	0.085	0.860	0.825	0.929	0.056	0.776	0.729	0.865	0.099	0.863	0.844	0.932	0.055
MMCI [7]	✓	✓	0.858	0.852	0.915	0.079	0.856	0.815	0.913	0.059	0.813	0.781	0.882	0.082	0.848	0.822	0.928	0.065
PCF [6]	✓	✓	0.877	0.872	0.924	0.059	0.874	0.841	0.925	0.044	0.841	0.807	0.894	0.062	0.842	0.804	0.893	0.049
TANet [10]	✓	✓	0.878	0.874	0.925	0.060	0.886	0.863	0.941	0.041	0.839	0.810	0.897	0.063	0.858	0.827	0.910	0.046
CPFP [133]	✓	✓	0.878	0.877	0.923	0.053	0.888	0.867	0.932	0.036	0.807	0.766	0.852	0.082	0.872	0.846	0.923	0.038
DMRA [81]	✓	✓	0.886	0.886	0.927	0.051	0.899	0.879	0.947	0.031	0.857	0.844	0.906	0.058	0.900	0.888	0.943	0.030
SSF [131]	✓	✓	**0.899**	0.896	0.935	**0.043**	0.914	0.896	0.953	0.026	0.790	0.762	0.867	0.084	0.904	0.884	0.941	0.026
UCNet [128]	✓	✓	0.897	0.895	0.936	**0.043**	**0.920**	0.903	0.956	**0.025**	**0.865**	**0.855**	**0.907**	**0.049**	**0.933**	**0.930**	**0.976**	**0.018**
JLDCF [26]	✓	✓	0.897	**0.899**	**0.939**	0.044	**0.920**	**0.907**	**0.959**	0.026	-	-	-	-	0.913	0.905	0.955	0.026

Figure 4.6 : Visual comparison between our model and state-of-the-art RGB and RGBD saliency models. Our model outperforms SOTA RGB saliency models and surprisingly achieve comparable or even better results than SOTA RGB-D saliency models.

Additive experimental results illustrate the universality of the model towards the high light images and low light images.

We separately adopt a CNN-based model to enhance the images from the Ex-DARK dataset[69], which includes over 7000 original low-light images. Then, we conduct ours and other seven different SOD methods on these images after the previous enhancement step. Figure 4.7 shows the CNN-based enhanced results. Visualization performance proves that our method outperforms other methods and

	Train	Test	DUT-RGBD [81]				STEREO [76]				LFSD [51]			
	w D	w D	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
RGB Saliency Detection Models														
Amulet [132]	✗	✗	0.813	0.792	0.875	0.089	0.867	0.854	0.919	0.053	0.804	0.808	0.865	0.100
DSS [33]	✗	✗	0.803	0.776	0.850	0.097	0.794	0.791	0.866	0.094	0.791	0.784	0.837	0.116
BMP [129]	✗	✗	0.855	0.843	0.890	0.069	0.891	0.880	0.931	0.049	0.802	0.790	0.844	0.103
PiCANet [62]	✗	✗	0.878	0.868	0.910	0.070	**0.896**	0.884	**0.932**	0.051	0.824	0.810	0.854	0.106
R3Net [16]	✗	✗	0.819	0.805	0.868	0.113	0.768	0.757	0.831	0.107	0.828	0.818	**0.871**	0.098
CPD [113]	✗	✗	0.875	0.865	0.911	**0.055**	0.893	0.886	0.929	**0.042**	0.822	0.811	0.860	**0.089**
EGNet [134]	✗	✗	0.872	0.853	0.905	0.059	0.859	0.844	0.903	0.063	**0.834**	**0.829****	0.869	0.090
MINet [78]	✗	✗	0.875	0.861	0.900	0.058	0.820	0.842	0.896	0.070	0.813	0.791	0.841	0.096
ITSD [136]	✗	✗	**0.881**	**0.873**	**0.918**	**0.055**	0.894	**0.887**	0.930	0.045	0.811	0.797	0.850	0.095
A2dele [82]	✓	✗	0.885	0.892	0.930	0.042	0.879	0.879	0.928	0.044	0.833	0.832	0.874	0.077
Ours	✓	✗	0.864	0.853	0.902	0.072	*0.899**	*0.887**	*0.933**	0.046	0.827	0.813	0.866	0.092
RGB-D Saliency Detection Models														
DF [84]	✓	✓	0.736	0.740	0.823	0.144	0.757	0.757	0.847	0.141	0.791	0.817	0.865	0.138
AFNet [98]	✓	✓	0.702	0.659	0.796	0.122	0.825	0.823	0.887	0.075	0.738	0.744	0.815	0.133
CTMF [29]	✓	✓	0.831	0.823	0.899	0.097	0.848	0.831	0.912	0.086	0.796	0.791	0.865	0.119
MMCI [7]	✓	✓	0.791	0.767	0.859	0.113	0.873	0.863	0.927	0.068	0.787	0.771	0.839	0.132
PCF [6]	✓	✓	0.801	0.771	0.856	0.100	0.875	0.860	0.925	0.064	0.794	0.779	0.835	0.112
TANet [10]	✓	✓	0.808	0.790	0.861	0.093	0.871	0.861	0.923	0.060	0.801	0.796	0.847	0.111
CPFP [133]	✓	✓	0.818	0.795	0.859	0.076	0.879	0.874	0.925	0.051	0.828	0.826	0.872	0.088
DMRA [81]	✓	✓	0.889	0.898	0.933	0.048	0.834	0.847	0.910	0.066	0.847	0.856	0.900	0.075
SSF [131]	✓	✓	**0.915**	**0.924**	**0.951**	**0.033**	0.837	0.840	0.912	0.065	0.859	**0.867**	0.900	**0.066**
UCNet [128]	✓	✓	0.871	0.866	0.910	0.059	**0.903**	**0.899**	**0.944**	**0.039**	**0.864**	0.864	**0.905**	**0.066**
JLDCF [26]	✓	✓	-	-	-	-	0.894	0.889	0.938	0.046	0.833	0.840	0.877	0.091

preserves the detailed information.

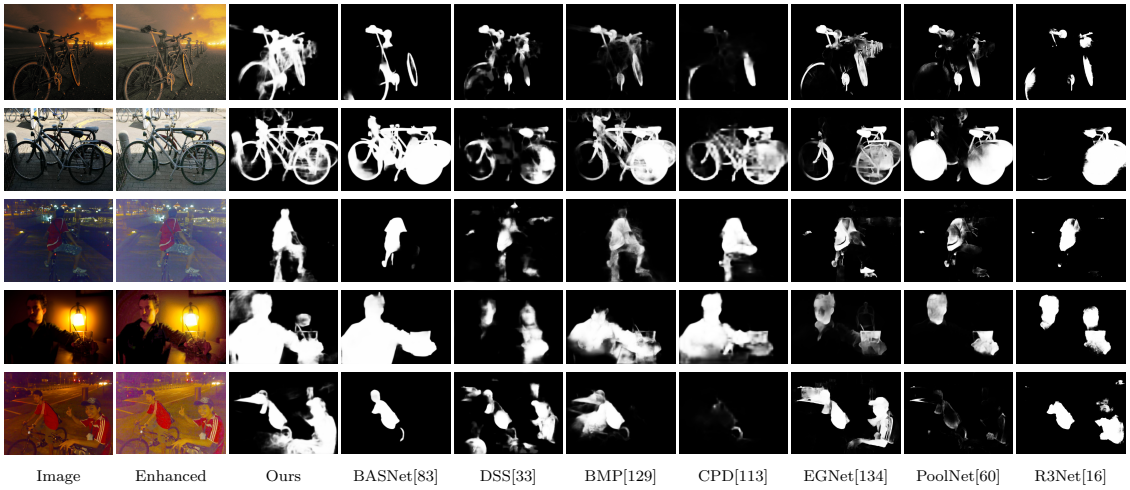


Figure 4.7 : Visualization comparison of different SOD methods conducted on enhanced low light images from ExDARK Dataset [69] with using our designed CNN-based model in chapter 2.

More surprisingly, our model also outperforms the contemporary RGB-D saliency detection methods, i.e., SSF, UCNet, and JLDCF, on four datasets, i.e., NLPR, SSD, RGBD135, and STEREO. These results show the potential of our proposed strategy that only involves depth data in training and without using it in testing for saliency detection. Aiming to achieve this same goal, our model outperforms the contemporary A2dele model on four datasets, i.e., , NJUD [42], SSD [139], RGBD135 [12], and STEREO [76], although A2dele also uses the training set of DUT-RGBD for training. Such a comparison clearly demonstrates the effectiveness of our mechanism in terms of leveraging the depth data.

In Figure 4.6, we show a visual comparison between our model and the state-of-the-art RGB and RGB-D saliency models. We can see that our model not only outperforms RGB saliency models, but also can achieve comparable results compared with RGB-D saliency models. It can leverage depth cues to more accurately localize salient objects and ignore the disturbance from background objects by com-

Figure 4.8 : Visual comparison between "RGB U-Net" and the "+Depth" setting. The GT depth maps and our predicted ones are also given.

paring the depths of the salient objects and other background ones. We can also see that our model can work well on various scenes, such as images with both simple and cluttered backgrounds, cartoon films, both indoor and outdoor scenes, well showing its robustness.

4.4.4 Ablation Study

To understand why our model performs well, we conduct ablation study experiments on four datasets, i.e., NJUD [42], NLPR [80], RGBD135 [12] and SSD [139]. The qualitative results can be found in Table 3.1. Row (a) means that we train the baseline U-Net [85] architecture by only using RGB images of the two datasets. Row (b) means that we add the depth branch and fuse its features with the RGB saliency branch using fusion decoding modules. Row (c) means that we further use DASPP for the depth branch and also use the proposed DMSF module for the RGB saliency branch to fuse the depth DASPP features, but without using the NL model. The last row means that we further use the NL model in DMSF, i.e., our whole network.

From the results, we can see that adding the depth branch and fusing its features

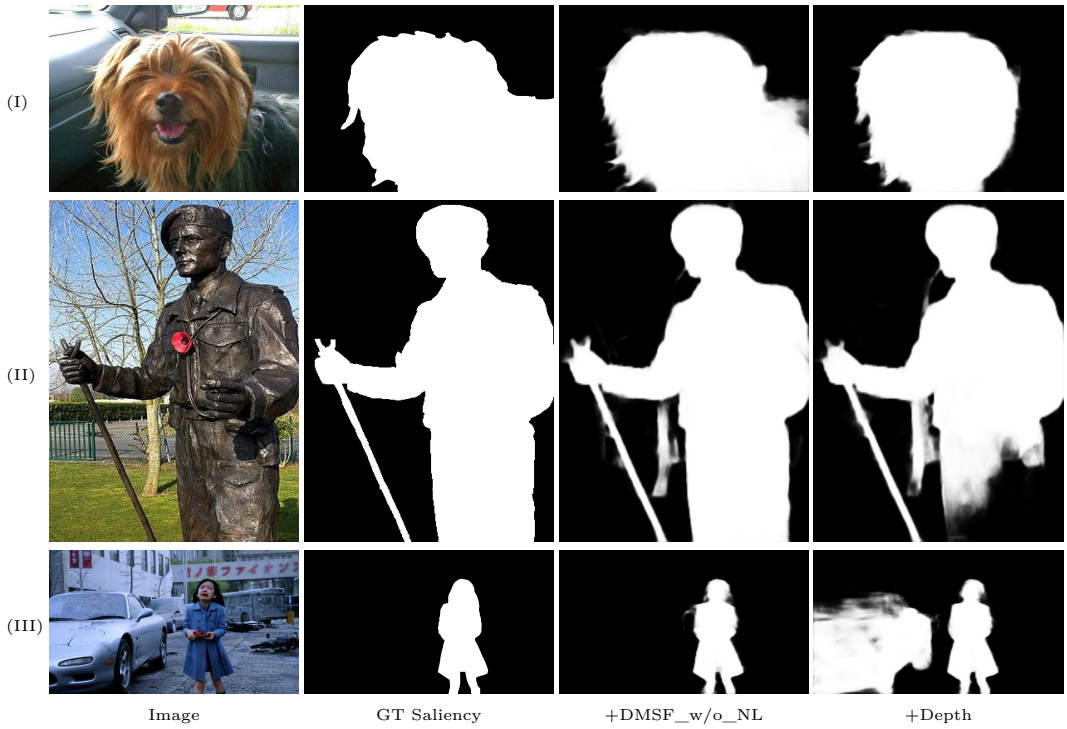


Figure 4.9 : Visual comparison between the "+DMSF_w/o_NL" and the "+Depth" setting.

for saliency detection can largely improve the model performance, especially for the MAE metric. Moreover, using the proposed DMSF model to fuse the multiscale DASPP features can bring further performance gains, especially on the SSD dataset. Finally, we can obtain the best performance on three out of four datasets by adding the NL module in DMSF to further incorporate global contexts. These results clearly demonstrate the effectiveness of our proposed ideas.

We also give qualitative results to show how our proposed model improves performance.

In Figure 4.8, we show the comparison of "RGB U-Net" and the "+Depth" settings. We can see that adding the depth cues can help our saliency model remove the distraction from backgrounds (rows (I) and (II)) or recovering the missing parts of salient objects (rows (III) and (IV)). We also show the GT depth maps and

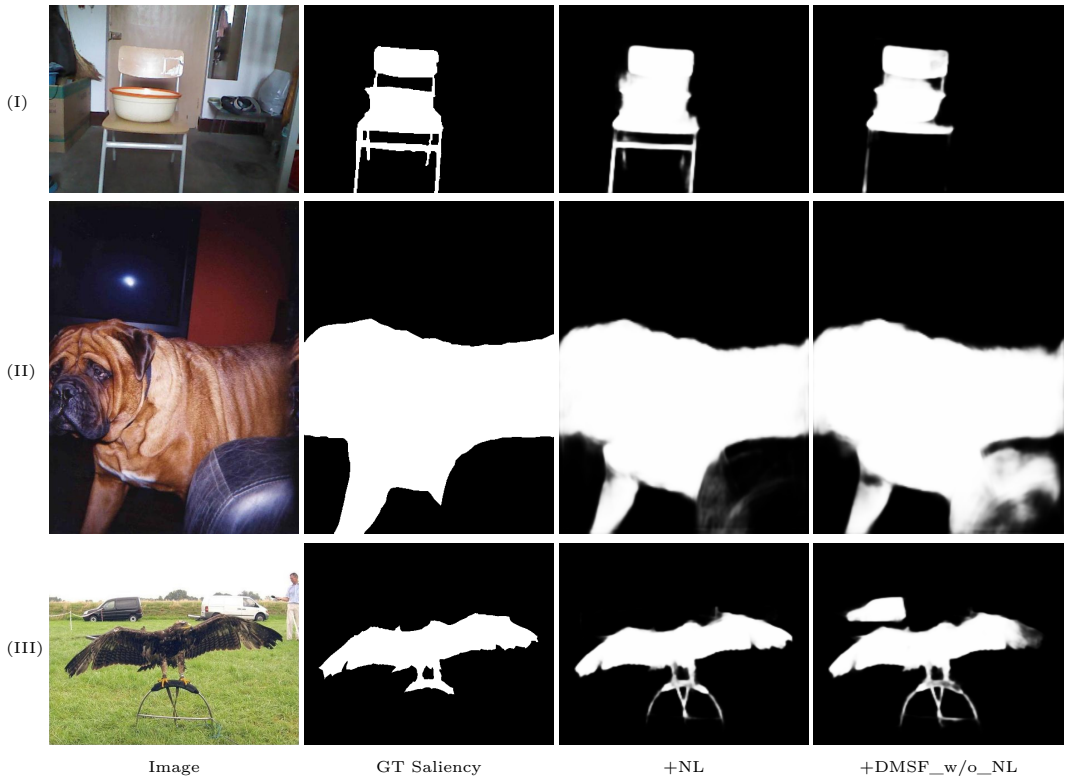


Figure 4.10 : Visual comparison between the "+NL" and the "+DMSF_w/o_NL" setting.

our predicted depth maps in columns (c) and (d). We can see that the depth information supplies complementary cues with effective discrimination. We also find that sometimes the GT depth maps are noisy but our model can estimate more accurate depth (rows (I) and (IV)). This may be the reason why our model can sometimes outperform the state-of-the-art RGB-D saliency models.

We also show the visual improvements of the "+DMSF_w/o_NL" and the "+NL" settings in Figure 4.9 and 4.10, respectively. The comparisons show that using the DMSF model and the NL branch can further help to discriminate and uniformly highlight the salient objects, thus demonstrating their effectiveness.

4.4.5 Discussion

In this section, we discuss whether our model can improve the performance of the RGB saliency detection and also its limitation.

Table 4.2 : Quantitative comparison among our proposed model, baseline RGB U-Net, and state-of-the-art RGB salient object detection models on six RGB saliency datasets. The number in **bold** indicates the best performance in each group.

	SOD [74]				DUT-O [121]				DUTS-TE [97]			
	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
Amulet [132]	0.755	0.808	0.812	0.145	0.781	0.743	0.834	0.098	0.803	0.778	0.851	0.085
DSS [33]	0.741	0.847	0.813	0.128	0.788	0.771	0.845	0.066	0.822	0.825	0.884	0.057
BMP [129]	0.784	**0.856**	0.847	0.112	0.809	0.774	0.848	0.064	0.861	0.851	0.907	0.049
PiCANet [62]	0.787	0.855	0.846	0.108	0.826	**0.794**	0.865	0.068	0.861	0.851	0.915	0.054
R3Net [16]	0.761	0.816	0.835	0.124	0.817	0.760	0.857	0.063	0.835	0.801	0.881	0.057
CPD [113]	0.765	0.853	0.849	0.119	0.818	**0.794**	0.868	0.057	0.866	0.864	0.914	0.043
EGNet [134]	0.807	0.844	**0.873**	0.097	**0.841**	0.777	0.878	**0.053**	**0.887**	0.866	0.927	0.039
MINet [78]	0.805	0.836	0.870	**0.092**	0.833	0.769	0.869	0.056	0.884	0.865	0.927	**0.037**
ITSD [136]	**0.809**	0.844	**0.873**	0.093	0.840	0.792	**0.880**	0.061	0.885	**0.868**	**0.929**	0.041
RGB U-Net	0.786	0.811	0.857	0.099	**0.821**	**0.753**	**0.856**	**0.065**	**0.862**	**0.831**	**0.906**	**0.050**
Ours	**0.795**	**0.814**	**0.867**	**0.094**	0.820	0.747	0.848	0.069	0.861	0.824	0.899	0.052

	ECSSD [120]				HKU-IS [49]				PASCAL-S [54]			
	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
Amulet [132]	0.894	0.915	0.932	0.059	0.883	0.896	0.933	0.052	0.821	0.857	0.862	0.103
DSS [33]	0.882	0.921	0.931	0.052	0.880	0.913	0.938	0.040	0.774	0.849	0.860	0.113
BMP [129]	0.911	0.928	0.944	0.045	0.907	0.921	0.950	0.039	0.831	0.877	0.892	0.086
PiCANet [62]	0.914	0.931	0.953	0.047	0.906	0.921	0.951	0.042	0.837	**0.880**	0.900	0.088
R3Net [16]	0.910	0.926	0.949	0.040	0.895	0.904	0.944	0.036	0.807	0.800	0.853	0.092
CPD [113]	0.910	0.936	0.951	0.040	0.904	0.924	0.950	0.033	0.824	**0.880**	0.891	0.087
EGNet [134]	**0.925**	0.936	0.955	0.037	0.918	0.923	0.956	0.031	0.852	0.841	0.892	0.074
MINet [78]	**0.925**	0.938	0.957	**0.033**	**0.919**	**0.926**	**0.960**	**0.029**	0.856	0.846	0.903	**0.064**
ITSD [136]	**0.925**	**0.939**	**0.959**	0.034	0.917	**0.926**	**0.960**	0.031	**0.859**	0.855	**0.908**	0.066
RGB U-Net	0.911	0.920	**0.946**	**0.044**	0.901	**0.907**	**0.946**	**0.039**	**0.849**	**0.839**	**0.897**	**0.073**
Ours	**0.914**	**0.921**	**0.946**	**0.044**	**0.904**	0.906	0.944	**0.039**	0.847	0.833	0.893	0.076

Model performance on RGB saliency datasets.

Since our model only requires RGB images as input during testing, it naturally raises a question of whether it can improve the performance of RGB saliency detection. To answer this question, we compare our model with state-of-the-art RGB saliency methods, as well as the baseline U-Net model that does not involve depth data in training. The results are given in Table 4.2. We can observe that, compared with the SOTA RGB saliency models, our model shows better results on two datasets, i.e., SOD and ECSSD, but not on all RGB saliency datasets. The possible reasons are two folds. First, RGB SOD has drawn extensive research interests for several years and many models have resorted to various elaborately designed methods to achieve precise saliency detection results, such as attention models, recurrent models, and complementary contour/edge features. In contrast, we only incorporate depth estimation into the U-Net model. Second, the current RGB saliency datasets and RGB-D ones have different data distribution and properties. Depth cues may be more important for the current RGB-D saliency datasets but they do not supply much informative cue for current RGB saliency datasets. Hence, the effectiveness of our proposed model depends on specific scenes. Our proposed model is not suitable to all visual scenes, nor are other SOTA saliency models.

Model limitations. As aforementioned, our model does not bring performance gains for current RGB saliency detection datasets. As for the RGB-D saliency datasets, it also does not outperform the state-of-the-art RGB-D saliency methods. However, it allows inferring RGB-D saliency without requiring depth input.

In Figure 4.11, we also show some failure cases. We find that our model fails mainly in two cases. The first case is when the model fails to predict accurate depth maps, it would produce incorrect saliency maps, as shown in the first two rows in Figure 4.11. The second case is when the model predicts accurate depth maps but may not combine both depth and appearance cues properly, it would also yield incorrect saliency detection results, as shown in the last two rows of this figure.

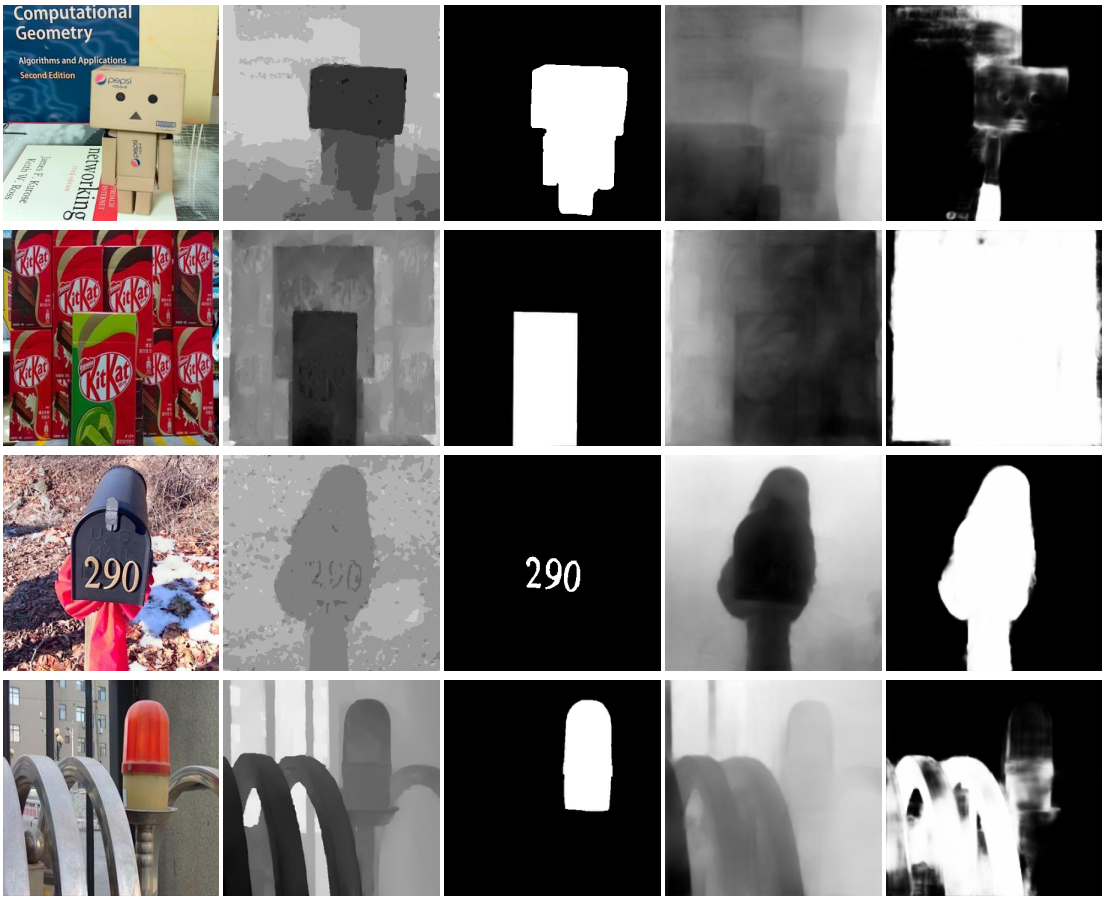


Figure 4.11 : Failure case analysis.

4.5 Summary

Depth information plays a very important role in the visual attention mechanism. However, directly collecting depth data for each image or video is expensive and impractical. In this chapter, we have proposed to simultaneously estimate the depth and detect saliency for RGB images in a unified deep CNN. Intermediate depth features can be fused with RGB saliency features to supply complementary information for improving the saliency detection performance. We have further proposed to fuse multiscale depth and RGB features and also introduced global contexts. Experimental results have clearly demonstrated the effectiveness of our proposed model on high light and low light image datasets, compared with both state-of-the-art RGB

and RGB-D saliency models. We hope our work can inspire further research on leveraging depth cues for RGB saliency detection.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we have presented three algorithms contributing to the development of low-light image enhancement and saliency object detection. The model aiss CNN based models towards the low-light image enhancement, and two saliency detection models utilizing feature aggregation and a no-depth module, respectively. We have provided our insights into some key issues in saliency object detection and discussed the promising solutions based on different methodologies.

In Chapter 1, we have focused on the overview of the whole architecture of the low light enhancement and saliency object detection, thus making an overall introduction for the thesis. In this chapter, we have presented an overview of the main theme, covering the early theoretical investigation in the 2010s, the more recent attempts to develop the techniques in the research area, and the latest breakthroughs of deep learning techniques with low light enhancement.

In Chapter 2, we have conducted research on low light image enhancement technologies. Low light environments and dark conditions are often overlooked in computer vision tasks and research subjects, and the conventional approaches' success in this field is severely limited. Established solutions can result in over-exposure and a halo effect in a low light scene and they need to be improved. We have suggested a broad semantic brightening network based on the cognitive perception paradigm of Retinex theory in this chapter to efficiently integrate the inception network with high-level semantic knowledge about foreground and context. We have

added a framework in order to train a model for this goal. In order to train a model towards this goal, we have proposed to add the structure loss and perceptual loss to boost the enhancement effect by incorporating high-level semantic content. The qualitative and quantitative comparative studies on benchmark databases have demonstrated that our method outperforms the current approaches by addressing the disadvantages of white and color distortion.

In Chapter 3, we have introduced that Deep RGB-D salient target recognition models often use UNet-based architectures, but UNet use only a top-down decoder network to gradually aggregate high-level functions with low-level ones. We have suggested in this chapter to boost the function of aggregation by using holistic aggregation paths and an additional bottom-up decoder network. The former aggregates multi-level features holistically in order to learn abundant function interactions, while the latter aggregates enhanced low-level features with high-level features, thus improving their representation ability. Experimental results have illustrated the universality of the model towards the high light images and low light images.

We have also suggested a factorized focus module for modulating the feature aggregation behavior for each feature node effectively. The findings of experiments on seven commonly used benchmark datasets have shown that all of the proposed components will steadily boost RGB-D salient object detection results. As a result, our final saliency model outperforms other state-of-the-art models.

In Chapter 4, we have suggested that depth information was calculated from monocular RGB images and that intermediate depth features are used to increase saliency detection efficiency in a deep neural network system. The experimental results have also shown the universality of the model towards the high light images and low light images.

To be more precise, we have used an encoder network to extract common features

from each RGB image before building two decoder networks for depth estimation and saliency detection. To increase their capability, the depth decoder features can be paired with the RGB saliency features. In addition, based on the dense ASPP model, we have suggested a novel dense multi-scale fusion model for densely fusing multi-scale depth and RGB functions. To enhance the multi-scale functionality, a new global context branch has been introduced. The experimental results have shown that both of the additional depth cues and the proposed fusion model can increase the saliency detection efficiency. Finally, our model has not only outperformed the cutting-edge RGB saliency models, but also produced similar performance compared to the cutting-edge RGB-D saliency models.

5.2 Future Work

Research on unsupervised depth saliency model.

Most of the night images on the Internet and night image data obtained from surveillance videos are raw data obtained through user upload or device acquisition, and these images do not contain any deterministic annotation information. However, the network training of the existing deep structure models requires a large amount of labeled data. Therefore, in order to make the existing algorithms better extend to those unlabeled night images and make full use of the existing resources to continuously increase the discriminative performance of the model, our future work will further study how to extend the deep learning model to semi-supervised or unsupervised model in learning. On one hand, it is considered to design an effective learning framework to better mine and utilize the information of the unlabeled night data. On the other hand, it is also considered to automatically label the salient areas of the existing night data with undefined labels.

The practical application of the night-time salient target detection model in related fields.

The approaches using night-time images proposed in this thesis are significant. The target detection model has achieved better detection performance than the existing cutting-edge models on some public datasets and our night image dataset. For some applications, such as target detection in night video surveillance, night pedestrian detection, and segmentation, *etc*, research is also expected to see their social effects. The further work of this research will also be devoted to extending the method of night-time salient target detection to more practical applications, promoting the security monitoring system to achieve full real-time monitoring and other related works.

Acquisition and Processing of Low Light and Narrow Dynamic Range Image Data

The collection and processing of a large number of low-light and other types of image data have always been hot research topics. How to select appropriate training data and data paradigms, how to reduce the workload and training time of manual labeling with a better, more stable and more accurate model method, and how to achieve more robust processing results are also research topics that researchers have been focusing on in recent years. In the selection of supervised and unsupervised models, the scope of applications of supervised models has been extended to the processing of more image types, while the problem of unsupervised or semi-supervised models is still under development. The next step is to consider dynamically designing a data augmentation processing framework.

Research on more image scenarios

This thesis mainly studies the scenes of low-light images. The research methodology can work on other image scenarios such as foggy days and it has a very important migration value. The next step is to study a variety of other types of images visual processing tasks.

Bibliography

[1] Z. Al-Ameen and G. Sulong, "A new algorithm for improving the low contrast of computed tomography images using tuned brightness controlled single-scale retinex," *Scanning*, vol. 37, no. 2, pp. 116–125, 2015.

[2] K. A. Bak C and E. E, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, vol. 20(7), pp. 1688–1698, 2017.

[3] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *IEEE International Conference on Computer Vision*, 2011, pp. 914–921.

[4] A. Chaudhry, P. K. Dokania, and P. H. S. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *British Machine Vision Conference*, 2017.

[5] H. Chen, Y.-F. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for rgb-d salient object detection," in *International Conference on Intelligent Robots and Systems*. IEEE, 2018, pp. 6821–6826.

[6] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[7] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection,"

Pattern Recognition, vol. 86, pp. 376–385, 2019.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[9] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *European Conference on Computer Vision*, 2018, pp. 234–250.

[10] Chen, Hao and Li, Youfu, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[11] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[12] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *International Conference on Internet Multimedia Computing and Service*. ACM, 2014, p. 23.

[13] A. Ciptadi, T. Hermans, and J. Rehg, "An in depth view of saliency," in *British Machine Vision Conference*, 2013.

[14] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*. JMLR. org, 2017, pp. 933–941.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[16] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.

[17] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang, "Depth-aware saliency detection using convolutional neural networks," *Journal of Visual Communication and Image Representation*, vol. 61, pp. 1–9, 2019.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[19] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[20] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE International Conference on Computer Vision*, 2017, pp. 4558–4567.

[21] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 698–704.

[22] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, 2020.

[23] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks," *arXiv preprint arXiv:1907.06781*, 2019.

[24] X. Fan, C. Xiang, C. Chen, P. Yang, L. Gong, X. Song, P. Nanda, and X. He, "Buildsensys: Reusing building sensing data for traffic prediction with cross-domain learning," *IEEE Transactions on Mobile Computing*, 2020.

[25] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.

[26] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3052–3062.

[27] GaoYuanyuan, HuHai-Miao, LiBo, and GuoQiang, "Naturalness preserved nonuniform illumination estimation for image enhancement based on retinex," *IEEE Transactions on Multimedia*, 2018.

[28] Guo, Xiaojie and Li, Yu and Ling, Haibin, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.

[29] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[30] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2009–2021, 2013.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern*

Recognition, 2016, pp. 770–778.

[32] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5300–5309.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[35] L. W. Imamoglu N and F. Y, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15(1), pp. 96–105, 2013.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[38] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International conference on Multimedia*, 2014, pp. 675–678.

[39] Jiang, Lincheng and Jing, Yumei and Hu, Shengze and Ge, Bin and Xiao, Weidong, "Deep refinement network for natural low-light image enhancement in symmetric pathways," *Symmetry*, vol. 10, no. 10, p. 491, 2018.

[40] H. Jin, L. Tu, and X. Deng, "Night image enhancement algorithm based on retinex theory," *International Journal of Advancements in Computing Technology*, vol. 3, no. 10, pp. 291–298, 2011.

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

[42] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *International Conference on Image Processing*. IEEE, 2014, pp. 1115–1119.

[43] S. Ko, S. Yu, W. Kang, C. Park, S. Lee, and J. Paik, "Artifact-free low-light video enhancement using temporal similarity and guide map," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6392–6401, 2017.

[44] L. Shen and Zihan Yue and Fan Feng and Quan Chen and S. Liu and J. Ma, "Msr-net: Low-light image enhancement using deep convolutional network," *ArXiv*, vol. abs/1711.02488, 2017.

[45] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *European Conference on Computer Vision*. Springer, 2012, pp. 101–115.

[46] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 965–968.

[47] S. Lee and C. Lee, "Multiscale morphology based illumination normalization with enhanced local textures for face recognition," *Expert Systems with Applications*, vol. 62, pp. 347–357, 2016.

[48] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[49] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[50] J. Li and C. Miao, "The conveyor belt longitudinal tear on-line detection based on improved ssr algorithm," *Optik*, vol. 127, no. 19, pp. 8002–8010, 2016.

[51] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[52] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[53] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *European Conference on Computer Vision*, 2018, pp. 355–370.

[54] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[55] Li, Guanbin and Yu, Yizhou, "Deep contrast learning for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[56] X. Lian, Y. Pang, Y. He, X. Li, and A. Yang, "Learning tone mapping function for dehazing," *Cognitive Computation*, vol. 9, no. 1, pp. 95–114, 2017.

[57] J. Lim, M. Heo, C. Lee, and C.-S. Kim, "Contrast enhancement of noisy low-light images based on structure-texture-noise decomposition," *Journal of Visual Communication and Image Representation*, vol. 45, pp. 107–121, 2017.

[58] W. Z. J. Lin X and M. L, "Saliency detection via multi-scale global cues[j]. ieee transactions on multimedia," *IEEE Transactions on Multimedia*, vol. 21(7), pp. 1646–1659, 2019.

[59] H. Ling, "Cross-modal weighting network for rgb-d salient object detection," 2020.

[60] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[61] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[62] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.

[63] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2010.

[64] Z. Liu, W. Zhang, and P. Zhao, "A cross-modal adaptive gated fusion generative adversarial network for rgb-d salient object detection," *Neurocomputing*, vol. 387, pp. 210–220, 2020.

[65] Liu, Nian and Han, Junwei and Yang, Ming-Hsuan, "Picanet: Pixel-wise contextual attention learning for accurate saliency detection," *IEEE Transactions on Image Processing*, 2020.

[66] Liu, Nian and Zhang, Ni and Han, Junwei, "Learning selective self-mutual attention for rgb-d saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[67] Liu, Zhengyi and Shi, Song and Duan, Quntao and Zhang, Wei and Zhao, Peng, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[68] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, 2019.

[69] ——, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[70] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

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

[72] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6593–6601.

[73] F. Lv, F. Lu, J. Wu, and C. Lim, "Mbllen: Low-light image/video enhancement using cnns." in *BMVC*, 2018, p. 220.

[74] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 49–56.

[75] A. Nikonorov, S. Bibikov, V. Myasnikov, Y. Yuzifovich, and V. Fursov, "Correcting color and hyperspectral images with identification of distortion model," *Pattern Recognition Letters*, vol. 83, pp. 178–187, 2016.

[76] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.

[77] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *International Conference on Pattern Recognition*, vol. 1. IEEE, 2000, pp. 375–378.

[78] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[79] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[80] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection:

A benchmark and algorithms," in *European Conference on Computer Vision*. Springer, 2014, pp. 92–109.

[81] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.

[82] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[83] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[84] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[85] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[86] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "Msr-net: Low-light image enhancement using deep convolutional network," *arXiv preprint arXiv:1711.02488*, 2017.

[87] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2749–2757.

[88] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[89] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *European Conference on Computer Vision*, 2018, pp. 715–731.

[90] M. W. Spratling, "A hierarchical predictive coding model of object recognition in natural images," *Cognitive computation*, vol. 9, no. 2, pp. 151–167, 2017.

[91] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1639–1649, 2015.

[92] X. Sun, H. Liu, S. Wu, Z. Fang, C. Li, and J. Yin, "Low-light image enhancement based on guided image filtering in gradient domain," *International journal of digital multimedia broadcasting*, vol. 2017, 2017.

[93] L. Tao and V. Asari, "Modified luminance based msr for fast and efficient image enhancement," in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.* IEEE, 2003, pp. 174–179.

[94] L. Tao, C. Zhu, J. Song, T. Lu, H. Jia, and X. Xie, "Low-light image enhancement using cnn and bright channel prior," in *2017 IEEE International Conference on Image Processing (ICIP).* IEEE, 2017, pp. 3215–3219.

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017, pp. 5998–6008.

[96] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[97] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[98] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[99] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[100] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[101] W. Wang, B. Li, J. Zheng, S. Xian, and J. Wang, "A fast multi-scale retinex algorithm for color image enhancement," in *2008 International Conference on Wavelet Analysis and Pattern Recognition*, vol. 1. IEEE, 2008, pp. 80–85.

[102] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[103] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5968–5977.

[104] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1711–1720.

[105] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2019.

[106] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[107] W. Wang, C. Wei, W. Yang, and J. Liu, "Gladnet: Low-light enhancement network with global awareness," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 751–755.

[108] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[109] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error measurement to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 1, 2004.

[110] Z.-y. Wang, J. Luo, K.-y. Qin, H.-b. Li, and G. Li, "Model based edge-preserving and guided filter for real-world hazy scenes visibility restoration," *Cognitive Computation*, vol. 9, no. 4, pp. 468–481, 2017.

[111] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic

segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2314–2320, 2017.

[112] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[113] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[114] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 311–320.

[115] J. Xiao, H. Peng, Y. Zhang, C. Tu, and Q. Li, "Fast image enhancement based on color space fusion," *Color Research and Application*, pp. 22–31, 2016.

[116] X. Xiao, Y. Zhou, and Y.-J. Gong, "Rgb-'d'saliency detection with pseudo depth," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2126–2139, 2018.

[117] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[118] X. Xu and J. Wang, "Extended non-local feature for visual saliency detection in low contrast images," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[119] R. Y. Xu M and W. Z, "Saliency detection in face videos: A data-driven approach," *IEEE Transactions on Multimedia*, vol. 20(6), pp. 1335–1349, 2017.

[120] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[121] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[122] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[123] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3015–3022.

[124] Ying, Zhenqiang and Li, Ge and Gao, Wen, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," *arXiv preprint arXiv:1711.00591*, 2017.

[125] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *IEEE International Conference on Computer Vision*.

[126] Yurui Ren and Zhenqiang Ying and Thomas H. Li and G. Li, "Lecarm: Low-light image enhancement using the camera response model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 968–981, 2019.

[127] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced

curriculum learning framework," vol. 127, no. 4, pp. 363–380, 2019.

[128] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.

[129] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[130] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6024–6033.

[131] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481.

[132] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[133] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[134] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8779–8788.

[135] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[136] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.

[137] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang, "Attention-guided rgbd saliency detection using appearance information," *Image and Vision Computing*, vol. 95, p. 103888, 2020.

[138] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *International Conference on Multimedia and Expo*. IEEE, 2019, pp. 199–204.

[139] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *International Conference on Computer Vision Workshops*, 2017, pp. 3008–3014.

[140] H. Zhu and G. Wan, "Local contrast preserving technique for the removal of thin cloud in aerial image," *Optik*, vol. 127, no. 2, pp. 742–747, 2016.